Supplementary Material for "Estimating Functionals of the Out-of-Sample Error Distribution in High-Dimensional Ridge Regression"

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

Carnegie Mellon University

This supplement contains additional details, proofs, and numerical experiments for the paper "Estimating Functionals of the Out-of-Sample Error Distribution in High-Dimensional Ridge Regression." All section, equation, and figure numbers in this document begin with the letter "S" to differentiate them from those appearing in the main paper that do not have such prefix.

The content of the supplement is organized as follows. In Sections S.1 to S.3, we first provide proofs related to Theorems 3 to 5, respectively, along with supporting lemmas used in the process, as they constitute building blocks for other theoretical results. Then Section S.4 contains proof of Theorem 1, while Section S.5 contains proofs related to Theorem 6, along with further theoretical results related to quantile estimation. Additional numerical results and experimental details are provided in Section S.6. Finally, Section S.7 collects statements of supplementary results from the literature that are used in various proofs throughout the supplement.

A note about constants throughout the supplement: we use the letter C (either standalone or with a subscript such as C_1) to denote a generic constant whose value can change from line to line. Additionally, some of the inequalities only hold almost surely for sufficiently large n. We will sometimes use the term eventually almost surely to indicate such statements.

Contents

S.1 PROOFS RELATED TO Theorem 3

As suggested in the proof overview in Section 4 of the paper, we will first show the second part of the theorem statement: $\widehat{T}_\lambda^{\text{loo}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, and use it to show the first part: $\widehat{T}_\lambda^{\text{gcv}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

- To prove $\widehat{T}_\lambda^{\text{loo}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, we introduce an intermediate quantity $\widetilde{T}_\lambda$ as in (19) and break the difference
$$T_\lambda - \widehat{T}_\lambda^{\text{loo}} = (T_\lambda - \widetilde{T}_\lambda) + (\widetilde{T}_\lambda - \widehat{T}_\lambda^{\text{loo}}). \tag{S.1}$$
We will show that both terms in the decomposition (S.1) almost surely vanish. Section S.1.1 shows the convergence for the first term, while Section S.1.2 shows the convergence for the second term.
- To prove $\widehat{T}_\lambda^{\text{gcv}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, we similarly break the difference
$$T_\lambda - \widehat{T}_\lambda^{\text{gcv}} = (T_\lambda - \widehat{T}_\lambda^{\text{loo}}) + (\widehat{T}_\lambda^{\text{loo}} - \widehat{T}_\lambda^{\text{gcv}}). \tag{S.2}$$
We have already dealt with the first term in the decomposition (S.2) in (S.1). We show the second term almost surely goes to zero in Section S.1.3.

We will show the three aforementioned converges first under a slight stronger assumption that the error function t is uniformly continuous. Using a truncation argument, we will then relax them to continuous error functions t in Section S.1.4. Let $\omega_t : [0, \infty] \to [0, \infty]$ denote a modulus of continuity of t. Without of loss of generality, we can assume ω_t to be non-decreasing and continuous. Since the error function is assumed to be uniformly continuous, such a modulus exits (see, e.g., Chapter 2 of DeVore and Lorentz, 1993). In addition, let $\overline{\omega}_t$ denote the least concave majorant of ω_t. From DeVore and Lorentz (1993, Lemma 6.1), $\overline{\omega}_t$ is also a modulus of continuity and satisfies $\overline{\omega}_t(r) \le 2\omega_t(r)$ for $r \ge 0$. We will make use of these properties below.

S.1.1 Functional to LOO Functional

Towards showing $T_\lambda - \widetilde{T}_\lambda \xrightarrow{\text{a.s.}} 0$, we begin by manipulating the desired difference using properties of conditional expectation as follows:

$$\begin{aligned}
T_\lambda - \widetilde{T}_\lambda &= \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big] \\
&= \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big] \\
&= \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}, x_i, y_i\big] \\
&= \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big] \\
&= \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big].
\end{aligned}$$

The second equality above uses independence of (y_0, x_0) and (X_{-i}, y_{-i}), while the third equality uses independence of (y_0, x_0), $\widehat{\beta}_{-i,\lambda}$, and (x_i, y_i). We will next show below that under proportional asymptotics absolute value of the right-hand side of the last display almost surely goes to zero; in other words, we will show

$$\left| \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big] \right| \xrightarrow{\text{a.s.}} 0. \tag{S.3}$$

Using the modulus of continuity of t and its least concave majorant, we first bound the summands in (S.3) for $i = 1, \ldots, n$ as

$$\begin{aligned}
\big|t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda})\big| &\le \omega_t\big(\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big|\big) \\
&\le \overline{\omega}_t\big(\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big|\big).
\end{aligned}$$

We can then bound the summation in (S.3) as

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\big[t(y_0 - x_0^\top\widehat{\beta}_\lambda) - t(y_0 - x_0^\top\widehat{\beta}_{-i,\lambda}) \mid X, y\big]\right| &\leq \frac{1}{n}\sum_{i=1}^{n}\Big|\mathbb{E}\big[t(y_0 - x_0^\top\widehat{\beta}_\lambda) - t(y_0 - x_0^\top\widehat{\beta}_{-i,\lambda}) \mid X, y\big]\Big| \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big|t(y_0 - x_0^\top\widehat{\beta}_\lambda) - t(y_0 - x_0^\top\widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\overline{\omega}_t\big(\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big|\big) \mid X, y\Big] \\
&\leq \frac{1}{n}\sum_{i=1}^{n}\overline{\omega}_t\Big(\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big]\Big) \\
&\leq \overline{\omega}_t\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big]\right) \\
&\leq 2\omega_t\left(\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big]\right).
\end{aligned}$$

In the above chain of inequalities, the second, forth, and fifth inequalities follow from repeated use of Jensen's inequality (on the absolute value function and the concave majorant function). To finish the proof, we will finally show below that

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big] \xrightarrow{\text{a.s.}} 0, \tag{S.4}$$

which along with the continuity of the modulus that vanishes at 0 shows (S.3), leading to the desired conclusion that $T_\lambda - \widetilde{T}_\lambda \xrightarrow{\text{a.s.}} 0$.

Towards showing (S.4), first note that under Assumption 1, we can bound the summands for each $i = 1, \ldots, n$ as

$$\begin{aligned}
\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big] &\leq \Big(\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big|^2 \mid X, y\Big]\Big)^{1/2} \\
&= \Big(\mathbb{E}\Big[\big|z_0^\top\Sigma^{1/2}(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big|^2 \mid X, y\Big]\Big)^{1/2} \\
&= \Big(\mathbb{E}\Big[(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})^\top\Sigma^{1/2}z_0z_0^\top\Sigma^{1/2}(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}) \mid X, y\Big]\Big)^{1/2} \\
&= \Big((\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\Sigma(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\Big)^{1/2} \\
&\leq \Big(r_{\max}(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\Big)^{1/2} \\
&= \sqrt{r_{\max}}\big\|(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big\|_2.
\end{aligned}$$

The inequality in the first line uses Jensen's inequality (on the square root function), and the inequality in the forth line follows since the maximum eigenvalue of Σ is upper bounded by $r_{\max}$. Hence, overall we can bound the left-hand side of (S.4) by

$$\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big] \leq \sqrt{r_{\max}}\left(\frac{1}{n}\sum_{i=1}^{n}\big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2\right). \tag{S.5}$$

We show in Lemma S.2 that the term in the parenthesis on the right-hand side of (S.5) almost surely goes to zero under Assumptions 1 and 2, proving (S.4) and completing the proof.

S.1.2 LOO Functional to LOOCV Estimator

To show $\widetilde{T}_\lambda - \widehat{T}_\lambda^{\text{loo}} \xrightarrow{\text{a.s.}} 0$, we start by breaking the difference into two pieces:

$$\begin{aligned}\big|\widetilde{T}_\lambda - \widehat{T}_\lambda^{\text{loo}}\big| &= \left|\widetilde{T}_\lambda - \frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) + \frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widehat{T}_\lambda^{\text{loo}}\right| \\ &\le \left|\widetilde{T}_\lambda - \frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\right| + \left|\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widehat{T}_\lambda^{\text{loo}}\right|. \end{aligned} \tag{S.6}$$

In the sequel, we will show that each of two pieces in (S.6) vanishes almost surely under proportional asymptotics. For the second piece in (S.6), using the modulus of t and its concave majorant, we can bound the difference as

$$\begin{aligned}\left|\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widetilde{T}_\lambda^{\text{loo}}\right| &= \left|\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \frac{1}{n}\sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right)\right| \\ &\le \frac{1}{n}\sum_{i=1}^n \left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right)\right| \\ &\le \frac{1}{n}\sum_{i=1}^n \overline{\omega}_t\left(\left|y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right|\right) \\ &\le \overline{\omega}_t\left(\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right|\right) \\ &\le 2\omega\left(\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-[L_\lambda]_{ii}}\right|\right), \end{aligned} \tag{S.7}$$

where line four uses Jensen's inequality (on the concave majorant). Note that the above is valid when $1-[L_\lambda]_{ii} \neq 0$ for any of $i = 1, \ldots, n$. For the case of min-norm estimator where $[L_0]_{ii} = 0$, we similarly bound

$$\left|\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,0}) - \widetilde{T}_\lambda^{\text{loo}}\right| \le 2\omega\left(\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_{-i,0} - \frac{[(XX^\top/n)^\dagger]_i}{[(XX^\top/n)^\dagger]_{ii}}\right|\right). \tag{S.8}$$

The argument of ω in either cases of (S.7) and (S.8) goes to 0 almost surely, and thus the continuity of ω provides the desired convergence of the second piece in (S.6) It is worth mentioning that the only reason we need to worry about (S.7) and (S.8) is the way we have defined ridge estimator in (1) where the leave-one-out estimator $\widehat{\beta}_{-i,\lambda}$ gets a dividing factor of $(n-1)$ instead of n, otherwise these terms would be exactly 0. It is a short straightforward calculation to show however that this does not make a difference as $n \to \infty$.

We now focus on the first piece in the decomposition (S.6). Note that we can express

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widetilde{T}_\lambda &= \frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\big|\; X_{-i}, y_{-i}\big] \\ &= \frac{1}{n}\sum_{i=1}^n \Big\{t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\big|\; X_{-i}, y_{-i}\big]\Big\}. \end{aligned} \tag{S.9}$$

For $i = 1, \ldots, n$, let $\mathcal{F}_i$ denote the increasing σ-field generated by $(x_1, y_1), \ldots, (x_i, y_i)$. Observe that

$$\Big\{t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\big|\; X_{-i}, y_{-i}\big]\Big\}_{i=1}^n$$

forms a martingale difference array with respect to the filtration $\{\mathcal{F}_i\}_{i=1}^n$. To see this, note that

$$\begin{aligned}
&\mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big] \;\Big|\; \mathcal{F}_{i-1}\Big] \\
&= \mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\Big|\; \mathcal{F}_{i-1}\Big] - \mathbb{E}\Big[\mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big] \;\Big|\; \mathcal{F}_{i-1}\Big] \\
&= \mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\Big|\; \mathcal{F}_{i-1}\Big] - \mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\Big|\; \mathcal{F}_{i-1}\Big] \\
&= 0,
\end{aligned}$$

where for the second equality we used the tower property of conditional expectation as $\mathcal{F}_{i-1}$ is a subset of the σ-field generated by (X_{-i}, y_{-i}). This observation allows us to use the Burkholder inequality (see Lemma S.8 for an exact statement) to bound q-th moment of the difference for $q \geq 2$.

Applying the Burkholder inequality to our martingale sequence, we can bound

$$\begin{aligned}
&\mathbb{E}\left[\left|\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \\
&\leq C\mathbb{E}\left[\left\{\sum_{i=1}^n \mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^2 \;\middle|\; \mathcal{F}_{i-1}\right]\right\}^{q/2}\right] \\
&\quad + C\mathbb{E}\left[\sum_{i=1}^n \left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \qquad \text{(S.10)}
\end{aligned}$$

for some constant $C > 0$. We next bound each of the terms in turn. Denote by X_{i+i}^n and y_{i+i}^n dataset consisting of observations $(x_{i+1}, y_{i+1}), \cdots, (x_n, y_n)$.

For the first term, from the law of total expectation observe that

$$\begin{aligned}
&\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^2 \;\middle|\; \mathcal{F}_{i-1}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left\{\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^2 \;\middle|\; \mathcal{F}_{i-1}, X_{i+1}^n, y_{i+1}^n\right\}\right] \\
&= \mathbb{E}\left[\mathbb{E}\left\{\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^2 \;\middle|\; X_{-i}, y_{-i}\right\}\right] \\
&\leq 4\mathbb{E}\left[\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^2 \;\Big|\; X_{-i}, y_{-i}\Big]\right],
\end{aligned}$$

where in the last step we used the inequality $\mathbb{E}[|a+b|^2] \leq 2\big(\mathbb{E}[|a|^2] + \mathbb{E}[|b|^2]\big)$.

For the second term, similarly note that

$$\begin{aligned}
&\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \\
&\leq \mathbb{E}\left[\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \;\middle|\; X_{-i}, y_{-i}\right] \\
&\leq 2^q\mathbb{E}\left[\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\right|^q \;\middle|\; X_{-i}, y_{-i}\right]\right],
\end{aligned}$$

where the last step follows from using the inequality $\mathbb{E}[|a+b|^q] \leq 2^{q-1}\big(\mathbb{E}[|a|^q] + \mathbb{E}[|b|^q]\big)$ for $q > 1$.

In addition, from Jensen's inequality, we have for $q \geq 2$

$$\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^2 \mid X_{-i}, y_{-i}\Big] \leq \mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \mid X_{-i}, y_{-i}\Big].$$

Hence, to bound both the terms, it is sufficient to control q-th moment of the functional. From Lemma S.1, for $q \leq 2 + \min\{\mu/2, \nu/2\}$,

$$\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \mid X_{-i}, y_{-i}\Big] \leq \big(C_1 + C_2\big\|\widehat{\beta}_{-i,\lambda}\big\|_2\big)^{2q}$$

for some positive constants C_1 and C_2. Combined Lemma S.3 that implies $\|\widehat{\beta}_{-i,\lambda}\|_2 \leq C$ almost surely for n large enough under Assumptions 1 and 2, we have

$$\mathbb{E}\left[\mathbb{E}\left[\left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\right|^q \mid X_{-i}, y_{-i}\right]\right] \leq C$$

for some constant $C > 0$ and $2 \leq q \leq 2 + \min\{\mu/2, \nu/2\}$.

Therefore, from (S.10) we can bound q-th moment of normalized sum (S.9) to get

$$\begin{aligned}
&\mathbb{E}\left[\left|\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right|^q\right] \\
&\leq \frac{(nC)^{q/2} + nC}{n^q} \\
&\leq C\frac{1}{n^{q/2}} + C\frac{1}{n^{q-1}}.
\end{aligned}$$

Finally, choosing $2 < q \leq 2 + \min\{\mu/2, \nu/2\}$ and applying Lemma S.14 provides the desired convergence for the first piece in (S.6). This concludes the proof.

S.1.3 LOOCV Estimator to GCV Estimator

To prove $\widehat{T}_\lambda^{\text{gcv}} - \widehat{T}_\lambda^{\text{loo}} \xrightarrow{\text{a.s.}} 0$, we start by bounding the absolute difference of interest by the average of absolute differences for $i = 1, \ldots, n$:

$$\begin{aligned}
\left|\widehat{T}_\lambda^{\text{gcv}} - \widehat{T}_\lambda^{\text{loo}}\right| &= \left|\frac{1}{n}\sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \text{tr}[L_\lambda]/n}\right) - \frac{1}{n}\sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n \left|t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \text{tr}[L_\lambda]/n}\right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right|. \qquad \text{(S.11)}
\end{aligned}$$

We will show below that the right-hand side of the expression (S.11) almost surely goes to zero. As with the proof of $\widetilde{T}_\lambda - \widehat{T}_\lambda \xrightarrow{\text{a.s.}} 0$, we will first assume $L_{ii} \neq 0$ so (S.11) is well defined. We will indicate the changes that we need to make when $L_{ii} = 0$ towards the end of the proof.

Using the modulus of continuity of t and it least concave majorant, we have

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^n \left|t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \text{tr}[L_\lambda]/n}\right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right| &\leq \frac{1}{n}\sum_{i=1}^n \omega_t\left(\left|\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \text{tr}[L_\lambda]/n} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\leq \frac{1}{n}\sum_{i=1}^n \overline{\omega}_t\left(\left|\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \text{tr}[L_\lambda]/n} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\leq \overline{\omega}_t\left(\frac{1}{n}\sum_{i=1}^n \left|\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \text{tr}[L_\lambda]/n} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\leq 2\omega_t\left(\frac{1}{n}\sum_{i=1}^n \left|\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \text{tr}[L_\lambda]/n} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\leq 2\omega_t\left(\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_\lambda\right| \left|\frac{1}{1 - \text{tr}[L_\lambda]/n} - \frac{1}{1 - [L_\lambda]_{ii}}\right|\right).
\end{aligned}$$

In the above chain on inequalities, we used Jensen's inequality on the concave majorant $\overline{\omega}_t$ for the third line, and monotonicity of ω_t on the fifth line.

Thus, from continuity of ω_t at 0, we will be done by showing

$$\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_\lambda\right| \left|\frac{1}{1 - \text{tr}[L_\lambda]/n} - \frac{1}{1 - [L_\lambda]_{ii}}\right| \xrightarrow{\text{a.s.}} 0. \qquad \text{(S.12)}$$

To build towards proving (S.12), let us denote by $r \in \mathbb{R}^n$ the vector of residuals $y_i - x_i^\top \widehat{\beta}^\lambda$ and by $d \in \mathbb{R}^n$ the vector of differences $(1 - \mathrm{tr}[L_\lambda]/n)^{-1} - (1 - [L_\lambda]_{ii})^{-1}$. Observe that

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_\lambda\right| \left|\frac{1}{1-\mathrm{tr}[L_\lambda]/n} - \frac{1}{1-[L_\lambda]_{ii}}\right| &= \frac{1}{n} r^\top d \\ &\leq \frac{1}{n}\|r\|_1 \|d\|_\infty \\ &\leq \frac{1}{\sqrt{n}}\|r\|_2 \|d\|_\infty,\end{aligned}$$

where we used Hölder's inequality in the second line and the the bound $\|a\|_1 \leq \sqrt{n}\|a\|_2$ for any $a \in \mathbb{R}^n$ in the last line. Since $r = (I - L_\lambda)y$, and the operator norm of $I - L_\lambda$ is bounded for $\lambda \in (\lambda_{\min}, 0)$ and $\|y\|_2/\sqrt{n}$ is almost surely bounded for sufficiently large n from the strong law of large numbers under Assumption 2, we have that $\|r\|_2/\sqrt{n}$ is eventually almost surely bounded. We now show in the sequel that $\|d\|_\infty \xrightarrow{\text{a.s.}} 0$ leading to the desired conclusion.

First for each $i = 1, \ldots, n$, by adding and subtracting $1 + \mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n$, and $\mathrm{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n$, we decompose the difference

$$\begin{aligned}&\left|\frac{1}{1-\mathrm{tr}[L_\lambda]/n} - \frac{1}{1-[L_\lambda]_{ii}}\right| \\ &= \Bigg|\frac{1}{1-\mathrm{tr}[L_\lambda]/n} - \big(1 + \mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n\big) + \mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n - \mathrm{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n \\ &\qquad + \big(1 + \mathrm{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n\big) - \frac{1}{1-[L_\lambda]_{ii}}\Bigg| \\ &\leq \left|\frac{1}{1-\mathrm{tr}[L_\lambda]/n} - \big(1 - \mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n\big)\right| \\ &\qquad + \left|\mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n - \mathrm{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n\right| \\ &\qquad\quad + \left|\big(1 - \mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n\big) - \frac{1}{1-[L_\lambda]_{ii}}\right|.\end{aligned}$$

This lets us decompose

$$\begin{aligned}\|d\|_\infty &= \max_{1\leq i\leq n}\left|\frac{1}{1-\mathrm{tr}[L_\lambda]/n} - \frac{1}{1-[L_\lambda]_{ii}}\right| \\ &\leq \left|\frac{1}{1-\mathrm{tr}[L_\lambda]/n} - \big(1 - \mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n\big)\right| \\ &\qquad + \max_{1\leq i\leq n}\left|\mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n - \mathrm{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n\right| \\ &\qquad\quad + \max_{1\leq i\leq n}\left|\big(1 - \mathrm{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n\big) - \frac{1}{1-[L_\lambda]_{ii}}\right|.\end{aligned}$$

Finally, we verify that each of the term in the decomposition almost surely vanishes. Using the $\lambda \neq 0$ case of Lemma S.11, we have for the first term

$$\left|\frac{1}{1-\mathrm{tr}[L_\lambda]/n} - \big(1 - \mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n\big)\right| \xrightarrow{\text{a.s.}} 0.$$

For the second term, following the proof of Lemma S.11, for $i = 1, \ldots, n$ we can bound

$$\left|\mathrm{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n - \mathrm{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n\right| \leq C/n,$$

almost surely for sufficiently large n. This uses the Sherman-Morrison-Woodbury formula with Moore-Penrose inverse to express the difference

$$(X^\top X/n + \lambda I)^\dagger - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger = -\frac{(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger x_i x_i^\top /n (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger}{1 + x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger x_i}. \tag{S.13}$$

The second term thus almost surely goes to zero. For the third term, note that from using the Sherman-Morrison-Woodbury formula again, we can simplify

$$\begin{aligned}1 - [L_\lambda]_{ii} &= 1 - x_i^\top (X^\top X/n + \lambda I)^\dagger x_i/n \\ &= 1 - x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I + x_i x_i^\top/n)^\dagger x_i/n \\ &= \frac{1}{1 + x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger x_i/n}.\end{aligned}$$

Therefore, for $q \geq 2$, we can now proceed to bound the q-th moment of the second term as

$$\begin{aligned}&\mathbb{E}\left[\left\{\max_{1\leq i\leq n}\left|1 + \mathrm{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n - \frac{1}{1-[L_\lambda]_{ii}}\right|\right\}^q\right] \\ &= \mathbb{E}\left[\left\{\max_{1\leq i\leq n}\left|1 + \mathrm{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n - \big(1 + x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger/n\big)\right|\right\}^q\right] \\ &= \mathbb{E}\left[\left\{\max_{1\leq i\leq n}\left|\mathrm{tr}\big[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\big]/n - x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger/n\right|\right\}^q\right] \\ &\leq \max_{1\leq i\leq n}\mathbb{E}\left[\left\{\left|\mathrm{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n - x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger/n\right|\right\}^q\right] \\ &\leq n\mathbb{E}\left[\left\{\mathrm{tr}\left[(X_{-j}^\top X_{-j}/n + \lambda I)^\dagger \Sigma\right]/n - x_j^\top (X_{-j}^\top X_{-j}/n + \lambda I)^\dagger x_j/n\right\}^q\right]\end{aligned}$$

for any $j = 1, \ldots, n$. Note that the last line follows from noting that $\mathrm{tr}\big[(X_{-j}^\top X_{-j}/n + \lambda I)^\dagger \Sigma\big]/n$, and $x_i^\top \big(X_{-i}^\top X_{-i}/n + \lambda I\big)^\dagger x_i$ are identically distributed for $i = 1, \ldots, n$. Since

$$\mathrm{tr}\big[(X_{-j}^\top X_{-j}/n + \lambda I)^\dagger\big]/n \leq C/n$$

almost surely for sufficiently large n, using Lemma S.10, the above quantity is of order $O(n/n^q)$. Choosing $q > 2$ and applying Lemma S.14 thus provides the desired almost sure convergence.

The above argument assumed that $L_{ii} \neq 0$. For the case of min-norm interpolator when $L_{ii} = 0$, we follow exactly similar steps as above using the modified errors defined in (13) and (14). (For more details on the λ cancellation for modified errors, see the proof of $\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{W}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$ in Section S.1.4.) This reduces to showing

$$\frac{1}{n}\sum_{i=1}^n \big|[(XX^\top/n)^\dagger y]_i\big| \left|\frac{1}{\mathrm{tr}[(XX^\top/n)^\dagger]/n} - \frac{1}{[(XX^\top/n)^\dagger]_{ii}}\right| \xrightarrow{\text{a.s.}} 0. \tag{S.14}$$

The same way we argued the almost sure boundedness of $\|r\|_2$, we can bound the norm of modified error vector $(XX^\top/n)^\dagger y$ as shown in Section S.1.4. Finally, analogous to the argument used to bound d, we can now use the case of $\lambda = 0$ equivalence in Lemma S.11 for the difference vector in the modified errors of (S.14). This takes care of both the cases and concludes the proof.

S.1.4 Truncation Arguments

We established the converges in Sections S.1.1 to S.1.3 under the the assumption that the error function t is uniformly continuous. In this section, we relax this assumption to t being only continuous by a truncation argument. Let $\mathbb{I}\{\mathcal{A}\}$ denote the indicator function for set $\mathcal{A}$.

Let t be a continuous error function. Define $w : \mathbb{R} \to \mathbb{R}$ to be the truncation of t on the compact interval $[-n, n]$, in other words, $w(r) = t(r)\mathbb{I}\{|r| \leq n\}$. Let W_λ denote the linear functional (5) corresponding to the error function w, and let $\widetilde{W}_\lambda$ be the intermediate averaged LOO functional defined analogously to (19) using w. Let $\widehat{W}_\lambda^{\mathrm{gcv}}$ and $\widehat{W}_\lambda^{\mathrm{loo}}$ denote the plug-in GCV and LOOCV estimators associated with w. The arguments in Sections S.1.1 to S.1.3 establish $W_\lambda - \widetilde{W}_\lambda \xrightarrow{\text{a.s.}} 0$, $\widetilde{W}_\lambda - \widehat{W}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} 0$, and $\widehat{W}_\lambda^{\mathrm{loo}} - \widehat{W}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$. We will now show that $T_\lambda - W_\lambda \xrightarrow{\text{a.s.}} 0$, $\widetilde{T}_\lambda - \widetilde{W}_\lambda \xrightarrow{\text{a.s.}} 0$, $\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{W}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$, $\widehat{T}_\lambda^{\mathrm{loo}} - \widehat{W}_\lambda^{\mathrm{loo}} \xrightarrow{\text{a.s.}} 0$ to finish the proof of Theorem 3. Since the proof of LOOCV mirrors that for GCV, we will only show the argument for GCV to avoid repetition.

Showing $T_\lambda - W_\lambda \xrightarrow{\text{a.s.}} 0$.

We can bound the absolute difference as follows:

$$\begin{aligned}
|T_\lambda - W_\lambda| &= \Big|\mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] - \mathbb{E}\big[w(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big]\Big| \\
&= \Big|\mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - w(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big]\Big| \\
&= \Big|\mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda)\mathbb{I}\{|y_0 - x_0^\top \widehat{\beta}| > n\} \mid X, y\big]\Big| \\
&\leq \sqrt{\mathbb{E}\big[|t(y_0 - x_0^\top \widehat{\beta}_\lambda)|^2 \mid X, y\big]}\sqrt{\mathbb{P}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda| > n \mid X, y\big]} \\
&\leq C\sqrt{\mathbb{P}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda| > n \mid X, y\big]} \\
&\leq C\sqrt{\frac{\mathbb{E}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda|^2 \mid X, y\big]}{n^2}} \\
&\leq \frac{C}{n} \to 0,
\end{aligned}$$

where the third line uses the Cauchy-Schwarz inequality, the fourth line uses Lemmas S.1 and S.3 with $q = 2$, the fifth line uses Chebychev's inequality, and the last line again uses Lemmas S.1 and S.3 with t as the identity function and $q = 2$.

Showing $\widetilde{T}_\lambda - \widetilde{W}_\lambda \xrightarrow{\text{a.s.}} 0$.

We can bound the absolute difference as follows:

$$\begin{aligned}
\Big|\widetilde{T}_\lambda - \widetilde{W}_\lambda\Big| &= \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big] - \frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[w(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right| \\
&= \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - w(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big]\right| \\
&\leq \left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\mathbb{I}\big\{|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\big\} \mid X_{-i}, y_{-i}\Big]\right| \\
&\leq \frac{1}{n}\sum_{i=1}^n \sqrt{\mathbb{E}\Big[|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})|^2 \mid X_{-i}, y_{-i}\Big]}\sqrt{\mathbb{P}\Big\{|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n \mid X_{-i}, y_{-i}\Big\}} \\
&\leq \frac{1}{n}\sum_{i=i}^n \sqrt{\mathbb{E}\big[|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})|^2 \mid X_{-i}, y_{-i}\big]}\sqrt{\mathbb{P}\left\{\max_{j=1}^n |y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n \;\middle|\; X, y\right\}} \\
&\leq \left|\frac{1}{n}\sum_{i=i}^n \sqrt{\mathbb{E}\big[|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})|^2 \mid X_{-i}, y_{-i}\big]}\right|\sqrt{\mathbb{P}\left\{\max_{j=1}^n |y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\right\}} \\
&\leq C\sqrt{\mathbb{P}\left\{\max_{j=1}^n |y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\right\}}.
\end{aligned}$$

Above, line four uses the Cauchy-Schwarz inequality, line five uses the fact that the event $|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n$ for any $i = 1, \ldots, n$ is contained inside the event $\max_{j=1}^n |y_j - x_j^\top \widehat{\beta}_{-j,\lambda}| > n$, and the last line follows from the q-th moment control as done in Section S.1.2 with $q = 2$. It therefore suffices to bound the probability of the event $\max_{j=1}^n |y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n$ which we do below.

Starting with union bound, we have that

$$\begin{aligned}\mathbb{P}\left\{\max_{j=1}^{n}|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\right\} &\le \sum_{i=1}^{n}\mathbb{P}\left\{|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}| > n\right\} \\ &\le \sum_{i=1}^{n}\frac{\mathbb{E}\big[|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}|^2\big]}{n^2} \\ &\le \sum_{i=1}^{n}\frac{C}{n^2} \\ &\le \frac{C}{n} \to 0.\end{aligned}$$

Showing $\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{W}_\lambda^{\mathrm{gcv}} \xrightarrow{\text{a.s.}} 0$.

By following similar argument used to bound $\big|\widetilde{T}_\lambda - \widetilde{W}_\lambda\big|$, it suffices to show that

$$\mathbb{P}\left\{\max_{j=1}^{n}\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-\mathrm{tr}[L_\lambda]/n} > n\right\} \to 0.$$

Using the union bound, it is thus enough to show that almost surely

$$\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-\mathrm{tr}[L_\lambda]/n}\right)^2 \le C.$$

Note that this is valid when $\lambda \neq 0$. To cover the case of min-norm interpolator, we start by rewriting the residuals in an alternate form as follows:

$$\begin{aligned}y_i - x_i^\top \widehat{\beta}_\lambda &= y_i - x_i^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n \\ &= y_i - [X^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n]_i \\ &= [y - X^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n]_i \\ &= [(I - X^\top (X^\top X/n + \lambda I)^\dagger X/n)y]_i \\ &= \lambda[(XX^\top/n + \lambda I)^\dagger y]_i \end{aligned} \tag{S.15}$$

Similarly, we rewrite the denominator of GCV using

$$\begin{aligned}1 - \mathrm{tr}[L_\lambda]/n &= 1 - \mathrm{tr}[X(XX^\top/n + \lambda I)^\dagger X^\top]/n \\ &= \mathrm{tr}[I - X(XX^\top/n + \lambda I)^\dagger X^\top]/n \\ &= \lambda\,\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n. \end{aligned} \tag{S.16}$$

This lets us rewrite the invidual GCV reweighted errors as

$$\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-\mathrm{tr}[L_\lambda]/n} = \frac{\lambda[(XX^\top/n + \lambda I)^\dagger y]_i}{\lambda\,\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n} = \frac{[(XX^\top/n + \lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n}.$$

Thus, we can now bound

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1-\mathrm{tr}[L_\lambda]/n}\right)^2 &= \frac{\big\|(XX^\top/n + \lambda I)^\dagger y\big\|_2^2/n}{\big(\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n\big)^2} \\ &\le \frac{\big\|(XX^\top/n + \lambda I)^\dagger\big\|_{\mathrm{op}}^2\|y\|_2^2/n}{\big(\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n\big)^2}.\end{aligned}$$

Each term in the above ratio is almost surely bounded for sufficiently large n under Assumption 1 and Assumption 2 as explained in the proof of Lemma S.3. This finishes the argument.

S.1.5 Auxiliary Lemmas

In this section, we gather supporting lemmas used in the proofs in Sections S.1.1 to S.1.3, along with their proofs.

Lemma S.1 (Bounding conditional q-th moment of the i-th LOO residual)**.** *Suppose Assumptions 1 and 2 hold, and the error function t satisfies Assumption 3. Then, for $q \le \min\{\mu/2, \nu/2\}$ and $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathbb{E}\Big[\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \,\big|X_{-i}, y_{-i}\Big] \le \big(C_1 + C_2 \|\widehat{\beta}_{-i,\lambda}\|_2\big)^{2q}$$

for some positive constants C_1 and C_2.

Proof. Note that under Assumption 3, $\big|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\big|^q \le a\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} + b\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^q + c$ for some positive constants a, b, c. Because $\mathbb{E}\big[Z^{q_l}\big] \le \mathbb{E}\big[Z^{q_h}\big]^{q_l/q_h}$ for $q_l \le q_h$ from Jensen's inequality, it suffices to bound $\mathbb{E}\big[\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\big]$, which we do below.

From the triangle inequality for the conditional L_q norm, observe that

$$\begin{aligned}\mathbb{E}\Big[\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q} &\le \mathbb{E}\Big[\big|y_i\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q} + \mathbb{E}\Big[\big|x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q} \\ &\le \mathbb{E}\Big[\big|y_i\big|^{2q}\Big]^{1/2q} + \mathbb{E}\Big[\big|x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q}.\end{aligned}$$

The first term is bounded for $q \le 2 + \mu/2$ under Assumption 2. For the second term, start by writing

$$\mathbb{E}\Big[\big|x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big] = \mathbb{E}\Big[\big|z_i^\top \Sigma^{1/2} \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big].$$

Note that conditional on X_{-i} and y_{-i}, $\Sigma^{1/2}\widehat{\beta}_{-i,\lambda}$ is a fixed vector in $\mathbb{R}^p$. For $q \le 2 + \nu/2$, Lemma S.9 then provides

$$\mathbb{E}\Big[\big|x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big]^{1/2q} \le C\|\Sigma^{1/2}\widehat{\beta}_{-i,\lambda}\|_2 \le C\sqrt{r_{\max}}\|\widehat{\beta}_{-i,\lambda}\|_2,$$

where the last inequality follows since the maximum eigenvalue of Σ is bounded by $r_{\max}$. Therefore, for $q \le 2 + \min\{\mu/2, \nu/2\}$, we get

$$\mathbb{E}\Big[\big|y_i - x_i^\top \widehat{\beta}_{-i,\lambda}\big|^{2q} \mid X_{-i}, y_{-i}\Big] \le \big(C_1 + C_2\|\widehat{\beta}_{-i,\lambda}\|_2\big)^{2q}$$

for some positive constants C_1 and C_2 as desired. This completes the proof. □

Lemma S.2 (Bounding norm of the difference of leave-one-out ridge estimators)**.** *Suppose Assumptions 1 and 2 hold. Then, for $\lambda \in (\lambda_{\min}, \infty)$,*

$$\frac{1}{n}\sum_{i=1}^n \big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2 \xrightarrow{\text{a.s.}} 0$$

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

Proof. For each $i = 1, \dots, n$, we start by breaking the difference

$$\begin{aligned}\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda} &= (X^\top X/n + \lambda I)^\dagger X^\top y/n - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger X_i^\top y_{-i}/(n-1) \\ &= (X^\top X/n + \lambda I)^\dagger X^\top y/n - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger X^\top y/n \\ &\quad + (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger X^\top y/n - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger X_{-i}^\top y_{-i}/(n-1) \\ &= \big\{(X^\top X/n + \lambda I)^\dagger - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\} X^\top y/n \\ &\quad + (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \big\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\}.\end{aligned}$$

Applying the triangle inequality, for each $i = 1, \dots, n$, we can then bound

$$\begin{aligned}\big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2 \le{}& \big\|\big\{(X^\top X/n + \lambda I)^\dagger - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger\big\} X^\top y/n\big\|_2 \\ &+ \big\|(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \big\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\}\big\|_2.\end{aligned}$$

Averaging the bounds above thus provides

$$\frac{1}{n}\sum_{i=1}^{n}\big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2 \le \frac{1}{n}\sum_{i=1}^{n}\big\|(X_{-i}^\top X_{-i}/n+\lambda I)^\dagger\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\}\big\| + \frac{1}{n}\sum_{i=1}^{n}\big\|\{(X^\top X/n+\lambda I)^\dagger - (X_{-i}^\top X_{-i}/n+\lambda I)^\dagger\}X^\top y/n\big\|. \tag{S.17}$$

We will see below that each of the two terms on the right-hand side of (S.17) almost surely goes to zero providing the desired convergence. Note that for each $i=1,\dots,n$, we can bound

$$\begin{aligned}\big\|(X_{-i}^\top X_{-i}/n+\lambda I)^\dagger\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\}\big\|_2 &\le \big\|(X_{-i}^\top X_{-i}/n+\lambda I)^\dagger\big\|_{\mathrm{op}}\big\|X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\|_2\\ &\le C\big\|X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\big\|_2\\ &= C\left\|\frac{x_i y_i}{n} - \sum_{j\neq i}\frac{x_j y_j}{(n-1)n}\right\|_2\\ &\le \frac{C}{\sqrt{n}}\frac{\|x_i y_i\|_2}{\sqrt{n}} + \frac{C}{(n-1)\sqrt{n}}\sum_{j\neq i}\frac{\|x_j y_j\|_2}{\sqrt{n}},\end{aligned}$$

where the second line follows from the fact that $\big\|(X_{-i}^\top X_{-i}/n+\lambda I)^\dagger\big\|_{\mathrm{op}}$ is almost surely bounded for n large enough (as explained in the proof of Lemma S.3), and last line uses triangle inequality. Now writing $x_i = \Sigma^{1/2}z_i$, note that for each $i=1,\dots,n$,

$$\big\|x_i y_i\big\|_2/\sqrt{n} = \big\|\Sigma^{1/2}z_i y_i\big\|_2/\sqrt{n} \le \big\|\Sigma^{1/2}\big\|_{\mathrm{op}} y_i\big\|z_i\big\|_2/\sqrt{n} \le y_i\big\|z_i\big\|_2/\sqrt{n} \le C y_i$$

almost surely for sufficiently large n since $\|z_i\|_2/\sqrt{n}$ is eventually almost surely bounded from the strong law of large numbers. Hence, we have

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\big\|(X_{-i}^\top X_{-i}+\lambda I)^\dagger\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\}\big\| &\le \frac{C}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}|y_i| + \frac{C}{(n-1)\sqrt{n}}\frac{1}{n}\sum_{i=i}^{n}\sum_{j\neq i}|y_j|\\ &\le \frac{C}{\sqrt{n}}\frac{(2n-1)}{(n-1)n}\sum_{i=1}^{n}|y_i|\\ &\le \frac{C}{\sqrt{n}} \to 0.\end{aligned} \tag{S.18}$$

Here the second inequality follows by adding $|y_i|$ to the second term, and the last inequality follows because $\sum_{i=1}^n |y_i|/n$ is eventually almost surely bounded from the strong law of large numbers under Assumption 2. Using the leave-one-out sample covariance difference (S.13), we can similarly show that the second term goes to zero almost surely. Hence, we have that (S.17) almost surely goes to zero. This completes the proof.

□

Lemma S.3 (Bounding norm of the ridge estimator)**.** *Suppose Assumption 1 and Assumption 2 hold. Then, for $\lambda \in (\lambda_{\min},\infty)$, $\|\widehat{\beta}_\lambda\|_2 \le C$ for some positive constant C eventually almost surely.*

Proof. We can bound the norm of ridge estimator as

$$\begin{aligned}\big\|\widehat{\beta}_\lambda\big\|_2 &= \big\|(X^\top X/n+\lambda I)^\dagger X^\top y/n\big\|_2\\ &\le \big\|(X^\top X/n+\lambda I)^\dagger X^\top/\sqrt{n}\big\|_{\mathrm{op}}\|y\|_2/\sqrt{n}\\ &\le \big\|(X^\top X/n+\lambda I)^\dagger\big\|_{\mathrm{op}}\big\|X^\top/\sqrt{n}\big\|_{\mathrm{op}}\|y\|_2/\sqrt{n}.\end{aligned} \tag{S.19}$$

Now for $\lambda \in (\lambda_{\min},\infty)$, the first two terms in the product (S.19) are almost surely bounded for n large enough. This is because the maximum eigenvalue of $X^\top X/n$ is upper bounded by $C(1+\sqrt{\gamma})^2 r_{\max}$ for some $C>1$ and the minimum non-zero eigenvalue is lower bounded by $c(1-\sqrt{\gamma})^2 r_{\min}$ for some $c<1$ almost surely for sufficiently large n under Assumption 1 (Bai and Silverstein, 1998). From the strong law of large numbers, the final term is eventually almost surely bounded as the second moment of the response is bouned under Assumption 2. Hence, the product is eventually almost surely bounded, finishing the proof. □

S.2 PROOFS RELATED TO Theorem 4

To show almost sure uniform convergence (in λ), we will appeal to Lemma S.12. A sufficient condition to establish strong stochastic equicontinuity in the current differentiable case is uniform boundness of the associated functions and their derivatives (with respect to λ) (e.g., Chpater 21 of Davidson, 1994). We will show that both T_λ and $\widehat{T}_\lambda^{\mathrm{gcv}}$ and their derivates are bounded over Λ, implying strong stochastic equicontinuity of the family of functions $\{T_\lambda - \widehat{T}_\lambda^{\mathrm{gcv}}\}_{\lambda\in\Lambda}$. Analogous analysis holds for $\{T_\lambda - \widehat{T}_\lambda^{\mathrm{loo}}\}_{\lambda\in\Lambda}$, which we omit due to its similarity with the GCV analysis. Recall that Λ is a compact set in $(\lambda_{\min}, \infty)$. In the following, let $\Lambda \subset [\underline{\lambda}, \overline{\lambda}]$ where $\lambda_{\min} < \underline{\lambda} \le \overline{\lambda} < \infty$.

Bounding T_λ. We start with T_λ. Using Lemma S.1 with $q = 1$, under Assumptions 1 and 2, for error function t satisfying Assumption 3, we can bound T_λ in terms of the norm of the ridge estimator $\widehat{\beta}_\lambda$ as

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \le \big(C_1 + C_2 \|\widehat{\beta}_\lambda\|_2\big)^2, \tag{S.20}$$

for some positive constants C_1 and C_2. Now following Lemma S.3, over Λ, we have that $\|\widehat{\beta}_\lambda\|_2$ is eventually almost surely bounded by $C\sqrt{r_{\max}}(\lambda_{\min} + \underline{\lambda})^{-1}$ for some positive constant C (independent of λ). This shows that T_λ is eventually almost surely bounded over $\lambda \in \Lambda$.

Bounding $\widehat{T}_\lambda^{\mathrm{gcv}}$. We next consider $\widehat{T}_\lambda^{\mathrm{gcv}}$. Using the alternate representation (S.15), for error function t satisfying Assumption 3, for some positive constants C, C_1, C_2, we can bound

$$\begin{aligned}
\widehat{T}_\lambda^{\mathrm{gcv}} &= \frac{1}{n}\sum_{i=1}^n t\left(\frac{[(XX^\top/n + \lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n}\right) \\
&\le \frac{C_2}{n}\sum_{i=1}^n \frac{\big\{[(XX^\top/n + \lambda I)^\dagger y]_i\big\}^2}{\big\{\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n\big\}^2} + \frac{C_1}{n}\sum_{i=1}^n \frac{\big|[(XX^\top/n + \lambda I)^\dagger y]_i\big|}{\big|\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n\big|} + C \\
&\le \frac{C_2}{n}\sum_{i=1}^n \big\{[(XX^\top/n + \lambda I)^\dagger y]_i\big\}^2 + \frac{C_1}{n}\sum_{i=1}^n \big|[(XX^\top/n + \lambda I)^\dagger y]_i\big| + C.
\end{aligned} \tag{S.21}$$

The last inequality above follows by noting that the map $\lambda \mapsto \mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n$ is non-increasing over $[\underline{\lambda}, \overline{\lambda}]$, so $\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n$ is lower bounded by $\mathrm{tr}[(XX^\top/n + \underline{\lambda} I)^\dagger/n]$. Since $\lambda_{\min} < \underline{\lambda}$, we then have that $\{\mathrm{tr}[(XX^\top/n + \lambda I)^\dagger]/n\}^{-1}$ is upper bounded by $(\lambda_{\min} + \underline{\lambda})^{-1}$. Now, observe that for the first term in (S.21):

$$\frac{1}{n}\sum_{i=i}^n \big\{[(XX^\top/n + \lambda I)^\dagger y]_i\big\}^2 = \frac{1}{n}\big\|(XX^\top/n + \lambda I)^\dagger y\big\|_2^2 \le \frac{1}{n}\big\|(XX^\top/n + \lambda I)^\dagger\big\|_{\mathrm{op}}^2 \|y\|_2^2.$$

Similarly, note that for the second term in (S.21):

$$\frac{1}{n}\sum_{i=1}^n \big|[(XX^\top/n + \lambda I)^\dagger y]_i\big| = \frac{1}{n}\big\|(XX^\top/n + \lambda I)^\dagger y\big\|_1 \le \frac{1}{\sqrt{n}}\big\|(XX^\top/n + \lambda I)^\dagger y\big\|_2 \le \frac{1}{\sqrt{n}}\big\|(XX^\top/n + \lambda I)^\dagger\big\|_{\mathrm{op}}\|y\|_2.$$

Since $\|(XX^\top/n + \lambda I)^\dagger\|_{\mathrm{op}}$ is uniformly bounded over $\lambda \in \Lambda$ under Assumption 1 as argued above, and $\|y\|_2^2/n$ is almost surely bouned for n large enough from the law of large numbers under Assumption 2, it follows that $\widehat{T}_\lambda^{\mathrm{gcv}}$ is almost surely bounded over $\lambda \in \Lambda$.

Bounding derivative of T_λ. We now turn to bounding the derivaties of the map $\lambda \mapsto T_\lambda$. First note that since $\mathbb{E}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda| \mid X, y\big] \le \mathbb{E}\big[|y_0 - x_0^\top \widehat{\beta}_\lambda|^2 \mid X, y\big]^{1/2}$, and since the latter is almost surely bounded as shown above, we can switch the order of differentiation and integration. The derivative of T_λ with respect to λ can then be bounded above by

$$T_\lambda' = \mathbb{E}\big[t'(y_0 - x_0^\top \widehat{\beta}_\lambda)\; x_0^\top \widehat{\beta}_\lambda' \mid X, y\big] \le \mathbb{E}\big[\{t'(y_0 - x_0^\top \widehat{\beta}_\lambda)\}^2 \mid X, y\big]^{1/2} \cdot \mathbb{E}\big[(\widehat{\beta}_\lambda')^\top x_0 x_0^\top \widehat{\beta}_\lambda' \mid X, y\big] \le C\sqrt{r_{\max}}\|\widehat{\beta}_\lambda'\|_2. \tag{S.22}$$

In the above chain, the first inequality follows from Cauchy-Schwarz inequality, and the second inequality follows from the bounding of T_λ per (S.20) above (because under Assumption 3, t' is bounded above by a linear function), and the fact that $\|\Sigma\|_{\mathrm{op}} \le r_{\max}$. Applying Lemma S.4 on the last term of (S.22), we thus conclude that the derivative of T_λ is almost surely uniformly bounded over $\lambda \in \Lambda$, as desired.

Bounding derivative of $\widehat{T}_\lambda^{\text{gcv}}$. Finally, we bound the derivative of the map $\lambda \mapsto \widehat{T}_\lambda^{\text{gcv}}$. From the chain rule, the derivative of $\widehat{T}_\lambda^{\text{gcv}}$ with respect to λ can be expressed as

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n t'\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right) \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^n\left\{t'\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2}\sqrt{\frac{1}{n}\sum_{i=1}^n\left\{\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2} \qquad \text{(S.23)} \\
&\leq C\sqrt{\sum_{i=1}^n\left\{\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2} \qquad \text{(S.24)}
\end{aligned}$$

The first inequality above again follows from the Cauchy-Schwarz inequalty. The second inequality follows since, from Assumption 3, t' is bouned above by a linear function, and the bounding of $\widehat{T}_\lambda^{\text{gcv}}$ per (S.21) above shows that the first term of (S.23) is almost surely bounded. Applying Lemma S.5, we can now upper bound the final term of (S.24). This leads the derivative of $\widehat{T}_\lambda^{\text{gcv}}$ to be almost surely bounded over $\lambda \in \Lambda$ and concludes the proof.

Lemma S.4 (Bounding norm of the derivative of ridge estimator)**.** *Suppose Assumptions 1 and 2 hold. Then, for $\lambda \in (\lambda_{\min}, \infty)$, $\|\widehat{\beta}'_\lambda\|_2 \leq C$ eventually almost surely for some positive constant C.*

Proof. The proof follows from a straightforward calculation. Expressing the ridge estimation in the gram form, observe that

$$\frac{d\widehat{\beta}_\lambda}{d\lambda} = \frac{dX^\top(XX^\top/n+\lambda I)^\dagger y/n}{d\lambda} = X^\top(XX^\top/n+I)^\dagger(XX^\top/n+\lambda I)^\dagger y/n.$$

In the above, we use the fact that for $\lambda \in (\lambda_{\min}, \infty)$, the map $\lambda \mapsto (XX^\top/n+\lambda I)^\dagger$ is almost surely differentiable for n large enough, with the derivative given by $(XX^\top/n+\lambda I)^\dagger(XX^\top/n+\lambda I)^\dagger$. The result then follows by noting that the opeator norms of $X/\sqrt{n}$ and $(XX^\top/n+\lambda I)^\dagger$ are uniformly bounded over Λ as argued above, and $\|y\|_2/\sqrt{n}$ is almost surely bounded for n large enough, as explained in the proof of Lemma S.3. □

Lemma S.5 (Bounding norm of the derivative of modified GCV residuals)**.** *Suppose Assumptions 1 and 2 hold. Then, for $\lambda \in (\lambda_{\min}, \infty)$, we have that*

$$\frac{1}{\sqrt{n}}\left\|\frac{d}{d\lambda}\left(\frac{(XX^\top/n+\lambda I)^\dagger y}{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\|_2 \leq C$$

eventually almost surely for some positive contant C.

Proof. The proof uses straightforward matrix calculus (Petersen et al., 2008). Using the chain rule, we can write

$$\begin{aligned}
\frac{d}{d\lambda}\left(\frac{(XX^\top/n+\lambda I)^\dagger y}{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right) &= -\frac{\text{tr}[(XX^\top/n+\lambda I)^\dagger(XX^\top/n+\lambda I)^\dagger]/n}{\{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n\}^2}(XX^\top/n+\lambda I)^\dagger y \\
&\quad + \frac{1}{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\frac{d}{d\lambda}\big((XX^\top/n+\lambda y)^\dagger y\big).
\end{aligned}$$

Note that $\{\text{tr}[(XX^\top/n+\lambda I)^\dagger]/n\}^{-1}$ is almost surely bounded for n sufficiently large as argued above. In addition, since the operator norm of $(XX^\top/n+\lambda I)^\dagger$ is uniformly upper bounded for $\lambda \in \Lambda$, we also have that $\text{tr}[(XX^\top/n+\lambda I)^\dagger(XX^\top/n+\lambda I)^\dagger]/n$ is uniformly upper bounded over Λ. Next, observe that

$$\frac{d}{d\lambda}\big((XX^\top/n+\lambda I)^\dagger y\big) = (XX^\top/n+\lambda I)^\dagger(XX^\top/n+\lambda I)^\dagger y.$$

As above, since the opeator norm of $(XX^\top/n+\lambda I)^\dagger$ is uniformly bounded for $\lambda \in \Lambda$, and $\|y\|_2/\sqrt{n}$ is almost surely bounded for n large enough, the result then follows from simple application of the triangle inequality (with respect to the ℓ_2 norm). This finishes the proof. □

S.3 PROOFS RELATED TO Theorem 5

The proof is similar to that of proof of Theorem 4. We will again use Lemma S.12. In the current the nonsmooth case, it is sufficient to show that the family of random functions under consideration is almost surely Lipschitz continuous, along with the almost sure uniform bounds as shown in the proof of Theorem 4 (see, e.g., Chpater 21 of Davidson, 1994). We will show in the two helper lemmas below that this holds for $\{T_\lambda\}_{\lambda\in\Lambda}$ and $\{\widehat{T}_\lambda^{\text{gcv}}\}_{\lambda\in\Lambda}$, assuming that the loss function t is Lipschitz continuous. This will show that $\{T_\lambda - \widehat{T}_\lambda^{\text{gcv}}\}_{\lambda\in\Lambda}$ is almost surely Lipschitz continuous from which the theorem follows. A similar analysis holds for $\{T_\lambda - \widehat{T}_\lambda^{\text{loo}}\}_{\lambda\in\Lambda}$.

Lemma S.6 (Lipschitz continuity of the out-of-sample functional)**.** *Suppose Assumption 1 and Assumption 2 hold, and the error function t is Lipschitz continuous. Let Λ be a compact set in $(\lambda_{\min}, \infty)$. Then, over Λ, the random map $\lambda \mapsto T_\lambda$ is almost surely Lipschitz continuous.*

Proof. Since Λ is compact, let $\Lambda \subseteq [\underline{\lambda}, \overline{\lambda}]$ where $\lambda_{\min} < \underline{\lambda} \le \overline{\lambda} < \infty$. For any $\lambda_1, \lambda_2 \in [\underline{\lambda}, \overline{\lambda}]$, using the Lipschitz continuity of the error function, we have

$$\left|t(y_0 - x_0^\top \widehat{\beta}_{\lambda_1}) - t(y_0 - x_0^\top \widehat{\beta}_{\lambda_2})\right| \le L\left|x_0^\top(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})\right|$$

for some $L \ge 0$. Now consider

$$\begin{aligned}
\left|T_{\lambda_1} - T_{\lambda_2}\right| &= \left|\mathbb{E}\left[t(y_0 - x_0^\top \widehat{\beta}_{\lambda_1}) - t(y_0 - x_0^\top \widehat{\beta}_{\lambda_2}) \mid X, y\right]\right| \\
&\le \mathbb{E}\left[\left|t(y_0 - x_0^\top \widehat{\beta}_{\lambda_1}) - t(y_0 - x_0^\top \widehat{\beta}_{\lambda_2})\right| \mid X, y\right] \\
&\le L\mathbb{E}\left[\left|x_0^\top(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})\right| \mid X, y\right] \\
&= L\mathbb{E}\left[\sqrt{\left|x_0^\top(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})\right|^2} \mid X, y\right] \\
&\le L\sqrt{\mathbb{E}\left[\left|x_0^\top(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})\right|^2 \mid X, y\right]} \\
&\le L\sqrt{\mathbb{E}\left[\left|(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})^\top x_0 x_0^\top(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})\right|^2 \mid X, y\right]} \\
&\le L\sqrt{(\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})^\top \Sigma (\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2})} \\
&\le L\sqrt{r_{\max}}\left\|\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}\right\|_2.
\end{aligned}$$

Above, the second and fourth lines follow from using Jensen's inequality (on the absolute and square root functions, respectively), the third line follows from the Lipschitz bound on the error function, and the last inequality follow since the operator norm of Σ is bounded above by $r_{\max}$.

To complete the proof, we show below that over $[\underline{\lambda}, \overline{\lambda}]$, $\|\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}\| \le C|\lambda_1 - \lambda_2|$ for some constant C that is eventually almost surely bounded. To see this, we start by writing the difference using equivalent gram representation for ridge estimator:

$$\begin{aligned}
\left\|\widehat{\beta}_{\lambda_1} - \widehat{\beta}_{\lambda_2}\right\|_2 &= \left\|X(XX^\top/n + \lambda_1)^\dagger y/n - X(XX^\top/n + \lambda_2)^\dagger y/n\right\|_2 \\
&\le \left\|X/\sqrt{n}\right\|_{\text{op}}\left\|(XX^\top/n + \lambda_1) - (XX^\top/n + \lambda_2)\right\|_{\text{op}}\left\|y\right\|_2/\sqrt{n}. \qquad \text{(S.25)}
\end{aligned}$$

As argued before, both the first and the last term in the product (S.25) are eventually almost surely bounded under Assumptions 1 and 2. For the middle term, note that on $[\underline{\lambda}, \overline{\lambda}]$, since $\lambda_{\min} < \underline{\lambda}$, the map $\lambda \mapsto (XX^\top/n + \lambda I)^\dagger$ is differentiable on $[\underline{\lambda}, \overline{\lambda}]$ with the derivative with respect to λ equal to $(XX^\top/n + \lambda I)^\dagger(XX^\top/n + \lambda I)^\dagger$. Thus, using the mean value theorem, for some $\lambda \in (\underline{\lambda}, \overline{\lambda})$, we can bound

$$\left|(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\right| \le \left|(XX^\top/n + \lambda I)^\dagger(XX^\top/n + \lambda I)^\dagger\right| |\lambda_1 - \lambda_2|.$$

Hence, we can bound the second term as

$$\begin{aligned}
\left\|(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\right\|_{\text{op}} &\le \left\|(XX^\top/n + \lambda I)^\dagger(XX^\top/n + \lambda I)^\dagger\right\|_{\text{op}} |\lambda_1 - \lambda_2| \\
&\le \left\|(XX^\top/n + \lambda I)^\dagger\right\|_{\text{op}}\left\|(XX^\top/n + \lambda I)^\dagger\right\|_{\text{op}} |\lambda_1 - \lambda_2| \\
&\le C\,|\lambda_1 - \lambda_2|, \qquad \text{(S.26)}
\end{aligned}$$

where the last inequality follows because $\lambda \geq \underline{\lambda} > \lambda_{\min}$ as explained in the proof of Lemma S.3. This concludes the proof. □

Lemma S.7 (Lipschitz continuity of the GCV functional)**.** *Suppose Assumption 1 and Assumption 2 hold, and the error function t is Lipschitz continuous. Let Λ be a compact set in $(\lambda_{\min}, \infty)$. Then, over Λ, the random map $\lambda \mapsto \widehat{T}_\lambda^{\mathrm{gcv}}$ is almost surely Lipschitz continuous.*

Proof. Let $\Lambda \subseteq [\underline{\lambda}, \overline{\lambda}]$, where $\lambda_{\min} < \underline{\lambda} \leq \overline{\lambda} < \infty$. Using the alternate representation (S.15) for the numerator and (S.16) for the denominator of GCV reweighted errors, we can rewrite the plug-in functional $\widehat{T}_\lambda^{\mathrm{gcv}}$ as

$$\widehat{T}_\lambda^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^n t\left(\frac{\left[(XX^\top/n + \lambda I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n+\lambda I)^\dagger\right]/n}\right).$$

For $\lambda_1, \lambda_2 \in \Lambda$ using the Lipschitz continuity of the error function, note that

$$\begin{aligned}
&\widehat{T}_{\lambda_1}^{\mathrm{gcv}} - \widehat{T}_{\lambda_2}^{\mathrm{gcv}} && \text{(S.27)}\\
&= \frac{1}{n}\sum_{i=1}^n t\left(\frac{\left[(XX^\top/n + \lambda_1 I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n+\lambda_1 I)^\dagger\right]/n}\right) - t\left(\frac{\left[(XX^\top/n + \lambda_2 I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n+\lambda_2 I)\right]/n}\right)\\
&\leq \frac{1}{n}\sum_{i=1}^n L\left|\frac{\left[(XX^\top/n + \lambda_1 I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n+\lambda_1 I)^\dagger\right]/n} - \frac{\left[(XX^\top/n + \lambda_2 I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n+\lambda_2 I)^\dagger\right]/n}\right|\\
&\leq L\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n+\lambda_1 I)^\dagger\right]/n} - \frac{1}{\mathrm{tr}\left[(XX^\top/n+\lambda_2 I)^\dagger\right]/n}\right| \frac{1}{n}\sum_{i=1}^n \left|\left[(XX^\top/n + \lambda_1 I)^\dagger y\right]_i - \left[(XX^\top/n + \lambda_2 I)^\dagger y\right]_i\right|\\
&\leq L\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n+\lambda_1 I)^\dagger\right]/n} - \frac{1}{\mathrm{tr}\left[(XX^\top/n+\lambda_2 I)^\dagger\right]/n}\right| \frac{1}{n}\sum_{i=1}^n \left|\left[\{(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\} y\right]_i\right|\\
&\leq L\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n+\lambda_1 I)^\dagger\right]/n} - \frac{1}{\mathrm{tr}\left[(XX^\top/n+\lambda_2 I)^\dagger\right]/n}\right| \frac{1}{n}\left\|\{(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\} y\right\|_1 && \text{(S.28)}
\end{aligned}$$

Since the map $\lambda \mapsto \mathrm{tr}\left[(XX^\top + \lambda I)^\dagger\right]/n$ is non-increasing over $[\underline{\lambda}, \overline{\lambda}]$, we can bound the first term of (S.28) using

$$\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n+\lambda_1 I)^\dagger\right]/n} - \frac{1}{\mathrm{tr}\left[(XX^\top/n+\lambda_2 I)^\dagger\right]/n}\right| \leq 2\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n+\underline{\lambda} I)^\dagger\right]/n}\right|. \tag{S.29}$$

For bounding the second term of (S.28), note that

$$\begin{aligned}
\left\|\{(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\} y\right\|_1 /n &\leq \left\|\{(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\} y\right\|_2 /\sqrt{n}\\
&\leq \left\|(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\right\|_{\mathrm{op}} \|y\|_2/\sqrt{n}\\
&\leq C\,|\lambda_1 - \lambda_2|\,, && \text{(S.30)}
\end{aligned}$$

where we used the bound from (S.26), along with the fact that $\|y\|_2/\sqrt{n}$ is almost surely bounded for n large enough from the strong law of large numbers under Assumption 2. Plugging (S.29) and (S.30) into (S.28) then finishes the proof. □

S.4 PROOF OF Theorem 1

Let $\widehat{F}_\lambda^{\mathrm{gcv}}$ and $\widehat{F}_\lambda^{\mathrm{loo}}$ denote the CDFs associated with the plug-in distributions $\widehat{P}_\lambda^{\mathrm{gcv}}$ and $\widehat{P}_\lambda^{\mathrm{loo}}$ of the GCV and LOOCV reweighted errors, respectively. Recall that F_λ denotes the CDF of the out-of-sample error distribution P_λ. To prove Theorem 1, for all $z \in \mathbb{R}$ that are continuity points of F_λ for n sufficiently large, we will sandwich $F_\lambda(z)$ such that, almost surely, $\limsup_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \leq F_\lambda(z)$ along with $F_\lambda(z) \leq \liminf_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z)$. This then yields the desired result that $\widehat{F}_\lambda^{\mathrm{gcv}}(z) - F_\lambda(z) \xrightarrow{\text{a.s.}} 0$. Similar argument shows $\widehat{F}_\lambda^{\mathrm{loo}}(z) - F_\lambda(z) \xrightarrow{\text{a.s.}} 0$. The idea of

the proof is similar to that used in the proof of the Portmanteau theorem, with the main difference being that the target distribution in our case is also a random distribution. We will make use of Theorem 3 to deduce the desired inequalities in each direction using suitably chosen error functions.

Fix $\epsilon > 0$ and $z \in \mathbb{R}$. For the first direction, let $t_{z,\epsilon}$ be an error function defined as

$$t_{z,\epsilon}(r) = \begin{cases} 1 & r \le z \\ 1 + (z - r)/\epsilon & z \le r \le z + \epsilon \\ 0 & r \ge z + \epsilon. \end{cases}$$

Observe that $\mathbb{I}\{r \le z\} \le t_{z,\epsilon}(r)$ for all $r \in \mathbb{R}$. Here $\mathbb{I}$ denotes the indicator function. This allow us to write

$$\widehat{F}^{\mathrm{gcv}}_{\lambda}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} \le z\right\} \le \frac{1}{n}\sum_{i=1}^{n} t_{z,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right). \tag{S.31}$$

Furthermore, $t_{r,\epsilon}$ is Lipschitz continuous and satisfies Assumption 3. Hence, invoking Theorem 3, we have that

$$\frac{1}{n}\sum_{i=1}^{n} t_{z,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) - \mathbb{E}\big[t_{z,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \xrightarrow{\text{a.s.}} 0. \tag{S.32}$$

In addition, observe that $t_{z,\epsilon}(r) \le \mathbb{I}\{r \le z + \epsilon\}$ for all $r \in \mathbb{R}$. This gives us

$$\mathbb{E}\big[t_{z,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \le \mathbb{E}\big[\mathbb{I}\{y_0 - x_0^\top \widehat{\beta}_\lambda \le z + \epsilon\} \mid X, y\big] = \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \le z + \epsilon \mid X, y\big]. \tag{S.33}$$

Thus, combining (S.31) to (S.33), we get that almost surely

$$\limsup_{n\to\infty} \widehat{F}^{\mathrm{gcv}}_{\lambda}(z) \le \limsup_{n\to\infty} \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \le z + \epsilon \mid X, y\big] = \limsup_{n\to\infty} F_\lambda(z + \epsilon). \tag{S.34}$$

Now sending $\epsilon \to 0$, we obtain the desired inequality $\limsup_{n\to\infty} \widehat{F}^{\mathrm{gcv}}_{\lambda}(z) \le F_\lambda(z)$ almost surely.

We proceed analogously on the other side. Again fix $\epsilon > 0$ and let $z \in \mathbb{R}$ be a continuity point of F_λ for n sufficiently large. We will now use the function $t_{z-\epsilon,\epsilon}$. Explicitly, the evaluation map of $t_{z-\epsilon,\epsilon}$ is given by

$$t_{z-\epsilon,\epsilon}(r) = \begin{cases} 1 & r \le z - \epsilon \\ (z - r)/\epsilon & z - \epsilon \le r \le z \\ 0 & r \ge z. \end{cases}$$

Noting that $t_{z-\epsilon,\epsilon}(r) \le \mathbb{I}\{r \le z\}$ for all $r \in \mathbb{R}$, we obtain

$$\widehat{F}^{\mathrm{gcv}}_{\lambda}(z) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{I}\left\{\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} \le z\right\} \ge \frac{1}{n}\sum_{i=1}^{n} t_{z-\epsilon,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right). \tag{S.35}$$

Again, since $t_{z-\epsilon,\epsilon}$ is Lipschitz continuous and satisfies Assumption 3, application of Theorem 3 yields

$$\frac{1}{n}\sum_{i=1}^{n} t_{z-\epsilon,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) - \mathbb{E}\big[t_{z-\epsilon,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \xrightarrow{\text{a.s.}} 0. \tag{S.36}$$

Finally, because $t_{z-\epsilon,\epsilon}(r) \ge \mathbb{I}\{r \le z - \epsilon\}$ for $r \in \mathbb{R}$, we have that

$$\mathbb{E}\big[t_{z-\epsilon,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \ge \mathbb{E}\big[\mathbb{I}\{y_0 - x_0^\top \widehat{\beta}_\lambda \le z - \epsilon\} \mid X, y\big] = \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \le z - \epsilon\big]. \tag{S.37}$$

Combining (S.35) to (S.37), we have almost surely,

$$\liminf_{n\to\infty} \widehat{F}^{\mathrm{gcv}}_{\lambda}(z) \ge \liminf_{n\to\infty} \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \le z - \epsilon\big] = \liminf_{n\to\infty} F_\lambda(z - \epsilon). \tag{S.38}$$

Since z is a continuity point of F_λ, sending $\varepsilon \to 0$, we get the desired inequality $\liminf_{n\to\infty} \widehat{F}^{\mathrm{gcv}}_{\lambda}(z) \ge F_\lambda(z)$ almost surely.

Combining (S.34) and (S.38), we conclude that almost surely $\limsup_{n\to\infty} \widehat{F}^{\mathrm{gcv}}_{\lambda}(z) - \liminf_{n\to\infty} \widehat{F}^{\mathrm{gcv}}_{\lambda}(z) \to 0$, and $\widehat{F}^{\mathrm{gcv}}_{\lambda}(z) - F(z) \to 0$, completing the proof.

S.5 PROOFS RELATED TO Theorem 6

S.5.1 Proof of Theorem 6

As hinted in the paper, the proof of Theorem 6 mainly builds on the result of Theorem 3. We will use Theorem 3 to certify pointwise convergence (in v) of $\widehat{T}_\lambda^{\mathrm{gcv}}(v)$ and $\widehat{T}_\lambda^{\mathrm{loo}}(v)$ to $T_\lambda(v)$. Then using the equicontinuity of $\mathcal{T}_\mathcal{V}$ and appealing to Lemma S.13, we will prove the convergence of the minimizers $\widehat{V}_\lambda^{\mathrm{gcv}}$ and V_λ^{loo} to V_λ.

First observe that each $t(\cdot, v) : \mathbb{R} \to \mathbb{R}$ is a continuos function since $\mathcal{T}_\mathcal{V}$ is an equicontinous family of functions. In addition, each $t(\cdot, v)$ satisfies Assumption 3. Thus, for each $v \in \mathcal{V}$, Theorem 3 implies

$$\widehat{T}_\lambda^{\mathrm{gcv}}(v) - T_\lambda(v) \xrightarrow{\text{a.s.}} 0.$$

Next note that for any $\delta > 0$,

$$\begin{aligned}
&\sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \big| T_\lambda(v_1) - T_\lambda(v_2) \big| \\
&= \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \Big| \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) \mid X, y \big] - \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \mid X, y \big] \Big| \\
&= \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \Big| \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \mid X, y \big] \Big| \\
&\le \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \mathbb{E}\Big[\big| t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \big| \;\Big|\; X, y \Big] \\
&\le \mathbb{E}\left[\sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \big| t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \big| \;\Big|\; X, y \right],
\end{aligned} \tag{S.39}$$

where the third line follows from Jensen's inequality, the last inequality follows because for any $v_1, v_2 \in \mathcal{V}$ such that $|v_1 - v_2| \le \delta$, we have that

$$\big| t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \big| \le \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \big| t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \big|,$$

which after taking expectation and taking sup gives the desired inequality. Similarly, for any $\delta > 0$,

$$\begin{aligned}
&\sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \big| \widehat{T}_\lambda^{\mathrm{gcv}}(v_1) - \widehat{T}_\lambda^{\mathrm{gcv}}(v_2) \big| \\
&= \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \left| \frac{1}{n} \sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1 \right) - \frac{1}{n} \sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2 \right) \right| \\
&\le \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^n \left| t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1 \right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2 \right) \right| \\
&\le \frac{1}{n} \sum_{i=1}^n \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \left| t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1 \right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2 \right) \right|.
\end{aligned} \tag{S.40}$$

Note that the exact argument holds for the case of $\lambda = 0$ by replacing replacing the first argument of t with the modified GCV errors. Since the family $\{t(\cdot, v) : v \in \mathcal{V}\}$ is pointwise equicontinous, (S.39) and (S.40) imply equicontinuity of $\{T_\lambda(v) : v \in \mathcal{V}\}$ and $\{\widehat{T}_\lambda^{\mathrm{gcv}}(v) : v \in \mathcal{V}\}$. Moreover, as $\mathcal{V}$ is compact and V_λ is assumed to be unique, Lemma S.13 yields

$$\widehat{V}_\lambda^{\mathrm{gcv}} - V_\lambda \xrightarrow{\text{a.s.}} 0.$$

Analogous argument shows the convergence for $\widehat{V}_\lambda^{\mathrm{loo}}$ by using the LOOCV part of Theorem 3.

S.5.2 Proof of Corollary 7

We verify that the conditions of Theorem 6 are satisfied. For $\tau \in (0, 1)$ and compact set $\mathcal{U} \subseteq \mathbb{R}$, the family of error functions under consideration is $\mathcal{T}_\mathcal{U} = \{t_\tau(\cdot, u) : u \in \mathcal{U}\}$, where each function $t_\tau(\cdot, u)$ is such that for $r \in \mathbb{R}$

$$t_\tau(r, u) = (r - u)(\tau - \mathbb{I}\{r - u < 0\}).$$

In other words, the evaluation map is given by

$$t_\tau(r,u) = \begin{cases} (r-u)\tau & \text{if } r \geq u \\ (u-r)(1-\tau) & \text{if } u > r. \end{cases}$$

A sufficient condition to establish equicontinuity of $\mathcal{T}_\mathcal{U}$ is to show that the functions in the family are Lipschitz continuous with uniformly bounded Lipschitz constant (see, e.g., Section 1.8 of Tao, 2010). It is easy to check that each function in the family $\mathcal{T}_\mathcal{U}$ is Lipschitz continuous with uniformly bounded constant $L = \max\{\tau, 1-\tau\}$. Thus, the family $\mathcal{T}_\mathcal{U}$ is equicontinous over compact set $\mathcal{U}$. Furthermore, since $\mathcal{U}$ is assumed to contain the true quantile, $Q_\lambda(\tau)$ is unique. Therefore, invoking Theorem 6 we obtain the desired conclusion.

S.6 ADDITIONAL NUMERICAL RESULTS

In this section, we provide additional numerical illustrations to complement those included in the main paper. The details of feature and response models used throughout different experiments are described next.

Feature model. The feature $x_i \in \mathbb{R}^p$ is generated according to

$$x_i = \Sigma^{1/2} z_i, \tag{S.41}$$

where $z_i \in \mathbb{R}^p$ contains independently sampled entires from a common distribution, and $\Sigma \in \mathbb{R}^{p\times p}$ is a positive semidefinite feature covariance matrix. The different distributions that we use for the components of z_i include: (1) Gaussian distribution, (2) Student's t-distribution, and (3) Bernoulli distribution. These represent a mix of both continuous and discrete, and light- and heavy-tailed distributions. We standardize the distributions so that the mean is zero and the variance is one. The different feature covariance matrix structures that we use include: (1) Identity ($\Sigma_{ij} = 1$ when $i = j$ and $\Sigma_{ij} = 0$ when $i \neq j$) and (2) Autoregressive with parameter ρ ($\Sigma_{ij} = \rho^{|i-j|}$ for all i, j).

Response model. Given x_i, the response $y_i \in \mathbb{R}$ is generated according to

$$y_i = \beta_0^\top x_i + \big(x_i^\top A x_i - \mathrm{tr}[A\Sigma]\big)/p + \varepsilon_i, \tag{S.42}$$

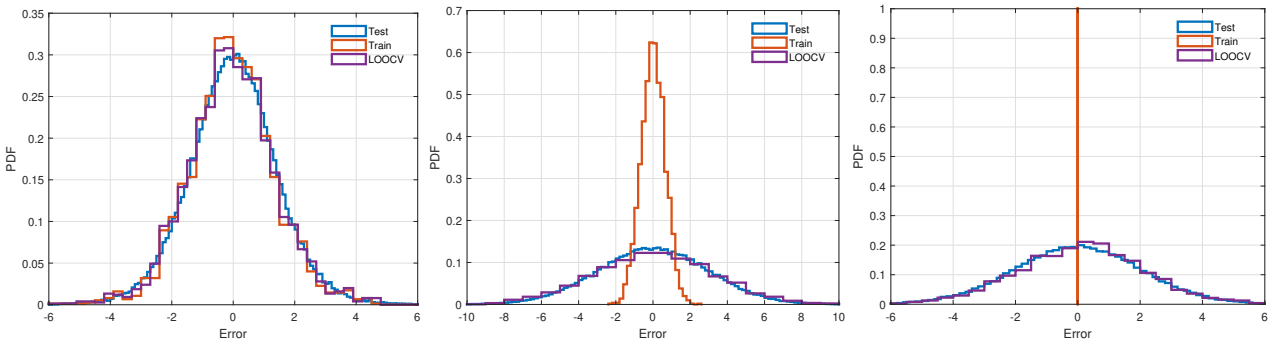
where $\beta_0 \in \mathbb{R}^p$ is a fixed signal vector, $A \in \mathbb{R}^{p\times p}$ is a fixed matrix, and $\varepsilon_i \in \mathbb{R}$ is a random noise variable. Note that we have subtracted the mean from the squared nonlinear component and scaled it to keep the variance of the nonlinear component at the same order as the noise variance (see Mei and Montanari (2019) for more details, for example). We again use either Gaussian, Student's t, or Bernoulli distribution for the random noise component, which is again standardized so that the mean is zero and the variance is one. We refer to the value of $\beta_0^\top \Sigma \beta_0$ as the effective signal energy.

Train and test set sizes. In all of our experiments, the sample size for the train set is fixed at $n = 2500$. To compute various out-of-sample quantities, we use a test set of 100000 indepedent observations. We use three feature sizes of $p = 100$, $p = 2000$, and $p = 5000$ that represent low, moderate, and high-dimensional settings (with aspect ratios p/n of 0.04, 0.8, and 2), respectively.

S.6.1 Distribution Estimation

As promised in the paper, we first present illustrations with LOOCV reweighted errors for Figures 1 and 2 in Figures S.1 and S.2, respectively.

Note that both in Figures 1 and 2 in the paper, as well as Figures S.1 and S.2, the out-of-sample error distributions and the associated GCV and LOOCV reweighted error distributions are all symmetric distributions. This need not be the case. In Figure S.3, we consider a case in which the out-of-sample error distribution and the estimated distributions based on GCV and LOOCV reweighted errors are negatively skewed.



(a) Low dimension ($p/n = 0.04$) (b) Moderate dimension ($p/n = 0.8$) (c) High dimension ($p/n = 2$)

Figure S.1: A simulation with $n = 2500$ and $p \in \{100, 2000, 5000\}$ features with a different p per panel above. In each setting, the feature vectors x_i are generated as in (S.41) with identity covariance with components of z_i sampled from a t-distribution with 5 degrees of freedom, and the responses y_i are generated as in (S.42). We fit the min-norm least squares solution, as in (1) with $\lambda = 0$. The blue curve in each panel is a histogram of the true prediction error distribution, computed from 10^5 independent test samples. The red curve is a histogram of the training errors; when $p > n$, this is just a point mass at zero. The purple curve is a histogram of LOOCV reweighted training errors, as in (12) (when $p < n$ in the first two panels) and (14) (when $p > n$ in the last panel). This tracks the blue curve very well in all three settings again. Empirical results for GCV are provided in Figure 1 of the paper.

S.6.2 Quantile Estimation

We first provide further details on the setup used in Figure 3 of the main paper. We use a special "latent" space data model, in which the true signal component lies in a small eigenspace of the feature covariance matrix. Such setup was investigated in the context of ridge regression by Kobak et al. (2020); Wu and Xu (2020); Richards et al. (2020); Hastie et al. (2019), who study the optimality of zero (or even negative) ridge regularization for expected squared out-of-sample error under special cases. We verify empirically that such behavior continues to hold even for general functionals of the out-of-sample error distribution and their plug-in estimators based on GCV and LOOCV such as the length of prediction intervals, and even under nonlinear model.

For numerical illustration, we consider an extreme case where the signal vector is aligned with the eigenvector of the covariance matrix corresponding to the largest eigenvalue. More precisely, let $\Sigma = WRW^\top$ denote the eigenvalue decomposition of the covariance matrix Σ, where $W \in \mathbb{R}^{p\times p}$ is a orthogonal matrix whose columns $w_1, \ldots, w_p$ are eigenvectors of Σ and $R \in \mathbb{R}^{p\times p}$ is a diagonal matrix whose entries $r_1 \geq \cdots \geq r_p$ are eigenvalues of Σ in descending order. We then let $\beta_0 = \zeta w_1$, where ζ controls the effective signal energy. Figure S.4 illustrate the coverage and length of prediction intervals (30) computed using the LOOCV reweighted error distribution.

Finally, as a contrast we consider a "regular" setting in Figure S.5 where the signal does not have any special structure, and the signal covariance is identity, where we see that regularization does in fact help indicating the subtle interplay between the signal vector and feature covariance that causes the near optimality of ridgeless estimator for various functionals of the out-of-sample error distribution.

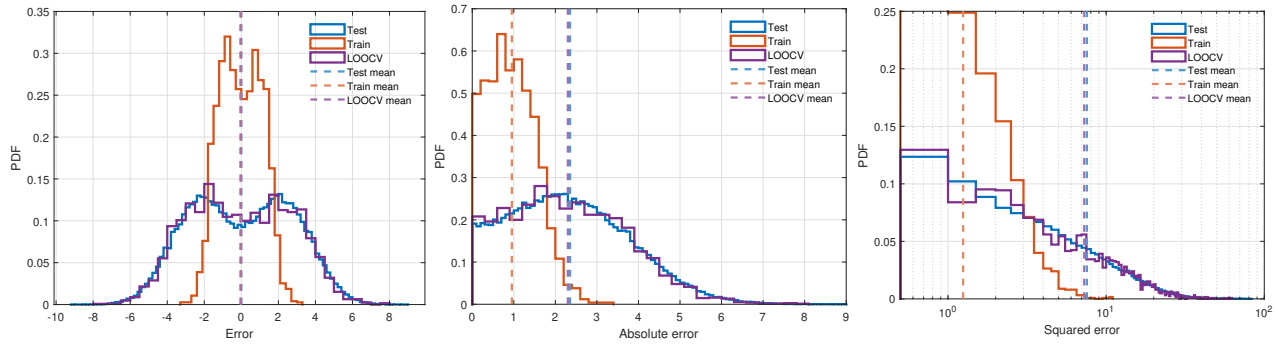
S.7 SUPPLEMENTARY RESULTS

In this section, we record statements of various results adapted from other sources that are used in the proofs throughout the supplement.

The following inequality bounding q-th moment of sum of random variables is by Burkholder (1973). See also Bai and Silverstein (2010, Lemma 2.13).

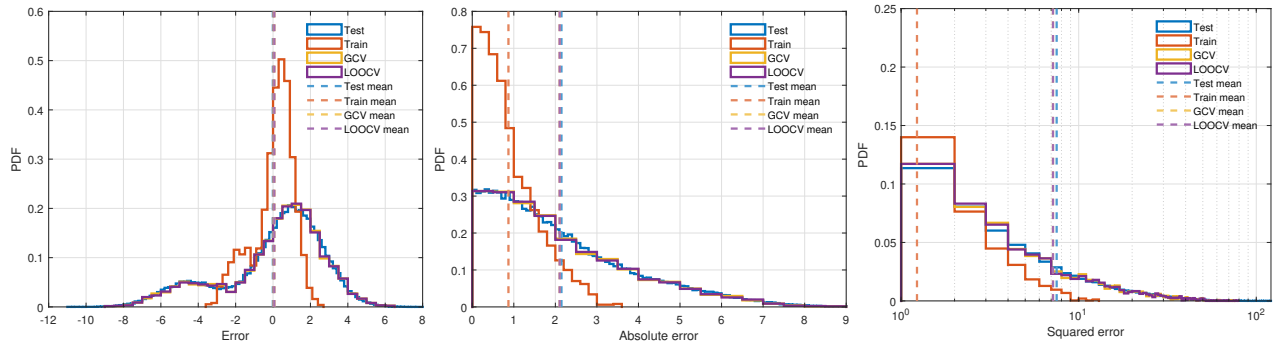
Lemma S.8 (Burkholder's inequality)**.** *Let $\{Z_k\}$ be a martingale difference sequence with respect to the increasing σ-field $\{\mathcal{F}_k\}$. Then, for $q \geq 2$,*

$$\mathbb{E}\Big[\Big|\sum_k Z_k\Big|^q\Big] \leq C_q \left\{\mathbb{E}\left[\left(\sum_k \mathbb{E}\Big[|Z_k|^2 \mid \mathcal{F}_{k-1}\Big]\right)^{q/2}\right] + \mathbb{E}\Big[\sum_k |Z_k|^q\Big]\right\}$$



(a) $h(e) = e$ (b) $h(e) = |e|$ (c) $h(e) = e^2$

Figure S.2: An example with $n = 2500$, $p = 5000$. We generated each x_i according to (S.41) with identity covariance with the components of z_i sampled from a symmetric Bernoulli distribution, and each response y_i is generated according to (S.42). The ridge parameter was fixed at $\lambda = 1$. Each panel above examines weak convergence per (17) for a different function h of the error variable (identity, absolute value, and square, from left to right). In each case, the LOOCV estimate (purple) tracks the true distribution (blue) closely. Empirical results for GCV are in Figure 2 of the paper.



(a) $h(e) = e$ (b) $h(e) = |e|$ (c) $h(e) = e^2$

Figure S.3: An example with $n = 2500$, $p = 5000$. We generated each x_i according to (S.41) with identity covariance and components of z_i sampled from a Gaussian distribution, and each response y_i according to (S.42) with noise variable ε_i distributed according to a Bernoulli random variable with success probability 0.8. The ridge parameter was fixed at $\lambda = 1$. Each panel above examines weak convergence per (17) for a different function h of the error variable (identity, absolute value, and square, from left to right). In each case, the GCV estimate (yellow) and LOOCV estimate (purple) track the true distribution (blue) closely.

for a constant C_q that only depends on q.

The following inequality bounding L_p norm of an inner product is from Erdos and Yau (2017, Lemma 7.8).

Lemma S.9 (L_q norm of an inner product)**.** *Let $u \in \mathbb{R}^p$ be a random vector consisting of independent entries u_i with $\mathbb{E}[u_i] = 0$, $\mathbb{E}[u_i^2] = 1$, and $\|u_i\|_{L_q} \le K_q$ for $i = 1, \dots, p$. Let $a \in \mathbb{R}^p$ be a deterministic vector. Then,*

$$\|a^\top u\|_{L_q} \le C_q K_q \|a\|_2$$

for a constant C_q depending only on q.

The following lemma bounding q-th moment of a quadratic form is from Bai and Silverstein (2010, Lemma B.26). See also Dobriban and Wager (2018, Lemma 7.10).

Lemma S.10 (Centered moment a quadratic form)**.** *Let $W \in \mathbb{R}^{p \times p}$ be a deterministic matrix. Let $v \in \mathbb{R}^p$ be a random vector of independent entries v_i for $i = 1, \dots, p$ with each $\mathbb{E}[v_i] = 0$, $\mathbb{E}[v_i^2] = 1$, and $\mathbb{E}[|v_i|^r] \le M_r$. Then, for any $q \ge 1$,*

$$\mathbb{E}\Big[\big|v^\top W v - \operatorname{tr}[W]\big|^q\Big] \le C_q \Big\{ \big(M_4 \operatorname{tr}[WW^\top]\big)^{q/2} + M_{2q} \operatorname{tr}\big[(WW^\top)^{q/2}\big] \Big\}$$

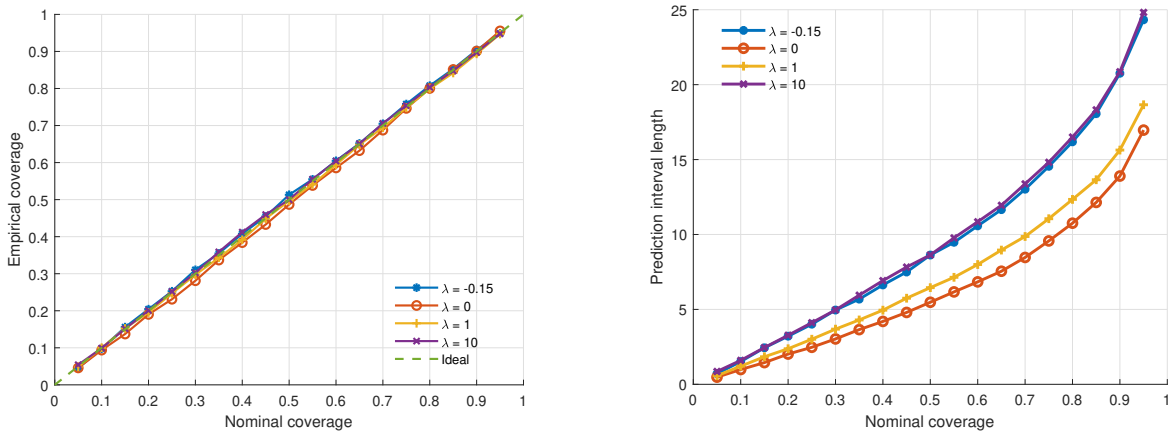


Figure S.4: Illustration of empirical coverage and length of LOOCV prediction intervals constructed using (30) against nominal coverage, where $n = 2500$, $p = 5000$. We generated features x_i according to (S.41) with autoregressive covariance structure (with $\rho = 0.25$) and t-distributed components of z_i with 5 degrees of freedom. The responses y_i are generated according to (S.42) where the signal β_0 is aligned with the top eigenvector of the covariance matrix and the effective signal energy is 50. We see that intervals for any λ have excellent finite-sample coverage (left), and the case of $\lambda = 0$ provides the smallest interval lengths (right). Empirical results for GCV prediction intervals are in Figure S.4 of the paper.

for a constant C_q that only depends on q.

The following equivalence lemma for the denominator arising from GCV is adapted from Patil et al. (2021, Lemma S.3.1).

Lemma S.11 (GCV denominator lemma)**.** *Suppose Assumption 1 holds. Then, for $\lambda \in (\lambda_{\min}, \infty) \setminus \{0\}$*

$$1 + \operatorname{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n - \frac{1}{1 - \operatorname{tr}\left[(X^\top X/n + \lambda I)^\dagger X^\top X/n\right]/n} \xrightarrow{\text{a.s.}} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$, and for the case of $\lambda = 0$,

$$\operatorname{tr}\left[\left(I - (X^\top X/n)^\dagger X^\top X/n\right)\Sigma\right]/n - \frac{1}{\operatorname{tr}\left[(X^\top X/n)^\dagger\right]/n} \xrightarrow{\text{a.s.}} 0,$$

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

The following results are standard results on stochastic uniform convergence. See, e.g., Chapter 21 of Davidson (1994).

Lemma S.12 (Stochastic uniform convergence)**.** *Let $f_n(\theta)$, $\theta \in \Theta$ be a family of stochastic functions. Suppose Θ is a compact, and for every $\theta \in \Theta$, $f_n(\theta) \xrightarrow{\text{a.s.}} f(\theta)$. Further, assume that $\{f_n(\theta)\}$ is strongly stochastic equicontinous. Then, as $n \to \infty$,*

$$\sup_{\theta \in \Theta} |f_n(\theta) - f(\theta)| \xrightarrow{\text{a.s.}} 0.$$

A corollary of Lemma S.12 is the following statement.

Lemma S.13 (Convergence of minimizers)**.** *Assume the setting of Lemma S.12. Let $\widehat{\xi}_n$ and ξ be minimizers of f_n and f over $\theta \in \Theta$, respectively. Moreover, assume that f has a unique minimizer over Θ. Then, as $n \to \infty$,*

$$\widehat{\xi} \xrightarrow{\text{a.s.}} \xi.$$

The following lemma is a simple application of Markov's inequality along with the Borel-Cantelli lemma.

Lemma S.14 (Moment version of the Borel-Cantelli lemma)**.** *Let $\{S_n\}$ be a sequence of random variables. Suppose $\left\{\mathbb{E}[|S_n|^p]\right\}$ forms a summable sequence for some $p > 0$. Then, as $n \to \infty$, $S_n \xrightarrow{\text{a.s.}} 0$.*

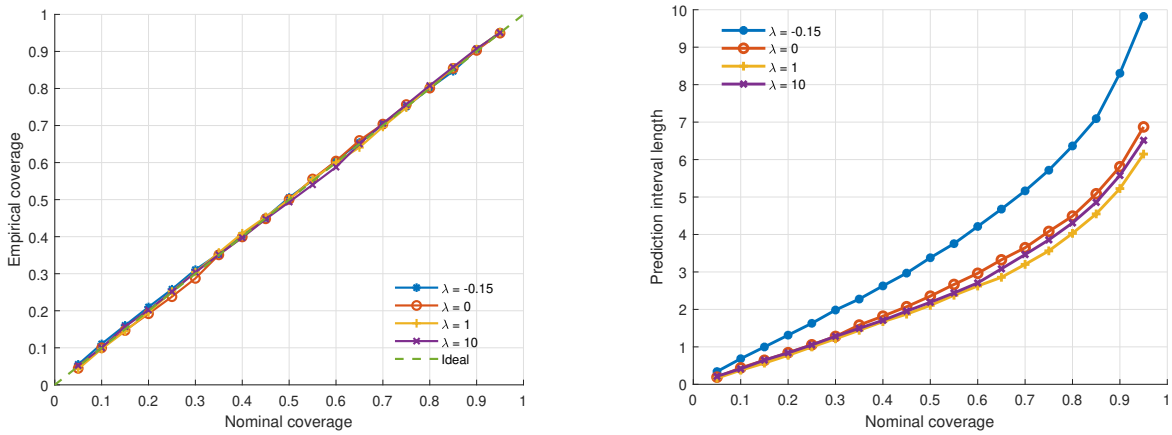


Figure S.5: Illustration of empirical coverage and length of LOOCV prediction intervals (30) against nominal coverage, where $n = 2500$, $p = 5000$. The features x_i are generated according to (S.41) with identity covariance and components of z_i having Gaussian distribution. The responses y_i are generated according to (S.42) with the nonlinearity component set to 0 (thus a well-specified linear model) and a random signal vector. We see again that the intervals for any λ have excellent finite-sample coverage (left) and now the case of $\lambda = 1$ provides the smallest interval lengths (right). Similar trend holds for GCV prediction intervals, and hence we do not present the corresponding figure for GCV.

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.

Donald L. Burkholder. Distribution function inequalities for martingales. *The Annals of Probability*, 1(1):19–42, 1973.

James Davidson. *Stochastic limit theory: An introduction for econometricians*. OUP Oxford, 1994.

Ronald A. DeVore and George G. Lorentz. *Constructive approximation*, volume 303. Springer Science & Business Media, 1993.

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

László Erdos and Horng-Tzer Yau. A dynamical approach to random matrix theory. *Courant Lecture Notes in Mathematics*, 28, 2017.

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

Dmitry Kobak, Jonathan Lomond, and Benoit Sanchez. The optimal ridge penalty for real-world high-dimensional data can be zero or negative due to the implicit ridge regularization. *Journal of Machine Learning Research*, 21 (169):1–16, 2020.

Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and the double descent curve. *Communications on Pure and Applied Mathematics*, 2019.

Pratik Patil, Yuting Wei, Alessandro Rinaldo, and Ryan J. Tibshirani. Uniform consistency of cross-validation estimators for high-dimensional ridge regression. In *International Conference on Artificial Intelligence and Statistics*, pages 3178–3186. PMLR, 2021.

Kaare Brandt Petersen, Michael Syskind Pedersen, et al. The matrix cookbook. *Technical University of Denmark*, 7(15):510, 2008.

Dominic Richards, Jaouad Mourtada, and Lorenzo Rosasco. Asymptotics of ridge(less) regression under general source condition. *arXiv preprint arXiv:2006.06386*, 2020.

Terence Tao. *An epsilon of room, I: real analysis*, volume 1. American Mathematical Soc., 2010.

Denny Wu and Ji Xu. On the optimal weighted ℓ_2 regularization in overparameterized linear regression. *arXiv preprint arXiv:2006.05800*, 2020.