

On Ridge Estimation in High-dimensional Rotationally Sparse Linear Regression

Libin Liang¹ and Zhiqiang Tan¹

May 3, 2024

Abstract. Recently, deep neural networks have been found to nearly interpolate training data but still generalize well in various applications. To help understand such a phenomenon, it has been of interest to analyze the ridge estimator and its interpolation limit in high-dimensional regression models. For this motivation, we study the ridge estimator in a rotationally sparse setting of high-dimensional linear regression, where the signal of a response is aligned with a small number, d , of covariates with large or spiked variances, compared with the remaining covariates with small or tail variances, *after* an orthogonal transformation of the covariate vector. We establish high-probability upper and lower bounds on the out-sample and in-sample prediction errors in two distinct regimes depending on the ratio of the effective rank of tail variances over the sample size n . The separation of the two regimes enables us to exploit relevant concentration inequalities and derive concrete error bounds without making any oracle assumption or independent components assumption on covariate vectors. Moreover, we derive sufficient and necessary conditions which indicate that the prediction errors of ridge estimation can be of the order $O(\frac{d}{n})$ if and only if the gap between the spiked and tail variances are sufficiently large. We also compare the orders of optimal out-sample and in-sample prediction errors and find that, remarkably, the optimal out-sample prediction error may be significantly smaller than the optimal in-sample one. Finally, we present numerical experiments which empirically confirm our theoretical findings.

Key words and phrases. High-dimensional regression; In-sample prediction error; Interpolating estimator; Linear regression; Out-sample prediction error; Ridge estimation.

¹Department of Statistics, Rutgers University. Address: 110 Frelinghuysen Road, Piscataway, NJ 08854. E-mails: ll866@stat.rutgers.edu, ztan@stat.rutgers.edu. The authors thank Qiyang Han for helpful comments related to error approximation formulas in Section 4.1.

1 Introduction

Over-parameterized models, in particular, deep neural networks, have been successful when trained without penalty or with a mild penalty in various applications. This phenomenon appears to be at odds with conventional statistical thinking that complex models tend to overfit the training data and generalize poorly without proper regularization (Belkin et al., 2019; C. Zhang et al., 2016). Considerable research has been devoted to studying such phenomena; see Bartlett et al. (2020), Bunea et al. (2022), Hastie et al. (2022), and Tsigler & Bartlett (2023) among others.

We consider high-dimensional linear regression as a basic example of over-parameterized models and study ridge estimation under a rotationally sparse setting, which will be introduced shortly. Suppose that the training data, $(y_1, x_1), \dots, (y_n, x_n)$, are i.i.d. from the following model:

$$y_i = x_i^T \theta^* + \epsilon_i, \quad i = 1, \dots, n, \quad (1)$$

where $\theta^* \in \mathbb{R}^p$, $y_i \in \mathbb{R}$ is a response variable, $x_i \in \mathbb{R}^p$ is a covariate vector satisfying $E(x_i) = 0$ and $\text{Var}(x_i) = \Sigma$ (assumed to be non-singular), and ϵ_i is a noise variable, independent of x_i and satisfying $E(\epsilon_i) = 0$ and $\text{Var}(\epsilon_i) = \sigma^2$ for $i = 1, \dots, n$. In addition, assume that $\Sigma^{-1/2}x_i$ is a sub-gaussian random vector with sub-gaussian norm σ_x for $i = 1, \dots, n$. See Supplement Section I for the definition of the sub-gaussian norm.

We are interested in the ridge estimator defined as

$$\hat{\theta}(\tau) = \operatorname{argmin}_{\theta \in \mathbb{R}^p} \left\{ \frac{1}{n} \|Y - X\theta\|^2 + \tau \|\theta\|^2 \right\}, \quad (2)$$

where $Y = (y_1, \dots, y_n)^T \in \mathbb{R}^p$, $X = (x_1, \dots, x_n)^T \in \mathbb{R}^{n \times p}$, $\tau \geq 0$ is a tuning parameter, and $\|\cdot\|$ denotes the L_2 norm. In the case of $\tau = 0$, the estimator $\hat{\theta}(0)$ is defined as the limit of $\hat{\theta}(\tau)$ as $\tau \rightarrow 0+$, and called a min-norm interpolator.

As a measure of prediction performance, we investigate both the out-sample and in-sample mean squared errors (MSE) of $\hat{\theta}(\tau)$, defined conditionally on X as follows.

- Out-sample error

$$\text{MSE}_{\text{out}} = E[(x_0^T (\hat{\theta}(\tau) - \theta^*))^2 | X] = E[\|\hat{\theta}(\tau) - \theta^*\|_{\Sigma}^2 | X],$$

where $x_0 \in \mathbb{R}^p$ is a new covariate vector independent of X , and $\|b\|_M = (b^T M b)^{1/2}$ for any positive semi-definite matrix M and any vector b of the suitable dimension.

- In-sample error

$$\text{MSE}_{\text{in}} = E\left[\frac{1}{n} \|X(\hat{\theta}(\tau) - \theta^*)\|^2 | X\right] = E[\|\hat{\theta}(\tau) - \theta^*\|_{\hat{\Sigma}}^2 | X],$$

where $\hat{\Sigma} = X^T X / n$ is the (uncentered) sample covariance matrix of X .

An important property of the ridge estimator is that its out-sample and in-sample prediction errors are invariant to an orthonormal transformation of the covariate vectors x_i . By this property, we assume without loss of generality that the covariance matrix Σ is diagonal, i.e., $\Sigma = \text{Diag}(\lambda_1, \dots, \lambda_p)$, where $\lambda_1 \geq \dots \geq \lambda_p > 0$ are the eigenvalues of Σ .

As motivated by the recent literature (Bartlett et al., 2020; Bunea et al., 2022; Hastie et al., 2022), we assume that only a few covariates (or a few directions of the covariate vector before orthogonalization) with large variances contain most of the information about the response. Formally, we assume that the mean response is aligned with a small number of covariates with large variances, for instance, the first $1 \leq d < p$ covariates:

$$\frac{\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2} \approx 0.$$

We denote as $\theta_{1:d}^*$ the first d entries of θ^* and as $\theta_{(d+1):p}^*$ the remaining $p - d$ entries of θ^* , which are called the spiked part and tail part respectively. In addition, we denote $\Sigma_{1:d} = \text{Diag}(\lambda_1, \dots, \lambda_d)$ and $\Sigma_{(d+1):p} = \text{Diag}(\lambda_{d+1}, \dots, \lambda_p)$. We refer to such a setting as a rotationally sparse setting, because a small number of coefficients $\theta_{1:d}^*$ are nonzero whereas the remaining ones $\theta_{(d+1):p}^*$ are zero or close to zero, *after* an orthogonal transformation of the covariates. From both the recent literature and our results later, a sufficiently large gap between λ_d and λ_{d+1} may be necessary and sufficient for ridge estimation to achieve meaningful prediction performance.

The rotationally sparse setting fundamentally differs from the directly sparse setting of regression models, where a small number of coefficients, for example, s , are nonzero whereas the remaining $p - s$ ones are zero or close to zero, with the original covariate vectors. Such a sparse structure depends on the particular coordinate system for the covariate vectors and would be lost after an orthogonal transformation. For the directly sparse setting, Lasso estimation is known to be effective in achieving (in-sample) prediction errors in the order $O(s \frac{\log p}{n})$, that is, $O(\frac{s}{n})$ up to a logarithmic factor of p under suitable conditions (including compatibility or restricted eigenvalue conditions) (e.g., Bickel et al. (2009); Bühlmann & van de Geer (2011)). Hence one of the interesting questions which motivate our work is to investigate plausible conditions for ridge estimation to achieve prediction errors in the order $O(\frac{d}{n})$ in the rotationally sparse linear regression.

Our work. We study the out-sample and in-sample prediction errors of the ridge estimator in high-dimensional rotationally sparse linear regression. Classical analysis of ridge estimation deals with the in-sample MSE in a fixed design (with fixed X), whereas recent work on the ridge estimator and the min-norm interpolator, as reviewed later, has been focused on the out-sample MSE in a random design (with random X). Nevertheless, it is also of interest to study the in-

Table 1: Summary of our main results in the rotationally sparse setting

Regime I (small or moderate TER): $r_d(\Sigma) \lesssim n$

Range of τ	Out-sample error [Theorem 1]	In-sample error [Theorem 2]
$\tau < \lambda_{d+1}$	$\text{MSE}_{\text{out}} \gtrsim \sigma^2(\frac{d}{n} + \frac{r_d(\Sigma^2)}{n})$	$\text{MSE}_{\text{in}} \gtrsim \sigma^2(\frac{d}{n} + \frac{r_d^2(\Sigma)}{n^2})$
$\lambda_{d+1} \leq \tau \leq \lambda_d$	$\text{MSE}_{\text{out}} \asymp \ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n})$	Assume $\tau \gg \lambda_{d+1}$: $\ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma)}{n})$ $\gtrsim \text{MSE}_{\text{in}} \gtrsim \ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2})$
$\tau > \lambda_d$	$\text{MSE}_{\text{out}} \gtrsim \ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 \lambda_d^2$	Assume $\tau \gg \lambda_{d+1}$: $\text{MSE}_{\text{in}} \gtrsim \ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 \lambda_d^2$

Regime II (large TER): $r_d(\Sigma) \geq c_x n$

Range of τ	Out-sample error [Theorem 3]	In-sample error [Theorem 4]
$\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$	$\text{MSE}_{\text{out}} \asymp \ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 (\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n})$	Assume $\tau \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$: $\text{MSE}_{\text{in}} \asymp \ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 (\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2})$
$\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_d$	$\text{MSE}_{\text{out}} \gtrsim \ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 \lambda_d^2$	Assume $\tau \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$: $\text{MSE}_{\text{in}} \gtrsim \ \theta_{1:d}^*\ _{\Sigma_{1:d}^{-1}}^2 \lambda_d^2$

Note: c_x is a constant depending only on the sub-gaussian norm σ_x of covariate vectors.

sample MSE in a random design. In fact, the standard analysis of Lasso estimation concerns the in-sample MSE either in a fixed design or in a random design (e.g., Bickel et al. (2009); Bühlmann & van de Geer (2011)). The in-sample MSE in a random design also plays an important role in analyzing de-biased Lasso estimation (C.-H. Zhang & Zhang, 2014; van de Geer et al., 2014) and high-dimensional estimation of average treatment effects (Chernozhukov et al., 2018; Tan, 2020).

The main findings of our work can be summarized as follows. First, we establish high-probability bounds on the out-sample and in-sample MSEs of the ridge estimator (see Table 1) in two distinct, albeit possibly overlapping, regimes, called small or moderate TER and large TER. The two regimes are defined by whether the ratio $\frac{r_d(\Sigma)}{n}$ is small or large, where $r_d(\Sigma)$ is the tail effective rank (TER)

$$r_d(\Sigma) = \frac{\sum_{j>d} \lambda_j}{\lambda_{d+1}}.$$

Such a quantity is also central in the related analyses of ridge estimation and min-norm interpolation (Bartlett et al., 2020; Tsigler & Bartlett, 2023). A main difference between the two regimes is that in the large TER regime, the prediction errors of the ridge estimator can sometimes be controlled for a small ridge parameter τ , including $\tau = 0$ corresponding to the min-norm interpolator. From a technical perspective, the separation of the two regimes enables us to exploit rerlevant concentration

Table 2: Summary of conditions on $\frac{\lambda_{d+1}}{\lambda_d}$ to achieve $O(\frac{d}{n})$ prediction errors in the rotationally sparse setting

		$\text{MSE}_{\text{out}} = O(\frac{d}{n})$	$\text{MSE}_{\text{in}} = O(\frac{d}{n})$
$r_d(\Sigma) \lesssim n$	Sufficient Condition	$\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$	$\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$
	Necessary Condition	Assume $n \gg d$ and $r_d(\Sigma^2) \gg d$: $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$	Assume $n \gg d$ and $\frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \gg 1$: $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$
$r_d(\Sigma) \geq c_x n$	Sufficient Condition	$\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$	$\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$
	Necessary Condition	Assume $n \gg d$: $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$	Assume $n \gg d$: $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$

Note: c_x is a constant depending only on the sub-gaussian norm σ_x of covariate vectors.

Table 3: Summary of optimal MSE in the rotationally sparse setting

	Out-sample error	In-sample error
$r_d(\Sigma) \lesssim n$	Assume $\lambda_d \gtrsim \lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}}$: $\text{MSE}_{\text{out}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{d}{n}\}$	Assume $r_d(\Sigma) \asymp n$ and $\lambda_d \gg \lambda_{d+1}$: $\text{MSE}_{\text{in}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d}, \frac{d}{n}\}$
	$\text{MSE}_{\text{out}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2}, \frac{d}{n}\}$	Assume $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$: $\text{MSE}_{\text{in}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}, \frac{d}{n}\}$

Note: c_x is a constant depending only on the sub-gaussian norm σ_x of covariate vectors. $\text{MSE}_{\text{out}}^*$ denotes the MSE_{out} with optimal τ and MSE_{in}^* denotes the MSE_{in} with optimal τ .

inequalities and derive concrete error bounds without making any oracle assumption or independent components assumption on covariate vectors as used in Tsigler & Bartlett (2023).

Second, from our error bounds, we derive sufficient and necessary conditions on the ratio $\frac{\lambda_{d+1}}{\lambda_d}$ together with the choice of ridge parameter τ such that the out-sample and in-sample MSEs is of the order $O(\frac{d}{n})$ respectively (see Table 2). All of these conditions are determined in the simple form that the ratio $\frac{\lambda_{d+1}}{\lambda_d}$ is sufficiently small, i.e., the gap between the spiked and tail variances is sufficiently large. In other words, our results indicate that ridge estimation can achieve prediction errors in the order $O(\frac{d}{n})$ for a suitable choice of τ if and only if the gap between the spiked and tail variances is sufficiently large in the rotationally sparse linear regression. These results can be seen to serve as a counterpart to existing theory for Lasso estimation to achieve prediction errors in the order $O(s \frac{\log p}{n})$ under suitable conditions including compatibility conditions on Σ .

Third, from our error bounds depending on the ridge parameter τ , we also derive the optimal orders of out-sample and in-sample MSEs obtained respectively with the optimal choices of τ (see Table 3). The optimal orders of prediction errors may be greater than $O(\frac{d}{n})$. Remarkably, we find that if $\frac{\lambda_{d+1}}{\lambda_d}$ is sufficiently small, then the optimal out-sample MSE is, up to a constant factor, smaller than the optimal in-sample MSE in both the regime of small or moderate TER (under

some technical conditions) and the regime of large TER. We also identify specific conditions under which the optimal out-sample MSE is significantly smaller than the optimal in-sample MSE (see Remarks 1 and 4 for details, and Figure 2 for numerical results). This phenomenon seems to be surprising: out-sample MSEs may be usually considered to be no smaller than in-sample MSEs.

Related works. There is a large and growing literature on prediction properties of ridge estimators and min-norm interpolators. See Tsigler & Bartlett (2023), Section 9, for a recent review. We discuss directly related works to ours, in addition to the earlier discussion.

Hsu et al. (2014) allowed sub-gaussian covariate vectors and studied the out-sample MSE of the ridge estimator when the ridge parameter τ is large enough such that the effective dimension, $\sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \tau}$, is small compared with the sample size. For this reason, their error bounds are not applicable to a small ridge parameter or a min-norm interpolator.

Hastie et al. (2022) derived out-sample error approximation formulas for the ridge estimator and the min-norm interpolator using random matrix theory. They also showed that the deviation between the out-sample error and the approximation formula is upper bounded by the order of $n^{-\frac{1}{2}}$ for the ridge estimator (with a ridge tuning parameter bounded away from 0) and is upper bounded by the order of $n^{-\frac{1}{7}}$ for the min-norm interpolator. Compared to our results, the independent components assumption and boundedness of $\frac{p}{n}$ are assumed in Hastie et al. (2022). Moreover, the orders of their deviation bounds may be much larger than $\frac{d}{n}$, so that combining the approximation formulas and the deviation bounds may lead to less sharp out-sample error bounds than ours in the rotationally sparse setting. Despite these differences, it can be shown that the orders of the approximation formulas in Hastie et al. (2022) match the orders of our error bounds, which are obtained without the independent components assumption or boundedness of $\frac{p}{n}$ in the rotationally sparse setting. See Section 4.1 for details.

Bartlett et al. (2020) studied the min-norm interpolator and gave upper bounds of the out-sample error variance and bias and a lower bound of the out-sample error variance (but not bias). However, their results rely on the independent components assumption. Moreover, although the tail effective rank is involved, their out-sample error bounds are obtained in terms of the overall $\|\theta^*\|^2$, regardless of how the mean response is aligned differently with the spiked and tail parts of covariate vectors, which are essential to the rotationally sparse setting.

Tsigler & Bartlett (2023) provided upper and lower bounds of both the out-sample error variance and bias, while exploiting the decomposition of the spiked and tail parts of covariate vectors. However, although the variance and bias upper bounds in Tsigler & Bartlett (2023) are obtained with sub-Gaussian covariate vectors instead of the independent components assumption, an oracle

assumption is required on some random matrix from covariate vectors. Moreover, their variance lower bound is obtained under the independent components assumption, and their bias lower bound is provided in terms of the expectation with respect to a prior distribution on θ^* under an extra oracle assumption on covariate vectors. By comparison, our error upper bounds match those in Tsigler & Bartlett (2023) for the ridge tuning parameter in suitable ranges, and all our upper and lower bounds are obtained with sub-Gaussian covariate vectors without making any oracle assumption or independent components assumption. See Section 4.2 for details.

Bunea et al. (2022) studied the min-norm interpolator in a latent factor model as follows:

$$y_i = \beta^\top z_i + \xi_i, \quad x_i = Az_i + e_i, \quad i = 1, \dots, n, \quad (3)$$

where $\beta \in \mathbb{R}^d$, $A \in \mathbb{R}^{p \times d}$, $z_i \in \mathbb{R}^d$ is a latent feature vector, $\xi_i \in \mathbb{R}$ and $e_i \in \mathbb{R}^p$ are mean-zero noises, and (z_i, ξ_i, e_i) are mutually independent for each i . The matrix $\Sigma = \text{Var}(x_i)$ can be expressed as $\Sigma = A\Sigma_Z A^\top + \Sigma_E$, where $\Sigma_Z = \text{Var}(z_i)$ and $\Sigma_E = \text{Var}(e_i)$. The latent factor model can be seen to share a similar structure as our rotationally sparse linear regression model. From our comparison in Section 4.3, the upper bound of the out-sample MSE in Bunea et al. (2022) is obtained for the min-norm interpolator in the large TER regime, and is less sharp than our result which gives the order of out-sample MSE (i.e., matching upper and lower bounds up to a constant factor) except in the trivial situation where the out-sample MSE is bounded away from zero. In addition, the analysis of Bunea et al. (2022) assumes that the whitened noises, $\Sigma_E^{-1/2} e_i$, has independent components. Such an assumption of independent components is avoided in our analysis.

2 Assumptions and notation

We formulate the following assumptions to facilitate our theoretical analysis. Let d be the dimension of the spike part satisfying $0 < d < p$.

Assumption 1 (Low dimension of spiked part). *Suppose that $d \leq n$ and $\frac{d}{n}$ is small enough such that*

$$\eta_1 = C_0 \sigma_x^2 \sqrt{\frac{d}{n}} < \frac{1}{2},$$

where C_0 is an absolute constant from Lemma S14–S18.

As shown in Bartlett et al. (2020), the tail effective rank (TER) is important for analyzing benign linear regression. For the covariance matrix Σ , define

$$r_d(\Sigma) = \frac{\sum_{j>d} \lambda_j}{\lambda_{d+1}},$$

where $\lambda_1 \geq \dots \geq \lambda_p > 0$ are the eigenvalues of Σ . We refer to $r_d(\Sigma)$ as TER, because it pertains to the tail eigenvalues of Σ . We also use the following related quantity:

$$r_d(\Sigma^2) = \frac{\sum_{j>d} \lambda_j^2}{\lambda_{d+1}^2}.$$

It can be easily verified that $r_d(\Sigma^2) \leq r_d(\Sigma)$.

The following two assumptions describe two regimes of TER, in terms of the ratio $\frac{r_d(\Sigma)}{n}$. The magnitude of $\frac{r_d(\Sigma)}{n}$ affects the behavior of the out-sample error and in-sample error.

Assumption 2 (Small or moderate TER). $\frac{r_d(\Sigma)}{n} \leq C_1$, where $C_1 > 0$ is a constant.

Assumption 3 (Large TER). $\frac{r_d(\Sigma)}{n}$ is large enough such that

$$\eta_2 = C_0 \sigma_x^2 \sqrt{\frac{4n^2}{r_d(\Sigma)^2} + \frac{2n}{r_d(\Sigma)}} \leq \frac{1}{2}.$$

Alternatively, it is sufficient to assume that $\frac{r_d(\Sigma)}{n} \geq c_x$ for some c_x depending only on σ_x . For instance, c_x can be $\max\{4\sqrt{2}C_0\sigma_x^2, 16C_0^2\sigma_x^4\}$.

Separating the two regimes above is desirable for theoretical analysis, because it enables us to establish concrete results and avoid making any oracle assumption or independent components assumption on covariate vectors as used in Tsigler & Bartlett (2023). See Section 4.2 for further information. The two regimes above are not contained by each other. For instance, $\frac{r_d(\Sigma)}{n}$ satisfies Assumption 2 but not Assumption 3 if $\frac{r_d(\Sigma)}{n} \ll 1$, whereas $\frac{r_d(\Sigma)}{n}$ satisfies Assumption 3 but not Assumption 2 if $\frac{r_d(\Sigma)}{n} \gg 1$. Overlapping of the two regimes is possible, for instance, $\frac{r_d(\Sigma)}{n} \asymp 1$ and $\frac{r_d(\Sigma)}{n}$ satisfies both Assumptions 2 and 3.

The following assumption describes the rotationally sparse setting in terms of the relative magnitudes of $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2$.

Assumption 4 (Rotational Sparsity).

(i) [Applied with small or moderate TER]. For some $0 < \delta_1 < 1$,

$$\frac{\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2} \leq \frac{\delta_1}{4(1 + \sigma_x^2)} \lambda_{d+1}^2.$$

(ii) [Applied with large TER]. For some $0 < \delta_2 < 1$,

$$\frac{\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2} \leq \frac{\delta_2}{4(1 + \sigma_x^2)} \left(\frac{1}{\lambda_d} + \frac{4n}{\sum_{j>d} \lambda_j} \right)^{-2},$$

which can be equivalently stated as

$$\frac{\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2} \leq \frac{\delta_2}{4(1 + \sigma_x^2)} \left(\frac{\lambda_{d+1}}{\lambda_d} + \frac{4n}{r_d(\Sigma)} \right)^{-2} \lambda_{d+1}^2.$$

Assumption 4(i) is specified for the small or moderate regime, whereas Assumption 4(ii) is specified for the regime large TER regime. Assumption 4(ii) provides a much weaker condition than Assumption 4(i) on $\frac{\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2}$ if $\lambda_d \gg \lambda_{d+1}$ and $r_d(\Sigma) \gg n$.

Notation. Given two positive sequence $\{a_k\}$ and $\{b_k\}$, $a_k \lesssim b_k$ ($a_k \gtrsim b_k$) indicates that there exist constants $c > 0$ and $K \geq 1$ such that $a_k \leq cb_k$ ($a_k \geq cb_k$) for all $k \geq K$. We also denote $a_k = O(b_k)$ if $a_k \lesssim b_k$. Moreover, $a_k \asymp b_k$ indicates that both $a_k \lesssim b_k$ and $a_k \gtrsim b_k$; and $a_k \ll b_k$ (or $a_k \gg b_k$) indicates $\lim_{k \rightarrow \infty} \frac{a_k}{b_k} = 0$ (or $\lim_{k \rightarrow \infty} \frac{a_k}{b_k} = \infty$). Finally, $\text{Poly}_{deg}(x)$ denotes a polynomial of x with positive bounded coefficients and the highest order equal to deg .

3 Main results

The standard formula of the ridge estimator for $n > p$ is

$$\hat{\theta}(\tau) = (X^T X + n\tau)^{-1} X^T Y.$$

However, this formula does not hold for $\tau = 0$ when $X^T X$ is not invertible with $n < p$. In the high-dimensional setting, the ridge estimator $\hat{\theta}(\tau)$ can be expressed as

$$\hat{\theta}(\tau) = X^T (X X^T + n\tau)^{-1} Y. \quad (4)$$

See, for example, Appendix B in Tsigler & Bartlett (2023).

With the expression (4) and the assumption that X and ϵ are independent, the out-sample and in-sample errors can be decomposed into bias and variance as follows:

- Out-sample error

$$\begin{aligned} \text{MSE}_{\text{out}} = & \underbrace{\|(I_p - X^T (X X^T + n\tau I_n)^{-1} X) \theta^*\|_{\Sigma}^2}_{\text{B}_{\text{out}}} \\ & + \underbrace{\sigma^2 \text{Tr}((X X^T + n\tau I_n)^{-1} X \Sigma X^T (X X^T + n\tau I_n)^{-1})}_{\text{V}_{\text{out}}}, \end{aligned} \quad (5)$$

- In-sample error

$$\begin{aligned} \text{MSE}_{\text{in}} = & \underbrace{\|(I_p - X^T (X X^T + n\tau I_n)^{-1} X) \theta^*\|_{\Sigma}^2}_{\text{B}_{\text{in}}} \\ & + \underbrace{\sigma^2 \text{Tr}((X X^T + n\tau I_n)^{-1} X \hat{\Sigma} X^T (X X^T + n\tau I_n)^{-1})}_{\text{V}_{\text{in}}}. \end{aligned} \quad (6)$$

We present our main results about the out-sample and in-sample errors in the small or moderate TER regime in Section 3.1 and the large TER regime in Section 3.2.

3.1 Regime I: Small or moderate TER

Consider the regime of small or moderate TER such that $r_d(\Sigma) \leq C_1 n$, as stated in Assumption 2. The following is our main result about MSE_{out} , the out-sample MSE. Let A_0 be a constant satisfying $A_0 \geq 1$ and $(1 + A_0)^2 < 4\delta_1^{-1}$, with $0 < \delta_1 < 1$ from Assumption 4(i). For $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$, we determine the order of MSE_{out} (including upper and lower bounds). For $\tau \leq A_0^{-1}\lambda_{d+1}$ and $\tau \geq A_0\lambda_d$, we give lower bounds of MSE_{out} through, respectively, the variance and bias terms.

Theorem 1 (Out-sample error with small or moderate TER). *Under Assumption 1, 2 and 4(i), for any ν satisfying $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ and any A_0 satisfying $A_0 \geq 1$ and $(1 + A_0)^2 < 4\delta_1^{-1}$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0 \sigma_x^4}\} - 18\exp\{-\frac{n}{C_0}\}$:*

$$\begin{aligned}
(i) \quad & \text{MSE}_{\text{out}} \geq \underbrace{M_1 \sigma^2 \left(\frac{d}{n} + \frac{r_d(\Sigma^2)}{n} \right)}_{V_{\text{out}}} \quad \text{for } \tau \leq A_0^{-1} \lambda_{d+1}, \\
(ii) \quad & \underbrace{M_2 (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2)}_{\bar{B}_{\text{out}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \right)}_{\bar{V}_{\text{out}}} \geq \text{MSE}_{\text{out}} \geq \\
& \underbrace{M_1 (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2)}_{\underline{B}_{\text{out}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \right)}_{V_{\text{out}}} \quad \text{for } A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d, \\
(iii) \quad & \text{MSE}_{\text{out}} \geq \underbrace{M_1 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2}_{\underline{B}_{\text{out}}} \quad \text{for } \tau \geq A_0 \lambda_d,
\end{aligned}$$

where $M_1, M_2 > 0$ are constants, depending only on $(\sigma_x, \eta_1, C_1, \delta_1, \nu, A_0)$, and $\underline{B}_{\text{out}}$ and V_{out} represent lower bounds of out-sample bias and variance and $\bar{B}_{\text{out}}$ and $\bar{V}_{\text{out}}$ represent upper bounds of out-sample bias and variance, up to the constant M_1 or M_2 .

The following corollary provides simple conditions for achieving $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ in the regime of small or moderate TER.

Corollary 1 (Conditions for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with small or moderate TER). *In the setting of Theorem 1, assume further that $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$.*

(i) *A sufficient condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability approaching 1 as $n \rightarrow \infty$ is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$ and the ridge parameter τ is chosen in the range $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_{d+1}$ if $r_d(\Sigma^2) \leq d$ or $A_0^{-1}\lambda_{d+1} \max\{\frac{1}{c} \sqrt{\frac{r_d(\Sigma^2)}{d}}, 1\} \leq \tau \leq A_0\lambda_d \min\{c \sqrt{\frac{d}{n}}, 1\}$ if $r_d(\Sigma^2) > d$, where c is a constant satisfying $c \geq 1$ and $\frac{\lambda_{d+1}}{\lambda_d} \leq c \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$.*

(ii) *Suppose that $n \gg d$ and $r_d(\Sigma^2) \gg d$. Then a necessary condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with*

a probability bounded away from 0 is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$ and the ridge parameter τ is chosen in the range $\sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} \lesssim \tau \lesssim \sqrt{\frac{d}{n}} \lambda_d$.

The sufficient and necessary conditions become matched, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$, in the case where $n \gg d$ and $r_d(\Sigma^2) \gg d$ in addition to the assumptions stated.

Next, we give our main result about MSE_{in} , the in-sample MSE, in the regime of small or moderate TER stated in Assumption 2. As before, let A_0 be a constant satisfying $A_0 \geq 1$ and $(1 + A_0)^2 < 4\delta_1^{-1}$, with $0 < \delta_1 < 1$ from Assumption 4(i). For $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$, we derive upper and lower bound of MSE_{in} . For $\tau \leq A_0^{-1}\lambda_{d+1}$, we give a lower bound through variance. For $\tau \geq A_0\lambda_d$, we give a lower bound of MSE_{in} through the sum of bias and variance terms.

Theorem 2 (In-sample error with small or moderate TER). *Under Assumption 1, 2 and 4(i), for any ν satisfying $0 < \nu < \frac{1}{4} \min\{1, \sigma_x^2\}$ and any A_0 satisfying $A_0 \geq 1$ and $(1 + A_0)^2 < 4\delta_1^{-1}$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0 \sigma_x^4}\} - 8\exp\{-\frac{n}{C_0}\}$:*

$$\begin{aligned}
(i) \quad & \text{MSE}_{\text{in}} \geq \underbrace{M_1 \sigma^2 \left(\frac{d}{n} + \frac{r_d^2(\Sigma)}{n^2} \right)}_{V_{\text{in}}} \quad \text{for } \tau \leq A_0^{-1} \lambda_{d+1}, \\
(ii) \quad & \underbrace{M_2 (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2)}_{\bar{B}_{\text{in}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2 r_d(\Sigma)}{\tau^2 n} \right)}_{\bar{V}_{\text{in}}} \geq \text{MSE}_{\text{in}} \geq \\
& \underbrace{M_1 (\kappa_1(\tau) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2)}_{\underline{B}_{\text{in}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2 r_d^2(\Sigma)}{\tau^2 n^2} \right)}_{V_{\text{in}}} \quad \text{for } A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d, \\
(iii) \quad & \text{MSE}_{\text{in}} \geq \underbrace{M_1 (\kappa_1(\tau) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2)}_{\underline{B}_{\text{in}}} + \underbrace{\sigma^2 \frac{\lambda_{d+1}^2 r_d^2(\Sigma)}{\tau^2 n^2}}_{V_{\text{in}}} \quad \text{for } \tau \geq A_0 \lambda_d,
\end{aligned}$$

where $\kappa_1(\tau) = \max\{1 - (\frac{2C_0\sigma_x^2(2+C_1)\lambda_{d+1}}{\tau} (1 + 16(2C_0\sigma_x^2 + 1)(1 + C_1)\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}) + 64\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}), 0\}$, and $M_1, M_2 > 0$ are constants depending only on $(\sigma_x, \eta_1, C_1, \delta_1, \nu, A_0)$. The terms $\underline{B}_{\text{in}}$ and V_{in} represent lower bounds of in-sample bias and variance and $\bar{B}_{\text{in}}$ and $\bar{V}_{\text{in}}$ represent upper bounds of in-sample bias and variance, up to the constant M_1 or M_2 .

By the definition of $\kappa_1(\tau)$, the bias term $\underline{B}_{\text{in}}$ is activated in the lower bound of MSE_{in} only when $\frac{\lambda_{d+1}}{\tau}$ and $\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}$ are small enough. This can be explained from our proof strategy as follows (see Section 6.2 for details). The in-sample bias $\|\hat{\theta}(\tau) - \theta^*\|_{\hat{\Sigma}}^2$ can be expressed as the sum of $\|\hat{\theta}(\tau)_{1:d} - \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}^2$, $2(\hat{\theta}(\tau)_{1:d} - \theta_{1:d}^*)^T \hat{\Sigma}_{(1:d), (d+1):p} (\hat{\theta}(\tau)_{(d+1):p} - \theta_{(d+1):p}^*)$ and $\|\hat{\theta}(\tau)_{(d+1):p} - \theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2$ and only the interaction term $2(\hat{\theta}(\tau)_{1:d} - \theta_{1:d}^*)^T \hat{\Sigma}_{(1:d), (d+1):p} (\hat{\theta}(\tau)_{(d+1):p} - \theta_{(d+1):p}^*)$ can be negative. When

$\frac{\lambda_{d+1}}{\tau}$ and $\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}$ are small enough, the bias from the spiked part, $\|\hat{\theta}(\tau)_{1:d} - \theta_{1:d}^*\|_{\Sigma_{1:d}}^2$, can be shown to dominate the interaction term. Then a lower bound on the bias from the spiked part, which can be deduced in a convenient manner, also provides a lower bound on the overall bias.

From Theorem 2, we deduce the following simple conditions for achieving $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ in the regime of small or moderate TER.

Corollary 2 (Conditions for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with small or moderate TER). *In the setting of Theorem 2, assume further that $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$.*

(i) *A sufficient condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability approaching 1 as $n \rightarrow \infty$ is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$ and the ridge parameter τ is chosen in the range $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_{d+1}$ if $r_d(\Sigma) \leq d$ or $A_0^{-1} \lambda_{d+1} \max\{\frac{1}{c} \sqrt{\frac{r_d(\Sigma)}{d}}, 1\} \leq \tau \leq A_0 \lambda_d \min\{c \sqrt{\frac{d}{n}}, 1\}$ if $r_d(\Sigma) > d$, where c is a constant satisfying $c \geq 1$ and $\frac{\lambda_{d+1}}{\lambda_d} \leq c \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$.*

(ii) *Suppose that $n \gg d$, $\frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \gg 1$ and $64 \frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}} < 1$. Then a necessary condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$ and the ridge parameter τ is chosen in the range $\lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau \lesssim \lambda_d \sqrt{\frac{d}{n}}$.*

The sufficient and necessary conditions become matched, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$, in the case where $n \gg d$ and $r_d(\Sigma) \asymp n$ in addition to the assumptions stated.

From Theorems 1 and 2, we derive the order of MSE_{out} with an optimal choice τ , denoted as $\text{MSE}_{\text{out}}^*$, and the order of MSE_{in} with an optimal choice τ , denoted as MSE_{in}^* . The following corollary gives the orders of $\text{MSE}_{\text{out}}^*$ and MSE_{in}^* in the small or moderate TER regime.

Corollary 3 (Optimal error orders with small or moderate TER). *Suppose that Assumption 1, 2 and 4(i) are satisfied and further $\sigma^2 \asymp 1$, $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, $r_d(\Sigma) \asymp n$, $\lambda_d \gtrsim \lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}}$, $\lambda_d \gg \lambda_{d+1}$, and $64 \frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}} < 1$. Then*

(i) $\text{MSE}_{\text{out}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{d}{n}\}$ with a probability approaching to 1 and the optimal τ is chosen as $\tau = \sqrt{\lambda_d \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{n}} \min\{\sqrt{c A_0^{-2}}, \frac{A_0 \lambda_d}{\sqrt{\lambda_d \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{n}}}}\}}$ where c is a constant satisfying

$$\lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}} \leq c \lambda_d.$$

(ii) $\text{MSE}_{\text{in}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d}, \frac{d}{n}\}$ with a probability approaching to 1 and the optimal τ is chosen as $\tau \asymp \sqrt{\lambda_{d+1} \lambda_d}$.

Therefore $\text{MSE}_{\text{out}}^* \lesssim \text{MSE}_{\text{in}}^*$ with a probability approaching to 1, by noting $r_d(\Sigma^2) \leq r_d(\Sigma) \asymp n$.

The additional conditions $r_d(\Sigma) \asymp n$, $\lambda_d \gtrsim \lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}}$ and $\lambda_d \gg \lambda_{d+1}$ can be explained as follows. First, $r_d(\Sigma) \asymp n$ and $\lambda_d \gg \lambda_{d+1}$ are important for determining the order of MSE_{in}^* . In fact, the order of V_{in} for $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$ can be determined from Theorem 2(ii) only under

$r_d(\Sigma) \asymp n$, due to the difference between $\underline{V}_{\text{in}}$ and $\bar{V}_{\text{in}}$. Note that the sum of the in-sample bias and the tail part of in-sample variance, $\|\theta_{1:d}^*\|_{\Sigma^{-1}}^2 \tau^2 + \sigma^2 \frac{\lambda_{d+1}^2}{\tau^2}$, reaches the minimum order of $\frac{\lambda_{d+1}}{\lambda_d}$ by the choice $\tau \asymp \sqrt{\lambda_d \lambda_{d+1}}$. The condition $\lambda_d \gg \lambda_{d+1}$ ensures that this choice of τ is large enough so that $\kappa_1(\tau)$ is activated. Second, $\lambda_d \gtrsim \lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}}$ is important for determining the order of $\text{MSE}_{\text{out}}^*$ because under $\lambda_d \gtrsim \lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}}$, the order of V_{out} for $\tau \leq A_0^{-1} \lambda_{d+1}$ can be shown to be larger than $\max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{d}{n}\}$ and then the range $\tau \leq A_0^{-1} \lambda_{d+1}$ can be ruled out when optimizing MSE_{out} . See the proof of Corollary 3 in Supplement Section III.2 for details.

Remark 1. In the setting of Corollary 3, we observe that the gap between MSE_{in}^* and $\text{MSE}_{\text{out}}^*$ can be significantly large, for example, if further $\frac{\lambda_{d+1}}{\lambda_d} \gtrsim \frac{d}{n} \sqrt{\frac{n}{r_d(\Sigma^2)}}$ and $\frac{n}{r_d(\Sigma^2)} \gg 1$. In this case, both MSE_{out} and MSE_{in} do not achieve $O(\frac{d}{n})$, and $\text{MSE}_{\text{out}}^* \asymp \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}$ and $\text{MSE}_{\text{in}}^* \asymp \frac{\lambda_{d+1}}{\lambda_d}$ by Corollary 3. With $\frac{n}{r_d(\Sigma^2)} \gg 1$, there can be a substantial gap between MSE_{in}^* and $\text{MSE}_{\text{out}}^*$.

In the setting of Corollary 3, we point out that the advantage of $\text{MSE}_{\text{out}}^*$ over MSE_{in}^* can be attributed to $V_{\text{out}} \lesssim V_{\text{in}}$ for $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$ under $r_d(\Sigma) \asymp n$. See the proof of Corollary 3 in Supplement Section III.2 for details. In fact, in the setting of Corollary 3, the optimal choices of τ for both MSE_{out} and MSE_{in} are chosen from the range $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$, because both MSE_{out} and MSE_{in} are lower bounded by large variance for $\tau \leq A_0^{-1} \lambda_{d+1}$ and by large bias for $\tau \geq A_0 \lambda_d$. Moreover, as seen from the proof, the optimal order of MSE_{in} , $\max\{\frac{\lambda_{d+1}}{\lambda_d}, \frac{d}{n}\}$, in the setting of Corollary 3 can be achieved only when $\underline{B}_{\text{in}}$ is activated. With $\underline{B}_{\text{in}}$ activated, the orders of B_{out} and B_{in} are the same for $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$. By comparison, V_{out} can be shown to be smaller than V_{in} up to a constant factor for $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$ under $r_d(\Sigma) \asymp n$:

$$\underbrace{\sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \right)}_{\text{order of } V_{\text{out}} \text{ in Theorem 1(ii)}} \lesssim \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \right)}_{\text{order of } V_{\text{in}} \text{ in Theorem 2(ii)}}.$$

Hence the advantage of $\text{MSE}_{\text{out}}^*$ stems from the smaller order of V_{out} in the setting of Corollary 3.

3.2 Regime II: Large TER

The second regime we investigate is when $\frac{r_d(\Sigma)}{n}$ is large enough such that Assumption 3 is satisfied. In this regime, as shown below, the smallest eigenvalue of $\tau I_n + n^{-1} X_{(d+1):p} X_{(d+1):p}^T$ is lower bounded away from 0 for any $\tau \geq 0$, so that MSE_{out} can sometimes be controlled even for $\tau = 0$. Let A_0 be any positive constant. For the ridge parameter $\tau \geq 0$ satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq A_0 \lambda_d$, we determine the order of MSE_{out} (including upper and lower bounds). For $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \geq A_0 \lambda_d$, we give a lower bound of MSE_{out} through the bias term.

Theorem 3 (Out-sample error with large TER). *Under Assumption 1, 3 and 4(ii), for any ν satisfying $0 < \nu < \frac{1}{4} \min\{1, \sigma_x^2\}$ and $\frac{r_d(\Sigma)\nu^2}{C_0^2\sigma_x^4} > 1$ and any $A_0 > 0$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 2\exp\{-\frac{\nu^2 n}{C_0^2\sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0\sigma_x^4}\} - 16\exp\{-\frac{n}{C_0}\}$:*

$$\begin{aligned}
(i) \quad & \underbrace{M_2(\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n})^2)}_{\bar{B}_{\text{out}}} + \underbrace{\sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n})^2}\frac{r_d(\Sigma^2)}{n})}_{\bar{V}_{\text{out}}} \geq \text{MSE}_{\text{out}} \geq \\
& \underbrace{M_1(\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n})^2)}_{\underline{B}_{\text{out}}} + \underbrace{\sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n})^2}\frac{r_d(\Sigma^2)}{n})}_{\underline{V}_{\text{out}}} \quad \text{for } \tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \leq A_0\lambda_d, \\
(ii) \quad & \text{MSE}_{\text{out}} \geq \underbrace{M_1\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2\lambda_d^2}_{\underline{B}_{\text{out}}}, \quad \text{for } \tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \geq A_0\lambda_d,
\end{aligned}$$

where $M_1, M_2 > 0$ are constants depending only on $(\sigma_x, \eta_2, \delta_2, \nu, A_0)$.

Remark 2. The error bounds in Theorems 1 and 3 are derived from the same set of algebraic bounds but then by applying relevant high-probability inequalities to control random quantities for different ranges of τ under the two regimes of TER (see Section 6.1). In fact, after ignoring the range choice of τ , the error bounds (i)–(ii) in Theorem 3 appear similar to (ii)–(iii) in Theorem 1 except with $\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n}$ in place of τ . In the overlapping case of small or moderate TER regime and large TER regime, i.e., $\frac{r_d(\Sigma)}{n} \asymp 1$ and $\frac{r_d(\Sigma)}{n}$ satisfies Assumption 2 and 3, the result from Theorem 1 (ii), $\text{MSE}_{\text{out}} \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2\tau^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2}\frac{r_d(\Sigma^2)}{n})$, and the result from Theorem 3 (i), $\text{MSE}_{\text{out}} \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n})^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n})^2}\frac{r_d(\Sigma^2)}{n})$, are equivalent to each other for τ in the range $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$, where $\tau \asymp \tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n}$ with $r_d(\Sigma) \asymp n$. However, the error bound in Theorem 3 (i) remains applicable, but that in Theorem 1 (ii) does not apply, to small τ satisfying $0 \leq \tau \leq A_0^{-1}\lambda_{d+1}$ including $\tau = 0$.

The following corollary provides simple conditions for achieving $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ in the regime of large TER.

Corollary 4 (Conditions for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with large TER). *In the setting of Theorem 3, assume further that $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2\lambda_d^2 \asymp 1$.*

(i) *A sufficient condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability approaching 1 as $n \rightarrow \infty$ is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$ and the ridge parameter τ is chosen satisfying $\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}}\lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$ or $\sqrt{\frac{r_d(\Sigma^2)}{d}}\lambda_{d+1} \lesssim \tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}}\lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$.*

(ii) Suppose that $n \gg d$. Then a necessary condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$ and τ is chosen satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$ or $\sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$. The sufficient and necessary conditions become matched, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$, in the case where $n \gg d$ in addition to the assumptions stated.

Next, we give our main result about MSE_{in} in the regime of large TER stated in Assumption 3. Let A_0 be any positive constant. We derive upper and lower bounds of MSE_{in} in the case of $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq A_0 \lambda_d$, and a lower bound of MSE_{in} in the case of $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \geq A_0 \lambda_d$, through the sum of bias and variance terms.

Theorem 4 (In-sample error with large TER). *Under Assumption 1, 3 and 4(ii), for any ν satisfying $0 < \nu < \frac{1}{4}$ and $\frac{r_d(\Sigma)\nu^2}{C_0^2\sigma_x^4} > 1$ and any $A_0 > 0$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2n\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 2\exp\{-\frac{\nu^2 n}{C_0^2\sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0\sigma_x^4}\} - 12\exp\{-\frac{n}{C_0}\}$:*

$$\begin{aligned}
(i) \quad & \underbrace{M_2(\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2)}_{\underline{B}_{\text{in}}} + \underbrace{\sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2})}_{\underline{V}_{\text{in}}} \geq \text{MSE}_{\text{in}} \geq \\
& \underbrace{M_1(\kappa_2(\tau)\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2)}_{\underline{B}_{\text{in}}} + \underbrace{\sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2})}_{\underline{V}_{\text{in}}} \quad \text{for } \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq A_0 \lambda_d, \\
(ii) \quad & \text{MSE}_{\text{in}} \geq \underbrace{M_1(\kappa_2(\tau)\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2)}_{\underline{B}_{\text{in}}} + \underbrace{\sigma^2 \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}}_{\underline{V}_{\text{in}}} \quad \text{for } \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \geq A_0 \lambda_d,
\end{aligned}$$

where $\kappa_2(\tau) = \max\{1 - (16 \frac{\lambda_{d+1} \frac{r_d(\Sigma)}{n}}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} (1 + 112 \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}) + 64 \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}), 0\}$ and $M_1, M_2 > 0$ are constants depending only on $(\sigma_x, \eta_2, \delta_2, \nu, A_0)$.

Similarly to $\kappa_1(\tau)$ in Theorem 3, the definition of $\kappa_2(\tau)$ indicates that the bias term $\underline{B}_{\text{in}}$ is activated in the lower bound of MSE_{in} only when $\frac{\lambda_{d+1} \frac{r_d(\Sigma)}{n}}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})}$ and $\frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}$ are small enough. In this case, the bias from the spiked part, $\|\hat{\theta}(\tau)_{1:d} - \theta_{1:d}^*\|_{\Sigma_{1:d}}^2$, can be shown to dominate the interaction term between the spiked part and the tail part.

Remark 3. Similarly to out-sample error bounds discussed in Remark 2, the error bounds in Theorems 2 and 4 are also derived from the same set of algebraic bounds but then by applying relevant high-probability inequalities to control random quantities for different ranges of τ under the two regimes of TER (see Section 6.2). After ignoring the range choice of τ , the error bounds

(i)–(ii) in Theorem 4 appear similar to (ii)–(iii) in Theorem 2 except with $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ in place of τ and with the additional difference that $\frac{r_d(\Sigma)}{n^2}$ are involved both $\bar{V}_{\text{in}}$ and $\underline{V}_{\text{in}}$ in Theorem 4 (i), but not in Theorem 2 (ii). In the overlapping case of small or moderate TER regime and large TER regime, i.e., $\frac{r_d(\Sigma)}{n} \asymp 1$ and $\frac{r_d(\Sigma)}{n}$ satisfies Assumption 2 and 3, the result from Theorem 2 (ii) with $\kappa_1(\tau)$ is activated reduces to $\text{MSE}_{\text{out}} \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2})$, and the result from Theorem 4 (i) with $\kappa_2(\tau)$ is activated reduces to $\text{MSE}_{\text{out}} \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2 + \sigma^2(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2})$, and the two results are equivalent to each other, for τ in the range $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$, where $\tau \asymp \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ with $r_d(\Sigma) \asymp n$. However, the error bound in Theorem 4 (i) remains applicable, but that in Theorem 2 (ii) does not apply, to small τ satisfying $0 \leq \tau \leq A_0^{-1} \lambda_{d+1}$ including $\tau = 0$.

From Theorem 4, we deduce the following simple conditions for achieving $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ in the regime of large TER.

Corollary 5 (Conditions for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with large TER). *In the setting of Theorem 4, assume further that $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$.*

(i) *A sufficient condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability approaching to 1 as $n \rightarrow \infty$ is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$ and the ridge parameter τ is chosen such that $\lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d \sqrt{\frac{d}{n}}$.*

(ii) *Suppose that $n \gg d$ and $64 \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}} < 1$. Then a necessary condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$ and the ridge parameter τ is chosen in the range $\lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d \sqrt{\frac{d}{n}}$.*

The sufficient and necessary conditions become matched, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$, in the case where $n \gg d$ in addition to the assumptions stated.

From Theorem 3 and 4, we derive the orders of $\text{MSE}_{\text{out}}^*$ and MSE_{in}^* , i.e., MSE_{out} and MSE_{in} with optimal choices of τ respectively, in large TER regime.

Corollary 6 (Optimal error orders with large TER). *Suppose that Assumption 1, 3 and 4(ii) are satisfied and further $\sigma^2 \asymp 1$, $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ and $64 \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}} < 1$. Then*

(i) *The order of $\text{MSE}_{\text{out}}^*$ is $\max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2}, \frac{d}{n}\}$ with a probability approaching to 1 and the optimal τ is chosen as $\tau = 0$ if $\frac{n \sqrt{r_d(\Sigma^2)}}{\sqrt{d} r_d(\Sigma)} \leq \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ or satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \asymp \sqrt{\lambda_d \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{n}}}$ if $\frac{n \sqrt{r_d(\Sigma^2)}}{\sqrt{d} r_d(\Sigma)} > \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$.*

(ii) *The order of MSE_{in}^* is $\max\{\frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}, \frac{d}{n}\}$ with a probability approaching to 1 and the optimal τ is chosen satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \asymp \sqrt{\lambda_d \lambda_{d+1} \frac{r_d(\Sigma)}{n}}$.*

Therefore $\text{MSE}_{\text{out}}^ \lesssim \text{MSE}_{\text{in}}^*$ with a probability approaching to 1 because $\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}} \lesssim \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ by noting $r_d(\Sigma^2) \leq r_d(\Sigma)$ and $r_d(\Sigma) \gtrsim n$ (Assumption 3) and because $\frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2} \lesssim \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ by noting $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$.*

The additional condition $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ is important for determining the order of MSE_{in}^* . In fact, the sum of the in-sample bias and the tail part of in-sample variance, $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2 + \sigma^2 \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}$, reach the minimum order of $\frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ by the choice τ satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \asymp \sqrt{\lambda_d \lambda_{d+1} \frac{r_d(\Sigma)}{n}}$. The condition $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ ensures that this choice of τ is large enough so that $\kappa_2(\tau)$ is activated. See the proof of Corollary 6 in Supplement Section III.2 for details.

Remark 4. In the setting of Corollary 6, we observe that the gap between MSE_{in}^* and $\text{MSE}_{\text{out}}^*$ can be significantly large, for example, in two cases, $\frac{\lambda_{d+1}}{\lambda_d} \gtrsim \frac{d}{\sqrt{nr_d(\Sigma^2)}} \min\{1, \frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}}\}$ and $\frac{\lambda_{d+1}}{\lambda_d} \gtrsim \frac{n\sqrt{nr_d(\Sigma^2)}}{r_d(\Sigma)^2}$ hold or $\frac{\lambda_{d+1}}{\lambda_d} \gtrsim \frac{d}{\sqrt{nr_d(\Sigma^2)}} \min\{1, \frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}}\}$, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{n\sqrt{nr_d(\Sigma^2)}}{r_d(\Sigma)^2}$ and $\frac{r_d(\Sigma)}{\sqrt{nr_d(\Sigma^2)}} \gg 1$ hold. In the first case, $\text{MSE}_{\text{in}}^* \asymp \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ and $\text{MSE}_{\text{out}}^* \asymp \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2}$ by Corollary 6, and hence $\frac{\text{MSE}_{\text{in}}^*}{\text{MSE}_{\text{out}}^*} = \frac{\lambda_d}{\lambda_{d+1}} \frac{n}{r_d(\Sigma)} \gg 1$ with $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$. In the second case, $\text{MSE}_{\text{in}}^* \asymp \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ and $\text{MSE}_{\text{out}}^* \asymp \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}$ by Corollary 6, and hence $\frac{\text{MSE}_{\text{in}}^*}{\text{MSE}_{\text{out}}^*} = \frac{r_d(\Sigma)}{\sqrt{nr_d(\Sigma^2)}} \gg 1$.

In the setting of Corollary 6, we point out that the advantage of $\text{MSE}_{\text{out}}^*$ over MSE_{in}^* can be attributed to $V_{\text{out}} \lesssim V_{\text{in}}$ for $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq A_0 \lambda_d$. See the proof of Corollary 6 in Supplement Section III.2 for details. In fact, in the setting of Corollary 6, the optimal τ for both MSE_{out} and MSE_{in} are chosen from the range $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq A_0 \lambda_d$, because both MSE_{out} and MSE_{in} are lower bounded by large bias for $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \geq A_0 \lambda_d$. Moreover, as seen from the proof, the optimal order of MSE_{in} , $\max\{\frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}, \frac{d}{n}\}$, in the setting of Corollary 6 can be achieved only when $\underline{B}_{\text{in}}$ is activated. With $\underline{B}_{\text{in}}$ activated, the orders of B_{out} and B_{in} are the same for $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq A_0 \lambda_d$. By comparison, V_{out} can be shown to be smaller than V_{in} up to a constant factor for $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq A_0 \lambda_d$ in the large TER regime:

$$\underbrace{\sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} \right)}_{\text{order of } V_{\text{out}} \text{ in Theorem 3(i)}} \lesssim \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2} \right)}_{\text{order of } V_{\text{in}} \text{ in Theorem 4(i)}}.$$

Hence the advantage of $\text{MSE}_{\text{out}}^*$ stems from the smaller order of V_{out} in the setting of Corollary 6.

4 Connection and comparison with existing results

4.1 Comparison with error approximation formulas

We compare our results to error approximation formulas, obtained under an independence assumption on whiten covariates in ridge linear regression (see Assumption 5 below). We first review

the results of Hastie et al. (2022) about approximation formulas for out-sample error. To facilitate the comparison, we also derive and justify the approximation formulas for in-sample error. Then we show that in the rotationally sparse setting, the orders of out-sample and in-sample error approximation formulas match those from our results for the ridge tuning parameter in suitable ranges.

For ridge linear regression, Hastie et al. (2022) gave the following approximation formulas for the out-sample bias and variance:

$$\mathcal{B}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) = \tau^2 \|\theta^*\|^2 (1 + \gamma m_{n,1}(-\tau)) \int \frac{s}{[\tau + (1 - \gamma + \gamma \lambda m_n(-\tau))s]^2} d\hat{G}_n(s), \quad (7)$$

$$\mathcal{V}_{\text{out}}(\tau, \hat{H}_n, \gamma) = \sigma^2 \gamma \int \frac{s^2 (1 - \gamma + \gamma \lambda^2 m'_n(-\tau))}{[\tau + s(1 - \gamma + \gamma \tau m_n(-\tau))]^2} d\hat{H}_n(s), \quad (8)$$

where $\gamma = \frac{p}{n}$, $\hat{H}_n(s) = p^{-1} \sum_{j=1}^p 1_{\{s \geq \lambda_j\}}$, $\hat{G}_n(s) = \sum_{i=j}^p (\langle \theta^*, v_j \rangle^2 / \|\theta^*\|^2) 1_{\{s \geq \lambda_j\}}$, $v_1, \dots, v_p$ are the eigenvectors of Σ , $m_n(z)$ is determined by solving the following equation

$$m_n(z) = \int \frac{1}{s[1 - \gamma - \gamma z m_n(z)] - z} d\hat{H}_n(s), \quad (9)$$

and $m_{n,1}(z)$ is calculated by

$$m_{n,1}(z) = \frac{\int \frac{s^2 [1 - \gamma - \gamma z m_n(z)]}{[s[1 - \gamma - \gamma z m_n(z)] - z]^2} d\hat{H}_n(s)}{1 - \gamma \int \frac{zs}{[s[1 - \gamma - \gamma z m_n(z)] - z]^2} d\hat{H}_n(s)}. \quad (10)$$

Consider the following assumption on the whitened covariates, defined as $z_i = \Sigma^{-1/2} x_i$, the variance matrix Σ , and the ratio $\frac{p}{n}$.

Assumption 5.

- (i) Each vector $z_i = (z_{i1}, \dots, z_{ip})^T$ has independent entries with $E(z_{ij}) = 0$, $E(z_{ij}^2) = 1$ and $E(|z_{ij}|^k) \leq C_k < \infty$ for all $k \geq 2$.
- (ii) $\lambda_1 \leq M$ and $\int s^{-1} \hat{H}_n(s) ds < M$.
- (iii) $|1 - \frac{p}{n}| \geq \frac{1}{M}$, $\frac{1}{M} \leq \frac{p}{n} \leq M$.

Hastie et al. (2022) showed that under Assumption 5 and assuming $\max\{\tau, \lambda_p\} > \frac{1}{M}$ and $n^{-2/3+1/M} < \tau < M$, for any constant $D > 0$ and $\delta > 0$, with probability at least $1 - C(M, D, \delta)n^{-D}$,

$$|\mathcal{B}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) - \mathcal{B}_{\text{out}}| < \frac{C(M) \|\theta^*\|^2}{\tau n^{(1-\delta)/2}}, \quad |\mathcal{V}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) - \mathcal{V}_{\text{out}}| < \frac{C(M)}{\tau^2 n^{(1-\delta)/2}},$$

where $C(M, D, \delta)$ is a constant depending only on (M, D, δ) , and $C(M)$ is a constant depending only on M .

To facilitate the comparison between our results and error approximation formulas, we also give approximation formulas for in-sample bias and variance as follows:

$$\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) = \tau^2 \|\theta^*\|^2 (\gamma \tau^2 m'_n(-\tau) + 1 - \gamma) \int \frac{s}{[\tau + (1 - \gamma + \gamma \lambda m_n(-\tau))s]^2} d\hat{G}_n(s), \quad (11)$$

$$\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) = \sigma^2 \gamma (1 - 2\tau m_n(-\tau) + \tau^2 m'_n(-\tau)). \quad (12)$$

We establish the convergence of $\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma)$ and $\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma)$ in the following theorem.

Theorem 5 (Convergence of in-sample error approximation formulas). *Under Assumption 5, further assume that $\tau > \frac{1}{M}$ and $n^{-2/3+1/M} < \tau < \frac{M}{2}$. Then for any $D > 0$, $\delta > 0$, with probability at least $1 - C(M, D, \delta)n^{-D}$,*

$$\begin{aligned} |\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) - \mathcal{B}_{\text{in}}| &\leq C(M) \max\left\{\frac{1}{\tau^{2/3}n^{(1-\delta)/3}}, \frac{8M}{\tau n^{(1-\delta)/2}}\right\}, \\ |\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) - \mathcal{V}_{\text{in}}| &\leq \sigma^2 C(M) \left(\max\left\{\frac{1}{\tau^{2/3}n^{(1-\delta)/3}}, \frac{8M}{\tau n^{(1-\delta)/2}}\right\} + \frac{1}{n^{(1-\delta)/2}}\right), \end{aligned}$$

where $C(M, D, \delta)$ is a constant depending only on (M, D, δ) , and $C(M)$ is a constant depending only on M .

Given any $\tau > 0$, $\gamma = \frac{p}{n} > 0$ and $\tilde{\lambda} = (\lambda_1, \dots, \lambda_p)$ with $\lambda_j > 0$ for $1 \leq j \leq p$, we define $\alpha > 1$ as a solution to the equation

$$\frac{1}{\alpha} = 1 - \gamma \frac{1}{p} \sum_{j=1}^p \frac{1}{1 + \frac{\alpha\tau}{\lambda_j}}. \quad (13)$$

The approximation formulas above can be equivalently expressed as follows. These formulas can also be calculated using a distributional approximation method in Han & Shen (2023) under the independent components assumption, for which the discussion is deferred to Supplement Section IV.3.

Corollary 7 (Equivalent expressions of error approximation formulas). *With α defined in (13), we have*

$$\mathcal{B}_{\text{out}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) = \left(1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}\right)^{-1} \frac{1}{n} \sum_{j=1}^p \frac{\alpha^2 \tau^2 \lambda_j \theta_j^{*2}}{(\lambda_j + \alpha\tau)^2}, \quad (14)$$

$$\mathcal{V}_{\text{out}}(\tau; \hat{H}_n, \gamma) = \left(1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}\right)^{-1} \left(\frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}\right) \sigma^2, \quad (15)$$

$$\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) = \frac{1}{\alpha^2} \left(1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}\right)^{-1} \frac{1}{n} \sum_{j=1}^p \frac{\alpha^2 \tau^2 \lambda_j \theta_j^{*2}}{(\lambda_j + \alpha\tau)^2}, \quad (16)$$

$$\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) = \left(1 - \frac{2}{\alpha} + \frac{\left(1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}\right)^{-1}}{\alpha^2}\right) \sigma^2. \quad (17)$$

Next, we study the orders of error approximation formulas (14)-(17) in the high-dimensional rotationally sparse setting, and compare them with our results, which are obtained without requiring independence of the whiten covariates. The first result is about the small or moderate TER regime.

Corollary 8 (Matching error approximation formulas with small or moderate TER).

(i) Suppose that $\frac{d}{n} < 1$, $r_d(\Sigma) \lesssim n$, and $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_{d+1}^2$. For $\lambda_{d+1} \lesssim \tau \lesssim \lambda_d$, we have

$$\mathcal{B}_{\text{out}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{out}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \right). \quad (18)$$

(ii) Suppose further that $r_d(\Sigma) \asymp n$. For $\lambda_{d+1} \lesssim \tau \lesssim \lambda_d$, we have

$$\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \right). \quad (19)$$

For comparison, we notice that the conditions, $\frac{d}{n} < 1$, $r_d(\Sigma) \lesssim n$ and $\|\theta_{(d+1):p}^*\|^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_{d+1}^2$ correspond to, respectively, Assumption 1, 2 and 4(i) used in Theorems 1 and 2. The order of the approximation formula (18) matches Theorem 1(ii). The order of the approximation formula (19) matches Theorem 2(ii) when $r_d(\Sigma) \asymp n$ and $\kappa_1(\tau)$ is activated.

Then we study the large TER regime and the results are summarized as follows.

Corollary 9 (Matching error approximation formulas with large TER).

(i) Suppose that $\frac{d}{n} < \frac{1}{5}$, $r_d(\Sigma) > cn$ for some $c > 10$, $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{\sum_{j>d} \lambda_j} \right)^{-2}$. For $\lambda_d \gtrsim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ and $\tau > 0$, we have

$$\mathcal{B}_{\text{out}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{out}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \right)^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\left(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \right)^2} \frac{r_d(\Sigma^2)}{n} \right). \quad (20)$$

(ii) Suppose further that $\tau > \lambda_{d+1} \frac{r_d(\Sigma)}{n}$. For $\lambda_d \gtrsim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ and $\tau > 0$, we have

$$\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \right)^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\left(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \right)^2} \frac{r_d^2(\Sigma)}{n^2} \right). \quad (21)$$

The conditions, $\frac{d}{n} < \frac{1}{5}$, $r_d(\Sigma) > cn$ for some $c > 10$, $\|\theta_{(d+1):p}^*\|^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{\sum_{j>d} \lambda_j} \right)^{-2}$ correspond to, respectively, Assumption 1, 3 and 4(ii) used in Theorems 3 and 4. The order of the approximation formula (20) matches Theorem 3(i). The order of the approximation formula (21) matches Theorem 4(i) when $\kappa_2(\tau)$ is activated.

4.2 Comparison with Tsigler & Bartlett (2023)

In this section, we compare our results to Tsigler & Bartlett (2023), where the non-asymptotic out-sample error bounds are studied for high-dimensional ridge regression. We make a comparison in both upper bounds and lower bounds. In the following, denote $A_d = X_{(d+1):p} X_{(d+1):p}^T + n\tau I_n$. The

conditional number of A_d is defined $\mu_1(A_d)/\mu_n(A_d)$, where $\mu_1(A_d)$ and $\mu_n(A_d)$ are the maximum eigenvalue and minimum eigenvalue of A_d respectively.

Upper bound of MSE_{out} . Our upper bounds of MSE_{out} match the results in Tsigler & Bartlett (2023) for most ridge tuning parameters, but our result avoids making any oracle condition on covariate vectors as used in Tsigler & Bartlett (2023). In fact, for any d small enough compared to n , given that the conditional number of A_d is controlled by L , it is shown in Tsigler & Bartlett (2023) that with a high probability:

$$\text{B}_{\text{out}}/c \leq \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2 + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2, \quad (22)$$

$$\text{V}_{\text{out}}/c \leq \frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n}, \quad (23)$$

where c is a constant depending on the σ_x and L . When $r_d(\Sigma) \lesssim n$ (i.e., in the small or moderate TER regime), our result (25)–(26) in Proposition 1 matches (22)–(23) for $\lambda_{d+1} \leq \tau \leq \lambda_1$. When $r_d(\Sigma) > c_x n$ for some c_x depending only on σ_x (i.e., in the large TER regime), our result (38)–(39) in Proposition 7 matches (22)–(23) for $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_1$.

To control the conditional number of A_d , Tsigler & Bartlett (2023) requires an oracle small-ball assumption on covariate vectors: $\sum_{j>d} x_{ij}^2 > c(\sum_{j>d} \lambda_j + n\tau)$ for all $i = 1, \dots, n$ and some c satisfying $0 < c < 1$ with a high probability.

Instead of requiring an extra oracle assumption, our analysis achieves the control of the conditional number of A_d using concentration inequalities specifically in the two TER regimes for certain ranges of τ , which are summarized below. See Section 6.1.3 and Supplement Section II.4 for details. We derive an upper bound of $\mu_1(A_d)$, using concentration properties of matrix operator norms based on sub-gaussian covariate vectors, for $\tau \geq \lambda_{d+1}$ in the small or moderate TER regime and for $\tau \geq 0$ in the large TER regime. To obtain a lower bound of $\mu_n(A_d)$, we handle the small or moderate TER regime and the large TER regime separately. In small or moderate TER regime, we use a trivial lower bound: $\mu_n(A_d) \geq n\tau$. In the large TER regime, we use concentration properties of quadratic forms of sub-gaussian random vectors (Zajkowski (2020), Corollary 2.8) to show that the oracle small-ball assumption in Tsigler & Bartlett (2023) is satisfied for $\tau \geq 0$. Then we derive a lower bound of $\mu_n(A_d)$ following similar reasoning as in Tsigler & Bartlett (2023).

Lower bound of V_{out} . Our lower bound of V_{out} also matches the result in Tsigler & Bartlett (2023) for a certain range of ridge tuning parameters, but our result does not require the independence of the components of the whitened covariate vector as assumed in Tsigler & Bartlett (2023). In fact, given that the components of whitened x_i are independent, for any d small enough compared

to n , it is shown in Tsigler & Bartlett (2023) that with a high probability,

$$V_{\text{out}}/c \geq \frac{1}{n} \sum_{j=1}^p \min\left\{1, \frac{\lambda_j^2}{\lambda_{d+1}^2 \left(\frac{\sum_{l \geq d} \lambda_l + n\tau}{n\lambda_{d+1}} + 1\right)^2}\right\}. \quad (24)$$

where c is a constant depending on σ_x . When $r_d(\Sigma) \lesssim n$ (i.e., in the small or moderate TER regime), for $\lambda_{d+1} \leq \tau \leq \lambda_d$, (24) can be shown to reduce to

$$V_{\text{out}} \gtrsim \frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n},$$

which matches our result in (30) of Proposition 3 for small or moderate TER regime. When $r_d(\Sigma) > c_x n$ for some c_x depending on σ_x (i.e., in the large TER regime), for $\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau \leq \lambda_d$, (24) can be shown to reduce to

$$V_{\text{out}} \gtrsim \frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n},$$

which matches our result in (42) of Proposition 9 for large TER regime.

Instead of requiring the assumption of independent components in whitened covariate vectors as in Tsigler & Bartlett (2023), we derive a lower bound of V_{out} using concentration properties of sub-gaussian random vectors (see Section 6.1 for further information).

Lower bound of B_{out} . The lower bound of B_{out} in Tsigler & Bartlett (2023) is provided as a probability lower bound on the expectation of B_{out} with respect to a prior distribution on θ^* under an extra oracle assumption on certain modifications of matrix $A = XX^T + n\tau I_n$. Our lower bound on B_{out} is a direct probability bound without assuming a prior distribution on θ^* and any extra oracle assumption on covariate vectors, but focuses on the rotationally sparse setting.

4.3 Comparison with Bunea et al. (2022)

For the min-norm interpolator ($\tau = 0$), we compare Theorem 3 with Theorem 16 in Bunea et al. (2022), which are both obtained in the large TER regime. See Supplement Section V for details. Note that models (1) and (3) are related via $\Sigma = A\Sigma_Z A^T + \Sigma_E$ and $\theta^* = (\Sigma_E + A\Sigma_Z A^T)^{-1} A\Sigma_Z \beta$. To facilitate the comparison, we let $\Sigma_Z = I_d$, $\Sigma_E = \text{Diag}(\underbrace{\lambda_{d+1}, \dots, \lambda_{d+1}}_{d \text{ entries}}, \lambda_{d+1}, \dots, \lambda_p) \in \mathbb{R}^{p \times p}$, and

$$A = \begin{pmatrix} \text{Diag}(\sqrt{\lambda_1 - \lambda_{d+1}}, \dots, \sqrt{\lambda_d - \lambda_{d+1}}) \\ 0_{(p-d) \times d} \end{pmatrix} \in \mathbb{R}^{p \times d},$$

such that $\Sigma = \text{Diag}(\lambda_1, \dots, \lambda_d) \in \mathbb{R}^{p \times p}$ and

$$\theta^* = \left(\text{Diag}\left(\frac{\sqrt{\lambda_1 - \lambda_d}}{\lambda_1}, \dots, \frac{\sqrt{\lambda_d - \lambda_{d+1}}}{\lambda_d}\right) \beta, \underbrace{0, \dots, 0}_{p-d \text{ entries}} \right)^T \in \mathbb{R}^p.$$

We further assume that $\lambda_1 \asymp \lambda_d$, $\lambda_d > c_1 \lambda_{d+1}$ and $r_d(\Sigma) > c_2 d$ for some $c_1 > 1$ and $c_2 > 1$. In this setting, Theorem 3 with $\tau = 0$ shows that with a high probability,

$$\begin{aligned} \text{MSE}_{\text{out}} &\asymp \underbrace{\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2 \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d^2(\Sigma)}{n^2}}_{\text{B}_{\text{out}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{n r_d(\Sigma^2)}{r_d^2(\Sigma)} \right)}_{\text{V}_{\text{out}}}, \quad \text{for } \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_d, \\ \text{MSE}_{\text{out}} &\gtrsim \underbrace{\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2}_{\text{B}_{\text{out}}}, \quad \text{for } \lambda_{d+1} \frac{r_d(\Sigma)}{n} > \lambda_d. \end{aligned}$$

After ignoring the $\log(n)$ factor, Theorem 16 in Bunea et al. (2022) gives that with a high probability,

$$\text{MSE}_{\text{out}} \lesssim \underbrace{\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2 \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}}_{\text{B}_{\text{out}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{n}{r_d(\Sigma)} \right)}_{\text{V}_{\text{out}}}.$$

Hence for $\lambda_d \geq \lambda_{d+1} \frac{r_d(\Sigma)}{n}$, our result gives the order of out-sample MSE which is sharper than the upper bound in Bunea et al. (2022). For $\lambda_d < \lambda_{d+1} \frac{r_d(\Sigma)}{n}$, our lower bound indicates that the out-sample MSE is larger than $\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2$ up to a constant, and accordingly the upper bound in Bunea et al. (2022) is larger than $\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2$ up to a constant.

5 Proofs of main results (Theorems 1–4)

We provide proofs of the main results (Theorems 1–4), depending on auxiliary bounds on B_{out} , B_{in} , V_{out} , V_{in} , for which the proofs are outlined in Section 6. Without loss of generality, we only consider the case of $A_0 = 1$ involved in the ranges of the ridge parameter τ .

5.1 Proof of Theorem 1

We provide auxiliary bounds for the out-sample squared bias and variance B_{out} and V_{out} under the small or moderate TER regime (Assumption 2).

Proposition 1 (Upper bound of out-sample error with small or moderate TER). *Under Assumption 1 and 2, for any ν satisfying $0 < \nu < \frac{1}{2}$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 18\exp\{-\frac{n}{C_0}\}$: for $\tau \geq \lambda_{d+1}$,*

$$\text{B}_{\text{out}} \leq \frac{(1 + C_1)^3 (1 + \nu + \eta_1)^2 \text{Poly}_6(\sigma_x)}{(1 - \nu - \eta_1)^4} (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau})^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2), \quad (25)$$

$$\text{V}_{\text{out}} \leq \frac{(1 + C_1)^2 \text{Poly}_6(\sigma_x)}{(1 - \nu - \eta_1)^4} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \right). \quad (26)$$

Further with Assumption 4(i),

$$\text{B}_{\text{out}} \leq \frac{(1 + C_1)^3 (1 + \nu + \eta_1)^2 \text{Poly}_6(\sigma_x)}{(1 - \nu - \eta_1)^4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau} \right)^{-2}. \quad (27)$$

Proposition 2 (Lower bound of B_{out} with small or moderate TER). *Under Assumption 1 and 4(i), for any ν satisfying $0 < \nu < \frac{1}{2}$, the following inequality holds uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 2\exp\{-\frac{n}{C_0}\}$: for $\tau \geq \lambda_{d+1}$,*

$$B_{\text{out}} \geq \frac{(1 - \sqrt{\delta_1})^2}{(1 + \nu + \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\tau}\right)^{-2}. \quad (28)$$

Proposition 3 (Lower bound of V_{out} with small or moderate TER). *Under Assumption 1 and 2, for any ν satisfying $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 12\exp\{-\frac{n}{C_0}\}$: for $\tau \leq \lambda_{d+1}$,*

$$V_{\text{out}} \geq \frac{(1 - \nu - \eta_1)^2 (\frac{1}{2} - \nu)}{(1 + C_1)^2 (1 + \nu + \eta_1)^4 \text{Poly}_4(\sigma_x)} \frac{1}{1 + \frac{2C_0 \sigma_x^2}{\frac{1}{2} - \eta_1}} \sigma^2 \left(\frac{d}{n} + \frac{r_d(\Sigma^2)}{n}\right), \quad (29)$$

and for $\lambda_d \geq \tau \geq \lambda_{d+1}$,

$$V_{\text{out}} \geq \frac{(1 - \nu - \eta_1)^2 (\frac{1}{2} - \nu)}{(1 + C_1)^2 (1 + \nu + \eta_1)^4 \text{Poly}_4(\sigma_x)} \frac{1}{1 + \frac{2C_0 \sigma_x^2}{\frac{1}{2} - \eta_1}} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n}\right). \quad (30)$$

Theorem 1 can be deduced by combining the bounds for B_{out} and V_{out} above. The probability control is determined from the intersection of the relevant events included in the propositions.

- If $\tau \leq \lambda_{d+1}$, then the lower bound for V_{out} in Theorem 1(i) is obtained from (29) in Proposition 3.
- If $\lambda_{d+1} \leq \tau \leq \lambda_d$, then the upper bounds for B_{out} and V_{out} in Theorem 1(ii) are obtained from (27) and (26) in Proposition 1, and the lower bounds for B_{out} and V_{out} in Theorem 1(ii) are obtained from (28) in Proposition 2 and (30) in Proposition 3.
- If $\tau \geq \lambda_d$, then the lower bound for B_{out} in Theorem 1(iii) is obtained from (28) in Proposition 2.

5.2 Proof of Theorem 2

We provide auxiliary bounds for the in-sample squared bias and variance B_{in} and V_{in} under the small or moderate TER regime (Assumption 2).

Proposition 4 (Upper bound of in-sample error with small or moderate TER). *Under Assumption 1 and 2, for any ν satisfying $0 < \nu < \frac{1}{2}$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 8\exp\{-\frac{n}{C_0}\}$: for $\tau \geq \lambda_{d+1}$,*

$$B_{\text{in}} \leq \frac{(1 + C_1)^4 \text{Poly}_8(\sigma_x)}{(1 - \nu - \eta_1)^2} (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau}\right)^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2), \quad (31)$$

$$V_{\text{in}} \leq (1 + C_1)^2 \text{Poly}_4(\sigma_x) \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma)}{n} \right). \quad (32)$$

Further with Assumption 4(ii),

$$B_{\text{in}} \leq \frac{(1 + C_1)^4 \text{Poly}_8(\sigma_x)}{(1 - \nu - \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau} \right)^{-2}. \quad (33)$$

Proposition 5 (Lower bound of B_{in} with small or moderate TER). *Under Assumption 1, 2 and 4(i), for any ν satisfying $0 < \nu < \frac{1}{4}$, the following inequality holds uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 8\exp\{-\frac{n}{C_0}\}$: for $\tau \geq \lambda_{d+1}$,*

$$B_{\text{in}} \geq \kappa_1(\tau) \frac{(1 - \sqrt{\delta_1})^2}{(1 + \nu + \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\tau} \right)^{-2}, \quad (34)$$

where $\kappa_1(\tau) = \max\{1 - (\frac{2C_0\sigma_x^2(2+C_1)\lambda_{d+1}}{\tau}(1 + 16(2C_0\sigma_x^2 + 1)(1 + C_1)\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}) + 64\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}), 0\}$.

Proposition 6 (Lower bound of V_{in} with small or moderate TER). *Under Assumption 1, 2, 4(i), then for any ν satisfying $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\} - 6\exp\{-\frac{n}{C_0}\}$: (i) for $\tau \leq \lambda_{d+1}$,*

$$V_{\text{in}} \geq \frac{(1 - \nu)^2}{\text{Poly}_4(\sigma_x)(1 + C_1)^2(1 + \frac{1}{(1-\nu-\eta_2)^2})^2} \sigma^2 \left(\frac{d}{n} + \frac{r_d^2(\Sigma)}{n^2} \right), \quad (35)$$

(ii) for $\lambda_{d+1} \leq \tau \leq \lambda_d$,

$$V_{\text{in}} \geq \frac{(1 - \nu)^2}{\text{Poly}_4(\sigma_x)(1 + C_1)^2(1 + \frac{1}{(1-\nu-\eta_2)^2})^2} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2} \right), \quad (36)$$

(iii) for $\tau \geq \lambda_d$,

$$V_{\text{in}} \geq \frac{(1 - \nu)^2}{\text{Poly}_4(\sigma_x)(1 + C_1)^2} \sigma^2 \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2}. \quad (37)$$

Theorem 2 can be deduced by combining the bounds for B_{in} and V_{in} above. The probability control is determined from the intersection of the relevant events included in the propositions.

- If $\tau \leq \lambda_{d+1}$, then the lower bound for V_{in} in Theorem 2(i) is obtained from (35) in Proposition 6.
- If $\lambda_{d+1} \leq \tau \leq \lambda_d$, then the upper bounds for B_{in} and V_{in} in Theorem 2(ii) are obtained from (33) and (32) in Proposition 4, and the lower bounds for B_{in} and V_{in} in Theorem 2(ii) are obtained from (34) in Proposition 5 and (36) in Proposition 6.
- If $\tau \geq \lambda_d$, then the lower bounds for B_{in} and V_{in} in Theorem 2(iii) are obtained from (34) in Proposition 5 and (37) in Proposition 6.

5.3 Proof of Theorem 3

We provide auxiliary bounds for the out-sample squared bias and variance B_{out} and V_{out} under the large TER regime (Assumption 3).

Proposition 7 (Upper bound of out-sample error with large TER). *Under Assumption 1, 3, for any ν satisfying $0 < \nu < \frac{1}{2}$ and $\frac{r_d(\Sigma)\nu^2}{C_0^2\sigma_x^4} > 1$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2\sigma_x^4}\} - 2n\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 16\exp\{-\frac{n}{C_0}\}$: for $\tau \geq 0$,*

$$B_{\text{out}} \leq \frac{(1 + \nu + \eta_1)^2(1 + \nu + \eta_2)^2 \text{Poly}_4(\sigma_x)}{(1 - \nu - \eta_1)^4(1 - \nu - \eta_2)^2} (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}})^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2), \quad (38)$$

$$V_{\text{out}} \leq \frac{(1 + \nu + \eta_2)^2 \text{Poly}_2(\sigma_x)}{(1 - \nu - \eta_1)^4(1 - \nu - \eta_2)^2} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} \right). \quad (39)$$

Further with Assumption 4(ii),

$$B_{\text{out}} \leq \frac{(1 + \nu + \eta_1)^2(1 + \nu + \eta_2)^2 \text{Poly}_4(\sigma_x)}{(1 - \nu - \eta_1)^4(1 - \nu - \eta_2)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \right)^{-2}. \quad (40)$$

Proposition 8 (Lower bound of B_{out} with large TER). *Under Assumption 1, 3 and 4(ii), for any ν satisfying $0 < \nu < \frac{1}{4}$ and $\frac{r_d(\Sigma)\nu^2}{C_0^2\sigma_x^4} > 1$, the following inequality holds uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu n}{C_0^2\sigma_x^4}\} - 2n\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 6\exp\{-\frac{n}{C_0}\}$: for $\tau \geq 0$,*

$$B_{\text{out}} \geq \frac{(1 - \sqrt{\delta_2})^2(1 - \nu - \eta_2)^2}{(1 + \nu + \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \right)^{-2}. \quad (41)$$

Proposition 9 (Lower bound of V_{out} with large TER). *Under Assumption 1 and 3, for any ν satisfying $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ and $\frac{r_d(\Sigma)\nu^2}{C_0^2\sigma_x^4} > 1$, the following inequality holds uniformly in the range of τ stated with probability $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2\sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0^2\sigma_x^4}\} - 2n\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 10\exp\{-\frac{n}{C_0}\}$: for $\tau + \frac{\sum_{j>d} \lambda_j}{n} \leq \lambda_d$,*

$$V_{\text{out}} \geq \frac{(1 - \nu - \eta_1)^2(1 - \nu - \eta_2)^2(\frac{1}{2} - \nu)}{4(1 + \nu + \eta_1)^4(1 + \nu + \eta_2)^2} \frac{8C_0\sigma_x^2(\frac{1}{2} - \eta_1)}{1 + 8C_0\sigma_x^2(\frac{1}{2} - \eta_1)} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} \right). \quad (42)$$

Theorem 3 can be deduced by combining the bounds for B_{out} and V_{out} above. The probability control is determined from the intersection of the relevant events included in the propositions.

- If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_d$, then the upper bound for B_{out} and V_{out} in Theorem 3(i) are obtained from (40) and (39) in Proposition 7, and the lower bounds for B_{out} and V_{out} in Theorem 3(i) are obtained from (41) in Proposition 8 and (42) in Proposition 9.
- If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \geq \lambda_d$, then the lower bound for B_{out} in Theorem 3(ii) is obtained from (41) in Proposition 8.

5.4 Proof of Theorem 4

We provide auxiliary bounds for the in-sample squared bias and variance B_{in} and V_{in} under the large TER regime (Assumption 3).

Proposition 10 (Upper bound of in-sample error with large TER). *Given Assumption 1 and 3, for any ν satisfying $0 < \nu < \frac{1}{2}$ and $\frac{r_d(\Sigma)\nu^2}{C_0^2\sigma_x^4} > 1$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0\sigma_x^4}\} - 2\exp\{-\frac{\nu^2 n}{C_0^2\sigma_x^4}\} - 2n\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 12\exp\{-\frac{n}{C_0}\}$: for $\tau \geq 0$,*

$$B_{\text{in}} \leq \frac{(1 + \nu + \eta_2)^2 \text{Poly}_6(\sigma_x)}{(1 - \nu - \eta_1)^2 (1 - \nu - \eta_2)^2} (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}})^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2), \quad (43)$$

$$V_{\text{in}} \leq \frac{(1 + \nu + \eta_2)^2 \text{Poly}_4(\sigma_x)}{(1 - \nu - \eta_2)^2} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2} \right). \quad (44)$$

Further with Assumption 4(ii),

$$B_{\text{in}} \leq \frac{(1 + \nu + \eta_2)^2 \text{Poly}_6(\sigma_x)}{(1 - \nu - \eta_1)^2 (1 - \nu - \eta_2)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \right)^{-2}. \quad (45)$$

Proposition 11 (Lower bound of B_{in} with large TER). *Under Assumption 1, 3 and 4(ii), for any ν satisfying $0 < \nu < \frac{1}{4}$ and $\frac{r_d(\Sigma)\nu^2}{C_0^2\sigma_x^4} > 1$, the following inequality holds uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2\sigma_x^4}\} - 2n\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 6\exp\{-\frac{n}{C_0}\}$: for $\tau \geq 0$,*

$$B_{\text{in}} \geq \frac{(1 - \sqrt{\delta_2})^2 (1 - \nu - \eta_2)^2}{(1 + \nu + \eta_1)^2} \kappa_2(\tau) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \right)^{-2}, \quad (46)$$

where $\kappa_2(\tau) = \max\{1 - (16 \frac{\lambda_{d+1} \frac{r_d(\Sigma)}{n}}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} (1 + 112 \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}) + 64 \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}), 0\}$.

Proposition 12 (Lower bound of V_{in} with large TER). *Under Assumption 1 and 3, for any ν satisfying $0 < \nu < \frac{1}{2}$ and $\frac{r_d(\Sigma)\nu^2}{C_0^2\sigma_x^4} > 1$, the following inequalities hold uniformly in the range of τ stated with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2\sigma_x^4}\} - 2n\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 4\exp\{-\frac{n}{C_0}\}$: for $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_d$,*

$$V_{\text{in}} \geq \frac{(1 - \nu - \eta_2)^2}{2(1 + \nu + \eta_2)^2 (1 + \frac{1}{(1 - \nu - \eta_1)^2})} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2} \right), \quad (47)$$

and for $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \geq \lambda_d$,

$$V_{\text{in}} \geq \frac{(1 - \nu - \eta_2)^2}{2(1 + \nu + \eta_2)^2} \sigma^2 \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}. \quad (48)$$

Theorem 4 can be deduced by combining the bounds for B_{in} and V_{in} above. The probability control is determined from the intersection of the relevant events included in the propositions.

- If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_d$, then the upper bounds for B_{in} and V_{in} in Theorem 4(i) are obtained from (45) and (44) of Proposition 10, and the lower bounds for B_{in} and V_{in} in Theorem 4(i) are obtained from (46) and (47) of Proposition 12.
- If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \geq \lambda_d$, then the lower bounds for B_{in} and V_{in} are obtained from (46) in Proposition 11 and (48) in Proposition 12.

6 Proof outlines of auxiliary results (Propositions 1–12)

We provide proof outlines of the auxiliary bounds (Proposition 1–12) used in the proofs in Section 5. In Section 6.1, we discuss the results for out-sample error including Propositions 1–3 and Propositions 7–9. In Section 6.2, we discuss the results for in-sample error including Propositions 4–6 and Propositions 10–12. We introduce the following notation.

- $A = XX^T + n\tau I_n$, $A_d = X_{(d+1):p} X_{(d+1):p}^T + n\tau I_n$.
- $X_{1:d}$ denotes the the matrices comprised of the first d columns of X and $X_{(d+1):p}$ denotes the the matrices comprised of the last $p - d$ columns of X .
- $\hat{\Sigma}_{1:d} = \frac{X_{1:d}^T X_{1:d}}{n}$, $\hat{\Sigma}_{(d+1):p} = \frac{X_{(d+1):p}^T X_{(d+1):p}}{n}$ and $\hat{\Sigma}_{1:d,(d+1):p} = \frac{X_{1:d}^T X_{(d+1):p}}{n}$.
- $H_d = \Sigma_{1:d}^{-1/2} X_{1:d}^T$, $\hat{H}_d = \hat{\Sigma}_{1:d}^{-1/2} X_{1:d}^T$.
- $M_d = X_{(d+1):p} \Sigma_{(d+1):p} X_{(d+1):p}^T$, $\hat{M}_d = X_{(d+1):p} \hat{\Sigma}_{(d+1):p} X_{(d+1):p}^T$.
- $\mu_j(M)$ is the j -th largest eigenvalue of symmetric semi-positive definite matrix M .
- c, c_1, c_2, c_3 are absolute constants that may differ from line to line.
- $c(\sigma_x), c_1(\sigma_x), c_2(\sigma_x), c_3(\sigma_x), c_4(\sigma_x)$ are constants related to σ_x which may differ from line to line.
- w.h.p. indicates that an event holds with probability at least $1 - \frac{n}{c} \exp(-c\sqrt{n})$ for a constant c and sample size n .

The out-sample squared bias and variance can be decomposed as follows:

$$\begin{aligned}
B_{\text{out}} &= \underbrace{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\Sigma_{1:d}}^2}_{B_{\text{out},1}} + \underbrace{\|\theta_{(d+1):p}^* - X_{(d+1):p}^T A^{-1} X \theta^*\|_{\Sigma_{(d+1):p}}^2}_{B_{\text{out},2}}, \\
V_{\text{out}} &= \underbrace{\sigma^2 \text{Tr}(A^{-1} X_{1:d} \Sigma_{1:d} X_{1:d}^T A^{-1})}_{V_{\text{out},1}} + \underbrace{\sigma^2 \text{Tr}(A^{-1} X_{(d+1):p} \Sigma_{(d+1):p} X_{(d+1):p}^T A^{-1})}_{V_{\text{out},2}}.
\end{aligned}$$

Similarly, the in-sample squared bias and variance can be shown to satisfy

$$\begin{aligned}
B_{\text{in}} &= \underbrace{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}^2}_{B_{\text{in},1}} + \underbrace{\|\theta_{(d+1):p}^* - X_{d+1:p}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{d+1:p}}^2}_{B_{\text{in},2}} \\
&\quad + \underbrace{2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \hat{\Sigma}_{1:d,(d+1):p} (\theta_{(d+1):p}^* - X_{d+1:p}^T A^{-1} X \theta^*)}_{B_{\text{in},12}}, \\
V_{\text{in}} &\leq \underbrace{2\sigma^2 \text{Tr}(A^{-1} X_{1:d} \hat{\Sigma}_{1:d} X_{1:d}^T A^{-1})}_{V_{\text{in},1}} + \underbrace{2\sigma^2 \text{Tr}(A^{-1} X_{(d+1):p} \hat{\Sigma}_{(d+1):p} X_{(d+1):p}^T A^{-1})}_{V_{\text{in},2}}.
\end{aligned}$$

where the Cauchy–Schwartz inequality is used to bound V_{in} .

6.1 Derivation of bounds for out-sample error

To derive the bounds for out-sample error, we first build algebraic bounds of the out-sample error. Then we provide intermediate bounds of the out-sample error by controlling some random quantities in the algebraic bounds. We deduce the final bounds mainly by further controlling the extreme eigenvalues $\mu_1(A_d)$ and $\mu_n(A_d)$ (from Lemma S7) in the intermediate bounds and incorporating Assumption 4 (rotational sparsity) to control quantities related to $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}$. The details of our derivation are presented in Supplement Section II.1.

6.1.1 Algebraic bounds of out-sample error

The first step of deriving the bounds is to build algebraic bounds for the out-sample bias and variance. The lemmas below are inspired by Lemma 27 and 28 in Tsigler & Bartlett (2023).

Lemma 1 (Algebraic upper bounds of out-sample error). *Given invertible $\Sigma_{1:d}$, we have*

$$\begin{aligned}
B_{\text{out}} &\leq 2\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n(H_d H_d^T)}{\mu_1(A_d)} \right)^{-2} + \frac{2\mu_1(H_d H_d^T)}{\mu_n^2(H_d H_d^T)} \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \\
&\quad + 3\left(\|M_d\| \frac{\mu_1(H_d H_d^T)}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n(H_d H_d^T)}{\mu_1(A_d)} \right)^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \right. \\
&\quad \left. + \frac{\|M_d\|}{\mu_n(A)^2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \right), \\
V_{\text{out}} &\leq \sigma^2 \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \frac{\text{Tr}(H_d^T H_d)}{\mu_d(H_d H_d^T)^2} + \sigma^2 \frac{\text{Tr}(M_d)}{\mu_n(A_d)^2}.
\end{aligned} \tag{49}$$

Lemma 2 (Algebraic lower bounds of out-sample error). *Given invertible $\Sigma_{1:d}$, we have*

$$V_{\text{out}} \geq \sigma^2 \frac{d\mu_d(H_d H_d^T)}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^{-2} + \max\{0, \sigma^2 \frac{\text{Tr}(M_d) - d\mu_1(M_d)}{\mu_1(A_d)^2}\}.$$

Further if $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{\mu_1(H_d H_d)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)}$, then

$$B_{\text{out}} \geq \left(1 - \frac{\mu_1(H_d H_d^T)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}}\right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^{-2}.$$

6.1.2 Intermediate bounds of out-sample error

Based on the sub-gaussianity of the covariate vectors, the following random quantities in Lemmas 1 and 2 can be controlled with high probability. See Supplement Lemma S6 (i), (ii), (iii) and (iv).

- Bounds of eigenvalues of $H_d H_d^T$: w.h.p.

$$\mu_1(H_d H_d^T) \leq c_1 n, \quad \mu_d(H_d H_d^T) \geq c_2 n.$$

- Bounds of traces of random matrices: w.h.p.

$$\text{Tr}(H_d^T H_d) = \text{Tr}(X_{1:d} \Sigma_{1:d}^{-1} X_{1:d}) \leq c_1(\sigma_x) n d,$$

$$\text{Tr}(M_d) = \text{Tr}(X_{(d+1):p} \Sigma_{(d+1):p} X_{(d+1):p}^T) \leq c_2(\sigma_x) n \sum_{j>d} \lambda_j^2,$$

$$\text{Tr}(M_d) = \text{Tr}(X_{(d+1):p} \Sigma_{(d+1):p} X_{(d+1):p}^T) \geq c_1 n \sum_{j>d} \lambda_j^2.$$

- Bounds of norms of random matrix and vector: w.h.p.

$$\mu_1(M_d) = \|M_d\| \leq c_1(\sigma_x)(n\lambda_{d+1}^2 + \sum_{j>d} \lambda_j^2), \quad \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \leq c_2(\sigma_x) n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2.$$

Substituting the probability bounds above into the algebraic bounds in Lemmas 1 and 2 yields the intermediate bounds below with high probability.

Upper bounds:

$$\begin{aligned} B_{\text{out}} &\leq c_1 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} + c_1(\sigma_x) \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\quad + c_2(\sigma_x) \frac{(n^2 \lambda_{d+1}^2 + n \sum_{j>d} \lambda_j^2)}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} \\ &\quad + (3 + c_3(\sigma_x) \frac{(n^2 \lambda_{d+1}^2 + n \sum_{j>d} \lambda_j^2)}{\mu_n(A_d)^2}) \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2, \\ V_{\text{out}} &\leq c_1(\sigma_x) \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \sigma^2 \frac{d}{n} + c_2(\sigma_x) \sigma^2 \frac{n \sum_{j>d} \lambda_j^2}{\mu_n(A_d)^2}. \end{aligned}$$

Lower bounds:

$$V_{\text{out}} \geq c_1 \sigma^2 \frac{n d}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2} + \sigma^2 \max\{0, c_2 \frac{n \sum_{j>d} \lambda_j^2}{\mu_1(A_d)^2} (1 - c_1(\sigma_x) (\frac{d}{r_d(\Sigma^2)} + \frac{d}{n}))\}.$$

Further if $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{\mu_1(H_d H_d)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)}$, then

$$B_{\text{out}} \geq c(1 - \frac{c_1(\sigma_x) n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2}.$$

6.1.3 Final bounds for out-sample error

The eigenvalues $\mu_1(A_d)$ and $\mu_n(A_d)$ in the intermediate bounds above can be controlled with high probability respectively in the small or moderate TER regime and the large TER regime. See Supplement Lemma S7.

- Under the small or moderate TER regime (Assumption 2), for $\tau \geq \lambda_{d+1}$,

$$\mu_1(A_d) \leq c(\sigma_x)n\tau, \quad \mu_n(A_d) \geq n\tau.$$

- Under the large TER regime (Assumption 3), for $\tau \geq 0$,

$$\mu_1(A_d) \leq c_1n(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n}), \quad \mu_n(A_d) \geq c_2n(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n}).$$

We deduce the final bounds for the out-sample error from the intermediate bounds as follows, mainly by applying the probability bounds on $\mu_1(A_d)$ and $\mu_n(A_d)$ and incorporating Assumption 4 (rotational sparsity).

Upper bounds of B_{out} , V_{out} (Proposition 1,7). We first substitute the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ into the intermediate bounds. Then we incorporate Assumption 4 to control quantities related to $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}$. This leads to the final upper bounds for B_{out} and V_{out} .

Lower bound of B_{out} (Proposition 2,8). We first show that, w.h.p.

$$\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{\mu_1(H_d H_d)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)},$$

if $\tau \geq \lambda_{d+1}$ under Assumption 2 (small or moderate TER) or if $\tau \geq 0$ under Assumption 3 (large TER). Then we substitute the bounds of $\mu_n(A_d)$ into the intermediate lower bound of B_{out} and incorporate Assumption 4 to control quantities related to $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2$. This leads to the final lower bound for B_{out} .

Lower bound of V_{out} (Proposition 3,9). The intermediate lower bound of the out-sample variance in Section 6.1.2 is

$$V_{\text{out}} \geq c\sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2} + \sigma^2 \max\{0, c_2 \frac{n \sum_{j>d} \lambda_j^2}{\mu_1(A_d)^2} (1 - c_1(\sigma_x) (\frac{d}{r_d(\Sigma^2)} + \frac{d}{n}))\}. \quad (50)$$

To deduce the final lower bound for V_{out} , our strategy is as follows. We first derive a lower bound for the first term $\sigma^2 \frac{nd}{\mu_1(A_d)^2} (\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)})^{-2}$. Then we discuss two complementary cases. See Supplement Section II.1.4 for details. The first case is that $\frac{r_d(\Sigma^2)}{d}$ is upper bounded by a constant (possibly depending on σ_x). The second case is that $\frac{r_d(\Sigma^2)}{d}$ is large enough such that $(1 - c_1(\sigma_x) (\frac{d}{r_d(\Sigma^2)} + \frac{d}{n})) > 0$. Lastly, we show that in these two cases, V_{out} satisfies lower bounds of the same order, which gives the final lower bound for V_{out} .

For small or moderate TER, after substituting the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ into the first term of the intermediate lower bound in (50), we have for $\tau \geq \lambda_{d+1}$, w.h.p.

$$\sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2} \geq c(\sigma_x) \sigma^2 \frac{d}{n\tau^2} \left(\frac{1}{\lambda_d} + \frac{1}{\tau} \right)^{-2},$$

which implies that for $\lambda_{d+1} \leq \tau \leq \lambda_d$, w.h.p.

$$\sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2} \geq c(\sigma_x) \sigma^2 \frac{d}{n}. \quad (51)$$

Then we discuss two cases which are complementary to each other. The first case is that $\frac{r_d(\Sigma^2)}{d}$ is upper bounded by a constant $c(\sigma_x)$. In this case, we have for $\tau \geq \lambda_{d+1}$,

$$d \geq \frac{1}{c(\sigma_x)} \frac{\sum_{j>d} \lambda_j^2}{\lambda_{d+1}^2} \geq \frac{1}{c(\sigma_x)} \frac{\sum_{j>d} \lambda_j^2}{\tau^2},$$

and hence (allowing that $c(\sigma_x)$ below may vary from the previous line)

$$\frac{d}{n} \geq c(\sigma_x) \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n\tau^2} \right). \quad (52)$$

Combining (50), (51) and (52) shows that for $\lambda_{d+1} \leq \tau \leq \lambda_d$, we have, w.h.p.

$$V_{\text{out}} \geq c\sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2} \geq c_1(\sigma_x) \sigma^2 \frac{d}{n} \geq c_2(\sigma_x) \sigma^2 \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n\tau^2} \right).$$

The second case is that $\frac{r_d(\Sigma^2)}{d}$ is large enough such that $1 - c_1(\sigma_x) \left(\frac{d}{r_d(\Sigma^2)} + \frac{d}{n} \right) > 0$. In this case, after substituting the upper bound of $\mu_1(A_d)$ into the second term of the intermediate lower bound in (50), we have for $\tau \geq \lambda_{d+1}$, w.h.p.

$$\sigma^2 \max\left\{0, \frac{n \sum_{j>d} \lambda_j^2}{\mu_1(A_d)^2} \left(1 - c_1(\sigma_x) \left(\frac{d}{r_d(\Sigma^2)} + \frac{d}{n} \right) \right) \right\} \geq c(\sigma_x) \frac{\sum_{j>d} \lambda_j^2}{n\tau^2}. \quad (53)$$

Then combining (50), (51) and (53) yields that for $\lambda_{d+1} \leq \tau \leq \lambda_d$, w.h.p.

$$V_{\text{out}} \geq c(\sigma_x) \sigma^2 \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n\tau^2} \right).$$

In conclusion of the two cases, it holds that for $\lambda_{d+1} \leq \tau \leq \lambda_d$, w.h.p.

$$V_{\text{out}} \geq c(\sigma_x) \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \right).$$

For $\tau \leq \lambda_{d+1}$, the lower bound of V_{out} follows by the monotonicity of variance: V_{out} for $\tau \leq \lambda_{d+1}$ is no smaller than V_{out} for $\tau = \lambda_{d+1}$. Hence for $\tau \leq \lambda_{d+1}$,

$$V_{\text{out}} \geq c(\sigma_x) \sigma^2 \left(\frac{d}{n} + \frac{r_d(\Sigma^2)}{n} \right).$$

The lower bound of V_{out} in the large TER regime can be derived similarly to the small or moderate TER regime. For succinctness, we omit the associated details.

6.2 Derivation of bounds for in-sample error

Our strategy for deriving the bounds for in-sample error is similar to that for out-sample error. The details of our derivation are presented in Supplement Section II.2.

6.2.1 Algebraic bounds of in-sample error

Similarly to the out-sample error, we first give the algebraic bounds for in-sample bias and variance.

Lemma 3 (Algebraic upper bounds of in-sample error). *Given invertible $\hat{\Sigma}_{1:d}$, we have*

$$\begin{aligned} B_{\text{in}} &\leq 2\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n^2}{\mu_1^2(A_d)} \right)^{-2} + \frac{2}{n} \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \\ &\quad + 3 \left(\|\hat{M}_d\| \frac{n}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} + \|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 + \frac{\|\hat{M}_d\|}{\mu_n(A)^2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \right), \\ V_{\text{in}} &\leq 2\sigma^2 \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \frac{d}{n} + 2\sigma^2 \frac{\frac{1}{n} \text{Tr}(X_{(d+1):p} X_{(d+1):p}^T) \mu_1(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_n(A_d)^2}. \end{aligned}$$

Lemma 4 (Algebraic lower bounds of in-sample error). *Given invertible $\hat{\Sigma}_{1:d}$, we have*

$$V_{\text{in}} \geq \sigma^2 \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^d \frac{\mu_i^2(X_{1:d} X_{1:d}^T)}{(\mu_i(X_{1:d} X_{1:d}^T) + n\tau)^2} + \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \right).$$

Further if $\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} \geq \frac{n^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)}$, then

$$B_{\text{in}} \geq \max\{0, 1 - \frac{|B_{\text{in},12}|}{B_{\text{in}}}\} \left(1 - \frac{n^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}} \right)^2 \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2}.$$

6.2.2 Intermediate bounds of in-sample error

In addition to the probability bounds in Section 6.1.2, the following probability bounds can be obtained about random quantities in Lemmas 3 and 4. See Supplement Lemma S6 (i), (v) and (vii).

- Bounds of $\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2$, w.h.p.

$$c_1 \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \leq \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \leq c_2 \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2.$$

- Bounds of $\mu_1(X_{(d+1):p} X_{(d+1):p}^T)$ and $\|\hat{M}_d\|$, w.h.p.

$$\mu_1(X_{(d+1):p} X_{(d+1):p}^T) \leq c_1(\sigma_x)(n\lambda_{d+1} + \sum_{j>d} \lambda_j), \quad \|\hat{M}_d\| \leq c_2(\sigma_x) \frac{(n\lambda_{d+1} + \sum_{j>d} \lambda_j)^2}{n}.$$

- Bounds of $\text{Tr}(X_{(d+1):p} X_{(d+1):p}^T)$, w.h.p.

$$\text{Tr}(X_{(d+1):p} X_{(d+1):p}^T) \leq cn \sum_{j>d} \lambda_j.$$

Substituting the probability bounds in Section 6.1.2 and above into the algebraic bounds in Lemmas 3 and 4 yields the intermediate bounds below, w.h.p.

Upper bounds:

$$\begin{aligned} B_{\text{in}} &\leq c_1 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} + c_1(\sigma_x) \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\quad + c_2(\sigma_x) \frac{(n\lambda_{d+1} + \sum_{j>d} \lambda_j)^2}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} \\ &\quad + (c_3(\sigma_x) + c_4(\sigma_x) \frac{(n\lambda_{d+1} + \sum_{j>d} \lambda_j)^2}{\mu_n(A_d)^2}) \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2, \\ V_{\text{in}} &\leq c_3 \sigma^2 \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \frac{d}{n} + c_3(\sigma_x) \sigma^2 \frac{(\sum_{j>d} \lambda_j)(n\lambda_{d+1} + \sum_{j>d} \lambda_j)}{\mu_n(A_d)^2}. \end{aligned}$$

Lower bounds:

Further if $\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} \geq \frac{n^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)}$, then

$$B_{\text{in}} \geq c_1 \max\{0, 1 - \frac{|B_{\text{in},12}|}{B_{\text{in}}}\} \left(1 - \frac{c_1(\sigma_x) n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2} \right) \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2}.$$

6.2.3 Final bounds for in-sample error

We deduce the final bounds for the in-sample error from the intermediate bounds as follows, mainly by applying the probability bounds on $\mu_1(A_d)$ and $\mu_n(A_d)$ (from Section 6.1.3) and incorporating Assumption 4 (rotational sparsity).

Upper bounds of B_{in} , V_{in} (Proposition 4,10). We first substitute the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ into the intermediate bounds. Then we incorporate Assumption 4 to control quantities related to $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2$. This leads to the final upper bounds for B_{in} and V_{in} .

Lower bound of B_{in} (Proposition 5,11). We first show that w.h.p.

$$\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} \geq \frac{n^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)},$$

if $\tau \geq \lambda_{d+1}$ under Assumption 2 (small or moderate TER) or if $\tau \geq 0$ under Assumption 3 (large TER). Then we show that w.h.p.

$$\max\{0, 1 - \frac{|B_{\text{in},2}|}{B_{\text{in},1}}\} \geq \kappa_1(\tau),$$

if $\tau \geq \lambda_{d+1}$ under Assumption 2 (small or moderate regime TER regime) and Assumption 4(i), or

$$\max\{0, 1 - \frac{|B_{\text{in},2}|}{B_{\text{in},1}}\} \geq \kappa_2(\tau),$$

if $\tau \geq 0$ under Assumption 3 (large TER regime) and Assumption 4(ii). See Theorems 2 and 4 for the definition of $\kappa_1(\tau)$ and $\kappa_2(\tau)$, and Supplement Lemma S5 for details. We substitute the

bounds of $\mu_n(A_d)$ into the intermediate lower bound of B_{in} and incorporate Assumption 4 to control quantities related to $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}$. This leads to the final lower bound of B_{in} .

Lower bound of V_{in} (Proposition 6,12). The algebraic lower bound of in-sample variance from Section 6.2.1 is

$$V_{\text{in}} \geq \sigma^2 \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^d \frac{\mu_i^2(X_{1:d} X_{1:d}^T)}{(\mu_i(X_{1:d} X_{1:d}^T) + n\tau)^2} + \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \right). \quad (54)$$

By the concentration of $X_{1:d}^T X_{1:d}$ (see Supplement Lemma S6 (i)), we have w.h.p.

$$\mu_1(X_{1:d}^T X_{1:d}) \geq \dots \geq \mu_d(X_{1:d}^T X_{1:d}) \geq c\lambda_d.$$

If $\tau \leq \lambda_d$, then the first term in the algebraic bound satisfies w.h.p.

$$\frac{1}{n} \sum_{i=1}^d \frac{\mu_i^2(X_{1:d} X_{1:d}^T)}{(\mu_i(X_{1:d} X_{1:d}^T) + n\tau)^2} \geq c \frac{d}{n}.$$

For the second term of the algebraic lower bound, we first give the algebraic bound

$$\sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \geq \sigma^2 \frac{1}{n^2} \frac{\text{Tr}(X_{(d+1):p} X_{(d+1):p}^T)^2}{\mu_1(A_d)^2}.$$

By the control of $\mu_1(A_d)$ in Section 6.1.3 and the fact that w.h.p. (see Supplement Lemma S6 (vii))

$$\text{Tr}(X_{(d+1):p} X_{(d+1):p}^T) \geq cn \sum_{j>d} \lambda_j,$$

the second term in the algebraic bound (54) can be lower bounded as follows.

- In small or moderate TER regime, under Assumption 2, for $\tau \geq \lambda_{d+1}$, w.h.p.

$$\sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \geq c(\sigma_x) \sigma^2 \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2}.$$

- In large TER regime, under Assumption 3, for $\tau \geq 0$, w.h.p.

$$\sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \geq c\sigma^2 \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}.$$

Combining the preceding bounds on the two terms of (54) gives the lower bound of V_{in} .

7 Numerical studies

We present numerical results in support of our theoretical results, including the sufficient conditions and necessary conditions for $\text{MSE} = O(\frac{d}{n})$ and the conditions for when $\text{MSE}_{\text{out}}^*$ can be much smaller than MSE_{in}^* as described in Remark 1 and 4.

7.1 Data generation and MSE calculation

We first generate the covariance matrix Σ and coefficient vector θ^* .

Generating Σ . Given $0 < d < p$ and $\rho < 1$, we generate a diagonal covariance matrix Σ as follows. We let $\lambda_i = 1$ for $i = 1, \dots, d$ and let $\lambda_i = \rho$ for $i = d + 1, \dots, p$ unless otherwise stated. Here ρ represents the gap between the spiked and tail eigenvalues of Σ .

Generating θ^* . Given the covariance matrix Σ , we generate $\theta^* \in \mathbb{R}^p$ as follows. We let $\theta_{1:d}^* = \frac{1}{\sqrt{d}}$. To generate $\theta_{(d+1):p}^*$, we first generate $\beta_{(d+1):p} \sim N_{p-d}(0, I)$ and then let

$$\theta_{(d+1):p}^* = \begin{cases} \frac{\beta_{(d+1):p}}{\|\beta_{(d+1):p}\|_{\Sigma_{(d+1):p}}} \sqrt{0.01 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_{d+1}^2}, & \text{if } r_d(\Sigma) < 10n, \\ \frac{\beta_{(d+1):p}}{\|\beta_{(d+1):p}\|_{\Sigma_{(d+1):p}}} \sqrt{0.01 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{\sum_{i>d} \lambda_i}\right)^{-2}}, & \text{if } r_d(\Sigma) \geq 10n. \end{cases}$$

In the numerical study, we consider $r_d(\Sigma) < 10n$ as the small or moderate TER regime and consider $r_d(\Sigma) \geq 10n$ as the large TER regime. From the generating process above, $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 = 1$ and $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2$ satisfies rotational sparsity Assumption 4(i) in small or moderate TER regime or satisfies rotational sparsity Assumption 4(ii) in large TER regime.

Generating x_i and y_i . Given Σ and θ^* from above, we generate data x_i and y_i for $i = 1, \dots, n$ as follows. We sample $z_{1i} \sim \text{unif}(\sqrt{p}S^{p-1})$ and $z_{2i} \sim N_p(0, I)$, where S^{p-1} is the spherical surface with radius 1 in $\mathbb{R}^p$. Then we let $z_i = \frac{\sqrt{2}}{2}z_{1i} + \frac{\sqrt{2}}{2}z_{2i}$ and $x_i = \Sigma^{1/2}z_i$. By the generating process, $z_i \in \mathbb{R}^p$ is an isotropic random vector with dependent components. Then we sample $\epsilon_i \sim N(0, 1)$ and generate $y_i = x_i^T \theta^* + \epsilon_i$ by model (1).

With x_i and y_i generated from above, for different ridge parameters τ , we calculate MSE_{out} or MSE_{in} according to (5) or (6), where y_i 's are averaged out. We report the MSE_{out} and MSE_{in} based on the average of 10 repeated runs of data generation.

7.2 Experiment settings

7.2.1 Study of conditions for $\text{MSE} = O(\frac{d}{n})$

We use the following settings to study the sufficient conditions and the necessary conditions for $\text{MSE} = O(\frac{d}{n})$ in Corollary 1, 2, 4 and 5. We focus on the scenarios where the sufficient condition matches the necessary condition up to a constant, that is, $\text{MSE} = O(\frac{d}{n})$ if and only if the ratio $\frac{\lambda_{d+1}}{\lambda_d}$ is smaller than or equal to a certain threshold.

Study of Corollary 1. Given the small or moderate TER regime, $n \gg d$ and $r_d(\Sigma^2) \gg d$, from Corollary 1, the sufficient condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$ and the necessary condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$. To embody this condition, we set $d = 5, n = 1500, p = 1500$ and $\rho = [0.1, 1, 10] \times \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$ such that $r_d(\Sigma^2) = 1495$.

Study of Corollary 2. Given the small or moderate TER regime, $n \gg d$ and $r_d(\Sigma) \asymp n$, from Corollary 2, the sufficient condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$ and the necessary condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$. To embody this condition, we set $d = 5, n = 1500, p = 1500$ and $\rho = [0.1, 1, 10] \times \frac{d}{r_d(\Sigma)}$ such that $r_d(\Sigma) = 1495$.

Study of Corollary 4. Given the large TER regime and $n \gg d$, from Corollary 4, the sufficient condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ and the necessary condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ are the same. The condition is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{\sqrt{nd}}{r_d(\Sigma)}$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$ and the condition is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{\sqrt{nr_d(\Sigma^2)}}$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$. To embody the first condition, we set $d = 5, n = 50, p = 1500$ such that $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} = 0.577 \leq 1$ and $\rho = [0.1, 1, 10] \times \frac{\sqrt{nd}}{r_d(\Sigma)}$. To embody the second condition, we set $d = 5, n = 150, p = 1500$ such that $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} = 1.732 > 1$ and $\rho = [0.1, 1, 10] \times \frac{d}{\sqrt{nr_d(\Sigma^2)}}$.

Study of Corollary 5. Given the large TER regime and $n \gg d$, from Corollary 5, the sufficient condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ and the necessary condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ are the same, and the condition is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$. To embody this condition, we set $d = 5, n = 150, p = 1500$ and $\rho = [0.1, 1, 10] \times \frac{d}{r_d(\Sigma)}$.

7.2.2 Study of conditions for $\text{MSE}_{\text{out}}^*$ much smaller than MSE_{in}^*

We use the following settings to study the conditions for when $\text{MSE}_{\text{out}}^*$ can be much smaller than MSE_{in}^* , which is discussed in Remark 1 and 4.

(i) In the small or moderate TER regime, $\text{MSE}_{\text{out}}^*$ can be much smaller than MSE_{in}^* if

$$r_d(\Sigma) \asymp n, \frac{\lambda_{d+1}}{\lambda_d} \gtrsim \frac{d}{n} \sqrt{\frac{n}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma^2)} \gg 1.$$

To embody this condition, we set $d = 2, n = 300, p = 15000, \lambda_1 = \dots = \lambda_d = 1, \lambda_{d+1} = \dots = \lambda_{11d} = \rho, \lambda_{11d+1} = \dots = \lambda_p = 0.02\rho$ and $\rho = \frac{d}{n} \sqrt{\frac{n}{r_d(\Sigma^2)}}$ such that $r_d(\Sigma) = 319.56, \frac{\lambda_{d+1}}{\lambda_d} = \frac{d}{n} \sqrt{\frac{n}{r_d(\Sigma^2)}}$ and $\frac{n}{r_d(\Sigma^2)} = 11.54$.

(ii) In the large TER regime, $\text{MSE}_{\text{out}}^*$ can be much smaller than MSE_{in}^* if

$$\frac{\lambda_{d+1}}{\lambda_d} \gtrsim \frac{d}{\sqrt{nr_d(\Sigma^2)}} \min\{1, \frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}}\}, \frac{\lambda_{d+1}}{\lambda_d} \gtrsim \frac{n\sqrt{nr_d(\Sigma^2)}}{r_d(\Sigma)^2}.$$

To embody this condition, we set $d = 2, n = 150, p = 15000$ and $\rho = \frac{d}{\sqrt{nr_d(\Sigma^2)}} \min\{1, \frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}}\}$ such that $\frac{\lambda_{d+1}}{\lambda_d} = \frac{d}{\sqrt{nr_d(\Sigma^2)}} \min\{1, \frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}}\}, \frac{\lambda_{d+1}}{\lambda_d} = 0.0011549$ and $\frac{n\sqrt{nr_d(\Sigma^2)}}{r_d(\Sigma)^2} = 0.0010002$.

(iii) In the large TER regime, $\text{MSE}_{\text{out}}^*$ can be much smaller than MSE_{in}^* if

$$\frac{\lambda_{d+1}}{\lambda_d} \gtrsim \frac{d}{\sqrt{nr_d(\Sigma^2)}} \min\{1, \frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}}\}, \frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{n\sqrt{nr_d(\Sigma^2)}}{r_d(\Sigma)^2}, \frac{r_d(\Sigma)^2}{nr_d(\Sigma^2)} \gg 1.$$

To embody this condition, we set $d = 2$, $n = 300$, $p = 15000$ and $\rho = \frac{d}{\sqrt{nr_d(\Sigma^2)}} \min\{1, \frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}}\}$ such that $\frac{\lambda_{d+1}}{\lambda_d} = \frac{d}{\sqrt{nr_d(\Sigma^2)}} \min\{1, \frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}}\}$, $\frac{\lambda_{d+1}}{\lambda_d} = 0.0009429$, $\frac{n\sqrt{nr_d(\Sigma^2)}}{r_d(\Sigma)^2} = 0.0028290$ and $\frac{r_d(\Sigma)^2}{nr_d(\Sigma^2)} = 50$.

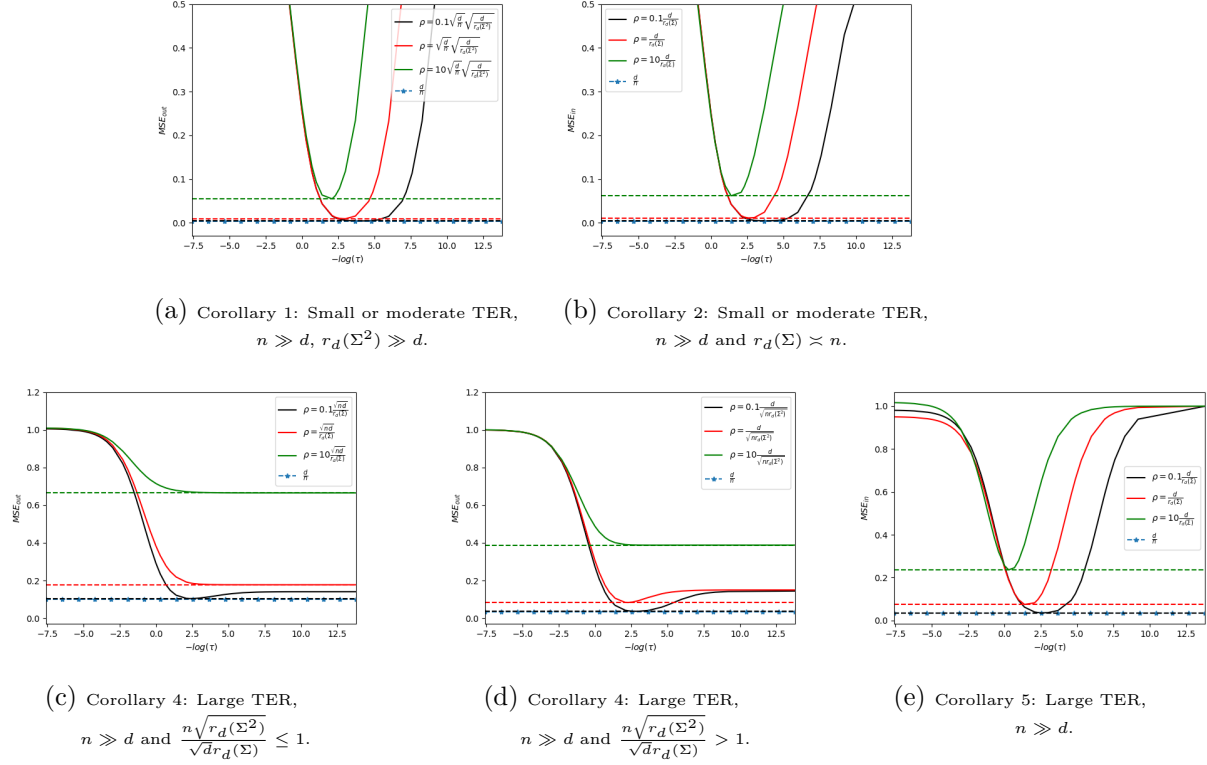


Figure 1: Study of conditions for $MSE = O(\frac{d}{n})$.

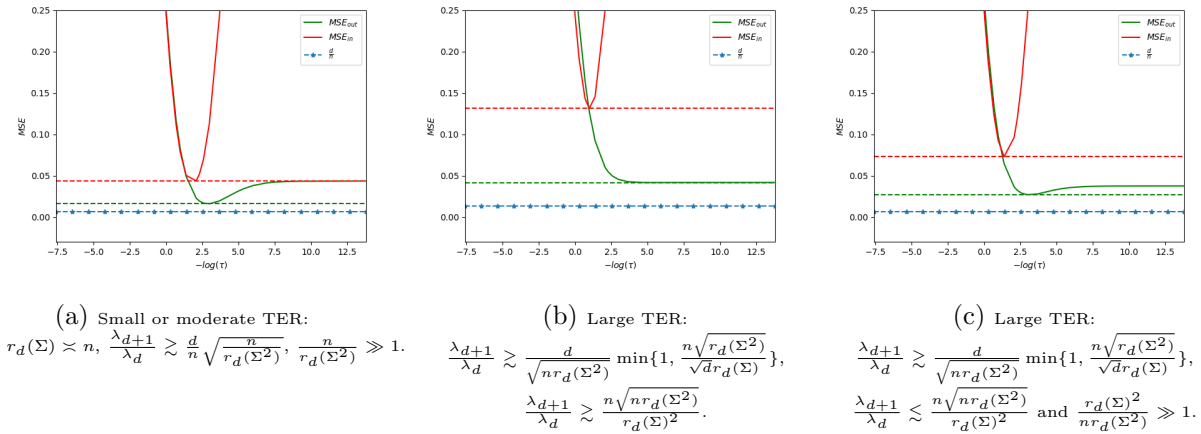


Figure 2: Study of conditions for MSE_{out}^* much smaller than MSE_{in}^* .

7.3 Results

The numerical results are summarized in Figure 1 and 2. From Figure 1, we see that when the ratio $\frac{\lambda_{d+1}}{\lambda_d}$ is related to the thresholds described in Section 7.2.1 by a pre-factor equal to 0.1 or 1, but not 10, the MSEs with near optimal choices of τ are close to $\frac{d}{n}$, which gives numerical support to our conditions for $\text{MSE} = O(\frac{d}{n})$ discussed in Corollary 1, 2, 4 and 5. From Figure 2, we see that $\text{MSE}_{\text{out}}^*$ is much smaller than MSE_{in}^* , each associated with the optimal choices of τ , in the settings described in Section 7.2.2 which embody the conditions from our theory for when $\text{MSE}_{\text{out}}^*$ can be much smaller than MSE_{in}^* in Section 3.1 and 3.2.

References

- Bartlett, P. L., Long, P. M., Lugosi, G., & Tsigler, A. (2020). Benign overfitting in linear regression. *Proceedings of the National Academy of Sciences*, 117(48), 30063–30070.
- Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849–15854.
- Bickel, P. J., Ritov, Y., & Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Annals of Statistics*, 37(4), 1705–1732.
- Bühlmann, P., & van de Geer, S. (2011). *Statistics for High-Dimensional Data*. Springer, Heidelberg.
- Bunea, F., Strimas-Mackey, S., & Wegkamp, M. (2022). Interpolating predictors in high-dimensional factor regression. *Journal of Machine Learning Research*, 23(10), 1–60.
- Chernozhukov, V., Chetverikov, D., Demirer, M., Duflo, E., Hansen, C., Newey, W., & Robins, J. (2018, 01). Double/debiased machine learning for treatment and structural parameters. *Econometrics Journal*, 21(1), C1–C68.
- Han, Q., & Shen, Y. (2023). Universality of regularized regression estimators in high dimensions. *Annals of Statistics*, 51(4), 1799–1823.
- Hastie, T., Montanari, A., Rosset, S., & Tibshirani, R. J. (2022). Surprises in high-dimensional ridgeless least squares interpolation. *Annals of Statistics*, 50(2), 949 – 986.
- Hsu, D., Kakade, S. M., & Zhang, T. (2014). Random design analysis of ridge regression. *Foundations of Computational Mathematics*, 14, 569–600.

- Knowles, A., & Yin, J. (2017). Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169.
- Marshall, A. W., Olkin, I., & Arnold, B. C. (2011). *Inequalities: Theory of Majorization and its Applications* (Second ed.). Springer.
- Tan, Z. (2020). Model-assisted inference for treatment effects using regularized calibrated estimation with high-dimensional data. *Annals of Statistics*, 48(2), 811 – 837.
- Tsigler, A., & Bartlett, P. L. (2023). Benign overfitting in ridge regression. *Journal of Machine Learning Research*, 24(123), 1–76.
- van de Geer, S., Bühlmann, P., Ritov, Y., & Dezeure, R. (2014). On asymptotically optimal confidence regions and tests for high-dimensional models. *Annals of Statistics*, 42(3), 1166–1202.
- Vershynin, R. (2012). *Introduction to the Non-asymptotic Analysis of Random Matrices*. Cambridge University Press.
- Zajkowski, K. (2020). Bounds on tail probabilities for quadratic forms in dependent sub-gaussian random variables. *Statistics and Probability Letters*, 167, 108898.
- Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *Communications of the ACM*, 64.
- Zhang, C.-H., & Zhang, S. S. (2014). Confidence intervals for low dimensional parameters in high dimensional linear models. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 76(1), 217–242.

Supplementary Material for “On Ridge Estimation in High-dimensional Rotationally Sparse Linear Regression”

I Definition of sub-gaussianity

A random variable $z \in \mathbb{R}$ is sub-gaussian if it has a finite sub-gaussian norm

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

The sub-gaussian norm of a random vector $Z \in \mathbb{R}^p$ is

$$\|Z\|_{\psi_2} = \sup_{s \neq 0} \left\| \frac{\langle s, Z \rangle}{\|s\|} \right\|_{\psi_2}.$$

II Proofs of main results

We provide proofs of Propositions 1–12 in Section 5. For convenience, we re-state the following notation from Section 6.

- $A = XX^T + n\tau I_n$, $A_d = X_{(d+1):p}X_{(d+1):p}^T + n\tau I_n$.
- $X_{1:d}$ denotes the the matrices comprised of the first d columns of X and $X_{(d+1):p}$ denotes the the matrices comprised of the last $p - d$ columns of X .
- $\hat{\Sigma}_{1:d} = \frac{X_{1:d}^T X_{1:d}}{n}$, $\hat{\Sigma}_{(d+1):p} = \frac{X_{(d+1):p}^T X_{(d+1):p}}{n}$ and $\hat{\Sigma}_{1:d,(d+1):p} = \frac{X_{1:d}^T X_{(d+1):p}}{n}$.
- $H_d = \Sigma_{1:d}^{-1/2} X_{1:d}^T$, $\hat{H}_d = \hat{\Sigma}_{1:d}^{-1/2} X_{1:d}^T$.
- $M_d = X_{(d+1):p} \Sigma_{(d+1):p} X_{(d+1):p}^T$, $\hat{M}_d = X_{(d+1):p} \hat{\Sigma}_{(d+1):p} X_{(d+1):p}^T$.
- $\mu_j(M)$ is the j -th largest eigenvalue of symmetric semi-positive definite matrix M .

II.1 Proof of the bounds for out-sample error

II.1.1 Algebraic bounds of the out-sample error

The bias and variance of the out-sample error can be decomposed as follows:

$$\begin{aligned} B_{\text{out}} &= \underbrace{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\Sigma_{1:d}}^2}_{B_{\text{out},1}} + \underbrace{\|\theta_{(d+1):p}^* - X_{(d+1):p}^T A^{-1} X \theta^*\|_{\Sigma_{(d+1):p}}^2}_{B_{\text{out},2}}, \\ V_{\text{out}} &= \underbrace{\sigma^2 \text{Tr}(A^{-1} X_{1:d} \Sigma_{1:d} X_{1:d}^T A^{-1})}_{V_{\text{out},1}} + \underbrace{\sigma^2 \text{Tr}(A^{-1} X_{(d+1):p} \Sigma_{(d+1):p} X_{(d+1):p}^T A^{-1})}_{V_{\text{out},2}}. \end{aligned}$$

The following algebraic bounds are the foundation of the upper bounds and lower bounds for the out-sample error. The algebraic upper bounds below are mainly inspired by Lemma 27 and 28 in Tsigler & Bartlett (2023). Moreover, we provide new algebraic lower bounds for the out-sample error.

Lemma S1 (Algebraic upper bounds of out-sample error). *Given invertible $\Sigma_{1:d}$, we have*

$$\begin{aligned} B_{\text{out}} &\leq 2\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n(H_d H_d^T)}{\mu_1(A_d)} \right)^{-2} + \frac{2\mu_1(H_d H_d^T)}{\mu_n^2(H_d H_d^T)} \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \\ &\quad + 3\left(\|M_d\| \frac{\mu_1(H_d H_d^T)}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\Sigma^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n(H_d H_d^T)}{\mu_1(A_d)} \right)^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \right. \\ &\quad \left. + \frac{\|M_d\|}{\mu_n(A)^2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \right), \\ V_{\text{out}} &\leq \sigma^2 \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \frac{\text{Tr}(H_d^T H_d)}{\mu_d(H_d H_d^T)^2} + \sigma^2 \frac{\text{Tr}(M_d)}{\mu_n(A_d)^2}. \end{aligned}$$

Proof.

Algebraic upper bound for the out-sample bias. From Section H.2 in Supplement of Tsigler & Bartlett (2023), we have

$$\begin{aligned} B_{\text{out},2} &\leq 3(\|X_{(d+1):p}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\Sigma_{(d+1):p}}^2 + \|X_{(d+1):p} A^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2) \\ &\leq 3(\|X_{(d+1):p}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\Sigma_{(d+1):p}}^2 + \frac{\|M_d\|}{\mu_n(A)^2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2). \end{aligned}$$

From Lemma S10 (ii), we have

$$\begin{aligned} &\|X_{(d+1):p}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\Sigma_{(d+1):p}}^2 \\ &= \|X_{(d+1):p}^T A_d^{-1} X_{1:d} (I_d + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} \theta_{1:d}^*\|_{\Sigma_{(d+1):p}}^2 \\ &= \|X_{(d+1):p}^T A_d^{-1} X_{1:d} \Sigma_{1:d}^{-1/2} (\Sigma_{1:d}^{-1} + \Sigma_{1:d}^{-1/2} X_{1:d}^T A_d^{-1} X_{1:d} \Sigma_{1:d}^{-1/2})^{-1} \Sigma_{1:d}^{-1/2} \theta_{1:d}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\leq \|M_d\| \frac{\mu_1(H_d H_d^T)}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\Sigma^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n(H_d H_d^T)}{\mu_1(A_d)} \right)^{-2}. \end{aligned}$$

Hence we have

$$\begin{aligned} B_{\text{out},2} &\leq 3\left(\|M_d\| \frac{\mu_1(H_d H_d^T)}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\Sigma^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n(H_d H_d^T)}{\mu_1(A_d)} \right)^{-2} \right. \\ &\quad \left. + \frac{\|M_d\|}{\mu_n(A)^2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \right). \end{aligned} \tag{S1}$$

It remains to give an upper bound of $B_{\text{out},1}$. From Lemma S10 (i), we have

$$\hat{\theta}(\tau, X\theta^*)_{1:d} + X_{1:d}^T A_d^{-1} X_{1:d} \hat{\theta}(\tau, X\theta^*)_{1:d} = X_{1:d}^T A_d^{-1} X\theta^*.$$

Denote $\zeta_{1:d} = \hat{\theta}(\tau, X\theta^*)_{1:d} - \theta_{1:d}^*$. Then

$$H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \Sigma_{1:d}^{-1/2} \theta_{1:d}^* = \Sigma_{1:d}^{-1/2} \zeta_{1:d} + \Sigma_{1:d}^{-1/2} X_{1:d}^T A_d^{-1} X_{1:d} \zeta_{1:d}$$

$$= (\Sigma_{1:d}^{-1} + H_d A_d^{-1} H_d^T) \Sigma_{1:d}^{1/2} \zeta_{1:d}. \quad (\text{S2})$$

By standard manipulations,

$$\begin{aligned} \|(\Sigma_{1:d}^{-1} + H_d A_d^{-1} H_d^T) \Sigma_{1:d}^{1/2} \zeta_{1:d}\|^2 &\geq \mu_n(\Sigma_{1:d}^{-1} + H_d A_d^{-1} H_d^T)^2 \|\Sigma_{1:d}^{1/2} \zeta_{1:d}\|^2 \\ &\geq B_{\text{out},1} \left(\frac{1}{\lambda_1} + \frac{\mu_n(H_d H_d^T)}{\mu_1(A_d)} \right)^2, \end{aligned} \quad (\text{S3})$$

$$\|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \Sigma_{1:d}^{-1/2} \theta_{1:d}^*\|^2 \leq (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} + \|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|)^2. \quad (\text{S4})$$

From (S2), (S3) and (S4), we have

$$B_{\text{out},1} \left(\frac{1}{\lambda_1} + \frac{\mu_n(H_d H_d^T)}{\mu_1(A_d)} \right)^2 \leq (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} + \|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|)^2.$$

That is, we have

$$\begin{aligned} B_{\text{out},1} &\leq (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} + \|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|)^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n^2(H_d H_d^T)}{\mu_1^2(A_d)} \right)^{-2} \\ &\leq 2\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n^2(H_d H_d^T)}{\mu_1^2(A_d)} \right)^{-2} + 2\|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \frac{\mu_1^2(A_d)}{\mu_n^2(H_d H_d^T)} \\ &\leq 2\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{\mu_n^2(H_d H_d^T)}{\mu_1^2(A_d)} \right)^{-2} + \frac{2\mu_1(H_d H_d^T)}{\mu_n^2(H_d H_d^T)} \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2. \end{aligned} \quad (\text{S5})$$

Combining (S1) and (S5) gives the upper bound of out-sample bias.

Algebraic upper bound for the out-sample variance. From Lemma 27 in Tsigler & Bartlett (2023), we have

$$V_{\text{out},1} \leq \sigma^2 \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \frac{\text{Tr}(H_d^T H_d)}{\mu_d(H_d H_d^T)^2}.$$

To upper bound $V_{\text{out},2}$, we have

$$\begin{aligned} V_{\text{out},2} &= \sigma^2 \text{Tr}(A^{-1} X_{(d+1):p} \Sigma_{(d+1):p} X_{(d+1):p}^T A^{-1}) \\ &\leq \sigma^2 \frac{\text{Tr}(M_d)}{\mu_n(A)^2} \\ &\leq \sigma^2 \frac{\text{Tr}(M_d)}{\mu_n(A_d)^2}. \end{aligned}$$

Combining the preceding two displays gives the upper bound of out-sample variance. $\square$

Lemma S2 (Algebraic lower bounds of out-sample error). *Given invertible $\Sigma_{1:d}$, we have*

$$V_{\text{out}} \geq \sigma^2 \frac{d\mu_d(H_d H_d^T)}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^{-2} + \max\{0, \sigma^2 \frac{\text{Tr}(M_d) - d\mu_1(M_d)}{\mu_1(A_d)^2}\}.$$

Further if $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{\mu_1(H_d H_d)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)}$, then

$$B_{\text{out}} \geq \left(1 - \frac{\mu_1(H_d H_d^T)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^{-2}. \quad (\text{S6})$$

Proof.

Algebraic lower bound for the out-sample variance: From Lemma S10 (i), we have

$$X_{1:d}^T A_d^{-1} \epsilon = \hat{\theta}_{1:d}(\tau, \epsilon) + X_{1:d}^T A_d^{-1} X_{1:d} \hat{\theta}_{1:d}(\tau, \epsilon).$$

Multiplying the two sides with $\Sigma^{-1/2}$, we have

$$H_d A_d^{-1} \epsilon = (\Sigma_{1:d}^{-1} + H_d A_d^{-1} \Sigma_{1:d}^{-1/2}) \Sigma_{1:d}^{1/2} \hat{\theta}(\tau, \epsilon),$$

and hence

$$\begin{aligned} \|H_d A_d^{-1} \epsilon\|^2 &= \|(\Sigma_{1:d}^{-1} + H_d A_d^{-1} \Sigma_{1:d}^{-1/2}) \Sigma_{1:d}^{1/2} \hat{\theta}(\tau, \epsilon)\|^2 \\ &\leq \mu_1(\Sigma_{1:d}^{-1} + H_d A_d^{-1} \Sigma_{1:d}^{-1/2})^2 \|\Sigma_{1:d}^{1/2} \hat{\theta}(\tau, \epsilon)\|^2. \end{aligned}$$

Taking the expectations of the two sides with respect to ϵ yields

$$\begin{aligned} \mu_1(\Sigma_{1:d}^{-1} + H_d A_d^{-1} \Sigma_{1:d}^{-1/2})^2 V_{\text{out},1} &\geq E_\epsilon[\|H_d A_d^{-1} \epsilon\|^2] \\ &= \sigma^2 \text{Tr}(A_d^{-1} H_d H_d^T A_d^{-1}). \end{aligned}$$

By simple manipulations, we have

$$\begin{aligned} \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)}\right)^2 V_{\text{out},1} &\geq \sigma^2 \text{Tr}(H_d A_d^{-2} H_d^T) \\ &\geq \sigma^2 \frac{\text{Tr}(H_d H_d^T)}{\mu_1(A_d)^2} \\ &\geq \sigma^2 \frac{d \mu_d(H_d H_d^T)}{\mu_1(A_d)^2}, \end{aligned}$$

and hence

$$V_{\text{out},1} \geq \sigma^2 \frac{d \mu_d(H_d H_d^T)}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)}\right)^{-2}. \quad (\text{S7})$$

To lower bound $V_{\text{out},2}$, we have

$$\begin{aligned} V_{\text{out},2} &= \sigma^2 \text{Tr}(A^{-1} M_d A^{-1}) \\ &\geq \sum_{i=1}^{n-d} \mu_i(A^{-2}) \mu_{n-i+1}(M_d) \\ &\geq \frac{\sum_{i=1}^{n-d} \mu_{n-i+1}(M_d)}{\mu_1(A_d)^2} \\ &\geq \frac{\text{Tr}(M_d) - d \mu_1(M_d)}{\mu_1(A_d)^2}. \end{aligned}$$

The first inequality above is from Lemma S13. For the second inequality, because the rank of $A - A_d$ is at most d ,

$$\mu_i(A - A_d) = 0 \quad \forall i \geq d + 1.$$

From Weyl's inequality (Lemma S12), we have

$$\begin{aligned}
\mu_i(A) - \mu_1(A_d) &\leq \mu_i(A - A_d) = 0 \quad \forall i \geq d+1 \\
\implies \mu_i(A) &\leq \mu_1(A_d) \quad \forall i \geq d+1 \\
\implies \frac{1}{\mu_1(A_d)} &\leq \frac{1}{\mu_i(A)} \quad \forall i \geq d+1 \\
\implies \frac{1}{\mu_1(A_d)} &\leq \mu_i(A^{-1}) \quad \forall i \leq n-d.
\end{aligned}$$

By requiring $V_{\text{out},2} \geq 0$, we have

$$V_{\text{out},2} \geq \max\{0, \frac{\text{Tr}(M_d) - d\mu_1(M_d)}{\mu_1(A_d)^2}\}. \quad (\text{S8})$$

Combining (S7) and (S8) gives the lower bound of out-sample variance.

Algebraic lower bound for the out-sample bias. The norms of the two sides of (S2) can be bounded as follows:

$$\begin{aligned}
\|(\Sigma_{1:d}^{-1} + H_d A_d^{-1} H_d^T) \Sigma_{1:d}^{1/2} \zeta_{1:d}\|^2 &\leq \mu_1(\Sigma_{1:d}^{-1} + H_d A_d^{-1} H_d^T)^2 \|\Sigma_{1:d}^{1/2} \zeta_{1:d}\|^2 \\
&\leq B_{\text{out},1} \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^2,
\end{aligned} \quad (\text{S9})$$

$$\|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \Sigma_{1:d}^{-1/2} \theta_{1:d}^*\|^2 \geq (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} - \|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|)^2. \quad (\text{S10})$$

From (S2), (S9) and (S10), we have

$$B_{\text{out},1} \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^2 \geq (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} - \|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|)^2,$$

and hence

$$\begin{aligned}
B_{\text{out}} &\geq B_{\text{out},1} \\
&\geq (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} - \|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|)^2 \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^{-2}.
\end{aligned}$$

With $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{\mu_1(H_d H_d^T)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)}$, we have

$$\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{\mu_1(H_d H_d^T)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)} \geq \|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|,$$

and hence

$$\begin{aligned}
B_{\text{out}} &\geq (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} - \|H_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|)^2 \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^{-2} \\
&\geq \left(1 - \frac{\mu_1(H_d H_d^T)^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{\mu_1(H_d H_d^T)}{\mu_n(A_d)} \right)^{-2}.
\end{aligned}$$

□

II.1.2 Intermediate bounds of the out-sample error

We give the intermediate bounds of out-sample error under the event that some random quantities in the algebraic bounds above are controlled. In the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_4$ for $0 < \nu < \frac{1}{2}$ defined in Lemma S6, substituting the bounds of $\mu_1(H_d H_d^T)$, $\mu_d(H_d H_d^T)$, $\|X_{(d+1):p} \theta_{(d+1):p}^*\|$ and $\|M_d\|$ into the algebraic upper bound of B_{out} yields

$$\begin{aligned} B_{\text{out}} &\leq \frac{2}{(1-\nu-\eta_1)^4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} + \frac{\text{Poly}_2(\sigma_x)(1+\nu+\eta_1)^2}{(1-\nu-\eta_1)^4} \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\quad + \frac{\text{Poly}_2(\sigma_x)(1+\nu+\eta_1)^2}{(1-\nu-\eta_1)^4} \frac{(n^2 \lambda_{d+1}^2 + n \sum_{j>d} \lambda_j^2)}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\Sigma^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} \\ &\quad + (3 + \text{Poly}_4(\sigma_x)) \frac{(n^2 \lambda_{d+1}^2 + n \sum_{j>d} \lambda_j^2)}{\mu_n(A_d)^2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2. \end{aligned} \quad (\text{S11})$$

In the event $\Omega_1(\nu) \cap \Omega_4$ for $0 < \nu < \frac{1}{2}$ defined in Lemma S6 and with

$$\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{(1+\nu+\eta_1)(1+\sigma_x^2)^{1/2} n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n(A_d)},$$

substituting the bounds of $\mu_1(H_d H_d^T)$ and $\|X_{(d+1):p} \theta_{(d+1):p}^*\|$ into the algebraic lower bound of B_{out} yields

$$B_{\text{out}} \geq \frac{1}{(1+\nu+\eta_1)^2} \left(1 - \frac{(1+\nu+\eta_1)(1+\sigma_x^2)^{1/2} n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2}. \quad (\text{S12})$$

In the event $\Omega_1(\nu) \cap \Omega_{31} \cap \Omega_{32}$ for $\nu < \frac{1}{2}$ defined in Lemma S6, substituting the bounds of $\mu_d(H_d H_d^T)$, $\text{Tr}(H_d^T H_d)$ and $\text{Tr}(M_d)$ into the algebraic upper bound of V_{out} yields

$$V_{\text{out}} \leq \frac{\text{Poly}_2(\sigma_x)}{(1-\nu-\eta_1)^4} \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \sigma^2 \frac{d}{n} + \text{Poly}_2(\sigma_x) \sigma^2 \frac{n \sum_{j>d} \lambda_j^2}{\mu_n(A_d)^2}. \quad (\text{S13})$$

In the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu)$ for $\nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6, substituting the bounds of $\mu_1(H_d H_d^T)$, $\mu_d(H_d H_d^T)$, $\text{Tr}(M_d)$ and $\mu_1(M_d)$ into the algebraic lower bound of V_{out} yields

$$V_{\text{out}} \geq \frac{(1-\nu-\eta_1)^2}{(1+\nu+\eta_1)^4} \sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2} + \max\{0, \frac{n \sum_{j>d} \lambda_j^2}{\mu_1(A_d)^2} (1-\nu - C_0 \sigma_x^2 (\frac{2d}{r_d(\Sigma^2)} + \frac{d}{n}))\}. \quad (\text{S14})$$

II.1.3 Final upper bounds of out-sample bias and variance

We give the final upper bounds of out-sample bias and variance as stated in Proposition 1 and 7. From the intermediate bounds in Section II.1.2, we derive the final bounds by further controlling $\mu_1(A_d)$ and $\mu_n(A_d)$ (from Lemma S7) in the intermediate bounds and incorporating Assumption

4 to control the terms related to $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2$. We first discuss the small or moderate TER regime and then the large TER regime.

(i) Small or moderate TER

From Lemma S7(i), (S11) and (S13), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{31} \cap \Omega_{32} \cap \Omega_4 \cap \Omega_5$ for $\nu < \frac{1}{2}$ defined in Lemma S6 and Assumption 2, substituting the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ in (S47)–(S48) into (S11) and (S13), we have for $\tau \geq \lambda_{d+1}$,

$$\begin{aligned} B_{\text{out}} &\leq \frac{\text{Poly}_4(\sigma_x)(1+C_1)^2}{(1-\nu-\eta_1)^4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau}\right)^{-2} + \frac{\text{Poly}_6(\sigma_x)(1+\nu+\eta_1)^2}{(1-\nu-\eta_1)^4} (1+C_1)^2 \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\quad + \frac{\text{Poly}_6(\sigma_x)(1+C_1)^3(1+\nu+\eta_1)^2}{(1-\nu-\eta_1)^4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau}\right)^{-2} + \text{Poly}_4(\sigma_x)(1+C_1) \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2, \\ V_{\text{out}} &\leq \frac{\text{Poly}_6(\sigma_x)(1+C_1)^2}{(1-\nu-\eta_1)^4} \left(\sigma^2 \frac{d}{n} + \sigma^2 \frac{\sum_{j>d} \lambda_j^2}{n\tau^2}\right). \end{aligned}$$

Hence, we obtain the upper bounds

$$\begin{aligned} B_{\text{out}} &\leq \frac{(1+C_1)^3(1+\nu+\eta_1)^2 \text{Poly}_6(\sigma_x)}{(1-\nu-\eta_1)^4} \left(\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau}\right)^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2\right), \\ V_{\text{out}} &\leq \frac{\text{Poly}_6(\sigma_x)(1+C_1)^2}{(1-\nu-\eta_1)^4} \left(\sigma^2 \frac{d}{n} + \sigma^2 \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n}\right). \end{aligned}$$

With Assumption 4(i), we have for $\tau \geq \lambda_{d+1}$,

$$\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \leq \frac{\delta_1}{4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_{d+1}^2 \leq \delta_1 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau}\right)^{-2}. \quad (\text{S15})$$

Further with (S15), we have for $\tau \geq \lambda_{d+1}$,

$$B_{\text{out}} \leq \frac{(1+C_1)^3(1+\nu+\eta_1)^2 \text{Poly}_6(\sigma_x)}{(1-\nu-\eta_1)^4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau}\right)^{-2}.$$

This gives Proposition 1.

(ii) Large TER

From Lemma S7(ii), (S11) and (S13), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{31} \cap \Omega_{32} \cap \Omega_4 \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{2}$ defined in Lemma S6 and Assumption 3, substituting the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ in (S49)–(S50) into (S11) and (S13), for $\tau \geq 0$,

$$\begin{aligned} B_{\text{out}} &\leq \frac{(1+\nu+\eta_2)^2}{(1-\nu-\eta_1)^4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}}\right)^{-2} \\ &\quad + \frac{\text{Poly}_2(\sigma_x)(1+\nu+\eta_1)^2}{(1-\nu-\eta_1)^4} \frac{(1+\nu+\eta_2)^2}{(1-\nu-\eta_2)^2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\quad + \frac{\text{Poly}_2(\sigma_x)(1+\nu+\eta_1)^2(1+\nu+\eta_2)^2}{(1-\nu-\eta_1)^4(1-\nu-\eta_2)^2} \|\theta_{1:d}^*\|_{\Sigma^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}}\right)^{-2} \\ &\quad + \frac{\text{Poly}_4(\sigma_x)}{(1-\nu-\eta_2)^2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2, \end{aligned}$$

$$V_{\text{out}} \leq \frac{\text{Poly}_2(\sigma_x)(1+\nu+\eta_2)^2}{(1-\nu-\eta_1)^4(1-\nu-\eta_2)^2} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} \right).$$

Hence, we obtain the upper bounds

$$\begin{aligned} B_{\text{out}} &\leq \frac{(1+\nu+\eta_1)^2(1+\nu+\eta_2)^2 \text{Poly}_4(\sigma_x)}{(1-\nu-\eta_1)^4(1-\nu-\eta_2)^2} (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau + \frac{\sum_{j>d} \lambda_j}{n}})^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2), \\ V_{\text{out}} &\leq \frac{(1+\nu+\eta_2)^2 \text{Poly}_2(\sigma_x)}{(1-\nu-\eta_1)^4(1-\nu-\eta_2)^2} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} \right). \end{aligned}$$

With Assumption 4(ii), we have for $\tau \geq 0$,

$$\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \leq \frac{\delta_2}{4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\lambda_{d+1} \frac{r_d(\Sigma)}{n}} \right)^{-2} \leq \frac{\delta_2}{4} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \right)^{-2}. \quad (\text{S16})$$

Further with (S16), we have for $\tau \geq 0$,

$$B_{\text{out}} \leq \frac{(1+\nu+\eta_1)^2(1+\nu+\eta_2)^2 \text{Poly}_4(\sigma_x)}{(1-\nu-\eta_1)^4(1-\nu-\eta_2)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \right)^{-2}.$$

This gives Proposition 7.

II.1.4 Final lower bounds of out-sample bias and variance

We give the final lower bounds of out-sample bias and variance. We first discuss the small or moderate TER regime and then the large TER regime.

(i) Small or moderate TER

Lower bound of out-sample bias. For $\tau \geq \lambda_{d+1}$, we have $\mu_n(A_d) \geq n\tau \geq n\lambda_{d+1}$. Then from Assumption 4(i), we have for $0 < \nu < \frac{1}{2}$ and $\tau \geq \lambda_{d+1}$,

$$\begin{aligned} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} &\geq \frac{2(1+\sigma_x^2)^{1/2} n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{n\lambda_{d+1}} \\ &\geq \frac{(1+\nu+\eta_1)(1+\sigma_x^2)^{1/2} n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n(A_d)}. \end{aligned} \quad (\text{S17})$$

From (S12), in the event $\Omega_1(\nu) \cap \Omega_4$ for $\nu < \frac{1}{2}$ defined in Lemma S6 and Assumption 4(i), substituting the lower bound of $\mu_n(A_d)$ in (S48) into (S12), we have for $\tau \geq \lambda_{d+1}$,

$$B_{\text{out}} \geq \frac{1}{(1+\nu+\eta_1)^2} \left(1 - \frac{(1+\nu+\eta_1)(1+\sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\tau \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\tau} \right)^{-2}.$$

With Assumption 4(i), we also have for $\tau \geq \lambda_{d+1}$,

$$\frac{(1+\nu+\eta_1)(1+\sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\tau \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \leq \frac{2(1+\sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\lambda_{d+1} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \leq \sqrt{\delta_1}. \quad (\text{S18})$$

By applying (S18), in the event $\Omega_1(\nu) \cap \Omega_4$ for $\nu < \frac{1}{2}$ defined in Lemma S6 and Assumption 4(i), we have for $\tau \geq \lambda_{d+1}$,

$$B_{\text{out}} \geq \frac{(1 - \sqrt{\delta_1})^2}{(1 + \nu + \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\tau}\right)^{-2}.$$

This gives the results of Proposition 2.

Lower bound of out-sample variance. To deduce the final lower bound of V_{out} , our strategy is as follows. We first derive a lower bound for the first term in the intermediate lower bound (S14). Then we discuss two complementary cases. The first case is that $\frac{r_d(\Sigma^2)}{d}$ is upper bounded. The second case is that $\frac{r_d(\Sigma^2)}{d}$ is large enough such that $1 - \nu - C_0\sigma_x^2(\frac{2d}{r_d(\Sigma^2)} + \frac{d}{n}) > 0$. Lastly, we show that in these two cases, V_{out} satisfies lower bounds of the same order, which gives the final lower bound for V_{out} .

As preparation, we derive an equivalence relationship, which is useful in the following analysis. Given Assumption 1, we have

$$\begin{aligned} \frac{2C_0\sigma_x^2d}{r_d(\Sigma^2)} + \frac{C_0\sigma_x^2d}{n} \geq \frac{1}{2} &\iff \frac{2C_0\sigma_x^2d}{r_d(\Sigma^2)} \geq \frac{1}{2} - \frac{C_0\sigma_x^2d}{n} \\ &\iff \frac{2C_0\sigma_x^2d}{r_d(\Sigma^2)} \geq \frac{1}{2} - \eta_1 \quad (\text{from Assumption 1}) \\ &\iff \frac{2C_0\sigma_x^2d}{\frac{1}{2} - \eta_1} \geq \frac{\sum_{j>d} \lambda_j^2}{\lambda_{d+1}^2}. \end{aligned} \quad (\text{S19})$$

Now, we are ready to give the lower bound of V_{out} . We first consider the first term of the right-hand side in (S14). From Lemma S7(i) and (S14), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_5$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6, substituting the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ in (S47)–(S48) into the first term in (S14), we have for $\tau \geq \lambda_{d+1}$,

$$\frac{(1 - \nu - \eta_1)^2}{(1 + \nu + \eta_1)^4} \sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)}\right)^{-2} \geq \frac{(1 - \nu - \eta_1)^2}{(1 + C_1)^2(1 + \nu + \eta_1)^4 \text{Poly}_4(\sigma_x)} \sigma^2 \frac{d}{n\tau^2} \left(\frac{1}{\lambda_d} + \frac{1}{\tau}\right)^{-2}.$$

Then in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_5$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6 and Assumption 2, we have for $\lambda_{d+1} \leq \tau \leq \lambda_d$,

$$\frac{(1 - \nu - \eta_1)^2}{(1 + \nu + \eta_1)^4} \sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)}\right)^{-2} \geq \frac{(1 - \nu - \eta_1)^2}{(1 + C_1)^2(1 + \nu + \eta_1)^4 \text{Poly}_4(\sigma_x)} \sigma^2 \frac{d}{n}. \quad (\text{S20})$$

Then we discuss two complementary cases. The first cases is that $\frac{2C_0\sigma_x^2d}{r_d(\Sigma^2)} + \frac{C_0\sigma_x^2d}{n} \geq \frac{1}{2}$. From the equivalence in (S19) and with Assumption 1, for $\tau \geq \lambda_{d+1}$,

$$\frac{2C_0\sigma_x^2d}{\frac{1}{2} - \eta_1} \geq \frac{\sum_{j>d} \lambda_j^2}{\lambda_{d+1}^2} \geq \frac{\sum_{j>d} \lambda_j^2}{\tau^2},$$

which implies that

$$\frac{d}{n} \geq \frac{\frac{1}{2} - \eta_1}{2C_0\sigma_x^2} \frac{\sum_{j>d} \lambda_j^2}{n\tau^2}$$

$$\iff \frac{d}{n} \geq \frac{\frac{\frac{1}{2}-\eta_1}{2C_0\sigma_x^2}}{1 + \frac{\frac{1}{2}-\eta_1}{2C_0\sigma_x^2}} \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n\tau^2} \right).$$

Then from (S14) and (S20), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_5$ for $\nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6, we have for $\lambda_{d+1} \leq \tau \leq \lambda_d$,

$$\begin{aligned} V_{\text{out}} &\geq \frac{(1-\nu-\eta_1)^2}{(1+C_1)^2(1+\nu+\eta_1)^4 \text{Poly}_4(\sigma_x)} \sigma^2 \frac{d}{n} \\ &\geq \frac{(1-\nu-\eta_1)^2}{(1+C_1)^2(1+\nu+\eta_1)^4 \text{Poly}_4(\sigma_x)} \frac{\frac{\frac{1}{2}-\eta_1}{2C_0\sigma_x^2}}{1 + \frac{\frac{1}{2}-\eta_1}{2C_0\sigma_x^2}} \sigma^2 \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n\tau^2} \right). \end{aligned}$$

The second case is that $\frac{2C_0\sigma_x^2 d}{r_d(\Sigma^2)} + \frac{C_0\sigma_x^2 d}{n} < \frac{1}{2}$. From Lemma S7(i) and (S14), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_5$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6 and Assumption 2, for $\tau \geq \lambda_{d+1}$, the second term in (S14) satisfies

$$\max\{0, \frac{n \sum_{j>d} \lambda_j^2}{\mu_1(A_d)^2} (1-\nu - C_0\sigma_x^2 (\frac{2d}{r_d(\Sigma^2)} + \frac{d}{n}))\} \geq \frac{1}{(1+C_1)^2 \text{Poly}_4(\sigma_x)} \frac{\sum_{j>d} \lambda_j^2}{n\tau^2} (\frac{1}{2} - \nu). \quad (\text{S21})$$

Then from (S14), (S20) and (S21), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_5$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6 and Assumption 2, we have for $\tau \geq \lambda_{d+1}$,

$$V_{\text{out}} \geq \frac{(1-\nu-\eta_1)^2 (\frac{1}{2} - \nu)}{(1+C_1)^2(1+\nu+\eta_1)^4 \text{Poly}_4(\sigma_x)} \sigma^2 \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n\tau^2} \right).$$

In conclusion, in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_5$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6 and Assumption 2, we have for $\lambda_{d+1} \leq \tau \leq \lambda_d$,

$$V_{\text{out}} \geq \frac{(1-\nu-\eta_1)^2 (\frac{1}{2} - \nu)}{(1+C_1)^2(1+\nu+\eta_1)^4 \text{Poly}_4(\sigma_x)} \frac{\frac{\frac{1}{2}-\eta_1}{2C_0\sigma_x^2}}{1 + \frac{\frac{1}{2}-\eta_1}{2C_0\sigma_x^2}} \sigma^2 \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n\tau^2} \right).$$

or equivalently

$$V_{\text{out}} \geq \frac{(1-\nu-\eta_1)^2 (\frac{1}{2} - \nu)}{(1+C_1)^2(1+\nu+\eta_1)^4 \text{Poly}_4(\sigma_x)} \frac{\frac{\frac{1}{2}-\eta_1}{2C_0\sigma_x^2}}{1 + \frac{\frac{1}{2}-\eta_1}{2C_0\sigma_x^2}} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \right).$$

The result for $\tau \leq \lambda_{d+1}$ follows from the monotonicity of variance in Lemma S11. This gives Proposition 3.

(ii) Large TER

Lower bound of out-sample bias. In the event $\Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$ defined in Lemma S6 and Assumption 3 and 4(ii), incorporating the bound of $\mu_n(A_d)$, we have for $\tau \geq 0$,

$$\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{2(1+\sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\frac{1}{4}(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)}$$

$$\geq \frac{(1 + \nu + \eta_1)(1 + \sigma_x^2)^{1/2} n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n(A_d)}. \quad (\text{S22})$$

Then from (S12) and Lemma S7(ii), in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$ defined in Lemma S6 and Assumption 3 and 4(ii), substituting the lower bounds of $\mu_n(A_d)$ in (S50) into (S12), we have for $\tau \geq 0$,

$$B_{\text{out}} \geq \frac{(1 - \nu - \eta_2)^2}{(1 + \nu + \eta_1)^2} \left(1 - \frac{(1 + \nu + \eta_1)(1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{(1 - \nu - \eta_2)(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)} \right)^{-2}.$$

With Assumption 4(ii), we have for $\tau \geq 0$,

$$\frac{(1 + \nu + \eta_1)(1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{(1 - \nu - \eta_2)(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \leq \frac{2(1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\frac{1}{4} \lambda_{d+1} \frac{r_d(\Sigma)}{n} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \leq \sqrt{\delta_2}. \quad (\text{S23})$$

Moreover, by applying (S23), in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$ defined in Lemma S6 and Assumption 4(ii), we have for $\tau \geq 0$,

$$B_{\text{out}} \geq \frac{(1 - \nu - \eta_2)^2}{(1 + \nu + \eta_1)^2} (1 - \sqrt{\delta_2})^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})} \right)^{-2}.$$

This gives Proposition 8.

Lower bound of out-sample variance. Our strategy for deriving the lower bound of out-sample variance in large TER regime is similar to that in small or moderate TER regime. With Assumption 3, we have

$$\frac{1}{2} > C_0 \sigma_x^2 \sqrt{\frac{4n^2}{r_d(\Sigma)^2} + \frac{n}{r_d(\Sigma)}} > C_0 \sigma_x^2 \frac{2n}{r_d(\Sigma)},$$

and hence

$$\frac{1}{4C_0 \sigma_x^2} \frac{\sum_{j>d} \lambda_j}{n} > \lambda_{d+1}. \quad (\text{S24})$$

Moreover, with the (S19) under Assumption 1 and (S24) under Assumption 3, we have for $\tau \geq 0$,

$$\begin{aligned} & \frac{2C_0 \sigma_x^2 d}{r_d(\Sigma^2)} + \frac{C_0 \sigma_x^2 d}{n} \geq \frac{1}{2} \\ \implies & \frac{2C_0 \sigma_x^2 d}{\frac{1}{2} - \eta_1} \geq \frac{\sum_{j>d} \lambda_j^2}{\lambda_{d+1}^2} \\ \implies & \frac{d}{8C_0 \sigma_x^2 (\frac{1}{2} - \eta_1)} \geq \frac{\sum_{j>d} \lambda_j^2}{(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)^2}. \end{aligned} \quad (\text{S25})$$

We first consider the first term of the right-hand side in (S14). From Lemma S7(ii) and (S14), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6, substituting the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ in (S49)–(S50) into the first term in (S14), we have for $\tau \geq 0$,

$$\frac{(1 - \nu - \eta_1)^2}{(1 + \nu + \eta_1)^4} \sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2}$$

$$\geq \frac{(1-\nu-\eta_2)^2(1-\nu-\eta_1)^2}{(1+\nu+\eta_2)^2(1+\nu+\eta_1)^4} \sigma^2 \frac{d}{n(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)^2} \left(\frac{1}{\lambda_d} + \frac{1}{(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)} \right)^{-2}.$$

Then in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6, we have for $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_d$,

$$\frac{(1-\nu-\eta_1)^2}{(1+\nu+\eta_1)^4} \sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2} \geq \frac{(1-\nu-\eta_2)^2(1-\nu-\eta_1)^2}{4(1+\nu+\eta_2)^2(1+\nu+\eta_1)^4} \sigma^2 \frac{d}{n}. \quad (\text{S26})$$

Then we discuss two complementary cases. The first cases is that $\frac{2C_0\sigma_x^2d}{r_d(\Sigma^2)} + \frac{C_0\sigma_x^2d}{n} \geq \frac{1}{2}$. From (S25), with Assumption 1 and 3, we have for $\tau \geq 0$,

$$\begin{aligned} \frac{d}{8C_0\sigma_x^2(\frac{1}{2}-\eta_1)} &\geq \frac{\sum_{j>d} \lambda_j^2}{(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)^2} \\ \Rightarrow \frac{d}{n} &\geq \frac{8C_0\sigma_x^2(\frac{1}{2}-\eta_1)}{1+8C_0\sigma_x^2(\frac{1}{2}-\eta_1)} \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)^2} \right). \end{aligned}$$

Combining with (S26), we have for $\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau \leq \lambda_d$,

$$\begin{aligned} V_{\text{out}} &\geq \frac{(1-\nu-\eta_1)^2}{(1+\nu+\eta_1)^4} \sigma^2 \frac{nd}{\mu_1(A_d)^2} \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2} \\ &\geq \frac{(1-\nu-\eta_1)^2(1-\nu-\eta_1)^2}{4(1+\nu+\eta_1)^4(1+\nu+\eta_2)^2} \sigma^2 \frac{d}{n} \\ &\geq \frac{(1-\nu-\eta_1)^2(1-\nu-\eta_1)^2}{4(1+\nu+\eta_1)^4(1+\nu+\eta_2)^2} \frac{8C_0\sigma_x^2(\frac{1}{2}-\eta_1)}{1+8C_0\sigma_x^2(\frac{1}{2}-\eta_1)} \sigma^2 \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \right). \end{aligned}$$

The second case is that $\frac{2C_0\sigma_x^2d}{r_d(\Sigma^2)} + \frac{C_0\sigma_x^2d}{n} < \frac{1}{2}$. From (S14) and Lemma S7(ii), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6, we have for $\tau \geq 0$,

$$\max\{0, \frac{n \sum_{j>d} \lambda_j^2}{\mu_1(A_d)^2} (1-\nu - C_0\sigma_x^2(\frac{2d}{r_d(\Sigma^2)} + \frac{d}{n}))\} \geq \frac{\sum_{j>d} \lambda_j^2}{n(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)^2} \left(\frac{1}{2} - \nu \right). \quad (\text{S27})$$

From (S14), (S26) and (S27), in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_6(\nu)$ for $\nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6, we have for $\tau \geq 0$,

$$V_{\text{out}} \geq \frac{(1-\nu-\eta_1)^2(1-\nu-\eta_1)^2(\frac{1}{2}-\nu)}{4(1+\nu+\eta_1)^4(1+\nu+\eta_2)^2} \sigma^2 \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \right).$$

In conclusion, in the event $\Omega_1(\nu) \cap \Omega_2 \cap \Omega_{33}(\nu) \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6 and Assumption 1,3, we have for $\tau \geq 0$,

$$V_{\text{out}} \geq \frac{(1-\nu-\eta_1)^2(1-\nu-\eta_1)^2(\frac{1}{2}-\nu)}{4(1+\nu+\eta_1)^4(1+\nu+\eta_2)^2} \frac{8C_0\sigma_x^2(\frac{1}{2}-\eta_1)}{1+8C_0\sigma_x^2(\frac{1}{2}-\eta_1)} \sigma^2 \left(\frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n(\tau + \frac{\sum_{j>d} \lambda_j}{n})^2} \right),$$

or equivalently

$$V_{\text{out}} \geq \frac{(1-\nu-\eta_1)^2(1-\nu-\eta_1)^2(\frac{1}{2}-\nu)}{4(1+\nu+\eta_1)^4(1+\nu+\eta_2)^2} \frac{8C_0\sigma_x^2(\frac{1}{2}-\eta_1)}{1+8C_0\sigma_x^2(\frac{1}{2}-\eta_1)} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} \right).$$

This gives Proposition 9.

II.2 Proof of the bounds for in-sample error

II.2.1 Algebraic bounds of the in-sample error

The bias and variance of the in-sample error can be decomposed or upper bounded as follows:

$$\begin{aligned}
B_{\text{in}} &= \underbrace{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}^2}_{B_{\text{in},1}} + \underbrace{\|\theta_{(d+1):p}^* - X_{(d+1):p}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{(d+1):p}}^2}_{B_{\text{in},2}} \\
&\quad + \underbrace{2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \hat{\Sigma}_{1:d, (d+1):p} (\theta_{(d+1):p}^* - X_{(d+1):p}^T A^{-1} X \theta^*)}_{B_{\text{in},12}}, \\
V_{\text{in}} &\leq \underbrace{2\sigma^2 \text{Tr}(A^{-1} X_{1:d} \hat{\Sigma}_{1:d} X_{1:d}^T A^{-1})}_{V_{\text{in},1}} + \underbrace{2\sigma^2 \text{Tr}(A^{-1} X_{(d+1):p} \hat{\Sigma}_{(d+1):p} X_{(d+1):p}^T A^{-1})}_{V_{\text{in},2}}. \tag{S28}
\end{aligned}$$

The following algebraic bounds are the foundation of the upper bounds and lower bounds for the in-sample error.

Lemma S3 (Algebraic upper bounds of in-sample error). *Given $\hat{\Sigma}_{1:d}$ is invertible, we have*

$$\begin{aligned}
B_{\text{in}} &\leq 2\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n^2}{\mu_1^2(A_d)} \right)^{-2} + \frac{2}{n} \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \\
&\quad + 3(\|\hat{M}_d\| \frac{n}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} + \|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 + \frac{\|\hat{M}_d\|}{\mu_n(A)^2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|^2), \tag{S29}
\end{aligned}$$

$$V_{\text{in}} \leq 2\sigma^2 \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \frac{d}{n} + 2\sigma^2 \frac{\frac{1}{n} \text{Tr}(X_{(d+1):p} X_{(d+1):p}^T) \mu_1(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_n(A_d)^2}.$$

Proof.

Algebraic upper bound for the in-sample bias. Given invertible $\hat{\Sigma}_{1:d}$, (S29) can be derived similarly as (49) .

Algebraic upper bound for the in-sample variance. Similarly as the derivation in Lemma 27 in Tsigler & Bartlett (2023), given invertible $\hat{\Sigma}_{1:d}$, we have

$$\begin{aligned}
V_{\text{in},1} &= 2\sigma^2 \text{Tr}(A^{-1} X_{1:d} \hat{\Sigma}_{1:d} X_{1:d}^T A^{-1}) \leq 2 \frac{\sigma^2 \mu_1(A_d)^2 \text{Tr}(X_{1:d} \hat{\Sigma}_{1:d}^{-1} X_{1:d})}{\mu_n(A_d)^2 \mu_d(\hat{H}_d \hat{H}_d^T)^2} \\
&= 2\sigma^2 \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \frac{d}{n}.
\end{aligned}$$

Moreover, we have

$$\begin{aligned}
V_{\text{in},2} &= 2\sigma^2 \text{Tr}(A^{-1} X_{(d+1):p} \hat{\Sigma}_{(d+1):p} X_{(d+1):p}^T A^{-1}) \leq 2\sigma^2 \frac{\text{Tr}(X_{(d+1):p} \hat{\Sigma}_{(d+1):p}^{-1} X_{(d+1):p})}{\mu_n(A_d)^2} \\
&= 2\sigma^2 \frac{\frac{1}{n} \text{Tr}((X_{(d+1):p} X_{(d+1):p}^T)^2)}{\mu_n(A_d)^2}
\end{aligned}$$

$$\leq 2\sigma^2 \frac{\frac{1}{n} \text{Tr}(X_{(d+1):p} X_{(d+1):p}^T) \mu_1(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_n(A_d)^2}.$$

By (S28), combining the preceding bounds gives the algebraic upper bound of in-sample variance.

□

Lemma S4 (Algebraic lower bounds of in-sample error). *Given invertible $\hat{\Sigma}_{1:d}$, we have*

$$V_{\text{in}} \geq \sigma^2 \frac{1}{2} \left(\frac{1}{n} \sum_{i=1}^d \frac{\mu_i^2(X_{1:d} X_{1:d}^T)}{(\mu_i(X_{1:d} X_{1:d}^T) + n\tau)^2} + \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \right), \quad (\text{S30})$$

Further if $\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} \geq \frac{n^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d)}$, then

$$B_{\text{in}} \geq \max\{0, 1 - \frac{|B_{\text{in},12}|}{B_{\text{in}}}\} \left(1 - \frac{n^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}}\right)^2 \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)}\right)^{-2}.$$

Proof.

Algebraic lower bound for the in-sample variance. The in-sample variance is

$$V_{\text{in}} = \sigma^2 \text{Tr}(A^{-1} X \hat{\Sigma} X^T A^{-1}) = \sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X X^T)}{(\mu_i(X X^T) + n\tau)^2}.$$

From Weyl's inequality (Lemma S12), we have

$$\begin{aligned} \mu_i(X X^T) &\geq \mu_i(X_{1:d} X_{1:d}^T), \quad i = 1, \dots, d, \\ \mu_i(X X^T) &\geq \mu_i(X_{(d+1):p} X_{(d+1):p}^T), \quad i = 1, \dots, n \end{aligned}$$

Then V_{in} can be lower bounded by

$$\begin{aligned} V_{\text{in}} &\geq \sigma^2 \max\left\{ \frac{1}{n} \sum_{i=1}^d \frac{\mu_i^2(X_{1:d} X_{1:d}^T)}{(\mu_i(X_{1:d} X_{1:d}^T) + n\tau)^2}, \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \right\} \\ &\geq \frac{\sigma^2}{2} \left(\frac{1}{n} \sum_{i=1}^d \frac{\mu_i^2(X_{1:d} X_{1:d}^T)}{(\mu_i(X_{1:d} X_{1:d}^T) + n\tau)^2} + \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \right). \end{aligned}$$

Algebraic lower bound for the in-sample bias. Note that

$$B_{\text{in}} \geq \max\{0, 1 - \frac{|B_{\text{in},12}|}{B_{\text{in},1}}\} B_{\text{in},1}.$$

The result follows from the algebraic lower bound for $B_{\text{in},1}$:

$$B_{\text{in},1} \geq \left(1 - \frac{n^{1/2} \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}}\right)^2 \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)}\right)^{-2},$$

which can be derived similarly as (S6). □

II.2.2 Intermediate bounds of the in-sample error

We give the intermediate bounds of in-sample error under the event that some random quantities in the algebraic bounds above are controlled. In the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_5$ for $0 < \nu < \frac{1}{2}$ defined in Lemma S6, substituting the bounds of $\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}$, $\|X_{(d+1):p}\theta_{(d+1):p}^*\|$ and $\|\hat{M}_d\|$ into the algebraic upper bound of B_{in} yields

$$\begin{aligned} B_{\text{in}} &\leq \frac{1}{(1 - \nu - \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} + \text{Poly}_2(\sigma_x) \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\quad + \frac{1}{(1 - \nu - \eta_1)^2} \text{Poly}_4(\sigma_x) \frac{(n\lambda_{d+1} + \sum_{j>d} \lambda_j)^2}{\mu_n^2(A_d)} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{n}{\mu_1(A_d)} \right)^{-2} \\ &\quad + (\text{Poly}_2(\sigma_x) + \text{Poly}_6(\sigma_x)) \frac{(n\lambda_{d+1} + \sum_{j>d} \lambda_j)^2}{\mu_n(A_d)^2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2. \end{aligned} \quad (\text{S31})$$

In the event $\Omega_1(\nu) \cap \Omega_4$ for $0 < \nu < \frac{1}{2}$ defined in Lemma S6 and with

$$\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}} \geq \frac{(1 + \nu + \eta_1)(1 + \sigma_x^2)^{1/2} n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n(A_d)},$$

substituting the bound of $\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}$ and $\|X_{(d+1):p}\theta_{(d+1):p}^*\|$ into the algebraic lower bound of B_{in} yields

$$B_{\text{in}} \geq \frac{\max\{0, 1 - \frac{|B_{\text{in},2}|}{B_{\text{in},1}}\}}{(1 + \nu + \eta_1)^2} \left(1 - \frac{(1 + \nu + \eta_1)(1 + \sigma_x^2)^{1/2} n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)} \right)^{-2}. \quad (\text{S32})$$

In the event $\Omega_5 \cap \Omega_7(\nu)$ for $0 < \nu < \min\{1, \sigma_x^2\}$, substituting the $\text{Tr}(X_{(d+1):p} X_{(d+1):p}^T)$ and $\mu_1(X_{(d+1):p} X_{(d+1):p}^T)$ into the algebraic upper bound of V_{in} yields

$$V_{\text{in}} \leq 2\sigma^2 \frac{\mu_1^2(A_d)}{\mu_n^2(A_d)} \frac{d}{n} + \text{Poly}_4(\sigma_x) \sigma^2 \frac{(\sum_{j>d} \lambda_j)(n\lambda_{d+1} + \sum_{j>d} \lambda_j)}{\mu_n(A_d)^2}. \quad (\text{S33})$$

II.2.3 Final upper bounds of in-sample bias and variance

We give the final upper bounds of in-sample bias and variance. We first discuss the small or moderate TER regime and then the large TER regime.

(i) Small or moderate TER

From Lemma S7(i) and (S31),(S33), in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_5 \cap \Omega_7(\nu)$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6 and Assumption 2, substituting the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ in (S47)–(S48) into (S31) and (S33), we have for $\tau \geq \lambda_{d+1}$,

$$\begin{aligned} B_{\text{in}} &\leq \frac{(1 + C_1)^2 \text{Poly}_4(\sigma_x)}{(1 - \nu - \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau} \right)^{-2} + \text{Poly}_6(\sigma_x) (1 + C_1)^2 \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\quad + \frac{(1 + C_1)^2 \text{Poly}_8(\sigma_x)}{(1 - \nu - \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_1} + \frac{1}{\tau} \right)^{-2} \end{aligned}$$

$$\begin{aligned}
& + (\text{Poly}_2(\sigma_x) + \text{Poly}_6(\sigma_x)) \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2, \\
V_{\text{in}} & \leq \text{Poly}_4(\sigma_x)(1 + C_1)^2 \frac{d}{n} + \text{Poly}_4(\sigma_x) \sigma^2 \frac{(\sum_{j>d} \lambda_j)(n\lambda_{d+1} + \sum_{j>d} \lambda_j)}{n^2 \tau^2}.
\end{aligned}$$

Hence, we obtain the upper bounds

$$\begin{aligned}
B_{\text{in}} & \leq \frac{(1 + C_1)^4 \text{Poly}_8(\sigma_x)}{(1 - \nu - \eta_1)^2} (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau})^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2), \\
V_{\text{in}} & \leq \text{Poly}_4(\sigma_x)(1 + C_1)^2 \sigma^2 (\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma)}{n}).
\end{aligned}$$

Further with (S15), we have for $\tau \geq \lambda_{d+1}$,

$$B_{\text{in}} \leq \frac{(1 + C_1)^4 \text{Poly}_8(\sigma_x)}{(1 - \nu - \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau})^{-2}.$$

This gives Proposition 4.

(ii) Large TER

From Lemma S7(ii) and (S11),(S13), in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_5 \cap \Omega_6(\nu) \cap \Omega_7(\nu)$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$ defined in Lemma S6 and Assumption 3, substituting the bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ in (S49)–(S50) into (S31) and (S33), we have

$$\begin{aligned}
B_{\text{in}} & \leq \frac{(1 + \nu + \eta_2)^2}{(1 - \nu - \eta_1)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}})^{-2} + \text{Poly}_2(\sigma_x) \frac{(1 + \nu + \eta_2)^2}{(1 - \nu - \eta_2)^2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\
& + \frac{\text{Poly}_4(\sigma_x)(1 + \nu + \eta_2)^2}{(1 - \nu - \eta_1)^2 (1 - \nu - \eta_2)^2} \|\theta_{1:d}^*\|_{\Sigma^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}})^{-2} \\
& + (\text{Poly}_2(\sigma_x) + \frac{\text{Poly}_6(\sigma_x)}{(1 - \nu - \eta_2)^2}) \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2, \\
V_{\text{in}} & \leq \frac{2(1 + \nu + \eta_2)^2}{(1 - \nu - \eta_2)^2} \sigma^2 \frac{d}{n} + \frac{\text{Poly}_4(\sigma_x)}{(1 - \nu - \eta_2)^2} \sigma^2 \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\lambda_{d+1} \frac{r_d(\Sigma)}{n} + \tau)^2}.
\end{aligned}$$

Hence, we obtain the upper bound

$$\begin{aligned}
B_{\text{in}} & \leq \frac{(1 + \nu + \eta_2)^2 \text{Poly}_6(\sigma_x)}{(1 - \nu - \eta_1)^2 (1 - \nu - \eta_2)^2} (\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}})^{-2} + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2), \\
V_{\text{in}} & \leq \frac{(1 + \nu + \eta_2)^2 \text{Poly}_4(\sigma_x)}{(1 - \nu - \eta_2)^2} \sigma^2 (\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}).
\end{aligned}$$

Further with (S16), we have for $\tau \geq 0$,

$$B_{\text{in}} \leq \frac{(1 + \nu + \eta_2)^2 \text{Poly}_6(\sigma_x)}{(1 - \nu - \eta_1)^2 (1 - \nu - \eta_2)^2} \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_1} + \frac{1}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}})^{-2}. \quad (\text{S34})$$

This gives Proposition 10.

II.2.4 Final lower bounds of in-sample bias and variance

We give the final lower bounds of in-sample bias and variance. As preparation, we give the following lemma to compare $B_{\text{in},1}$ and $|B_{\text{in},12}|$.

Lemma S5 (Comparison between $B_{\text{in},1}$ and $|B_{\text{in},12}|$).

(i) Given Assumption 2 and 4(i) and in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_5$ for $0 < \nu < \frac{1}{2}$, we have for $\tau \geq \lambda_{d+1}$,

$$\max\{1 - \frac{|B_{\text{in},12}|}{B_{\text{in},1}}, 0\} \geq \kappa_1(\tau), \quad (\text{S35})$$

where $\kappa_1(\tau) = \max\{1 - (\frac{2C_0\sigma_x^2(2+C_1)\lambda_{d+1}}{\tau}(1 + 16(2C_0\sigma_x^2 + 1)(1 + C_1)\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}) + 64\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}), 0\}$.

(ii) Given Assumption 3 and 4(ii) and in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$, we have for $\tau \geq 0$,

$$\max\{1 - \frac{|B_{\text{in},12}|}{B_{\text{in},1}}, 0\} \geq \kappa_2(\tau), \quad (\text{S36})$$

where $\kappa_2(\tau) = \max\{1 - (16\frac{\lambda_{d+1}\frac{r_d(\Sigma)}{n}}{\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n}}(1 + 112\frac{\sqrt{\delta_2}}{1-\sqrt{\delta_2}}) + 64\frac{\sqrt{\delta_2}}{1-\sqrt{\delta_2}}), 0\}$.

The proof of Lemma S5 is left to Section II.5. In the following, we first discuss the small or moderate TER regime and then the large TER regime.

(i) Small or moderate TER

Lower bound of in-sample bias. From (S17) and Lemma S5(i), in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_5$ for $0 < \nu < \frac{1}{2}$ defined in Lemma S6 and Assumption 4(i), substituting the lower bound of $\mu_n(A_d)$ and lower bound of $\max\{1 - \frac{|B_{\text{in},12}|}{B_{\text{in},1}}, 0\}$ in (S35) into (S32), we have for $\tau \geq \lambda_{d+1}$,

$$B_{\text{in}} \geq \frac{\kappa_1(\tau)}{(1 + \nu + \eta_1)^2} \left(1 - \frac{2(1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2}{\tau \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\tau}\right)^{-2}.$$

Moreover, by applying (S18), we have for $\tau \geq \lambda_{d+1}$,

$$B_{\text{in}} \geq \frac{\kappa_1(\tau)}{(1 + \nu + \eta_1)^2} (1 - \sqrt{\delta_1})^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{1}{\tau}\right)^{-2}.$$

This gives the Proposition 5.

Lower bound of in-sample variance. We first study the first term in (S30). In the event $\Omega_1(\nu)$ for $0 < \nu < \frac{1}{2}$, substituting the lower bound of $\mu_d(X_{1:d}X_{1:d}^T) = n\mu_d(\hat{\Sigma}_{1:d})$ in (S41), we have for $\tau \leq \lambda_d$,

$$\mu_d(X_{1:d}^T X_{1:d}) \geq \lambda_d n (1 - \nu - \eta_1)^2,$$

$$\geq \tau n(1 - \nu - \eta_1)^2,$$

and hence

$$\frac{1}{(1 - \nu - \eta_1)^2} \mu_d(X_{1:d}^T X_{1:d}) \geq \tau n.$$

Then in the event $\Omega_1(\nu)$ for $0 < \nu < \frac{1}{2}$, we have for $\tau \leq \lambda_d$,

$$\frac{1}{n} \sum_{i=1}^d \frac{\mu_i^2(X_{1:d} X_{1:d}^T)}{(\mu_i(X_{1:d} X_{1:d}^T) + n\tau)^2} \geq \left(\frac{1}{1 + \frac{1}{(1-\nu-\eta_1)^2}} \right)^2 \frac{d}{n}. \quad (\text{S37})$$

Then we study the second term in (S30). Given Assumption 2 and in the event $\Omega_5 \cap \Omega_7(\nu)$ for $0 < \nu < \sigma_x^2$, substituting the bounds of $\mu_1(A_d)$ and $\text{Tr}(X_{(d+1):p} X_{(d+1):p}^T)$ in (S46)–(S47) into the second term in (S30), we have for $\tau \geq \lambda_{d+1}$,

$$\begin{aligned} \sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} &\geq \sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_1^2(A_d)} \\ &\geq \sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(2C_0\sigma_x^2 + 1)^2(1 + C_1)^2 n^2 \tau^2} \\ &\geq \sigma^2 \frac{1}{n} \frac{\text{Tr}(X_{(d+1):p} X_{(d+1):p}^T)^2}{(2C_0\sigma_x^2 + 1)^2(1 + C_1)^2 n^3 \tau^2} \\ &\geq \frac{(1 - \nu)^2}{(2C_0\sigma_x^2 + 1)^2(1 + C_1)^2} \sigma^2 \frac{(\frac{\sum_{j>d} \lambda_j}{n})^2}{\tau^2}. \end{aligned}$$

By combining the two terms, in the event $\Omega_1(\nu) \cap \Omega_5 \cap \Omega_7(\nu)$ for $0 < \nu < \frac{1}{2} \min\{1, \sigma_x^2\}$, we have for $\lambda_{d+1} \leq \tau \leq \lambda_d$,

$$V_{\text{in}} \geq \frac{(1 - \nu)^2}{2(2C_0\sigma_x^2 + 1)^2(1 + C_1)^2(1 + \frac{1}{(1-\nu-\eta_2)^2})^2} \sigma^2 \left(\frac{d}{n} + \frac{(\frac{\sum_{j>d} \lambda_j}{n})^2}{\tau^2} \right),$$

or equivalently,

$$V_{\text{in}} \geq \frac{(1 - \nu)^2}{2(2C_0\sigma_x^2 + 1)^2(1 + C_1)^2(1 + \frac{1}{(1-\nu-\eta_2)^2})^2} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2} \right).$$

For $\tau > \lambda_d$, we have

$$\begin{aligned} V_{\text{in}} &\geq \sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \\ &\geq \frac{(1 - \nu)^2}{2(2C_0\sigma_x^2 + 1)^2(1 + C_1)^2} \sigma^2 \frac{(\frac{\sum_{j>d} \lambda_j}{n})^2}{\tau^2}, \end{aligned}$$

or equivalently

$$V_{\text{in}} \geq \frac{(1 - \nu)^2}{2(2C_0\sigma_x^2 + 1)^2(1 + C_1)^2} \sigma^2 \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2}.$$

The result for $\tau \leq \lambda_{d+1}$ follows from the monotonicity of variance in Lemma S11. This gives Proposition 6.

(ii) Large TER

Lower bound of in-sample bias. From (S22) and Lemma S5(ii), in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$ defined in Lemma S6 and Assumption 4(ii), substituting the bounds of $\mu_n(A_d)$ in (S50) and the bound of $\max\{0, 1 - \frac{|B_{\text{in},2}|}{B_{\text{in},1}}\}$ in (S53) into (S32), we have for $\tau \geq 0$,

$$\begin{aligned} B_{\text{in}} \geq & \frac{\kappa_2(\tau)(1-\nu-\eta_2)^2}{(1+\nu+\eta_1)^2} \left(1 - \frac{(1+\nu+\eta_1)(1+\sigma_x^2)^{1/2}n\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{(1-\nu-\eta_2)(\frac{\sum_{j>d}\lambda_j}{n} + \tau)\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}}\right)^2 \\ & \cdot \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{(\frac{\sum_{j>d}\lambda_j}{n} + \tau)}\right)^{-2}. \end{aligned} \quad (\text{S38})$$

Moreover, by applying (S23), we have for $\tau \geq 0$,

$$B_{\text{in}} \geq \frac{\kappa_2(\tau)(1-\nu-\eta_2)^2}{(1+\nu+\eta_1)^2} (1 - \sqrt{\delta_2})^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \frac{n}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})}\right)^{-2}.$$

This gives Proposition 11.

Lower bound of in-sample variance. The first term in (S30) can be studied for $\tau \leq \lambda_d$ similarly as (S37) in the small or moderate TER regime. Then we study the second term in (S30). In the event $\Omega_6(\nu)$ for $0 < \nu < \frac{1}{2}$, we have for $\tau \geq 0$,

$$(1-\nu-\eta_2) \sum_{j>d} \lambda_j \leq \mu_n(X_{(d+1):p} X_{(d+1):p}^T) \leq (1+\nu+\eta_2) \sum_{j>d} \lambda_j,$$

and hence

$$\begin{aligned} \sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} & \geq \sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_1^2(A_d)} \\ & \geq \sigma^2 \frac{1}{n^2} \frac{\text{Tr}(X_{(d+1):p} X_{(d+1):p}^T)^2}{\mu_1^2(A_d)} \\ & \geq \frac{(1-\nu-\eta_2)^2}{(1+\nu+\eta_2)^2} \sigma^2 \frac{(\frac{\sum_{j>d}\lambda_j}{n})^2}{(\tau + \frac{\sum_{j>d}\lambda_j}{n})^2} \\ & = \sigma^2 \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}. \end{aligned} \quad (\text{S39})$$

By combining the two terms, in the event $\Omega_1(\nu) \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{2}$, we have for $0 \leq \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_d$,

$$V_{\text{in}} \geq \frac{(1-\nu-\eta_2)^2}{2(1+\nu+\eta_2)^2(1 + \frac{1}{(1-\nu-\eta_1)^2})} \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}\right).$$

For $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} > \lambda_d$, we have

$$\begin{aligned} V_{\text{in}} &\geq \sigma^2 \frac{1}{n} \sum_{i=1}^n \frac{\mu_i^2(X_{(d+1):p} X_{(d+1):p}^T)}{(\mu_i(X_{(d+1):p} X_{(d+1):p}^T) + n\tau)^2} \\ &\geq \frac{(1 - \nu - \eta_2)^2}{2(1 + \nu + \eta_2)^2} \sigma^2 \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}. \end{aligned}$$

This gives Proposition 12.

II.3 Bounds of random quantities

We give the probability bounds of some random quantities used in the proofs in Supplement Sections II.1 and II.2. We define the related events and give the probability bounds for the events.

Lemma S6 (Bounds of random quantities).

(i) [Bounding $\mu_1(H_d H_d^T)$ and $\mu_d(H_d H_d^T)$] For η_1 defined in Assumption 1 and $0 < \nu < \frac{1}{2}$, denote by $\Omega_1(\nu)$ the event that

$$(1 - \nu - \eta_1)^2 n \leq \mu_d(H_d H_d^T) \leq \mu_1(H_d H_d^T) \leq n(1 + \nu + \eta_1)^2. \quad (\text{S40})$$

In the event $\Omega_1(\nu)$, we have

$$\lambda_d(1 - \nu - \eta_1)^2 \leq \mu_d(\hat{\Sigma}_{1:d}), \quad (\text{S41})$$

$$\frac{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2}{(1 + \nu + \eta_1)^2} \leq \|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}^2 \leq \frac{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2}{(1 - \nu - \eta_1)^2}. \quad (\text{S42})$$

Under the sub-gaussianity of $\Sigma^{-\frac{1}{2}} x_i$ and Assumption 1, $P(\Omega_1(\nu)) \geq 1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\}$.

(ii) [Bounding $\|M_d\|$] Denote by Ω_2 the event that

$$\|M_d\| \leq C_0 \sigma_x^2 (2n \lambda_{d+1}^2 + \sum_{j>d} \lambda_j^2).$$

Under the sub-gaussianity of $\Sigma^{-\frac{1}{2}} x_i$, $P(\Omega_2) \geq 1 - 6\exp\{-\frac{n}{C_0^2}\}$.

(iii) [Bounding the trace of $\text{Tr}(H_d^T H_d)$ and $\text{Tr}(M_d)$] Denote by Ω_{31} the event that

$$\text{Tr}(H_d^T H_d) \leq (1 + \sigma_x^2) n d.$$

Denote by Ω_{32} the event that

$$\text{Tr}(M_d) \leq (1 + \sigma_x^2) n \sum_{j>d} \lambda_j^2.$$

For $0 < \nu < \sigma_x^2$, denote by $\Omega_{33}(\nu)$ the event that

$$\text{Tr}(M_d) \geq (1 - \nu)n \sum_{j>d} \lambda_j^2.$$

Under the sub-gaussianity of $\Sigma^{-\frac{1}{2}}x_i$, $\text{P}(\Omega_{31}) \geq 1 - 2\exp\{-\frac{n}{C_0}\}$, $\text{P}(\Omega_{32}) \geq 1 - 2\exp\{-\frac{n}{C_0}\}$ and $\text{P}(\Omega_{33}(\nu)) \geq 1 - 2\exp\{-\frac{\nu^2 n}{C_0 \sigma_x^4}\}$.

(iv) [Bounding the $\|X_{(d+1):p}\theta_{(d+1):p}^*\|^2$] Denote by Ω_4 the event that

$$\|X_{(d+1):p}\theta_{(d+1):p}^*\|^2 \leq (1 + \sigma_x^2)n\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2. \quad (\text{S43})$$

Under the sub-gaussianity of $\Sigma^{-\frac{1}{2}}x_i$, $\text{P}(\Omega_4) \geq 1 - 2\exp\{-\frac{n}{C_0}\}$.

(v) [Bounding the $\mu_1(X_{(d+1):p}X_{(d+1):p}^T)$ with $r_d(\Sigma) \leq C_1 n$] Denote by Ω_5 the event that

$$\mu_1(X_{(d+1):p}X_{(d+1):p}^T) \leq C_0 \sigma_x^2 (2n\lambda_{d+1} + \sum_{j>d} \lambda_j). \quad (\text{S44})$$

In the event Ω_5 , we have

$$\begin{aligned} \|\hat{M}_d\| &\leq \frac{1}{n} \|X_{(d+1):p}X_{(d+1):p}^T\|^2 \\ &\leq \frac{4C_0^2 \sigma_x^4}{n} (n\lambda_{d+1} + \sum_{j>d} \lambda_j)^2. \end{aligned}$$

Under the sub-gaussianity of $\Sigma^{-\frac{1}{2}}x_i$, $\text{P}(\Omega_5) \geq 1 - 6\exp\{-\frac{n}{C_0}\}$.

(vi) [Bounding the $\mu_1(X_{(d+1):p}X_{(d+1):p}^T)$ and $\mu_n(X_{(d+1):p}X_{(d+1):p}^T)$ with Assumption 3] For η_2 defined in Assumption 3 and $0 < \nu < \frac{1}{2}$, denote by $\Omega_6(\nu)$ the event that

$$(1 - \nu - \eta_2) \sum_{j>d} \lambda_j \leq \mu_n(X_{d+1:p}X_{d+1:p}^T) \leq \mu_1(X_{d+1:p}X_{d+1:p}^T) \leq (1 + \nu + \eta_2) \sum_{j>d} \lambda_j. \quad (\text{S45})$$

Under the sub-gaussianity of $\Sigma^{-\frac{1}{2}}x_i$ and Assumption 3, let $\frac{\nu^2 r_d(\Sigma)}{C_0^2 \sigma_x^4} > 1$, $\text{P}(\Omega_6(\nu)) \geq 1 - 2n\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0 \sigma_x^2}\} - 4\exp\{-\frac{n}{C_0}\}$.

(vii) [Bounding of $\text{Tr}(X_{(d+1):p}X_{(d+1):p}^T)$] For $0 < \nu < \min\{1, \sigma_x^2\}$, denote by $\Omega_7(\nu)$ the event that

$$(1 - \nu)n \sum_{j>d} \lambda_j \leq \text{Tr}(X_{(d+1):p}X_{(d+1):p}^T) \leq (1 + \nu)n \sum_{j>d} \lambda_j. \quad (\text{S46})$$

Under the sub-gaussianity of $\Sigma^{-1/2}x_i$, $\text{P}(\Omega_7(\nu)) \geq 1 - 2\exp\{-\frac{\nu^2 n}{C_0 \sigma_x^4}\}$.

Proof.

(i) Given the sub-gaussianity of $\Sigma^{-\frac{1}{2}}x_i$, from Lemma S17, we have with probability $1 - 2\exp\{-\frac{t}{C_0^2 \sigma_x^2}\}$,

$$(\sqrt{n} - C_0 \sigma_x^2 \sqrt{d} - \sqrt{t})^2 \leq \mu_d(H_d H_d^T) \leq \mu_1(H_d H_d^T) \leq (\sqrt{n} + C_0 \sigma_x^2 \sqrt{d} + \sqrt{t})^2,$$

which implies that

$$n(1 - C_0\sigma_x^2\sqrt{\frac{d}{n}} - \sqrt{\frac{t}{n}})^2 \leq \mu_d(H_d H_d^T) \leq \mu_1(H_d H_d^T) \leq n(1 + C_0\sigma_x^2\sqrt{\frac{d}{n}} + \sqrt{\frac{t}{n}})^2.$$

Under Assumption 1, with $0 < \nu < \frac{1}{2}$, we have $C_0\sigma_x^2\sqrt{\frac{d}{n}} + \nu \leq \nu + \eta_1 < 1$. Taking $t = \nu^2 n$, we have with probability at least $1 - 2\exp\{-\frac{\nu^2 n}{C_0^2 \sigma_x^4}\}$,

$$n(1 - \nu - \eta_1)^2 \leq \mu_d(H_d H_d^T) \leq \mu_1(H_d H_d^T) \leq n(1 + \nu + \eta_1)^2.$$

In the event $\Omega_1(\nu)$ for $0 < \nu < \frac{1}{2}$, we also have

$$\begin{aligned} \mu_d(X_{1:d}^T X_{1:d}) &= \mu_d(\Sigma_{1:d}^{1/2} (\Sigma_{1:d}^{-1/2} X_{1:d}^T X_{1:d} \Sigma_{1:d}^{-1/2}) \Sigma_{1:d}^{1/2}) \\ &\geq \mu_d(\Sigma_{1:d}) \mu_d(\Sigma_{1:d}^{-1/2} X_{1:d}^T X_{1:d} \Sigma_{1:d}^{-1/2}) \\ &\geq \lambda_d n (1 - \nu - \eta_1)^2, \end{aligned}$$

or equivalently

$$\mu_d(\hat{\Sigma}_{1:d}) = \mu_d\left(\frac{X_{1:d}^T X_{1:d}}{n}\right) \geq \lambda_d (1 - \nu - \eta_1)^2.$$

Moreover, in the event $\Omega_1(\nu)$ for $\nu < \frac{1}{2}$,

$$\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 = \theta_{1:d}^{*T} \left(\frac{X_{1:d}^T X_{1:d}}{n}\right)^{-1} \theta_{1:d}^* = \theta_{1:d}^{*T} \Sigma_{1:d}^{-1/2} \left(\frac{H_d H_d^T}{n}\right)^{-1} \Sigma_{1:d}^{-1/2} \theta_{1:d}^*,$$

and hence

$$\frac{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2}{(1 + \nu + \eta_1)^2} \leq \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \mu_1^{-1}\left(\frac{H_d H_d^T}{n}\right) \leq \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \leq \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \mu_d^{-1}\left(\frac{H_d H_d^T}{n}\right) \leq \frac{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2}{(1 - \nu - \eta_1)^2}.$$

(ii) Given the sub-gaussianity of $\Sigma^{-\frac{1}{2}} x_i$, from Lemma S16, we have with probability at least $1 - 6\exp\{-\frac{n}{C_0}\}$,

$$\|X_{(d+1):p} \Sigma_{(d+1):p} X_{(d+1):p}^T\| \leq C_0 \sigma_x^2 (2n \lambda_{d+1}^2 + \sum_{j>d} \lambda_j^2).$$

(iii) Given the sub-gaussianity of $\Sigma^{-\frac{1}{2}} x_i$, from Lemma S18, with probability at least $1 - 2\exp\{-\frac{n}{C_0}\}$,

$$\begin{aligned} \text{Tr}(X_{1:d} \Sigma_{1:d}^{-1} X_{1:d}^T) &< (1 + \sigma_x^2) n d, \\ \text{Tr}(X_{d+1:p} \Sigma_{d+1:p} X_{d+1:p}^T) &< (1 + \sigma_x^2) n \sum_{j>d} \lambda_j^2. \end{aligned}$$

(iv) Given the sub-gaussianity of $\Sigma^{-\frac{1}{2}} x_i$, from Lemma S18, with probability at least $1 - 2\exp\{-\frac{n}{C_0}\}$,

$$\|X_{(d+1):p} \theta_{(d+1):p}^*\|^2 \leq (1 + \sigma_x^2) n \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2.$$

(v) Given the sub-gaussianity of $\Sigma^{-\frac{1}{2}}x_i$, from Lemma S16, with probability at least $1 - 6\exp\{-\frac{n}{C_0}\}$,

$$\mu_1(X_{(d+1):p}X_{(d+1):p}^T) \leq C_0\sigma_x^2(2n\lambda_{d+1} + \sum_{j>d} \lambda_j).$$

(vi) Given the sub-gaussianity of $\Sigma^{-\frac{1}{2}}x_i$, from Lemma S14, with probability at least

$$1 - 2\exp\{-\min\{r_d(\Sigma)\frac{\delta^2}{C_0^2\sigma_x^4}, \sqrt{r_d(\Sigma)}\frac{\delta}{C_0\sigma_x^2}\}\},$$

$$(1 - \delta) \sum_{j>d} \lambda_j \leq \mu_n(\text{Diag}(X_{d+1:p}X_{d+1:p}^T)) \leq \mu_1(\text{Diag}(X_{d+1:p}X_{d+1:p}^T)) \leq (1 + \delta) \sum_{j>d} \lambda_j.$$

From Lemma S15, we have with probability at least $1 - 4\exp\{-n/C_0\}$,

$$\|X_{d+1:p}X_{d+1:p}^T - \text{Diag}(X_{d+1:p}X_{d+1:p}^T)\| \leq C_0\sigma_x^2 \sqrt{4n^2\lambda_{d+1}^2 + 2n \sum_{j>d} \lambda_j^2}.$$

Then with probability at least $1 - 2n\exp\{-\min\{r_d(\Sigma)\frac{t^2}{C_0^2\sigma_x^4}, \sqrt{r_d(\Sigma)}\frac{t}{C_0\sigma_x^2}\}\} - 4\exp\{-\frac{n}{C_0}\}$,

$$\begin{aligned} (1 - t) \sum_{j>d} \lambda_j - C_0\sigma_x^2 \sqrt{4n^2\lambda_{d+1}^2 + 2n \sum_{i>d} \lambda_i^2} &\leq \mu_n(X_{d+1:p}X_{d+1:p}^T) \\ &\leq \mu_1(X_{d+1:p}X_{d+1:p}^T) \leq (1 + t) \sum_{j>d} \lambda_j + C_0\sigma_x^2 \sqrt{4n^2\lambda_{d+1}^2 + 2n \sum_{j>d} \lambda_j^2}, \end{aligned}$$

which can be equivalently written as

$$\begin{aligned} \sum_{j>d} \lambda_j (1 - t - C_0\sigma_x^2 \sqrt{\frac{4n^2}{r_d^2(\Sigma)} + \frac{2n \sum_{j>d} \lambda_j^2}{(\sum_{j>d} \lambda_j)^2}}) &\leq \mu_n(X_{d+1:p}X_{d+1:p}^T) \\ &\leq \mu_1(X_{d+1:p}X_{d+1:p}^T) \leq \sum_{j>d} \lambda_j (1 + t + C_0\sigma_x^2 \sqrt{\frac{4n^2}{r_d(\Sigma)^2} + \frac{2n \sum_{j>d} \lambda_j^2}{(\sum_{j>d} \lambda_j)^2}}). \end{aligned}$$

Hence

$$\begin{aligned} \sum_{j>d} \lambda_j (1 - t - C_0\sigma_x^2 \sqrt{\frac{4n^2}{r_d^2(\Sigma)} + \frac{2n}{r_d(\Sigma)}}) &\leq \mu_n(X_{d+1:p}X_{d+1:p}^T) \\ &\leq \mu_1(X_{d+1:p}X_{d+1:p}^T) \leq \sum_{j>d} \lambda_j (1 + t + C_0\sigma_x^2 \sqrt{\frac{4n^2}{r_d(\Sigma)^2} + \frac{2n}{r_d(\Sigma)}}). \end{aligned}$$

Under Assumption 3, with $0 < \nu < \frac{1}{2}$ and $\frac{\nu^2 r_d(\Sigma)}{C_0^2 \sigma_x^4} > 1$, we have $C_0\sigma_x^2 \sqrt{\frac{4n^2}{r_d(\Sigma)^2} + \frac{2n}{r_d(\Sigma)}} + \nu \leq \nu + \eta_2 < 1$.

Taking $t = \nu$, we have with probability at least $1 - 2\exp\{-\frac{\nu\sqrt{r_d(\Sigma)}}{C_0\sigma_x^2}\} - 4\exp\{-\frac{n}{C_0}\}$,

$$(1 - \nu - \eta_2) \sum_{j>d} \lambda_j \leq \mu_n(X_{d+1:p}X_{d+1:p}^T) \leq \mu_1(X_{d+1:p}X_{d+1:p}^T) \leq (1 + \nu + \eta_2) \sum_{j>d} \lambda_j.$$

(vii) From Lemma S18, let $t = \frac{\nu^2 n}{\sigma_x^4}$ for $0 < \nu < \min\{\sigma_x^2, 1\}$, with probability at least $1 - \exp\{-\frac{\nu^2 n}{C_0\sigma_x^4}\}$,

$$(1 - \nu)n \sum_{j>d} \lambda_j \leq \text{Tr}(X_{d+1:p}X_{d+1:p}^T) \leq (1 + \nu)n \sum_{j>d} \lambda_j.$$

□

II.4 Bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$

We give the following lemma to control $\mu_1(A_d)$ and $\mu_n(A_d)$ under the small or moderate TER regime and the large TER regime.

Lemma S7 (Bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$).

(i) (Small or moderate TER regime) Given Assumption 2 and in the event Ω_5 defined in Lemma S6, we have for $\tau \geq \lambda_{d+1}$,

$$\mu_1(A_d) \leq (2C_0\sigma_x^2 + 1)(1 + C_1)n\tau, \quad (\text{S47})$$

$$\mu_n(A_d) \geq n\tau. \quad (\text{S48})$$

(ii) (Large TER regime) Given Assumption 3 and in the event $\Omega_6(\nu)$ for $0 < \nu < \frac{1}{2}$ defined in Lemma S6, we have for $\tau \geq 0$,

$$\mu_1(A_d) \leq (1 + \nu + \eta_2) \sum_{j>d} \lambda_j + n\tau, \quad (\text{S49})$$

$$\mu_n(A_d) \geq (1 - \nu - \eta_2) \sum_{j>d} \lambda_j + n\tau. \quad (\text{S50})$$

Proof.

(i) By the definition of A_d ,

$$\mu_n(A_d) \geq n\tau.$$

Given Assumption 2 and in the event Ω_5 , we have

$$\mu_1(X_{(d+1):p} X_{(d+1):p}^T) \leq C_0\sigma_x^2(2n\lambda_{d+1} + \sum_{j>d} \lambda_j).$$

Hence

$$\begin{aligned} \mu_1(A_d) &\leq 2C_0\sigma_x^2(n\lambda_{d+1} + \sum_{j>d} \lambda_j) + n\tau \\ &\leq 2C_0\sigma_x^2(1 + C_1)n\lambda_{d+1} + n\tau. \end{aligned}$$

Further if $\tau \geq \lambda_{d+1}$, then

$$\begin{aligned} \mu_1(A_d) &\leq 2C_0\sigma_x^2(1 + C_1)n\lambda_{d+1} + n\tau \\ &\leq (2C_0\sigma_x^2 + 1)(1 + C_1)n\tau. \end{aligned} \quad (\text{S51})$$

(ii) Given Assumption 3 and in the event $\Omega_6(\nu)$, we have

$$(1 - \nu - \eta_2) \sum_{j>d} \lambda_j \leq \mu_1(X_{d+1:p} X_{d+1:p}^T) \leq \mu_n(X_{d+1:p} X_{d+1:p}^T) \leq (1 + \nu + \eta_2) \sum_{j>d} \lambda_j.$$

Then we have for $\tau \geq 0$,

$$(1 - \nu - \eta_2) \sum_{j>d} \lambda_j + n\tau \leq \mu_n(A_d) \leq \mu_1(A_d) \leq (1 + \nu + \eta_2) \sum_{j>d} \lambda_j + n\tau.$$

□

In the large TER regime, an upper bound on $\mu_1(A_d)$ similar to (S49) can be obtained by using (S51) and then Assumption 3, $\frac{r_d(\Sigma)}{n} \geq c_x$:

$$\begin{aligned} \mu_1(A_d) &\leq 2C_0\sigma_x^2(1 + C_1)n\lambda_{d+1} + n\tau \\ &\leq 2C_0\sigma_x^2(1 + C_1)c_x \sum_{j>d} \lambda_j + n\tau. \end{aligned}$$

This inequality can be used to achieve a similar purpose as (S49) in our proofs.

II.5 Comparison between $B_{\text{in},1}$ and $|B_{\text{in},12}|$

We give a proof of Lemma S5, which is re-stated as follows.

Lemma S5 (Comparison between $B_{\text{in},1}$ and $|B_{\text{in},12}|$).

(i) Given Assumption 2 and 4(i) and in the event $\Omega_1(\nu) \cap \Omega_4$ for $0 < \nu < \frac{1}{2}$, we have for $\tau \geq \lambda_{d+1}$,

$$\max\{1 - \frac{|B_{\text{in},12}|}{B_{\text{in},1}}, 0\} \geq \kappa_1(\tau), \quad (\text{S52})$$

where $\kappa_1(\tau) = \max\{1 - (\frac{2C_0\sigma_x^2(2+C_1)\lambda_{d+1}}{\tau}(1 + 16(2C_0\sigma_x^2 + 1)(1 + C_1)\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}) + 64\frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}}), 0\}$.

(ii) Given Assumption 3 and 4(ii) and in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$, we have for $\tau \geq 0$,

$$\max\{1 - \frac{|B_{\text{in},12}|}{B_{\text{in},1}}, 0\} \geq \kappa_2(\tau), \quad (\text{S53})$$

where $\kappa_2(\tau) = \max\{1 - (16\frac{\lambda_{d+1}\frac{r_d(\Sigma)}{n}}{\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n}}(1 + 112\frac{\sqrt{\delta_2}}{1-\sqrt{\delta_2}}) + 64\frac{\sqrt{\delta_2}}{1-\sqrt{\delta_2}}), 0\}$.

We first give an algebraic bound of $\frac{|B_{\text{in},12}|}{B_{\text{in},1}}$.

Lemma S8 (Algebraic bound of $\frac{|B_{\text{in},12}|}{B_{\text{in},1}}$). Given $\frac{\mu_1(X_{(d+1):p}X_{(d+1):p}^T)}{\mu_n(A_d)} \leq 1$, we have

$$\begin{aligned} \frac{|B_{\text{in},12}|}{B_{\text{in},1}} &\leq \frac{2\mu_1(X_{(d+1):p}X_{(d+1):p}^T)}{\mu_n(A_d)} \frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} \\ &\quad + 4 \frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}. \end{aligned}$$

Proof. Given $\frac{\mu_1(X_{(d+1):p}X_{(d+1):p}^T)}{\mu_n(A_d)} \leq 1$, we have from the Cauchy–Schwartz inequality and simple manipulations,

$$\begin{aligned}
& \|\theta_{(d+1):p}^* - X_{d+1:p}^T A^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 \\
& \leq 2\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 + 2\|X_{d+1:p}^T A^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 \\
& = 2\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 + 2\left\|\frac{1}{n} X_{(d+1):p} X_{(d+1):p}^T A^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\right\|^2 \\
& \leq 2\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 + 2\mu_1(X_{(d+1):p} X_{d+1:p}^T A^{-1}) \left\|\frac{1}{n} X_{(d+1):p} \theta_{(d+1):p}^*\right\|^2 \\
& \leq (2 + 2\mu_1(X_{(d+1):p} X_{d+1:p}^T A^{-1})) \|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 \\
& \leq (2 + 2\frac{\mu_1(X_{(d+1):p} X_{d+1:p}^T)}{\mu_n(A_d)}) \|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2 \\
& \leq 4\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}^2.
\end{aligned} \tag{S54}$$

Then we apply the triangle inequality to $|B_{\text{in},12}|$:

$$\begin{aligned}
|B_{\text{in},12}| &= |2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \hat{\Sigma}_{1:d,(d+1):p} (\theta_{(d+1):p}^* - X_{d+1:p}^T A^{-1} X \theta^*)| \\
&\leq |2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \hat{\Sigma}_{1:d,(d+1):p} X_{d+1:p}^T A^{-1} X_{1:d} \theta_{1:d}^*| \\
&\quad + 2|(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \hat{\Sigma}_{1:d,(d+1):p} (\theta_{(d+1):p}^* - X_{d+1:p}^T A^{-1} X_{(d+1):p} \theta_{(d+1):p}^*)|.
\end{aligned}$$

The two terms on the right-hand side of the inequality above can be bounded as follows. First,

$$\begin{aligned}
& |2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \hat{\Sigma}_{1:d,(d+1):p} X_{d+1:p}^T A^{-1} X_{1:d} \theta_{1:d}^*| \\
&= |2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \frac{X_{1:d}^T X_{(d+1):p}}{n} X_{d+1:p}^T A^{-1} X_{1:d} \theta_{1:d}^*| \\
&= |2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \frac{X_{1:d}^T X_{(d+1):p}}{n} X_{d+1:p}^T A_d^{-1} X_{1:d} (I_d + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} \theta_{1:d}^*| \\
&\quad (\text{from Lemma S10(ii)}) \\
&= |2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \frac{X_{1:d}^T X_{(d+1):p}}{n} X_{d+1:p}^T A_d^{-1} X_{1:d} (I_d - X_{1:d}^T A^{-1} X_{1:d}) \theta_{1:d}^*| \\
&\quad (\text{from Lemma S10(iii)}) \\
&= |2(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \frac{X_{1:d}^T X_{(d+1):p}}{n} X_{d+1:p}^T A_d^{-1} X_{1:d} (\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*)| \\
&\leq \frac{2\mu_1(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_n(A_d)} \|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}} \|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}},
\end{aligned}$$

Second,

$$\begin{aligned}
& 2|(\theta_{1:d}^{*T} - \theta^{*T} X^T A^{-1} X_{1:d}) \hat{\Sigma}_{1:d,(d+1):p} (\theta_{(d+1):p}^* - X_{d+1:p}^T A^{-1} X_{(d+1):p} \theta_{(d+1):p}^*)| \\
& \leq 2\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}} \|\theta_{(d+1):p}^* - X_{d+1:p}^T A^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}
\end{aligned}$$

(Cauchy–Schwartz inequality)

$$\leq 4\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}} \|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}. \quad (\text{From (S54)})$$

Combining the preceding three displays yields

$$\begin{aligned} |B_{\text{in},12}| &\leq \frac{2\mu_1(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_n(A_d)} \|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}} \|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}} \\ &\quad + 4\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}} \|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}, \end{aligned}$$

or equivalently

$$\frac{|B_{\text{in},12}|}{B_{\text{in},1}} \leq \frac{2\mu_1(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_n(A_d)} \frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} + 4 \frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}.$$

□

Next, it is desired to control the quantities $\frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}$ and $\frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}$, which is stated in Lemma S9 below. We discuss the small or moderate TER regime and the large TER regime, respectively.

Lemma S9 (Bound of $\frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}$ and $\frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}$).

(i) Given Assumption 2 and 4(i), in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_5$ for $0 < \nu < \frac{1}{2}$, we have for $\tau \geq \lambda_{d+1}$,

$$\begin{aligned} \frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} &\leq 1 + \frac{(2C_0\sigma_x^2 + 1)(1 + C_1)}{(1 - \nu - \eta_1)^2} \frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}, \\ \frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} &\leq \frac{1}{(1 - \nu - \eta_1)^2} \frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}. \end{aligned}$$

(ii) Given Assumption 3 and 4(ii) and in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$, we have for $\tau \geq 0$,

$$\begin{aligned} \frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} &\leq 1 + \frac{(1 + \nu + \eta_2)}{(1 - \nu - \eta_2)(1 - \nu - \eta_1)^2} \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}, \\ \frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} &\leq \frac{1}{(1 - \nu - \eta_1)^2} \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}. \end{aligned}$$

Proof. As preparation, we derive a useful identity. We have

$$\begin{aligned} \theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^* &= \theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^* - X_{1:d}^T A^{-1} X_{(d+1):p} \theta_{(d+1):p}^* \\ &= (I_d + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} \theta_{1:d}^* - (I_d + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} X_{1:d}^T A_d X_{(d+1):p} \theta_{(d+1):p}^* \\ &\quad (\text{from Lemma S10(ii) and (iii)}) \end{aligned}$$

$$= -(I_d + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} (X_{1:d}^T A_d X_{(d+1):p} \theta_{(d+1):p}^* - \theta_{1:d}^*). \quad (\text{S55})$$

(i) We first discuss the bound of $\frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}$. We have,

$$\begin{aligned} & \frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} \\ & \leq 1 + \frac{\|X_{1:d}^T A^{-1} X \theta^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|(I_d + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} (X_{1:d}^T A_d X_{(d+1):p} \theta_{(d+1):p}^* - \theta_{1:d}^*)\|_{\hat{\Sigma}_{1:d}}} \quad (\text{Using (S55)}) \\ & = 1 + \frac{\|X_{1:d}^T A^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|_{\hat{\Sigma}_{1:d}}}{\|\hat{\Sigma}_{1:d}^{1/2} (I_{1:d} + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} (X_{1:d}^T A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \theta_{1:d}^*)\|} \\ & = 1 + \frac{\|\hat{\Sigma}_{1:d}^{1/2} (I_{1:d} + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} X_{1:d}^T A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|}{\|\hat{\Sigma}_{1:d}^{1/2} (I_{1:d} + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} (X_{1:d}^T A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \theta_{1:d}^*)\|} \quad (\text{from Lemma S10(ii)}) \\ & = 1 + \frac{\|(\hat{\Sigma}_{1:d}^{-1} + \hat{H}_d A_d^{-1} \hat{H}_d^T)^{-1} \hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|}{\|(\hat{\Sigma}_{1:d}^{-1} + \hat{H}_d A_d^{-1} \hat{H}_d^T)^{-1} (\hat{H}_d T A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \hat{\Sigma}_{1:d}^{-1/2} \theta_{1:d}^*)\|} \\ & \leq 1 + \frac{\mu_1(\hat{\Sigma}_{1:d}^{-1} + \hat{H}_d A_d^{-1} \hat{H}_d^T) n^{1/2} \mu_1(A_d^{-1}) \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{\mu_n(\hat{\Sigma}_{1:d}^{-1} + \hat{H}_d A_d^{-1} \hat{H}_d^T) \|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \hat{\Sigma}_{1:d}^{-1/2} \theta_{1:d}^*\|} \\ & \leq 1 + \frac{\mu_1(\hat{\Sigma}_{1:d}^{-1} + \hat{H}_d A_d^{-1} \hat{H}_d^T) n^{1/2} \mu_1(A_d^{-1}) \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{n \mu_n(A_d^{-1}) \|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \hat{\Sigma}_{1:d}^{-1/2} \theta_{1:d}^*\|} \\ & \leq 1 + \frac{(\frac{1}{\mu_d(\hat{\Sigma}_{1:d})} + \frac{n}{\mu_n(A_d)}) n^{1/2} \mu_1(A_d^{-1}) \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{n \mu_n(A_d^{-1}) \|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \hat{\Sigma}_{1:d}^{-1/2} \theta_{1:d}^*\|} \\ & \leq 1 + \frac{(\frac{1}{\mu_d(\hat{\Sigma}_{1:d})} + \frac{n}{\mu_n(A_d)}) n^{1/2} \mu_1(A_d^{-1}) \|X_{(d+1):p} \theta_{(d+1):p}^*\|}{n \mu_n(A_d^{-1}) \|\theta_{1:d}^* - \hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|}. \quad (\text{S56}) \end{aligned}$$

In the event $\Omega_1(\nu) \cap \Omega_4$ and given $\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} > \|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|$, substituting (S40)–(S43) into (S56), we have

$$\begin{aligned} & \frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} \\ & \leq 1 + \frac{1}{(1 - \nu - \eta_1)^2} \frac{\mu_1(A_d) (\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)}) (1 + \nu + \eta_1) (1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n A_d \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \\ & \quad \times \frac{\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}}{\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} - \|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|} \\ & \leq 1 + \frac{1}{(1 - \nu - \eta_1)^2} \frac{\mu_1(A_d) (\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)}) (1 + \nu + \eta_1) (1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n A_d \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \\ & \quad \times \frac{1}{|1 - \frac{n(1+\nu+\eta_1)(1+\sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}}|}. \quad (\text{S57}) \end{aligned}$$

Then we discuss $\frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}$. Substituting (S55) in the denominator, we have

$$\begin{aligned}
& \frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} \\
&= \frac{\left\| \frac{X_{(d+1):p}}{n} \theta_{(d+1):p}^* \right\|}{\|(I_d + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} (X_{1:d}^T A_d X_{(d+1):p} \theta_{(d+1):p}^* - \theta_{1:d}^*)\|_{\hat{\Sigma}_{1:d}}} \\
&\leq \frac{\left\| \frac{X_{(d+1):p}}{n} \theta_{(d+1):p}^* \right\|}{\|\hat{\Sigma}^{1/2} (I_{1:d} + X_{1:d}^T A_d^{-1} X_{1:d})^{-1} (X_{1:d}^T A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \theta_{1:d}^*)\|} \\
&= \frac{\left\| \frac{X_{(d+1):p}}{n} \theta_{(d+1):p}^* \right\|}{\|(\hat{\Sigma}_{1:d}^{-1} + \hat{H}_d^T A_d^{-1} X_{1:d} \hat{\Sigma}^{-1/2})^{-1} (\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \hat{\Sigma}_{1:d}^{-1/2} \theta_{1:d}^*)\|} \\
&\leq \frac{\mu_1(\hat{\Sigma}_{1:d}^{-1} + \hat{H}_d A_d^{-1} \hat{H}_d^T) \left\| \frac{X_{(d+1):p}}{n} \theta_{(d+1):p}^* \right\|}{\|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \hat{\Sigma}_{1:d}^{-1/2} \theta_{1:d}^*\|}. \tag{S58}
\end{aligned}$$

In the event $\Omega_1(\nu) \cap \Omega_4$ and given $\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} > \|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|$, substituting (S40)–(S43) into (S58), we have

$$\begin{aligned}
\frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} &\leq \frac{1}{(1 - \nu - \eta_1)^2} \frac{(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)})(1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}} \\
&\quad \times \frac{\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}}}{\|\hat{H}_d^T A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^* - \hat{\Sigma}^{-1/2} \theta_{1:d}^*\|} \\
&\leq \frac{1}{(1 - \nu - \eta_1)^2} \frac{(\frac{1}{\lambda_d} + \frac{n}{\mu_n(A_d)})(1 + \nu + \eta_1)(1 + \sigma_x^2) \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \\
&\quad \times \frac{1}{\left| 1 - \frac{n(1+\nu+\eta_1)(1+\sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\mu_n(A_d) \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \right|}. \tag{S59}
\end{aligned}$$

We control $\frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}$ and $\frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}}$ in small or moderate TER and large TER respectively.

In small or moderate TER regime, given Assumption 2 and 4(i), and in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_5$ for $0 < \nu < \frac{1}{4}$, we have for $\tau \geq \lambda_{d+1}$,

$$\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} > \|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|.$$

Substituting bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ in Lemma S7(i) into (S57) and (S59), we have

$$\begin{aligned}
& \frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} \\
&\leq 1 + \frac{(2C_0 \sigma_x^2 + 1)(1 + C_1)}{(1 - \nu - \eta_1)^2} \frac{(1 + \nu + \eta_1)(\frac{1}{\lambda_d} + \frac{1}{\lambda_{d+1}})(1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \frac{1}{1 - \sqrt{\delta_1}}
\end{aligned}$$

$$\leq 1 + \frac{(2C_0\sigma_x^2 + 1)(1 + C_1)}{(1 - \nu - \eta_1)^2} \frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}.$$

For $\tau \geq \lambda_{d+1}$,

$$\begin{aligned} \frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} &\leq \frac{1}{(1 - \nu - \eta_1)^2} \frac{(\frac{1}{\lambda_d} + \frac{1}{\lambda_{d+1}})(1 + \nu + \eta_1)(1 + \sigma_x^2) \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \frac{1}{1 - \sqrt{\delta_1}} \\ &\leq \frac{1}{(1 - \nu - \eta_1)^2} \frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}. \end{aligned}$$

In large TER regime, under Assumption 3 and 4(ii) and in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_6(\nu)$ for $\nu < \frac{1}{4}$, we have for $\tau \geq 0$,

$$\|\theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}^{-1}} > \|\hat{H}_d A_d^{-1} X_{(d+1):p} \theta_{(d+1):p}^*\|.$$

Substituting bounds of $\mu_1(A_d)$ and $\mu_n(A_d)$ in Lemma S7(ii) into (S57) and (S59), we have

$$\begin{aligned} &\frac{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X_{1:d} \theta_{1:d}^*\|_{\hat{\Sigma}_{1:d}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} \\ &\leq 1 + \frac{(1 + \nu + \eta_2)}{(1 - \nu - \eta_1)^2(1 - \nu - \eta_2)} \frac{(1 + \nu + \eta_1)(\frac{1}{\lambda_d} + \frac{n}{(1 - \nu - \eta_2) \sum_{j>d} \lambda_j})(1 + \sigma_x^2)^{1/2} \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \frac{1}{1 - \sqrt{\delta_2}} \\ &\leq \frac{(1 + \nu + \eta_2)}{(1 - \nu - \eta_1)^2(1 - \nu - \eta_2)} \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}, \end{aligned}$$

and

$$\begin{aligned} \frac{\|\theta_{(d+1):p}^*\|_{\hat{\Sigma}_{(d+1):p}}}{\|\theta_{1:d}^* - X_{1:d}^T A^{-1} X \theta^*\|_{\hat{\Sigma}_{1:d}}} &\leq \frac{1}{(1 - \nu - \eta_1)^2} \frac{(1 + \nu + \eta_1)(\frac{1}{\lambda_d} + \frac{n}{(1 - \nu - \eta_2) \sum_{j>d} \lambda_j})(1 + \sigma_x^2) \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}}{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}} \\ &\quad \times \frac{1}{1 - \sqrt{\delta_2}} \\ &\leq \frac{1}{(1 - \nu - \eta_1)^2} \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}. \end{aligned}$$

□

Proof of Lemma S5. We obtain Lemma S5 by combining Lemma S8 and S9. We discuss the small or moderate TER regime and the large TER regime, respectively.

Small or moderate TER. We substitute the bounds from Lemma S9(i) into Lemma S8. Given

Assumption 2, 4(i) and in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_5$ for $0 < \nu < \frac{1}{4}$, we have for $\tau \geq \lambda_{d+1}$ and $\frac{\mu_1(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_n(A_d)} \leq 1$,

$$\frac{|\text{B}_{\text{in},12}|}{\text{B}_{\text{in},1}} \leq \frac{2\mu_1(X_{(d+1):p} X_{(d+1):p}^T)}{\mu_n(A_d)} (1 + 16(2C_0\sigma_x^2 + 1)(1 + C_1) \frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}) + 64 \frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}.$$

Under Assumption 2 and in the event Ω_5 , from (S44), we have

$$\mu_1(X_{(d+1):p} X_{(d+1):p}^T) \leq C_0 \sigma_x^2 (2n\lambda_{d+1} + \sum_{j>d} \lambda_j) \leq C_0 \sigma_x^2 (2 + C_1) n \lambda_{d+1}.$$

Note that $\mu_n(A_d) \geq n\tau$. We have for $\tau \geq \lambda_{d+1}$ and $\frac{C_0\sigma_x^2(2+C_1)n\lambda_{d+1}}{n\tau} \leq 1$,

$$\frac{|B_{\text{in},12}|}{B_{\text{in},1}} \leq \frac{2C_0\sigma_x^2(2+C_1)\lambda_{d+1}}{\tau} (1 + 16(2C_0\sigma_x^2 + 1)(1 + C_1)\frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}) + 64\frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}.$$

Hence we have for $\tau \geq \lambda_{d+1}$,

$$\frac{\max\{B_{\text{in},1} - |B_{\text{in},12}|, 0\}}{B_{\text{in},1}} \geq \kappa_1(\tau),$$

where $\kappa_1 = \max\{1 - (\frac{2C_0\sigma_x^2(2+C_1)\lambda_{d+1}}{\tau} (1 + 16(2C_0\sigma_x^2 + 1)(1 + C_1)\frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}) + 64\frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}}, 0\}$.

Large TER. We substitute the bounds from Lemma S9(ii) into Lemma S8. Given Assumption 3 and 4(ii) and in the event $\Omega_1(\nu) \cap \Omega_4 \cap \Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$, we have for $\tau \geq 0$ and $\frac{\mu_1(X_{(d+1):p}X_{(d+1):p}^T)}{\mu_n(A_d)} \leq 1$,

$$\frac{|B_{\text{in},12}|}{B_{\text{in},1}} \leq \frac{2\mu_1(X_{(d+1):p}X_{(d+1):p}^T)}{\mu_n(A_d)} (1 + 112\frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}} + 64\frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}).$$

In the event $\Omega_6(\nu)$ for $0 < \nu < \frac{1}{4}$, from (S45) and (S50), we have

$$\begin{aligned} \mu_1(X_{d+1:p}X_{d+1:p}^T) &\leq (1 + \nu + \eta_2) \sum_{j>d} \lambda_j \leq 2 \sum_{j>d} \lambda_j \\ \mu_n(A_d) &\geq (1 - \nu - \eta_2) (\sum_{j>d} \lambda_j + n\tau) \geq \frac{1}{4} (\sum_{j>d} \lambda_j + n\tau). \end{aligned}$$

Then for $\tau \geq 0$ and $\frac{8(\sum_{j>d} \lambda_j)}{\tau + \frac{n}{\sum_{j>d} \lambda_j}} \leq 1$, we have

$$\frac{|B_{\text{in},12}|}{B_{\text{in},1}} \leq \frac{16(\frac{\sum_{j>d} \lambda_j}{n})}{\tau + \frac{n}{\sum_{j>d} \lambda_j}} (1 + 112\frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}} + 64\frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}).$$

Hence we have for $\tau \geq 0$,

$$\frac{\max\{B_{\text{in},1} - |B_{\text{in},12}|, 0\}}{B_{\text{in},1}} \geq \kappa_2(\tau),$$

where $\kappa_2(\tau) = \max\{1 - (16\frac{\lambda_{d+1}\frac{r_d(\Sigma)}{n}}{\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n}} (1 + 112\frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}) + 64\frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}}, 0\}$.

II.6 Useful identities and inequalities

Lemma S10 (Identities from Tsigler & Bartlett (2023)).

Let $\hat{\theta}(\tau, y) = X^T(XX^T + n\tau I_n)^{-1}y$ and $\hat{\theta}(\tau, y)^T = [\hat{\theta}(\tau, y)_{1:d}^T, \hat{\theta}(\lambda, y)_{d+1:p}^T]$. Then

(i)

$$\hat{\theta}(\tau, y)_{1:d} + X_{1:d}^T A_d^{-1} X_{1:d} \hat{\theta}(\tau, y)_{1:d} = X_{1:d}^T A_d^{-1} y,$$

(ii)

$$A^{-1}X_{1:d} = A_d^{-1}X_{1:d}(I_d + X_{1:d}^T A_d^{-1}X_{1:d})^{-1},$$

(iii)

$$(I_d + X_{1:d}^T A_d^{-1}X_{1:d})^{-1} = I_d - X_{1:d}^T A^{-1}X_{1:d}.$$

Proof.

Lemma S10(i) is from Section F in Tsigler & Bartlett (2023). Lemma S10(ii) is from H.2 in Tsigler & Bartlett (2023). To show Lemma S10(iii), from Lemma S10(ii),

$$A^{-1}X_{1:d} = A_d^{-1}X_{1:d}(I_d + X_{1:d}^T A_d^{-1}X_{1:d})^{-1}.$$

Then we have

$$\begin{aligned} X_{1:d}^T A^{-1}X_{1:d} &= X_{1:d}^T A_d^{-1}X_{1:d}(I_d + X_{1:d}^T A_d^{-1}X_{1:d})^{-1} \\ &= (X_{1:d}^T A_d^{-1}X_{1:d} + I_d - I_d)(I_d + X_{1:d}^T A_d^{-1}X_{1:d})^{-1} \\ &= I_d - (I_d + X_{1:d}^T A_d^{-1}X_{1:d})^{-1}, \end{aligned}$$

which gives

$$(I_d + X_{1:d}^T A_d^{-1}X_{1:d})^{-1} = I_d - X_{1:d}^T A^{-1}X_{1:d}.$$

□

Lemma S11 (Monotonicity of variance). *Denote by $V_{\text{out}}(\tau)$ the V_{out} in (5) with the ridge parameter τ . If $0 \leq \tau_1 \leq \tau_2$, then $V_{\text{out}}(\tau_2) \leq V_{\text{out}}(\tau_1)$.*

Proof. From the definition, we have

$$\begin{aligned} V_{\text{out}}(\tau_1) &= \sigma^2 \text{Tr}((n\tau_1 I_n + XX^T)^{-1} X \Sigma X^T (n\tau_1 I_n + XX^T)^{-1}) \\ &= \sigma^2 \text{Tr}(X \Sigma X^T (n\tau_1 I_n + XX^T)^{-2}) \\ &\geq \sigma^2 \text{Tr}(X \Sigma X^T (n\tau_2 I_n + XX^T)^{-2}) \\ &= \sigma^2 \text{Tr}((n\tau_2 I_n + XX^T)^{-1} X \Sigma X^T (n\tau_2 I_n + XX^T)^{-1}) \\ &= V_{\text{out}}(\tau_2). \end{aligned}$$

The inequality follows because $(n\tau_1 I_n + XX^T)^{-2} - (n\tau_2 I_n + XX^T)^{-2}$ is semi-positive definite. □

Lemma S12 (Weyl's inequality). *For $M, N, R \in \mathbb{C}^{n \times n}$, suppose that $M = N + R$, R is Hermitian matrices, and their respective eigenvalues are ordered as follows:*

$$(M) \quad \mu_1 \geq \mu_2 \geq \dots \geq \mu_n,$$

$$(N) \quad \nu_1 \geq \nu_2 \geq \dots \geq \nu_n,$$

$$(R) \quad \rho_1 \geq \rho_2 \geq \dots \geq \rho_n.$$

Then for $i = 1, 2, \dots, n$,

$$\nu_i + \rho_n \leq \mu_i \leq \nu_i + \rho_1.$$

Lemma S13 (Ruhe's trace inequality in Marshall et al. (2011)). *If U and V are $n \times n$ positive semidefinite Hermitian matrices, then*

$$\text{Tr}(UV) \geq \sum_{i=1}^n \lambda_i(U) \lambda_{n-i+1}(V)$$

In Lemma S14–S18 below, C_0 is an absolute constant which may vary from lemma to lemma. For simplicity, we treat C_0 as a common absolute constant, by taking the maximum of such constants from the individual lemmas.

Lemma S14 (Corollary 2.8 in Zajkowski (2020)). *Suppose that $z \in \mathbb{R}^p$, with $\text{Cov}(z, z) = I_p$, is a sub-gaussian vector with norm σ_x . Let $x = z \text{Diag}(\sqrt{\lambda_1}, \dots, \sqrt{\lambda_p})$. Then*

$$P(|\sum_{j=1}^p x_j^2 - \sum_{j=1}^p \lambda_j| \geq \sum_{j=1}^p \lambda_j \delta) \leq 2 \exp\{-\min\{r_0(\Sigma) \frac{\delta^2}{C_0^2 \sigma_x^4}, \sqrt{r_0(\Sigma)} \frac{\delta}{C_0 \sigma_x^2}\}\}.$$

That is, with probability at least $1 - 2 \exp\{-\min\{r_0(\Sigma) \frac{\delta^2}{C_0^2 \sigma_x^4}, \sqrt{r_0(\Sigma)} \frac{\delta}{C_0 \sigma_x^2}\}\}$,

$$(1 - \delta) \sum_{j=1}^p \lambda_j \leq \sum_{j=1}^p x_j^2 \leq (1 + \delta) \sum_{j=1}^p \lambda_j.$$

Lemma S15 (Lemma 23 in Tsigler & Bartlett (2023)). *Suppose that $z_1, \dots, z_n$ are independent sub-gaussian vectors in $\mathbb{R}^p$, each with sub-gaussian norm σ_x . Let $\Sigma = \text{Diag}(\lambda_1, \dots, \lambda_p)$ for some positive non-increasing sequence $\{\lambda_i\}_{i=1}^p$. Denote Z to be the matrix with rows $\{z_i \Sigma^{1/2}\}_{i=1}^n$ and $A = ZZ^T$. Denote also $\mathring{A}$ to be the matrix A with zeroed out diagonal elements: $\mathring{A}_{i,j} = (1 - \delta_{i,j}) A_{i,j}$, where $\delta_{i,j} = 0$ if $i \neq j$ or $\delta_{i,j} = 1$ if $i = j$. Then for any $t > 0$ with probability at least $1 - 4 \exp\{-t/C_0\}$,*

$$\|\mathring{A}\| \leq C_0 \sigma_x^2 \sqrt{(t+n)(\lambda_1^2(t+n) + \sum_{j=1}^p \lambda_j^2)}.$$

Lemma S16 (Lemma 24 in Tsigler & Bartlett (2023)). *In the same setting as Lemma S15, we have with probability at least $1 - 6\exp\{-t/C_0\}$,*

$$\|A\| \leq C_0 \sigma_x^2 (\lambda_1(t+n) + \sum_{j=1}^p \lambda_j).$$

Lemma S17 (Theorem 5.39 in Vershynin (2012)). *Let $d \leq n$ and $Z \in \mathbb{R}^{n \times d}$ whose rows z_i are independent sub-gaussian isotropic vectors in $\mathbb{R}^d$ with sub-gaussian norm σ_x . Then for $t \geq 0$, with probability at least $1 - 2\exp\{-\frac{t^2}{C_0^2 \sigma_x^4}\}$,*

$$\sqrt{n} - C_0 \sigma_x^2 \sqrt{d} - t \leq s_{\min}(Z) \leq s_{\max}(Z) \leq \sqrt{n} + C_0 \sigma_x^2 \sqrt{d} + t.$$

Lemma S18 (Lemma 21 in Tsigler & Bartlett (2023)). *Let $Z \in \mathbb{R}^{n \times p}$ whose rows z_i are independent isotropic sub-gaussian vectors in $\mathbb{R}^p$ with sub-gaussian norm σ_x . Let $\Sigma = \text{Diag}(\lambda_1, \dots, \lambda_p)$ for some positive non-increasing sequence $\{\lambda_i\}_{i=1}^p$. Then for any $t \in (0, n)$ with probability at least $1 - 2\exp\{-t/C_0\}$,*

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

III Proofs of additional results in Section 3

We provide proofs of Corollaries 1, 2, 4 and 5 which are re-stated below for convenience.

III.1 Sufficient and necessary conditions for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ and $\text{MSE}_{\text{in}} = O(\frac{d}{n})$

Corollary 1 (Conditions for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with small or moderate TER). *In the setting of Theorem 1, assume further that $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$.*

(i) *A sufficient condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability approaching 1 as $n \rightarrow \infty$ is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$ and the ridge parameter τ is chosen in the range $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_{d+1}$ if $r_d(\Sigma^2) \leq d$ or $A_0^{-1} \lambda_{d+1} \max\{\frac{1}{c} \sqrt{\frac{r_d(\Sigma^2)}{d}}, 1\} \leq \tau \leq A_0 \lambda_d \min\{c \sqrt{\frac{d}{n}}, 1\}$ if $r_d(\Sigma^2) > d$, where c is a constant satisfying $c \geq 1$ and $\frac{\lambda_{d+1}}{\lambda_d} \leq c \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$.*

(ii) *Suppose that $n \gg d$ and $r_d(\Sigma^2) \gg d$. Then a necessary condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$ and the ridge parameter τ is chosen in the range $\sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} \lesssim \tau \lesssim \sqrt{\frac{d}{n}} \lambda_d$.*

The sufficient and necessary conditions become matched, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$, if in the case where $n \gg d$ and $r_d(\Sigma^2) \gg d$ in addition to the assumptions stated.

Proof. From Theorem 1, for any $0 < \epsilon < 1$, the bounds in Theorem 1(i)(ii)(iii) hold with probability at least $1 - \epsilon$ for $n \geq N$ if N is large enough. From the bounds in Theorem 1(i)(ii)(iii), $\sigma^2 \asymp 1$ and

$\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, we have

$$\text{MSE}_{\text{out}} \gtrsim \frac{d}{n} + \frac{r_d(\Sigma^2)}{n}, \quad \text{for } \tau \leq A_0^{-1} \lambda_{d+1}, \quad (\text{S60})$$

$$\frac{\tau^2}{\lambda_d^2} + \frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \gtrsim \text{MSE}_{\text{out}} \gtrsim \frac{\tau^2}{\lambda_d^2} + \frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n}, \quad \text{for } A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d, \quad (\text{S61})$$

$$\text{MSE}_{\text{out}} \gtrsim 1, \quad \text{for } \tau \geq A_0 \lambda_d. \quad (\text{S62})$$

① Proof of Corollary 1(i):

Suppose $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$. Then there exists a constant $c \geq 1$, such that

$$\frac{\lambda_{d+1}}{\lambda_d} \leq c \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}. \quad (\text{S63})$$

Then we prove the sufficiency of the condition, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$, in two cases $d \geq r_d(\Sigma^2)$ and $d < r_d(\Sigma^2)$.

- If $d \geq r_d(\Sigma^2)$, from (S63), then

$$\frac{\lambda_{d+1}}{\lambda_d} \leq c \sqrt{\frac{d}{n}}. \quad (\text{S64})$$

If we let $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_{d+1}$, from (S64) and $d \geq r_d(\Sigma^2)$, then

$$\begin{aligned} \frac{\tau^2}{\lambda_d^2} &\leq A_0^2 \frac{\lambda_{d+1}^2}{\lambda_d^2} \leq c^2 A_0^2 \frac{d}{n}, \\ \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} &\leq A_0^2 \frac{d}{n}. \end{aligned}$$

From the upper bound in (S61), we have $\text{MSE}_{\text{out}} = O(\frac{d}{n})$.

- If $d < r_d(\Sigma^2)$, from (S63), then

$$\frac{\lambda_{d+1}}{\lambda_d} \leq c \frac{d}{\sqrt{n r_d(\Sigma^2)}}.$$

Let τ be in the range $A_0^{-1} \lambda_{d+1} \leq A_0^{-1} \lambda_{d+1} \max\{\frac{1}{c} \sqrt{\frac{r_d(\Sigma^2)}{d}}, 1\} \leq \tau \leq A_0 \lambda_d \min\{c \sqrt{\frac{d}{n}}, 1\} \leq A_0 \lambda_d$, then

$$\begin{aligned} \frac{\tau^2}{\lambda_d^2} &\leq c^2 A_0^2 \frac{d}{n}, \\ \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{\tau^2} &\leq c^2 A_0^2 \frac{d}{n}. \end{aligned}$$

From the upper bound in (S61), we have $\text{MSE}_{\text{out}} = O(\frac{d}{n})$.

In conclusion, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$ is a sufficient condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability approaching 1 as $n \rightarrow \infty$. The ridge parameter τ is chosen in the range $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_{d+1}$ if $r_d(\Sigma^2) \leq d$ or $A_0^{-1}\lambda_{d+1} \max\{\frac{1}{c}\sqrt{\frac{r_d(\Sigma^2)}{d}}, 1\} \leq \tau \leq A_0\lambda_d \min\{c\sqrt{\frac{d}{n}}, 1\}$ if $r_d(\Sigma^2) > d$, where c is a constant satisfying $c \geq 1$ and $\frac{\lambda_{d+1}}{\lambda_d} \leq c\sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma^2)}}\}$.

② Proof of Corollary 1(ii):

We first show that $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 only when $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$ by method of exclusion.

- If $\tau \geq A_0\lambda_d$, then from lower bound in (S62), we have $\text{MSE}_{\text{out}} \gtrsim 1$, which is contradictory to $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ and $n \gg d$.
- If $\tau \leq A_0^{-1}\lambda_{d+1}$, then from lower bound in (S60), we have $\text{MSE}_{\text{out}} \gtrsim \frac{r_d(\Sigma^2)}{n} \gg \frac{d}{n}$ (because $r(\Sigma^2) \gg d$), which is contradictory to $\text{MSE}_{\text{out}} = O(\frac{d}{n})$.

By excluding the above two possibilities, $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 only when $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$. From the lower bound in (S61), we have

$$\begin{aligned} \frac{\tau^2}{\lambda_d^2} &= O(\frac{d}{n}), \\ \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{\tau^2} &= O(\frac{d}{n}). \end{aligned}$$

That is,

$$\begin{aligned} \sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} &\lesssim \tau \lesssim \sqrt{\frac{d}{n}} \lambda_d, \\ \frac{\lambda_{d+1}}{\lambda_d} &\lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}. \end{aligned}$$

Hence a necessary condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \sqrt{\frac{d}{r_d(\Sigma^2)}}$ and $\sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} \lesssim \tau \lesssim \sqrt{\frac{d}{n}} \lambda_d$.

□

Corollary 2 (Conditions for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with small or moderate TER). *In the setting of Theorem 2, assume further that $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$.*

(i) *A sufficient condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability approaching 1 as $n \rightarrow \infty$ is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$ and the ridge parameter τ is chosen in the range $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_{d+1}$ if $r_d(\Sigma) \leq d$ or $A_0^{-1}\lambda_{d+1} \max\{\frac{1}{c}\sqrt{\frac{r_d(\Sigma)}{d}}, 1\} \leq \tau \leq A_0\lambda_d \min\{c\sqrt{\frac{d}{n}}, 1\}$ if $r_d(\Sigma) > d$, where c is a constant satisfying $c \geq 1$ and $\frac{\lambda_{d+1}}{\lambda_d} \leq c\sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$.*

(ii) Suppose that $n \gg d$, $\frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \gg 1$ and $64 \frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}} < 1$. Then a necessary condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$ and the ridge parameter τ is chosen in the range $\lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau \lesssim \lambda_d \sqrt{\frac{d}{n}}$.

The sufficient and necessary conditions become matched, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$, if in the case where $n \gg d$ and $r_d(\Sigma) \asymp n$ in addition to the assumptions stated.

Proof.

From Theorem 2, for $0 < \epsilon < 1$, Theorem 2(i)(ii)(iii) hold with probability at least $1 - \epsilon$ for $n \geq N$ if N is large enough. From the bounds in Theorem 2(i)(ii)(iii), $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, we have

$$\text{MSE}_{\text{in}} \gtrsim \frac{d}{n} + \frac{r_d^2(\Sigma)}{n^2}, \quad \text{for } \tau \leq A_0^{-1} \lambda_{d+1}, \quad (\text{S65})$$

$$\frac{\tau^2}{\lambda_d^2} + \frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma)}{n} \gtrsim \text{MSE}_{\text{in}} \gtrsim \kappa_1(\tau) \frac{\tau^2}{\lambda_d^2} + \frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2}, \quad \text{for } A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d, \quad (\text{S66})$$

$$\text{MSE}_{\text{in}} \gtrsim \kappa_1(\tau) + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2}, \quad \text{for } \tau \geq A_0 \lambda_d. \quad (\text{S67})$$

① Proof of Corollary 2(i):

Suppose $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$, then there exists a constant c such that $c \geq 1$ and

$$\frac{\lambda_{d+1}}{\lambda_d} \leq c \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}. \quad (\text{S68})$$

We prove the sufficiency of the condition, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$, in two cases, $d \geq r_d(\Sigma)$ and $d < r_d(\Sigma)$.

- If $d \geq r_d(\Sigma)$, from (S68), we have

$$\frac{\lambda_{d+1}}{\lambda_d} \leq c \sqrt{\frac{d}{n}}. \quad (\text{S69})$$

If we let $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_{d+1}$, from (S69) and $d \geq r_d(\Sigma)$, we have

$$\begin{aligned} \frac{\tau^2}{\lambda_d^2} &\leq A_0^2 \frac{\lambda_{d+1}^2}{\lambda_d^2} \leq c^2 A_0^2 \frac{d}{n}, \\ \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma)}{n} &\leq A_0^2 \frac{r_d(\Sigma)}{n} \leq A_0^2 \frac{d}{n}. \end{aligned}$$

Then from the upper bound in (S66), we have $\text{MSE}_{\text{in}} = O(\frac{d}{n})$.

- If $d < r_d(\Sigma)$, from (S68), we have

$$\frac{\lambda_{d+1}}{\lambda_d} \leq c \frac{d}{\sqrt{n r_d(\Sigma)}}.$$

Let τ be in the range $A_0^{-1}\lambda_{d+1} \leq A_0^{-1}\lambda_{d+1} \max\{\frac{1}{c}\sqrt{\frac{r_d(\Sigma)}{d}}, 1\} \leq \tau \leq A_0\lambda_d \min\{c\sqrt{\frac{d}{n}}, 1\} \leq A_0\lambda_d$, then we have

$$\begin{aligned} \frac{\tau^2}{\lambda_d^2} &\leq c^2 A_0^2 \frac{d}{n}, \\ \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma)}{n} &\leq c^2 A_0^2 \frac{d}{n}. \end{aligned}$$

From the upper bound in (S66), we have $\text{MSE}_{\text{in}} = O(\frac{d}{n})$.

In conclusion, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$ is a sufficient condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability approaching 1 as $n \rightarrow \infty$. The ridge parameter τ is chosen in the range $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_{d+1}$ if $r_d(\Sigma) \leq d$ or $A_0^{-1}\lambda_{d+1} \max\{\frac{1}{c}\sqrt{\frac{r_d(\Sigma)}{d}}, 1\} \leq \tau \leq A_0\lambda_d \min\{c\sqrt{\frac{d}{n}}, 1\}$ if $r_d(\Sigma) > d$, where c is a constant satisfying $c \geq 1$ and $\frac{\lambda_{d+1}}{\lambda_d} \leq c\sqrt{\frac{d}{n}} \min\{1, \sqrt{\frac{d}{r_d(\Sigma)}}\}$.

② Proof of Corollary 2(ii):

First, we point out that with $64 \frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}} < 1$, we have $\kappa_1(\tau) \gtrsim 1$ if $\frac{\lambda_{d+1}}{\tau} \ll 1$. Then we show that $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 only when $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$ by method of exclusion.

- If $\tau \leq A_0^{-1}\lambda_{d+1}$, from lower bound in (S65) and $\frac{r_d(\Sigma)}{n} \gg \sqrt{\frac{d}{n}}$, we have $\text{MSE}_{\text{in}} \gg \frac{d}{n}$, which is contradictory to $\text{MSE}_{\text{in}} = O(\frac{d}{n})$.
- If $\tau \geq A_0\lambda_d$ and $\text{MSE}_{\text{in}} = O(\frac{d}{n})$, from lower bound of (S67), we have

$$\frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2} = O(\frac{d}{n}),$$

hence there exists a constant $c \geq 1$ such that

$$\frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2} \leq c \frac{d}{n}.$$

With $\frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \gg 1$, we have

$$\frac{\lambda_{d+1}}{\tau} \leq \sqrt{c} \sqrt{\frac{d}{n}} \frac{n}{r_d(\Sigma)} \ll 1.$$

Hence we have

$$\begin{aligned} \kappa_1(\tau) &\gtrsim 1 \\ \implies \text{MSE}_{\text{in}} &\gtrsim 1, \quad (\text{from (S67)}) \end{aligned}$$

which is contradictory to $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ if $n \gg d$.

By excluding the above two possibilities, we know that $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 only when $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$. From lower bound in (S66) and $\text{MSE}_{\text{in}} = O(\frac{d}{n})$, we have

$$\frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2} = O(\frac{d}{n}).$$

Similarly to the derivation in the case of $\tau \geq A_0\lambda_d$, we have $\kappa_1(\tau) \gtrsim 1$. Hence $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ only when $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$ and $\kappa_1(\tau) \gtrsim 1$. From the lower bound in (S66), we have $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ only when $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$ and

$$\begin{aligned} \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d^2(\Sigma)}{n^2} &= O(\frac{d}{n}), \\ \frac{\tau^2}{\lambda_d^2} &= O(\frac{d}{n}). \end{aligned}$$

That is,

$$\begin{aligned} A_0^{-1}\lambda_{d+1} &\ll \lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau \lesssim \lambda_d \sqrt{\frac{d}{n}} \ll A_0\lambda_d, \\ \frac{\lambda_{d+1}}{\lambda_d} &\lesssim \frac{d}{r_d(\Sigma)}. \end{aligned}$$

Hence a necessary condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$ and $\lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau \lesssim \lambda_d \sqrt{\frac{d}{n}}$. $\square$

Corollary 4 (Conditions for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with large TER). *In the setting of Theorem 3, assume further that $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$.*

(i) *A sufficient condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability approaching 1 as $n \rightarrow \infty$ is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$ and the ridge parameter τ is chosen satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$ or $\sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$.*

(ii) *Suppose that $n \gg d$. Then a necessary condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 is that $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$ and τ is chosen satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$ or $\sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$. The sufficient and necessary conditions become matched, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$, if in the case where $n \gg d$ in addition to the assumptions stated.*

Proof.

From Theorem 3, for $0 < \epsilon < 1$, Theorem 3(i)(ii) hold with probability at least $1 - \epsilon$ for $n \geq N$ if N is large enough. Because A_0 can be any unbounded positive value in Theorem 3, from the

bounds in Theorem 3(i)(ii), $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, we have

$$\begin{aligned} \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} &\gtrsim \text{MSE}_{\text{out}} \gtrsim \\ \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n}, &\text{ for } \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d, \end{aligned} \quad (\text{S70})$$

$$\text{MSE}_{\text{out}} \gtrsim 1, \quad \text{for } \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_d. \quad (\text{S71})$$

① Proof of Corollary 4(i):

We prove the sufficiency of the condition, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$, in two cases $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$ and $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$.

- If $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$, from $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$, we have

$$\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{\sqrt{nd}}{r_d(\Sigma)}.$$

Because $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{\sqrt{nd}}{r_d(\Sigma)}$, we have $\lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$. Then we can choose non-negative τ such that $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ and

$$\frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} \lesssim \frac{d}{n}. \quad (\text{S72})$$

Moreover, with $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$, we have $\frac{nr_d(\Sigma^2)}{r_d^2(\Sigma)} \leq \frac{d}{n}$. Hence we have for $\tau \geq 0$,

$$\begin{aligned} \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} &\leq \frac{n^2}{r_d^2(\Sigma)} \frac{r_d(\Sigma^2)}{n} \\ &\leq \frac{d}{n}. \end{aligned} \quad (\text{S73})$$

Then from (S72)–(S73) and the upper bound in (S70), we have $\text{MSE}_{\text{out}} = O(\frac{d}{n})$.

- If $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$, from $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$, we have

$$\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{\sqrt{n} r_d(\Sigma^2)}. \quad (\text{S74})$$

With $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$, we have

$$\sqrt{r_d(\Sigma^2)} > \frac{\sqrt{dr_d(\Sigma)}}{n}. \quad (\text{S75})$$

Now we prove $\sqrt{\frac{d}{n}} \lambda_d \gtrsim \lambda_{d+1} \frac{r_d(\Sigma)}{n}$. From (S74), we have

$$\sqrt{\frac{d}{n}} \lambda_d \gtrsim \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{d}}.$$

Further with (S75), we have

$$\begin{aligned}\sqrt{\frac{d}{n}}\lambda_d &\gtrsim \lambda_{d+1}\sqrt{\frac{r_d(\Sigma^2)}{d}}, \\ &> \lambda_{d+1}\frac{r_d(\Sigma)}{n}.\end{aligned}\tag{S76}$$

From (S76), we let $\tau \geq 0$ such that

$$\sqrt{\frac{r_d(\Sigma^2)}{d}}\lambda_{d+1} \lesssim \tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}}\lambda_d \leq \lambda_d.\tag{S77}$$

Then we have

$$\begin{aligned}\frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{(\tau + \frac{\sum_{j>d}\lambda_j}{n})^2} &\lesssim \frac{d}{n}, \\ \frac{(\tau + \frac{\sum_{j>d}\lambda_j}{n})^2}{\lambda_d^2} &\lesssim \frac{d}{n}.\end{aligned}$$

From upper bound in (S70), we have $\text{MSE}_{\text{out}} = O(\frac{d}{n})$.

In conclusion, we have $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a high probability approaching 1 if $n \rightarrow \infty$ given $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$. The ridge parameter τ is chosen such that $\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}}\lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$ or $\sqrt{\frac{r_d(\Sigma^2)}{d}}\lambda_{d+1} \lesssim \tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}}\lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$.

② Proof of Corollary 4(ii):

We first show that $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 only when $\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \lambda_d$ by method of exclusion.

If $\tau \gtrsim \lambda_d$, from lower bound (S71), we have $\text{MSE}_{\text{out}} \gtrsim 1$, which is contradictory to $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ and $n \gg d$.

By excluding the above possibility, $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 only when $\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \lambda_d$. Then we prove the necessity of the condition in two cases, $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$ and $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > 1$.

- If $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq 1$, from $\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \lambda_d$ and lower bound in (S70), we have

$$\frac{(\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} = O(\frac{d}{n}).$$

Then we have

$$\tau + \lambda_{d+1}\frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}}\lambda_d,$$

and

$$\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \frac{\lambda_{d+1}}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}}. \quad (\text{S78})$$

Obviously, we have

$$\sqrt{\frac{d}{n}} \frac{\lambda_{d+1}}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \leq \frac{\sqrt{nd}}{r_d(\Sigma)}. \quad (\text{S79})$$

Combining (S78) and (S79), we have

$$\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{\sqrt{nd}}{r_d(\Sigma)}.$$

- If $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{d}r_d(\Sigma)} > 1$, from $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$ and the lower bound in (S70), we have

$$\begin{aligned} \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{(\tau + \frac{\sum_{j>d} \lambda_j}{n})^2} &= O\left(\frac{d}{n}\right), \\ \frac{(\tau + \frac{\sum_{j>d} \lambda_j}{n})^2}{\lambda_d^2} &= O\left(\frac{d}{n}\right). \end{aligned}$$

Hence

$$\begin{aligned} \sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} &\lesssim \tau + \frac{\sum_{j>d} \lambda_j}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d, \\ \frac{\lambda_{d+1}}{\lambda_d} &\lesssim \frac{d}{\sqrt{n} r_d(\Sigma^2)}. \end{aligned}$$

In conclusion, a necessary condition for $\text{MSE}_{\text{out}} = O(\frac{d}{n})$ with a probability bounded away from 0 is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \sqrt{\frac{d}{n}} \min\{\sqrt{\frac{d}{r_d(\Sigma^2)}}, \frac{n}{r_d(\Sigma)}\}$ and the ridge parameter τ is chosen such that $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{d}r_d(\Sigma)} \leq 1$ or $\sqrt{\frac{r_d(\Sigma^2)}{d}} \lambda_{d+1} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\frac{d}{n}} \lambda_d$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{d}r_d(\Sigma)} > 1$. $\square$

Corollary 5 (Conditions for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with large TER). *In the setting of Theorem 4, assume further that $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$.*

(i) *A sufficient condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability approaching to 1 as $n \rightarrow \infty$ is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$ and the ridge parameter τ is chosen such that $\lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d \sqrt{\frac{d}{n}}$.*

(ii) *Suppose that $n \gg d$ and $64 \frac{\sqrt{\delta_2}}{1-\sqrt{\delta_2}} < 1$. Then a necessary condition for $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 is $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$ and the ridge parameter τ is chosen in the range $\lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d \sqrt{\frac{d}{n}}$.*

The sufficient and necessary conditions become matched, $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$, in the case where $n \gg d$ in addition to the assumptions stated.

Proof.

From Theorem 4, for $0 < \epsilon < 1$, Theorem 4(i)(ii) hold with probability at least $1 - \epsilon$ for $n \geq N$ if N is large enough. Because A_0 can be any unbounded positive value in Theorem 4, from the bounds in Theorem 4(i)(ii), $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, we have

$$\begin{aligned} \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2} &\gtrsim \text{MSE}_{\text{in}} \gtrsim \\ \kappa_2(\tau) \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2}, &\text{ for } \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d, \end{aligned} \quad (\text{S80})$$

$$\text{MSE}_{\text{in}} \gtrsim \kappa_2(\tau) + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n}, \quad \text{ for } \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_d. \quad (\text{S81})$$

① Proof of Corollary 5(i):

If $\frac{\lambda_{d+1}}{\lambda_d} \lesssim \frac{d}{r_d(\Sigma)}$, we have $\lambda_d \sqrt{\frac{d}{n}} \gtrsim \lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}}$. Hence we can choose $\tau \geq 0$ such that

$$\lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d \sqrt{\frac{d}{n}}. \quad (\text{S82})$$

Then

$$\begin{aligned} \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} &= O\left(\frac{d}{n}\right), \\ \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} &= O\left(\frac{d}{n}\right). \end{aligned}$$

From the upper bound in (S80), we have $\text{MSE}_{\text{in}} = O(\frac{d}{n})$.

② Proof of Corollary 5(ii):

We point out that with $64 \frac{\sqrt{\delta_2}}{1-\sqrt{\delta_2}} < 1$, we have $\kappa_2(\tau) \gtrsim 1$ if $\frac{\lambda_{d+1} \frac{r_d(\Sigma)}{n}}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \ll 1$. If $\text{MSE}_{\text{in}} = O(\frac{d}{n})$, from lower bound in (S80)–(S81), we have

$$\frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} = O\left(\frac{d}{n}\right).$$

With $n \gg d$, we have

$$\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}.$$

Equivalently, we have

$$\frac{\lambda_{d+1} \frac{r_d(\Sigma)}{n}}{\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \ll 1,$$

then

$$\kappa_2(\tau) \gtrsim 1.$$

Hence $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ only when $\kappa_2(\tau) \gtrsim 1$. Then we show that $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 only when $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$ by method of exclusion.

If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_d$, from the lower bound in (S81), we have

$$\text{MSE}_{\text{in}} \gtrsim \kappa_2(\tau) \gtrsim 1,$$

which is contradictory to $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ and $n \gg d$.

By excluding the above possibility, $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ with a probability bounded away from 0 only when $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$. From the lower bound in (S80), $\text{MSE}_{\text{in}} = O(\frac{d}{n})$ only when

$$\begin{aligned} \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} &= O(\frac{d}{n}), \\ \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} &= O(\frac{d}{n}). \end{aligned}$$

Then

$$\begin{aligned} \lambda_{d+1} \frac{r_d(\Sigma)}{n} \sqrt{\frac{n}{d}} &\lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d \sqrt{\frac{d}{n}}, \\ \frac{\lambda_{d+1}}{\lambda_d} &\lesssim \frac{d}{r_d(\Sigma)}. \end{aligned}$$

□

III.2 Out-sample and in-sample errors with optimal ridge parameters

Corollary 3 (Optimal error orders with small or moderate TER). *Suppose that Assumption 1, 2 and 4(i) are satisfied and further $\sigma^2 \asymp 1$, $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, $r_d(\Sigma) \asymp n$, $\lambda_d \gtrsim \lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}}$, $\lambda_d \gg \lambda_{d+1}$, and $64 \frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}} < 1$. Then*

(i) $\text{MSE}_{\text{out}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{d}{n}\}$ with a probability approaching to 1 and the optimal τ is chosen as $\tau = \sqrt{\lambda_d \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{n}}} \min\{\sqrt{c A_0^{-2}}, \frac{A_0 \lambda_d}{\sqrt{\lambda_d \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{n}}}}\}$ where c is a constant satisfying

$$\lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}} \leq c \lambda_d.$$

(ii) $\text{MSE}_{\text{in}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d}, \frac{d}{n}\}$ with a probability approaching to 1 and the optimal τ is chosen satisfying $\tau \asymp \sqrt{\lambda_{d+1} \lambda_d}$.

Therefore $\text{MSE}_{\text{out}}^* \lesssim \text{MSE}_{\text{in}}^*$ with a probability approaching to 1, by noting $r_d(\Sigma^2) \leq r_d(\Sigma) \asymp n$.

Proof.

Proof of $\text{MSE}_{\text{out}}^*$: For $0 < \epsilon < 1$, Theorem 1(i)(ii)(iii) hold with probability at least $1 - \epsilon$ for $n \geq N$ if N is large enough under Assumption 1, 2 and 4(i). From the bounds in Theorem

1(i)(ii)(iii), $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, (S60)–(S62) hold. Denote MSE_{out} with optimal τ chosen from $\tau \leq A_0^{-1} \lambda_{d+1}$ as $\text{MSE}_{1,\text{out}}^*$, MSE_{out} with optimal τ chosen from $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$ as $\text{MSE}_{2,\text{out}}^*$ and MSE_{out} with optimal τ chosen from $\tau \geq A_0 \lambda_d$ as $\text{MSE}_{3,\text{out}}^*$. Then we give orders of bounds of $\text{MSE}_{1,\text{out}}^*$, $\text{MSE}_{2,\text{out}}^*$ and $\text{MSE}_{3,\text{out}}^*$ and give the order of $\text{MSE}_{\text{out}}^*$ by the integrating the orders of bounds of $\text{MSE}_{1,\text{out}}^*$, $\text{MSE}_{2,\text{out}}^*$ and $\text{MSE}_{3,\text{out}}^*$.

- If $\tau \leq A_0^{-1} \lambda_{d+1}$, from lower bound in (S60), we have $\text{MSE}_{1,\text{out}}^* \gtrsim (\frac{d}{n} + \frac{r_d(\Sigma^2)}{n}) \asymp \max\{\frac{d}{n}, \frac{r_d(\Sigma^2)}{n}\}$. Moreover, we have $\max\{\frac{d}{n}, \frac{r_d(\Sigma^2)}{n}\} \gtrsim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}\}$ because $\lambda_d \gtrsim \lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}}$.
- If $\tau \geq A_0 \lambda_d$, from lower bound in (S62), we have $\text{MSE}_{3,\text{out}}^* \gtrsim 1$.
- If $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$, from (S61), we have $\text{MSE}_{\text{out}} \asymp \frac{\tau^2}{\lambda_d^2} + \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{\tau^2} + \frac{d}{n}$ and $\frac{\tau^2}{\lambda_d^2} + \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{\tau^2} \geq \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}$. Let

$$\tau = \sqrt{\lambda_d \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{n}}} \min\left\{\sqrt{c A_0^{-2}}, \frac{A_0 \lambda_d}{\sqrt{\lambda_d \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{n}}}}\right\},$$

where c is a constant satisfying $\lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}} \leq c \lambda_d$, then $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$ because $\lambda_{d+1} \sqrt{\frac{n}{r_d(\Sigma^2)}} \leq c \lambda_d$. With the above choice of τ , we have $\text{MSE}_{\text{out}} \asymp \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}$. Therefore, if $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$, we have $\text{MSE}_{2,\text{out}}^* \asymp \frac{d}{n} + \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}} \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}\}$.

By integrating the orders of bounds of $\text{MSE}_{1,\text{out}}^*$, $\text{MSE}_{2,\text{out}}^*$ and $\text{MSE}_{3,\text{out}}^*$, we have $\text{MSE}_{\text{out}}^* \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}\}$.

Proof of MSE_{in}^* : For $0 < \epsilon < 1$, Theorem 2(i)(ii)(iii) hold with probability at least $1 - \epsilon$ for $n \geq N$ if N is large enough under Assumption 1, 2 and 4(i). From the bounds in Theorem 2(i)(ii)(iii), $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, (S65)–(S67) hold. Denote MSE_{in} with optimal τ chosen from $\tau \leq A_0^{-1} \lambda_{d+1}$ as $\text{MSE}_{1,\text{in}}^*$ and MSE_{in} with optimal τ chosen from $A_0^{-1} \lambda_{d+1} \leq \tau \leq A_0 \lambda_d$ as $\text{MSE}_{2,\text{in}}^*$ and MSE_{in} with optimal τ chosen from $\tau \geq A_0 \lambda_d$ as $\text{MSE}_{3,\text{in}}^*$. Then we give orders of bounds of $\text{MSE}_{1,\text{in}}^*$, $\text{MSE}_{2,\text{in}}^*$ and $\text{MSE}_{3,\text{in}}^*$ and give the order of MSE_{in}^* by the integrating the orders of bounds of $\text{MSE}_{1,\text{in}}^*$, $\text{MSE}_{2,\text{in}}^*$ and $\text{MSE}_{3,\text{in}}^*$.

- If $\tau \leq A_0^{-1} \lambda_{d+1}$, from lower bound in (S65) and $r_d(\Sigma) \asymp n$, we have

$$\text{MSE}_{1,\text{in}}^* \gtrsim \frac{d}{n} + \frac{r_d^2(\Sigma)}{n^2} \asymp 1.$$

- If $\tau \geq A_0 \lambda_d$, then $\tau \gg \lambda_{d+1}$ because $\lambda_d \gg \lambda_{d+1}$. Further with $64 \frac{\sqrt{\delta_1}}{1 - \sqrt{\delta_1}} < 1$, we have $\kappa_1(\tau) \gtrsim 1$. Then from the lower bound in (S67), we have for $\tau \geq A_0 \lambda_d$,

$$\text{MSE}_{3,\text{in}}^* \gtrsim \kappa_1(\tau) \gtrsim 1.$$

- If $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$, we first show that $\text{MSE}_{2,\text{in}}^* \lesssim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$ and then show that $\text{MSE}_{2,\text{in}}^* \gtrsim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$. These give $\text{MSE}_{2,\text{in}}^* \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$.

We first show that $\text{MSE}_{2,\text{in}}^* \lesssim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$. From the upper bound in (S66), we have $\text{MSE}_{\text{in}} \lesssim \frac{\tau^2}{\lambda_d^2} + \frac{\lambda_{d+1}^2}{\tau^2} + \frac{d}{n}$. Let $\tau \asymp \sqrt{\lambda_{d+1}\lambda_d}$, then $A_0^{-1}\lambda_{d+1} \ll \tau \ll A_0\lambda_d$ and $\frac{\tau^2}{\lambda_d^2} + \frac{\lambda_{d+1}^2}{\tau^2} = \frac{\lambda_{d+1}}{\lambda_d}$. Hence we have $\text{MSE}_{\text{in}} \lesssim \frac{\lambda_{d+1}}{\lambda_d} + \frac{d}{n} \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$.

Next we show that $\text{MSE}_{2,\text{in}}^* \gtrsim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$. From lower bound in (S66), we have $\text{MSE}_{\text{in}} \gtrsim \kappa_1(\tau) \frac{\tau^2}{\lambda_d^2} + \frac{\lambda_{d+1}^2}{\tau^2} + \frac{d}{n}$. Then we show that $\kappa_1(\tau) \frac{\tau^2}{\lambda_d^2} + \frac{\lambda_{d+1}^2}{\tau^2} \gtrsim \frac{\lambda_{d+1}}{\lambda_d}$ for $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$. For $A_0^{-1}\lambda_{d+1} \leq \tau \lesssim \sqrt{\lambda_{d+1}\lambda_d}$, we have

$$\kappa_1(\tau) \frac{\tau^2}{\lambda_d^2} + \frac{\lambda_{d+1}^2}{\tau^2} \geq \frac{\lambda_{d+1}^2}{\tau^2} \geq \frac{\lambda_{d+1}}{\lambda_d}.$$

For $\sqrt{\lambda_{d+1}\lambda_d} \lesssim \tau \leq A_0\lambda_d$, from $\lambda_d \gg \lambda_{d+1}$, we have $\tau \gg \lambda_{d+1}$. Further with $64 \frac{\sqrt{\delta_1}}{1-\sqrt{\delta_1}} < 1$, we have $\kappa_1(\tau) \gtrsim 1$, and then

$$\kappa_1(\tau) \frac{\tau^2}{\lambda_d^2} + \frac{\lambda_{d+1}^2}{\tau^2} \geq \kappa_1(\tau) \frac{\tau^2}{\lambda_d^2} \gtrsim \frac{\lambda_{d+1}}{\lambda_d}.$$

Hence we have $\kappa_1(\tau) \frac{\tau^2}{\lambda_d^2} + \frac{\lambda_{d+1}^2}{\tau^2} \gtrsim \frac{\lambda_{d+1}}{\lambda_d}$ for $A_0^{-1}\lambda_{d+1} \leq \tau \leq A_0\lambda_d$. That is, $\text{MSE}_{2,\text{in}}^* \gtrsim \frac{d}{n} + \frac{\lambda_{d+1}}{\lambda_d} \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$.

From $\text{MSE}_{2,\text{in}}^* \lesssim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$ and $\text{MSE}_{2,\text{in}}^* \gtrsim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$, we have $\text{MSE}_{2,\text{in}}^* \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$.

By integrating the orders of bounds of $\text{MSE}_{1,\text{in}}^*$, $\text{MSE}_{2,\text{in}}^*$ and $\text{MSE}_{3,\text{in}}^*$, we have $\text{MSE}_{\text{in}}^* \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d}\}$.

□

Corollary 6 (Optimal error orders with large TER). *Suppose that Assumption 1, 3 and 4(ii) are satisfied, and further $\sigma^2 \asymp 1$, $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ and $64 \frac{\sqrt{\delta_2}}{1-\sqrt{\delta_2}} < 1$. Then*

(i) *The order of $\text{MSE}_{\text{out}}^*$ is $\max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2}, \frac{d}{n}\}$ with a probability approaching to 1 and the optimal τ is chosen as $\tau = 0$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ or satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \asymp \sqrt{\lambda_d \lambda_{d+1} \sqrt{\frac{r_d(\Sigma^2)}{n}}}$ if $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$.*

(ii) *The order of MSE_{in}^* is $\max\{\frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}, \frac{d}{n}\}$ with a probability approaching to 1 and the optimal τ is chosen satisfying $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \asymp \sqrt{\lambda_d \lambda_{d+1} \frac{r_d(\Sigma)}{n}}$.*

Therefore $\text{MSE}_{\text{out}}^ \lesssim \text{MSE}_{\text{in}}^*$ with a probability approaching to 1 because $\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}} \lesssim \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ by noting $r_d(\Sigma^2) \leq r_d(\Sigma)$ and $r_d(\Sigma) \gtrsim n$ (by Assumption 3) and $\frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2} \lesssim \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ by noting $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$.*

Proof.

Proof of $\text{MSE}_{\text{out}}^*$: For $0 < \epsilon < 1$, Theorem 3(i)(ii) hold with probability at least $1 - \epsilon$ for $n \geq N$

if N is large enough under Assumption 1, 3 and 4(ii). Because A_0 can be any unbounded positive value in Theorem 3, from the bounds in Theorem 3(i)(ii), $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp 1$, we have (S70)–(S71) hold. Denote MSE_{out} with optimal τ chosen from $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$ as $\text{MSE}_{1,\text{out}}^*$ and MSE_{out} with optimal τ chosen from $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_d$ as $\text{MSE}_{2,\text{out}}^*$. Then we give orders of bounds of $\text{MSE}_{1,\text{out}}^*$ and $\text{MSE}_{2,\text{out}}^*$ and give the order of $\text{MSE}_{\text{out}}^*$ by the integrating the orders of bounds of $\text{MSE}_{1,\text{out}}^*$ and $\text{MSE}_{2,\text{out}}^*$.

- If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_d$, from lower bound in (S71), we have $\text{MSE}_{2,\text{out}}^* \gtrsim 1$.
- If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$, from (S70), we have $\text{MSE}_{\text{out}} \asymp \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} + \frac{d}{n}$. Then we discuss the order of $\text{MSE}_{1,\text{out}}^*$ in two cases, $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ and $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} > \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$.
If $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$, we have

$$\frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \leq \frac{nr_d(\Sigma^2)}{r_d(\Sigma)^2} \leq \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2} \leq \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2}.$$

Hence

$$\begin{aligned} \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} &\asymp \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} \\ &\geq \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2}. \end{aligned}$$

If we let $\tau = 0$, we have

$$\frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \asymp \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2}.$$

If $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \geq \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$, we have

$$\frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \geq \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}.$$

We also have

$$\frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{r_d(\Sigma^2)}{n} \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \asymp \frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}},$$

when $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \asymp \sqrt{\lambda_d \lambda_{d+1}} \sqrt{\frac{r_d(\Sigma^2)}{n}} \lesssim \lambda_d$.

By combining two cases, $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \leq \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ and $\frac{n\sqrt{r_d(\Sigma^2)}}{\sqrt{dr_d(\Sigma)}} \geq \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$, we have $\text{MSE}_{1,\text{out}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2}, \frac{d}{n}\}$.

By integrating the order of bounds of $\text{MSE}_{1,\text{out}}^*$ and $\text{MSE}_{2,\text{out}}^*$, we have $\text{MSE}_{\text{out}}^* \asymp \max\{\frac{\lambda_{d+1}}{\lambda_d} \sqrt{\frac{r_d(\Sigma^2)}{n}}, \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d(\Sigma)^2}{n^2}, \frac{d}{n}\}$.

Proof of MSE_{in}^* : For $0 < \epsilon < 1$, we have Theorem 4(i)(ii) hold with probability at least $1 - \epsilon$ for $n \geq N$ if N is large enough under Assumption 1, 3 and 4(ii). Because A_0 can be any unbounded positive value in Theorem 4, from the bounds in Theorem 4(i)(ii), $\sigma^2 \asymp 1$ and $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \asymp 1$, we have (S80)–(S81) hold. Denote MSE_{in} with optimal τ chosen from $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$ as $\text{MSE}_{1,\text{in}}^*$ and MSE_{in} with optimal τ chosen from $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_d$ as $\text{MSE}_{2,\text{in}}^*$. Then we give orders of bounds of $\text{MSE}_{1,\text{in}}^*$ and $\text{MSE}_{2,\text{in}}^*$ and give the order of MSE_{in}^* by the integrating the orders of bounds of $\text{MSE}_{1,\text{in}}^*$ and $\text{MSE}_{2,\text{in}}^*$.

- If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gtrsim \lambda_d$, then $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ because $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$. Further with $64 \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}} < 1$, we have $\kappa_2(\tau) \gtrsim 1$. From the lower bound in (S81), we have $\text{MSE}_{2,\text{in}}^* \gtrsim \kappa_2(\tau) \gtrsim 1$.
- If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$, we first show that $\text{MSE}_{1,\text{in}}^* \lesssim \max\{\frac{d}{n} + \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$ and then show that $\text{MSE}_{1,\text{in}}^* \gtrsim \max\{\frac{d}{n} + \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$, which gives the order of $\text{MSE}_{1,\text{in}}^*$.

We first show that $\text{MSE}_{1,\text{in}}^* \lesssim \max\{\frac{d}{n} + \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$. From the upper bound in (S80), we have $\text{MSE}_{\text{in}} \lesssim \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} + \frac{d}{n}$. Note that

$$\frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \sum_{j>d} \frac{r_d(\Sigma)}{n})^2} \geq \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$$

and the equality can be achieved when $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \asymp \sqrt{\lambda_d \lambda_{d+1} \frac{r_d(\Sigma)}{n}} \lesssim \lambda_d$. Hence we have $\text{MSE}_{1,\text{in}}^* \lesssim \frac{d}{n} + \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n} \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$.

Next we show that $\text{MSE}_{1,\text{in}}^* \gtrsim \max\{\frac{d}{n} + \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$. From the lower bound in (S80), we have $\text{MSE}_{\text{in}} \gtrsim \kappa_2(\tau) \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} + \frac{d}{n}$. Then we show $\kappa_2(\tau) \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \gtrsim \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ if $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$. If $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \sqrt{\lambda_{d+1} \lambda_d \frac{r_d(\Sigma)}{n}}$, we have

$$\kappa_2(\tau) \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \geq \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \gtrsim \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}.$$

If $\sqrt{\lambda_{d+1} \lambda_d \frac{r_d(\Sigma)}{n}} \lesssim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$, we have $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ because $\lambda_d \gg \lambda_{d+1} \frac{r_d(\Sigma)}{n}$. Further with $64 \frac{\sqrt{\delta_2}}{1 - \sqrt{\delta_2}} < 1$, we have $\kappa_2(\tau) \gtrsim 1$. Then

$$\kappa_2(\tau) \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \geq \kappa_2(\tau) \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} \gtrsim \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}.$$

Hence we have $\kappa_2(\tau) \frac{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{\lambda_d^2} + \frac{(\lambda_{d+1} \frac{r_d(\Sigma)}{n})^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \gtrsim \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}$ if $\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \lesssim \lambda_d$ and $\text{MSE}_{1,\text{in}}^* \gtrsim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$.

From $\text{MSE}_{1,\text{in}}^* \lesssim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$ and $\text{MSE}_{1,\text{in}}^* \gtrsim \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$, we have $\text{MSE}_{1,\text{out}}^* \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$.

By integrating the orders of bounds of $\text{MSE}_{1,\text{in}}^*$ and $\text{MSE}_{2,\text{in}}^*$, we have $\text{MSE}_{\text{in}}^* \asymp \max\{\frac{d}{n}, \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}\}$.

□

IV Error approximation formulas

IV.1 Convergence of in-sample error approximation formulas

We provide a proof of Theorem 5 in Section 4.1, which is re-stated below for convenience.

Theorem 5 (Convergence of in-sample error approximation formulas). *Under Assumption 5, further assume that $\tau > \frac{1}{M}$ and $n^{-2/3+1/M} < \tau < \frac{M}{2}$. Then for any $D > 0$, $\delta > 0$, with probability at least $1 - C(M, D, \delta)n^{-D}$,*

$$|\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) - \text{B}_{\text{in}}| \leq C(M) \max\left\{\frac{1}{\tau^{2/3}n^{(1-\delta)/3}}, \frac{8M}{\tau n^{(1-\delta)/2}}\right\},$$

$$|\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) - \text{V}_{\text{in}}| \leq \sigma^2 C(M) \left(\max\left\{\frac{1}{\tau^{2/3}n^{(1-\delta)/3}}, \frac{8M}{\tau n^{(1-\delta)/2}}\right\} + \frac{1}{n^{(1-\delta)/2}}\right).$$

where $C(M, D, \delta)$ is a constant depending only on (M, D, δ) and $C(M)$ is a constant depending only on M .

Proof. Our proof is inspired by the proof of Theorem 5 in Hastie et al. (2022). We first give the proof for bias and then the proof for variance. In the following proof, $C(M)$ is a constant depending on M and may differ from line to line.

Bias. We first define two functions, $\bar{F}_n^\tau(\eta, \nu)$ and $F_n^\tau(\eta, \nu)$, and control the quantity $|\bar{F}_n^\tau(\eta, \nu) - F_n^\tau(\eta, \nu)|$. Moreover, we show that $-\frac{\partial \bar{F}_n^\tau}{\partial \eta}(0, \tau) = \text{B}_{\text{in}}$ and $-\frac{\partial F_n^\tau}{\partial \eta}(0, \tau) = \mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma)$. Then our objective is to control the quantity $|\frac{\partial \bar{F}_n^\tau}{\partial \eta}(0, \tau) - \frac{\partial F_n^\tau}{\partial \eta}(0, \tau)|$ so that we can control $|\text{B}_{\text{in}} - \mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma)|$.

Without loss of generality, we let $\|\theta^*\|_2^2 = 1$. For $\tau \in \mathbb{R}$, $\frac{1}{M} < \tau < \frac{M}{2}$ and $\tau > n^{-2/3+1/M}$, we define

$$\begin{aligned} \bar{F}_n^\tau(\eta, \nu) &= \nu \langle \theta^*, (\hat{\Sigma} + \nu I + \tau \eta \hat{\Sigma})^{-1} \theta^* \rangle \\ &= \nu \langle \theta^*, ((1 + \tau \eta) \hat{\Sigma} + \nu I)^{-1} \theta^* \rangle. \end{aligned}$$

Define $\mathbb{D} = \{(\eta, \nu) \in \mathbb{R} \times \mathbb{C} : \eta > -\frac{1}{2M}, \text{Re}(\nu) > 0 \text{ and } \text{Im}(-\nu) \geq 0\}$. Because $0 < \tau < M$, $\bar{F}_n^\tau(\eta, \nu)$ is analytical in $\mathbb{D}$ and it can be easily verified that

$$\begin{aligned} -\frac{\partial \bar{F}_n^\tau}{\partial \eta}(0, \tau) &= \tau^2 \langle \theta^*, (\hat{\Sigma} + \tau I)^{-2} \hat{\Sigma} \theta^* \rangle \\ &= \langle \theta^*, (\tau(\hat{\Sigma} + \tau I)^{-1}) \hat{\Sigma} (\tau(\hat{\Sigma} + \tau I)^{-1}) \theta^* \rangle \\ &= \langle \theta^*, (I - \hat{\Sigma}(\hat{\Sigma} + \tau I)^{-1}) \hat{\Sigma} (I - \hat{\Sigma}(\hat{\Sigma} + \tau I)^{-1}) \theta^* \rangle \\ &= B_{\text{in}}. \end{aligned}$$

Define $\mathbb{D}_0 = \{(\eta, \nu) \in \mathbb{R} \times \mathbb{C} : \eta > -\frac{1}{2M}, 0 < \text{Re}(\nu) < M, 0 < \text{Im}(-\nu) < M\}$. By using the anisotropic local law for covariance matrices in Theorem 3.16(i) of Knowles & Yin (2017), we obtain that for any $\delta > 0, \epsilon_0 > 0, D > 0$, with probability at least $1 - C(\epsilon_0, D, \delta)n^{-D}$, we have for $(\eta, \nu) \in \mathbb{D}_0$ and $\text{Re}(\nu) > n^{-2/3+\epsilon_0}$,

$$|\bar{F}_n^\tau(\eta, \nu) - F_n^\tau(\eta, \nu)| \leq \sqrt{\frac{\text{Im}(\tilde{r}_n(\eta, -\nu))}{\text{Im}(-\nu)}} n^{-1+\delta}, \quad (\text{S83})$$

$$F_n^\tau(\eta, \nu) = \langle \theta^*, (I + \tilde{r}_n(\eta, -\nu)(1 + \tau\eta)\Sigma)^{-1} \theta^* \rangle,$$

where $\tilde{r}_n(\eta, z)$ is defined in the domain $\mathbb{D}_1 = \{(\eta, z) \in \mathbb{R} \times \mathbb{C} : \eta > -\frac{1}{2M}, \text{Re}(z) < 0 \text{ and } \text{Im}(z) \geq 0\}$ and it is defined as the unique solution satisfying $\text{Im}(\tilde{r}_n(\eta, z)) > 0$ if $\text{Im}(z) > 0$ or $\tilde{r}_n(\eta, z) > 0$ if $\text{Im}(z) = 0$ of

$$\frac{1}{\tilde{r}_n} = -z + \gamma \frac{1}{p} \sum_{j=1}^p \frac{(1 + \tau\eta)\lambda_j}{1 + (1 + \tau\eta)\lambda_j \tilde{r}_n}.$$

Following a similar process as in Section A.1.2 of Hastie et al. (2022), we have

$$|\text{Im}(\tilde{r}_n(\eta, -\nu))| \leq \frac{|\text{Im}(\nu)|}{\text{Re}(\nu)^2}.$$

Then taking the limit $\text{Im}(\nu) \rightarrow 0$ in (S83) and let $\epsilon_0 = \frac{1}{M}$, we obtain, with probability at least $1 - C(M, D, \delta)n^{-D}$, for $n^{-2/3+1/M} < \nu < M$ and $\eta > -\frac{1}{2M}$,

$$|\bar{F}_n^\tau(\eta, \nu) - F_n^\tau(\eta, \nu)| \leq \frac{1}{n^{(1-\delta)/2}\nu}.$$

Let $\nu = \tau$, we have

$$|\bar{F}_n^\tau(\eta, \tau) - F_n^\tau(\eta, \tau)| \leq \frac{1}{n^{(1-\delta)/2}\tau}. \quad (\text{S84})$$

Then we reformulate $F_n^\tau(\eta, \tau)$ by substituting $\tilde{r}_n(\eta, -\nu)$ with other quantity and show that $-\frac{\partial F_n^\tau}{\partial \eta}(0, \tau) = \mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma)$. Define $\mathbb{D}_2 = \{z \in \mathbb{C} : \text{Re}(z) < 0, \text{Im}(z) \geq 0\}$. For $z \in \mathbb{D}_2$, we define $r_n(z) \in \mathbb{R}$ as the unique solution satisfying $\text{Im}(r_n(z)) > 0$ if $\text{Im}(z) > 0$ or $r_n(z) > 0$ if $\text{Im}(z) = 0$ of

$$\frac{1}{r_n} = -z + \gamma \frac{1}{p} \sum_{j=1}^p \frac{\lambda_j}{1 + \lambda_j r_n}. \quad (\text{S85})$$

In the following discussion, we consider $(\eta, \nu) \in \mathbb{D} \cap \mathbb{R}^2$, that is $\eta > -\frac{1}{2M}$ and $\nu > 0$. Then we have,

$$\begin{aligned}\tilde{r}_n(\eta, -\nu) &= \frac{1}{1 + \tau\eta} r_n(-\frac{\nu}{1 + \tau\eta}), \\ F_n^\tau(\eta, \nu) &= \langle \theta^*, (I + r_n(-\frac{\nu}{1 + \tau\eta})\Sigma)^{-1} \theta^* \rangle.\end{aligned}\tag{S86}$$

Let $m_n(z) = \frac{1 - \gamma + z r_n(z)}{\gamma z}$, which by (S85) is the unique solution of

$$m_n = \frac{1}{p} \sum_{j=1}^p \frac{1}{\lambda_j(1 - \gamma - \gamma z m_n) - z}.$$

Then we have

$$F_n^\tau(\eta, \nu) = \langle \theta^*, (I + (m_n(-\frac{\nu}{1 + \tau\eta})\gamma + \frac{(1 - \gamma)(1 + \tau\eta)}{\nu})\Sigma)^{-1} \theta^* \rangle,$$

and

$$\begin{aligned}-\frac{\partial F_n^\tau(0, \tau)}{\partial \eta} &= (\gamma \tau^2 m'_n(-\tau) + 1 - \gamma) \langle \theta^*, (I + (m_n(-\tau)\gamma + \frac{1 - \gamma}{\tau})\Sigma)^{-2} \Sigma \theta^* \rangle \\ &= \tau^2 (\gamma \tau^2 m'_n(-\tau) + 1 - \gamma) \|\theta^*\|^2 \int \frac{s}{[\tau + (1 - \gamma + \gamma \lambda m_n(-\tau))s]^2} d\hat{G}_n(s) \\ &= \mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma).\end{aligned}$$

In the following discussion, we give upper bounds on $|\frac{\partial \bar{F}_n^\tau}{\partial \eta}(0, \tau) - \frac{\partial F_n^\tau}{\partial \eta}(0, \tau)|$, equivalently $|\mathcal{B}_{\text{in}} - \mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma)|$. Our strategy is to control $|\frac{\partial^k \bar{F}_n^\tau}{\partial \eta^k}(\eta, \tau)|$, $|\frac{\partial^k F_n^\tau}{\partial \eta^k}(\eta, \tau)|$, then $|\mathcal{B}_{\text{in}} - \mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma)|$, can be controlled by $|\bar{F}_n^\tau(\eta, \tau) - F_n^\tau(\eta, \tau)|$, $|\frac{\partial^k \bar{F}_n^\tau}{\partial \eta^k}(\eta, \tau)|$ and $|\frac{\partial^k F_n^\tau}{\partial \eta^k}(\eta, \tau)|$ based on the Lemma A.1 in Hastie et al. (2022).

We first give upper bound on $|\frac{\partial^k \bar{F}_n^\tau}{\partial \eta^k}(\eta, \tau)|$. For $0 < \nu < M, \eta > -\frac{1}{2M}$, we have for $k \geq 1$,

$$\frac{\partial^k \bar{F}_n^\tau}{\partial \eta^k}(\eta, \nu) = k!(-1)^{k+1} \tau^k \nu \langle \theta^*, R^k \hat{\Sigma}^k R \theta^* \rangle,$$

where $k! = k \times \cdots \times 1$ and $R = ((1 + \tau\eta)\hat{\Sigma} + \nu I)^{-1}$. Then we have for $\nu = \tau, \eta > -\frac{1}{2M}$ and $k \geq 1$,

$$|\frac{\partial^k \bar{F}_n^\tau}{\partial \eta^k}(\eta, \tau)| \leq k! 2^k \tau^k \|\theta^*\|^2 \leq k! 2^k M^k.\tag{S87}$$

Next we give the upper bound on $|\frac{\partial^k F_n^\tau}{\partial \eta^k}(\eta, \tau)|$. From (S86), it is sufficient to upper bound $|\frac{\tau^{l+1}}{(1 + \tau\eta)^{l+1}}|$, $|r_n^{(l)}(-\frac{\tau}{1 + \tau\eta})|$ and $|\langle \theta^*, (I + r_n(-\frac{\tau}{1 + \tau\eta})\Sigma)^{-(l+1)} \Sigma^l \theta^* \rangle|$ for $1 \leq l \leq k$. We give their upper bounds separately as follows.

- **Upper bound of $|\frac{\tau^{l+1}}{(1 + \tau\eta)^{l+1}}|$.** Because $\frac{1}{M} < \tau < \frac{M}{2}$, for $-\frac{1}{2M} < \eta < \frac{1}{2M}$ and $1 \leq l \leq k$, we have

$$|\frac{\tau^{l+1}}{(1 + \tau\eta)^{l+1}}| \leq M^{l+1}.$$

- **Upper bound of $|\langle \theta^*, (I + r_n(-\frac{\tau}{1+\tau\eta})\Sigma)^{-(l+1)}\Sigma^l\theta^* \rangle|$.** We show $r_n(-\frac{\tau}{1+\tau\eta}) > 0$. Let $u_n(z) = \frac{1}{r_n(-z)}$. Then for $z > 0$, $u_n(z)$ is the unique solution of

$$u_n = z + \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j u_n}{u_n + \lambda_j}.$$

From Lemma A.2(a)(b) in Hastie et al. (2022), we have for $\frac{1}{2M} < z < M$,

$$u_n(z) > \frac{1}{4M} (> 0), \quad (\text{S88})$$

$$0 < |u'_n(z)| < C(M). \quad (\text{S89})$$

By (S88) and the definition $u_n(z) = \frac{1}{r_n(-z)}$, we have for $\frac{1}{2M} < z < M$,

$$r_n(-z) > 0. \quad (\text{S90})$$

Because $\frac{1}{M} < \tau < \frac{M}{2}$, then for $-\frac{1}{2M} < \eta < \frac{1}{2M}$, we have $\frac{1}{2M} < \frac{\tau}{1+\tau\eta} < M$ and $r_n(-\frac{\tau}{1+\tau\eta}) > 0$. Hence for $-\frac{1}{2M} < \eta < \frac{1}{2M}$ and $1 \leq l \leq k$,

$$\begin{aligned} |\langle \theta^*, (I + r_n(-\frac{\tau}{1+\tau\eta})\Sigma)^{-(l+1)}\Sigma^l\theta^* \rangle| &\leq \|(I + r_n(-\frac{\tau}{1+\tau\eta})\Sigma)^{-1}\|_{\text{op}}^{l+1} \|\Sigma\|_{\text{op}}^l \|\theta^*\|^2 \\ &\leq \|\Sigma\|_{\text{op}}^l \|\theta^*\|^2 \quad (\text{because } r_n(-\frac{\tau}{1+\tau\eta}) > 0) \\ &\leq \lambda_1^l \|\theta^*\|^2 \\ &\leq M^l \|\theta^*\|^2 \quad (\text{from Assumption 5}) \end{aligned}$$

- **Upper bound of $|r_n^{(l)}(-\frac{\tau}{1+\tau\eta})|$.** Let $u_n(z) = \frac{1}{r_n(-z)}$. Then $|r_n^{(l)}(-z)|$ can be controlled by the polynomial of $|u_n^{(m)}(z)|$ for $1 \leq m \leq l$ and $|\frac{1}{u_n(z)}|$.

From (S88), we have for $\frac{1}{2M} < z < M$,

$$|u_n^{-1}(z)| \leq 4M. \quad (\text{S91})$$

The upper bound of $|u'_n(z)|$ is provided in (S89). Then we give the upper bounds of $|u_n^{(m)}(z)|$ for $2 \leq m \leq l$. We consider the following function

$$f(u_n, z) = u_n + z - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j u_n}{u_n + \lambda_j}.$$

From implicit function theorem and (S89), we have for $\frac{1}{2M} < z < M$,

$$|\frac{\partial f}{\partial u_n}| = |u'_n(z)|^{-1} \geq \frac{1}{C(M)}. \quad (\text{S92})$$

To give upper the bounds of $|u_n^{(m)}(z)|$ for $2 \leq m \leq l$, from the implicit function theorem, it is sufficient to further give the upper bounds of $|\frac{\partial^{s+t}f}{\partial u_n^s \partial z^t}|$ for all $s+t \leq m$ and $s \geq 1, t \geq 1$ or $s=0, t \geq 2$. Denote

$$f_1(u_n) = \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j u_n}{u_n + \lambda_j}.$$

Then it is sufficient to give upper bounds of $|\frac{\partial^s f_1(u_n)}{\partial u_n^s}|$ for $1 \leq s \leq m$. From (S88), we have for $\frac{1}{2M} < z < M$,

$$u_n(z) > \frac{1}{4M}.$$

Then

$$\begin{aligned} \left| \frac{\partial^s f_1(u_n)}{\partial u_n^s} \right| &= \left| \frac{s!}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + u_n)^{s+1}} \right| \\ &\leq s! \frac{p}{n} \frac{\lambda_1^2}{(\lambda_p + u_n)^{s+1}} \\ &\leq s! \frac{p}{n} \frac{\lambda_1^2}{u_n^{s+1}} \\ &\leq C(M). \quad (\text{from } \gamma = \frac{p}{n} < M, \lambda_1 < M \text{ and } u_n(z) > \frac{1}{4M}) \end{aligned}$$

Hence from the implicit function theorem, for $\frac{1}{2M} < z < M$, $s+t \leq m$ and $s \geq 0, t \geq 1$ or $s=0, t \geq 2$,

$$\left| \frac{\partial^{s+t}f}{\partial u_n^s \partial z^t} \right| \leq C(M).$$

Combining the above with (S92), from implicit function theorem, we have for $\frac{1}{2M} < z < M$ and $2 \leq m \leq l$,

$$|u_n^{(m)}(z)| \leq C(M). \tag{S93}$$

With the upper bound of $|u_n^{-1}(z)|$ in (S91), upper bound of $|u_n'(z)|$ in (S89) and upper bound of $|u_n^{(m)}(z)|$ for $2 \leq m \leq l$ in (S93), we have for $\frac{1}{2M} < z < M$ and for $1 \leq l \leq k$,

$$|r_n^{(l)}(-z)| \leq C(M). \tag{S94}$$

Because $\frac{1}{M} < \tau < \frac{M}{2}$, for $-\frac{1}{2M} < \eta < \frac{1}{2M}$, we have $\frac{1}{2M} < \frac{\tau}{1+\eta\tau} < M$. From (S94), we have for $-\frac{1}{2M} < \eta < \frac{1}{2M}$ and $1 \leq l \leq k$,

$$|r_n^{(l)}(-\frac{\tau}{1+\eta\tau})| \leq C(M).$$

From the upper bounds of $|\frac{\tau^{l+1}}{(1+\tau\eta)^{l+1}}|$, $|\langle \theta^*, (I + r_n(-\frac{\tau}{1+\tau\eta})\Sigma)^{-(l+1)}\Sigma^l\theta^* \rangle|$ and $|r_n^{(l)}(-\frac{\tau}{1+\tau\eta})|$ for $1 \leq l \leq k$ above, we have for $-\frac{1}{2M} < \eta < \frac{1}{2M}$ and $k \geq 1$,

$$|\frac{\partial^k F_n^\tau}{\partial \eta^k}(\eta, \tau)| \leq C(M). \quad (\text{S95})$$

In the following discussion, we combine the upper bounds of $|\bar{F}_n^\tau(\eta, \tau) - F_n^\tau(\eta, \tau)|$, $|\frac{\partial^k \bar{F}_n^\tau}{\partial \eta^k}(\eta, \tau)|$ and $|\frac{\partial^k F_n^\tau}{\partial \eta^k}(\eta, \tau)|$ from above and apply Lemma A.1 in Hastie et al. (2022) to control $|\frac{\partial \bar{F}_n^\tau}{\partial \eta}(0, \tau) - \frac{\partial F_n^\tau}{\partial \eta}(0, \tau)|$. Combining (S84), (S87) and (S95), from Lemma A.1 in Hastie et al. (2022), and letting $k = 3$, we have for $0 < \xi < \frac{1}{4M}$, $D > 0$ and $\delta > 0$, with probability at least $1 - C(\delta, M, D)n^{-D}$,

$$|\frac{\partial \bar{F}_n^\tau}{\partial \eta}(0, \tau) - \frac{\partial F_n^\tau}{\partial \eta}(0, \tau)| \leq C(M)(\frac{1}{\tau n^{(1-\delta)/2}} \frac{1}{\xi} + \xi^2).$$

That is, we have

$$|B_{in} - \mathcal{B}_{in}(\tau; \hat{H}_n, \hat{G}_n, \gamma)| \leq C(M)(\frac{1}{\tau n^{(1-\delta)/2}} \frac{1}{\xi} + \xi^2).$$

Letting $\xi = \min\{\frac{1}{8M}, \tau^{1/3}n^{(1-\delta)/6}\}$, we have

$$|\mathcal{B}_{in}(\tau; \hat{H}_n, \hat{G}_n, \gamma) - B_{in}| \leq C(M) \max\{\frac{1}{\tau^{2/3}n^{(1-\delta)/3}}, \frac{8M}{\tau n^{(1-\delta)/2}}\}.$$

Variance. We have

$$\begin{aligned} V_{in} &= \frac{\sigma^2}{n} \text{Tr}(\hat{\Sigma}^2(\hat{\Sigma}^2 + \tau I)^{-2}) \\ &= \sigma^2(\gamma - 2\tau\gamma \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau I)^{-1}) + \tau^2\gamma \frac{1}{p} \text{Tr}(\hat{\Sigma} + \tau I)^{-2}). \end{aligned}$$

From (12) and $m_n(-\tau) = \frac{\gamma - 1 + \tau r_n(-\tau)}{\gamma \tau}$,

$$\mathcal{V}_{in}(\tau; \hat{H}_n, \gamma) = \sigma^2(1 - 2\tau r_n(-\tau) + \tau^2 r_n'(-\tau)).$$

We first give the equations and inequality below for the following analysis. From (S85), we have

$$\begin{aligned} \tau r_n(-\tau) + \gamma - 1 &= \gamma \frac{1}{p} \sum_{j=1}^n \frac{1}{1 + r_n(-\tau)\lambda_j} \\ &= \gamma \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1}). \end{aligned} \quad (\text{S96})$$

From (S96), we have

$$\tau \frac{\partial \gamma \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau} = -\tau^2 r_n'(-\tau) + \tau r_n(-\tau).$$

Hence we have

$$\tau \frac{\partial \gamma \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau I)^{-1})}{\partial \tau} - \tau \frac{\partial \gamma \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau}$$

$$\begin{aligned}
&= \tau\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) - \tau^2\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-2}) + \tau^2r'_n(-\tau) - \tau r_n(-\tau) \\
&= (\tau^2r'_n(-\tau) + \gamma - 1 - \tau^2\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-2}) + (\tau\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) - (\tau r_n(-\tau) + \gamma - 1))).
\end{aligned}$$

From above, we have

$$\begin{aligned}
&|\tau^2r'_n(-\tau) + \gamma - 1 - \tau^2\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-2})| \\
&\leq |\tau\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) - (\tau r_n(-\tau) + \gamma - 1)| + |\tau\gamma\frac{\partial\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1})}{\partial\tau} - \tau\gamma\frac{\partial\frac{1}{p}\text{Tr}((I + r_n(-\tau)\Sigma)^{-1})}{\partial\tau}| \\
&= |\tau\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) - \gamma\frac{1}{p}\text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})| + |\tau\gamma\frac{\partial\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1})}{\partial\tau} - \tau\gamma\frac{\partial\frac{1}{p}\text{Tr}((I + r_n(-\tau)\Sigma)^{-1})}{\partial\tau}|.
\end{aligned} \tag{S97}$$

We give the upper bound of $|\mathcal{V}_{\text{in}} - \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma)|$ by controlling $|\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau)^{-1}) - \frac{1}{p}\text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})|$ and $|\tau\frac{\partial\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1})}{\partial\tau} - \tau\frac{\partial\frac{1}{p}\text{Tr}((I + r_n(-\tau)\Sigma)^{-1})}{\partial\tau}|$. In fact, we have

$$\begin{aligned}
&|\mathcal{V}_{\text{in}} - \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma)| \\
&= \sigma^2|(\gamma - 2\tau\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) + \tau^2\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-2})) - (1 - 2\tau r_n(-\tau) + \tau^2r'_n(-\tau))| \\
&\leq \sigma^2|2\tau r_n(-\tau) + 2\gamma - 2 - 2\tau\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) + (\tau^2\gamma\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-2}) - (\tau^2r'_n(-\tau) + \gamma - 1))| \\
&\leq \sigma^2(2|\gamma\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) - (\tau r_n(-\tau) + \gamma - 1)| + |\gamma\tau^2\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-2}) - (\tau^2r'_n(-\tau) + \gamma - 1)|) \\
&\leq \sigma^2\gamma(3|\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) - \frac{1}{p}\text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})| \\
&\quad + |\tau\frac{\partial\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1})}{\partial\tau} - \tau\frac{\partial\frac{1}{p}\text{Tr}((I + r_n(-\tau)\Sigma)^{-1})}{\partial\tau}|) \quad (\text{By (S96) and (S97)}).
\end{aligned} \tag{S98}$$

We first give the upper bound on $|\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau)^{-1}) - \frac{1}{p}\text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})|$. Using Theorem 3.16(i) in Knowles & Yin (2017), we have for $D > 0$, $\delta > 0$, $\text{Im}(-\tau) > 0$ and $\text{Re}(\tau) > n^{-2/3+1/M}$, with probability at least $1 - C(M, D, \delta)n^{-D}$,

$$|\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau)^{-1}) - \frac{1}{p}\text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})| \leq \sqrt{\frac{\text{Im}(r_n(-\tau))}{\text{Im}(-\tau)}}.$$

Following a similar process in Section A.1.2 of Hastie et al. (2022), we have $|\text{Im}(r_n(-\tau))| \leq |\text{Im}(\tau)|/\text{Re}(\tau)^2$. Letting $\text{Im}(\tau) \rightarrow 0$ shows that for $D > 0$, $\delta > 0$, $\frac{1}{M} < \tau < M$ and $\tau > n^{-2/3+(1/M)}$, with probability at least $1 - C(M, D, \delta)n^{-D}$,

$$|\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1}) - \frac{1}{p}\text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})| \leq \frac{1}{\tau n^{(1-\delta)/2}} \leq \frac{M}{n^{(1-\delta)/2}}. \tag{S99}$$

Next, we give the upper bound on $|\tau\frac{\partial\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau I)^{-1})}{\partial\tau} - \tau\frac{\partial\frac{1}{p}\text{Tr}((I + r_n(-\tau)\Sigma)^{-1})}{\partial\tau}|$. Our strategy is to upper bound $|\frac{\partial^k\tau\frac{1}{p}\text{Tr}((\hat{\Sigma} + \tau)^{-1})}{\partial\tau^k}|$ and $|\frac{\partial^k\frac{1}{p}\text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial\tau^k}|$ for $k \geq 1$, so that Lemma A.1 in Hastie

et al. (2022) can be applied again to bound $|\frac{\partial \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1})}{\partial \tau} - \frac{\partial \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau}|$ by $|\tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1}) - \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})|$, $|\frac{\partial^k \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1})}{\partial \tau^k}|$ and $|\frac{\partial^k \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau^k}|$ for $k \geq 1$. The upper bound of $|\tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1}) - \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})|$ is given in (S99). Then we give the upper bound of $|\frac{\partial^k \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1})}{\partial \tau^k}|$ and $|\frac{\partial^k \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau^k}|$ as below.

We give the upper bound on $|\frac{\partial^k \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau I)^{-1})}{\partial \tau^k}|$. For $k \geq 1$ and $\frac{1}{M} < \tau < M$,

$$\begin{aligned} |\frac{\partial^k \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau I)^{-1})}{\partial \tau^k}| &= |\frac{k!}{p} \text{Tr}(\hat{\Sigma}(\hat{\Sigma} + \tau I)^{-(k+1)})| \\ &\leq \frac{k!}{\tau^k} \leq C(M). \end{aligned} \quad (\text{S100})$$

Then we give the upper bound to $|\frac{\partial^k \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau^k}|$. It is sufficient to upper bound $|r_n^{(l)}(-\tau)|$ and $|\frac{1}{p} \text{Tr}(\Sigma^l(I + r_n(-\tau)\Sigma)^{-(l+1)})|$ for $1 \leq l \leq k$. From (S90), we have for $\frac{1}{2M} < \tau < M$,

$$r_n(-\tau) > 0,$$

and

$$\begin{aligned} |\frac{1}{p} \text{Tr}(\Sigma^l(I + r_n(-\tau)\Sigma)^{-(l+1)})| &\leq \|(I + r_n(-\tau)\Sigma)^{-1}\|_{\text{op}}^{l+1} \|\Sigma\|_{\text{op}}^l \\ &\leq \|\Sigma\|_{\text{op}}^l \quad (\text{because } r_n(-\tau) > 0) \\ &\leq \lambda_1^l \\ &\leq M^l \quad (\text{from Assumption 5}) \end{aligned}$$

From (S94), for $\frac{1}{2M} < \tau < M$ and $1 \leq l \leq k$,

$$|r_n^{(l)}(-\tau)| < C(M).$$

Hence for $\frac{1}{2M} < \tau < M$ and $k \geq 1$,

$$|\frac{\partial^k \frac{1}{p} \text{Tr}((I + r_n(-\tau)\Sigma)^{-1})}{\partial \tau^k}| < C(M). \quad (\text{S101})$$

We combine the upper bounds of $|\tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1}) - \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})|$, $|\frac{\partial^k \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1})}{\partial \tau^k}|$ and $|\frac{\partial^k \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau^k}|$ from above and apply Lemma A.1 in Hastie et al. (2022) to control $|\tau \frac{\partial \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1})}{\partial \tau} - \tau \frac{\partial \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau}|$. From (S99)–(S101) and Lemma A.1 in Hastie et al. (2022), and letting $k = 3$, we have for $D > 0$, $\delta > 0$, $\xi < \frac{1}{M}$, for $\tau > n^{-2/3+(1/M)}$ and $\frac{1}{M} < \tau < M$, with probability at least $1 - C(D, \delta, M)n^{-D}$,

$$|\tau \frac{\partial \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1})}{\partial \tau} - \tau \frac{\partial \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau}| \leq C(M) (\frac{1}{\tau n^{(1-\delta)/2}} \frac{1}{\xi} + \xi^2). \quad (\text{S102})$$

With the upper bounds of $|\tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau I)^{-1}) - \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})|$ and $|\tau \frac{\partial \tau \frac{1}{p} \text{Tr}((\hat{\Sigma} + \tau)^{-1})}{\partial \tau} - \tau \frac{\partial \frac{1}{p} \text{Tr}((1 + r_n(-\tau)\Sigma)^{-1})}{\partial \tau}|$ from above, we give the upper bound of $|\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) - V_{\text{in}}|$. From (S98)–(S99) and (S102), we have for any $D > 0$, $\delta > 0$, $0 < \xi < \frac{1}{M}$, for $\tau > n^{-2/3+(1/M)}$ and $\frac{1}{M} < \tau < M$, with probability at least $1 - C(D, \delta, M)n^{-D}$,

$$|\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) - V_{\text{in}}| \leq \sigma^2 C(M) \left(\frac{1}{\tau n^{(1-\epsilon)/2}} \frac{1}{\xi} + \xi^2 + \frac{1}{n^{(1-\delta)/2}} \right).$$

Letting $\xi = \min\{\frac{1}{8M}, \tau^{1/3} n^{(1-\delta)/6}\}$, we have

$$|\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) - V_{\text{in}}| \leq \sigma^2 C(M) \left(\max\left\{ \frac{1}{\tau^{2/3} n^{(1-\epsilon)/3}}, \frac{8M}{\tau n^{(1-\epsilon)/2}} \right\} + \frac{1}{n^{(1-\epsilon)/2}} \right).$$

□

IV.2 Orders of error approximation formulas

We provide proofs of Corollaries 8 and 9 in Section 4.1, which are re-stated below for convenience.

Corollary 7 (Matching error approximation formulas with small or moderate TER).

(i) Suppose that $\frac{d}{n} < 1$, $r_d(\Sigma) \lesssim n$, and $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_{d+1}^2$. For $\lambda_{d+1} \lesssim \tau \lesssim \lambda_d$, we have

$$\mathcal{B}_{\text{out}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{out}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \frac{r_d(\Sigma^2)}{n} \right).$$

(ii) Suppose further that $r_d(\Sigma) \asymp n$. For $\lambda_{d+1} \lesssim \tau \lesssim \lambda_d$, we have

$$\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \right).$$

Proof.

We first show that $\lambda_d \gtrsim \alpha \tau$. From (13), $\tau \gtrsim \lambda_{d+1}$ and $n > d$, we have

$$\begin{aligned} & \frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d}^p \lambda_j}{\alpha \tau} \asymp 1 \\ \implies \alpha \tau & \asymp \left(\tau + \frac{\sum_{j>d}^p \lambda_j}{n} \right) \\ \implies \alpha \tau & \asymp \tau \quad (\text{from } \lambda_{d+1} \lesssim \tau \text{ and } r_d(\Sigma) \lesssim n) \\ \implies \lambda_d & \gtrsim \alpha \tau \quad (\text{from } \tau \lesssim \lambda_d). \end{aligned} \tag{S103}$$

Then we show $\alpha \asymp 1$. From (13), $\tau \gtrsim \lambda_{d+1}$ and $n > d$, we have

$$\begin{aligned} & \frac{1}{\alpha} \left(1 + \frac{\sum_{j>d} \lambda_j}{n \tau} \right) \asymp 1 \\ & \left(1 + \frac{\sum_{j>d} \lambda_j}{n \tau} \right) \asymp \alpha \end{aligned}$$

$$\implies \alpha \asymp 1 \quad (\text{from } \lambda_{d+1} \lesssim \tau \text{ and } r_d(\Sigma) \lesssim n). \quad (\text{S104})$$

Then we give the orders of $1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}$, $\frac{1}{n} \sum_{j=1}^p \frac{\alpha^2 \tau^2 \lambda_j \theta_j^{*2}}{(\lambda_j + \alpha\tau)^2}$ and $\frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}$, which are important in the formulas (14)–(17). We have

$$\begin{aligned} 1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} &\geq 1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j}{\lambda_j + \alpha\tau} = \frac{1}{\alpha} \asymp 1 \quad (\text{from (S104)}), \\ 1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} &\leq 1. \end{aligned}$$

Hence

$$1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} \asymp 1. \quad (\text{S105})$$

From (S103)–(S104) and $\tau \gtrsim \lambda_{d+1}$, we have

$$\frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} \asymp \frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n\tau^2}, \quad (\text{S106})$$

$$\frac{1}{n} \sum_{j=1}^p \frac{\alpha^2 \tau^2 \lambda_j \theta_j^{*2}}{(\lambda_j + \alpha\tau)^2} \asymp \sum_{j=1}^d \frac{\tau^2 \theta_j^{*2}}{\lambda_j} + \sum_{j>d} \lambda_j \theta_j^{*2}. \quad (\text{S107})$$

Substituting (S105)–(S107) into (14)–(15), we have

$$\begin{aligned} \mathcal{B}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) &\asymp \sum_{j=1}^d \frac{\tau^2 \theta_j^{*2}}{\lambda_j} + \sum_{j>d} \lambda_j \theta_j^{*2} \\ &\asymp \tau^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\ &\asymp \tau^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \quad (\text{from } \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_{d+1}^2), \\ \mathcal{V}_{\text{out}}(\tau, \gamma, \tilde{\lambda}) &\asymp \sigma^2 \left(\frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n\tau^2} \right). \end{aligned}$$

Hence we have

$$\mathcal{B}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{out}}(\tau; \hat{H}_n, \gamma) \asymp \tau^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 + \sigma^2 \left(\frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n\tau^2} \right).$$

Note that $\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma)$ can be also expressed as

$$\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) = \left(1 - \frac{1}{\alpha}\right)^2 \sigma^2 + \frac{1}{\alpha^2} \frac{\frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}}{\left(1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}\right)} \sigma^2. \quad (\text{S108})$$

From (13), (S103), (S104) and $\tau \gtrsim \lambda_{d+1}$, we have

$$1 - \frac{1}{\alpha} \asymp \frac{d}{n} + \frac{\sum_{j>d} \lambda_j}{n\tau}. \quad (\text{S109})$$

Substituting (S104)–(S107) into (16) and (S108), we have

$$\begin{aligned}
\mathcal{B}_{\text{in}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) &\asymp \sum_{j=1}^d \frac{\tau^2 \theta_j^{*2}}{\lambda_j} + \sum_{j>d} \lambda_j \theta_j^{*2} \\
&\asymp \tau^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\
&\asymp \tau^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \quad (\text{from } \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_{d+1}^2), \\
\mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) &\asymp \sigma^2 \left(\frac{d}{n} + \frac{(\sum_{j>d} \lambda_j)^2}{\tau^2} \right) \\
&\asymp \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \right) \quad (\text{from } r_d(\Sigma) \asymp n).
\end{aligned}$$

Hence we have

$$\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \tau^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{\tau^2} \right). \quad (\text{S110})$$

□

Corollary 8 (Matching error approximation formulas with large TER).

(i) Suppose that $\frac{d}{n} < \frac{1}{5}$, $r_d(\Sigma) > cn$ for some $c > 10$, $\|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_d} + \sum_{j>d} \frac{n}{\lambda_j})^{-2}$. For $\lambda_d \gtrsim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ and $\tau > 0$, we have

$$\mathcal{B}_{\text{out}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{out}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d(\Sigma^2)}{n} \right).$$

(ii) Suppose further that $\tau > \lambda_{d+1} \frac{r_d(\Sigma)}{n}$. For $\lambda_d \gtrsim \tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n}$ and $\tau > 0$, we have

$$\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2} \right).$$

Proof.

We first prove that $\alpha\tau > (c-1)\lambda_{d+1}$. From (13),

$$\begin{aligned}
n &> \sum_{j>d} \frac{\lambda_j}{\lambda_j + \alpha\tau} \\
&> \frac{\sum_{j>d} \lambda_j}{\lambda_{d+1} + \alpha\tau} \\
&> \frac{\sum_{j>d} \lambda_j}{\lambda_{d+1}} \frac{1}{(1 + \frac{\alpha\tau}{\lambda_{d+1}})} \\
&> cn \frac{1}{(1 + \frac{\alpha\tau}{\lambda_{d+1}})}.
\end{aligned}$$

Hence we have

$$\alpha\tau > (c-1)\lambda_{d+1}. \quad (\text{S111})$$

Then we prove that $\frac{c}{c-1} > \frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\alpha\tau} > \frac{4}{5}$. We have,

$$\begin{aligned}
\frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\alpha\tau} &> \frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\lambda_j + \alpha\tau} \\
&= 1 - \frac{1}{n} \frac{\sum_{i=1}^d \lambda_j}{\lambda_j + \alpha\tau} \quad (\text{from (13)}) \\
&\geq 1 - \frac{d}{n} \\
&> \frac{4}{5} \quad (\text{from } \frac{d}{n} < \frac{1}{5}).
\end{aligned}$$

From (13) and (S111), we have

$$\begin{aligned}
\frac{c-1}{c} \left(\frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\alpha\tau} \right) &= \frac{1}{(\frac{1}{c-1} + 1)\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{(\frac{1}{c-1} + 1)\alpha\tau} \\
&< \frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\lambda_{d+1} + \alpha\tau} \\
&\leq \frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\lambda_j + \alpha\tau} \\
&= 1 - \frac{1}{n} \frac{\sum_{i=1}^d \lambda_j}{\lambda_j + \alpha\tau} \\
&\leq 1.
\end{aligned}$$

That is, we have

$$\frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\alpha\tau} < \frac{c}{c-1}.$$

Hence we have

$$\frac{c}{c-1} > \frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\alpha\tau} > \frac{4}{5}, \tag{S112}$$

and

$$\frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\alpha\tau} \asymp 1. \tag{S113}$$

Now we prove that $\lambda_d \gtrsim \alpha\tau$. From (S113), we have

$$\begin{aligned}
\frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d}^p \lambda_j}{\alpha\tau} &\asymp 1 \\
\implies \alpha\tau &\asymp \left(\tau + \frac{\sum_{j>d}^p \lambda_j}{n} \right) \\
\implies \lambda_d &\gtrsim \alpha\tau \quad (\text{from } \lambda_d \gtrsim \tau + \frac{\sum_{j>d} \lambda_j}{n}).
\end{aligned} \tag{S114}$$

We prove the order matching in two cases, $\sum_{j>d} \lambda_j < n\tau$ and $\sum_{j>d} \lambda_j \geq n\tau$. We discuss the two cases separately.

$$\textcircled{1} \sum_{j>d} \lambda_j < n\tau.$$

From (S113) and $\sum_{j>d} \lambda_j < n\tau$, we have

$$\alpha \asymp 1. \quad (\text{S115})$$

From (S111), (S114) and (S115), by a similar process as the proof in Corollary 8, (S105), (S106), (S107) and (S109) hold. Substituting (S105), (S106) and (S107) into (14), (15), we have

$$\begin{aligned} \mathcal{B}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) &\asymp \tau^2 \|\theta_{1:d}\|_{\Sigma_{1:d}^{-1}}^2 + \sigma^2 \left(\frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n\tau^2} \right) \\ &\asymp \left(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 + \sigma^2 \left(\frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \right) \\ &\quad (\text{from } \sum_{j>d} \lambda_j < n\tau). \end{aligned}$$

Substituting (S105), (S106), (S107), (S109) and (S115) into (16) and (S108), we have

$$\begin{aligned} \mathcal{B}_{\text{in}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) &\asymp \sum_{j=1}^d \frac{\tau^2 \theta_j^{*2}}{\lambda_j} + \sum_{j>d} \lambda_j \theta_j^{*2} \\ &\asymp \left(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \quad (\text{from } \sum_{j>d} \lambda_j < n\tau) \\ &\asymp \left(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \right)^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \quad (\text{from } \|\theta_{(d+1):p}^*\|^2 \lesssim \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{1}{\lambda_d} + \lambda_{d+1} \frac{1}{\frac{\sum_{j>d} \lambda_j}{n}} \right)^{-2}), \\ \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) &\asymp \sigma^2 \left(\frac{d}{n} + \frac{(\sum_{j>d} \lambda_j)^2}{\tau^2} \right) \\ &\asymp \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2} \right) \quad (\text{from } \sum_{j>d} \lambda_j < n\tau). \end{aligned}$$

Hence

$$\mathcal{B}_{\text{in}}(\tau; \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{in}}(\tau; \hat{H}_n, \gamma) \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n} \right)^2 + \sigma^2 \left(\frac{d}{n} + \frac{\lambda_{d+1}^2}{(\tau + \lambda_{d+1} \frac{r_d(\Sigma)}{n})^2} \frac{r_d^2(\Sigma)}{n^2} \right).$$

$$\textcircled{2} \sum_{j>d} \lambda_j \geq n\tau.$$

We first give the order of the term $1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}$, which is important in the following analysis. We have

$$\begin{aligned} \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} &\leq \frac{d}{n} + \frac{\sum_{j>d} \lambda_j^2}{n\alpha^2\tau^2} \\ &= \frac{d}{n} + \frac{n \sum_{j>d} \lambda_j^2}{n^2\alpha^2\tau^2} \\ &= \frac{d}{n} + \frac{(\sum_{j>d} \lambda_j)^2}{n^2\alpha^2\tau^2} \frac{n \sum_{j>d} \lambda_j^2}{(\sum_{j>d} \lambda_j)^2} \end{aligned}$$

$$\begin{aligned}
&\leq \frac{d}{n} + \frac{c^2}{(c-1)^2} \frac{n \sum_{j>d} \lambda_j^2}{(\sum_{j>d} \lambda_j)^2} \quad (\text{from (S112)}) \\
&\leq \frac{d}{n} + \frac{c^2}{(c-1)^2} \frac{n}{r_d(\Sigma)} \\
&\leq \frac{d}{n} + \frac{c}{(c-1)^2} \quad (\text{from } r_d(\Sigma) > cn) \\
&< 2/5 \quad (\text{from } \frac{d}{n} < \frac{1}{5} \text{ and } c > 10).
\end{aligned} \tag{S116}$$

$$< 2/5 \quad (\text{from } \frac{d}{n} < \frac{1}{5} \text{ and } c > 10). \tag{S117}$$

From (S117), we have

$$1 > 1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} > \frac{3}{5}, \tag{S118}$$

and

$$1 - \frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} \asymp 1. \tag{S119}$$

Now we give the orders of $\frac{1}{n} \sum_{j=1}^p \frac{\alpha^2 \tau^2 \lambda_j \theta_j^{*2}}{(\lambda_j + \alpha\tau)^2}$ and $\frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2}$, which are important in formulas (14) and (15). From (S113), we have

$$\begin{aligned}
&\frac{1}{\alpha} + \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\alpha\tau} \asymp 1 \\
&\Rightarrow \frac{1}{n} \frac{\sum_{j>d} \lambda_j}{\alpha\tau} \asymp 1 \quad (\text{from } \frac{\sum_{j>d} \lambda_j}{n} \geq n\tau) \\
&\Rightarrow \alpha\tau \asymp \frac{\sum_{j>d} \lambda_j}{n}.
\end{aligned} \tag{S120}$$

From (S111), (S114), (S120) and $\sum_{j>d} \lambda_j \geq n\tau$, we have

$$\frac{1}{n} \sum_{j=1}^p \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} \asymp \frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n(\frac{\sum_{j>d} \lambda_j}{n})^2}, \tag{S121}$$

$$\frac{1}{n} \sum_{j=1}^p \frac{\alpha^2 \tau^2 \lambda_j \theta_j^{*2}}{(\lambda_j + \alpha\tau)^2} \asymp \sum_{j=1}^d \frac{(\frac{\sum_{j>d} \lambda_j}{n})^2 \theta_j^{*2}}{\lambda_j} + \sum_{j>d} \lambda_j \theta_j^{*2}. \tag{S122}$$

Substituting (S118), (S120), (S121) and (S122) into (14)–(15), we have

$$\begin{aligned}
\mathcal{B}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) &\asymp \sum_{j=1}^d \frac{(\frac{\sum_{j>d} \lambda_j}{n})^2 \theta_j^{*2}}{\lambda_j} + \sum_{j>d} \lambda_j \theta_j^{*2} \\
&\asymp (\frac{\sum_{j>d} \lambda_j}{n})^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 + \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \\
&\asymp (\frac{\sum_{j>d} \lambda_j}{n})^2 \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \quad (\text{from } \|\theta_{(d+1):p}^*\|_{\Sigma_{(d+1):p}}^2 \ll \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 (\frac{1}{\lambda_d} + \frac{1}{\frac{\sum_{j>d} \lambda_j}{n}})^{-2}), \\
\mathcal{V}_{\text{out}}(\tau; \hat{H}_n, \gamma) &\asymp \sigma^2 (\frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n(\frac{\sum_{j>d} \lambda_j}{n})^2}).
\end{aligned}$$

That is,

$$\begin{aligned}
\mathcal{B}_{\text{out}}(\tau, \hat{H}_n, \hat{G}_n, \gamma) + \mathcal{V}_{\text{out}}(\tau; \hat{H}_n, \gamma) &\asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{\sum_{j>d} \lambda_j}{n} \right)^2 + \sigma^2 \left(\frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n(\frac{\sum_{j>d} \lambda_j}{n})^2} \right) \\
&\asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \left(\frac{\sum_{j>d} \lambda_j}{n} + \tau \right)^2 + \sigma^2 \left(\frac{d}{n} + \sum_{j>d} \frac{\lambda_j^2}{n(\frac{\sum_{j>d} \lambda_j}{n} + \tau)^2} \right) \\
&\quad (\text{from } \sum_{j>d} \lambda_j \geq n\tau).
\end{aligned}$$

□

IV.3 Alternative calculation of error approximation formulas

The asymptotic out-sample and in-sample errors can also be calculated using a distributional approximation method in Han & Shen (2023) under the independent components assumption. By letting $\mu = \sqrt{n}\Sigma^{1/2}\theta$, the ridge estimator in (2) can be equivalently formulated as $\hat{\theta}(\tau) = \frac{1}{\sqrt{n}}\Sigma^{-1/2}\hat{\mu}(\tau)$ with

$$\hat{\mu}(\tau) = \arg \max_{\mu \in \mathbb{R}^p} \left\{ \|Y - Z\mu\|^2 + \tau \|\Sigma^{-1/2}\mu\|^2 \right\},$$

where $Z = \frac{1}{\sqrt{n}}X\Sigma^{-1/2}$. The rows in Z are covariate vectors with covariance matrix $\frac{1}{n}I_p$. Let $\mu^* = \sqrt{n}\Sigma^{1/2}\theta^*$. Following Han & Shen (2023), for $z \in \mathbb{R}^p, \tau > 0$, we define

$$\psi_{\tilde{\lambda}}(z, \tau) = \operatorname{argmin}_{x \in \mathbb{R}^p} \left\{ \frac{1}{2} \|x - z\|^2 + \frac{\tau}{2} \|\Sigma^{-1/2}x\|^2 \right\} = \left(\frac{z_1}{1 + \frac{\tau}{\lambda_1}}, \dots, \frac{z_p}{1 + \frac{\tau}{\lambda_p}} \right)^T.$$

For an isotropic random vector $z_0 \in \mathbb{R}^p$, suppose (α, β) is a unique solution in $(0, \infty)^2$ to the following equations:

$$\begin{aligned}
\beta^2 - \sigma^2 &= \frac{1}{n} \mathbb{E} \|\psi_{\tilde{\lambda}}(\mu^* + \beta z_0, \alpha\tau) - \mu^*\|^2 = \frac{1}{n} \sum_j \frac{\alpha^2 \tau^2 \mu_j^{*2}}{(\lambda_j + \alpha\tau)^2} + \left(\frac{1}{n} \sum_j \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} \right) \beta^2, \quad (\text{S123}) \\
\frac{1}{\alpha} &= 1 - \frac{1}{n} \mathbb{E} \operatorname{div} \psi_{\tilde{\lambda}}(\mu^* + \beta z_0, \alpha\tau) = \gamma \frac{1}{p} \sum_{j=1}^p \frac{1}{1 + \frac{\alpha\tau}{\lambda_j}}.
\end{aligned}$$

where $\operatorname{div} f(x_1, \dots, x_p) = \sum_{j=1}^p \frac{\partial f}{\partial x_j}$. Note that from (S123), we have

$$\begin{aligned}
\beta^2 &= \left(1 - \frac{1}{n} \sum_j \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} \right)^{-1} \left(\sigma^2 + \frac{1}{n} \sum_j \frac{\alpha^2 \tau^2 \mu_j^{*2}}{(\lambda_j + \alpha\tau)^2} \right) \\
&= \left(1 - \frac{1}{n} \sum_j \frac{\lambda_j^2}{(\lambda_j + \alpha\tau)^2} \right)^{-1} \left(\sigma^2 + \sum_j \frac{\alpha^2 \tau^2 \lambda_j^2 \theta_j^{*2}}{(\lambda_j + \alpha\tau)^2} \right). \quad (\text{S124})
\end{aligned}$$

Then under suitable conditions (see conditions (R1)-(R3) in Han & Shen (2023)), the distributions of $\hat{\mu} - \mu^*$ and $Z(\hat{\mu} - \mu^*)$ can be approximated as follows:

$$\begin{aligned}\hat{\mu} - \mu^* &\stackrel{d}{\approx} \psi_{\tilde{\lambda}}(\mu^* + \beta z_0, \alpha\tau) - \mu^*, \\ Z(\hat{\mu} - \mu^*) &\stackrel{d}{\approx} \left(1 - \frac{1}{\alpha}\right)\xi + \frac{\sqrt{\gamma^2 - \sigma^2}}{\alpha}h,\end{aligned}\tag{S125}$$

where $\xi \sim N(0, \sigma^2)$ and $h \sim N(0, \sigma^2 I_n)$.

Given the approximation results (S125), the asymptotic error formulas in Corollary 7 can also be calculated as follows. The out-sample error can be approximated by

$$\begin{aligned}\|\hat{\theta}(\tau) - \theta^*\|_{\Sigma}^2 &= \frac{1}{n}\|\hat{\mu} - \mu^*\|^2 \\ &\approx \frac{1}{n}\|\psi_{\tilde{\lambda}}(\mu^* + \beta z_0, \alpha\tau) - \mu^*\|^2 \\ &\approx \frac{1}{n}\mathbb{E}\|\psi_{\tilde{\lambda}}(\mu^* + \beta z_0, \alpha\tau) - \mu^*\|^2 \\ &= \beta^2 - \sigma^2.\end{aligned}\tag{S126}$$

Substituting (S124) into (S126) yields the sum of (14) and (15). The in-sample error can be approximated by

$$\begin{aligned}\|\hat{\theta}(\tau) - \theta^*\|_{\Sigma}^2 &= \frac{1}{n}\|Z(\hat{\mu} - \mu^*)\|^2 \\ &\approx \left(1 - \frac{1}{\alpha}\right)^2 \sigma^2 + \frac{\beta^2 - \sigma^2}{\alpha^2}.\end{aligned}\tag{S127}$$

Substituting (S124) into (S127) yields the sum of (16) and (17).

V Comparison with Bunea et al. (2022)

V.1 Approximations of terms

In this section, we give the approximations of certain terms used in the comparison in Section 4.3 between our Theorem 3 and Theorem 16 of Bunea et al. (2022). In the setting described in Section 4.3, we show that the following approximations hold: $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2$, $\|\beta\|_{\Sigma_Z}^2 \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2$, $\frac{\lambda_{d+1}}{\lambda_d - \lambda_{d+1}} \asymp \frac{\lambda_{d+1}}{\lambda_d}$ and $r_0(\Sigma_E) \asymp r_d(\Sigma)$.

$$\textcircled{1} \quad \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2$$

First, we have

$$\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2 \geq \|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \geq \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2 \frac{\lambda_d^2}{\lambda_1^2}.$$

Then from $\lambda_1 \asymp \lambda_d$, we have $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2$.

$$\textcircled{2} \quad \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2 \asymp \|\beta\|_{\Sigma_Z}^2 \quad \text{and} \quad \frac{\lambda_{d+1}}{\lambda_d - \lambda_{d+1}} \asymp \frac{\lambda_{d+1}}{\lambda_d}$$

From $\lambda_1 \geq \dots \geq \lambda_d \geq c_1 \lambda_{d+1}$ for some $c_1 > 1$, we have for $i = 1, \dots, d$,

$$1 \geq \frac{\lambda_i - \lambda_{d+1}}{\lambda_i} \geq 1 - \frac{1}{c_1},$$

$$\frac{\lambda_i}{\lambda_{d+1}} \geq \frac{\lambda_i - \lambda_{d+1}}{\lambda_{d+1}} \geq (1 - \frac{1}{c_1}) \frac{\lambda_i}{\lambda_{d+1}}.$$

Then for $i = 1, \dots, d$,

$$\frac{\lambda_i - \lambda_{d+1}}{\lambda_i} \asymp 1, \tag{S128}$$

$$\frac{\lambda_i - \lambda_{d+1}}{\lambda_i} \asymp \frac{\lambda_i}{\lambda_{d+1}}. \tag{S129}$$

From (S129), we have $\frac{\lambda_{d+1}}{\lambda_d - \lambda_{d+1}} \asymp \frac{\lambda_{d+1}}{\lambda_d}$. From (S128), we have

$$\begin{aligned} \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2 &= \beta^T \text{Diag}\left(\frac{\lambda_1 - \lambda_d}{\lambda_1}, \dots, \frac{\lambda_d - \lambda_{d+1}}{\lambda_d}\right) \beta \\ &\asymp \beta^T \beta \\ &= \|\beta\|_{\Sigma_Z}^2 \quad (\text{with } \Sigma_Z = I_d). \end{aligned}$$

$$\textcircled{3} \quad r_0(\Sigma_E) \asymp r_d(\Sigma)$$

By definition, $r_0(\Sigma_E) = d + r_d(\Sigma)$. With $r_d(\Sigma) > c_2 d$, we have

$$(1 + c_2)r_d(\Sigma) \geq r_0(\Sigma_E) \geq r_d(\Sigma)$$

That is, $r_d(\Sigma) \asymp r_0(\Sigma_E)$.

V.2 Error bounds incorporated with approximations

In this section, we give the error order of our Theorem 3 and the error upper bound in Theorem 16 of Bunea et al. (2022) for the min-norm interpolator and incorporate the approximations of terms in Supplement Section V.1.

Based on Theorem 3, for $\tau = 0$, we have

$$\begin{aligned} \text{MSE}_{\text{out}} &\asymp \underbrace{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d^2(\Sigma)}{n^2}}_{\text{B}_{\text{out}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{n r_d(\Sigma^2)}{r_d^2(\Sigma)} \right)}_{\text{V}_{\text{out}}}, \quad \text{for } \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_d, \\ \text{MSE}_{\text{out}} &\gtrsim \underbrace{\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2}_{\text{B}_{\text{out}}}, \quad \text{for } \lambda_{d+1} \frac{r_d(\Sigma)}{n} > \lambda_d. \end{aligned}$$

Incorporating $\|\theta_{1:d}^*\|_{\Sigma_{1:d}^{-1}}^2 \lambda_d^2 \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2$, we have

$$\text{MSE}_{\text{out}} \asymp \underbrace{\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2 \frac{\lambda_{d+1}^2}{\lambda_d^2} \frac{r_d^2(\Sigma)}{n^2}}_{\text{B}_{\text{out}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{n r_d(\Sigma^2)}{r_d^2(\Sigma)} \right)}_{\text{V}_{\text{out}}}, \quad \text{for } \lambda_{d+1} \frac{r_d(\Sigma)}{n} \leq \lambda_d,$$

$$\text{MSE}_{\text{out}} \gtrsim \underbrace{\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2}_{\text{B}_{\text{out}}}, \quad \text{for } \lambda_{d+1} \frac{r_d(\Sigma)}{n} > \lambda_d.$$

Based on Theorem 16 of Bunea et al. (2022), we have

$$\text{MSE}_{\text{out}} \lesssim \underbrace{\|\beta\|_{\Sigma_Z}^2 \frac{\lambda_{d+1}}{\lambda_d(A\Sigma_Z A^T)} \frac{r_0(\Sigma_E)}{n}}_{\text{B}_{\text{out}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{n}{r_0(\Sigma_E)} \right)}_{\text{V}_{\text{out}}}.$$

Incorporating $\|\beta\|_{\Sigma_Z}^2 \asymp \|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2$, $\frac{\lambda_{d+1}}{\lambda_d - \lambda_{d+1}} \asymp \frac{\lambda_{d+1}}{\lambda_d}$ and $r_0(\Sigma_E) \asymp r_d(\Sigma)$, we have

$$\text{MSE}_{\text{out}} \lesssim \underbrace{\|\theta_{1:d}^*\|_{\Sigma_{1:d}}^2 \frac{\lambda_{d+1}}{\lambda_d} \frac{r_d(\Sigma)}{n}}_{\text{B}_{\text{out}}} + \underbrace{\sigma^2 \left(\frac{d}{n} + \frac{n}{r_d(\Sigma)} \right)}_{\text{V}_{\text{out}}}.$$