Benign Overfitting in Out-of-Distribution Generalization of Linear Models

Anonymous Author(s)
Affiliation
Address
email

Abstract

Benign overfitting refers to the phenomenon where a over-paramterized model fits the training data perfectly, including noise in the data, but still generalizes well to the unseen test data. While prior work provide a solid theoretical understanding of this phenomenon under the in-distribution setup, modern machine learning often operates in a more challenging Out-of-Distribution (OOD) regime, where the target (test) distribution can be rather different from the source (training) distribution. In this work, we take an initial step towards understanding benign overfitting in the OOD regime by focusing on the basic setup of over-parameterized linear models under covariate shift. We provide non-asymptotic guarantees proving that, when the target covariance satisfies certain structural conditions, benign overfitting occurs in standard ridge regression even under the OOD regime. We identify a number of key quantities relating source and target covariance, which govern the performance of OOD generalization. Our result is sharp, which provably recovers prior in-distribution benign overfitting guarantee (Tsigler & Bartlett, 2023), as well as under-parameterized OOD guarantee (Ge et al., 2024) when specializing to each setup. Moreover, we also present theoretical results for a more general family of target covariance matrix, where standard ridge regression only achieves a slow statistical rate of $\mathcal{O}(1/\sqrt{n})$ for the excess risk, while Principal Component Regression (PCR) is guaranteed to achieve the fast rate $\mathcal{O}(1/n)$, where n is the number of samples.

1 Introduction

In modern machine learning, distribution shift has become a ubiquitous challenge where models trained on a source data distribution are tested on a different target distribution (Zou et al., 2018; Hendrycks & Dietterich, 2019; Guan & Liu, 2021; Koh et al., 2021). Generalization under distribution shift, known as Out-of-Distribution (OOD) generalization, remains a fundamental issue in the practical application of machine learning (Recht et al., 2019; Hendrycks et al., 2021; Miller et al., 2021; Wenzel et al., 2022). While there has been extensive work on the theoretical understanding of OOD generalization, most of it has focused on under-parameterized models (Shimodaira, 2000; Lei et al., 2021; Ge et al., 2024; Zhang et al., 2022). However, over-parameterized models, such as deep neural networks and large language models (LLMs), which have more parameters than training samples, are widely used in modern machine learning. Surprisingly, despite the classic bias-variance tradeoff for under-parameterized models, over-parameterized models tend to overfit the data while still achieving strong in-distribution generalization, a phenomenon known as benign overfitting (Hastie et al., 2022; Shamir, 2023) or harmless interpolation (Muthukumar et al., 2020). Therefore, it is crucial to theoretically understand how benign overfitting shapes OOD generalization in over-parameterized models.

Submitted to the Mathematics of Modern Machine Learning Workshop at NeurIPS 2024. Do not distribute.

It is established in overparameterized models that "benign overfitting" occurs when the data essentially resides on a low-dimensional manifold. The manifold assumption (Belkin & Niyogi, 2003) is widely applicable across image, speech and language data, where although features are embedded in a high-dimensional ambient space, their generation is governed by a few degrees of freedom imposed by physical constraints (Niyogi, 2013). Specifically, the covariance matrix of the data should be characterized by several major directions with large eigenvalues while the remaining directions are high-dimensional but have smaller scale. In this setting, even though the estimator may overfit the noise, it can still capture the signal in the major directions while the noise is dampened in the minor directions. Recent non-asymptotic analyses have provided upper bounds on the excess risk for the minimum-norm interpolant and over-parameterized ridge estimator under this framework (Bartlett et al., 2020; Hastie et al., 2022; Tsigler & Bartlett, 2023).

However, theoretical characterization of OOD generalization in over-parameterized models remains elusive. In this paper, we take an initial step toward characterizing OOD generalization in over-parameterized models under *general* covariate shift, a standard assumption for OOD generalization (Ben-David et al., 2006), where the conditional distribution of the outcome given the covariates remains invariant. We derive the first vanishing, non-asymptotic excess risk bound for ridge regression and minimum-norm interpolation, assuming that the source covariance is dominated by a few major eigenvalues, which satisfies the benign overfitting condition. But we allow the target covariance to be arbitrary. This result contrasts with recent work that either addresses only a restrictive form of covariate shift (Hao et al., 2024; Mallinar et al., 2024) or provides excess risk bounds that asymptotically remain above a constant (Tripuraneni et al., 2021b; Hao et al., 2024).

In summary, our excess risk bound identifies several key quantities that relate the source and target covariance, suggesting that "benign overfitting" occurs when these quantities are well controlled. In such cases, the target distribution data lies on the low-dimensional manifold of the source distribution. Otherwise, ridge regression may incur excess risk, lower bounded by the slow statistical rate of $\mathcal{O}(1/\sqrt{n})$. In contrast, we show that principal component regression (PCR) achieves the fast rate of $\mathcal{O}(1/n)$ in such scenarios.

Our contributions.

1. We provide a sharp, instance-dependent excess risk bound for ridge regression (Theorem 2). Our result applies to any target distribution, requiring only that the source covariance be dominated by a few major eigenvectors and that the minor components are high-dimensional. We show that ridge regression exhibits "benign overfitting," achieving excess risk comparable to the in-distribution case, provided that certain key quantities relating the source and target distributions are bounded. Importantly, this condition requires that the *overall magnitude* of the target covariance along the minor directions scales similarly to, or smaller than, that of the source, but it does not depend on the spectral structure of the target covariance. Our results recover the in-distribution bound from Tsigler & Bartlett (2023) when the source and target match, and also recover the sharp bound from Ge et al. (2024) for under-parameterized linear regression under covariate shift when the minor components vanish.

2. We extend our analysis by examining the scenario where the target distribution has significant components in the minor directions. In this scenario, ridge regression incurs a higher error rate compared to the in-distribution setting, specifically the slow statistical rate of $\mathcal{O}(1/\sqrt{n})$ in some instances (Theorem 4). However, we demonstrate that principal component regression ensures a fast rate of $\mathcal{O}(1/n)$ in these cases, provided that the true signal primarily lies in the major directions of the source (Theorem 5). Additionally, PCR does not rely on the minor directions of the source distribution being high-dimensional, highlighting its advantage over ridge regression in such settings.

1.1 Related work

Over-parameterization. The success of over-parameterized models in machine learning has sparked significant research on their theoretical foundations. Harmless interpolation (Muthukumar et al., 2020) or benign overfitting (Shamir, 2023) describes cases where linear models interpolate noise yet still generalize well. Double descent in prediction error is also observed as the ambient dimension surpasses the number of training samples (Nakkiran, 2019; Xu & Hsu, 2019).

Research in this field can be divided into two categories based on assumptions about the spectral structure of the sample covariance. The first category assumes an almost isotropic sample covariance matrix with a bounded condition number or an isotropic prior distribution of parameters (Belkin et al., 2020). In this case, a limiting covariance spectral structure may emerge when $n \asymp d$ and both tend to infinity, allowing for asymptotic risk bounds (Dobriban & Wager, 2018; Richards et al., 2021). However, ridgeless regression is sub-optimal in this setting unless the signal-to-noise ratio is infinite (Wu & Xu, 2020), and non-asymptotic error bounds are lacking. Our work falls into the second category, focusing on covariance model where a small number of eigenvalues dominate the sample covariance, and the signal is concentrated in the subspace spanned by the leading eigenvectors (Bibas et al., 2019; Chinot & Lerasle, 2022; Hastie et al., 2022). Linear regression can be optimal without regularization under this covariance structure (Kobak et al., 2020), which is of practical interest because ridgeless regression is equivalent as gradient descent from zero initialization (Zhou et al., 2020). Sharp non-asymptotic bounds for variance and bias in ridge regression have been derived (Bartlett et al., 2020; Tsigler & Bartlett, 2023).

Extending the analysis of ridgeless estimators (i.e., minimum norm interpolants), uniform convergence bounds for generalization error have been studied for all interpolants with arbitrary norms. However, uniformly bounding the difference between population and empirical errors generally fails to ensure a consistent predictor (Zhou et al., 2020), necessitating strong assumptions on distributions (Koehler et al., 2021) or hypothesis classes (Negrea et al., 2020). Over-parameterization theory for linear models has also been applied to two-layer neural networks approximated via kernel ridge regression (Liang et al., 2020; Ghorbani et al., 2020, 2021; Bartlett et al., 2021; Mei & Montanari, 2022; Mei et al., 2022; Montanari & Zhong, 2022; Simon et al., 2023), though this lies beyond the scope of the present work.

Out-of-Distribution generalization. Out-of-Distribution generalization is well studied for under-parameterized models, particularly in transfer learning between two distributions, where labeled source data is combined with unlabeled target data to train models. For covariate shift, importance weighting (Cortes et al., 2010; Agapiou et al., 2017) is asymptotically optimal when using density ratio as weights (Shimodaira, 2000). More generally, the theoretical limits of transfer learning are explored through minimax lower bounds for bounded distribution shifts, measured by divergence metrics (Mousavi Kalan et al., 2020; Zhang et al., 2022). A number of algorithms are proposed to achieve matching upper bounds (Lei et al., 2021). However, Ge et al. (2024) shows that even without target data, vanilla MLE (Empirical Risk Minimization, ERM) is minimax optimal for well-specified models under covariate shift, with a sharp $1/n$ excess risk bound based on Fisher information.

Research on over-parameterized models under distribution shift has largely focused on covariate shift in linear regression. Importance weighting for over-parameterized models (Chen et al., 2024) and general sample reweighting offer no advantage over ERM since both converge to the same estimator via gradient descent (Zhai et al., 2022). Consequently, much literature focuses on minimum-norm interpolation as the natural ERM solution. For isotropic signals, Tripuraneni et al. (2021a) prove that over-parameterization improves robustness to covariate shift, deriving an asymptotic generalization bound decreasing with d/n. Under the essentially low-rank covariance model, Hao et al. (2024) derive a non-asymptotic bound for a specific covariate shift where features are translated by a constant but the covariance matrix is preserved. However, a constant excess risk remains in their bound due to estimation variance. Kausik et al. (2024) study a linear model with additive noise on covariates when data strictly lies in a low-dimensional subspace, also showing a non-vanishing bound. Mallinar et al. (2024) investigate minimum-norm interpolation with independent covariates and simultaneously diagonalizable source and target covariance matrices, allowing them to directly extend in-distribution bounds of Bartlett et al. (2020); Tsigler & Bartlett (2023). Still, their estimation bias bound is looser than ours due to a gap compared to Tsigler & Bartlett (2023)'s sharp bound even when the source matches the target. In contrast, our work achieves the first vanishing non-asymptotic error bound for general covariate shift, assuming only finite second moments for the target covariance matrix.

There also exist a line of work that considers non-parametric models under covariate shift (Kpotufe & Martinet, 2018; Hanneke & Kpotufe, 2019; Pathak et al., 2022; Ma et al., 2023), presenting minimax results controlling by a transfer-exponent that measures the similarity between source and target, though this lies beyond the scope of our work.

Principal component regression. Principal component regression (PCR) has been designed as a method of treating multicollinearity problems in high-dimensional linear regression, where the covariates have a latent, low-dimensional representation (Massy, 1965; Jeffers, 1967; Jolliffe, 1982; Jeffers, 1981). PCR has been widely used in statistics (Liu et al., 2003), chemometrics (Næs & Martens, 1988; Sun, 1995; Vigneau et al., 1997; Depczynski et al., 2000; Keithley et al., 2009), construction management (Chan & Park, 2005), environmental science (Kumar & Goyal, 2011; Hidalgo et al., 2000), signal processing (Huang & Yang, 2012) and etc.

Regarding the theory for PCR, Hadi & Ling (1998) give conditions under which PCR will fail. Bair et al. (2006) suggest selecting principal components based on their association with the outcome, and provide corresponding asymptotic consistency results. Xu & Hsu (2019) give asymptotic risk bounds for PCR, under different number of selected components k. They show that the "double descent" behaviour also happens in PCR when k/d grows, where d is the data dimension. Most related to our work, Agarwal et al. (2019) provide non-asymptotic error bounds of PCR, and show that the error will decay as $\mathcal{O}(1/\sqrt{n})$ (n is the sample size) given that all the singular values of the data matrix are of the same order. Agarwal et al. (2020) further improves the rate to $\mathcal{O}(1/n)$. However, the aforementioned two results both consider fixed design with strict low-rank assumptions, therefore not applicable to our setting of OOD-generalization.

2 Covariate shift setup under over-parameterization

2.1 Data with covariate shift

We address the out-of-distribution (OOD) generalization of over-parameterized models under covariate shift, where the covariates, denoted by a random vector $x \in \mathbb{R}^d$, follow different distributions during training and evaluation. Specifically, we assume that the training data is sampled from a source distribution $\mathcal{P}_S$, and the learned model is subsequently applied to data from an unknown target distribution $\mathcal{P}_T$. Let the covariates be zero-mean on the source distribution, and define the covariance matrix as $\Sigma_S := \mathbb{E}_{x\sim\mathcal{P}_S}\left[xx^T\right]$. Since we can always choose an orthonormal basis such that Σ_S becomes diagonal, we express $\Sigma_S = \mathrm{diag}(\lambda_1, \cdots, \lambda_d)$ without loss of generality, where the eigenvalues are arranged in non-increasing order: $\lambda_1 \geq \cdots \geq \lambda_d \geq 0$. Moreover, we assume sub-gaussianity of the source covariates, i.e., $\Sigma_S^{-1/2}x$ is σ-sub-gaussian where the precise definition of sub-gaussian norm is given in section A. We consider a general covariate distribution for the target, assuming only that it has a finite second moment, denoted by $\Sigma_T := \mathbb{E}_{x\sim\mathcal{P}_T}\left[xx^T\right]$, which is not necessarily diagonal.

We consider a linear response model that remains consistent across the source and target distributions. The outcome follows $y = x^T\beta^* + \epsilon$, where $\beta^* \in \mathbb{R}^d$ represents the true parameter, and ϵ is an independent noise with zero-mean and variance v^2.

2.2 Learning procedure and evaluation

The learning procedure involves training a linear model with n i.i.d. samples $\{(x_i, y_i)\}_{i=1}^n$ drawn from the source distribution. Define $X := (x_1, ..., x_n)^T \in \mathbb{R}^{n\times d}$, $Y := (y_1, ..., y_n)^T$ and $\boldsymbol{\epsilon} := (\epsilon_1, ..., \epsilon_n)^T$. We focus on models $\widehat{\beta}(Y)$ that are linear in Y, allowing us to write $\widehat{\beta}(Y) = \widehat{\beta}(X\beta^\star) + \widehat{\beta}(\boldsymbol{\epsilon})$. We consider ridge regression and principal component regression as two instances of such algorithms. With a regularization coefficient $\lambda \geq 0$, the ridge estimator in the over-parameterized setting, where $n < d$, is defined as:

$$\widehat{\beta}(Y) = X^T(XX^T + \lambda I_n)^{-1}Y.$$

The algorithm is assessed on the target distribution by its excess risk relative to the true model, expressed as the following equation:

$$\mathcal{R}\big(\widehat{\beta}(Y)\big) := \mathbb{E}_{(x,y)\sim\mathcal{P}_T}\left[\big(y - x^T\widehat{\beta}(Y)\big)^2 - \big(y - x^T\beta^*\big)^2\right] = \big\|\widehat{\beta}(Y) - \beta^*\big\|_{\Sigma_T}^2,$$

where we define $\|x\|_A := \sqrt{x^TAx}$ for any positive semi-definite matrix A. The metric of interest is the expected excess risk with respect to the noise, given by $\mathbb{E}_{\boldsymbol{\epsilon}}\left[\mathcal{R}\big(\widehat{\beta}(Y)\big)\right]$. Following from the linearity of the model, the expected excess risk can be decomposed into bias and variance components:

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\mathcal{R}\big(\widehat{\beta}(Y)\big)\right] = \mathbb{E}_{\boldsymbol{\epsilon}}\big\|\widehat{\beta}(\boldsymbol{\epsilon})\big\|_{\Sigma_T}^2 + \big\|\widehat{\beta}(X\beta^\star) - \beta^\star\big\|_{\Sigma_T}^2,$$

where we define the variance as $V := \mathbb{E}_{\boldsymbol{\epsilon}} \left\| \widehat{\beta}(\boldsymbol{\epsilon}) \right\|_{\Sigma_T}^2$ and the bias as $B := \left\| \widehat{\beta}(X\beta^\star) - \beta^\star \right\|_{\Sigma_T}^2$.

2.3 The structure of covariance in benign overfitting

Throughout this paper, we follow the convention of Tsigler & Bartlett (2023), consider the *source* covariance matrix Σ_S that has only a few number of high variance directions but a very large number of low variance directions with similar magnitude. We will also refer to those high variance directions of the source as "major directions", and those low variance directions as "minor directions". We denote the number of major directions as k. For remaining $d-k$ minor directions, we use the following notions of effective ranks to approximately capture the number of directions that have a similar scale. Let the ridge regularization coefficient be $\lambda \geq 0$, we define:

$$r_k := \frac{\lambda + \sum_{j>k} \lambda_j}{\lambda_{k+1}}, \quad R_k := \frac{\left(\lambda + \sum_{j>k} \lambda_j\right)^2}{\sum_{j>k} \lambda_j^2}.$$

We have $1 \leq r_k \leq R_k$. When $\lambda = 0$, we further have $R_k \leq d-k$. We denote the first k columns of X as X_k and the remaining $d-k$ columns as X_{-k}. Correspondingly, we partion $\beta^\star$ into $\beta^\star_k$ and $\beta^\star_{-k}$. The covariance matrix blocks along the diagonals are denoted by $\Sigma_{S,k}$, $\Sigma_{S,-k}$, $\Sigma_{T,k}$ and $\Sigma_{T,-k}$. To facilitate our presentation, we define

$$\mathcal{T} = \Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}, \quad \mathcal{U} = \Sigma_{S,-k}\Sigma_{T,-k}, \quad \mathcal{V} = \Sigma_{S,-k}^2. \tag{1}$$

These quantities turn out to be crucial in the analysis.

3 Over-parameterized ridge regression

In the context of in-distribution generalization for overparameterized linear models, Bartlett et al. (2020) and Tsigler & Bartlett (2023) demonstrate that the ridge estimator (minimum-norm interpolate estimator as a special case) can effectively learn the signal from the subspace of data spanned by the major eigenvectors, while benignly overfitting noise from the minor directions under certain scenarios. They argue that, when the true signal mainly lies in the major directions, and the minor directions have small scale but high effective rank, benign overfitting is possible. In this section, we explore whether this mechanism still holds under covariate shift. We derive upper bounds (Theorem 2) for the excess risk of the ridge estimator in the context of overparameterized OOD-generation, demonstrating that "benign overfitting" also happens under covariate shift, given that the target distribution's covariance structure remains dominated by the first k dimensions. To be specific, we show that $\mathcal{T}$ characterizes the shift in the major directions; the *overall magnitude* of $\Sigma_{T,-k}$, which characterizes the shift in the minor directions, is crucial for benign overfitting. When the the *overall magnitude* of $\Sigma_{T,-k}$, scales similarly to or smaller than those of the source, ridge regression achieves the same non-asymptotic error rate under covariate shift as in the in-distribution setting. Surprisingly, although high effective rank in the minor directions of source is essential for benign overfitting, for target distribution only the overall magnitude matters.

3.1 Warm-up: in-distribution benign overfitting

As a warm-up, we introduce Tsigler & Bartlett (2023)'s in-distribution result on benign overfitting in ridge regression. When the data dimension d exceeds the sample size n, the ridge estimator interpolates the training data, fitting the noise. In this case, the estimator $\widehat{\beta}$ lies in the subspace spanned by the n samples. If d is much larger than n, a new test point will likely be orthogonal to this subspace, preventing noise from affecting the prediction. The minor components of the covariance matrix actually provide implicit regularization in this case. Tsigler & Bartlett (2023) assume the data lies in a space with k major directions and $d-k$ weak, but essentially high-dimensional minor directions, allowing benign overfitting. This intuition is formalized through an assumption that controls the condition number of the Gram matrix for the remaining $d-k$ dimensions.

Assumption 1 (CondNum(k, δ, L), (Tsigler & Bartlett, 2023))**.** Define a matrix $A_k = \lambda I_n + X_{-k}X_{-k}^T$. With probability at least $1-\delta$, A_k is positive definite and has a condition number no greater than L, i.e.,

$$\frac{\mu_1(A_k)}{\mu_n(A_k)} \leq L.$$

Remark 1. This assumption is essentially assuming the minor directions have effective rank significantly larger than n. As an evidence, Tsigler & Bartlett (2023) prove that if CondNum holds, then the effective rank r_k is lower bounded by n/L. On the other hand, a lower bound on the effective rank r_k can also imply an upper bound of the condition number of A_k. See Tsigler & Bartlett (2023, Lemma 3) for further detail.

Assuming CondNum, Tsigler & Bartlett (2023) obtain sharp upper bounds for both the variance and bias of the ridge estimator, with matching lower bounds (see their Theorem 2). To facilitate the presentation, we use $\widetilde{\lambda} := \lambda + \sum_{j>k} \lambda_j$ to denote the overall regularization term.

Theorem 1 (Tsigler & Bartlett (2023))**.** There exists a constant c that only depends on σ, L, such that for any $n > ck$, if the assumption condNum(k, δ, L) (Assumption 1) is satisfied, then it holds that $n < cr_k$, and with probability at least $1 - \delta - ce^{-n/c}$,

$$\frac{V}{cv^2} \leq \frac{k}{n} + \frac{n}{R_k}, \qquad \frac{B}{c} \leq B_{\text{ID}} := \|\beta_k^\star\|^2_{\Sigma_{S,k}^{-1}} (\frac{\widetilde{\lambda}}{n})^2 + \left\|\beta_{-k}^\star\right\|^2_{\Sigma_{S,-k}}.$$

The first variance term arises from estimating the k major signal dimensions, corresponding to the classic variance for k-dimensional ordinary least squares. The second variance term, n/R_k, vanishes when the minor directions are sufficiently high-dimensional, i.e., when $R_k \gg n$. However, the signal in the minor directions, $\left\|\beta_{-k}^\star\right\|^2_{\Sigma_{S,-k}}$, is nearly lost when projected from the high-dimensional ambient space onto the low-dimensional sample space, contributing to the second bias term. Finally, the first bias term relates to the signal estimation in the first k dimensions and is introduced by the overall regularization induced by both ridge and implicit regularization from the minor components.

3.2 Out-of-Distribution benign overfitting

We now investigate the out-of-distribution performance of ridge estimator. Intuitively, when all the minor components vanish (both on the source and the target), the over-parameterized ridge regression is actually reduced to the usual ridge regression on the major directions, thus achieving a rate of $\widetilde{\mathcal{O}}(\text{tr}[\mathcal{T}]/n)$ as Ge et al. (2024) demonstrate. When the minor components do not vanish, high effective rank of minor components on the source is essential for "benign overfitting", as Tsigler & Bartlett (2023) demonstrate. However, we argue that, regarding the target distribution, only the *overall magnitude* of those minor components is crucial for benign overfitting. The reason is that, when the minor directions of source have effective rank much larger than n, the n-dimensional subspace spanned by training samples is already almost orthogonal to any test point, with a high probability. Therefore, no special spectral structure of the target is needed for benign overfitting. Only small overall magnitude of those minor components on target is required.

We formalize those intuitive claims, by deriving upper bounds for both the variance and bias of ridge regression under covariate shift, assuming a source distribution similar to the in-distribution case. Our upper bound is sharp, and can be applied to any target distributions, reducing to Tsigler & Bartlett (2023)'s bound (Theorem 1) when the target and source distributions are aligned. Additionally, we recover Ge et al. (2024)'s sharp bound for under-parameterized linear regression under covariate shift when the high-dimensional minor components vanish.

Theorem 2. There exists a constant $c > 2$ depending only on σ, L, such that for any $cN < n < r_k$, if the assumption condNum(k, δ, L) (Assumption 1) is satisfied, then with probability at least $1-3\delta$,

$$\frac{V}{cv^2} \leq \frac{k}{n} \cdot \frac{\text{tr}[\mathcal{T}]}{k} + \frac{n}{R_k} \cdot \frac{\text{tr}[\mathcal{U}]}{\text{tr}[\mathcal{V}]}.$$
$$\frac{B}{c} \leq B_{\text{ID}} \cdot \left(\|\mathcal{T}\| + \frac{n}{r_k} \frac{\|\Sigma_{T,-k}\|}{\|\Sigma_{S,-k}\|}\right).$$

where $\mathcal{T}, \mathcal{U}, \mathcal{V}$ are defined in Equation (1).

$N = \text{Poly}(k + \ln(1/\delta), \lambda_1 \lambda_k^{-1}, 1 + \widetilde{\lambda} \lambda_k^{-1})$. Poly$(\cdot)$ denotes a polynomial function.

Recall B_{ID} is the upper bound for bias given by Theorem 1, we can see that Theorem 2 establishes an upper bound for the excess risk of ridge regression under general covariate shift, expressed as a multiplicative form of Theorem 1's results. This formulation enables a direct analysis of the impact of covariate shift on the bias and variance of ridge estimators, compared to the in-distribution case. The first conclusion is that Theorem 2 well reduces to the corresponding result in Theorem 1 when no distribution shift occurs–i.e., $\Sigma_S = \Sigma_T$. This connection follows directly from the condition $n < r_k$.

The second conclusion is that covariate shift in the first k dimensions and last $d-k$ dimensions introduce multiplicative factors of $\frac{\mathrm{tr}[\mathcal{T}]}{k}$, $\|\mathcal{T}\|$ and $\frac{\mathrm{tr}[\mathcal{U}]}{\mathrm{tr}[\mathcal{V}]}$, $nr_k^{-1}\frac{\|\Sigma_{T,-k}\|}{\|\Sigma_{S,-k}\|}$, respectively, on the excess risk. Therefore, as long as these factors are bounded by constants, over-parameterized ridge regression achieves the same non-asymptotic rate of excess risk under covariate shift as the in-distribution setting. This scenario, well addressed by ridge regression, occurs when the target distribution's covariance structure remains dominated by the first k dimensions. In the following, we analyze the impact of the factors introduced by covariate shift on both the major and minor directions.

1. $\mathcal{T}$ **characterizes the shift in the major directions.** Under covariate shift within the first k dimensions, we obtain the same non-asymptotic error rate as in Theorem 1, only if $\|\mathcal{T}\|$ is bounded by a constant, as $\mathrm{tr}[\mathcal{T}]/k \le \|\mathcal{T}\|$. The matrix $\mathcal{T}$ plays a central role in Theorem 2 to quantify covariate shift within the first k dimensions, matching our intuition. This echoes with Ge et al. (2024)'s finding that $\mathrm{tr}[\mathcal{T}]$ captures the difficulty of covariate shift for under-parameterized ridgeless regression (MLE). They establish a sharp upper bound on excess risk using Fisher information (see their Theorem 3.1), which simplifies to a rate of $\widetilde{\mathcal{O}}(\mathrm{tr}[\mathcal{T}]/n)$ for linear models. Theorem 2 recovers this result when applied to a k-dimensional under-parameterized setting where all high-dimensional minor components vanish, specifically when $\Sigma_{S,-k} = \Sigma_{T,-k} = \mathbf{0}$. Under the same condition as Theorem 2, for a constant c depending only on σ, L, with high probability the variance and bias terms are bounded by:
$$\frac{V}{cv^2} \le \frac{\mathrm{tr}[\mathcal{T}]}{n}, \quad \frac{B}{c} \le \|\beta_k^\star\|^2_{\Sigma_{S,k}^{-1}}\left(\frac{\lambda}{n}\right)^2\|\mathcal{T}\|.$$
The variance bound aligns with Ge et al. (2024)'s result while the bias vanishes as $\lambda \to 0$.

2. **The overall magnitude of** $\Sigma_{T,-k}$ **is crucial for benign overfitting.** Under covariate shift within the last $d-k$ dimensions, when both $\frac{\mathrm{tr}[\mathcal{U}]}{\mathrm{tr}[\mathcal{V}]}$ and $nr_k^{-1}\frac{\|\Sigma_{T,-k}\|}{\|\Sigma_{S,-k}\|}$ are bounded by constants, we achieve the same non-asymptotic error rate as in Theorem 1. Note that $\frac{\mathrm{tr}[\mathcal{U}]}{\mathrm{tr}[\mathcal{V}]} \le \frac{\|\Sigma_{T,-k}\|_{\mathrm{F}}}{\|\Sigma_{S,-k}\|_{\mathrm{F}}}$. In other words, matching our intuition, if the *overall magnitude* of the minor components of target covariance scales similarly to or smaller than those of the source, in terms of the covariance norms, "benign overfitting" also happens under covariate shift. Importantly, this condition does not impose constraints on the internal spectral structure of the minor components of target covariance. For example, we do not force each eigenvalue of $\Sigma_{T,-k}$ to scale with its corresponding eigenvalue of $\Sigma_{S,-k}$ in decreasing order, as assumed in prior work (Mallinar et al., 2024). Surprisingly, for benign overfitting to happen, it is essential for the source distribution to have high effective rank in the minor directions; however for target distribution, only the overall magnitude matters.
   Another observation is that the bias scales with $nr_k^{-1}\frac{\|\Sigma_{T,-k}\|}{\|\Sigma_{S,-k}\|}$, meaning that we only require $\frac{\|\Sigma_{T,-k}\|}{\|\Sigma_{S,-k}\|} = \mathcal{O}(r_k/n)$, which is a less restrictive condition for larger r_k. Thus, over-parameterization improves robustness of the estimation bias against covariate shift in the minor direction.

Remark 2 (Sample complexity)**.** We have assumed $n \ge c_x N$ in Theorem 2. The explicit formula for N is deferred to Theorem 25 and Remark 8. Here we summarize the sample complexity required for the bound to hold. The dependence on k varies between $\Omega(k)$ and $\Omega(k^3)$, depending on the degree of covariate shift. The optimal case, aligning with the sample complexity of classic linear regression, occurs when $\Sigma_{S,k} \approx \Sigma_{T,k}$. The worst case arises when there is significant covariate shift in the first k dimensions, such as when the test data lies predominantly in the subspace of the first dimension. This variation in sample complexity under covariate shift parallels the analysis of Ge et al. (2024) (see theire Theorem 4.2) for the under-parameterized setting. Additionally, we require $n \gg \lambda + \sum_{j>k}\lambda_j$, ensuring that the regularization is not too strong to introduce a bias exceeding a constant (as reflected in the first bias term). On the other hand, we assume $n < r_k$ in the theorem, consistent with the over-parameterized regime and Assumption 1, where the last $d-k$ components are considered to be essentially high-dimensional.

Remark 3 (Dependence on L)**.** Theorem 2 does not explicitly show how the excess risk depends on the condition number L of A_k. However, we demonstrate in Theorem 25 that out bounds scale at most as L^2. Notably, we maintain the same order of dependence on L in each term of the upper bounds as in the analysis by Tsigler & Bartlett (2023) (see their Theorem 5).

Finally, Theorem 2 suggests an $\mathcal{O}(1/n)$ vanishing error under several conditions that naturally follow from the previous discussions, which we now state rigorously. First, the covariate space decomposes into subspaces spanned by low-dimensional major directions and high-dimensional minor directions, with $k = \mathcal{O}(1)$ and $R_k = \Omega(n^2)$. Second, the low-rank covariance structure is preserved after covariate shift, such that $\|\mathcal{T}\|$, $\frac{\mathrm{tr}[\mathcal{U}]}{\mathrm{tr}[\mathcal{V}]}, nr_k^{-1}\frac{\|\Sigma_{T,-k}\|}{\|\Sigma_{S,-k}\|} = \mathcal{O}(1)$. Third, the signal lies predominantly in the major directions, with $\|\beta^\star_k\|_{\Sigma_{S,k}^{-1}} = \mathcal{O}(1)$ and $\left\|\beta^\star_{-k}\right\|_{\Sigma_{S,-k}} = \mathcal{O}(1/\sqrt{n})$. Lastly, the regularization is not excessively strong to introduce a significant bias, with $\widetilde{\lambda} = \lambda + \sum_{j>k}\lambda_j = \mathcal{O}(\sqrt{n})$.

4 Large shift in minor directions

In the previous section, we established an upper bound for overparameterized ridge regression under covariate shift. We showed that when the shift in the minor directions is controlled—specifically, when the overall magnitude of $\Sigma_{T,-k}$ is small—"benign overfitting" also occurs under covariate shift. However, when the shift in minor directions is significant, meaning the target covariance matrix has many large eigenvalues with corresponding eigenvectors outside the major directions, the excess risk for ridge regression deteriorates. In this section, we further illustrate the limitations of ridge regression in such cases by providing a lower bound for its performance for large distribution shift in the minor directions, showing that it can only achieve the slow rate of $\mathcal{O}(1/\sqrt{n})$ for the excess risk. On the other hand, it is natural to consider alternative algorithms to ridge regression in this scenario. We demonstrate that even with a large shift in the minor directions, principal component regression (PCR) is guaranteed to achieve the fast rate $\mathcal{O}(1/n)$, provided that the signal $\beta^\star$ lies primarily within the subspace spanned by the major directions. Moreover, PCR does not require the minor directions to have a high effective rank in the source distribution, highlighting its advantage over ridge regression in such cases.

4.1 Slow rate for ridge regression

In this subsection, we demonstrate the limitations of ridge regression when the overall magnitude of $\Sigma_{T,-k}$ is large. Consider an instance where Σ_S has its first k components as $\Theta(1)$, while the minor directions have eigenvalues of $o(1)$. If we set $\Sigma_T = I_d$, in contrast to the "benign overfitting" regime described in Theorem 2, ridge regression will have a large excess risk for this instance. Although the signal from the major directions is effectively captured, the signal in the minor directions is nearly lost. Unlike the case in Section 3, here the estimation error in the minor directions is crucial because the target distribution has significant components in these directions. We formalize this intuitive example through the following theorems:

Corollary 3. For some absolute constants C_1, C_2, consider the following instance of Σ_S:

$$\lambda_1 = \cdots = \lambda_k = 1, \quad \lambda_{k+1} = \cdots = \lambda_{k+\lfloor\frac{\sqrt{n}}{C_2}\rfloor} = \frac{C_1}{\sqrt{n}}, \quad \lambda_{k+\lfloor\frac{\sqrt{n}}{C_2}\rfloor+1} = \cdots = \lambda_d = 0.$$

Assume $\Sigma_{T,-k} = \mathbf{0}, \Sigma_{T,k} = I_k$, and $\beta^\star_{-k} = 0$. By choosing $\lambda = \sqrt{n}$, under the same conditions of Theorem 2, we can bound the excess risk of the ridge estimator with probability at least $1 - 3\delta$:

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\mathcal{R}\big(\widehat{\beta}(Y)\big)\right] \le \mathcal{O}\Big(\frac{v^2 k + \|\beta^\star\|^2}{n}\Big).$$

Remark 4. Corollary 3 is a direct application of Theorem 2.

Theorem 4. Consider the same instance of Σ_S as in Corollary 3. Assume $\Sigma_T = I_d$ and $\lambda = \sqrt{n}$. There exists an absolute constant $C > 0$, such that for some $0 < \delta < 1$, $N_2 > 0$ and for any $n > N_2$, with probability at least $1 - \delta$, we have $V \ge Cv^2$.

Furthermore, for any $\lambda > 0$, we can bound the excess risk of the ridge estimator with probability at least $1 - \delta$:

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\mathcal{R}\big(\widehat{\beta}(Y)\big)\right] \ge C\frac{\|\beta^\star\|^2 \wedge v^2}{\sqrt{n}}.$$

From Theorem 4, we observe that when $\Sigma_T = I_d$, the performance of ridge regression deteriorates compared to the case where $\Sigma_{T,-k} = \mathbf{0}$. If we set $\lambda = \sqrt{n}$ as in Corollary 3, ridge regression incurs a constant excess risk under covariate shift, while achieving an in-distribution error rate of $\mathcal{O}(1/n)$. Furthermore, Theorem 4 shows no matter how we choose the regularization parameter λ, the excess risk is always lower bounded by the slow statistical rate $\mathcal{O}(1/\sqrt{n})$, which is worse than the fast rate of $\mathcal{O}(1/n)$. However, as we will prove in the next subsection, principal component regression (PCR) can achieve an excess risk of $\mathcal{O}(1/n)$ under this instance, even with $\Sigma_T = I_d$.

4.2 Fast rate for principal component regression

As discussed earlier, ridge regression faces significant limitations when there is a large shift in the minor directions. In Section 3.1, it was shown that the signal in the minor directions, $\beta^\star_{-k}$, is nearly lost when projected from the high-dimensional ambient space onto the low-dimensional sample space. In other words, learning the true signal from the minor directions is essentially impossible. Therefore, in this subsection, we continue to focus on the scenario where the true signal $\beta^\star$ primarily resides in the major directions. In this case, principal component regression (PCR) emerges as a natural algorithm which estimates the space spanned by the major directions and performs regression on that subspace.

Principal Component Regression (PCR).

- **Step 1: Obtain an estimator $\widehat{U}$ of the top-k subspace of Σ_S.** For simplicity, we assume a sample size of $2n$ and use the first half of the data to compute $\widehat{U}$ by principal component analysis (PCA) on the sample covariance matrix $\widehat{\Sigma}_S := \frac{1}{n}X^TX$. Specifically, $\widehat{U} = (\widehat{u}_1, \cdots, \widehat{u}_k)$ where $\widehat{u}_i$ is the i-th eigenvector of $\widehat{\Sigma}_S$.
- **Step 2: Use the data projected on $\widehat{U}$ to conduct linear regression.** With a little abuse of notation, we use $X \in \mathbb{R}^{n\times d}$ to denote the data matrix $(x_{n+1}, \cdots, x_{2n})^T$, and $Y \in \mathbb{R}^n$ to denote $(y_{n+1}, \cdots, y_{2n})^T$. If we let $Z := X\widehat{U} \in \mathbb{R}^{n\times k}$ be the projected data matrix, the estimator $\widehat{\beta}$ we obtained is given by

$$\widehat{\beta} = \widehat{U}(Z^TZ)^{-1}Z^TY = \widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TY.$$

Consider the scenario where the last $d-k$ components of the true signal β^* is exactly zero, namely $\beta^\star_{-k} = 0$. We can imagine that if the subspace represented by $\widehat{U}$ is exactly the same as the subspace represented by $U = \begin{pmatrix} I_k \\ 0 \end{pmatrix} \in \mathbb{R}^{d\times k}$, (i.e., the first k components), then PCR is actually doing linear regression using only the first k components of the samples, therefore will only have a excess risk induced by the usual variance of linear regression in the major directions. Under this scenario, no matter how large $\|\Sigma_{T,-k}\|$ is, the PCR estimator have zero estimates on the last $d-k$ components, therefore avoid inducing large excess risk. Further more, if the distance between $\widehat{U}$ and U is not zero, there should be another term in the excess risk induced by the estimation error of $\widehat{U}$. We formalize this intuitive claim as the following upper bound for the excess risk of PCR. To facilitate the presentation, we introduce the following quantity for measuring the estimation accuracy of $\widehat{U}$. We define $\Delta = \mathrm{dist}(\widehat{U}, U) := \|UU^T - \widehat{U}\widehat{U}^T\|$, the distance between the subspace spanned by the columns of $\widehat{U}$ and U. Then we have the following theorem:

Theorem 5. Assume $\beta^\star_{-k} = 0$. If $\Delta \le \Theta$, for any $0 < \delta < 1$ and any $n \ge N_1$, we can bound the excess risk of PCR estimator $\widehat{\beta}$ with probability $1-\delta$:

$$\mathbb{E}_{\boldsymbol{\epsilon}}\big[\mathcal{R}\big(\widehat{\beta}(Y)\big)\big] \le \mathcal{O}\left(v^2\frac{\mathrm{tr}(\mathcal{T})}{n} + \|\beta^\star\|^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\|\Delta^2\right),$$

where Θ, N_1 is defined as follows:

$\Theta^{-1} = \mathrm{Poly}(\lambda_1\lambda_k^{-1}, \|\Sigma_T\|\lambda_k^{-1}, k\,\mathrm{tr}(\mathcal{T})^{-1}),$

$N_1 = \mathrm{Poly}(\sigma, \lambda_1\lambda_k^{-1}, \|\Sigma_T\|\lambda_k^{-1}, k\ln(1/\delta), k\,\mathrm{tr}(\mathcal{T})^{-1}).$

Remark 5. Theorem 5 is a special case of Lemma 31. For detailed characterization of Θ and N_1, as well as an upper bound for cases where $\beta^\star_{-k} \ne 0$, one can refer to Lemma 31 for detail.

The excess risk upper bound given by Theorem 5 consists of two terms. The variance term $\frac{\mathrm{tr}(\mathcal{T})}{n}$ is incurred by the nature of linear regression on the major directions, and is unavoidable even if the subspace estimation is accurate (i.e., $\Delta = 0$). This term also appears in the first term of variance in Theorem 2, and exactly matches the sharp rate $\mathrm{tr}[\Sigma_S^{-1}\Sigma_T]/n$ for under-parameterized linear regression under covariate shift (Ge et al., 2024). The second term $\|\beta^\star\|^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\|\Delta^2$ is the bias term induced by the estimation error of the subspace in the first step. We can see that it has a quadratic dependence on Δ. If we combine Theorem 5 with a control of Δ, we can get the end-to-end excess risk upper bound of PCR. For controlling Δ, we have the following lemma:

Lemma 6. With probability at least $1-\delta$, if $n \geq r + \ln(1/\delta)$, we have

$$\Delta \leq \mathcal{O}\left(\sigma^4 \frac{\lambda_1}{\lambda_k - \lambda_{k+1}} \sqrt{\frac{r + \ln\frac{1}{\delta}}{n}}\right),$$

where $r = \lambda_1^{-1}\sum_{i=1}^d \lambda_i$ is the effective rank of the entire Σ_S.

Remark 6. Lemma 6 shows that Δ depends on several quantities: the eigenvalue gap between the major directions and the minor directions, i.e., $\lambda_k - \lambda_{k+1}$, and the effective rank r. We can see that Δ will be small, if the major directions and the minor directions are well separated, i.e., $\lambda_k - \lambda_{k+1}$ is large, and the minor directions are relatively small compared to λ_1.

Combining Theorem 5 and Lemma 6, an end-to-end error bound for PCR can be derived (for a detailed theorem, one can refer to Theorem 29), suggesting that PCR will achieve a small excess risk, as long as the major directions and the minor directions are well separated, and the effective rank of the entire source covariance matrix is small. Contrast to ridge regression, PCR does not require the minor components to have high-effective rank. This shows the superiority of PCR compared with ridge regression under certain scenarios.

As an example, consider the instance in Theorem 4, where $k, \|\Sigma_T\|, \lambda_1, \lambda_k$ are all $\Theta(1)$. In this case, the variance term will scale as $1/n$, and the bias term scales as $\mathcal{O}(\Delta^2)$. Notice that in this instance, $r = \Theta(1)$, therefore $\Delta \leq \mathcal{O}(1/\sqrt{n})$. We conclude that in this instance, PCR will achieve a $\mathcal{O}(1/n)$ rate even when $\Sigma_T = I_d$. Comparing with the excess risk for ridge regression, which is at least $1/\sqrt{n}$, PCR shows its superiority against ridge regression under the scenario where the shift in minor directions is large.

5 Conclusion and discussion

In conclusion, we provide an instance-dependent characterization of the excess risk for ridge regression under general covariate shift. Our findings demonstrate that "benign overfitting" also happens in OOD generalization when the shift in the minor directions is well controlled. We also explore the "large shift in the minor directions" regime, under which ridge regression may incur a large excess risk, whereas principal component regression (PCR) exhibits superior performance.

Our work opens up several future research directions. First, while we have established a lower bound for ridge regression in certain instances, a key challenge remains in deriving a general lower bound that matches our upper bounds, offering a precise characterization of the excess risk under covariate shift. Second, our analysis has been focused on linear models as a first step in understanding overparameterized OOD problems. Extending this investigation to more complex, nonlinear models would be a interesting direction for future exploration.

References

Sergios Agapiou, Omiros Papaspiliopoulos, Daniel Sanz-Alonso, and Andrew M Stuart. Importance sampling: Intrinsic dimension and computational cost. *Statistical Science*, pp. 405–431, 2017.

Anish Agarwal, Devavrat Shah, Dennis Shen, and Dogyoon Song. On robustness of principal component regression. *Advances in Neural Information Processing Systems*, 32, 2019.

Anish Agarwal, Devavrat Shah, and Dennis Shen. On model identification and out-of-sample prediction of principal component regression: Applications to synthetic controls. *arXiv preprint arXiv:2010.14449*, 2020.

Eric Bair, Trevor Hastie, Debashis Paul, and Robert Tibshirani. Prediction by supervised principal components. *Journal of the American Statistical Association*, 101(473):119–137, 2006.

Peter L Bartlett, Philip M Long, Gábor Lugosi, and Alexander Tsigler. Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48):30063–30070, 2020.

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

Mikhail Belkin and Partha Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural computation*, 15(6):1373–1396, 2003.

Mikhail Belkin, Daniel Hsu, and Ji Xu. Two models of double descent for weak features. *SIAM Journal on Mathematics of Data Science*, 2(4):1167–1180, 2020.

Shai Ben-David, John Blitzer, Koby Crammer, and Fernando Pereira. Analysis of representations for domain adaptation. *Advances in neural information processing systems*, 19, 2006.

Koby Bibas, Yaniv Fogel, and Meir Feder. A new look at an old problem: A universal learning approach to linear regression. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pp. 2304–2308. IEEE, 2019.

Swee Lean Chan and Moonseo Park. Project cost estimation using principal component regression. *Construction Management and Economics*, 23(3):295–304, 2005.

Yihang Chen, Fanghui Liu, Taiji Suzuki, and Volkan Cevher. High-dimensional kernel methods under covariate shift: Data-dependent implicit regularization. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=bBzlapzeR1.

Yuxin Chen, Yuejie Chi, Jianqing Fan, Cong Ma, et al. Spectral methods for data science: A statistical perspective. *Foundations and Trends® in Machine Learning*, 14(5):566–806, 2021.

Geoffrey Chinot and Matthieu Lerasle. On the robustness of the minimim l2 interpolator. *Bernoulli*, 2022.

Corinna Cortes, Yishay Mansour, and Mehryar Mohri. Learning bounds for importance weighting. *Advances in neural information processing systems*, 23, 2010.

Uwe Depczynski, VJ Frost, and K Molt. Genetic algorithms applied to the selection of factors in principal component regression. *Analytica Chimica Acta*, 420(2):217–227, 2000.

Edgar Dobriban and Stefan Wager. High-dimensional asymptotics of prediction: Ridge regression and classification. *The Annals of Statistics*, 46(1):247–279, 2018.

Jiawei Ge, Shange Tang, Jianqing Fan, and Chi Jin. On the provable advantage of unsupervised pretraining. *arXiv preprint arXiv:2303.01566*, 2023.

Jiawei Ge, Shange Tang, Jianqing Fan, Cong Ma, and Chi Jin. Maximum likelihood estimation is all you need for well-specified covariate shift. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=eoTCKKOgIs.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. When do neural networks outperform kernel methods? *Advances in Neural Information Processing Systems*, 33: 14820–14830, 2020.

Behrooz Ghorbani, Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Linearized two-layers neural networks in high dimension. *The Annals of Statistics*, 49(2), 2021.

Hao Guan and Mingxia Liu. Domain adaptation for medical image analysis: a survey. *IEEE Transactions on Biomedical Engineering*, 69(3):1173–1185, 2021.

Ali S Hadi and Robert F Ling. Some cautionary notes on the use of principal components regression. *The American Statistician*, 52(1):15–19, 1998.

Steve Hanneke and Samory Kpotufe. On the value of target data in transfer learning. *Advances in Neural Information Processing Systems*, 32, 2019.

Yifan Hao, Yong Lin, Difan Zou, and Tong Zhang. On the benefits of over-parameterization for out-of-distribution generalization. *CoRR*, abs/2403.17592, 2024. doi: 10.48550/ARXIV.2403.17592. URL https://doi.org/10.48550/arXiv.2403.17592.

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *Annals of statistics*, 50(2):949, 2022.

Dan Hendrycks and Thomas Dietterich. Benchmarking neural network robustness to common corruptions and perturbations. *arXiv preprint arXiv:1903.12261*, 2019.

Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, et al. The many faces of robustness: A critical analysis of out-of-distribution generalization. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8340–8349, 2021.

Hugo G Hidalgo, Thomas C Piechota, and John A Dracup. Alternative principal components regression procedures for dendrohydrologic reconstructions. *Water Resources Research*, 36(11): 3241–3249, 2000.

Shih-Ming Huang and Jar-Ferr Yang. Improved principal component regression for face recognition under illumination variations. *IEEE signal processing letters*, 19(4):179–182, 2012.

JNR Jeffers. Investigation of alternative regressions: Some practical examples. *Journal of the Royal Statistical Society. Series D (The Statistician)*, 30(2):79–88, 1981.

John NR Jeffers. Two case studies in the application of principal component analysis. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 16(3):225–236, 1967.

Ian T Jolliffe. A note on the use of principal components in regression. *Journal of the Royal Statistical Society Series C: Applied Statistics*, 31(3):300–303, 1982.

Chinmaya Kausik, Kashvi Srivastava, and Rishi Sonthalia. Double descent and overfitting under noisy inputs and distribution shift for linear denoisers. *Trans. Mach. Learn. Res.*, 2024, 2024. URL https://openreview.net/forum?id=HxfqTdLIRF.

Richard B Keithley, R Mark Wightman, and Michael L Heien. Multivariate concentration determination using principal component regression with residual analysis. *TrAC Trends in Analytical Chemistry*, 28(9):1127–1136, 2009.

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. *Journal of Machine Learning Research*, 21(169):1–16, 2020.

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

Pang Wei Koh, Shiori Sagawa, Henrik Marklund, Sang Michael Xie, Marvin Zhang, Akshay Balsubramani, Weihua Hu, Michihiro Yasunaga, Richard Lanas Phillips, Irena Gao, et al. Wilds: A benchmark of in-the-wild distribution shifts. In *International conference on machine learning*, pp. 5637–5664. PMLR, 2021.

Samory Kpotufe and Guillaume Martinet. Marginal singularity, and the benefits of labels in covariate-shift. In *Conference On Learning Theory*, pp. 1882–1886. PMLR, 2018.

Anikender Kumar and Pramila Goyal. Forecasting of air quality in delhi using principal component regression technique. *Atmospheric Pollution Research*, 2(4):436–444, 2011.

Qi Lei, Wei Hu, and Jason Lee. Near-optimal linear regression under distribution shift. In *International Conference on Machine Learning*, pp. 6164–6174. PMLR, 2021.

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*, pp. 2683–2711. PMLR, 2020.

RX Liu, J Kuang, Qiong Gong, and XL Hou. Principal component regression analysis with spss. *Computer methods and programs in biomedicine*, 71(2):141–147, 2003.

Cong Ma, Reese Pathak, and Martin J Wainwright. Optimally tackling covariate shift in rkhs-based nonparametric regression. *The Annals of Statistics*, 51(2):738–761, 2023.

Neil Mallinar, Austin Zane, Spencer Frei, and Bin Yu. Minimum-norm interpolation under covariate shift. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Zw7TcnTmHj.

William F Massy. Principal components regression in exploratory statistical research. *Journal of the American Statistical Association*, 60(309):234–256, 1965.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 75(4):667–766, 2022.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Generalization error of random feature and kernel methods: hypercontractivity and kernel matrix concentration. *Applied and Computational Harmonic Analysis*, 59:3–84, 2022.

John P Miller, Rohan Taori, Aditi Raghunathan, Shiori Sagawa, Pang Wei Koh, Vaishaal Shankar, Percy Liang, Yair Carmon, and Ludwig Schmidt. Accuracy on the line: on the strong correlation between out-of-distribution and in-distribution generalization. In *International conference on machine learning*, pp. 7721–7735. PMLR, 2021.

Andrea Montanari and Yiqiao Zhong. The interpolation phase transition in neural networks: Memorization and generalization under lazy training. *The Annals of Statistics*, 50(5):2816–2847, 2022.

Mohammadreza Mousavi Kalan, Zalan Fabian, Salman Avestimehr, and Mahdi Soltanolkotabi. Minimax lower bounds for transfer learning with linear and one-hidden layer neural networks. *Advances in Neural Information Processing Systems*, 33:1959–1969, 2020.

Vidya Muthukumar, Kailas Vodrahalli, Vignesh Subramanian, and Anant Sahai. Harmless interpolation of noisy data in regression. *IEEE Journal on Selected Areas in Information Theory*, 1(1): 67–83, 2020.

Tormod Næs and Harald Martens. Principal component regression in nir analysis: viewpoints, background details and selection of components. *Journal of chemometrics*, 2(2):155–167, 1988.

Preetum Nakkiran. More data can hurt for linear regression: Sample-wise double descent. *arXiv preprint arXiv:1912.07242*, 2019.

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*, pp. 7263–7272. PMLR, 2020.

Partha Niyogi. Manifold regularization and semi-supervised learning: Some theoretical analyses. *Journal of Machine Learning Research*, 14(5), 2013.

Reese Pathak, Cong Ma, and Martin Wainwright. A new similarity measure for covariate shift with applications to nonparametric regression. In *International Conference on Machine Learning*, pp. 17517–17530. PMLR, 2022.

Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do imagenet classifiers generalize to imagenet? In *International conference on machine learning*, pp. 5389–5400. PMLR, 2019.

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

Ohad Shamir. The implicit bias of benign overfitting. *Journal of Machine Learning Research*, 24 (113):1–40, 2023.

Hidetoshi Shimodaira. Improving predictive inference under covariate shift by weighting the log-likelihood function. *Journal of statistical planning and inference*, 90(2):227–244, 2000.

James B Simon, Dhruva Karkada, Nikhil Ghosh, and Mikhail Belkin. More is better in modern machine learning: when infinite overparameterization is optimal and overfitting is obligatory. *arXiv preprint arXiv:2311.14646*, 2023.

Jianguo Sun. A correlation principal component regression analysis of nir data. *Journal of Chemometrics*, 9(1):21–29, 1995.

Nilesh Tripuraneni, Ben Adlam, and Jeffrey Pennington. Overparameterization improves robustness to covariate shift in high dimensions. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pp. 13883–13897, 2021a. URL https://proceedings.neurips.cc/paper/2021/hash/73fed7fd472e502d8908794430511f4d-Abstract.html.

Nilesh Tripuraneni, Chi Jin, and Michael Jordan. Provable meta-learning of linear representations. In *International Conference on Machine Learning*, pp. 10434–10443. PMLR, 2021b.

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *J. Mach. Learn. Res.*, 24:123:1–123:76, 2023. URL http://jmlr.org/papers/v24/22-1398.html.

J Leo van Hemmen and Tsuneya Ando. An inequality for trace ideals. *Communications in Mathematical Physics*, 76:143–148, 1980.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *CoRR*, abs/1011.3027, 2010. URL http://arxiv.org/abs/1011.3027.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

Evelyne Vigneau, MF Devaux, EM Qannari, and P Robert. Principal component regression, ridge regression and ridge principal component regression in spectroscopy calibration. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 11(3):239–249, 1997.

Per-Ake Wedin. Perturbation theory for pseudo-inverses. *BIT Numerical Mathematics*, 13:217–232, 1973.

Florian Wenzel, Andrea Dittadi, Peter Gehler, Carl-Johann Simon-Gabriel, Max Horn, Dominik Zietlow, David Kernert, Chris Russell, Thomas Brox, Bernt Schiele, et al. Assaying out-of-distribution generalization in transfer learning. *Advances in Neural Information Processing Systems*, 35:7181–7198, 2022.

Denny Wu and Ji Xu. On the optimal weighted l2 regularization in overparameterized linear regression. *Advances in Neural Information Processing Systems*, 33:10112–10123, 2020.

Ji Xu and Daniel J Hsu. On the number of variables to use in principal component regression. *Advances in neural information processing systems*, 32, 2019.

Runtian Zhai, Chen Dan, Zico Kolter, and Pradeep Ravikumar. Understanding why generalized reweighting does not improve over erm. *arXiv preprint arXiv:2201.12293*, 2022.

Xuhui Zhang, Jose Blanchet, Soumyadip Ghosh, and Mark S Squillante. A class of geometric structures in transfer learning: Minimax bounds and optimality. In *International Conference on Artificial Intelligence and Statistics*, pp. 3794–3820. PMLR, 2022.

Lijia Zhou, Danica J Sutherland, and Nati Srebro. On uniform convergence and low-norm interpolation learning. *Advances in Neural Information Processing Systems*, 33:6867–6877, 2020.

Yang Zou, Zhiding Yu, BVK Kumar, and Jinsong Wang. Unsupervised domain adaptation for semantic segmentation via class-balanced self-training. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 289–305, 2018.

A Ridge regression

Let $X = (x_1, ..., x_n)^T \in \mathbb{R}^{n \times d}$, $Y = (y_1, ..., y_n)^T \in \mathbb{R}^n$ and $\boldsymbol{\epsilon} = (\epsilon_1, ..., \epsilon_n)^T \in \mathbb{R}^n$. We denote the first k columns of X as X_k and the remaining $d-k$ columns as X_{-k}. Similarly, $\beta_k^\star$ and $\beta_{-k}^\star$ represent the corresponding components of $\beta^\star$. $\Sigma_{S,k}$, $\Sigma_{S,-k}$ are the corresponding blocks on the diagonal of Σ_S. The i-th eigenvalue of a matrix is denoted by $\mu_i(\cdot)$. Define $Z = X\Sigma_S^{-1/2}$, where the rows of Z are i.i.d. centered isotropic random vectors. Additionally, we assume the rows of Z are σ-sub-gaussian, where the sub-gaussian norm is defined as follows.

For a random variable s, the sub-gaussian norm $\|s\|_{\psi_2}$ is given by:

$$\|s\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}\left[\exp\frac{s^2}{t^2}\right] \le 2\right\}.$$

For a random vector S, the sub-gaussian norm $\|S\|_{\psi_2}$ is given by:

$$\|S\|_{\psi_2} = \sup_{v \neq 0} \frac{\|\langle S, v\rangle\|_{\psi_2}}{\|v\|}.$$

For $\lambda \ge 0$, consider the ridge estimator:

$$\begin{aligned}\widehat{\beta}(Y) &= X^T(XX^T + \lambda I_n)^{-1}Y \\ &= X^T(XX^T + \lambda I_n)^{-1}X\beta^\star + X^T(XX^T + \lambda I_n)^{-1}\boldsymbol{\epsilon} \\ &= \widehat{\beta}(X\beta^\star) + \widehat{\beta}(\boldsymbol{\epsilon}),\end{aligned}$$

where we define $\widehat{\beta}(X\beta^\star) = X^T(XX^T+\lambda I_n)^{-1}X\beta^\star$ and $\widehat{\beta}(\boldsymbol{\epsilon}) = X^T(XX^T+\lambda I_n)^{-1}\boldsymbol{\epsilon}$. Additionally, we define $\widetilde{\Sigma}_S = \Sigma_S + \frac{\lambda}{n} I_d$. The effective rank of $\widetilde{\Sigma}_{S,k}$ is defined as $r_k = \lambda_{k+1}^{-1}(\lambda + \sum_{j>k}\lambda_j)$.

Assumption 2 (CondNum(k, δ, L))**.** Define a matrix $A_k = \lambda I_n + X_{-k}X_{-k}^T$. With probability at least $1-\delta$, A_k is positive definite and has a condition number no greater than L, i.e.,

$$\frac{\mu_1(A_k)}{\mu_n(A_k)} \le L.$$

A.1 Concentration inequalities

Denote the element of a matrix X in the i-th row and the j-th column as $X[i,j]$, and the i-th row of the matrix X as $X[i,*]$.

Lemma 7 (Lemma 20 of Tsigler & Bartlett (2023))**.** Let z be a sub-gaussian vector in $\mathbb{R}^p$ with $\|z\|_{\psi_2} \le \sigma$, and consider $\Sigma = \operatorname{diag}(\lambda_1, \ldots, \lambda_p)$ where the sequence $\{\lambda_j\}_{j=1}^p$ is positive and non-increasing. Then there exists some absolute constant c, for any $t > 0$, with probability at least $1 - 2e^{-t/c}$:

$$\|\Sigma^{1/2}z\|^2 \le c\sigma^2\left(t\lambda_1 + \sum_{j=1}^p \lambda_j\right).$$

Lemma 8 (Lemma 23 of Tsigler & Bartlett (2023))**.** Let $\mathring{A}_k$ represent the matrix $X_{-k}X_{-k}^T$ with its diagonal elements set to zero:

$$\mathring{A}_k[i,j] = (1-\delta_{i,j})(X_{-k}X_{-k}^T)[i,j].$$

Then there exists some absolute constant c, for any $t > 0$, with probability at least $1 - 4e^{-t/c}$:

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

Lemma 9 (Lemma 21 of Tsigler & Bartlett (2023))**.** Suppose $\{z_i\}_{i=1}^n$ is a sequence of independent isotropic sub-gaussian random vectors, where $\|z_i\|_{\psi_2} \le \sigma$. Let $\Sigma = \text{diag}(\lambda_1, \dots, \lambda_p)$ represent a diagonal matrix with a positive, non-increasing sequence $\{\lambda_i\}_{i=1}^p$. Then there exists some absolute constant c, for any $t \in (0, n)$, with probability at least $1 - 2e^{-ct}$:

$$(n - \sqrt{nt}\sigma^2) \sum_{j=1}^p \lambda_j \le \sum_{i=1}^n \|\Sigma^{1/2} z_i\|^2 \le (n + \sqrt{nt}\sigma^2) \sum_{j=1}^p \lambda_j.$$

Lemma 10. There exists a constant c_x, depending only on σ, such that for any n satisfying $n\lambda_{k+1} \le \left(\lambda + \sum_{j>k} \lambda_j\right)$, under the assumption CondNum(k, δ, L) (Assumption 2), with probability at least $1 - \delta - c_x e^{-n/c_x}$:

$$\frac{1}{c_x L}\left(\lambda + \sum_{j>k} \lambda_j\right) \le \mu_n(A_k) \le \mu_1(A_k) \le c_x \left(\lambda + \sum_{j>k} \lambda_j\right).$$

$$\mu_1(X_{-k} X_{-k}^T) \le c_x \left(n\lambda_{k+1} + \sum_{j>k} \lambda_j\right).$$

Proof. This result follows from the proof of Lemma 3 in Tsigler & Bartlett (2023), which establishes both upper and lower bounds of $\mu_1(A_k)$. By combining the lower bound with the assumption CondNum, we derive a lower bound of $\mu_n(A_k)$. For completeness, we restate the entire proof here.

According to lemma 7 and lemma 8, there exists an absolute constant c, such that for any $t > 0$:

1. for all $1 \le i \le n$, with probability at least $1 - 2e^{-t/c}$:

$$\|X_{-k}[i, *]\|^2 \le c\sigma^2 \left(t\lambda_{k+1} + \sum_{j>k} \lambda_j\right).$$

2. with probability at least $1 - 4e^{-t/c}$:

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

Since $\mu_1(A_k) \le \lambda + \|\mathring{A}_k\| + \max_i \|X_{-k}[i, *]\|^2$, by setting $t = n$, we have with probability at least $1 - (2n+4)e^{-n/c}$:

$$\begin{aligned}
\mu_1(A_k) &\le \lambda + c\sigma^2 \left(n\lambda_{k+1} + \sum_{j>k} \lambda_j + \sqrt{(2n\lambda_{k+1})^2 + 2n \sum_{j>k} \lambda_j^2}\right) \\
&\le \lambda + c\sigma^2 \left(n\lambda_{k+1} + \sum_{j>k} \lambda_j + 2n\lambda_{k+1} + \sqrt{2n \sum_{j>k} \lambda_j^2}\right) \\
&\le \lambda + c\sigma^2 \left(n\lambda_{k+1} + \sum_{j>k} \lambda_j + 2n\lambda_{k+1} + \sqrt{2n\lambda_{k+1} \sum_{j>k} \lambda_j}\right) \\
&\le \lambda + c\sigma^2 \left(n\lambda_{k+1} + \sum_{j>k} \lambda_j + 2n\lambda_{k+1} + n\lambda_{k+1} + \frac{1}{2}\sum_{j>k} \lambda_j\right) \\
&\le \lambda + 4c\sigma^2 \left(n\lambda_{k+1} + \sum_{j>k} \lambda_j\right)
\end{aligned}$$

$$\leq \max\left\{1, 4c\sigma^2\right\}\left(\lambda + \sum_{j>k}\lambda_j + n\lambda_{k+1}\right)$$

$$\leq 2\max\left\{1, 4c\sigma^2\right\}\left(\lambda + \sum_{j>k}\lambda_j\right). \tag{2}$$

The last inequality follows from $n\lambda_{k+1} \leq \left(\lambda + \sum_{j>k}\lambda_j\right)$. Similarly,

$$\mu_1(X_{-k}X_{-k}^T) \leq 4c\sigma^2\left(n\lambda_{k+1} + \sum_{j>k}\lambda_j\right). \tag{3}$$

On the other hand, by applying Lemma 9 with $t = \frac{n}{4\sigma^4}$, there exists an absolute constant c', such that with probability at least $1 - 2\exp\left\{-\frac{c'}{4\sigma^4}n\right\}$:

$$\sum_{i=1}^{n}\|X_{-k}[i,*]\|^2 \geq \frac{1}{2}n\sum_{j>k}\lambda_j.$$

On this event,

$$\begin{aligned}
\mu_1(A_k) &\geq \lambda + \frac{1}{n}\operatorname{tr}(X_{-k}X_{-k}^T) \\
&= \lambda + \frac{1}{n}\sum_{i=1}^{n}\|X_{-k}[i,*]\|^2 \\
&\geq \lambda + \frac{1}{2}\sum_{j>k}\lambda_j \\
&\geq \frac{1}{2}\left(\lambda + \sum_{j>k}\lambda_j\right).
\end{aligned}$$

By the assumption CondNum(k,δ,L), with probability at least $1 - \delta - 2\exp\left\{-\frac{c'}{4\sigma^4}n\right\}$:

$$\mu_n(A_k) \geq \frac{1}{L}\mu_1(A_k) \geq \frac{1}{2L}\left(\lambda + \sum_{j>k}\lambda_j\right). \tag{4}$$

Combining Equation 2, 3 and 4, there exists a constant c_x depending only on σ, such that with probability at least $1 - \delta - c_x e^{-n/c_x}$:

$$\frac{1}{c_x L}\left(\lambda + \sum_{j>k}\lambda_j\right) \leq \mu_n(A_k) \leq \mu_1(A_k) \leq c_x\left(\lambda + \sum_{j>k}\lambda_j\right).$$

$$\mu_1(X_{-k}X_{-k}^T) \leq c_x\left(n\lambda_{k+1} + \sum_{j>k}\lambda_j\right).$$

□

Lemma 11. There exists a constant c_x depending only on σ, such that with probability at least $1-\delta$, if $n > k + \ln(1/\delta)$,

$$\left\|\frac{1}{n}X_k^T X_k - \Sigma_{S,k}\right\| \leq c_x\lambda_1\sqrt{\frac{k + \ln\frac{1}{\delta}}{n}}.$$

Proof. This follows directly from Theorem 5.39 and Remark 5.40 of Vershynin (2010), which shows there exists a constant c'_x depending only on σ, such that for any $t \geq 0$, with probability at least $1 - 2\exp\{-t^2/c'_x\}$:

$$\left\|\frac{1}{n}X_k^T X_k - \Sigma_{S,k}\right\| \leq \lambda_1 \max\left\{c'_x\sqrt{\frac{k}{n}} + \frac{t}{\sqrt{n}}, \left(c'_x\sqrt{\frac{k}{n}} + \frac{t}{\sqrt{n}}\right)^2\right\}.$$

Taking $t = \sqrt{c'_x \ln(2/\delta)}$ completes the proof. $\square$

Corollary 12. Under the same conditions as in Lemma 11, and on the same event, the following holds:

$$\left\|\left(X_k^T X_k\right)^{\frac{1}{2}} - \sqrt{n}\Sigma_{S,k}^{\frac{1}{2}}\right\| \leq c_x\sqrt{k + \ln\frac{1}{\delta}}\lambda_1\lambda_k^{-\frac{1}{2}}.$$

Proof. According to Proposition 3.2 of van Hemmen & Ando (1980), for any positive semi-definite matrix $A, B \in \mathbb{R}^k$, we have

$$\|A - B\| \geq \left(\mu_k\left(A^{\frac{1}{2}}\right) + \mu_k\left(B^{\frac{1}{2}}\right)\right)\left\|A^{\frac{1}{2}} - B^{\frac{1}{2}}\right\|.$$

Therefore,

$$\begin{aligned}\left\|\left(X_k^T X_k\right)^{\frac{1}{2}} - \sqrt{n}\Sigma_{S,k}^{\frac{1}{2}}\right\| &\leq \frac{1}{\mu_k\left(\sqrt{n}\Sigma_{S,k}^{\frac{1}{2}}\right)}\left\|X_k^T X_k - n\Sigma_{S,k}\right\| \\ &= \sqrt{n}\lambda_k^{-\frac{1}{2}}\left\|\frac{1}{n}X_k^T X_k - \Sigma_{S,k}\right\|.\end{aligned}$$

By applying Lemma 11, the proof is complete. $\square$

Lemma 13. There exists a constant c_x depending only on σ, such that for any $n > c_x k$, with probability at least $1 - 2e^{-n/c_x}$:

$$\frac{1}{c_x}n \leq \mu_k\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) \leq \mu_1\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) \leq c_x n.$$

Proof. According to Theorem 5.39 of Vershynin (2010), there exists a constant c'_x depending only on σ, such that for any $t \geq 0$, with probability at least $1 - 2\exp\{-t^2/c'_x\}$:

$$\begin{aligned}\mu_k\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) &\geq \left(\sqrt{n} - c'_x\sqrt{k} - t\right)^2. \\ \mu_1\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) &\leq \left(\sqrt{n} + c'_x\sqrt{k} + t\right)^2.\end{aligned}$$

Let $t = \frac{1}{2}\sqrt{n}$. For $n > 16(c'_x)^2 k$, with probability at least $1 - 2\exp\left\{-n/(4c'_x)\right\}$:

$$\begin{aligned}\mu_k\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) &\geq \left(\sqrt{n} - \frac{1}{4}\sqrt{n} - \frac{1}{2}\sqrt{n}\right)^2 = \frac{1}{16}n. \\ \mu_1\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) &\leq \left(\sqrt{n} + \frac{1}{4}\sqrt{n} + \frac{1}{2}\sqrt{n}\right)^2 = \frac{49}{16}n.\end{aligned}$$

By taking $c_x = \max\left\{16(c'_x)^2, 4c'_x, 16\right\}$, the proof is complete. $\square$

Remark 7. On the same event, the following inequalities also hold:

$$\begin{aligned}\mu_1(X_k^T X_k) &\leq \|\Sigma_{S,k}\|\left\|\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right\| \leq c_x\lambda_1 n. \\ \mu_k(X_k^T X_k) &\geq \mu_k(\Sigma_{S,k})\mu_k\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) \geq \frac{1}{c_x}\lambda_k n.\end{aligned}$$

Lemma 14. There exists a constant c_x depending only on σ, with probability at least $1 - 2e^{-n/c_x}$:

$$\operatorname{tr}\left(X_{-k}\Sigma_{T,-k}X_{-k}^T\right) \leq c_x n \operatorname{tr}\left(\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}\right).$$

Proof. According to Hanson-Wright Inequality (Vershynin, 2018), there exists an absolute constant c, such that for any $1 \le i \le n$,

$$\left\| Z_{-k}[i,*]\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}Z_{-k}[i,*]^T \right\|_{\psi_1} \le c\sigma^2 \left\| \Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}} \right\|_{\mathrm{F}}$$
$$\le c\sigma^2 \operatorname{tr}\left(\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}\right).$$

By Bernstein Inequality (Proposition 5.16 of Vershynin (2010)), there exists an absolute constant c', for any $t \ge 0$,

$$\mathbb{P}\left\{ \frac{1}{n}\left| \sum_{i=1}^{n} \left[Z_{-k}[i,*]\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}Z_{-k}[i,*]^T - \operatorname{tr}\left(\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}\right)\right]\right| \ge t \right\}$$
$$\le 2\exp\left\{ -c'n \min\left\{ \frac{t^2}{K^2}, \frac{t}{K} \right\} \right\},$$

where $K = \max_i \left\| Z_{-k}[i,*]\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}Z_{-k}[i,*]^T \right\|_{\psi_1}$.

Let $t = c\sigma^2 \operatorname{tr}\left(\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}\right)$. Then, with probability at least $1 - 2e^{-c'n}$:

$$\operatorname{tr}\left(X_{-k}\Sigma_{T,-k}X_{-k}^T\right) = \sum_{i=1}^{n} Z_{-k}[i,*]\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}Z_{-k}[i,*]^T$$
$$\le (1 + c\sigma^2)n \operatorname{tr}\left(\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}\right).$$

By taking $c_x = \max\left\{1 + c\sigma^2, \frac{1}{c'}\right\}$, the proof is complete. □

Lemma 15. There exists a constant c_x depending only on σ, with probablity at least $1 - 2e^{-n/c_x}$:

$$(\beta_{-k}^\star)^T X_{-k}^T X_{-k}\beta_{-k}^\star \le c_x n (\beta_{-k}^\star)^T \Sigma_{S,-k}\beta_{-k}^\star.$$

Proof. The result follows from the proof of Lemma 3 in Tsigler & Bartlett (2023), which we restate here for completeness. Consider the isotropic vector $\left[(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star\right]^{-1/2} X_{-k}\beta_{-k}^\star$. For the i-th component,

$$\left\| \left[(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star\right]^{-\frac{1}{2}} X_{-k}[i,*]\beta_{-k}^\star \right\|_{\psi_2} = \left[(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star\right]^{-\frac{1}{2}} \left\| Z_{-k}[i,*]\Sigma_{S,-k}^{\frac{1}{2}}\beta_{-k}^\star \right\|_{\psi_2}$$
$$\le \left[(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star\right]^{-\frac{1}{2}} \sigma \left\| \Sigma_{S,-k}^{\frac{1}{2}}\beta_{-k}^\star \right\|$$
$$= \sigma.$$

By applying Lemma 9 for the sequence $\left\{ \left[(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star\right]^{-1/2} X_{-k}[i,*]\beta_{-k}^\star \right\}_{i=1}^{n}$, there exists an absolute constant c, for any $t \in (0, n)$, with probability at least $1 - 2e^{-ct}$:

$$\frac{(\beta_{-k}^\star)^T X_{-k}^T X_{-k}\beta_{-k}^\star}{(\beta_{-k}^\star)^T \Sigma_{S,-k}\beta_{-k}^\star} \le n + \sqrt{nt}\sigma^2.$$

Let $t = n/4$, with probability at least $1 - 2e^{-cn/4}$:

$$(\beta_{-k}^\star)^T X_{-k}^T X_{-k}\beta_{-k}^\star \le (1 + \frac{1}{2}\sigma^2)n \cdot (\beta_{-k}^\star)^T \Sigma_{S,-k}\beta_{-k}^\star.$$

By taking $c_x = \max\left\{1 + \frac{1}{2}\sigma^2, \frac{4}{c}\right\}$, the proof is complete. □

A.2 Block decomposition of $X_{-k}X_{-k}^T$

Let $X_k = U\widetilde{M}^{\frac{1}{2}}V$, where $U \in \mathbb{R}^{n\times n}$ and $V \in \mathbb{R}^{d\times d}$ are orthogonal matrices representing the left and right singular vectors, respectively. The matrix $\widetilde{M}^{\frac{1}{2}}$ is defined as:

$$\widetilde{M}^{\frac{1}{2}} = \begin{pmatrix} m_1^{\frac{1}{2}} & & \\ & \ddots & \\ & & m_k^{\frac{1}{2}} \\ & \mathbf{0}^{(n-k)\times k} & \end{pmatrix} \in \mathbb{R}^{n\times k}.$$

Therefore, we have $X_k X_k^T = UMU^T$, where $M = \mathrm{diag}(m_1, ..., m_k, 0, ..., 0) \in \mathbb{R}^{n\times n}$. Similarly, $X_k^T X_k = V^T M_k V$, where $M_k = \mathrm{diag}(m_1, ..., m_k) \in \mathbb{R}^{k\times k}$.

Let $\Delta = U^T X_{-k} X_{-k}^T U$, and write Δ in block matrix form as:

$$\Delta = \begin{pmatrix} \Delta_{11} & \Delta_{12} \\ \Delta_{12}^T & \Delta_{22} \end{pmatrix},$$

where $\Delta_{11} \in \mathbb{R}^{k\times k}$, $\Delta_{12} \in \mathbb{R}^{k\times(n-k)}$, and $\Delta_{22} \in \mathbb{R}^{(n-k)\times(n-k)}$.

We will repeatedly use the first k rows of $(M + \lambda I_n + \Delta)^{-1}$, which we compute here. Because $M + \lambda I_n + \Delta$ and $\lambda I_{n-k} + \Delta_{22}$ are invertible when A_k is positive definite, by block matrix inverse,

$$\begin{aligned} &(M + \lambda I_n + \Delta)^{-1}[k,*] \\ =& \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right)^{-1} \left(I_k, -\Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\right). \end{aligned} \tag{5}$$

Corollary 16 (Corollary of Lemma 10)**.** There exists a constant depending only on σ, such that for any $n < \lambda_{k+1}^{-1}\left(\lambda + \sum_{j>k}\lambda_j\right)$, if the assumption condNum(k, δ, L) is satisfied, the following inequalities hold with probability at least $1 - \delta - c_x e^{-n/c_x}$, on the same event as in Lemma 10.

$$\|\Delta_{11}\|, \|\Delta_{12}\| \le \|\Delta\| \le c_x \left(\lambda + \sum_{j>k}\lambda_j\right).$$

$$\|(\lambda I_{n-k} + \Delta_{22})^{-1}\| \le \|\Delta^{-1}\| \le c_x L \left(\lambda + \sum_{j>k}\lambda_j\right)^{-1}.$$

$$\left\|\Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-2}\Delta_{12}^T\right\| \le c_x^4 L^2.$$

$$\left\|\Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right\| \le c_x^3 L \left(\lambda + \sum_{j>k}\lambda_j\right).$$

$$\left\|\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right\| \le c_x \left(\lambda + \sum_{j>k}\lambda_j\right).$$

Proof.

1. The first inequality.

$$\|\Delta_{11}\|, \|\Delta_{12}\| \le \|\Delta\| = \|X_{-k}X_{-k}^T\| \le \|A_k\| \le c_x \left(\lambda + \sum_{j>k}\lambda_j\right).$$

2. The second inequality.

$$\|(\lambda I_{n-k} + \Delta_{22})^{-1}\| \le \|(\lambda I_n + \Delta)^{-1}\| = \|A_k^{-1}\| \le c_x L \left(\lambda + \sum_{j>k}\lambda_j\right)^{-1},$$

where the first inequality holds because $\lambda I_n + \Delta$ is positive definite.

3. The third inequality.

$$\left\|\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-2}\Delta_{12}^T\right\| \leq \|\Delta_{12}\|^2\|(\lambda I_{n-k}+\Delta_{22})^{-1}\|^2 \leq c_x^4 L^2.$$

4. The fourth inequality.

$$\left\|\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right\| \leq \|\Delta_{12}\|^2\|(\lambda I_{n-k}+\Delta_{22})^{-1}\| \leq c_x^3 L\left(\lambda+\sum_{j>k}\lambda_j\right).$$

5. The last inequality.

$$\begin{aligned}
&\left\|\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right\| \\
&= \left\|\Delta_{11}+\lambda I_k-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right\|-\lambda \\
&\leq \|\Delta_{11}+\lambda I_k\|-\lambda \\
&= \|\Delta_{11}\| \\
&\leq c_x\left(\lambda+\sum_{j>k}\lambda_j\right).
\end{aligned}$$

The first inequality holds because $\Delta_{11}+\lambda I_k-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T$ is the Schur complement of the block $\Delta_{11}+\lambda I_k$ of the matrix $\Delta+\lambda I_n$, which is positive definite. Therefore, we have

$$\Delta_{11}+\lambda I_k \succcurlyeq \Delta_{11}+\lambda I_k-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T.$$

$\square$

Lemma 17. There exists a constant $c_x > 2$ depending only on σ, such that for any $N_1 < n < N_2$, if the assumption condNum(k,δ,L) is satisfied, the following holds with probability at least $1-2\delta-c_xe^{-n/c_x}$, on both events from Lemma 10 and Lemma 11,

$$\begin{aligned}
&\left\|\left[X_k^TX_k+\lambda I_k+V^T\left(\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right]^{-1}-\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&\leq \frac{c_x^2\left(\sqrt{n(k+\ln\frac{1}{\delta})}\lambda_1+c_x^2L\left(\lambda+\sum_{j>k}\lambda_j\right)\right)}{(\lambda+n\lambda_k)^2}.
\end{aligned}$$

where

$$N_1 = \max\left\{4c_x^4(k+\ln(1/\delta))\frac{\lambda_1^2}{\lambda_k^2}, 2c_x^4L\lambda_k^{-1}\left(\lambda+\sum_{j>k}\lambda_j\right)\right\}.$$

$$N_2 = \frac{1}{\lambda_{k+1}}\left(\lambda+\sum_{j>k}\lambda_j\right).$$

Proof.

$$\begin{aligned}
&\left\|\left[X_k^TX_k+\lambda I_k+V^T\left(\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right]^{-1}-\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&\leq \left\|\left[X_k^TX_k+\lambda I_k+V^T\left(\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right]^{-1}\right\| \\
&\quad\cdot\left\|\left[X_k^TX_k+\lambda I_k+V^T\left(\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right]-\left(n\widetilde{\Sigma}_{S,k}\right)\right\| \\
&\quad\cdot\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&= \frac{1}{\lambda+n\lambda_k}\left\|\left[X_k^TX_k+\lambda I_k+V^T\left(\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right]^{-1}\right\|
\end{aligned}$$

$$\cdot \left\| X_k^T X_k - n\Sigma_{S,k} + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V \right\|.$$

According to Lemma 11, Corollary 16, there exists a constant $c_x > 2$ depending only on σ, such that for any $k + \ln(1/\delta) < N_1 < n < N_2 = \lambda_{k+1}^{-1}(\lambda + \sum_{j>k} \lambda_j)$, with probability at least $1 - 2\delta - c_x e^{-n/c_x}$, on both events in Lemma 10 and Lemma 11,

$$\left\| \frac{1}{n} X_k^T X_k - \Sigma_{S,k} \right\| \le c_x \lambda_1 \sqrt{\frac{k + \ln\frac{1}{\delta}}{n}}.$$

$$\left\| \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T \right\| \le c_x^3 L \left(\lambda + \sum_{j>k} \lambda_j \right).$$

1. $\left\| X_k^T X_k - n\Sigma_{S,k} + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V \right\|$.

$$\begin{aligned}
&\left\| X_k^T X_k - n\Sigma_{S,k} + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V \right\| \\
&\le \left\| X_k^T X_k - n\Sigma_{S,k} \right\| + \left\| \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) \right\| \\
&\le c_x \sqrt{n(k + \ln\frac{1}{\delta})}\lambda_1 + c_x^3 L \left(\lambda + \sum_{j>k} \lambda_j \right).
\end{aligned}$$

2. $\left\| \left[X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V \right]^{-1} \right\|$

$$\begin{aligned}
&\frac{1}{\lambda + n\lambda_k} \left\| X_k^T X_k - n\Sigma_{S,k} + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V \right\| \\
&\le \frac{1}{\lambda + n\lambda_k} \left(c_x \sqrt{n(k + \ln\frac{1}{\delta})}\lambda_1 + c_x^3 L \left(\lambda + \sum_{j>k} \lambda_j \right) \right).
\end{aligned}$$

Since $n > 4c_x^4 (k + \ln(1/\delta)) \frac{\lambda_1^2}{\lambda_k^2}$,

$$\begin{aligned}
\frac{1}{\lambda + n\lambda_k} c_x \sqrt{n(k + \ln\frac{1}{\delta})}\lambda_1 &\le \frac{c_x \sqrt{n(k + \ln\frac{1}{\delta})}\lambda_1}{n\lambda_k} \\
&= \frac{c_x \sqrt{(k + \ln\frac{1}{\delta})}\lambda_1}{\sqrt{n}\lambda_k} \\
&< \frac{1}{2c_x}.
\end{aligned}$$

Since $n > 2c_x^4 L \lambda_k^{-1}(\lambda + \sum_{j>k} \lambda_j)$,

$$\begin{aligned}
\frac{1}{\lambda + n\lambda_k} c_x^3 L \left(\lambda + \sum_{j>k} \lambda_j \right) &\le \frac{c_x^3 L \left(\lambda + \sum_{j>k} \lambda_j \right)}{n\lambda_k} \\
&< \frac{1}{2c_x}.
\end{aligned}$$

Therefore, we have

$$\frac{1}{\lambda + n\lambda_k} \left\| X_k^T X_k - n\Sigma_{S,k} + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V \right\| < \frac{1}{c_x}.$$

Now we derive the upper bound for our target.

$$\begin{aligned}
&\left\| \left[X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V \right]^{-1} \right\| \\
&= \left\| \left[n\widetilde{\Sigma}_{S,k} + X_k^T X_k - n\Sigma_{S,k} + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V \right]^{-1} \right\|
\end{aligned}$$

$$
\begin{aligned}
&\leq \left\| \left(n\widetilde{\Sigma}_{S,k}\right)^{-1} \right\| \left[1 - \left\| \left(n\widetilde{\Sigma}_{S,k}\right)^{-1} \right\| \right. \\
&\qquad \left. \cdot \left\| \left[X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T \right) V \right]^{-1} \right\| \right]^{-1} \\
&\leq \frac{1}{\lambda + n\lambda_k} \left(1 - \frac{1}{c_x} \right)^{-1} \\
&\leq \frac{c_x}{\lambda + n\lambda_k}.
\end{aligned}
$$

The first inequality follows from the result $\|(A+T)^{-1}\| \leq \|A^{-1}\| \left(1 - \|A^{-1}\|\|T\|\right)^{-1}$, provided that both A and $A+T$ are invertible and $\|A^{-1}\|\|T\| < 1$ (see Lemma 3.1 in Wedin (1973)).

Combining the above two inequalities,

$$
\begin{aligned}
&\left\| \left[X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T \right) V \right]^{-1} - \left(n\widetilde{\Sigma}_{S,k}\right)^{-1} \right\| \\
\leq\ & \frac{1}{\lambda + n\lambda_k} \frac{c_x}{\lambda + n\lambda_k} \left(c_x \sqrt{n(k + \ln\frac{1}{\delta})} \lambda_1 + c_x^3 L \left(\lambda + \sum_{j>k} \lambda_j \right) \right) \\
=\ & \frac{c_x^2 \left(\sqrt{n(k + \ln\frac{1}{\delta})} \lambda_1 + c_x^2 L \left(\lambda + \sum_{j>k} \lambda_j \right) \right)}{(\lambda + n\lambda_k)^2}.
\end{aligned}
$$

□

A.3 Bias variance decomposition

We consider the expection of the excess risk $\mathcal{R}\left(\widehat{\beta}(Y)\right) = \mathcal{R}\left(\widehat{\beta}(X\beta^\star) + \widehat{\beta}(\boldsymbol{\epsilon})\right)$ with respect to the distribution of the noise $\boldsymbol{\epsilon}$.

$$
\begin{aligned}
\mathbb{E}_{\boldsymbol{\epsilon}}\left[\mathcal{R}\left(\widehat{\beta}(Y)\right) \right] &= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\left(\widehat{\beta}(Y) - \beta^\star \right)^T \Sigma_T \left(\widehat{\beta}(Y) - \beta^\star \right) \right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\beta}(\boldsymbol{\epsilon})^T \Sigma_T \widehat{\beta}(\boldsymbol{\epsilon}) \right] + \left(\widehat{\beta}(X\beta^\star) - \beta^\star \right)^T \Sigma_T \left(\widehat{\beta}(X\beta^\star) - \beta^\star \right).
\end{aligned}
$$

We decompose the expected excess risk into variance and bias terms.

$$
\begin{aligned}
V &= \mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\beta}(\boldsymbol{\epsilon})^T \Sigma_T \widehat{\beta}(\boldsymbol{\epsilon}) \right] \\
&\leq 2\mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\beta}(\boldsymbol{\epsilon})_k^T \Sigma_{T,k} \widehat{\beta}(\boldsymbol{\epsilon})_k \right] + 2\mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\beta}(\boldsymbol{\epsilon})_{-k}^T \Sigma_{T,-k} \widehat{\beta}(\boldsymbol{\epsilon})_{-k} \right]. \\
B &= \left(\widehat{\beta}(X\beta^\star) - \beta^\star \right)^T \Sigma_T \left(\widehat{\beta}(X\beta^\star) - \beta^\star \right) \\
&\leq 2 \left(\widehat{\beta}(X\beta^\star)_k - \beta_k^\star \right)^T \Sigma_{T,k} \left(\widehat{\beta}(X\beta^\star)_k - \beta_k^\star \right) \\
&\quad + 2 \left(\widehat{\beta}(X\beta^\star)_{-k} - \beta_{-k}^\star \right)^T \Sigma_{T,-k} \left(\widehat{\beta}(X\beta^\star)_{-k} - \beta_{-k}^\star \right).
\end{aligned}
$$

The inequalities follow from the result for a positive definite block quadratic form:

$$
(x_1^T, x_2^T) \begin{pmatrix} A & B \\ B^T & D \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1^T A x_1 + 2x_1^T B x_2 + x_1^T D x_1,
$$

where the positive definiteness implies $x_1^T A x_1 + x_1^T D x_1 \geq 2x_1^T B x_2$.

Lemma 18. There exists a constant $c_x > 2$ depending only on σ, such that for any $N_1 < n < N_2$, if the assumption condNum(k, δ, L) (Assumption 2) is satisfied, then with probability at least $1 - 2\delta - c_x e^{-n/c_x}$, the following inequalities hold simultaneously:

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

$$\mu_1(A_k) \le c_x \left(\lambda + \sum_{j>k}\lambda_j\right).$$

$$\mu_1(X_{-k}X_{-k}^T) \le c_x \left(n\lambda_{k+1} + \sum_{j>k}\lambda_j\right).$$

$$\left\|\frac{1}{n}X_k^T X_k - \Sigma_{S,k}\right\| \le c_x\lambda_1\sqrt{\frac{k+\ln\frac{1}{\delta}}{n}}.$$

$$\left\|\left(X_k^T X_k\right)^{\frac{1}{2}} - \sqrt{n}\Sigma_{S,k}^{\frac{1}{2}}\right\| \le c_x\sqrt{k+\ln\frac{1}{\delta}}\lambda_1\lambda_k^{-\frac{1}{2}}.$$

$$\mu_k\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) \ge \frac{1}{c_x}n.$$

$$\mu_1\left(\Sigma_{S,k}^{-\frac{1}{2}}X_k^T X_k\Sigma_{S,k}^{-\frac{1}{2}}\right) \le c_x n.$$

$$\mu_1(X_k^T X_k) \le c_x\lambda_1 n.$$

$$\mu_k(X_k^T X_k) \ge \frac{1}{c_x}\lambda_k n.$$

$$\operatorname{tr}\left(X_{-k}\Sigma_{T,-k}X_{-k}^T\right) \le c_x n \operatorname{tr}\left(\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}\right).$$

$$(\beta_{-k}^\star)^T X_{-k}^T X_{-k}\beta_{-k}^\star \le c_x n(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star.$$

$$\|\Delta_{11}\|, \|\Delta_{12}, \|\Delta\|\| \le c_x\left(\lambda + \sum_{j>k}\lambda_j\right).$$

$$\|(\lambda I_{n-k} + \Delta_{22})^{-1}\|, \|\Delta^{-1}\| \le c_x L\left(\lambda + \sum_{j>k}\lambda_j\right)^{-1}.$$

$$\left\|\Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-2}\Delta_{12}^T\right\| \le c_x^4 L^2.$$

$$\left\|\Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right\| \le c_x^3 L\left(\lambda + \sum_{j>k}\lambda_j\right).$$

$$\left\|\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right\| \le c_x\left(\lambda + \sum_{j>k}\lambda_j\right).$$

And,

$$\left\|\left[X_k^T X_k + \lambda I_k + V^T\left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right)V\right]^{-1} - \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\|$$
$$\le \frac{c_x^2\left(\sqrt{n(k+\ln\frac{1}{\delta})}\lambda_1 + c_x^2 L\left(\lambda + \sum_{j>k}\lambda_j\right)\right)}{(\lambda + n\lambda_k)^2}.$$

N_1 and N_2 are defined as follows:

$$N_1 = \max\left\{4c_x^4(k+\ln(1/\delta))\frac{\lambda_1^2}{\lambda_k^2}, 2c_x^4 L\lambda_k^{-1}\left(\lambda + \sum_{j>k}\lambda_j\right)\right\}.$$

$$N_2 = \frac{1}{\lambda_{k+1}}\left(\lambda + \sum_{j>k}\lambda_j\right).$$

Proof. The lemma is a direct corollary from Lemma 10, Lemma 11, Corollary 12, Lemma 13, Lemma 14, Lemma 15, Corollary 16, Lemma 17. ☐

A.3.1 Variance in the first k dimensions

Lemma 19. Under the same conditions as in Lemma 18, and on the same event, for any $N_1 < n < N_2$,

$$\mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\beta}(\boldsymbol{\epsilon})_k^T \Sigma_{T,k} \widehat{\beta}(\boldsymbol{\epsilon})_k\right] \leq 16 v^2 (1 + c_x^4 L^2) \frac{1}{n} \operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right],$$

where

$$\begin{aligned} N_1 = \max \Bigg\{ & 4c_x^4 \left(k + \ln \frac{1}{\delta}\right) \lambda_1^4 \lambda_k^{-4}, \\ & 2c_x^4 L \lambda_1 \lambda_k^{-2} \left(\lambda + \sum_{j>k} \lambda_j\right), \\ & 4c_x^4 \left(k + \ln \frac{1}{\delta}\right) \lambda_1^6 \lambda_k^{-8} \|\Sigma_{T,k}\|^2 k^2 \left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-2}, \\ & 2c_x^4 L \lambda_1^2 \lambda_k^{-4} \left(\lambda + \sum_{j>k} \lambda_j\right) \|\Sigma_{T,k}\| k \left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-1} \Bigg\}, \\ N_2 = & \frac{1}{\lambda_{k+1}} \left(\lambda + \sum_{j>k} \lambda_j\right). \end{aligned}$$

Proof.

$$\begin{aligned}
& \mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\beta}(\boldsymbol{\epsilon})_k^T \Sigma_{T,k} \widehat{\beta}(\boldsymbol{\epsilon})_k\right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \operatorname{tr}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T (XX^T + \lambda I_n)^{-1} X_k \Sigma_{T,k} X_k^T (XX^T + \lambda I_n)^{-1}\right] \\
&= v^2 \operatorname{tr}\left[(XX^T + \lambda I_n)^{-1} X_k \Sigma_{T,k} X_k^T (XX^T + \lambda I_n)^{-1}\right] \\
&= v^2 \operatorname{tr}\Big[(UMU^T + U\Delta U^T + \lambda I_n)^{-1} U \widetilde{M}^{\frac{1}{2}} V \Sigma_{T,k} \\
&\quad \cdot V^T \left(\widetilde{M}^{\frac{1}{2}}\right)^T U^T (UMU^T + U\Delta U^T + \lambda I_n)^{-1}\Big] \\
&= v^2 \operatorname{tr}\left[U(M + \Delta + \lambda I_n)^{-1} \widetilde{M}^{\frac{1}{2}} V \Sigma_{T,k} V^T \left(\widetilde{M}^{\frac{1}{2}}\right)^T (M + \Delta + \lambda I_n)^{-1} U^T\right] \\
&= v^2 \operatorname{tr}\left[\left(\widetilde{M}^{\frac{1}{2}}\right)^T (M + \Delta + \lambda I_n)^{-1} (M + \Delta + \lambda I_n)^{-1} \widetilde{M}^{\frac{1}{2}} V \Sigma_{T,k} V^T\right] \\
&= v^2 \operatorname{tr}\Big[M_k^{\frac{1}{2}} \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1} \left(I_k, -\Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\right) \\
&\quad \cdot \left(I_k, -\Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\right)^T \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1} M_k^{\frac{1}{2}} \\
&\quad \cdot V \Sigma_{T,k} V^T\Big] \\
&= v^2 \operatorname{tr}\Big[M_k^{\frac{1}{2}} \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1} \\
&\quad \cdot \left(I_k + \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-2} \Delta_{12}^T\right) \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1} M_k^{\frac{1}{2}} \\
&\quad \cdot V \Sigma_{T,k} V^T\Big] \\
&= v^2 \operatorname{tr}\Big[\left(I_k + \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-2} \Delta_{12}^T\right) \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1} \\
&\quad \cdot M_k^{\frac{1}{2}} V \Sigma_{T,k} V^T M_k^{\frac{1}{2}} \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1}\Big] \\
&\leq v^2 \left\|I_k + \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-2} \Delta_{12}^T\right\| \operatorname{tr}\Big[\left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1} \\
&\quad \cdot M_k^{\frac{1}{2}} V \Sigma_{T,k} V^T M_k^{\frac{1}{2}} \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1}\Big]
\end{aligned}$$

$$
\begin{aligned}
&\leq v^2(1+c_x^4L^2)\operatorname{tr}\Big[\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)^{-1}\\
&\cdot M_k^{\frac{1}{2}}V\Sigma_{T,k}V^TM_k^{\frac{1}{2}}\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)^{-1}\Big]\\
&=v^2(1+c_x^4L^2)\operatorname{tr}\Big[\left(V^T\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right)^{-1}\\
&\cdot V^TM_k^{\frac{1}{2}}V\cdot\Sigma_{T,k}\cdot V^TM_k^{\frac{1}{2}}V\cdot\left(V^T\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right)^{-1}\Big]\\
&=v^2(1+c_x^4L^2)\operatorname{tr}\Big[\left(X_k^TX_k+\lambda I_k+V^T\left(\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right)^{-1}\\
&\cdot\left(X_k^TX_k\right)^{\frac{1}{2}}\Sigma_{T,k}\left(X_k^TX_k\right)^{\frac{1}{2}}\\
&\cdot\left(X_k^TX_k+\lambda I_k+V^T\left(\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right)^{-1}\Big].
\end{aligned}
$$

The sixth equation follows from Equation 5. The first inequality follows from the result $\operatorname{tr}[AB]\leq\|A\|\operatorname{tr}[B]$ where the matrix B is positive semi-definite.

We define two quantities that represent concentration error terms:

$$
\begin{aligned}
E_1&=\left\|\left[X_k^TX_k+\lambda I_k+V^T\left(\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)V\right]^{-1}-\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\|.\\
E_2&=\left(X_k^TX_k\right)^{\frac{1}{2}}-\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}.
\end{aligned}
$$

Since $n>4c_x^4\left(k+\ln\frac{1}{\delta}\right)\lambda_1^6\lambda_k^{-8}\|\Sigma_{T,k}\|^2k^2\left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-2}$,

and $n>2c_x^4L\left(\lambda+\sum_{j>k}\lambda_j\right)\lambda_1^2\lambda_k^{-4}\|\Sigma_{T,k}\|k\left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-1}$,

$$
\begin{aligned}
&\|E_1\|\left\|n\widetilde{\Sigma}_{S,k}\right\|\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\|\left\|(n\Sigma_{S,k})^{\frac{1}{2}}\right\|\|\Sigma_{T,k}\|\left\|(n\Sigma_{S,k})^{\frac{1}{2}}\right\|\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\|\\
&\leq\frac{c_x^2\left(\sqrt{n(k+\ln\frac{1}{\delta})}\lambda_1+c_x^2L\left(\lambda+\sum_{j>k}\lambda_j\right)\right)}{(\lambda+n\lambda_k)^2}(\lambda+n\lambda_1)\frac{n\lambda_1}{(\lambda+n\lambda_k)^2}\|\Sigma_{T,k}\|\\
&\leq\frac{c_x^2\left(\sqrt{n(k+\ln\frac{1}{\delta})}\lambda_1+c_x^2L\left(\lambda+\sum_{j>k}\lambda_j\right)\right)}{n^2}\frac{\lambda_1^2}{\lambda_k^4}\|\Sigma_{T,k}\|\\
&=\frac{c_x^2\sqrt{(k+\ln\frac{1}{\delta})}}{n\sqrt{n}}\frac{\lambda_1^3}{\lambda_k^4}\|\Sigma_{T,k}\|+\frac{c_x^4L\left(\lambda+\sum_{j>k}\lambda_j\right)}{n^2}\frac{\lambda_1^2}{\lambda_k^4}\|\Sigma_{T,k}\|\\
&<\frac{1}{2nk}\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]+\frac{1}{2nk}\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\\
&=\frac{1}{nk}\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right].
\end{aligned}
$$

Since $n>4c_x^4\left(k+\ln\frac{1}{\delta}\right)\lambda_1^4\lambda_k^{-4}$ and $n>2c_x^4L\left(\lambda+\sum_{j>k}\lambda_j\right)\lambda_1\lambda_k^{-2}$,

$$
\begin{aligned}
&\|E_1\|\left\|n\widetilde{\Sigma}_{S,k}\right\|\\
&\leq\frac{c_x^2\left(\sqrt{n(k+\ln\frac{1}{\delta})}\lambda_1+c_x^2L\left(\lambda+\sum_{j>k}\lambda_j\right)\right)}{(\lambda+n\lambda_k)^2}(\lambda+n\lambda_1)\\
&\leq\frac{c_x^2\left(\sqrt{n(k+\ln\frac{1}{\delta})}\lambda_1+c_x^2L\left(\lambda+\sum_{j>k}\lambda_j\right)\right)}{n}\frac{\lambda_1}{\lambda_k^2}\\
&=\frac{c_x^2\sqrt{(k+\ln\frac{1}{\delta})}}{\sqrt{n}}\frac{\lambda_1^2}{\lambda_k^2}+\frac{c_x^4L\left(\lambda+\sum_{j>k}\lambda_j\right)}{n}\frac{\lambda_1}{\lambda_k^2}\\
&<\frac{1}{2}+\frac{1}{2}\\
&=1.
\end{aligned}
\tag{6}
$$

Since $n > c_x^2 \left(k + \ln \frac{1}{\delta}\right) \lambda_1^4 \lambda_k^{-6} \|\Sigma_{T,k}\|^2 k^2 \left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-2}$,

$$
\begin{aligned}
&\|E_2\| \left\| \left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}} \right\| \left\| \left(n\widetilde{\Sigma}_{S,k}\right)^{-1} \right\| \left\| (n\Sigma_{S,k})^{\frac{1}{2}} \right\| \|\Sigma_{T,k}\| \left\| (n\Sigma_{S,k})^{\frac{1}{2}} \right\| \left\| \left(n\widetilde{\Sigma}_{S,k}\right)^{-1} \right\| \\
&\le c_x \sqrt{k + \ln \frac{1}{\delta}} \lambda_1 \lambda_k^{-\frac{1}{2}} (n\lambda_k)^{-\frac{1}{2}} \frac{n\lambda_1}{(\lambda + n\lambda_k)^2} \|\Sigma_{T,k}\| \\
&\le \frac{c_x \sqrt{k + \ln \frac{1}{\delta}}}{n\sqrt{n}} \frac{\lambda_1^2}{\lambda_k^3} \|\Sigma_{T,k}\| \\
&\le \frac{1}{nk} \operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right].
\end{aligned}
$$

Since $n > c_x^2 \left(k + \ln \frac{1}{\delta}\right) \lambda_1^2 \lambda_k^{-2}$,

$$
\begin{aligned}
&\|E_2\| \left\| \left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}} \right\| \\
&\le c_x \sqrt{k + \ln \frac{1}{\delta}} \lambda_1 \lambda_k^{-\frac{1}{2}} (n\lambda_k)^{-\frac{1}{2}} \\
&= \frac{c_x \sqrt{k + \ln \frac{1}{\delta}}}{\sqrt{n}} \frac{\lambda_1}{\lambda_k} \\
&< 1.
\end{aligned} \tag{7}
$$

Combing the above four inequalities, we have

$$
\begin{aligned}
&\operatorname{tr}\Big[\left(X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) V\right)^{-1} \\
&\quad \cdot \left(X_k^T X_k\right)^{\frac{1}{2}} \Sigma_{T,k} \left(X_k^T X_k\right)^{\frac{1}{2}} \\
&\quad \cdot \left(X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) V\right)^{-1}\Big] \\
&= \operatorname{tr}\left[\left(n\widetilde{\Sigma}_{S,k}\right)^{-1} (n\Sigma_{S,k})^{\frac{1}{2}} \Sigma_{T,k} (n\Sigma_{S,k})^{\frac{1}{2}} \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right] \\
&\quad + 2\operatorname{tr}\left[E_1 (n\Sigma_{S,k})^{\frac{1}{2}} \Sigma_{T,k} (n\Sigma_{S,k})^{\frac{1}{2}} \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right] \\
&\quad + 2\operatorname{tr}\left[\left(n\widetilde{\Sigma}_{S,k}\right)^{-1} E_2 \Sigma_{T,k} (n\Sigma_{S,k})^{\frac{1}{2}} \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right] \\
&\quad + \operatorname{tr}\left[E_1 (n\Sigma_{S,k})^{\frac{1}{2}} \Sigma_{T,k} (n\Sigma_{S,k})^{\frac{1}{2}} E_1\right] \\
&\quad + \operatorname{tr}\left[\left(n\widetilde{\Sigma}_{S,k}\right)^{-1} E_2 \Sigma_{T,k} E_2 \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right] \\
&\quad + 2\operatorname{tr}\left[E_1 (n\Sigma_{S,k})^{\frac{1}{2}} \Sigma_{T,k} E_2 \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right] \\
&\quad + 2\operatorname{tr}\left[E_1 E_2 \Sigma_{T,k} (n\Sigma_{S,k})^{\frac{1}{2}} \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right] \\
&\quad + 2\operatorname{tr}\left[E_1 E_2 \Sigma_{T,k} (n\Sigma_{S,k})^{\frac{1}{2}} E_1\right] \\
&\quad + 2\operatorname{tr}\left[E_1 E_2 \Sigma_{T,k} E_2 \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right] \\
&\quad + \operatorname{tr}\left[E_1 E_2 \Sigma_{T,k} E_2 E_1\right].
\end{aligned}
$$

In particular,

$$
\operatorname{tr}\left[\left(n\widetilde{\Sigma}_{S,k}\right)^{-1} (n\Sigma_{S,k})^{\frac{1}{2}} \Sigma_{T,k} (n\Sigma_{S,k})^{\frac{1}{2}} \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right]
$$

$$
\begin{aligned}
&= \frac{1}{n} \operatorname{tr}\left[\widetilde{\Sigma}_{S,k}^{-1} \Sigma_{S,k}^{\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{\frac{1}{2}} \widetilde{\Sigma}_{S,k}^{-1}\right] \\
&\le \frac{1}{n} \operatorname{tr}\left[\Sigma_{S,k}^{-1} \Sigma_{S,k}^{\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{\frac{1}{2}} \Sigma_{S,k}^{-1}\right] \\
&= \frac{1}{n} \operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right].
\end{aligned}
$$

The inequality follows from the fact that $\operatorname{tr}[BAB] = \operatorname{tr}[A^{\frac{1}{2}} B A^{\frac{1}{2}}] \le \operatorname{tr}[A^{\frac{1}{2}} C A^{\frac{1}{2}}] = \operatorname{tr}[CAC]$, where A, B, C are positive semi-definite matrices, and $C \succcurlyeq B$, which implies that $A^{\frac{1}{2}} C A^{\frac{1}{2}} \succcurlyeq A^{\frac{1}{2}} B A^{\frac{1}{2}}$.

$$
\begin{aligned}
&\operatorname{tr}\left[E_1 E_2 \Sigma_{T,k} E_2 E_1\right] \\
&= \operatorname{tr}\Bigg[E_1 n\widetilde{\Sigma}_{S,k} \left(n\widetilde{\Sigma}_{S,k}\right)^{-1} E_2 \left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}} \left(n\widetilde{\Sigma}_{S,k}\right)^{\frac{1}{2}} \\
&\quad \cdot \Sigma_{T,k} E_2 \left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}} \left(n\widetilde{\Sigma}_{S,k}\right)^{\frac{1}{2}} E_1 n\widetilde{\Sigma}_{S,k} \left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\Bigg] \\
&\le k \left(\|E_1\| \left\|n\widetilde{\Sigma}_{S,k}\right\|\right)^2 \left(\|E_2\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}}\right\|\right) \\
&\quad \cdot \|E_2\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}}\right\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \left\|(n\Sigma_{S,k})^{\frac{1}{2}}\right\| \|\Sigma_{T,k}\| \left\|(n\Sigma_{S,k})^{\frac{1}{2}}\right\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&\le \frac{1}{n} \operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right].
\end{aligned}
$$

The other terms can be similarly bounded. Therefore,

$$
\begin{aligned}
&\operatorname{tr}\Big[\left(X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) V\right)^{-1} \\
&\quad \cdot \left(X_k^T X_k\right)^{\frac{1}{2}} \Sigma_{T,k} \left(X_k^T X_k\right)^{\frac{1}{2}} \\
&\quad \cdot \left(X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) V\right)^{-1}\Big] \\
&\le \frac{16}{n} \operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right].
\end{aligned}
$$

The proof is complete by combing all the inequalities above. □

A.3.2 Variance in the last $d - k$ dimensions

Lemma 20. Under the same conditions as in Lemma 18, and on the same event, for any $N_1 < n < N_2$,

$$
\mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\beta}(\boldsymbol{\epsilon})_{-k}^T \Sigma_{T,-k} \widehat{\beta}(\boldsymbol{\epsilon})_{-k}\right] \le v^2 c_x^3 L^2 n \left(\lambda + \sum_{j>k} \lambda_j\right)^{-2} \operatorname{tr}\left[\Sigma_{S,-k}^{\frac{1}{2}} \Sigma_{T,-k} \Sigma_{S,-k}^{\frac{1}{2}}\right].
$$

where N_1, N_2 are defined as in Lemma 18.

Proof.

$$
\begin{aligned}
&\mathbb{E}_{\boldsymbol{\epsilon}}\left[\widehat{\beta}(\boldsymbol{\epsilon})_{-k}^T \Sigma_{T,-k} \widehat{\beta}(\boldsymbol{\epsilon})_{-k}\right] \\
&= \mathbb{E}_{\boldsymbol{\epsilon}} \operatorname{tr}\left[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T (XX^T + \lambda I_n)^{-1} X_{-k} \Sigma_{T,-k} X_{-k}^T (XX^T + \lambda I_n)^{-1}\right] \\
&= v^2 \operatorname{tr}\left[(XX^T + \lambda I_n)^{-1} X_{-k} \Sigma_{T,-k} X_{-k}^T (XX^T + \lambda I_n)^{-1}\right] \\
&\le v^2 \left\|(XX^T + \lambda I_n)^{-2}\right\| \operatorname{tr}\left[X_{-k} \Sigma_{T,-k} X_{-k}^T\right] \\
&\le v^2 \left\|(X_{-k} X_{-k}^T + \lambda I_n)^{-2}\right\| \operatorname{tr}\left[X_{-k} \Sigma_{T,-k} X_{-k}^T\right] \\
&\le v^2 \left(\frac{1}{c_x L} \left(\lambda + \sum_{j>k} \lambda_j\right)\right)^{-2} c_x n \operatorname{tr}\left[\Sigma_{S,-k}^{\frac{1}{2}} \Sigma_{T,-k} \Sigma_{S,-k}^{\frac{1}{2}}\right]
\end{aligned}
$$

$$= v^2 c_x^3 L^2 n \left(\lambda + \sum_{j>k} \lambda_j\right)^{-2} \operatorname{tr}\left[\Sigma_{S,-k}^{\frac{1}{2}} \Sigma_{T,-k} \Sigma_{S,-k}^{\frac{1}{2}}\right].$$

The first inequality follows from the result $\operatorname{tr}[ABA] = \operatorname{tr}[A^2B] \leq \|A^2\| \operatorname{tr}[B]$ where the matrix B is positive semi-definite. The second inequality follows from $XX^T + \lambda I_n \succcurlyeq X_{-k}X_{-k}^T + \lambda I_n$. $\square$

A.3.3 Bias in the first k dimensions

The bias in the first k dimensions can be decomposed into two terms.

$$\begin{aligned}
&\left(\widehat{\beta}(X\beta^\star)_k - \beta_k^\star\right)^T \Sigma_{T,k} \left(\widehat{\beta}(X\beta^\star)_k - \beta_k^\star\right) \\
&= \left(X_k^T(XX^T + \lambda I_n)^{-1} X\beta^\star - \beta_k^\star\right)^T \Sigma_{T,k} \left(X_k^T(XX^T + \lambda I_n)^{-1} X\beta^\star - \beta_k^\star\right) \\
&\leq 2\left(X_k^T(XX^T + \lambda I_n)^{-1} X_k\beta_k^\star - \beta_k^\star\right)^T \Sigma_{T,k} \left(X_k^T(XX^T + \lambda I_n)^{-1} X_k\beta_k^\star - \beta_k^\star\right) \\
&\quad + 2\left(X_k^T(XX^T + \lambda I_n)^{-1} X_{-k}\beta_{-k}^\star\right)^T \Sigma_{T,k} \left(X_k^T(XX^T + \lambda I_n)^{-1} X_{-k}\beta_{-k}^\star\right).
\end{aligned}$$

The inequality follows from the result $x_1^T A x_1 + x_2^T A x_2 \geq 2x_1^T A x_2$ where A is positive semi-definite.

Lemma 21. Under the same conditions as in Lemma 18, and on the same event, for any $N_1 < n < N_2$,

$$\begin{aligned}
&\left(X_k^T(XX^T + \lambda I_n)^{-1} X_k\beta_k^\star - \beta_k^\star\right)^T \Sigma_{T,k} \left(X_k^T(XX^T + \lambda I_n)^{-1} X_k\beta_k^\star - \beta_k^\star\right) \\
&\leq \frac{16c_x^4}{n^2} \left(\lambda + \sum_{j>k} \lambda_j\right)^2 (\beta_k^\star)^T \Sigma_{S,k}^{-1} \beta_k^\star \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\|.
\end{aligned}$$

where

$$\begin{aligned}
N_1 &= \max\left\{2c_x^3(\lambda + \sum_{j>k}\lambda_j)\lambda_1\lambda_k^{-2},\right. \\
&\qquad 4c_x^4(k + \ln(1/\delta))\lambda_1^2\lambda_k^{-2}, \\
&\qquad \left. 2c_x^4 L \lambda_k^{-1}\left(\lambda + \sum_{j>k}\lambda_j\right)\right\}, \\
N_2 &= \frac{1}{\lambda_{k+1}} \left(\lambda + \sum_{j>k}\lambda_j\right).
\end{aligned}$$

Proof.

$$\begin{aligned}
&\left(X_k^T(XX^T + \lambda I_n)^{-1} X_k\beta_k^\star - \beta_k^\star\right)^T \Sigma_{T,k} \left(X_k^T(XX^T + \lambda I_n)^{-1} X_k\beta_k^\star - \beta_k^\star\right) \\
&= (\beta_k^\star)^T \left(X_k^T(XX^T + \lambda I_n)^{-1} X_k - I_k\right)^T \Sigma_{T,k} \left(X_k^T(XX^T + \lambda I_n)^{-1} X_k - I_k\right) \beta_k^\star \\
&= (\beta_k^\star)^T \Sigma_{S,k}^{-\frac{1}{2}} \left(\Sigma_{S,k}^{\frac{1}{2}} X_k^T(XX^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{\frac{1}{2}} - \Sigma_{S,k}\right)^T \Sigma_{S,k}^{-\frac{1}{2}} \\
&\quad \cdot \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}} \left(\Sigma_{S,k}^{\frac{1}{2}} X_k^T(XX^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{\frac{1}{2}} - \Sigma_{S,k}\right)^T \Sigma_{S,k}^{-\frac{1}{2}} \beta_k^\star \\
&\leq (\beta_k^\star)^T \Sigma_{S,k}^{-1} \beta_k^\star \cdot \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\| \\
&\quad \cdot \left\|\Sigma_{S,k}^{\frac{1}{2}} X_k^T(XX^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{\frac{1}{2}} - \Sigma_{S,k}\right\|^2.
\end{aligned}$$

Subsequently,

$$\left\|\Sigma_{S,k}^{\frac{1}{2}} X_k^T(XX^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{\frac{1}{2}} - \Sigma_{S,k}\right\|$$

$$
\begin{aligned}
&= \left\| \Sigma_{S,k}^{\frac{1}{2}} V^T \left(\widetilde{M}^{\frac{1}{2}}\right)^T U^T U (M + \lambda I_n + \Delta)^{-1} U^T U \widetilde{M}^{\frac{1}{2}} V \Sigma_{S,k}^{\frac{1}{2}} - \Sigma_{S,k} \right\| \\
&= \left\| \Sigma_{S,k}^{\frac{1}{2}} V^T M_k^{\frac{1}{2}} \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right)^{-1} M_k^{\frac{1}{2}} V \Sigma_{S,k}^{\frac{1}{2}} - \Sigma_{S,k} \right\| \\
&= \left\| \Sigma_{S,k}^{\frac{1}{2}} \left(V^T M_k^{-\frac{1}{2}} \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1} \Sigma_{S,k}^{\frac{1}{2}} \right. \\
&\quad \left. - \Sigma_{S,k} \right\| \\
&= \left\| \Sigma_{S,k}^{\frac{1}{2}} \left(I_k + V^T M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1} \Sigma_{S,k}^{\frac{1}{2}} \right. \\
&\quad \left. - \Sigma_{S,k} \right\| \\
&= \left\| \Sigma_{S,k}^{\frac{1}{2}} \left(\left(I_k + V^T M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1} - I_k \right) \right. \\
&\quad \left. \cdot \Sigma_{S,k}^{\frac{1}{2}} \right\| \\
&= \left\| \Sigma_{S,k}^{\frac{1}{2}} V^T M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V \right. \\
&\quad \left. \cdot \left(I_k + V^T M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1} \Sigma_{S,k}^{\frac{1}{2}} \right\| \\
&\leq \left\| \Sigma_{S,k}^{\frac{1}{2}} V^T M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V \Sigma_{S,k}^{\frac{1}{2}} \right\| \\
&\quad + \left\| \Sigma_{S,k}^{\frac{1}{2}} V^T M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V \right. \\
&\quad \left. \cdot \left[\left(I_k + V^T M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1} - I_k \right] \Sigma_{S,k}^{\frac{1}{2}} \right\|.
\end{aligned}
$$

The second equation follows from Equation 5.

We will derive upper bounds for both terms in the last equation above.

1. The first term.

$$
\begin{aligned}
&\left\| \Sigma_{S,k}^{\frac{1}{2}} V^T M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} V \Sigma_{S,k}^{\frac{1}{2}} \right\| \\
&\leq \left\| \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T \right\| \left\| \Sigma_{S,k}^{\frac{1}{2}} V^T M_k^{-1} V \Sigma_{S,k}^{\frac{1}{2}} \right\| \\
&= \left\| \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T \right\| \left\| \left(\Sigma_{S,k}^{-\frac{1}{2}} \left(X_k^T X_k\right)^{-1} \Sigma_{S,k}^{-\frac{1}{2}} \right)^{-1} \right\| \\
&\leq \left(\lambda + c_x \left(\lambda + \sum_{j>k} \lambda_j \right) \right) \frac{c_x}{n} \\
&\leq \frac{2c_x^2}{n} \left(\lambda + \sum_{j>k} \lambda_j \right).
\end{aligned}
$$

   The inequality follows from $c_x > 2$.

2. The second term.

   Since $n > 2c_x^3(\lambda + \sum_{j>k} \lambda_j)\lambda_k^{-1}$,

$$
\begin{aligned}
&\left\| M_k^{-\frac{1}{2}} \left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T\right) M_k^{-\frac{1}{2}} \right\| \\
&\leq \left\| M_k^{-1} \right\| \left\| \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1} \Delta_{12}^T \right\| \\
&\leq \frac{c_x}{n\lambda_k} \cdot 2c_x \left(\lambda + \sum_{j>k} \lambda_j \right)
\end{aligned}
$$

$$< \frac{1}{c_x}.$$

Therefore,

$$\begin{aligned}
&\left\|\left(I_k + V^T M_k^{-\frac{1}{2}}\left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1} - I_k\right\| \\
&\le \left\|\left(I_k + V^T M_k^{-\frac{1}{2}}\left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1}\right\| \\
&\quad \cdot \left\|V^T M_k^{-\frac{1}{2}}\left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right\| \\
&\le \left(1 - \frac{1}{c_x}\right)^{-1} \left\|V^T M_k^{-\frac{1}{2}}\left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right\| \\
&\le c_x \cdot \frac{c_x}{n\lambda_k} \cdot 2c_x \left(\lambda + \sum_{j>k}\lambda_j\right) \\
&= \frac{2c_x^3}{n}\frac{\lambda + \sum_{j>k}\lambda_j}{\lambda_k}.
\end{aligned}$$

The second inequality follows from $\|(A+T)^{-1}\| \le \|A^{-1}\|\left(1 - \|A^{-1}\|\|T\|\right)^{-1}$, where both A and $A+T$ are invertible and $\|A^{-1}\|\|T\| < 1$. Note that $c_x > 2$.

Since $n > 2c_x^3(\lambda + \sum_{j>k}\lambda_j)\lambda_1\lambda_k^{-2}$,

$$\begin{aligned}
&\Bigg\|\Sigma_{S,k}^{\frac{1}{2}} V^T M_k^{-\frac{1}{2}}\left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) M_k^{-\frac{1}{2}} V \\
&\quad \cdot \left[\left(I_k + V^T M_k^{-\frac{1}{2}}\left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1} - I_k\right]\Sigma_{S,k}^{\frac{1}{2}}\Bigg\| \\
&\le \|\Sigma_{S,k}\|\left\|M_k^{-1}\right\|\left\|\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right\| \\
&\quad \cdot \left\|\left(I_k + V^T M_k^{-\frac{1}{2}}\left(\lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) M_k^{-\frac{1}{2}} V\right)^{-1} - I_k\right\| \\
&\le \lambda_1 \cdot \frac{c_x}{n\lambda_k} \cdot 2c_x \left(\lambda + \sum_{j>k}\lambda_j\right) \cdot \frac{2c_x^3}{n}\frac{\lambda + \sum_{j>k}\lambda_j}{\lambda_k} \\
&= \frac{1}{n} \cdot \frac{4c_x^5}{n}\lambda_1\lambda_k^{-2}\left(\lambda + \sum_{j>k}\lambda_j\right)^2 \\
&< \frac{2c_x^2}{n}\left(\lambda + \sum_{j>k}\lambda_j\right).
\end{aligned}$$

Combining both terms above, we have

$$\left\|\Sigma_{S,k}^{\frac{1}{2}} X_k^T (XX^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{\frac{1}{2}} - \Sigma_{S,k}\right\| \le \frac{4c_x^2}{n}\left(\lambda + \sum_{j>k}\lambda_j\right).$$

Therefore,

$$\begin{aligned}
&\left(X_k^T (XX^T + \lambda I_n)^{-1} X_k \beta_k^\star - \beta_k^\star\right)^T \Sigma_{T,k}\left(X_k^T (XX^T + \lambda I_n)^{-1} X_k \beta_k^\star - \beta_k^\star\right) \\
&\le (\beta_k^\star)^T \Sigma_{S,k}^{-1}\beta_k^\star \cdot \left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\| \\
&\quad \cdot \left\|\Sigma_{S,k}^{\frac{1}{2}} X_k^T (XX^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{\frac{1}{2}} - \Sigma_{S,k}\right\|^2
\end{aligned}$$

$$\leq \frac{16c_x^4}{n^2}\left(\lambda+\sum_{j>k}\lambda_j\right)^2 (\beta_k^\star)^T\Sigma_{S,k}^{-1}\beta_k^\star \left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|.$$

$\square$

Lemma 22. Under the same conditions as in Lemma 18, and on the same event, for any $N_1 < n < N_2$,

$$\begin{aligned}&\left(X_k^T(XX^T+\lambda I_n)^{-1}X_{-k}\beta_{-k}^\star\right)^T\Sigma_{T,k}\left(X_k^T(XX^T+\lambda I_n)^{-1}X_{-k}\beta_{-k}^\star\right)\\ &\leq 16c_x(1+c_x^4L^2)\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star.\end{aligned}$$

where

$$\begin{aligned}N_1 = \max\Bigg\{&4c_x^4\left(k+\ln\frac{1}{\delta}\right)\lambda_1^4\lambda_k^{-4},\\ &2c_x^4L\lambda_1\lambda_k^{-2}\left(\lambda+\sum_{j>k}\lambda_j\right),\\ &4c_x^4\left(k+\ln\frac{1}{\delta}\right)\lambda_1^6\lambda_k^{-8}\|\Sigma_{T,k}\|^2\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|^{-2},\\ &2c_x^4L\left(\lambda+\sum_{j>k}\lambda_j\right)\lambda_1^2\lambda_k^{-4}\|\Sigma_{T,k}\|\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|^{-1}\Bigg\},\\ N_2 = \frac{1}{\lambda_{k+1}}&\left(\lambda+\sum_{j>k}\lambda_j\right).\end{aligned}$$

Proof.

$$\begin{aligned}&\left(X_k^T(XX^T+\lambda I_n)^{-1}X_{-k}\beta_{-k}^\star\right)^T\Sigma_{T,k}\left(X_k^T(XX^T+\lambda I_n)^{-1}X_{-k}\beta_{-k}^\star\right)\\ &\leq \left\|(XX^T+\lambda I_n)^{-1}X_k\Sigma_{T,k}X_k^T(XX^T+\lambda I_n)^{-1}\right\|\cdot(\beta_{-k}^\star)^TX_{-k}^TX_{-k}\beta_{-k}^\star.\end{aligned}$$

From Lemma 18,

$$(\beta_{-k}^\star)^TX_{-k}^TX_{-k}\beta_{-k}^\star \leq c_xn(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star.$$

In the following, we derive an upper bound for the other term.

$$\begin{aligned}&\left\|(XX^T+\lambda I_n)^{-1}X_k\Sigma_{T,k}X_k^T(XX^T+\lambda I_n)^{-1}\right\|\\ &=\left\|(M+\lambda I_n+\Delta)^{-1}\widetilde{M}^{\frac{1}{2}}V\Sigma_{T,k}V^T\left(\widetilde{M}^{\frac{1}{2}}\right)^T(M+\lambda I_n+\Delta)^{-1}\right\|\\ &=\left\|\Sigma_{T,k}^{\frac{1}{2}}V^T\left(\widetilde{M}^{\frac{1}{2}}\right)^T(M+\lambda I_n+\Delta)^{-2}\widetilde{M}^{\frac{1}{2}}V\Sigma_{T,k}^{\frac{1}{2}}\right\|\\ &=\Big\|\Sigma_{T,k}^{\frac{1}{2}}V^TM_k^{\frac{1}{2}}\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)^{-1}\\ &\quad\cdot\left(I_k+\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-2}\Delta_{12}^T\right)\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)^{-1}\\ &\quad\cdot M_k^{\frac{1}{2}}V\Sigma_{T,k}^{\frac{1}{2}}\Big\|\\ &\leq\left\|I_k+\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-2}\Delta_{12}^T\right\|\\ &\quad\cdot\Big\|\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)^{-1}M_k^{\frac{1}{2}}V\Sigma_{T,k}\\ &\quad\cdot V^TM_k^{\frac{1}{2}}\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)^{-1}\Big\|\\ &\leq(1+c_x^4L^2)\Big\|\left(M_k+\lambda I_k+\Delta_{11}-\Delta_{12}(\lambda I_{n-k}+\Delta_{22})^{-1}\Delta_{12}^T\right)^{-1}M_k^{\frac{1}{2}}V\Sigma_{T,k}\end{aligned}$$

$$
\begin{aligned}
&\cdot V^T M_k^{\frac{1}{2}} \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right)^{-1}\Big\| \\
&= (1 + c_x^4 L^2) \left\| \left(V^T \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V\right)^{-1} \right. \\
&\cdot V^T M_k^{\frac{1}{2}} V \Sigma_{T,k} V^T M_k^{\frac{1}{2}} V \\
&\left. \cdot \left(V^T \left(M_k + \lambda I_k + \Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V\right)^{-1} \right\| \\
&= (1 + c_x^4 L^2) \left\| \left(X_k^T X_k + \lambda I_k + V^T(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T)V\right)^{-1} \right. \\
&\cdot \left(X_k^T X_k\right)^{\frac{1}{2}} \Sigma_{T,k} \left(X_k^T X_k\right)^{\frac{1}{2}} \\
&\left. \cdot \left(X_k^T X_k + \lambda I_k + V^T(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T)V\right)^{-1} \right\|.
\end{aligned}
$$

The third equation follows from Equation 5.

We define two quantities that represent concentration error terms:

$$
\begin{aligned}
E_1 &= \left\| \left[X_k^T X_k + \lambda I_k + V^T \left(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T\right) V\right]^{-1} - \left(n\widetilde{\Sigma}_{S,k}\right)^{-1} \right\|. \\
E_2 &= \left(X_k^T X_k\right)^{\frac{1}{2}} - \left(n\Sigma_{S,k}\right)^{\frac{1}{2}}.
\end{aligned}
$$

Since $n > 4c_x^4 \left(k + \ln\frac{1}{\delta}\right) \lambda_1^6 \lambda_k^{-8} \|\Sigma_{T,k}\|^2 \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\|^{-2}$,

and $n > 2c_x^4 L \left(\lambda + \sum_{j>k} \lambda_j\right) \lambda_1^2 \lambda_k^{-4} \|\Sigma_{T,k}\| \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\|^{-1}$,

$$
\begin{aligned}
&\|E_1\| \left\|n\widetilde{\Sigma}_{S,k}\right\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \left\|(n\Sigma_{S,k})^{\frac{1}{2}}\right\| \|\Sigma_{T,k}\| \left\|(n\Sigma_{S,k})^{\frac{1}{2}}\right\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&\le \frac{c_x^2 \left(\sqrt{n(k + \ln\frac{1}{\delta})}\lambda_1 + c_x^2 L \left(\lambda + \sum_{j>k} \lambda_j\right)\right)}{(\lambda + n\lambda_k)^2} (\lambda + n\lambda_1) \frac{n\lambda_1}{(\lambda + n\lambda_k)^2} \|\Sigma_{T,k}\| \\
&\le \frac{c_x^2 \left(\sqrt{n(k + \ln\frac{1}{\delta})}\lambda_1 + c_x^2 L \left(\lambda + \sum_{j>k} \lambda_j\right)\right)}{n^2} \frac{\lambda_1^2}{\lambda_k^4} \|\Sigma_{T,k}\| \\
&= \frac{c_x^2 \sqrt{(k + \ln\frac{1}{\delta})}}{n\sqrt{n}} \frac{\lambda_1^3}{\lambda_k^4} \|\Sigma_{T,k}\| + \frac{c_x^4 L \left(\lambda + \sum_{j>k} \lambda_j\right)}{n^2} \frac{\lambda_1^2}{\lambda_k^4} \|\Sigma_{T,k}\| \\
&< \frac{1}{2n} \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\| + \frac{1}{2n} \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\| \\
&= \frac{1}{n} \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\|.
\end{aligned}
$$

Similar to Equation 6, since $n > 4c_x^4 \left(k + \ln\frac{1}{\delta}\right) \lambda_1^4 \lambda_k^{-4}$ and $n > 2c_x^4 L \left(\lambda + \sum_{j>k} \lambda_j\right) \lambda_1 \lambda_k^{-2}$,

$$
\|E_1\| \left\|n\widetilde{\Sigma}_{S,k}\right\| < 1.
$$

Since $n > c_x^2 \left(k + \ln\frac{1}{\delta}\right) \lambda_1^4 \lambda_k^{-6} \|\Sigma_{T,k}\|^2 \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\|^{-2}$,

$$
\begin{aligned}
&\|E_2\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}}\right\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \left\|(n\Sigma_{S,k})^{\frac{1}{2}}\right\| \|\Sigma_{T,k}\| \left\|(n\Sigma_{S,k})^{\frac{1}{2}}\right\| \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&\le c_x \sqrt{k + \ln\frac{1}{\delta}} \lambda_1 \lambda_k^{-\frac{1}{2}} (n\lambda_k)^{-\frac{1}{2}} \frac{n\lambda_1}{(\lambda + n\lambda_k)^2} \|\Sigma_{T,k}\| \\
&\le \frac{c_x \sqrt{k + \ln\frac{1}{\delta}}}{n\sqrt{n}} \frac{\lambda_1^2}{\lambda_k^3} \|\Sigma_{T,k}\| \\
&\le \frac{1}{n} \left\|\Sigma_{S,k}^{-\frac{1}{2}} \Sigma_{T,k} \Sigma_{S,k}^{-\frac{1}{2}}\right\|.
\end{aligned}
$$

Similar to Equation 7, since $n > c_x^2\left(k + \ln\frac{1}{\delta}\right)\lambda_1^2\lambda_k^{-2}$,

$$\|E_2\|\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}}\right\| < 1.$$

Combining the four inequalities above,

$$\begin{aligned}
&\Big\|\left(X_k^TX_k + \lambda I_k + V^T(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T)V\right)^{-1} \\
&\cdot\left(X_k^TX_k\right)^{\frac{1}{2}}\Sigma_{T,k}\left(X_k^TX_k\right)^{\frac{1}{2}} \\
&\cdot\left(X_k^TX_k + \lambda I_k + V^T(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T)V\right)^{-1}\Big\| \\
\le &\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\Sigma_{T,k}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&+ 2\left\|E_1\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\Sigma_{T,k}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&+ 2\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}E_2\Sigma_{T,k}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&+ \left\|E_1\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\Sigma_{T,k}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}E_1\right\| \\
&+ \left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}E_2\Sigma_{T,k}E_2\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&+ 2\left\|E_1\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\Sigma_{T,k}E_2\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&+ 2\left\|E_1E_2\Sigma_{T,k}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&+ 2\left\|E_1E_2\Sigma_{T,k}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}E_1\right\| \\
&+ 2\left\|E_1E_2\Sigma_{T,k}E_2\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
&+ \left\|E_1E_2\Sigma_{T,k}E_2E_1\right\|.
\end{aligned}$$

In particular,

$$\begin{aligned}
&\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\Sigma_{T,k}\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\| \\
= &\frac{1}{n}\left\|\widetilde{\Sigma}_{S,k}^{-1}\Sigma_{S,k}^{\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{\frac{1}{2}}\widetilde{\Sigma}_{S,k}^{-1}\right\| \\
\le &\frac{1}{n}\left\|\Sigma_{S,k}^{-1}\Sigma_{S,k}^{\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{\frac{1}{2}}\Sigma_{S,k}^{-1}\right\| \\
= &\frac{1}{n}\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|.
\end{aligned}$$

The inequality follows from the fact that $\|BAB\| = \|A^{\frac{1}{2}}BA^{\frac{1}{2}}\| \le \|A^{\frac{1}{2}}CA^{\frac{1}{2}}\| = \|CAC\|$, where A, B, C are positive semi-definite matrices, and $C \succcurlyeq B$, which implies that $A^{\frac{1}{2}}CA^{\frac{1}{2}} \succcurlyeq A^{\frac{1}{2}}BA^{\frac{1}{2}}$.

$$\begin{aligned}
&\left\|E_1E_2\Sigma_{T,k}E_2E_1\right\| \\
= &\Big\|E_1n\widetilde{\Sigma}_{S,k}\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}E_2\left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}}\left(n\widetilde{\Sigma}_{S,k}\right)^{\frac{1}{2}} \\
&\cdot\Sigma_{T,k}E_2\left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}}\left(n\widetilde{\Sigma}_{S,k}\right)^{\frac{1}{2}}E_1n\widetilde{\Sigma}_{S,k}\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\Big\| \\
\le &\left(\|E_1\|\left\|n\widetilde{\Sigma}_{S,k}\right\|\right)^2\left(\|E_2\|\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}}\right\|\right) \\
&\cdot\|E_2\|\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-\frac{1}{2}}\right\|\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\|\left\|\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\right\|\left\|\Sigma_{T,k}\right\|\left\|\left(n\Sigma_{S,k}\right)^{\frac{1}{2}}\right\|\left\|\left(n\widetilde{\Sigma}_{S,k}\right)^{-1}\right\|
\end{aligned}$$

$$\leq \frac{1}{n}\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|.$$

The other terms can be similarly bounded. Therefore,

$$\begin{aligned}
&\left\|\left(X_k^T X_k + \lambda I_k + V^T(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T)V\right)^{-1}\right.\\
&\cdot \left(X_k^T X_k\right)^{\frac{1}{2}} \Sigma_{T,k} \left(X_k^T X_k\right)^{\frac{1}{2}}\\
&\left.\cdot \left(X_k^T X_k + \lambda I_k + V^T(\Delta_{11} - \Delta_{12}(\lambda I_{n-k} + \Delta_{22})^{-1}\Delta_{12}^T)V\right)^{-1}\right\|\\
\leq{}& \frac{16}{n}\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|.
\end{aligned}$$

□

A.3.4 Bias in the last $d-k$ dimensions

The upper bound for the bias in the last $d-k$ dimensions is extended from Tsigler & Bartlett (2023)'s Lemma 28. The bias can be decomposed into three terms.

$$\begin{aligned}
&\left(\widehat{\beta}(X\beta^\star)_{-k} - \beta^\star_{-k}\right)^T \Sigma_{T,-k} \left(\widehat{\beta}(X\beta^\star)_{-k} - \beta^\star_{-k}\right)\\
\leq{}& 3(\beta^\star_{-k})^T\Sigma_{T,-k}\beta^\star_{-k}\\
&+ 3(\beta^\star_{-k})^T X_{-k}^T (XX^T + \lambda I_n)^{-1} X_{-k}\Sigma_{T,-k}X_{-k}^T(XX^T + \lambda I_n)^{-1}X_{-k}\beta^\star_{-k}\\
&+ 3(\beta^\star_{k})^T X_{k}^T (XX^T + \lambda I_n)^{-1} X_{-k}\Sigma_{T,-k}X_{-k}^T(XX^T + \lambda I_n)^{-1}X_{k}\beta^\star_{k}.
\end{aligned}$$

Lemma 23. Under the same conditions as in Lemma 18, and on the same event, for any $N_1 < n < N_2$,

$$\begin{aligned}
&(\beta^\star_{-k})^T X_{-k}^T (XX^T + \lambda I_n)^{-1} X_{-k}\Sigma_{T,-k}X_{-k}^T(XX^T + \lambda I_n)^{-1}X_{-k}\beta^\star_{-k}\\
\leq{}& c_x^2 L\left(\lambda + \sum_j \lambda_j\right)^{-1} n\|\Sigma_{T,-k}\|(\beta^\star_{-k})^T\Sigma_{S,-k}\beta^\star_{-k}.
\end{aligned}$$

where N_1, N_2 are defined as in Lemma 18.

Proof.

$$\begin{aligned}
&(\beta^\star_{-k})^T X_{-k}^T (XX^T + \lambda I_n)^{-1} X_{-k}\Sigma_{T,-k}X_{-k}^T(XX^T + \lambda I_n)^{-1}X_{-k}\beta^\star_{-k}\\
\leq{}& \|\Sigma_{T,-k}\|(\beta^\star_{-k})^T X_{-k}^T (XX^T + \lambda I_n)^{-1} X_{-k}X_{-k}^T(XX^T + \lambda I_n)^{-1}X_{-k}\beta^\star_{-k}\\
\leq{}& \|\Sigma_{T,-k}\|(\beta^\star_{-k})^T X_{-k}^T (XX^T + \lambda I_n)^{-1} (XX^T + \lambda I_n)(XX^T + \lambda I_n)^{-1}X_{-k}\beta^\star_{-k}\\
\leq{}& \|\Sigma_{T,-k}\|\left\|(XX^T + \lambda I_n)^{-1}\right\|(\beta^\star_{-k})^T X_{-k}^T X_{-k}\beta^\star_{-k}\\
\leq{}& \|\Sigma_{T,-k}\|\left\|(X_{-k}X_{-k}^T + \lambda I_n)^{-1}\right\|(\beta^\star_{-k})^T X_{-k}^T X_{-k}\beta^\star_{-k}\\
\leq{}& \|\Sigma_{T,-k}\|\left(\frac{1}{c_x L}\left(\lambda + \sum_j \lambda_j\right)\right)^{-1} c_x n(\beta^\star_{-k})^T\Sigma_{S,-k}\beta^\star_{-k}\\
={}& c_x^2 L\left(\lambda + \sum_j \lambda_j\right)^{-1} n\|\Sigma_{T,-k}\|(\beta^\star_{-k})^T\Sigma_{S,-k}\beta^\star_{-k}.
\end{aligned}$$

The fourth inequality follows from $XX^T + \lambda I_n \succcurlyeq X_{-k}X_{-k}^T + \lambda I_n$. □

Lemma 24. Under the same conditions as in Lemma 18, and on the same event, for any $N_1 < n < N_2$,

$$(\beta^\star_{k})^T X_{k}^T (XX^T + \lambda I_n)^{-1} X_{-k}\Sigma_{T,-k}X_{-k}^T(XX^T + \lambda I_n)^{-1}X_{k}\beta^\star_{k}$$

$$\leq \frac{c_x^6}{n} L \left(\lambda + \sum_{j>k} \lambda_j \right) \|\Sigma_{T,-k}\| (\beta_k^\star)^T \Sigma_{S,k}^{-1} \beta_k^\star.$$

where N_1, N_2 are defined as in Lemma 18.

Proof. It can be verified by Woodbury matrix identity that:

$$(XX^T + \lambda I_n)^{-1} X_k = (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \left(I_k + X_k^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \right)^{-1}.$$

Therefore,

$$\begin{aligned}
&(\beta_k^\star)^T X_k^T (XX^T + \lambda I_n)^{-1} X_{-k} \Sigma_{T,-k} X_{-k}^T (XX^T + \lambda I_n)^{-1} X_k \beta_k^\star \\
&= \left\| \Sigma_{T,-k}^{\frac{1}{2}} X_{-k}^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \left(I_k + X_k^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \right)^{-1} \beta_k^\star \right\|^2 \\
&\leq \|\Sigma_{T,-k}\| \left\| (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_{-k} X_{-k}^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} \right\| \\
&\quad \cdot \left\| X_k \left(I_k + X_k^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \right)^{-1} \beta_k^\star \right\|^2 \\
&= \|\Sigma_{T,-k}\| \left\| (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_{-k} X_{-k}^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} \right\| \\
&\quad \cdot \left\| X_k \Sigma_{S,k}^{-\frac{1}{2}} \left(\Sigma_{S,k}^{-1} + \Sigma_{S,k}^{-\frac{1}{2}} X_k^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{-\frac{1}{2}} \right)^{-1} \Sigma_{S,k}^{-\frac{1}{2}} \beta_k^\star \right\|^2 \\
&\leq \|\Sigma_{T,-k}\| \left\| (X_{-k} X_{-k}^T + \lambda I_n)^{-1} \right\| \left\| \Sigma_{S,k}^{-\frac{1}{2}} X_k^T X_k \Sigma_{S,k}^{-\frac{1}{2}} \right\| \\
&\quad \cdot \left\| \left(\Sigma_{S,k}^{-1} + \Sigma_{S,k}^{-\frac{1}{2}} X_k^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{-\frac{1}{2}} \right)^{-2} \right\| (\beta_k^\star)^T \Sigma_{S,k}^{-1} \beta_k^\star.
\end{aligned}$$

In particular,

$$\begin{aligned}
&\left\| \left(\Sigma_{S,k}^{-1} + \Sigma_{S,k}^{-\frac{1}{2}} X_k^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{-\frac{1}{2}} \right)^{-1} \right\| \\
&\leq \left\| \left(\Sigma_{S,k}^{-\frac{1}{2}} X_k^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{-\frac{1}{2}} \right)^{-1} \right\| \\
&\leq \left\| X_{-k} X_{-k}^T + \lambda I_n \right\| \left\| \left(\Sigma_{S,k}^{-\frac{1}{2}} X_k^T X_k \Sigma_{S,k}^{-\frac{1}{2}} \right)^{-1} \right\| \\
&\leq c_x \left(\lambda + \sum_{j>k} \lambda_j \right) \frac{c_x}{n} \\
&= \frac{c_x^2}{n} \left(\lambda + \sum_{j>k} \lambda_j \right).
\end{aligned}$$

The second inequality follows from $\mu_{\min}(ABA^T) \geq \mu_{\min}(B) \mu_{\min}(AA^T)$ where the matrix B is positive definite.

Therefore,

$$\begin{aligned}
&(\beta_k^\star)^T X_k^T (XX^T + \lambda I_n)^{-1} X_{-k} \Sigma_{T,-k} X_{-k}^T (XX^T + \lambda I_n)^{-1} X_k \beta_k^\star \\
&\leq \|\Sigma_{T,-k}\| \left\| (X_{-k} X_{-k}^T + \lambda I_n)^{-1} \right\| \left\| \Sigma_{S,k}^{-\frac{1}{2}} X_k^T X_k \Sigma_{S,k}^{-\frac{1}{2}} \right\| \\
&\quad \cdot \left\| \left(\Sigma_{S,k}^{-1} + \Sigma_{S,k}^{-\frac{1}{2}} X_k^T (X_{-k} X_{-k}^T + \lambda I_n)^{-1} X_k \Sigma_{S,k}^{-\frac{1}{2}} \right)^{-2} \right\| (\beta_k^\star)^T \Sigma_{S,k}^{-1} \beta_k^\star \\
&\leq \|\Sigma_{T,-k}\| \cdot \left(\frac{1}{c_x L} \left(\lambda + \sum_{j>k} \lambda_j \right) \right)^{-1} \cdot c_x n \cdot \frac{c_x^4}{n^2} \left(\lambda + \sum_{j>k} \lambda_j \right)^2 \cdot (\beta_k^\star)^T \Sigma_{S,k}^{-1} \beta_k^\star \\
&= \frac{c_x^6}{n} L \left(\lambda + \sum_{j>k} \lambda_j \right) \|\Sigma_{T,-k}\| (\beta_k^\star)^T \Sigma_{S,k}^{-1} \beta_k^\star.
\end{aligned}$$

$\square$

A.4 Main results

Theorem 25. Let $\mathcal{T} = \Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}$ and $\mathcal{U} = \Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}$. There exists a constant $c > 2$ depending only on σ, such that for any $cN < n < r_k$, if the assumption condNum(k, δ, L) (Assumption 2) is satisfied, then with probability at least $1 - 2\delta - ce^{-n/c}$,

$$\frac{V}{cv^2} \le L^2\frac{\operatorname{tr}[\mathcal{T}]}{n} + L^2\frac{n\operatorname{tr}[\mathcal{U}]}{\left(\lambda + \sum_{j>k}\lambda_j\right)^2}.$$

$$\frac{B}{c} \le \|\beta_k^\star\|^2_{\Sigma_{S,k}^{-1}}\left(\frac{\lambda + \sum_{j>k}\lambda_j}{n}\right)^2\left[\|\mathcal{T}\| + L\frac{n\|\Sigma_{T,-k}\|}{\lambda + \sum_{j>k}\lambda_j}\right]$$
$$+ \left\|\beta_{-k}^\star\right\|^2_{\Sigma_{S,-k}}\left[L^2\|\mathcal{T}\| + L\frac{n\|\Sigma_{T,-k}\|}{\lambda + \sum_{j>k}\lambda_j}\right].$$

N is defined as follows:

$$N = \max\left\{(k + \ln\frac{1}{\delta})\lambda_1^6\lambda_k^{-8}\|\Sigma_{T,k}\|^2k^2\left(\operatorname{tr}[\mathcal{T}]\right)^{-2},\right.$$
$$\left. L\lambda_1^2\lambda_k^{-4}(\lambda + \sum_{j>k}\lambda_j)\|\Sigma_{T,k}\|k\left(\operatorname{tr}[\mathcal{T}]\right)^{-1}\right\}.$$

Remark 8 (Sample complexity)**.** We have assumed $n \ge c_x N$ in the theorem. The first condition on N indicates $n \gg k$. From the inequality $\lambda_k^2 \le \|\Sigma_{T,k}\|^2k^2\left(\operatorname{tr}[\mathcal{T}]\right)^{-2} \le k^2\lambda_1^2$, it follows that $n = \Omega(k)$ in the best case, consistent with the sample complexity of classic linear regression. This optimal case occurs when $\Sigma_{S,k} \approx \Sigma_{T,k}$. In the worst case, $n = \Omega(k^3)$ where covariate shift is significant in the first k dimensions–e.g., when the test data lies predominantly in the subspace of the first dimension. This shift in sample complexity under varying degrees of covariate shift parallels the analysis of Ge et al. (2024) (see theire Theorem 4.2) for the under-parameterized setting. The second condition implies $n \gg \lambda + \sum_{j>k}\lambda_j$, such that the regularization is not too strong to introduce a bias greater than a constant (as shown in the first bias term). On the other hand, we assume $n < r_k$ in the theorem, which is consistent with the over-parameterized regime and Assumption 1, where the last $d - k$ components are considered to be essentially high-dimensional.

Proof. The theorem follows from Lemma 18, Lemma 19, Lemma 20, Lemma 21, Lemma 22, Lemma 23 and Lemma 24. For a constant $c_x' > 2$ depending only on σ, these lemmas hold for values of n that satisfy the following inequalities:

$$n > 4c_x'^4(k + \ln(1/\delta))\lambda_1^2\lambda_k^{-2},$$
$$n > 2c_x'^4L\lambda_k^{-1}\left(\lambda + \sum_{j>k}\lambda_j\right),$$
$$n > 4c_x'^4\left(k + \ln\frac{1}{\delta}\right)\lambda_1^4\lambda_k^{-4},$$
$$n > 2c_x'^4L\lambda_1\lambda_k^{-2}\left(\lambda + \sum_{j>k}\lambda_j\right),$$
$$n > 4c_x'^4\left(k + \ln\frac{1}{\delta}\right)\lambda_1^6\lambda_k^{-8}\|\Sigma_{T,k}\|^2k^2\left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-2},$$
$$n > 2c_x'^4L\lambda_1^2\lambda_k^{-4}\left(\lambda + \sum_{j>k}\lambda_j\right)\|\Sigma_{T,k}\|k\left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-1},$$
$$n > 2c_x'^3(\lambda + \sum_{j>k}\lambda_j)\lambda_1\lambda_k^{-2},$$

$$
\begin{aligned}
&n > 4c_x'^4\left(k+\ln\frac{1}{\delta}\right)\lambda_1^6\lambda_k^{-8}\|\Sigma_{T,k}\|^2\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|^{-2},\\
&n > 2c_x'^4 L\left(\lambda+\sum_{j>k}\lambda_j\right)\lambda_1^2\lambda_k^{-4}\|\Sigma_{T,k}\|\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|^{-1},\\
&n < \lambda_{k+1}^{-1}\left(\lambda+\sum_{j>k}\lambda_j\right).
\end{aligned}
$$

A sufficient condition for all the inequalities above is given by $4c_x'^4 N_1 < n < r_k$. This follows from the following facts:

$$
\begin{aligned}
\lambda_1\lambda_k^{-1} &\geq 1,\\
c_x' &> 2,\\
L &\geq 1,\\
k\left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-1} &\geq \left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|^{-1},\\
k\|\Sigma_{T,k}\|\left(\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\right)^{-1} &\geq \lambda_k.
\end{aligned}
$$

Then, with probability at least $1-2\delta-c_x'e^{-n/c_x'}$:

$$
\begin{aligned}
V/2 &\leq 16v^2(1+c_x'^4L^2)\frac{1}{n}\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\\
&\quad + v^2c_x'^3L^2n\left(\lambda+\sum_{j>k}\lambda_j\right)^{-2}\operatorname{tr}\left[\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}\right]\\
&\leq 32v^2c_x'^4L^2\frac{1}{n}\operatorname{tr}\left[\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right]\\
&\quad + v^2c_x'^3L^2n\left(\lambda+\sum_{j>k}\lambda_j\right)^{-2}\operatorname{tr}\left[\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{\frac{1}{2}}\right],\\
B/2 &\leq \frac{16c_x'^4}{n^2}\left(\lambda+\sum_{j>k}\lambda_j\right)^2(\beta_k^\star)^T\Sigma_{S,k}^{-1}\beta_k^\star\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|\\
&\quad + 32c_x'(1+c_x'^4L^2)\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star\\
&\quad + 3c_x'^2L\left(\lambda+\sum_{j}\lambda_j\right)^{-1}n\|\Sigma_{T,-k}\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star\\
&\quad + 3\frac{c_x'^6}{n}L\left(\lambda+\sum_{j>k}\lambda_j\right)\|\Sigma_{T,-k}\|(\beta_k^\star)^T\Sigma_{S,k}^{-1}\beta_k^\star\\
&\quad + 3(\beta_{-k}^\star)^T\Sigma_{T,-k}\beta_{-k}^\star\\
&\leq 16c_x'^4\frac{1}{n^2}\left(\lambda+\sum_{j>k}\lambda_j\right)^2\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|(\beta_k^\star)^T\Sigma_{S,k}^{-1}\beta_k^\star\\
&\quad + 64c_x'^5L^2\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star\\
&\quad + 3c_x'^2Ln\left(\lambda+\sum_{j}\lambda_j\right)^{-1}\|\Sigma_{T,-k}\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star
\end{aligned}
$$

$$
\begin{aligned}
&+ 3c_x'^6 L \frac{1}{n}\left(\lambda + \sum_{j>k}\lambda_j\right)\|\Sigma_{T,-k}\|(\beta_k^\star)^T\Sigma_{S,k}^{-1}\beta_k^\star \\
&+ 3(\beta_{-k}^\star)^T\Sigma_{T,-k}\beta_{-k}^\star \\
\le\ & 16c_x'^4\frac{1}{n^2}\left(\lambda + \sum_{j>k}\lambda_j\right)^2\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|(\beta_k^\star)^T\Sigma_{S,k}^{-1}\beta_k^\star \\
&+ 64c_x'^5 L^2\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star \\
&+ 3c_x'^2 Ln\left(\lambda + \sum_{j>k}\lambda_j\right)^{-1}\|\Sigma_{T,-k}\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star \\
&+ 3c_x'^6 L \frac{1}{n}\left(\lambda + \sum_{j>k}\lambda_j\right)\|\Sigma_{T,-k}\|(\beta_k^\star)^T\Sigma_{S,k}^{-1}\beta_k^\star \\
&+ 3c_x'^5 L^2\left\|\Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}\right\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star.
\end{aligned}
$$

The last inequality follows from:

$$
\begin{aligned}
(\beta_{-k}^\star)^T\Sigma_{T,-k}\beta_{-k}^\star &= (\beta_{-k}^\star)^T\Sigma_{S,-k}^{\frac{1}{2}}\Sigma_{S,-k}^{-\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{-\frac{1}{2}}\Sigma_{S,-k}^{\frac{1}{2}}\beta_{-k}^\star \\
&\le \left\|\Sigma_{S,-k}^{-\frac{1}{2}}\Sigma_{T,-k}\Sigma_{S,-k}^{-\frac{1}{2}}\right\|(\beta_{-k}^\star)^T\Sigma_{S,-k}\beta_{-k}^\star.
\end{aligned}
$$

By taking $c = 134c_x'^6$, the proof is complete. $\square$

Corollary 26. Let $\mathcal{T} = \Sigma_{S,k}^{-\frac{1}{2}}\Sigma_{T,k}\Sigma_{S,k}^{-\frac{1}{2}}$, $\mathcal{U} = \Sigma_{S,-k}\Sigma_{T,-k}$ and $\mathcal{V} = \Sigma_{S,-k}^2$. There exists a constant $c > 2$ depending only on σ, L, such that for any $cN < n < r_k$, if the assumption condNum(k,δ,L) (Assumption 2) is satisfied, then with probability at least $1-3\delta$,

$$
\begin{aligned}
\frac{V}{cv^2} &\le \frac{k}{n}\frac{\mathrm{tr}[\mathcal{T}]}{k} + \frac{n}{R_k}\frac{\mathrm{tr}[\mathcal{U}]}{\mathrm{tr}[\mathcal{V}]}. \\
\frac{B}{c} &\le \left(\|\beta_k^\star\|^2_{\Sigma_{S,k}^{-1}}\left(\frac{\lambda + \sum_{j>k}\lambda_j}{n}\right)^2 + \left\|\beta_{-k}^\star\right\|^2_{\Sigma_{S,-k}}\right)\left[\|\mathcal{T}\| + \frac{n}{r_k}\frac{\|\Sigma_{T,-k}\|}{\|\Sigma_{S,-k}\|}\right].
\end{aligned}
$$

N is a polynomial function of $k + \ln(1/\delta)$, $\lambda_1\lambda_k^{-1}$, $1 + \left(\lambda + \sum_{j>k}\lambda_j\right)\lambda_k^{-1}$.

Proof. The first variance term follows directly from Theorem 25.

For the second variance term, by plugging in the definition of R_k,

$$
\begin{aligned}
L^2\frac{n\,\mathrm{tr}\left[\mathcal{U}\right]}{\left(\lambda + \sum_{j>k}\lambda_j\right)^2} &= L^2\frac{n}{R_k}\frac{\mathrm{tr}\left[\Sigma_{S,-k}\Sigma_{T,-k}\right]}{\sum_{j>k}\lambda_j^2} \\
&= L^2\frac{n}{R_k}\frac{\mathrm{tr}[\mathcal{U}]}{\mathrm{tr}[\mathcal{V}]}.
\end{aligned}
$$

For the first bias term, by plugging in the definition of r_k,

$$
\begin{aligned}
&\|\beta_k^\star\|^2_{\Sigma_{S,k}^{-1}}\left(\frac{\lambda + \sum_{j>k}\lambda_j}{n}\right)^2\left[\|\mathcal{T}\| + L\frac{n\|\Sigma_{T,-k}\|}{\lambda + \sum_{j>k}\lambda_j}\right] \\
=\ &\|\beta_k^\star\|^2_{\Sigma_{S,k}^{-1}}\left(\frac{\lambda + \sum_{j>k}\lambda_j}{n}\right)^2\left[\|\mathcal{T}\| + L\frac{n}{r_k}\frac{\|\Sigma_{T,-k}\|}{\lambda_{k+1}}\right].
\end{aligned}
$$

Similarly, the second bias term can be transformed into:

$$
\left\|\beta_{-k}^\star\right\|^2_{\Sigma_{S,-k}}\left[L^2\|\mathcal{T}\| + L\frac{n\|\Sigma_{T,-k}\|}{\lambda + \sum_{j>k}\lambda_j}\right] = \left\|\beta_{-k}^\star\right\|^2_{\Sigma_{S,-k}}\left[L^2\|\mathcal{T}\| + L\frac{n}{r_k}\frac{\|\Sigma_{T,-k}\|}{\lambda_{k+1}}\right].
$$

Since the statement of Theorem 25 holds with probability at least $1-2\delta-ce^{-n/c}$, we only require $ce^{-n/c}<\delta$, which is equivalent as $n>c\ln c+c\ln(1/\delta)$. Combining the lower bounds of n in Theorem 25, we should have:

$$n>\max\Bigg\{c\ln c+c\ln\frac{1}{\delta},\\ c\big(k+\ln\frac{1}{\delta}\big)\lambda_1^6\lambda_k^{-8}\|\Sigma_{T,k}\|^2k^2\left(\mathrm{tr}\left[\mathcal{T}\right]\right)^{-2},\\ cL\lambda_1^2\lambda_k^{-4}\big(\lambda+\sum_{j>k}\lambda_j\big)\|\Sigma_{T,k}\|k\left(\mathrm{tr}\left[\mathcal{T}\right]\right)^{-1}\Bigg\}.$$

For the first term in the maximum argument,

$$\begin{aligned}c\ln c+c\ln\frac{1}{\delta}&\le c^2+c\ln\frac{1}{\delta}\\&\le c^2\left(k+\ln\frac{1}{\delta}\right).\end{aligned}$$

The second term:

$$\begin{aligned}&c\big(k+\ln\frac{1}{\delta}\big)\lambda_1^6\lambda_k^{-8}\|\Sigma_{T,k}\|^2k^2\left(\mathrm{tr}\left[\mathcal{T}\right]\right)^{-2}\\ \le{}& c\big(k+\ln\frac{1}{\delta}\big)\lambda_1^6\lambda_k^{-8}\|\Sigma_{T,k}\|^2k^2\left(\mu_k(\Sigma_{S,k}^{-1})\,\mathrm{tr}[\Sigma_{T,k}]\right)^{-2}\\ \le{}& c\big(k+\ln\frac{1}{\delta}\big)\lambda_1^8\lambda_k^{-8}\|\Sigma_{T,k}\|^2k^2\|\Sigma_{T,k}\|^{-2}\\ ={}& c\big(k+\ln\frac{1}{\delta}\big)^3\lambda_1^8\lambda_k^{-8}.\end{aligned}$$

The first inequality follows from $\mathrm{tr}[MN]\ge\mu_{\min}(M)\,\mathrm{tr}[N]$ for postive semi-definite matrices M,N.

Similar, for the third term:

$$\begin{aligned}&cL\lambda_1^2\lambda_k^{-4}\big(\lambda+\sum_{j>k}\lambda_j\big)\|\Sigma_{T,k}\|k\left(\mathrm{tr}\left[\mathcal{T}\right]\right)^{-1}\\ \le{}& cL\lambda_1^2\lambda_k^{-4}\big(\lambda+\sum_{j>k}\lambda_j\big)\|\Sigma_{T,k}\|k\lambda_1\|\Sigma_{T,k}\|^{-1}\\ \le{}& cL\big(k+\ln\frac{1}{\delta}\big)\lambda_1^3\lambda_k^{-4}\big(\lambda+\sum_{j>k}\lambda_j\big).\end{aligned}$$

The proof is complete by taking c as c^2L^2 and $N=\big(k+\ln\frac{1}{\delta}\big)^3\big(\lambda_1\lambda_k^{-1}\big)^8\Big[1+\big(\lambda+\sum_{j>k}\lambda_j\big)\lambda_k^{-1}\Big]$.

$\square$

B Large shift in minor directions

In this section, we consider the scenario where the signal $\beta^\star$ mainly concentrate on the first k components (here we choose the basis to be the eigenvectors of Σ_S), but the target covariance Σ_T may not be small on the last $d-k$ components.

B.1 Lower bound for ridge regression

In this subsection, we will show that the original ridge regression algorithm will not work under this scenario.

Recall our model:

$$y=\beta^{\star T}x+\epsilon, \tag{8}$$

We can write our data as

$$Y = X\beta^\star + \boldsymbol{\epsilon}, \tag{9}$$

where $Y = (y_1, \cdots, y_n)^T \in \mathbb{R}^{n\times 1}$, $X = (x_1, \cdots, x_n)^T \in \mathbb{R}^{n\times d}$, $\boldsymbol{\epsilon} = (\epsilon_1, \cdots, \epsilon_n)^T \in \mathbb{R}^{n\times 1}$. We denote by $\widehat{\Sigma}_S := \frac{1}{n}X^T X$ the sample covariance matrix.

Assume the same assumptions as in our previous section still holds. We let $\Sigma_S = \mathbb{E}[x_i x_i^T]$ be the following: its eigenvalues $\lambda_1, \cdots, \lambda_d$ satisfies $\lambda_1 = \cdots = \lambda_k = 1$, $\lambda_{k+1} = \cdots = \lambda_{k+\lfloor\sqrt{n}/C_2\rfloor} = C_1/\sqrt{n}$ for sufficiently large constants C_1, C_2, and the remaining eigenvalues are all set to zero. We let $\Sigma_T = I_d$. Then the excess risk is $\mathbb{E}_\epsilon[(\widehat{\beta}-\beta^\star)^T\Sigma_T(\widehat{\beta}-\beta^\star)] = \mathbb{E}_\epsilon\|\widehat{\beta}-\beta^\star\|^2$. We will show that under this scenario, ridge regression can not obtain an error rate of $\mathcal{O}(\frac{1}{n})$. To see this, we explicitly write out the ridge solution:

$$\begin{aligned}
\widehat{\beta} &= (X^T X + \lambda I_d)^{-1}X^T Y \\
&= (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}(\frac{1}{n}X^T Y) \\
&= (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}(\frac{1}{n}X^T(X\beta^\star + \boldsymbol{\epsilon})) \\
&= (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}(\frac{1}{n}X^T X\beta^\star + \frac{1}{n}X^T\boldsymbol{\epsilon}) \\
&= (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}(\widehat{\Sigma}_S\beta^\star + \frac{1}{n}X^T\boldsymbol{\epsilon}) \\
&= (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\widehat{\Sigma}_S\beta^\star + (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\frac{1}{n}X^T\boldsymbol{\epsilon}.
\end{aligned} \tag{10}$$

Therefore

$$\begin{aligned}
\widehat{\beta} - \beta^\star &= (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\widehat{\Sigma}_S\beta^\star - \beta^\star + (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\frac{1}{n}X^T\boldsymbol{\epsilon} \\
&= (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\widehat{\Sigma}_S\beta^\star - (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)\beta^\star + (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\frac{1}{n}X^T\boldsymbol{\epsilon} \\
&= -\frac{\lambda}{n}(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\beta^\star + (\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\frac{1}{n}X^T\boldsymbol{\epsilon}
\end{aligned}$$

Taking expectation with respect to ϵ,

$$\begin{aligned}
\mathbb{E}_\epsilon\|\widehat{\beta} - \beta^\star\|^2 &= \frac{\lambda^2}{n^2}\|(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\beta^\star\|^2 + \frac{1}{n^2}\operatorname{tr}(\boldsymbol{\epsilon}^T X(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-2}X^T\boldsymbol{\epsilon}) \\
&= \frac{\lambda^2}{n^2}\|(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\beta^\star\|^2 + v^2\frac{1}{n}\operatorname{tr}((\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-2}\widehat{\Sigma}_S) \\
&:= B + V
\end{aligned} \tag{11}$$

where $B = \frac{\lambda^2}{n^2}\|(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\beta^\star\|^2$ is the bias, $V = \frac{v^2}{n}\operatorname{tr}((\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-2}\widehat{\Sigma}_S)$ is the variance. We state the formal version of Theorem 4 in the following:

Theorem 27. Under the instance we consider, namely $\lambda_1, \cdots, \lambda_d$ satisfies $\lambda_1 = \cdots = \lambda_k = 1$, $\lambda_{k+1} = \cdots = \lambda_{k+\lfloor\sqrt{n}/C_2\rfloor} = C_1/\sqrt{n}$, $\lambda_{k+\lfloor\sqrt{n}/C_2\rfloor+1} = \cdots = \lambda_d = 0$. WLOG assume $\sigma = 1$, $C_2 \geq C_1((\frac{C_1}{4C})^2 - k - \log\frac{1}{\delta})^{-1}$ for some absolute constant C, and $n \geq (\frac{3C_1}{2})^4$. With probability $1-\delta$, when $\lambda = c\sqrt{n}$, we have $\frac{V}{v^2} \geq C'$, where $C' > 0$ is some absolute constant. When $\lambda \leq n^{3/4}$, we have $\frac{V}{v^2} \geq C'\frac{1}{\sqrt{n}}$. When $\lambda \geq n^{3/4}$, $B \geq \frac{\|\beta^\star\|^2}{9\sqrt{n}}$.

Proof. We will use the following concentration lemma modified from (Vershynin, 2018, Exercise 9.2.5):

Lemma 28. Let $\{x_i\}_{i=1}^n$ be i.i.d. $d-$dimensional random vectors, satisfying: x_i is mean zero, $\mathbb{E}[xx^T] = \Sigma$ and is $\sigma^2\Sigma$-sub-gaussian, in the sense that

$$\mathbb{E}[\exp(v^T x_i)] \leq \exp\left(\frac{\|\sigma\Sigma^{1/2}v\|^2}{2}\right).$$

$X = (x_1, \cdots, x_n)^T \in \mathbb{R}^{n\times d}$. Then with probability $1-\delta$,

$$\|\widehat{\Sigma} - \Sigma\| \le C\sigma^4 \left(\sqrt{\frac{r+\log\frac{1}{\delta}}{n}} + \frac{r+\log\frac{1}{\delta}}{n}\right)\|\Sigma\|$$

where $r := \mathrm{tr}(\Sigma)/\|\Sigma\|$ is the stable rank of Σ, C is an absolute constant.

Applying Lemma 28, we have

$$\|\widehat{\Sigma}_S - \Sigma_S\| \le C\left(\sqrt{\frac{r+\log\frac{1}{\delta}}{n}} + \frac{r+\log\frac{1}{\delta}}{n}\right)$$

where $r = \sum_{i=1}^d \lambda_i = k + \lfloor\sqrt{n}/C_2\rfloor\frac{C_1}{\sqrt{n}} \le k + C_1/C_2$. When $n \ge C_1/C_2 + k + \log\frac{1}{\delta}$, we have

$$\|\widehat{\Sigma}_S - \Sigma_S\| \le 2C\sqrt{\frac{C_1/C_2 + k + \log\frac{1}{\delta}}{n}}.$$

We denote by $\widehat{\lambda}_1 \ge \cdots \ge \widehat{\lambda}_d$ the eigenvalues of $\widehat{\Sigma}_S$. Then by Weyl's inequality (Chen et al., 2021, Lemma 2.2), $\|\widehat{\lambda}_i - \lambda_i\| \le \|\widehat{\Sigma}_S - \Sigma_S\|$. Combining with previous inequalities, we have $1 - 2C\sqrt{\frac{C_1/C_2+k+\log\frac{1}{\delta}}{n}} \le \widehat{\lambda}_i \le 1 + 2C\sqrt{\frac{C_1/C_2+k+\log\frac{1}{\delta}}{n}}$ for $1 \le i \le k$, $\frac{C_1}{\sqrt{n}} - 2C\sqrt{\frac{C_1/C_2+k+\log\frac{1}{\delta}}{n}} \le \widehat{\lambda}_i \le \frac{C_1}{\sqrt{n}} + 2C\sqrt{\frac{C_1/C_2+k+\log\frac{1}{\delta}}{n}}$ for $k+1 \le i \le k + \lfloor\sqrt{n}/C_2\rfloor$. If we take $C_2 \ge C_1((\frac{C_1}{4C})^2 - k - \log\frac{1}{\delta})^{-1}$ then $2C\sqrt{\frac{C_1/C_2+k+\log\frac{1}{\delta}}{n}} \le \frac{C_1}{2\sqrt{n}}$. Therefore we have $\frac{C_1}{2\sqrt{n}} \le \widehat{\lambda}_i \le \frac{3C_1}{2\sqrt{n}}$ for $k+1 \le i \le k + \lfloor\sqrt{n}/C_2\rfloor$. When $\lambda = c\sqrt{n}$, we have

$$\begin{aligned}
\frac{V}{v^2} &= \frac{1}{n}\mathrm{tr}((\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-2}\widehat{\Sigma}_S) \\
&= \frac{1}{n}\sum_{i=1}^d (\widehat{\lambda}_i + \frac{\lambda}{n})^{-2}\widehat{\lambda}_i \\
&\ge \frac{1}{n}\sum_{i=k+1}^{k+\lfloor\sqrt{n}/C_2\rfloor} (\widehat{\lambda}_i + \frac{\lambda}{n})^{-2}\widehat{\lambda}_i \\
&= \frac{1}{n}\sum_{i=k+1}^{k+\lfloor\sqrt{n}/C_2\rfloor} (\widehat{\lambda}_i + \frac{c}{\sqrt{n}})^{-2}\widehat{\lambda}_i \\
&\ge \frac{1}{n}\sum_{i=k+1}^{k+\lfloor\sqrt{n}/C_2\rfloor} (\frac{3C_1}{2\sqrt{n}} + \frac{c}{\sqrt{n}})^{-2}\frac{C_1}{2\sqrt{n}} \\
&= \frac{1}{n}\lfloor\sqrt{n}/C_2\rfloor\frac{C_1}{2}(\frac{3C_1}{2} + c)^{-2}\sqrt{n} \\
&\ge \frac{C_1}{4C_2}(\frac{3C_1}{2} + c)^{-2}.
\end{aligned} \tag{12}$$

Similarly, if $\lambda \le n^{3/4}$,

$$\begin{aligned}
\frac{V}{v^2} &\ge \frac{1}{n}\sum_{i=k+1}^{k+\lfloor\sqrt{n}/C_2\rfloor} (\widehat{\lambda}_i + \frac{\lambda}{n})^{-2}\widehat{\lambda}_i \\
&\ge \frac{1}{n}\sum_{i=k+1}^{k+\lfloor\sqrt{n}/C_2\rfloor} (\widehat{\lambda}_i + n^{-1/4})^{-2}\widehat{\lambda}_i \\
&\ge \frac{1}{n}\sum_{i=k+1}^{k+\lfloor\sqrt{n}/C_2\rfloor} (\frac{3C_1}{2\sqrt{n}} + n^{-1/4})^{-2}\frac{C_1}{2\sqrt{n}}
\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{n}\lfloor\sqrt{n}/C_2\rfloor\frac{C_1}{2}(\frac{3C_1}{2}+n^{1/4})^{-2}\sqrt{n} \\
&\geq \frac{C_1}{16C_2}n^{-1/2},
\end{aligned} \tag{13}
$$

when $n \geq (\frac{3C_1}{2})^4$.

As for the bias term, assume $\lambda \geq n^{3/4}$. Using the same concentration argument, we have $2 > \widehat{\lambda}_i > 1/2$, for $1 \leq i \leq k$. When $\lambda \leq n$, $\lambda_{\max}(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d) \leq 2 + \lambda/n \leq 3$, therefore $\lambda_{\min}((\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}) \geq \frac{1}{3}$. This implies

$$
\begin{aligned}
B &= \frac{\lambda^2}{n^2}\|(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\beta^\star\|^2 \\
&\geq \frac{n^{3/2}}{n^2}\|(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\beta^\star\|^2 \\
&\geq \frac{1}{\sqrt{n}}\lambda_{\min}^2((\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1})\|\beta^\star\|^2 \\
&\geq \frac{\|\beta^\star\|^2}{9\sqrt{n}}.
\end{aligned}
$$

When $\lambda > n$, $\lambda_{\max}(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d) \leq 2 + \lambda/n \leq \frac{3\lambda}{n}$, which means $\lambda_{\min}((\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}) \geq \frac{n}{3\lambda}$ This implies

$$
\begin{aligned}
B &= \frac{\lambda^2}{n^2}\|(\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1}\beta^\star\|^2 \\
&\geq \frac{\lambda^2}{n^2}\lambda_{\min}^2((\widehat{\Sigma}_S + \frac{\lambda}{n}I_d)^{-1})\|\beta^\star\|^2 \\
&\geq \frac{\lambda^2}{n^2}\frac{n^2}{9\lambda^2}\|\beta^\star\|^2 \\
&\geq \frac{\|\beta^\star\|^2}{9}.
\end{aligned}
$$

□

B.2 Upper bound for PCR

In this subsection, we will give the following upper bound for principal component regression.

Theorem 29. When $n \gtrsim \sigma^8(r + \log\frac{1}{\delta})(\frac{\lambda_1}{\lambda_k - \lambda_{k+1}})^2\frac{\lambda_1^2k^2\|\Sigma_T\|^2}{\lambda_k^4\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k})^2}$,

$$
\begin{aligned}
\mathbb{E}_\epsilon\|\widehat{\beta} - \beta^\star\|_{\Sigma_T}^2 \leq \mathcal{O}(&\sigma^8(\frac{\lambda_1}{\lambda_k - \lambda_{k+1}})^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\|(\frac{r + \log\frac{1}{\delta}}{n})\|\beta_k^\star\|^2 + \frac{1}{n}v^2\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k}) \\
&+ \frac{\|\Sigma_{T,k}\|\|\beta_{-k}^\star\|^2\|\Sigma_{S,-k}\|}{\lambda_k} + \beta_{-k}^{\star T}\Sigma_{T,-k}\beta_{-k}^\star)
\end{aligned}
$$

where $r = \frac{\sum_{i=1}^d \lambda_i}{\lambda_1}$.

Proof. For simplicity, we assume we have a sample size of $2n$, and in the first step we obtain an estimator $\widehat{U} \in \mathbb{R}^{d\times k}$ of the top-k subspace $U = \begin{pmatrix} I_k \\ 0 \end{pmatrix} \in \mathbb{R}^{d\times k}$, by using principal component analysis on the sample covariance matrix $\widehat{\Sigma}_S := \frac{1}{n}X^TX = \frac{1}{n}\sum_{i=1}^n x_ix_i^T$, namely $\widehat{U} = (\widehat{u}_1, \cdots, \widehat{u}_k)$ where $\widehat{u}_i$ is the i-th eigenvector of $\widehat{\Sigma}_S$. We denote the distance between the estimated subspace and the original one by $\Delta := \operatorname{dist}(U, \widehat{U}) = \|UU^T - \widehat{U}\widehat{U}^T\|$. For controlling Δ, we have the following lemma (Lemma 6):

Lemma 30. With probability at least $1-\delta$,

$$\Delta \leq C\sigma^4 \left(\sqrt{\frac{r+\log\frac{1}{\delta}}{n}} + \frac{r+\log\frac{1}{\delta}}{n} \right) \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}$$

where $r = \frac{\sum_{i=1}^n \lambda_i}{\lambda_1}$.

In the second step, we do linear regression on the projected (second half) data. With a little abuse of notation, we still use $X \in \mathbb{R}^{n\times d}$ to denote the data matrix indexed from $n+1$ to $2n$. The data here is independent from the data in step 1, and therefore independent of Δ. If we let $Z := X\widehat{U} \in \mathbb{R}^{n\times k}$ be the projected data matrix, the estimator $\widehat{\beta}$ we obtained is given by

$$\begin{aligned} \widehat{\beta} &= \widehat{U}(Z^TZ)^{-1}Z^TY \\ &= \widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TY. \end{aligned} \tag{14}$$

We aim to bound the excess risk on target, which is given by $\|\widehat{\beta}-\beta^\star\|^2_{\Sigma_T} := \|\Sigma_T^{\frac{1}{2}}(\widehat{\beta}-\beta^\star)\|^2$. We introduce the following notations: suppose $\beta^\star = (\beta^*_1, \cdots, \beta^*_d)^T$. We let $\beta^\star_U := (\beta^*_1, \cdots, \beta^\star_k, 0, \cdots, 0)^T$, $\beta^\star_\perp := (0, \cdots, 0, \beta^*_{k+1}, \cdots, \beta^\star_d)^T = \beta^\star - \beta^\star_U$. Here we present an intermediate result for bounding the excess risk:

Lemma 31. Assume $\Delta \leq \frac{\lambda_k^2 \operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k})}{4\lambda_1 k \|\Sigma_T\|}$. When $n \gtrsim \frac{\sigma^4\lambda_1^2\|\Sigma_T\|^2k^3\log(1/\delta)}{\lambda_k^4 \operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k})^2}$, then with probability $1-\delta$,

$$\begin{aligned} \mathbb{E}_\epsilon\|\widehat{\beta}-\beta^\star\|^2_{\Sigma_T} \leq \mathcal{O}(&\|\beta^\star_U\|^2\Delta^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\| + \frac{1}{n}v^2\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k}) \\ &+ \frac{\|\Sigma_{T,k}\|\|\beta^\star_{-k}\|^2\|\Sigma_{S,-k}\|}{\lambda_k} + \beta^{\star T}_{-k}\Sigma_{T,-k}\beta^\star_{-k}) \end{aligned}$$

If further $n \gtrsim \sigma^4\Delta^{-2}k\log(1/\delta)$,

$$\begin{aligned} \mathbb{E}_\epsilon\|\widehat{\beta}-\beta^\star\|^2_{\Sigma_T} \leq \mathcal{O}(&\|\beta^\star_U\|^2(\Delta^4(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\| + \Delta^2\|\Sigma_{T,-k}\| + \Delta^3\|\Sigma_T\|) \\ &+ \frac{1}{n}v^2\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k}) + \frac{\|\Sigma_{T,k}\|\|\beta^\star_{-k}\|^2\|\Sigma_{S,-k}\|}{\lambda_k} + \beta^{\star T}_{-k}\Sigma_{T,-k}\beta^\star_{-k}) \end{aligned}$$

From Lemma 30, when $n \geq r + \log\frac{1}{\delta} = \frac{\sum_{i=1}^n \lambda_i}{\lambda_1} + \log\frac{1}{\delta}$, we have

$$\Delta \leq 2C\frac{\lambda_1}{\lambda_k-\lambda_{k+1}}\sigma^4\sqrt{\frac{r+\log\frac{1}{\delta}}{n}}$$

Therefore when $n \gtrsim (r+\log\frac{1}{\delta})\sigma^8(\frac{\lambda_1}{\lambda_k-\lambda_{k+1}})^2\frac{\lambda_1^2k^2\|\Sigma_T\|^2}{\lambda_k^4\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k})^2}$, the assumption for Δ and n in Lemma 31 will be both satisfied. We can thus apply Lemma 31 to get

$$\begin{aligned} \mathbb{E}_\epsilon\|\widehat{\beta}-\beta^\star\|^2_{\Sigma_T} \leq \mathcal{O}(&\sigma^8(\frac{\lambda_1}{\lambda_k-\lambda_{k+1}})^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\|\frac{r+\log\frac{1}{\delta}}{n}\|\beta^\star_U\|^2 + \frac{1}{n}v^2\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k}) \\ &+ \frac{\|\Sigma_{T,k}\|\|\beta^\star_{-k}\|^2\|\Sigma_{S,-k}\|}{\lambda_k} + \beta^{\star T}_{-k}\Sigma_{T,-k}\beta^\star_{-k}) \end{aligned}$$

where $r = \frac{\sum_{i=1}^d \lambda_i}{\lambda_1}$.

□

B.3 Proofs for Lemma 31

In the following we will prove Lemma 31.

Proof for Lemma 31. The proof idea is similar to (Ge et al., 2023, Theorem 4.4) and (Tripuraneni et al., 2021b, Theorem 4).

We can decompose $\widehat{\beta} - \beta^\star$ as

$$\begin{aligned}\widehat{\beta} - \beta^\star &= \widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1}\widehat{U}^T X^T Y - \beta^\star \\ &= \widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1}\widehat{U}^T X^T (X\beta^\star + \boldsymbol{\epsilon}) - \beta^\star \\ &= \widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1}\widehat{U}^T X^T (X\beta^\star_U + X\beta^\star_\perp + \boldsymbol{\epsilon}) - (\beta^\star_U + \beta^\star_\perp) \\ &= A_1 + A_2 + A_3 - \beta^\star_\perp,\end{aligned}$$

where $A_1 := \widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1}\widehat{U}^T X^T X\beta^\star_U - \beta^\star_U$, $A_2 := \widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1}\widehat{U}^T X^T X\beta^\star_\perp$, $A_3 := \widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1}\widehat{U}^T X^T \boldsymbol{\epsilon}$. Therefore

$$\|\widehat{\beta} - \beta^\star\|^2_{\Sigma_T} \le \|A_1\|^2_{\Sigma_T} + \|A_2\|^2_{\Sigma_T} + \|A_3\|^2_{\Sigma_T} + \|\beta^\star_\perp\|^2_{\Sigma_T} \tag{15}$$

We give three lemmas for bounding the related terms. The first lemma considers the bias term A_1:

Lemma 32. If $\Delta \le \frac{\lambda_k}{4\lambda_1}$ and $n \gtrsim \max\{\sigma^4(\frac{\lambda_1}{\lambda_k})^2 k\log(1/\delta), \sigma^4 k\log(1/\delta)\}$, then with probability at least $1-\delta$,

$$\|A_1\|^2_{\Sigma_T} \le \mathcal{O}(\|\beta^\star_U\|^2\Delta^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\|)$$

If we further have $n \gtrsim \sigma^4\Delta^{-2}k\log(1/\delta)$, then with probability at least $1-\delta$,

$$\|A_1\|^2_{\Sigma_T} \le \mathcal{O}(\|\beta^\star_U\|^2(\Delta^4(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\| + \Delta^2\|\Sigma_{T,-k}\| + \Delta^3\|\Sigma_T\|)) \le \mathcal{O}(\|\beta^\star_U\|^2\Delta^2\|\Sigma_T\|)$$

The second lemma considers the variance term A_3:

Lemma 33. If $\Delta \le \frac{\lambda_k^2 \operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k})}{4\lambda_1 k\|\Sigma_T\|}$ and $n \gtrsim \frac{\sigma^4\|\Sigma_S\|^2\|\Sigma_T\|^2 k^3\log(1/\delta)}{\lambda_k^4 \operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k})^2}$, then with probability at least $1-\delta$,

$$\mathbb{E}_\epsilon[\|A_3\|^2_{\Sigma_T}] \le \mathcal{O}(\frac{1}{n}v^2\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k})).$$

For bounding A_2, we actually have a similar result to bounding A_3:

Lemma 34. If $n \gtrsim \sigma^4(\frac{\lambda_1}{\lambda_k})^2 k\log(1/\delta)$ and $\Delta \le \min\{\frac{\|\Sigma_{T,k}\|}{2\|\Sigma_T\|}, \frac{\lambda_k}{4\lambda_1}\}$, then with probability at least $1-\delta$

$$\|A_2\|^2_{\Sigma_T} \le \mathcal{O}(\frac{\|\Sigma_{T,k}\|\|\beta^\star_{-k}\|^2\|\Sigma_{S,-k}\|}{\lambda_k}) \tag{16}$$

By Lemma 32, 33, 34, together with the decomposition (15), we have with probability $1-\delta$, when $n \gtrsim N_1$,

$$\mathbb{E}_\epsilon\|\widehat{\beta} - \beta^\star\|^2_{\Sigma_T} \le \mathcal{O}(\|\beta^\star_U\|^2\Delta^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\| + \frac{1}{n}v^2\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k}) \tag{17}$$

$$+ \frac{\|\Sigma_{T,k}\|\|\beta^\star_{-k}\|^2\|\Sigma_{S,-k}\|}{\lambda_k} + \beta^{\star T}_{-k}\Sigma_{T,-k}\beta^\star_{-k}) \tag{18}$$

If further $n \gtrsim \sigma^4\Delta^{-2}k\log(1/\delta)$,

$$\mathbb{E}_\epsilon\|\widehat{\beta} - \beta^\star\|^2_{\Sigma_T} \le \mathcal{O}(\|\beta^\star_U\|^2(\Delta^4(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\| + \Delta^2\|\Sigma_{T,-k}\| + \Delta^3\|\Sigma_T\|) \tag{19}$$

$$+ \frac{1}{n}v^2\operatorname{tr}((\Sigma_{S,k})^{-1}\Sigma_{T,k}) + \frac{\|\Sigma_{T,k}\|\|\beta^\star_{-k}\|^2\|\Sigma_{S,-k}\|}{\lambda_k} + \beta^{\star T}_{-k}\Sigma_{T,-k}\beta^\star_{-k}) \tag{20}$$

□

B.4 Technical proofs

In the sequel, we give the proofs of Lemma 32, 33, 34 and 30. We first prove some additional technical lemmas. The following lemma, which is a simple corollary of (Tripuraneni et al., 2021b, Lemma 20), shows the concentration property of empirical covariance matrix.

Lemma 35. Let $\{x_i\}_{i=1}^n$ be i.i.d. $d-$dimensional random vectors, satisfying: x_i is mean zero, $\mathbb{E}[xx^T] = \Sigma$ such that $\sigma_{\max}(\Sigma) \le C_{\max}$ and is $\sigma^2\Sigma$-sub-gaussian, in the sense that

$$\mathbb{E}[\exp(v^T x_i)] \le \exp\left(\frac{\|\sigma\Sigma^{1/2}v\|^2}{2}\right).$$

$X = (x_1, \cdots, x_n)^T \in \mathbb{R}^{n\times d}$. Then for any $A, B \in \mathbb{R}^{d\times k}$, we have with probability at least $1-\delta$

$$\|A^T(\frac{X^TX}{n})B - A^T\Sigma B\|_2 \le \mathcal{O}(\sigma^2\|A\|\|B\|\|\Sigma\|(\sqrt{\frac{k}{n}} + \frac{k}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}). \quad (21)$$

Proof. We write the SVD of A and B: $A = U_1\Lambda_1V_1^T$, $B = U_2\Lambda_2V_2^T$, where $U_1, U_2 \in \mathbb{R}^{d\times k}$, $\Lambda_1, \Lambda_2, V_1, V_2 \in \mathbb{R}^{k\times k}$. Then

$$\begin{aligned}\|A^T(\frac{X^TX}{n})B - A^T\Sigma B\|_2 &= \|V_1\Lambda_1U_1^T(\frac{X^TX}{n})U_2\Lambda_2V_2^T - V_1\Lambda_1U_1^T\Sigma U_2\Lambda_2V_2^T\|_2 \\ &\le \|V_1\Lambda_1\|\|U_1^T(\frac{X^TX}{n})U_2 - U_1^T\Sigma U_2\|\|\Lambda_2V_2^T\| \\ &\le \|A\|\|B\|\|U_1^T(\frac{X^TX}{n})U_2 - U_1^T\Sigma U_2\|. \end{aligned} \quad (22)$$

Now since $U_1, U_2 \in \mathbb{R}^{d\times k}$ are projection matrices, we can apply Tripuraneni et al. (2021b) Lemma 20, therefore

$$\|U_1^T(\frac{X^TX}{n})U_2 - U_1^T\Sigma U_2\| \le \mathcal{O}(\sigma^2\|\Sigma\|(\sqrt{\frac{k}{n}} + \frac{k}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n})) \quad (23)$$

which gives what we want. □

The following lemma is a basic matrix perturbation result (see Tripuraneni et al. (2021b) Lemma 25).

Lemma 36. Let A be a positive definite matrix and E another matrix which satisfies $\|EA^{-1}\| \le \frac{1}{4}$, then $F := (A+E)^{-1} - A^{-1}$ satisfies $\|F\| \le \frac{4}{3}\|A^{-1}\|\|EA^{-1}\|$.

With these two technical lemmas, we are able to prove Lemma 32, 33.

Proof of Lemma 32. Notice that by the definition of U and $\beta_U^\star$, we have $UU^T\beta_U^\star = \beta_U^\star$. We denote $\alpha^\star := U^T\beta_U^\star$, then we also have $\beta_U^\star = U\alpha^\star$. Therefore

$$\begin{aligned}A_1 &= \widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TX\beta_U^\star - \beta_U^\star \\ &= \widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TXU\alpha^\star - U\alpha^\star \\ &= (\widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TXU - U)\alpha^\star\end{aligned}$$

We consider $\widehat{U} \in \mathbb{R}^{d\times k}$ and $\widehat{U}_\perp^T \in \mathbb{R}^{d\times(d-k)}$ be orthonormal projection matrices spanning orthogonal subspaces which are rank k and rank $d-k$ respectively, so that $\mathrm{range}(\widehat{U}) \oplus \mathrm{range}(\widehat{U}_\perp) = \mathbb{R}^d$. Then $\Delta = dist(\widehat{U}, U^\star) = \|\widehat{U}_\perp^TU^\star\|_2$. Notice that $I_d = \widehat{U}\widehat{U}^T + \widehat{U}_\perp\widehat{U}_\perp^T$, we have

$$\begin{aligned}&\widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TXU^\star - U^\star \\ &= \widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TX(\widehat{U}\widehat{U}^T + \widehat{U}_\perp\widehat{U}_\perp^T)U^\star - U^\star \\ &= \widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TX\widehat{U}\widehat{U}^TU^\star + \widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TX\widehat{U}_\perp\widehat{U}_\perp^TU^\star - U^\star \\ &= \widehat{U}(\widehat{U}^TX^TX\widehat{U})^{-1}\widehat{U}^TX^TX\widehat{U}_\perp\widehat{U}_\perp^TU^\star + \widehat{U}\widehat{U}^TU^\star - U^\star\end{aligned}$$

$$= \widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1} \widehat{U}^T X^T X \widehat{U}_\perp \widehat{U}_\perp^T U^\star - \widehat{U}_\perp \widehat{U}_\perp^T U^\star \tag{24}$$

Thus

$$\begin{aligned}
\|A_1\|_{\Sigma_T}^2 &= A_1^T \Sigma_T A_1 \\
&= \alpha^{\star T}(\widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1} \widehat{U}^T X^T X U - U)^T \Sigma_T (\widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1} \widehat{U}^T X^T X U - U)\alpha^\star \\
&= \alpha^{\star T}(\widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1} \widehat{U}^T X^T X \widehat{U}_\perp \widehat{U}_\perp^T U^\star - \widehat{U}_\perp \widehat{U}_\perp^T U^\star)^T \Sigma_T \\
&\quad\ (\widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1} \widehat{U}^T X^T X \widehat{U}_\perp \widehat{U}_\perp^T U^\star - \widehat{U}_\perp \widehat{U}_\perp^T U^\star)\alpha^\star \\
&\le \|\alpha^\star\|^2 \|\widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1} \widehat{U}^T X^T X \widehat{U}_\perp \widehat{U}_\perp^T U^\star - \widehat{U}_\perp \widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2 \\
&\le \|\alpha^\star\|^2 (\|\widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1} \widehat{U}^T X^T X \widehat{U}_\perp \widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2 + \|\widehat{U}_\perp \widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2). \qquad (25)
\end{aligned}$$

Here we use the notation $\|M\|_{\Sigma_T} := \sqrt{\|M^T \Sigma_T M\|}$ for matrix M.

For the second term,

$$\|\widehat{U}_\perp \widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2 \le \|\widehat{U}_\perp^T \Sigma_T \widehat{U}_\perp\| \|\widehat{U}_\perp^T U^\star\|^2 \le \Delta^2 \|\widehat{U}_\perp^T \Sigma_T \widehat{U}_\perp\|. \tag{26}$$

For the first term,

$$\begin{aligned}
&\|\widehat{U}(\widehat{U}^T X^T X \widehat{U})^{-1} \widehat{U}^T X^T X \widehat{U}_\perp \widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2 \\
&= \|\widehat{U}(\widehat{U}^T \frac{X^T X}{n} \widehat{U})^{-1} \widehat{U}^T \frac{X^T X}{n} \widehat{U}_\perp \widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2 \\
&= \|\widehat{U}((\widehat{U}^T \Sigma_S \widehat{U})^{-1} + F)(\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star + E_1)\|_{\Sigma_T}^2 \\
&= \|(\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star + E_1)^T ((\widehat{U}^T \Sigma_S \widehat{U})^{-1} + F)^T \widehat{U}^T \Sigma_T \widehat{U} ((\widehat{U}^T \Sigma_S \widehat{U})^{-1} + F)(\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star + E_1)\| \\
&\le \|\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star + E_1\|^2 \|(\widehat{U}^T \Sigma_S \widehat{U})^{-1} + F\|^2 \|\widehat{U}^T \Sigma_T \widehat{U}\| \\
&\le (\|\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star\| + \|E_1\|)^2 (\|(\widehat{U}^T \Sigma_S \widehat{U})^{-1}\| + \|F\|)^2 \|\widehat{U}^T \Sigma_T \widehat{U}\| \qquad (27)
\end{aligned}$$

where $E_1 = \widehat{U}^T \frac{X^T X}{n} \widehat{U}_\perp \widehat{U}_\perp^T U^\star - \widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star$, $F = (\widehat{U}^T \frac{X^T X}{n} \widehat{U})^{-1} - (\widehat{U}^T \Sigma_S \widehat{U})^{-1}$. We aim to show that $\|E_1\| \le \|\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star\|$ and $\|F\| \le \|(\widehat{U}^T \Sigma_S \widehat{U})^{-1}\| = C_{\text{min}}^{-1}$ for sufficiently large n, therefore the term in (27) can be bounded well. First we need a careful analysis of $\|\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star\|$. It is obvious that

$$\|\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star\| \le \|\widehat{U}^T \Sigma_S \widehat{U}_\perp\| \|\widehat{U}_\perp^T U^\star\| \le \Delta \|\widehat{U}^T \Sigma_S \widehat{U}_\perp\|. \tag{28}$$

As for $\|\widehat{U}^T \Sigma_S \widehat{U}_\perp\|$, notice that if without the "hat", we have $U^T \Sigma_S U_\perp = 0$ by the definition of U and Σ_S is diagonal. By definition of distance between two subspaces, there exist $R \in \mathcal{O}^{k \times k}$ and $Q \in \mathcal{O}^{(d-k)\times(d-k)}$, such that $\|\widehat{U}R - U\| = \Delta = \|\widehat{U}_\perp Q - U_\perp\|$. Then we have

$$\begin{aligned}
\|\widehat{U}^T \Sigma_S \widehat{U}_\perp\| &= \|R^T \widehat{U}^T \Sigma_S \widehat{U}_\perp Q\| \\
&= \|U^T \Sigma_S U_\perp + R^T \widehat{U}^T \Sigma_S \widehat{U}_\perp Q - U^T \Sigma_S U_\perp\| \\
&= \|R^T \widehat{U}^T \Sigma_S \widehat{U}_\perp Q - U^T \Sigma_S U_\perp\| \\
&= \|R^T \widehat{U}^T \Sigma_S \widehat{U}_\perp Q - U^T \Sigma_S \widehat{U}_\perp Q + U^T \Sigma_S \widehat{U}_\perp Q - U^T \Sigma_S U_\perp\| \\
&\le \|R^T \widehat{U}^T \Sigma_S \widehat{U}_\perp Q - U^T \Sigma_S \widehat{U}_\perp Q\| + \|U^T \Sigma_S \widehat{U}_\perp Q - U^T \Sigma_S U_\perp\| \\
&\le \|R^T \widehat{U}^T - U^T\| \|\Sigma_S \widehat{U}_\perp Q\| + \|U^T \Sigma_S\| \|\widehat{U}_\perp Q - U_\perp\| \\
&\le 2\Delta \|\Sigma_S\|. \qquad (29)
\end{aligned}$$

Combine (28) and (29), we have

$$\|\widehat{U}^T \Sigma_S \widehat{U}_\perp \widehat{U}_\perp^T U^\star\| \le \mathcal{O}(\Delta^2 \|\Sigma_S\|) \tag{30}$$

In order to bound $\|F\|$, let $E = \widehat{U}^T \frac{X^T X}{n} \widehat{U} - \widehat{U}^T \Sigma_S \widehat{U}$, then by Lemma 35, with probability at least $1 - \delta$,

$$\|E\| \le \mathcal{O}(\sigma^2 \|\Sigma_S\| (\sqrt{\frac{k}{n}} + \frac{k}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n})). \tag{31}$$

Therefore,

$$\begin{aligned}\|E(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\| &\leq \|E\|\|(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\| \\ &\leq \|E\|C_{\text{min}}^{-1} \\ &\leq \mathcal{O}(\sigma^2 C_{\text{min}}^{-1}\|\Sigma_S\|(\sqrt{\frac{k}{n}}+\frac{k}{n}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n})),\end{aligned} \tag{32}$$

where $C_{\text{min}} := \lambda_{\text{min}}(\widehat{U}^T\Sigma_S\widehat{U})$. Notice that $n \gtrsim \sigma^4 C_{\text{min}}^{-2}\|\Sigma_S\|^2 k\log(1/\delta)$ implies $\sqrt{\frac{k}{n}}+\frac{k}{n}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n} \lesssim \sigma^{-2}C_{\text{min}}\|\Sigma_S\|^{-1}$. Thus, we show that when n is large enough, we have $\|E(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\| \leq \frac{1}{4}$. Therefore we can apply Lemma 36, which gives

$$\begin{aligned}\|F\| &\leq \frac{4}{3}\|E(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\|\|(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\| \\ &\leq \frac{4}{3}\times\frac{1}{4}\|(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\| \\ &\leq \frac{1}{3}C_{\text{min}}^{-1}.\end{aligned} \tag{33}$$

As for $\|E_1\|$, directly applying Lemma 35, when $n \gtrsim \sigma^4\Delta^{-2}k\log(1/\delta)$ we get

$$\begin{aligned}\|E_1\| &\leq \mathcal{O}(\sigma^2\|\Sigma_S\|\|\widehat{U}_\perp\widehat{U}_\perp^T U^\star\|(\sqrt{\frac{k}{n}}+\frac{k}{n}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n})) \\ &\leq \mathcal{O}(\sigma^2\|\Sigma_S\|\Delta(\sqrt{\frac{k}{n}}+\frac{k}{n}+\sqrt{\frac{\log(1/\delta)}{n}}+\frac{\log(1/\delta)}{n}))\end{aligned} \tag{34}$$

when $n \gtrsim \sigma^4 k\log(1/\delta)$ we have

$$\|E_1\| \leq \mathcal{O}(\Delta\|\Sigma_S\|) \tag{35}$$

, if further we have $n \gtrsim \sigma^4\Delta^{-2}k\log(1/\delta)$, then

$$\|E_1\| \leq \mathcal{O}(\Delta^2\|\Sigma_S\|). \tag{36}$$

Combining (27), (30), (33) and (36), we have

$$\begin{aligned}&\|\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T X\widehat{U}_\perp\widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2 \\ &\leq (\|\widehat{U}^T\Sigma_S\widehat{U}_\perp\widehat{U}_\perp^T U^\star\|+\|E_1\|)^2(\|(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\|+\|F\|)^2\|\widehat{U}^T\Sigma_T\widehat{U}\| \\ &\leq \mathcal{O}(\Delta^4\|\Sigma_S\|^2 C_{\text{min}}^{-2}\|\widehat{U}^T\Sigma_T\widehat{U}\|) \\ &\leq \mathcal{O}(\Delta^4\|\Sigma_S\|^2 C_{\text{min}}^{-2}\|\Sigma_T\|)\end{aligned} \tag{37}$$

Combining (25),(26) and (37), we get

$$\begin{aligned}\|A_1\|_{\Sigma_T}^2 &\leq \|\alpha^\star\|^2(\|\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T X\widehat{U}_\perp\widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2+\|\widehat{U}_\perp\widehat{U}_\perp^T U^\star\|_{\Sigma_T}^2) \\ &\leq \mathcal{O}(\|\alpha^\star\|^2(\Delta^4\|\Sigma_S\|^2 C_{\text{min}}^{-2}\|\Sigma_T\|+\Delta^2\|\widehat{U}_\perp^T\Sigma_T\widehat{U}_\perp\|))\end{aligned} \tag{38}$$

with probability at least $1-\delta$. Also, similar to (29), we have

$$\begin{aligned}\|\widehat{U}_\perp^T\Sigma_T\widehat{U}_\perp\| &= \|Q^T\widehat{U}_\perp^T\Sigma_T\widehat{U}_\perp Q\| \\ &\leq \|U_\perp^T\Sigma_T U_\perp\|+\|Q^T\widehat{U}_\perp^T\Sigma_T\widehat{U}_\perp Q-U_\perp^T\Sigma_T U_\perp\| \\ &\leq \|U_\perp^T\Sigma_T U_\perp\|+2\Delta\|\Sigma_T\|\end{aligned} \tag{39}$$

Similarly, we can further know that C_{min} is close to λ_k:

$$\begin{aligned}C_{\text{min}} &= \lambda_k(\widehat{U}^T\Sigma_S\widehat{U}) \\ &= \lambda_k(R^T\widehat{U}^T\Sigma_S\widehat{U}R) \\ &= \lambda_k(U^T\Sigma_S U+R^T\widehat{U}^T\Sigma_S\widehat{U}R-U^T\Sigma_S U)\end{aligned}$$

$$
\begin{aligned}
&\geq \lambda_k(U^T\Sigma_S U) - \|R^T\widehat{U}^T\Sigma_S\widehat{U}R - U^T\Sigma_S U\| \\
&\geq \lambda_k(U^T\Sigma_S U)2\Delta\|\Sigma_S\| \\
&\geq \lambda_k - 2\lambda_1\Delta \\
&\geq \frac{1}{2}\lambda_k, \qquad (40)
\end{aligned}
$$

where the last inequality holds when $\Delta \leq \frac{\lambda_k}{4\lambda_1}$. Finally, combining (38), (39), (40), we have

$$
\begin{aligned}
\|A_1\|_{\Sigma_T}^2 &\leq \mathcal{O}(\|\alpha^\star\|^2(\Delta^4(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\| + \Delta^2\|U_\perp^T\Sigma_T U_\perp\| + \Delta^3\|\Sigma_T\|)) \\
&\leq \mathcal{O}(\|\beta_U^\star\|^2(\Delta^4(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\| + \Delta^2\|U_\perp^T\Sigma_T U_\perp\| + \Delta^3\|\Sigma_T\|)) \qquad (41)
\end{aligned}
$$

when $\Delta \leq \frac{\lambda_k}{4\lambda_1}$ and $n \gtrsim \max\{\sigma^4(\frac{\lambda_1}{\lambda_k})^2 k\log(1/\delta), \sigma^4\Delta^{-2}k\log(1/\delta)\}$. If in the previous proofs we replace (36) by (35), we have

$$
\begin{aligned}
\|A_1\|_{\Sigma_T}^2 &\leq \mathcal{O}(\|\beta_U^\star\|^2(\Delta^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\| + \Delta^2\|U_\perp^T\Sigma_T U_\perp\| + \Delta^3\|\Sigma_T\|)) \qquad (42) \\
&\leq \mathcal{O}(\|\beta_U^\star\|^2\Delta^2(\frac{\lambda_1}{\lambda_k})^2\|\Sigma_T\|) \qquad (43)
\end{aligned}
$$

when $\Delta \leq \frac{\lambda_k}{4\lambda_1}$ and $n \gtrsim \max\{\sigma^4(\frac{\lambda_1}{\lambda_k})^2 k\log(1/\delta), \sigma^4 k\log(1/\delta)\}$. Notice that by definition of U, $U_\perp^T\Sigma_T U_\perp = \Sigma_{T,-k}$, therefore the result is exactly what we want. □

Proof of Lemma 33. Recall $A_3 := \widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T\boldsymbol{\epsilon}$. Therefore

$$
\begin{aligned}
\|A_3\|_{\Sigma_T}^2 &= \boldsymbol{\epsilon}^T X\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T\boldsymbol{\epsilon} \\
&= \mathrm{tr}(\boldsymbol{\epsilon}^T X\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T\boldsymbol{\epsilon}) \\
&= \mathrm{tr}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T X\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T)
\end{aligned}
$$

Taking expectation with respect to ϵ, using $\mathbb{E}[\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T] = v^2 I_n$, we have

$$
\begin{aligned}
\mathbb{E}_\epsilon[\|A_3\|_{\Sigma_T}^2] &= \mathbb{E}[\mathrm{tr}(\boldsymbol{\epsilon}\boldsymbol{\epsilon}^T X\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T)] \\
&= v^2\,\mathrm{tr}(X\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T) \\
&= v^2\,\mathrm{tr}((\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T X\widehat{U}) \\
&= v^2\,\mathrm{tr}((\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}) \\
&= \frac{1}{n}v^2\,\mathrm{tr}(((\widehat{U}^T\Sigma_S\widehat{U})^{-1} + F)\widehat{U}^T\Sigma_T\widehat{U}) \qquad (44)
\end{aligned}
$$

Here we actually need a bound stronger than (33) for $\|F\|$: recall (32), we have with probability $1-\delta$

$$
\|E(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\| \leq \mathcal{O}(\sigma^2 C_{\text{min}}^{-1}\|\Sigma_S\|(\sqrt{\frac{k}{n}} + \frac{k}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n})). \qquad (45)
$$

Applying Lemma 36, which gives

$$
\begin{aligned}
\|F\| &\leq \frac{4}{3}\|E(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\|\|(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\| \\
&\leq \mathcal{O}(\sigma^2 C_{\text{min}}^{-2}\|\Sigma_S\|(\sqrt{\frac{k}{n}} + \frac{k}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n})) \\
&\leq \mathcal{O}(\frac{1}{k\|\Sigma_T\|}\,\mathrm{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)) \qquad (46)
\end{aligned}
$$

when $n \gtrsim \sigma^4 C_{\min}^{-4} \|\Sigma_S\|^2 \|\Sigma_T\|^2 \operatorname{tr}((U^T\Sigma_S U)^{-1} U^T \Sigma_T U)^{-2} k^3 \log(1/\delta)$. Therefore we have

$$\begin{aligned}
\mathbb{E}_\epsilon[\|A_3\|_{\Sigma_T}^2] &= \frac{1}{n} v^2 \operatorname{tr}(((\widehat{U}^T\Sigma_S\widehat{U})^{-1} + F)\widehat{U}^T\Sigma_T\widehat{U}) \\
&= \frac{1}{n} v^2 (\operatorname{tr}((\widehat{U}^T\Sigma_S\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}) + \operatorname{tr}(F\widehat{U}^T\Sigma_T\widehat{U})) \\
&\le \frac{1}{n} v^2 (\operatorname{tr}((\widehat{U}^T\Sigma_S\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U})) + \frac{1}{n} v^2 \|F\| \operatorname{tr}(\widehat{U}^T\Sigma_T\widehat{U}) \\
&\le \frac{1}{n} v^2 (\operatorname{tr}((\widehat{U}^T\Sigma_S\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U})) + \frac{1}{n} v^2 k \|F\| \|\Sigma_T\| \\
&\le \frac{1}{n} v^2 (\operatorname{tr}((\widehat{U}^T\Sigma_S\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U})) + \frac{1}{n} v^2 \mathcal{O}(\operatorname{tr}((U^T\Sigma_S U)^{-1} U^T\Sigma_T U)) \qquad (47)
\end{aligned}$$

The remaining thing is to show that indeed $\operatorname{tr}((\widehat{U}^T\Sigma_S\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U})$ is close to $\operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)$. In fact, $\operatorname{tr}((\widehat{U}^T\Sigma_S\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}) = \operatorname{tr}((R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1}R^T\widehat{U}^T\Sigma_T R\widehat{U})$. Notice that

$$\|R^T\widehat{U}^T\Sigma_T\widehat{U}R - U^T\Sigma_T U\| \le 2\|\Delta\|\|\Sigma_T\|,$$

we have

$$\begin{aligned}
&\operatorname{tr}((R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1}R^T\widehat{U}^T\Sigma_T\widehat{U}R) \qquad (48) \\
\le\ &\operatorname{tr}((R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1}U^T\Sigma_T U) + \|R^T\widehat{U}^T\Sigma_T\widehat{U}R - U^T\Sigma_T U\| \operatorname{tr}((\widehat{U}^T\Sigma_S\widehat{U})^{-1}) \\
\le\ &\operatorname{tr}((R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1}U^T\Sigma_T U) + 2\|\Delta\|\|\Sigma_T\| \operatorname{tr}((\widehat{U}^T\Sigma_S\widehat{U})^{-1}) \\
\le\ &\operatorname{tr}((R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1}U^T\Sigma_T U) + 2\|\Delta\|\|\Sigma_T\| k C_{\min}^{-1} \\
\le\ &\operatorname{tr}((R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1}U^T\Sigma_T U) + \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U) \qquad (49)
\end{aligned}$$

when $\Delta \le \frac{\lambda_k \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)}{4k\|\Sigma_T\|}$. Also, we have

$$\begin{aligned}
\|(R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1} - (U^T\Sigma_S U)^{-1}\| &\le \|(R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1}\|\|(U^T\Sigma_S U)^{-1}\|\|R^T\widehat{U}^T\Sigma_S\widehat{U}R - U^T\Sigma_S U\| \\
&\le 4\lambda_k^{-2}\lambda_1\Delta,
\end{aligned}$$

therefore

$$\begin{aligned}
\operatorname{tr}((R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1}U^T\Sigma_T U) &\le \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U) + \|(R^T\widehat{U}^T\Sigma_S\widehat{U}R)^{-1} - (U^T\Sigma_S U)^{-1}\| \operatorname{tr}(U^T\Sigma_T U) \\
&\le \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U) + 4\lambda_k^{-2}\lambda_1\Delta \operatorname{tr}(U^T\Sigma_T U) \\
&\le 2\operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U), \qquad (50)
\end{aligned}$$

if $\Delta \le \frac{\lambda_k^2 \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)}{4\lambda_1 \operatorname{tr}(U^T\Sigma_T U)}$. Combining (47), (48) and (50) we have

$$\mathbb{E}_\epsilon[\|A_3\|_{\Sigma_T}^2] \le \mathcal{O}(\frac{1}{n} v^2 \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)),$$

whenever $\Delta \le \frac{\lambda_k^2 \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)}{4\lambda_1 k\|\Sigma_T\|} \le \min\{\frac{\lambda_k^2 \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)}{4\lambda_1 \operatorname{tr}(U^T\Sigma_T U)}, \frac{\lambda_k \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)}{4k\|\Sigma_T\|}\}$ and $n \gtrsim \sigma^4 C_{\min}^{-4}\|\Sigma_S\|^2\|\Sigma_T\|^2 \operatorname{tr}((U^T\Sigma_S U)^{-1}U^T\Sigma_T U)^{-2} k^3 \log(1/\delta)$, with probability at least $1-\delta$. Notice that $U^T\Sigma_S U = \Sigma_{S,k}$ and $U^T\Sigma_T U = \Sigma_{T,k}$, therefore the result is exactly what we want. $\square$

Proof of Lemma 34. Recall $A_2 := \widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T X\beta_\perp^\star$. Also we have

$$\begin{aligned}
\|\widehat{U}^T\Sigma_T\widehat{U}\| &= \|R^T\widehat{U}^T\Sigma_T\widehat{U}R\| \\
&\le \|U^T\Sigma_T U\| + \|R^T\widehat{U}^T\Sigma_T\widehat{U}R - U^T\Sigma_T U\| \\
&\le \|U^T\Sigma_T U\| + 2\Delta\|\Sigma_T\| \qquad (51)
\end{aligned}$$

Therefore

$$\begin{aligned}\|A_2\|_{\Sigma_T}^2 &= \|\beta_\perp^{\star T} X^T X\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T\Sigma_T\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T X\beta_\perp^\star\| \\ &\le \|X\widehat{U}(\widehat{U}^T X^T X\widehat{U})^{-1}(\widehat{U}^T X^T X\widehat{U})^{-1}\widehat{U}^T X^T\|\|\widehat{U}^T\Sigma_T\widehat{U}\|\|X\beta_\perp^\star\|^2 \\ &\le \|A\|(\|U^T\Sigma_T U\| + 2\Delta\|\Sigma_T\|)\|X\beta_\perp^\star\|^2 \\ &\le 2\|A\|\|U^T\Sigma_T U\|\|X\beta_\perp^\star\|^2 \end{aligned} \tag{52}$$

when $\Delta \le \frac{\|U^T\Sigma_T U\|}{2\|\Sigma_T\|}$, where we let $A = \frac{1}{n}\frac{X\widehat{U}}{\sqrt{n}}(\widehat{U}^T\frac{X^T X}{n}\widehat{U})^{-2}\frac{\widehat{U}^T X^T}{\sqrt{n}}$. If we define $B = \frac{X\widehat{U}}{\sqrt{n}} \in \mathbb{R}^{n\times r}$, then $A = \frac{1}{n}B(B^T B)^{-2}B^T$. Let the SVD of B be $B = PMO^T$, where $P \in \mathbb{R}^{n\times k}$, $M, O \in \mathbb{R}^{k\times k}$, then

$$\begin{aligned}\|A\|_2 &= \frac{1}{n}\|B(B^T B)^{-2}B^T\|_2 \\ &= \frac{1}{n}\|PMO^T(OM^2O^T)^{-2}OMP^T\|_2 \\ &= \frac{1}{n}\|PM^{-2}P^T\|_2 \\ &\le \frac{1}{n}\|M^{-2}\|_2 \\ &= \frac{1}{n}\|(B^T B)^{-1}\|_2 \end{aligned} \tag{53}$$

Let $F = (\widehat{U}^T\frac{X^T X}{n}\widehat{U})^{-1} - (\widehat{U}^T\Sigma\widehat{U})^{-1}$. Recall (33), which states that with probability at least $1-\delta$, we have $\|F\| \le \frac{1}{3}C_{\min}^{-1} \le \frac{2}{3}\lambda_k^{-1}$ when $n \gtrsim \sigma^4 C_{\min}^{-2}\|\Sigma_S\|^2 k\log(1/\delta)$ and $\Delta \le \frac{\lambda_k}{4\lambda_1}$. Therefore

$$\begin{aligned}\|A\| &\le \frac{1}{n}\|(\widehat{U}^T\frac{X^T X}{n}\widehat{U})^{-1}\| \\ &= \|(\widehat{U}^T\Sigma_S\widehat{U})^{-1} + F\| \\ &\le \frac{1}{n}\|(\widehat{U}^T\Sigma_S\widehat{U})^{-1}\| + \|F\| \\ &\le \mathcal{O}(\frac{1}{n}\lambda_k^{-1}). \end{aligned} \tag{54}$$

Thus $\|A\| \le \mathcal{O}(\lambda_k^{-1})$. As for $\|X\beta_\perp^\star\|^2$, notice that the first-k entries of $\beta_\perp^\star$ are zero, therefore $X\beta_\perp^\star = X_{-k}\beta_{-k}^\star$. by Lemma 35,

$$\|\beta_{-k}^{\star T}(\frac{X_{-k}^T X_{-k}}{n})\beta_{-k}^\star - \beta_{-k}^{\star T}\Sigma_{S,-k}\beta_{-k}^\star\| \le \mathcal{O}(\sigma^2\|\beta_{-k}^\star\|^2\|\Sigma_{S,-k}\|(\sqrt{\frac{1}{n}} + \frac{1}{n} + \sqrt{\frac{\log(1/\delta)}{n}} + \frac{\log(1/\delta)}{n}). \tag{55}$$

Therefore we have

$$\begin{aligned}\|X\beta_\perp^\star\|^2 &= n\beta_{-k}^{\star T}(\frac{X_{-k}^T X_{-k}}{n})\beta_{-k}^\star \\ &\le n(\beta_{-k}^{\star T}\Sigma_{S,-k}\beta_{-k}^\star + \|\beta_{-k}^{\star T}(\frac{X_{-k}^T X_{-k}}{n})\beta_{-k}^\star - \beta_{-k}^{\star T}\Sigma_{S,-k}\beta_{-k}^\star\|) \\ &\le \mathcal{O}(n\|\beta_{-k}^\star\|^2\|\Sigma_{S,-k}\|). \end{aligned} \tag{56}$$

Combining (52)(54) and (56), we have

$$\|A_2\|_{\Sigma_T}^2 \le \mathcal{O}(\frac{\|U^T\Sigma_T U\|\|\beta_{-k}^\star\|^2\|\Sigma_{S,-k}\|}{\lambda_k}) \tag{57}$$

when $n \gtrsim \sigma^4 C_{\min}^{-2}\|\Sigma_S\|^2 k\log(1/\delta)$ and $\Delta \le \min\{\frac{\|U^T\Sigma_T U\|}{2\|\Sigma_T\|}, \frac{\lambda_k}{4\lambda_1}\}$. □

Finally we prove Lemma 30 in the following.

Proof of Lemma 30. In the first step, we obtain $\widehat{U} \in \mathbb{R}^{d\times k}$ by selecting the top$-k$ eigenvectors of the sample covariance matrix $\widehat{\Sigma}_S := \frac{1}{n}XX^T = \frac{1}{n}\sum_{i=1}^n x_i x_i^T$ using PCA. Then by Davis-Kahan theorem (Chen et al., 2021, Corollary 2.8),

$$\Delta \leq \frac{2\|\widehat{\Sigma}_S - \Sigma_S\|}{\lambda_k - \lambda_{k+1}}. \tag{58}$$

Therefore it remains to bound $\|\widehat{\Sigma}_S - \Sigma_S\|$. Applying Lemma 28, we immediately have

$$\|\widehat{\Sigma}_S - \Sigma_S\| \leq C\sigma^4 \left(\sqrt{\frac{r + \log\frac{1}{\delta}}{n}} + \frac{r + \log\frac{1}{\delta}}{n} \right) \lambda_1$$

where $r = \frac{\sum_{i=1}^n \lambda_i}{\lambda_1}$. Together with (58), we have with probability at least $1-\delta$,

$$\Delta \leq C\sigma^4 \left(\sqrt{\frac{r + \log\frac{1}{\delta}}{n}} + \frac{r + \log\frac{1}{\delta}}{n} \right) \frac{\lambda_1}{\lambda_k - \lambda_{k+1}}.$$

$\square$