Supplementary Material for "Uniform Consistency of Cross-Validation Estimators for High-Dimensional Ridge Regression"

Pratik Patil **Yuting Wei** **Alessandro Rinaldo** **Ryan J. Tibshirani**

Carnegie Mellon University

This supplementary document contains proofs of the theorems and lemmas in the paper "Uniform Consistency of Cross-Validation Estimators for High-Dimensional Ridge Regression." All section and equation numbers in this document begin with the letter "S" to differentiate them from those appearing in the main paper.

The content of this supplement is organized as follows. In Section S.1, we provide proofs of the constituent Lemmas 5.1 to 5.4 related to Theorem 4.1 in the main paper, along with the remaining steps to complete the proof of Theorem 4.1. In Section S.2, we provide proof of the constituent Lemma 5.6 related to Theorem 4.2 in the main paper, along with the remaining steps to complete the proof of Theorem 4.2. In Section S.3, we list and prove auxiliary lemmas that we need in other proofs. Finally, in Section S.4, we list useful concentration results that are used in the proofs throughout.

A table of content for this supplement is collected below for ease of referring.

Contents

S.1 Proofs related to Theorem 4.1

S.1.1 Proof of Lemma 5.1

Recall from Equation (2) that the expected out-of-sample prediction error of the ridge estimator $\widehat{\beta}_\lambda$ is defined as

$$\mathrm{Err}(\widehat{\beta}_\lambda) = \mathbb{E}_{x_0,y_0}\left[(y_0 - x_0^T\widehat{\beta}_\lambda)^2 \mid X, y\right].$$

Under a well-specified linear response $y_0 = x_0^T\beta_0 + \varepsilon_0$, the prediction error can be decomposed as

$$\begin{aligned}\mathrm{Err}(\widehat{\beta}_\lambda) &= \mathbb{E}\left[(\beta_0 - \widehat{\beta}_\lambda)^T x_0 x_0^T(\beta_0 - \widehat{\beta}_\lambda) \mid X, y\right] + \mathbb{E}\left[(\beta_0 - \widehat{\beta}_\lambda)^T x_0\varepsilon_0 \mid X, y\right] + \mathbb{E}\left[\varepsilon_0^2 \mid X, y\right] \\ &= (\beta_0 - \widehat{\beta}_\lambda)^T\Sigma(\beta_0 - \widehat{\beta}_\lambda) + \sigma^2. \end{aligned} \tag{S.1}$$

Here we used the fact that $\mathbb{E}\big[x_0\varepsilon_0\big] = 0$ as ε_0 is independent of x_0. Using the expression of $\widehat{\beta}_\lambda$ from Equation (1), the deviation $\beta_0 - \widehat{\beta}_\lambda$ can be expressed as

$$\begin{aligned}\beta_0 - \widehat{\beta}_\lambda &= \beta_0 - (X^TX/n + \lambda I_p)^+X^Ty/n \\ &= \beta_0 - (X^TX/n + \lambda I_p)^+X^T(X\beta_0 + y - X\beta_0)/n \\ &= \big(I_p - (X^TX/n + \lambda I_p)^+X^TX/n\big)\beta_0 - (X^TX/n + \lambda I_p)^+X^T\varepsilon/n.\end{aligned}$$

Note that the first component depends on the signal parameter β_0 and the second depends on the error vector ε. Plugging this into (S.1), and denoting X^TX/n by $\widehat{\Sigma}$ and $\mathrm{Err}(\widehat{\beta}(\lambda))$ by $\mathrm{err}(\lambda)$, we have the following decomposition of the prediction error for any $\lambda \in \mathbb{R}$:

$$\mathrm{err}(\lambda) = \mathrm{err}_b(\lambda) + \mathrm{err}_c(\lambda) + \mathrm{err}_v(\lambda), \tag{S.2}$$

where $\mathrm{err}_b(\lambda)$, $\mathrm{err}_v(\lambda)$, and $\mathrm{err}_c(\lambda)$ are the bias, variance, and cross components in the decomposition given by

$$\begin{aligned}\mathrm{err}_b(\lambda) &= \beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0, \\ \mathrm{err}_c(\lambda) &= -2\beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+X^T\varepsilon/n, \\ \mathrm{err}_v(\lambda) &= \varepsilon^T\big(X(\widehat{\Sigma} + \lambda I_p)^+\Sigma(\widehat{\Sigma} + \lambda I_p)^+X^T/n\big)\varepsilon/n + \sigma^2.\end{aligned}$$

For any $\lambda \in (\lambda_{\min}, \infty)$, we establish below that

$$\mathrm{err}_c(\lambda) \xrightarrow{\text{a.s.}} 0 \tag{S.3}$$

under proportional asymptotic limit. The desired decomposition in Lemma 5.1 then follows by plugging convergence in (S.3) into (S.2).

To establish the convergence in (S.3), let us write $\mathrm{err}_c(\lambda) = a_n^T\varepsilon/n$ where $a_n \in \mathbb{R}^n$ is a function of X and β_0 given by

$$a_n = -2X(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0.$$

We note that for $\lambda \in (\lambda_{\min}, \infty)$,

$$\begin{aligned}\|a_n\|^2/n &= 4\beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0 \\ &\le C\left\|\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\right\| \\ &\le C,\end{aligned}$$

where the first inequality uses bound on the signal energy from Assumption 4 and the second inequality holds almost surely for large n by using the facts that $\|\widehat{\Sigma}\| \le C(\sqrt{\gamma} + 1)^2\|\Sigma\|$, $\|(\widehat{\Sigma} + \lambda I_p)^+\| \le (\lambda - \lambda_{\min})^{-1}$ almost surely for n large enough from Assumption 2 and $\|\Sigma\| \le r_{\max}$ from Assumption 3. In addition, ε has i.i.d. entries satisfying Assumption 1. The desired result then follows from application of Lemma S.4.1.

S.1.2 Proof of Lemma 5.2

We start by writing the GCV risk estimate $\mathrm{gcv}(\lambda)$ for the ridge estimator from Equation (5) as

$$\mathrm{gcv}(\lambda) = \frac{y^T(I_n - L_\lambda)^2 y/n}{\big(1 - \mathrm{tr}[L_\lambda]/n\big)^2} \tag{S.4}$$

where L_λ is the ridge smoothing matrix. Note that (S.4) is of the form $\frac{0}{0}$ when $L_\lambda = I_n$ (which happens when $\lambda = 0$ and X has rank n). In this case, we define the GCV risk estimate as the corresponding limit as $\lambda \to 0$. We handle this case separately below.

The denominator of (S.4) can be expressed as

$$\begin{aligned} 1 - \mathrm{tr}[L_\lambda]/n &= 1 - \mathrm{tr}\left[X(X^TX/n + \lambda I_p)^+ X^T/n\right]/n \\ &= 1 - \mathrm{tr}\left[(X^TX/n + \lambda I_p)^+ X^TX/n\right]/n. \end{aligned}$$

The numerator of (S.4) can be expressed as

$$\begin{aligned} y^T(I_n - L_\lambda)^2 y/n &= (X\beta_0 + \varepsilon)^T(I_n - L_\lambda)^2(X\beta_0 + \varepsilon)/n \\ &= \beta_0^T X^T(I_n - L_\lambda)^2 X\beta_0/n + 2\beta_0^T X^T(I_n - L_\lambda)^2\varepsilon/n + \varepsilon^T(I_n - L_\lambda)^2\varepsilon/n. \end{aligned}$$

Consider the first term of the numerator expression. The factor $X^T(I_n - L_\lambda)^2 X$ can be expressed as

$$\begin{aligned} X^T(I_n - L_\lambda)^2 X &= X^T\big(I_n - X(X^TX/n + \lambda I_p)^+ X^T/n\big)^2 X \\ &= \big(X^T - X^TX/n(X^TX/n + \lambda I_p)^+ X^T\big)\big(X - X(X^TX/n + \lambda I_p)^+ X^TX/n\big) \\ &= \big(I_p - X^TX/n(X^TX/n + \lambda I)^+\big)X^TX\big(I_p - (X^TX/n + \lambda I_p)^+ X^TX/n\big). \end{aligned}$$

Consider the second term of the numerator expression. The factor $X^T(I_n - L_\lambda)^2$ can be expressed as

$$\begin{aligned} X^T(I_n - L_\lambda)^2 &= X^T\big(I_n - X(X^TX/n + \lambda I_p)^+ X^T/n\big)^2 \\ &= \big(X^T - X^TX/n(X^TX/n + \lambda I_p)^+ X^T\big)\big(I_n - X(X^TX/n + \lambda I_p)^+ X^T/n\big) \\ &= \big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)X^T\big(I_n - X(X^TX/n + \lambda I_p)^+ X^T/n\big) \\ &= \big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)\big(X^T - X^TX/n(X^TX/n + \lambda I_p)^+ X^T\big) \\ &= \big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)\big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)X^T \end{aligned}$$

Consider the third term of the numerator expansion. The factor $(I_n - L_\lambda)^2$ can be expressed as

$$(I_n - L_\lambda)^2 = \big(I_n - X(X^TX/n + \lambda I_p)^+ X^T/n\big)^2$$

Case when $\lambda \neq 0$. The GCV denominator $1 - \mathrm{tr}\left[(X^TX/n + \lambda I_p)^+ X^TX/n\right]/n \neq 0$ when $\lambda \neq 0$. Thus plugging the denominator and numerator expansions into (S.4) and denoting X^TX/n by $\widehat{\Sigma}$, the GCV risk estimate can be decomposed as

$$\mathrm{gcv}(\lambda) = \frac{\mathrm{gcv}_b(\lambda) + \mathrm{gcv}_c(\lambda) + \mathrm{gcv}_v(\lambda)}{\mathrm{gcv}_d(\lambda)}, \tag{S.5}$$

where $\mathrm{gcv}_b(\lambda)$, $\mathrm{gcv}_v(\lambda)$, and $\mathrm{gcv}_c(\lambda)$ are the bias-like, variance-like, and cross components in the decomposition given by

$$\begin{aligned} \mathrm{gcv}_b(\lambda) &= \beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0, \\ \mathrm{gcv}_c(\lambda) &= 2\beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)^2 X^T\varepsilon/n, \\ \mathrm{gcv}_v(\lambda) &= \varepsilon^T\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big)^2\varepsilon/n, \end{aligned}$$

and $\mathrm{gcv}_d(\lambda)$ is the normalization factor given by

$$\mathrm{gcv}_d(\lambda) = \big(1 - \mathrm{tr}[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+]/n\big)^2.$$

Similar to the proof of Lemma 5.1, we now establish that

$$\mathrm{gcv}_c(\lambda) \xrightarrow{\text{a.s.}} 0 \tag{S.6}$$

under proportional asymptotic limit. Let us write $\mathrm{gcv}_c(\lambda) = b_n^T \varepsilon/n$ where $b_n \in \mathbb{R}^n$ is a function of X and β_0 given by

$$b_n = 2X\big(I_p - (\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big)^2 \beta_0.$$

As argued in the proof of Lemma 5.1, for $\lambda \in (\lambda_{\min}, \infty)$,

$$\begin{aligned}
\|b_n\|^2/n &= 4\beta_0^T \big(I_p - (\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big)^2 \widehat{\Sigma} \big(I_p - (\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big)^2 \beta_0 \\
&\le C \left\| \big(I_p - (\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big)^2 \widehat{\Sigma} \big(I_p - (\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big)^2 \right\| \\
&\le C
\end{aligned}$$

almost surely for large n, and since ε has i.i.d. entries satisfying Assumption 1, the convergence in (S.6) follow from application of Lemma S.4.1.

Limiting case when $\lambda = 0$. To handle the case when $\mathrm{gcv}_d(\lambda)$ can be zero, we note that when $\lambda \neq 0$ using Lemma S.3.2 the components in the decomposition (S.5) can be alternately expressed as

$$\mathrm{gcv}_b(\lambda) = \beta_0^T \lambda^2 (\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma} (\widehat{\Sigma} + \lambda I_p)^+ \beta_0,$$

$$\mathrm{gcv}_b(\lambda) = 2\lambda^2 \beta_0^T (\widehat{\Sigma} + \lambda I_p)^+ (\widehat{\Sigma} + \lambda I_p)^+ X^T \varepsilon/n,$$

$$\mathrm{gcv}_v(\lambda) = \lambda^2 \varepsilon^T (XX^T/n + \lambda I_n)^+ (XX^T/n + \lambda I_n)^+ \varepsilon,$$

$$\mathrm{gcv}_d(\lambda) = \lambda^2 \big(\mathrm{tr}[(XX^T/n + \lambda I_n)^+]/n\big)^2.$$

We can then cancel the factor of λ^2 and take the limit $\lambda \to 0$ to get the limiting GCV decomposition as

$$\mathrm{gcv}(0) = \frac{\mathrm{gcv}_b(0) + \mathrm{gcv}_b(0) + \mathrm{gcv}_v(0)}{\mathrm{gcv}_d(0)}, \tag{S.7}$$

where the limiting bias-like, variance-like and cross components in the decomposition are given by

$$\mathrm{gcv}_b(0) = \beta_0^T \widehat{\Sigma}^+ \widehat{\Sigma} \widehat{\Sigma}^+ \beta_0 = \beta_0^T \widehat{\Sigma}^+ \beta_0,$$

$$\mathrm{gcv}_c(0) = 2\beta_0^T \widehat{\Sigma}^{+2} X^T \varepsilon/n,$$

$$\mathrm{gcv}_v(0) = \varepsilon^T (XX^T/n)^{+2} \varepsilon/n,$$

and the limiting normalization can be written as

$$\mathrm{gcv}_d(0) = \big(\mathrm{tr}[\widehat{\Sigma}^+]/n\big)^2$$

by noting that $\mathrm{tr}[(XX^T/n)^+] = \mathrm{tr}[(X^TX/n)^+]$. As before, let us establish that

$$\mathrm{gcv}_c(0) \xrightarrow{\text{a.s.}} 0 \tag{S.8}$$

under proportional asymptotics. We write $\mathrm{gcv}_c(0) = b_n^T \varepsilon/n$ where $b_n \in \mathbb{R}^n$ is a function of X and β_0 given by

$$b_n = 2X\widehat{\Sigma}^{+2} \beta_0.$$

We note that $\|b_n\|^2/n$ is almost surely bounded for large n and ε contains i.i.d. entries satisfying Assumption 1. Using Lemma S.4.1, we conclude the convergence.

The desired decomposition in Lemma 5.2 then follows by using the convergences in (S.6) and (S.8) into (S.5) and (S.7), respectively.

S.1.3 Proof of Lemma 5.3

We start with $\mathrm{gcv}_b(\lambda)$ and first establish that

$$\beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0-\frac{\beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0}{\Big(1+\mathrm{tr}\left[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\right]/n\Big)^2}\xrightarrow{\text{a.s.}}0. \quad \text{(S.9)}$$

To that end, let $B:=\beta_0\beta_0^T$ and break the left-hand side into sum of quadratic forms evaluated at the n observations as follows:

$$\begin{aligned}
\beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0 &= \mathrm{tr}\left[B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\right]\\
&= \mathrm{tr}\left[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\right]\\
&= \mathrm{tr}\left[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\sum_{i=1}^n x_ix_i^T/n\right]\\
&= \frac{1}{n}\sum_{i=1}^n \mathrm{tr}\left[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)x_ix_i^T\right]\\
&= \frac{1}{n}\sum_{i=1}^n x_i^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)x_i.
\end{aligned}$$

The summands $x_i^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)x_i$ are quadratic forms where the point of evaluation x_i and the matrix $\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)$ are dependent. To break the dependence, we use the standard leave-one-out trick and the Sherman-Morrison-Woodbury formula with Moore-Penrose pseudo-inverse (Meyer, 1973). Let us temporarily call $w_i:=B(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+)x_i$ and proceed as follows:

$$\begin{aligned}
&x_i^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)x_i\\
&= w_i^T(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+)x_i\\
&= w_i^T\big(I_p-(\widehat{\Sigma}_{-i}+x_ix_i^T/n)(\widehat{\Sigma}_{-i}+\lambda I_p+x_ix_i^T/n)^+\big)x_i\\
&= w_i^T\left(I_p-(\widehat{\Sigma}_{-i}+x_ix_i^T/n)\left((\widehat{\Sigma}_{-i}+\lambda I_p)^+-\frac{(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_ix_i^T/n(\widehat{\Sigma}_{-i}+\lambda I_p)^+}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}\right)\right)x_i\\
&= w_i^Tx_i-w_i^T\big(\widehat{\Sigma}_{-i}+x_ix_i^T/n\big)\left((\widehat{\Sigma}_{-i}+\lambda I_p)^+-\frac{(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_ix_i^T/n(\widehat{\Sigma}_{-i}+\lambda I_p)^+}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}\right)x_i\\
&= w_i^Tx_i-w_i^T\big(\widehat{\Sigma}_{-i}+x_ix_i^T/n\big)\left((\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i-\frac{(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_ix_i^T/n(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}\right)\\
&= w_i^Tx_i-w_i^T\big(\widehat{\Sigma}_{-i}+x_ix_i^T/n\big)\left(\frac{(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i+(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_ix_i^T(\widehat{\Sigma}+\lambda I_p)^+x_i/n-(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_ix_i^T/n(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}\right)\\
&= w_i^Tx_i-\frac{w_i^T(\widehat{\Sigma}_{-i}+x_ix_i^T/n)(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}\\
&= \frac{w_i^Tx_i+w_i^Tx_ix_i^T(\widehat{\Sigma}+\lambda I_p)^+x_i/n-w_i^T\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i-w_i^Tx_ix_i^T/n(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}\\
&= \frac{w_i^Tx_i-w_i^T\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}\\
&= \frac{w_i^T(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+)x_i}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}\\
&= \frac{x_i^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}.
\end{aligned}$$

By carrying our similar leave-one-out strategy on the other side, we can further simplify

$$\frac{x_i^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i}{1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n}=\frac{x_i^T\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i}{\big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n\big)^2}.$$

We now split the error to the target in (S.9) as follows:

$$\begin{aligned}
&\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\Big]-\frac{\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\big]/n\Big)^2}\\
&=\frac{1}{n}\sum_{i=1}^n\frac{x_i^T\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i}{\big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n\big)^2}-\frac{\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\big]/n\Big)^2}\\
&=e_1+e_2,\ \text{where}
\end{aligned}$$

$$e_1:=\frac{1}{n}\sum_{i=1}^n\left(\frac{x_i^T\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i}{\Big(1+x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+x_i/n\Big)^2}-\frac{\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^+\Sigma\big]/n\Big)^2}\right),$$

$$e_2:=\frac{1}{n}\sum_{i=1}^n\left(\frac{\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1+\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^+\Sigma\big]/n\Big)^2}-\frac{\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\big]/n\Big)^2}\right).$$

In Section S.1.6, we show that both terms e_1 and e_2 almost surely approach 0 under proportional asymptotics.

Let us provide some intuition as follows. On one hand, in the error term e_1, conditional on X_{-i}, expected value of $x_i^T\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)x_i$ is $\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)\Sigma\Big]$ and the expected value of $x_i^T(\widehat{\Sigma}_{-i}+\lambda I)^+x_i/n$ is $\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I)^+\Sigma\big]/n$. Because of concentration of these quantities around their respective expectations rapid enough, the error term e_1 is almost surely 0. On the other hand, for e_2, $\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+\big)\Sigma\Big]$ and $\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]$, and $\operatorname{tr}\big[(\widehat{\Sigma}_{-i}+\lambda I_p)^+\Sigma\big]/n$ and $\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\big]/n$, the matrices involved differ by rank-1 component. The difference is almost surely 0 in the proportional asymptotic limit. We note that this strategy is similar to the ones used by, for example, Rubio and Mestre (2011); Ledoit and Peche (2009) to obtain expressions for certain functionals involving Σ and $\widehat{\Sigma}$ in terms of Σ. The main difference is that the eventual target in our case is defined solely in terms of $\widehat{\Sigma}$ rather than Σ.

We have so far established that

$$\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\Big]-\frac{\operatorname{tr}\Big[\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)B\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\Big]}{\Big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I)^+\Sigma\big]/n\Big)^2}\xrightarrow{\text{a.s.}}0,$$

which after expressing B in terms of β_0 and moving the denominator across yields

$$\Big(1+\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I)^+\Sigma\big]/n\Big)^2\beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0-\beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0\xrightarrow{\text{a.s.}}0. \tag{S.10}$$

Case when $\lambda\neq 0$. We now use the $\lambda\neq 0$ case of Lemma S.3.1 to get

$$\frac{\beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\widehat{\Sigma}\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0}{\Big(1-\operatorname{tr}\big[(\widehat{\Sigma}+\lambda I_p)^+\widehat{\Sigma}\big]/n\Big)^2}-\beta_0^T\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\big(I_p-\widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0\xrightarrow{\text{a.s.}}0$$

under proportional asymptotics as desired.

Limiting case when $\lambda = 0$. To handle the $\lambda = 0$ case, we first express $I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+ = \lambda(\widehat{\Sigma} + \lambda I_p)^+$ when $\lambda \neq 0$ using Lemma S.3.2. We can then move factor of λ^2 from $\beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0$ to $\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I)^+\Sigma\big]/n\Big)^2$ such that

$$\begin{aligned}
&\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I)^+\Sigma\big]/n\Big)^2 \beta_0^T\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\beta_0 \\
&= \Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I)^+\Sigma\big]/n\Big)^2 \lambda^2 \beta_0^T(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\beta_0 \\
&= \Big(\lambda + \lambda \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I)^+\Sigma\big]/n\Big)^2 \beta_0^T(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\beta_0 \\
&= \Big(\lambda + \operatorname{tr}\big[\lambda(\widehat{\Sigma} + \lambda I)^+\Sigma\big]/n\Big)^2 \beta_0^T(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\beta_0 \\
&= \Big(\lambda + \operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma\Big]/n\Big)^2 \beta_0^T(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\beta_0.
\end{aligned}$$

Using the above expression in (S.10) and sending $\lambda \to 0$ thus yields

$$\Big(\operatorname{tr}\big[(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma\big]/n\Big)^2 \beta_0^T\widehat{\Sigma}^+\widehat{\Sigma}\widehat{\Sigma}^+\beta_0 - \beta_0^T(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\beta_0 \xrightarrow{\text{a.s.}} 0,$$

or in other words,

$$\Big(\operatorname{tr}\big[(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma\big]/n\Big)^2 \beta_0^T\widehat{\Sigma}^+\beta_0 - \beta_0^T(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\beta_0 \xrightarrow{\text{a.s.}} 0.$$

Using Lemma S.3.1 for this case, we then have

$$\frac{\beta_0^T\widehat{\Sigma}^+\beta_0}{\big(\operatorname{tr}[\widehat{\Sigma}^+]/n\big)^2} - \beta_0^T(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\beta_0 \xrightarrow{\text{a.s.}} 0$$

under proportional asymptotics, completing both the cases in Lemma 5.3.

S.1.4 Proof of Lemma 5.4

Case when $\lambda \neq 0$. Under proportional asymptotic limit, our goal is to show that

$$\varepsilon^T\big(X(\widehat{\Sigma} + \lambda I_p)^+\Sigma(\widehat{\Sigma} + \lambda I_p)^+X^T/n\big)\varepsilon/n + \sigma^2 - \frac{\varepsilon^T\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+X^T/n\big)^2\varepsilon/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\big)^2} \xrightarrow{\text{a.s.}} 0.$$

We first note that $\varepsilon^T\varepsilon/n$ almost surely approaches σ^2 from the strong law of large numbers. Thus we can slightly rephrase our goals to show as

$$\varepsilon^T\left[\big(X(\widehat{\Sigma} + \lambda I_p)^+\Sigma(\widehat{\Sigma} + \lambda I_p)^+X^T/n\big) + I_n - \frac{\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+X^T/n\big)^2}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\big)^2}\right]\varepsilon/n \xrightarrow{\text{a.s.}} 0.$$

Our main strategy is to show that under proportional asymptotic limit

$$\operatorname{tr}\big[X(\widehat{\Sigma} + \lambda I_p)^+\Sigma(\widehat{\Sigma} + \lambda I_p)^+X^T/n\big]/n + 1 - \frac{\operatorname{tr}\Big[\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+X^T/n\big)^2\Big]/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\big)^2} \xrightarrow{\text{a.s.}} 0. \tag{S.11}$$

The desired convergence then follows by using Lemma S.4.2.

We proceed by decomposing the first component of (S.11) as follows:

$$\begin{aligned}
\operatorname{tr}\big[X(\widehat{\Sigma} + \lambda I_p)^+\Sigma(\widehat{\Sigma} + \lambda I_p)X^T/n\big]/n &= \operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\Sigma(\widehat{\Sigma} + \lambda I_p)^+\big]/n \\
&= \operatorname{tr}\big[\Sigma(\widehat{\Sigma} + \lambda I_p)^+\big]/n - \operatorname{tr}\big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+\big]/n.
\end{aligned}$$

For the numerator of the second component of (S.11), we note that

$$
\begin{aligned}
&\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big)^2 \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big)\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big) \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big) - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big) \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big) - X(X^TX/n + \lambda I_p)^+ X^T/n\big(I_n - X(X^TX/n + \lambda I_p)^+ X^T/n\big) \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big) - X(X^TX/n + \lambda I_p)^+\big(X^T/n - X^TX/n(X^TX/n + \lambda I_p)^+ X^T/n\big) \\
&= \big(I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big) - X(X^TX/n + \lambda I_p)^+\big(I_p - X^TX/n(X^TX/n + \lambda I_p)^+\big)X^T/n.
\end{aligned}
$$

Thus we have

$$
\begin{aligned}
&\frac{\operatorname{tr}\big[I_n - X(\widehat{\Sigma} + \lambda I_p)^+ X^T/n\big]^2/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\big)^2} \\
&= \frac{1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n - \operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\big]/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\big)^2} \\
&= \frac{1}{1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n} - \frac{\operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+)\big]/n}{\big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\big)^2}.
\end{aligned}
$$

To establish the desired equivalence, we now use the following two individual equivalences:

$$
\operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\big]/n - \frac{1}{1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n} + 1 \xrightarrow{\text{a.s.}} 0,
$$

which follows from Lemma S.3.1, and

$$
\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\Sigma(\widehat{\Sigma} + \lambda I_p)^+\Big]/n - \frac{\operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\Big]/n}{\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n\Big)^2} \xrightarrow{\text{a.s.}} 0,
$$

which follows analogously from the equivalence established in the proof of Lemma 5.3 with $B = I_p$.

Limiting case when $\lambda = 0$. To handle the case when $\lambda = 0$, we observe that when $\lambda \neq 0$, we can write

$$
\operatorname{tr}\big[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big)\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\big]/n = 1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n + \lambda^2 \operatorname{tr}\big[(XX^T/n + \lambda I_n)^{+2}\big]/n,
$$

along with

$$
1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+\widehat{\Sigma}\big]/n = \lambda \operatorname{tr}\big[(XX^T/n + \lambda I_n)^+\big]/n,
$$

which follow from Lemma S.3.2. This allows us to cancel the factor of λ^2 to write

$$
\operatorname{tr}\big[\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\Sigma(\widehat{\Sigma} + \lambda I_p)^+\big]/n - \frac{\operatorname{tr}\big[(XX^T/n + \lambda I_n)^{+2}\big]/n}{\big(\operatorname{tr}\big[(XX^T/n + \lambda I_n)^+\big]/n\big)^2} + 1 \xrightarrow{\text{a.s.}} 0,
$$

which in the limiting case by sending $\lambda \to 0$ provides the equivalence

$$
\operatorname{tr}[\widehat{\Sigma}^+\Sigma]/n - \frac{\operatorname{tr}[\widehat{\Sigma}^{+2}]/n}{\big(\operatorname{tr}[\widehat{\Sigma}^+]/n\big)^2} + 1 \xrightarrow{\text{a.s.}} 0
$$

under proportional asymptotic limit. Note that we have written the final expression in terms $\widehat{\Sigma}$ instead of XX^T/n simply for consistency with the $\lambda \neq 0$ case. Combining the two cases, we have the desired limiting equivalences in Lemma 5.4.

S.1.5 Completing the proof of Theorem 4.1

Lemmas 5.1 to 5.4 establish the almost sure pointwise convergence of $\mathrm{gcv}(\lambda)$ to $\mathrm{err}(\lambda)$ under proportional asymptotics for $\lambda \in (\lambda_{\min}, \infty)$. To complete the proof of Theorem 4.1, we now show that the convergence holds uniformly over compact subintervals of $(\lambda_{\min}, \infty)$ and subsequently show the convergence of tuned risks over such intervals.

The strategy is show that, on any compact subinterval $I \subseteq (\lambda_{\min}, \infty)$, $\mathrm{gcv}(\lambda)$ and $\mathrm{err}(\lambda)$, and their derivatives, as functions of λ are bounded over I. This provides equicontinuity of family as functions of λ over I. The Arzela-Ascoli theorem then provides the desired uniform convergence. The convergence of tuned risks subsequently follows from a standard argument.

We start by writing the GCV estimate (S.4) for the ridge estimator as

$$\mathrm{gcv}(\lambda) = \frac{y^T(I_n - L_\lambda)^2 y/n}{\big(\mathrm{tr}[I_n - L_\lambda]/n\big)^2}.$$

It is convenient to first assume $\lambda \neq 0$ and express $I_n - L_\lambda$ as $\lambda(XX^T/n + \lambda I_n)^+$ using Lemma S.3.2 and then cancel the factor of λ^2 from both the numerator and denominator, which also covers the limiting $\lambda \to 0$ case. This lets us write the GCV estimate as

$$\mathrm{gcv}(\lambda) = \frac{u_n(\lambda)}{v_n(\lambda)}, \tag{S.12}$$

where $u_n(\lambda) = y^T(XX^T/n + \lambda I_n)^{+2} y/n$, and the denominator $v_n(\lambda) = \Big(\mathrm{tr}\left[(XX^T/n + \lambda I_n)^+\right]/n\Big)^2$. We first bound the numerator and denominator appropriately. Let $s_{\min}$ and $s_{\max}$ denote the minimum non-zero and maximum eigenvalues of XX^T/n, respectively. We can upper bound the numerator as

$$|u_n(\lambda)| \leq \frac{\|y\|^2}{n} \frac{1}{(s_{\min} + \lambda)^2}, \tag{S.13}$$

and we can lower bound the denominator as

$$|v_n(\lambda)| \geq \frac{1}{(s_{\max} + \lambda)^2}. \tag{S.14}$$

Using the two bounds in (S.13) and (S.14) into (S.12), we have the following upper bound on the GCV estimate:

$$|\mathrm{gcv}(\lambda)| \leq \frac{\|y\|^2}{n} \left(\frac{s_{\max} + \lambda}{s_{\min} + \lambda}\right)^2.$$

From the strong law of large numbers we note that $\|y\|^2/n$ is almost surely upper bounded for sufficiently large n. From Bai and Silverstein (1998), we have that $s_{\max} \leq C(1 + \sqrt{\gamma})^2 r_{\max}$ for any $C > 1$ and $s_{\min} \geq c(1 - \sqrt{\gamma})^2 r_{\min}$ for any $c < 1$ almost surely for sufficiently large n, where $r_{\min}$ and $r_{\max}$ denote the bounds on the minimum and maximum eigenvalues of Σ from Assumption 3. Thus, over any compact subinterval I of $(\lambda_{\min}, \infty)$, $\mathrm{gcv}(\lambda)$ is bounded almost surely for sufficiently large n.

We next bound the derivative of $\mathrm{gcv}(\lambda)$ as a function of λ. We start with the quotient rule of the derivatives to write:

$$\mathrm{gcv}'(\lambda) = \frac{u_n'(\lambda)v_n(\lambda) - u_n(\lambda)v_n'(\lambda)}{v_n(\lambda)^2}. \tag{S.15}$$

We now upper bound the derivatives of $u_n(\lambda)$ and $v_n(\lambda)$, and additionally obtain an upper bound on $v_n(\lambda)$. From short calculations, we can upper bound the derivative of the numerator as

$$|u_n'(\lambda)| \leq \frac{2\|y\|^2}{n} \left|\frac{1}{(s_{\min} + \lambda)^3}\right|, \tag{S.16}$$

and the derivative of the denominator as

$$|v_n'(\lambda)| \leq \left|\frac{2}{(s_{\min} + \lambda)^3}\right|. \tag{S.17}$$

In addition, we can upper bound the denominator as

$$|v_n(\lambda)| \leq \frac{1}{(s_{\min}+\lambda)^2}. \tag{S.18}$$

Combining the bounds in (S.16) to (S.18), along with the bounds in (S.13) and (S.14), into (S.15), we get the following upper bound on the derivative:

$$|\mathrm{gcv}'(\lambda)| \leq \frac{4\|y\|^2}{n}\left|\frac{(s_{\max}+\lambda)^4}{(s_{\min}+\lambda)^5}\right|. \tag{S.19}$$

As before, we note that $\|y\|^2/n$ is almost surely upper bounded for sufficiently large n, and $s_{\max}$ is upper bounded and $s_{\min}$ lower bounded above $(\sqrt{\gamma}-1)^2 r_{\min}$ for sufficiently large n. Thus, over any compact subinterval I of $(\lambda_{\min},\infty)$, $|\mathrm{gcv}'(\lambda)|$ is almost surely upper bounded for sufficiently large n.

By similar arguments, we can bound the $\mathrm{err}(\lambda)$ and its derivative as a function of λ. Together, we have that the function $\mathrm{err}(\lambda) - \mathrm{gcv}(\lambda)$ forms an equicontinous family of functions of λ over any compact subinterval of $(\lambda_{\min},\infty)$. Applying the Arzela-Ascoli theorem, we conclude uniform convergence for a subsequence, and since the difference converges pointwise to 0, the uniform convergence holds for the entire sequence.

Finally, we use the uniform convergence to establish the convergence of the tuned risks by a standard argument. We start with the observation that $\mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) \leq \mathrm{gcv}(\lambda)$ for any $\lambda \in I$ using the optimality of $\widehat{\lambda}_I^{\mathrm{gcv}}$. Using the specific $\lambda = \lambda_I^\star$, we thus have that $\mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) \leq \mathrm{gcv}(\lambda_I^\star)$. We next note that

$$\begin{aligned}
\mathrm{err}(\widehat{\lambda}_I^{\mathrm{gcv}}) - \mathrm{err}(\lambda_I^\star) &= \mathrm{err}(\widehat{\lambda}_I^{\mathrm{gcv}}) - \mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) + \mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) - \mathrm{gcv}(\lambda_I^\star) + \mathrm{gcv}(\lambda_I^\star) - \mathrm{err}(\lambda_I^\star) \\
&\leq \mathrm{err}(\widehat{\lambda}_I^{\mathrm{gcv}}) - \mathrm{gcv}(\widehat{\lambda}_I^{\mathrm{gcv}}) + \mathrm{gcv}(\lambda_I^\star) - \mathrm{err}(\lambda_I^\star) \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

where the inequality follows from the optimality of $\widehat{\lambda}_I^{\mathrm{gcv}}$ for $\mathrm{gcv}(\lambda)$ and the two almost sure convergences follow from the uniform convergence. This concludes the proof of Theorem 4.1.

S.1.6 Error terms in the proof of Lemma 5.3

It is convenient to further split $e_1 = e_{11} + e_{12}$ where the suberror terms e_{11} and e_{12} are defined as follows:

$$e_{11} := \frac{1}{n}\sum_{i=1}^n \left(\frac{x_i^T (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+) x_i}{\left(1 + x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+ x_i/n\right)^2} - \frac{\mathrm{tr}\left[(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+)\Sigma\right]}{\left(1 + x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+ x_i/n\right)^2} \right),$$

$$e_{12} := \frac{1}{n}\sum_{i=1}^n \left(\frac{\mathrm{tr}\left[(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+)\Sigma\right]}{\left(1 + x_i^T(\widehat{\Sigma}_{-i}+\lambda I_p)^+ x_i/n\right)^2} - \frac{\mathrm{tr}\left[(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+)\Sigma\right]}{\left(1 + \mathrm{tr}\left[(\widehat{\Sigma}_{-i}+\lambda I_p)^+\Sigma\right]/n\right)^2} \right).$$

We similarly split $e_2 = e_{21} + e_{22}$ where the suberror terms e_{21} and e_{22} are defined as follows:

$$e_{21} := \frac{1}{n}\sum_{i=1}^n \left(\frac{\mathrm{tr}\left[(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+)\Sigma\right]}{\left(1 + \mathrm{tr}\left[(\widehat{\Sigma}_{-i}+\lambda I_p)^+\Sigma\right]/n\right)^2} - \frac{\mathrm{tr}\left[(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+)\Sigma\right]}{\left(1 + \mathrm{tr}\left[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\right]/n\right)^2} \right)$$

$$e_{22} := \frac{1}{n}\sum_{i=1}^n \left(\frac{\mathrm{tr}\left[(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i}+\lambda I_p)^+)\Sigma\right]}{\left(1 + \mathrm{tr}\left[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\right]/n\right)^2} - \frac{\mathrm{tr}\left[(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+) B (I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+)\Sigma\right]}{\left(1 + \mathrm{tr}\left[(\widehat{\Sigma}+\lambda I_p)^+\Sigma\right]/n\right)^2} \right).$$

Below we show that for $\lambda \in (\lambda_{\min},\infty)$ all the suberror terms almost surely approach 0 as $n, p \to \infty$ with $p/n \to \gamma \in (0,\infty)$. Note that we use a generic letter C to denote a constant (that does not depend on n or p) whose value can change from line to line and the inequality sign is used in an asymptotic sense which holds almost surely for sufficiently large n.

Error term e_{11}

We bound the error term e_{11} as follows:

$$
\begin{aligned}
|e_{11}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \frac{x_i^T \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) x_i - \operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) \Sigma\Big]}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\big)^2} \right| \\
&\leq C \left| \frac{1}{n} \sum_{i=1}^{n} x_i^T \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) x_i - \operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) \Sigma\Big] \right| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the first inequality follows by noting that from Lemma S.4.2 the quadratic form $x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n$ converges almost surely to $\operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma]/n$ (as operator norm of $(\widehat{\Sigma}_{-i} + \lambda I_p)^+$ is almost surely bounded for large n) and the fact that $\left|1/\big(1 + \operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma]/n\big)\right|$ is bounded by viweing $\operatorname{tr}[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma]/n$ as a Stieljes transform of a measure with bounded total mass (see, for example, Paul and Silverstein (2009); Couillet and Hachem (2014)). The convergence in the final step follows from application of Lemma S.4.4 since $\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)$ has trace norm almost surely bounded for large n (as trace norm of B is bounded and the operator norm of $\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)$ is almost surely bounded for large n).

Error term e_{12}

We bound the error term e_{12} as follows:

$$
\begin{aligned}
|e_{12}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \operatorname{tr}\Big[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) \Sigma\Big] \left(\frac{1}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\big)^2} - \frac{1}{\big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n\big)^2} \right) \right| \\
&\leq C \left| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\big)^2} - \frac{1}{\big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n\big)^2} \right| \\
&= C \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n\big)^2 - \big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\big)^2}{\big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\big)^2 \big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n\big)^2} \right| \\
&\leq C \left| \frac{1}{n} \sum_{i=1}^{n} \big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n\big)^2 - \big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\big)^2 \right| \\
&\leq C \max_{i=1,\ldots,n} \left| \big(1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\big)^2 - \big(1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n\big)^2 \right| \\
&\leq C \max_{i=1,\ldots,n} \left| x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n - \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n \right| \left| 2 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n \right| \\
&\leq C \max_{i=1,\ldots,n} \left| x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n - \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n \right| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the first inequality bound follows from noting that the matrix $\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) \Sigma$ almost surely has bounded trace norm for large n (since trace norm of $\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^+\big)$ is bounded almost surely for large n as argued for the error term e_{11} above and the operator norm of Σ is bounded) and the final convergence follows from using Lemma S.4.3 by noting that the operator norm of $(\widehat{\Sigma}_{-i} + \lambda I_p)^+$ is almost surely bounded for large n.

Error term e_{21}

We bound the error term e_{21} as follows:

$$
\begin{aligned}
|e_{21}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \operatorname{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) \Sigma \right] \left(\frac{1}{\big(1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n\big)^2} - \frac{1}{\big(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n\big)^2} \right) \right| \\
&\leq C \left| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{\big(1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n\big)^2} - \frac{1}{\big(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n\big)^2} \right| \\
&= \frac{C}{n} \left| \sum_{i=1}^{n} \frac{\big(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n\big)^2 - \big(1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n\big)^2}{\big(1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n\big)^2 \big(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n\big)^2} \right| \\
&\leq \frac{C}{n} \sum_{i=1}^{n} \left| \Big(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n\Big)^2 - \Big(1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n\Big)^2 \right| \\
&\leq \frac{C}{n} \sum_{i=1}^{n} \left| \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n - \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n \right| \left| 2 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n \right| \\
&\leq \frac{C}{n} \sum_{i=1}^{n} \left| \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n - \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n \right| \\
&\leq \frac{C}{n} \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the final convergence follows by noting that

$$
(\widehat{\Sigma} + \lambda I_p)^{+} - (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} = -\frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T / n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i / n},
$$

which after multiplying by Σ, taking the trace, and normalizing by n gives

$$
\begin{aligned}
\left| \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n - \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]/n \right| &= \frac{1}{n} \left| \frac{\operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T / n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma\right]}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i / n} \right| \\
&= \frac{1}{n} \left| \frac{x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i / n}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i / n} \right| \\
&\leq \frac{C}{n},
\end{aligned}
$$

where the last bound follows by noting that operator norm of $(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\Sigma$ is almost surely bounded for large n.

Error term e_{22}

We bound the error term e_{22} as follows:

$$
\begin{aligned}
|e_{22}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \frac{\operatorname{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) \Sigma \right] - \operatorname{tr}\left[(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}) B (I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}) \Sigma \right]}{\Big(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+}\Sigma\right]/n\Big)^2} \right| \\
&\leq \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) \Sigma \right] - \operatorname{tr}\left[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \Sigma \right] \right| \\
&\leq \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) \Sigma \right] - \operatorname{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \Sigma \right] \right| \\
&\quad + \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\left[\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \Sigma \right] - \operatorname{tr}\left[\big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \Sigma \right] \right| \\
&\leq \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\left[\Sigma \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \Big\{ \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) - \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \Big\} \right] \right| \\
&\quad + \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\left[\Big\{ \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) - \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \Big\} B \big(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+}\big) \Sigma \right] \right| \\
&\leq \frac{C}{n} \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}
$$

where the last inequality bound follows by noting that

$$
\begin{aligned}
&\widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+} - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \\
&= (\widehat{\Sigma}_{-i} + x_i x_i^T/n)(\widehat{\Sigma}_{-i} + x_i x_i^T/n + \lambda I_p)^{+} - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \\
&= (\widehat{\Sigma}_{-i} + x_i x_i^T/n) \left((\widehat{\Sigma}_{-i} + \lambda I_p)^{+} - \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + \frac{1}{n} x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i} \right) - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \\
&= \frac{x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} - \frac{\widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \\
&= \frac{\big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+}}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n},
\end{aligned}
$$

which after multiplying by $\Sigma(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+})B$ and taking the trace can be bounded as follows:

$$
\begin{aligned}
&\left| \operatorname{tr}\left[\Sigma \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \Big\{ \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^{+} - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \Big\} \right] \right| \\
&= \left| \frac{\operatorname{tr}\left[\Sigma \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \right]}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \right| \\
&= \frac{1}{n} \left| \frac{x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \Sigma \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) B \big(I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\big) x_i}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^{+} x_i/n} \right| \\
&\leq \frac{C}{n},
\end{aligned}
$$

where the last bound follows by noting that the matrix $(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \Sigma (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}) B (I_p - \widehat{\Sigma}_{-i}(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}) \Sigma$ has almost surely bounded trace norm for large n (since trace norm of B is bounded and the operator norm of the remaining matrix component is almost surely bounded for large n). The second term can be bounded analogously.

S.2 Proofs related to Theorem 4.2

S.2.1 Proof of Lemma 5.6

We start by writing the leave-one-out risk estimate $\mathrm{loo}(\lambda)$ from Equation (4) as

$$\mathrm{loo}(\lambda) = y^T(I_n - L_\lambda)^2 D_\lambda^{-2} y/n,$$

where L_λ is the ridge smoothing matrix and $D_\lambda \in \mathbb{R}^{n\times n}$ is a diagonal matrix with entries $1-[L_\lambda]_{ii}$ for $i = 1, \ldots, n$. Under proportional asymptotic limit, we show below that for any $\lambda \in (\lambda_{\min}, \infty)$,

$$\mathrm{loo}(\lambda) - y^T(I_n - L_\lambda)^2\Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\right]/n\Big)^2 y/n \xrightarrow{\text{a.s.}} 0, \tag{S.20}$$

which after substituting back for L_λ proves the desired convergence.

Observe that for any $i = 1, \ldots, n$,

$$\begin{aligned}[D_\lambda^{-1}]_{ii} = \frac{1}{1-[L_\lambda]_{ii}} &= \frac{1}{1 - \left[X(X^TX/n + \lambda I_p)^+X^T/n\right]_{ii}} \\ &= \frac{1}{1 - x_i^T/\sqrt{n}(X^TX/n + \lambda I_p)^+ x_i/\sqrt{n}}.\end{aligned}$$

Denoting X^TX/n by $\widehat{\Sigma}$ and using the Woodbury matrix identity as explained in the proof of Lemma S.3.1, we have that

$$\frac{1}{1 - x_i^T(\widehat{\Sigma} + \lambda I_p)^+ x_i/n} = 1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n.$$

The diagonal entries of the matrix D_λ^{-1} are thus $1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n$ for $i = 1, \ldots, n$.

We proceed to bound the difference in the two quantities of (S.20) as follows:

$$\begin{aligned}&\left|\mathrm{loo}(\lambda) - y^T(I_n - L_\lambda)^2\Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\right]/n\Big)^2 y/n\right| \\ &= \left|y^T(I_n - L_\lambda)^2 D_\lambda^{-2} y/n - y^T(I_n - L_\lambda)^2\Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\right]/n\Big)^2 y/n\right| \\ &\le y^T(I_n - L_\lambda)^2 y/n \max_{i=1,\ldots,n}\left|\Big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\Big)^2 - \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\right]/n\Big)^2\right| \\ &\le C \max_{i=1,\ldots,n}\left|\Big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\Big)^2 - \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\right]/n\Big)^2\right|,\end{aligned}$$

where the bound in the last inequality holds almost surely for sufficiently large n by noting that $y^T(I_n - L_\lambda)^2 y/n$ is almost surely bounded for sufficiently large n as explained in the proof of Theorem 4.1 Note that we do not require that the response y is well-specified. Finally, similar to the proof of Lemma 5.3, we decompose the error as

$$\max_{i=1,\ldots,n}\left|\Big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\Big)^2 - \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\right]/n\Big)^2\right| \le \xi_1 + \xi_2,$$

where the error terms ξ_1 and ξ_2 are defined as follows:

$$\xi_1 := \max_{i=1,\ldots,n}\left|\Big(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\Big)^2 - \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\right]/n\Big)^2\right|, \tag{S.21}$$

$$\xi_2 := \max_{i=1,\ldots,n}\left|\Big(1 + \mathrm{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\right]/n\Big)^2 - \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+\Sigma\right]/n\Big)^2\right|. \tag{S.22}$$

Both of the error terms approach 0 under proportional asymptotic limit using the final parts of the arguments used for e_{12} and e_{21} in the proof of Lemma 5.3.

S.2.2 Completing the proof of Theorem 4.2

Case when $\lambda \neq 0$. Recall from Equation (S.4) that the GCV risk estimate $\mathrm{gcv}(\lambda)$ in this case can be expressed as

$$\mathrm{gcv}(\lambda) = \frac{y^T(I_n - L_\lambda)^2 y/n}{\Big(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\Big)^2}.$$

On the other hand, from Lemma 5.6, under proportional asymptotics we have that

$$\mathrm{loo}(\lambda) - y^T(I_n - L_\lambda)^2 \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big)^2 y/n \xrightarrow{\text{a.s.}} 0.$$

The result then follows by noting that

$$\begin{aligned}
&\left| y^T(I_n - L_\lambda)^2 \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big)^2 y/n - \mathrm{gcv}(\lambda) \right| \\
&= \left| y^T(I_n - L_\lambda)^2 \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big)^2 y/n - \frac{y^T(I_n - L_\lambda)^2 y/n}{\Big(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\Big)^2} \right| \\
&\leq y^T(I_n - L_\lambda)^2 y/n \left| \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big)^2 - \frac{1}{\Big(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\Big)^2} \right| \\
&\leq C \left| \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big)^2 - \frac{1}{\Big(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\Big)^2} \right| \\
&\leq C \left| \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big) - \frac{1}{\Big(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\Big)} \right| \left| \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big) + \frac{1}{\Big(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\Big)} \right| \\
&\leq C \left| \Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big) - \frac{1}{\Big(1 - \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n\Big)} \right| \\
&\xrightarrow{\text{a.s.}} 0
\end{aligned}$$

under proportional asymptotics using the first part of Lemma S.3.1. Note that the bound in the second inequality again follows from the fact that $\|y\|^2/n$ is almost surely upper bounded for sufficiently large n, and the operator norm of $I_n - L_\lambda$ is bounded almost surely for large n for $\lambda \in (\lambda_{\min}, \infty)$.

Limiting case when $\lambda = 0$ Similar to the proofs of Lemma 5.3 and Lemma 5.4, to handle the case when $\lambda = 0$, we observe that for $\lambda \neq 0$, we can extract a factor of λ^2 from $(I_n - L_\lambda)^2$ and absorb into $\Big(1 + \mathrm{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right]/n\Big)^2$ and take $\lambda \to 0$ to write the limiting LOOCV risk estimate under proportional asymptotics as

$$\mathrm{loo}(0) - y^T(XX^T/n)^{+2} \Big(\mathrm{tr}\left[(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma\right]/n\Big)^2 y/n \xrightarrow{\text{a.s.}} 0,$$

while the limiting GCV estimate is given by

$$\mathrm{gcv}(0) = \frac{y^T(XX^T/n)^{+2} y/n}{\big(\mathrm{tr}[\widehat{\Sigma}^+]/n\big)^2}.$$

As above, we can then bound the difference to get

$$\begin{aligned}
&\left| y^T (XX^T/n)^{+2} \Big(\operatorname{tr}\Big[(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+) \Sigma \Big] / n \Big)^2 y/n - \frac{y^T (XX^T/n)^{+2} y/n}{\big(\operatorname{tr}[\widehat{\Sigma}^+]/n\big)^2} \right| \\
&\leq C \left| \operatorname{tr}\Big[(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+) \Sigma \Big] / n - \frac{1}{\operatorname{tr}[\widehat{\Sigma}^+]/n} \right| \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

where the convergence follows from the second part of Lemma S.3.1.

Putting things together, this establishes the almost sure pointwise convergence of $\mathrm{loo}(\lambda)$ to $\mathrm{gcv}(\lambda)$. To show uniform convergence and the convergence of tuned risks, we similarly bound the estimate $\mathrm{loo}(\lambda)$ and its derivative as a function of λ to establish equicontinuity as done in the proof of Theorem 4.1. We omit the details due to similarity.

S.3 Auxiliary lemmas

In this section, we state and prove auxiliary lemmas that we often make use of in other proofs. Note that Lemma 5.5 in the main paper is a special case of Lemma 5.3 and its proof follows analogous steps as the proof of Lemma 5.3 in Section S.1.3 and is omitted.

Lemma S.3.1 (Basic GCV denominator lemma)**.** *Under Assumption 2 and Assumption 3, for* $\lambda \in (\lambda_{\min}, \infty) \setminus \{0\}$*,*

$$1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n - \frac{1}{1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n} \xrightarrow{a.s.} 0 \tag{S.23}$$

as $n, p \to \infty$ *with* $p/n \to \gamma \in (0, \infty)$*. In the case when* $\lambda = 0$*,*

$$\operatorname{tr}\big[(I_p - \widehat{\Sigma}^+ \widehat{\Sigma}) \Sigma\big]/n - \frac{1}{\operatorname{tr}\big[\widehat{\Sigma}^+\big]/n} \xrightarrow{a.s.} 0 \tag{S.24}$$

as $n, p \to \infty$ *with* $p/n \to \gamma \in (0, \infty)$*.*

Proof. We start with the the GCV denominator (the denominator of the second term of (S.23)) and establish that under proportional asymptotics

$$1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n - \frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n} \xrightarrow{\text{a.s.}} 0.$$

To that end, we use the standard leave-one-out trick to break the trace functional $1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n$ into random quadratic forms where the point of evaluation is independent of the inner matrix as follows:

$$\begin{aligned}
1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n &= 1 - \frac{1}{n} \operatorname{tr}\Big[(\widehat{\Sigma} + \lambda I_p)^+ \sum_{i=1}^n x_i x_i^T / n\Big] \\
&= 1 - \frac{1}{n} \sum_{i=1}^n \operatorname{tr}\Big[(\widehat{\Sigma} + \lambda I_p)^+ x_i x_i^T / n\Big] \\
&= 1 - \frac{1}{n} \sum_{i=1}^n x_i^T (\widehat{\Sigma} + \lambda I_p)^+ x_i / n \\
&= \frac{1}{n} \sum_{i=1}^n \big(1 - x_i^T (\widehat{\Sigma} + \lambda I_p)^+ x_i / n\big) \\
&= \frac{1}{n} \sum_{i=1}^n \frac{1}{1 + x_i^T (\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i / n}.
\end{aligned}$$

Here the last equality follows from the following simplification using the Sherman-Morrison-Woodbury formula with Moore-Penrose inverse (Meyer, 1973):

$$
\begin{aligned}
&1 - x_i^T(\widehat{\Sigma} + \lambda I_p)^+ x_i/n \\
&= 1 - x_i^T\big(\widehat{\Sigma}_{-i} + \lambda I_p + x_i x_i^T/n\big)^+ x_i/n \\
&= 1 - x_i^T \left((\widehat{\Sigma}_{-i} + \lambda I_p)^+ - \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^+}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \right) x_i/n \\
&= 1 - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n + x_i^T \frac{(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^+}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} x_i/n \\
&= 1 - \frac{x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n x_i^T(\widehat{\Sigma} + \lambda I_p)^+ x_i/n + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i x_i^T/n (\widehat{\Sigma}_{-i} + \lambda I_p)^+}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \\
&= 1 - \frac{x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \\
&= \frac{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} \\
&= \frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n}.
\end{aligned}
$$

We now break the error in (S.23) as

$$
\begin{aligned}
1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n - \frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n} &= \frac{1}{n}\sum_{i=1}^n \frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} - \frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n} \\
&= \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} - \frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n} \right) \\
&= \delta_1 + \delta_2,
\end{aligned}
$$

where the error terms δ_1 and δ_2 are defined as follows:

$$
\begin{aligned}
\delta_1 &:= \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} - \frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n} \right), \\
\delta_2 &:= \frac{1}{n}\sum_{i=1}^n \left(\frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma}_{-i} + \lambda I_p)^+ \Sigma\big]/n} - \frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n} \right),
\end{aligned}
$$

In Section S.3.1, we show that both the error terms δ_1 and δ_2 almost surely approach 0 under proportional asymptotics for $\lambda \in (\lambda_{\min}, \infty)$ under Assumption 2 and Assumption 3.

We now finish the final step by considering the two cases of $\lambda \neq 0$ and $\lambda = 0$.

Case when $\lambda \neq 0$. We so far have that

$$
1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n - \frac{1}{1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n} \xrightarrow{\text{a.s.}} 0,
$$

which we can rewrite as

$$
\Big(1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n\Big)\Big(1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n\Big) - 1 \xrightarrow{\text{a.s.}} 0.
$$

When $\lambda \neq 0$, the GCV denominator $1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n \neq 0$, and we can safely take the inverse to get

$$
1 + \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\big]/n - \frac{1}{1 - \operatorname{tr}\big[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\big]/n} \xrightarrow{\text{a.s.}} 0
$$

under proportional asymptotic limit as desired.

Limiting case when $\lambda = 0$. In this case, $1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n$ can be zero (in particular, it is zero when $p \geq n$ and X has rank n). As before, we start with $\lambda \neq 0$ and using Lemma S.3.2, express

$$1 - \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \widehat{\Sigma}\right]/n = \lambda \operatorname{tr}\left[(XX^T/n + \lambda I_n)^+\right]/n,$$

along with

$$\lambda \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^+ \Sigma\right] = \operatorname{tr}\left[\left(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\right)\Sigma\right]/n.$$

This allows us to move λ across to write

$$\left(\operatorname{tr}\left[(XX^T/n + \lambda I_n)^+\right]/n\right)\left(\lambda + \operatorname{tr}\left[\left(I_p - \widehat{\Sigma}(\widehat{\Sigma} + \lambda I_p)^+\right)\Sigma\right]/n\right) - 1 \xrightarrow{\text{a.s.}} 0.$$

Sending $\lambda \to 0$, writing $\operatorname{tr}[(XX^T/n^+)]/n = \operatorname{tr}[\widehat{\Sigma}^+]/n$, and inverting safely, we have

$$\operatorname{tr}\left[(I_p - \widehat{\Sigma}\widehat{\Sigma}^+)\Sigma\right]/n - \frac{1}{\operatorname{tr}[\widehat{\Sigma}^+]/n} \xrightarrow{\text{a.s.}} 0$$

under proportional asymptotic limit as desired. □

Lemma S.3.2 (Gram and sample covariance matrix simplifications)**.** *Suppose $X^TX/n + \lambda I_p$ and $XX^T/n + \lambda I_n$ are invertible. Then it holds that*

$$I_n - X(X^TX/n + \lambda I_p)^+ X^T/n = \lambda(XX^T/n + \lambda I_n)^+,$$
$$I_p - \left(X^TX/n + \lambda I_p\right)^+ X^TX/n = \lambda(X^TX/n + \lambda I_p)^+.$$

Proof. Recall the Woodbury matrix identity

$$A^{-1} - A^{-1}U(VA^{-1}U + C^{-1})^{-1}VA^{-1} = (UCV + A)^{-1}.$$

Letting $A = I_n$, $U = X/\sqrt{n}$, $C = 1/\lambda I_p$, $V = X^T/\sqrt{n}$, we get

$$\begin{aligned} I_n - X(X^TX/n + \lambda I_p)^{-1}X^T/n &= (X/\sqrt{n}\, 1/\lambda I_p\, X^T/\sqrt{n} + I_n)^{-1} \\ &= \lambda(XX^T/n + \lambda I_n)^{-1}. \end{aligned}$$

On the other hand, letting $A = I_p$, $U = I_p$, $V = X^TX/n$, $C = 1/\lambda I_p$, we get

$$\begin{aligned} I_p - \left(X^TX/n + \lambda I_p\right)^{-1} X^TX/n &= \left(1/\lambda I_p\, X^TX/n + I_p\right)^{-1} \\ &= \lambda\left(X^TX/n + \lambda I_p\right)^{-1}. \end{aligned}$$

□

S.3.1 Error terms in the proof of Lemma S.3.1

Below we show that for $\lambda \in (\lambda_{\min}, \infty)$ both the error terms δ_1 and δ_2 almost surely approach 0 as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. The arguments mirror parts of the error analysis for terms e_{12} and e_{21} in Section S.1.6.

Error term δ_1

$$\begin{aligned} |\delta_1| &= \left|\frac{1}{n}\sum_{i=1}^n \frac{1}{1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n} - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\right]/n}\right| \\ &= \left|\frac{1}{n}\sum_{i=1}^n \frac{\operatorname{tr}\left[\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\right]/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n}{\left(1 + x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\right)\left(1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\right]/n\right)}\right| \\ &\leq C\left|\frac{1}{n}\sum_{i=1}^n \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\right]/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\right| \\ &\leq C \max_{i=1,\ldots,n}\left|\operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^+\Sigma\right]/n - x_i^T(\widehat{\Sigma}_{-i} + \lambda I_p)^+ x_i/n\right| \\ &\xrightarrow{\text{a.s.}} 0, \end{aligned}$$

where the final convergence follows from using Lemma S.4.4 as argued for the suberror term e_{12} in Section S.1.6.

Error term δ_2

$$\begin{aligned}
|\delta_2| &= \left| \frac{1}{n} \sum_{i=1}^{n} \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \Sigma\right]/n} - \frac{1}{1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+} \Sigma\right]/n} \right| \\
&= \frac{1}{n} \left| \sum_{i=1}^{n} \frac{\operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+} \Sigma\right]/n - \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \Sigma\right]/n}{\left(1 + \operatorname{tr}\left[(\widehat{\Sigma}_{-i} + \lambda I_p)^{+} \Sigma\right]/n\right)\left(1 + \operatorname{tr}\left[(\widehat{\Sigma} + \lambda I_p)^{+} \Sigma\right]/n\right)} \right| \\
&\leq \frac{C}{n} \left| \sum_{i=1}^{n} \operatorname{tr}\left[\Sigma(\widehat{\Sigma} + \lambda I_p)^{+}\right]/n - \operatorname{tr}\left[\Sigma(\widehat{\Sigma}_{-i} + \lambda I_p)^{+}\right]/n \right| \\
&\leq \frac{C}{n} \\
&\xrightarrow{\text{a.s.}} 0,
\end{aligned}$$

where the last inequality follows analogous simplification as done for the suberror term e_{21} in Section S.1.6.

S.4 Useful concentration results

The following lemma is a standard concentration of linear combination of i.i.d. entries.

Lemma S.4.1 (Concentration of linear form with independent components)**.** *Let ε be a random vector in R^n that satisfy conditions of error vector in Assumption 1. Let b_n be a sequence of random vectors in $\mathbb{R}^n$ independent of ε such that $\sup_n \|b_n\|^2/n < \infty$ almost surely. Then as $n \to \infty$,*

$$b_n^T \varepsilon / n \xrightarrow{a.s.} 0.$$

The following lemma is adapted from Dobriban and Wager (2018, Lemma 7.6).

Lemma S.4.2 (Concentration of quadratic form with independent components)**.** *Let $\varepsilon \in \mathbb{R}^n$ be a random vector that satisfy conditions of error vector in Assumption 1. Let D_n be a sequence of random matrices in $\mathbb{R}^{n \times n}$ that are independent of ε and have operator norm uniformly bounded in n. Then as $n \to \infty$,*

$$\varepsilon^T D_n \varepsilon / n - \sigma^2 \operatorname{tr}[D_n]/n \xrightarrow{a.s.} 0.$$

The following lemma is adapted from an argument in Hastie et al. (2019, Theorem 7) using union bound along with a lemma from Bai and Silverstein (2010, Lemma B.26).

Lemma S.4.3 (Concentration of maximum of quadratic forms with independent components)**.** *Let $x_1, \ldots, x_n$ be random vectors in $\mathbb{R}^p$ that satisfy Assumption 2 and Assumption 3. Let $G_1, \ldots, G_n$ be random matrices in $\mathbb{R}^{p \times p}$ such that G_i is independent of x_i (but may depend on all of X_{-i}) and have operator norm uniformly bounded in n. Then as $n \to \infty$,*

$$\max_{i=1,\ldots,n} \left| x_i^T G_i x_i / n - \operatorname{tr}[G_i \Sigma]/n \right| \xrightarrow{a.s.} 0.$$

The following lemma is adapted from Rubio and Mestre (2011, Lemma 4).

Lemma S.4.4 (Concentration of sum of quadratic forms with independent components)**.** *Let $x_1, \ldots, x_n$ be random vectors in $\mathbb{R}^p$ that satisfy Assumption 2 and Assumption 3. Let $H_1, \ldots, H_n$ be random matrices in $\mathbb{R}^{p \times p}$ such that H_i is independent of x_i (but may depend on all of X_{-i}) that have trace norm uniformly bounded in n. Then as $n \to \infty$,*

$$\left| \sum_{i=1}^{n} x_i^T H_i x_i / n - \operatorname{tr}[H_i \Sigma]/n \right| \xrightarrow{a.s.} 0.$$

References

Zhi-Dong Bai and Jack W. Silverstein. No eigenvalues outside the support of the limiting spectral distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998.

Zhidong Bai and Jack W. Silverstein. *Spectral Analysis of Large Dimensional Random Matrices*. Springer, 2010.

Romain Couillet and Walid Hachem. Analysis of the limiting spectral measure of large random matrices of the separable covariance type. *Random Matrices: Theory and Applications*, 3(04):1450016, 2014.

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

Trevor Hastie, Andrea Montanari, Saharon Rosset, and Ryan J. Tibshirani. Surprises in high-dimensional ridgeless least squares interpolation. *arXiv preprint arXiv:1903.08560*, 2019.

Olivier Ledoit and Sandrine Peche. Eigenvectors of some large sample covariance matrices ensembles. *SSRN Electronic Journal*, pages 233–264, 03 2009.

Carl D. Meyer, Jr. Generalized inversion of modified matrices. *SIAM Journal on Applied Mathematics*, 24(3): 315–323, 1973.

Debashis Paul and Jack W. Silverstein. No eigenvalues outside the support of the limiting empirical spectral distribution of a separable covariance matrix. *Journal of Multivariate Analysis*, 100(1):37–57, 2009.

Francisco Rubio and Xavier Mestre. Spectral convergence for a general class of random matrices. *Statistics & probability letters*, 81(5):592–602, 2011.