Estimating Functionals of the Out-of-Sample Error Distribution in High-Dimensional Ridge Regression

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

Carnegie Mellon University

Abstract

We study the problem of estimating the distribution of the out-of-sample prediction error associated with ridge regression. In contrast, the traditional object of study is the uncentered second moment of this distribution (the mean squared prediction error), which can be estimated using cross-validation methods. We show that both generalized and leave-one-out cross-validation (GCV and LOOCV) for ridge regression can be suitably extended to estimate the full error distribution. This is still possible in a high-dimensional setting where the ridge regularization parameter is zero. In an asymptotic framework in which the feature dimension and sample size grow proportionally, we prove that almost surely, with respect to the training data, our estimators (extensions of GCV and LOOCV) converge weakly to the true out-of-sample error distribution. This result requires mild assumptions on the response and feature distributions. We also establish a more general result that allows us to estimate certain functionals of the error distribution, both linear and nonlinear. This yields various applications, including consistent estimation of the quantiles of the out-of-sample error distribution, which gives rise to prediction intervals with asymptotically exact coverage conditional on the training data.

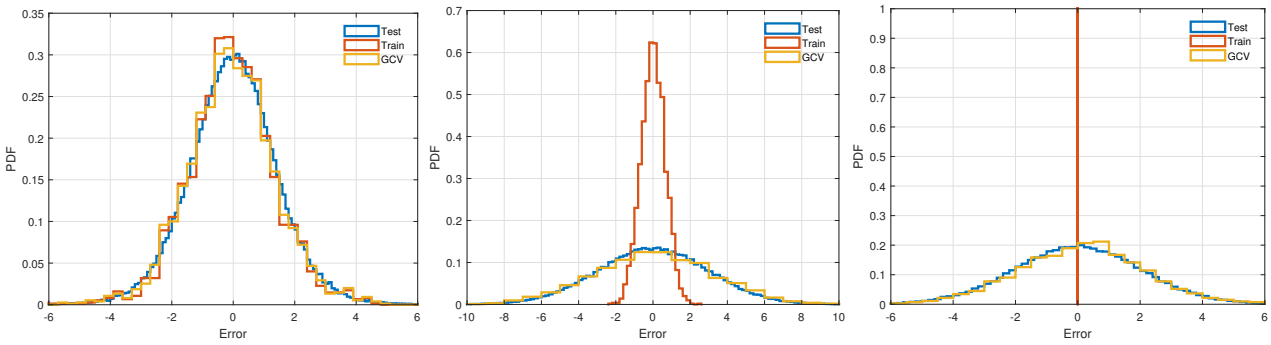
1 INTRODUCTION

The out-of-sample error associated with a predictive model is the difference between the true (unobserved) response and the predicted response at a new draw from the feature distribution. Being able to accurately estimate functionals of the out-of-sample error distribution is of critical importance in practice, both for model assessment and model selection purposes. By far the most common functional considered is the uncentered second moment of this error distribution—the mean squared error of the predictive model. Estimating this quantity has been the focus of many decades of research in the statistics and machine learning communities, which has yielded numerous advances in both theory and methodology. A central method in practice for estimating the mean squared prediction error is cross-validation (CV), which comes in many variants, including *generalized* and *leave-one-out* cross-validation (GCV and LOOCV, respectively). Classic references on CV include Allen (1974); Stone (1974, 1977); Geisser (1975); Golub et al. (1979); Wahba (1980, 1990); Li (1985, 1986, 1987). See Arlot and Celisse (2010) for a general review of CV.

In this paper, we study the problem of estimating the entire out-of-sample error distribution. Part of reason why so much past work in risk estimation has focused on mean squared out-of-sample error is undoubtedly the special analytical structure that it affords and the associated bias-variance decomposition. A main goal of this paper is to understand what other functionals of the out-of-sample error distribution can be reliably estimated using cross-validation. Such an understanding is useful for not only theoretical purposes (necessitating novel proof techniques to analyze generic functionals), but practical ones as well, since cross-validation estimators that work under such general settings then open up the possibility of employing a wider range of metrics for model evaluation and selection, which may be informative for the data analyst in any given problem setting at hand.

Throughout, we will focus on *ridge regression* (Hoerl and Kennard, 1970a,b) for the predictive model, a special form of Tikhonov regularization (Tikhonov, 1943, 1963), which is very widely used in statistics and machine learning. We choose to focus on ridge regression because GCV and LOOCV admit special forms for this

Proceedings of the 25[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2022, Valencia, Spain. PMLR: Volume 151.



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

Figure 1: A simulation with $n = 2500$ samples and $p \in \{100, 2000, 5000\}$ features (a different p per panel above). In each setting, we generated the feature vectors x_i to have independent components from a t-distribution with 5 degrees of freedom, and generated the responses y_i by adding t-distributed noise with 5 degrees of freedom to a nonlinear (quadratic) function of x_i. We then fit the minimum ℓ_2 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 yellow curve is a histogram of GCV-reweighted training errors, as in (11) (for $p < n$, in the first two panels) and (13) (for $p > n$, in the last panel). This tracks the blue curve very well in all settings. Empirical results for LOOCV are given in the supplement.

estimator, and also because ridge has recently attracted much attention—especially in the limiting case of zero regularization, often called the "ridgeless" limit—due to its somewhat exotic behavior in the overparametrized regime (see, e.g., Bartlett et al., 2020; Belkin et al., 2020; Hastie et al., 2019; Muthukumar et al., 2020, and references therein). Importantly, it has been recently shown that the ridgeless (minimum ℓ_2 norm) interpolator can be optimal for mean squared out-of-sample error, among all ridge models, for well-specified linear models with certain data geometries and high signal-to-noise ratios (Wu and Xu, 2020; Richards et al., 2020). This has been corroborated empirically using real data sets for ridge regression (Kobak et al., 2020) and kernel ridge regression (Liang and Rakhlin, 2020). Thus, providing theory that covers that ridgeless case is both of foundational and practical importance.

Before summarizing our main contributions, we give some empirical examples in Figure 1 to motivate our study.

1.1 Summary of Contributions

An overview of our main contributions is as follows.

- We define natural extensions of GCV and LOOCV in order to estimate the out-of-sample prediction error distribution associated with ridge regression. These are empirical distributions over reweighted training errors (where the reweighting is tied to GCV or LOOCV).
- Under an asymptotic framework where the feature dimension p and sample size n grow proportionally, $p/n \to \gamma \in (0, \infty)$, we prove that, almost surely with respect to the training data, these extensions of GCV and LOOCV converge weakly to the true out-of-sample error distribution of ridge regression. This result requires mild assumptions; we do not need the true regression model to be linear.
- The GCV and LOOCV extensions and the theory we prove about them all accommodate the choice of zero (or even negative) ridge regularization in high dimensions, where $p > n$.
- For certain linear functionals of the error distribution P, which take the form $\int t\, dP$ for a function t, we prove that suitable plug-in estimators (based on the GCV and LOOCV estimators of the entire error distribution) are asymptotically consistent, almost surely. This result requires t to satisfy certain continuity and growth conditions, but it can be unbounded.
- Finally, we use a uniform convergence argument to handle certain nonlinear functionals of the error distribution (that can be written in a variational form involving linear functionals). This allows us to consistently estimate, as an application, quantiles of the ridge error distribution.

1.2 Related Work

Among the different CV variants to assess prediction accuracy, k-fold CV is widely used in practice (Györfi

et al., 2006; Hastie et al., 2009). However, in a high-dimensional regime where the feature dimension p is comparable to the sample size n, small values of k (such as $k = 5$ or 10) lead to bias in error estimation (see, e.g., Rad and Maleki, 2020). LOOCV (where $k = n$) mitigates these bias issues, and consequently LOOCV and various approximations to it (that circumvent its computational burden) have been of interest in recent work, including Meijer and Goeman (2013); Liu et al. (2014); Obuchi and Kabashima (2016); Beirami et al. (2017); Wang et al. (2018); Stephenson and Broderick (2020); Giordano et al. (2019); Wilson et al. (2020); Rad et al. (2020); Xu et al. (2021). For recent results on ridge regression in particular, where LOOCV can be done efficiently via a "shortcut" formula, see Patil et al. (2021).

On the inferential side, Bayle et al. (2020) prove central limit theorems for CV error and a derive a consistent estimator of its asymptotic variance under certain stability assumptions, similar to Kale et al. (2011); Kumar et al. (2013); Celisse and Guedj (2016). Their results yield asymptotic confidence intervals for the prediction error and apply to k-fold CV (for a fixed k) as well as LOOCV. See also Austern and Zhou (2020) for similar guarantees. A prominent and distinctive aspect of our work compared to these papers and others is the focus on properties of the entire empirical distribution of the CV errors, rather than specific functionals such as the mean squared CV error.

In a contribution that is quite relevant to this paper, Steinberger and Leeb (2016, 2018) construct prediction intervals from quantiles of the empirical distribution of the LOOCV errors and provide conditional coverage guarantees, which hold in expectation. Their key assumptions are algorithmic stability, as in Bousquet and Elisseeff (2002), along with a bound in probability on the prediction error at a new test point. Under a more restrictive asymptotic regime in which $p/n \to \gamma < 1$, they show that the Kolmogorov-Smirnov distance between the empirical distribution of LOOCV errors and the conditional prediction error distribution vanishes in expectation. This general result is then applied to yield corollaries for various predictive models, including ridge regression, by leveraging model-specific stability and error results from the literature.

In comparison, our paper focuses on ridge regression alone, but we deliver stronger and broader guarantees. To be specific, our results (1) accommodate the high-dimensional regime, $p/n \to \gamma \geq 1$; (2) assume quite weak conditions on the data (e.g., we do not require a well-specified linear model); (3) hold uniformly over the choice of regularization parameter (which includes no regularization—the ridgeless limit); (4) yield not only consistent estimation of the prediction error distribution itself, but of a broad class of functionals of this distribution (which includes unbounded and nonlinear ones); and (5) produces guarantees that hold almost surely—rather than in expectation or in probability—with respect to the training data.

2 PRELIMINARIES

We adopt a standard regression setting, with i.i.d. samples (x_i, y_i), for $i = 1, \ldots, n$, where each $x_i \in \mathbb{R}^p$ is a feature vector and $y_i \in \mathbb{R}$ is its corresponding response value. We will denote by $X \in \mathbb{R}^{n \times p}$ the feature matrix whose i^{th} row is $x_i^\top$, and by $y \in \mathbb{R}^n$ the response vector whose i^{th} entry is y_i.

2.1 Ridge Regression

The *ridge regression* estimator $\widehat{\beta}_\lambda \in \mathbb{R}^p$, based on X, y, is defined as the solution to the following problem:

$$\underset{\beta \in \mathbb{R}^p}{\text{minimize}} \ \frac{1}{n}\|y - X\beta\|_2^2 + \lambda\|\beta\|_2^2.$$

Here λ is a regularization parameter. When $\lambda > 0$, the above optimization problem is strictly convex and has a unique solution:

$$\widehat{\beta}_\lambda = (X^\top X/n + \lambda I_p)^{-1} X^\top y/n.$$

When $\lambda = 0$, and $X^\top X$ is rank deficient (which will always be the case when $p > n$), there will be infinitely many solutions, and we focus on the solution with the minimum ℓ_2 norm, which we refer to as the *min-norm solution* for short. By defining the ridge estimator as

$$\widehat{\beta}_\lambda = (X^\top X/n + \lambda I_p)^\dagger X^\top y/n, \tag{1}$$

where $A^\dagger$ denotes the Moore-Penrose pseudoinverse of a matrix A, we simultaneously accommodate the case of $\lambda > 0$, in which case (1) reduces to the second to last display, and the case of $\lambda = 0$, in which case (1) becomes the min-norm solution (it lies in the column space of $(X^\top X)^\dagger$, i.e., the row space of X, so it has the minimum ℓ_2 norm among all least squares solutions). In fact, the above display even accommodates the case of $\lambda < 0$, in which case (1) remains well-defined.

The case of zero regularization is of particular interest when $\text{rank}(X) = n$, because then any least squares solution interpolates the training data, and the min-norm solution $\widehat{\beta}_0$ (by construction) has the minimum ℓ_2 norm among all such interpolators.

2.2 Out-of-Sample Error

Let (x_0, y_0) denote a test point drawn independently from the same distribution as the training data (x_i, y_i),

$i = 1, \ldots, n$, and denote the out-of-sample prediction error of ridge regression at tuning parameter λ by

$$e_\lambda = y_0 - x_0^\top \widehat{\beta}_\lambda. \tag{2}$$

This is a scalar random variable, and we denote by P_λ its distribution conditional the training data:[1]

$$P_\lambda = \mathcal{L}\big(e_\lambda \mid X, y\big). \tag{3}$$

We are interested in estimating P_λ using the training data. A naive estimator would be to use the empirical distribution over the training errors expressed as

$$\widehat{P}_\lambda = \frac{1}{n} \sum_{i=1}^n \delta\big(y_i - x_i^\top \widehat{\beta}_\lambda\big). \tag{4}$$

Here we use $\delta(z)$ for a point mass at z. Of course, this can be very inaccurate in high dimensions (as we saw in Figure 1); at the extreme case of $\mathrm{rank}(X) = n$ and $\lambda = 0$, the naive estimator $\widehat{P}_\lambda$ trivially places all mass at zero. In the next subsection, we will introduce more sensible estimators based on cross-validation.

Aside from estimating P_λ itself, we may be interested in estimating a particular *functional* of P_λ, denoted by $\psi(P_\lambda)$. Recall, a functional ψ acting on distributions is such that $P \mapsto \psi(P) \in \mathbb{R}$ for all distributions P.

In the context of the out-of-sample error distribution P_λ, the most common functional of interest is its uncentered second moment,

$$\psi(P_\lambda) = \int z^2 \, dP_\lambda(z) = \mathbb{E}\big[e_\lambda^2 \mid X, y\big],$$

which is simply the mean squared prediction error. We will consider general linear functionals of the form

$$\psi(P_\lambda) = \int t(z) \, dP_\lambda(z) = \mathbb{E}\big[t(e_\lambda) \mid X, y\big], \tag{5}$$

for functions t (possibly nonlinear and unbounded, but subject to certain continuity and growth conditions). We will also consider certain nonlinear functionals such as the level-τ quantile, for $\tau \in (0, 1)$:

$$\psi(P_\lambda) = \mathrm{Quantile}(P_\lambda; \tau) = \inf\{z : F_\lambda(z) \geq \tau\}, \tag{6}$$

where F_λ denotes the cumulative distribution function (CDF) of P_λ.

2.3 Cross-Validation

GCV and LOOCV are two popular versions of cross-validation that are used to estimate the mean squared prediction error. GCV is traditionally defined for linear smoothers only, but LOOCV is fully general: it applies to any predictive model. In order to describe the details for ridge regression, we introduce the notation:

$$L_\lambda = X(X^\top X/n + \lambda I_p)^\dagger X^\top/n, \tag{7}$$

for the ridge smoother matrix at regularization level λ. Thus, by definition, we can express the fitted values (predicted values at the training points x_i, $i = 1, \ldots, n$) from ridge regression as $X\widehat{\beta}_\lambda = L_\lambda \, y$.

The LOOCV estimate for the mean squared prediction error of a given ridge model $\widehat{\beta}_\lambda$ can now be written as

$$\frac{1}{n} \sum_{i=1}^n \Big(y_i - x_i^T \widehat{\beta}_{-i,\lambda}\Big)^2 = \frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - x_i^T \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)^2, \tag{8}$$

where $\widehat{\beta}_{-i,\lambda}$ denotes the ridge estimate when the i^{th} pair (x_i, y_i) is excluded from the training data set, and $[L_\lambda]_{ii}$ denotes the i^{th} diagonal element of L_λ. The left-hand side in (8) is the usual definition of LOOCV for any predictive model; the right-hand side is a so-called "shortcut" formula that holds for ridge (and a handful of other special linear smoothers; see, e.g., Chapter 7 of Hastie et al., 2009).

The GCV estimate for the mean squared error of $\widehat{\beta}_\lambda$ is given by

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{y_i - x_i^T \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right)^2, \tag{9}$$

where $\mathrm{tr}[A]$ denotes the trace of a matrix A.

Caution needs to be taken in (8) and (9) when $\lambda = 0$ and $\mathrm{rank}(X) = n$, in which case $L_\lambda = I_n$, and both of the numerators and denominators in every summand of (8), (9) are zero. To avoid this problem we redefine them by their respective limits as $\lambda \to 0$, which gives (see the supplement for details):

$$\frac{1}{n} \sum_{i=1}^n \left(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\right)^2 \text{ and } \frac{1}{n} \sum_{i=1}^n \left(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\right)^2, \tag{10}$$

for LOOCV and GCV, respectively.

2.4 Proposed Estimators

We propose estimators for the out-of-sample prediction error distribution P_λ in (3), building off the empirical distributions of reweighted training errors, inspired by GCV in (9) and LOOCV in (8). Precisely, we define

$$\widehat{P}_\lambda^{\text{gcv}} = \frac{1}{n} \sum_{i=1}^n \delta\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right), \tag{11}$$

[1] To be clear, P_λ is itself a random quantity, because it depends on the training data X, y. However, we suppress this dependence notationally, for simplicity.

which we refer to as the GCV estimate of the out-of-sample error distribution, and

$$\widehat{P}_\lambda^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \delta\bigg(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\bigg), \tag{12}$$

which we refer to as the LOOCV estimate of the out-of-sample error distribution.

When $\lambda = 0$ and $\mathrm{rank}(X) = n$, the above expressions are ill-defined, and we redefine them based on the forms of GCV and LOOCV in (10):

$$\widehat{P}_0^{\mathrm{gcv}} = \frac{1}{n}\sum_{i=1}^{n} \delta\bigg(\frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}\bigg), \tag{13}$$

$$\widehat{P}_0^{\mathrm{loo}} = \frac{1}{n}\sum_{i=1}^{n} \delta\bigg(\frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}}\bigg). \tag{14}$$

To estimate a generic functional of $\psi(P_\lambda)$ of the error distribution, we simply use

$$\widehat{\psi}_\lambda^{\mathrm{gcv}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}) \quad \text{and} \quad \widehat{\psi}_\lambda^{\mathrm{loo}} = \psi(\widehat{P}_\lambda^{\mathrm{gcv}}). \tag{15}$$

For $\psi(P_\lambda) = \int z^2\, dP_\lambda(z)$, the plug-in estimates above reduce to the standard GCV and LOOCV estimates of the mean squared prediction error.

3 DISTRIBUTION ESTIMATION

We first cover distributional convergence results. We impose the following mild structural and moment assumptions on the feature and response distributions.

Assumption 1 (Feature distribution)**.** Each feature vector can be decomposed as $x_i = \Sigma^{1/2} z_i$, for a deterministic symmetric matrix $\Sigma \in \mathbb{R}^{p\times p}$ whose maximum eigenvalue is bounded above by $r_{\max} < \infty$, and minimum eigenvalue is bounded below by $r_{\min} > 0$, where $r_{\max}$ and $r_{\min}$ are constants, and for a random vector $z_i \in \mathbb{R}^p$ whose entries are i.i.d. with mean zero, unit variance, and $\mathbb{E}[|z_{ij}|^{4+\mu}] \le M_z < \infty$, where $\mu > 0$ and M_z are constants.

The maximum eigenvalue bound for the feature covariance matrix Σ is used to control the magnitude of ridge predictions; the minimum eigenvalue bound is used in the analysis of the min-norm interpolator. Both of these can be relaxed further for some of our results, but we do not pursue such refinements here.

Assumption 2 (Response distribution)**.** Each y_i has mean zero and satisfies $\mathbb{E}[|y_i|^{4+\nu}] \le M_y < \infty$, where $\nu > 0$ and M_y are constants.

The condition that each y_i is centered is only used for simplicity. When y_i does not have mean zero, we would simply include an intercept in the model defined in (1), and all of our results would translate accordingly.

We work in an asymptotic regime where the number the samples n and the number of features p both diverge to ∞, and yet their ratio p/n converges to $\gamma \in (0, \infty)$. Such asymptotic regime has received considerable attention recently in high-dimensional statistics and machine learning theory, which is commonly referred to as proportional asymptotics. The range of regularization parameter values λ over which our results will hold is a function of γ and $r_{\min}$. In preparation for the coming theorem statements, we define $\lambda_{\min} = -(1 - \sqrt{\gamma})^2 r_{\min}$.

We are now ready to state the result concerning weak convergence of the empirical distributions (11)–(14) to the true out-of-sample error distribution (3).

Theorem 1 (Distribution estimation)**.** *Suppose Assumptions 1 and 2 hold. Then, for $\lambda > \lambda_{\min}$,*

$$\widehat{P}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} P_\lambda \quad \text{and} \quad \widehat{P}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} P_\lambda, \tag{16}$$

almost surely (which means, here and henceforth, almost surely with respect to the distribution of X, y), as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

In (16), note the left- and right-hand sides both depend on n, p. To explain what we mean by convergence in distribution here: if $\widehat{P}_n$ and P_n are univariate distributions depending on n (where we make the notational dependence explicit for concreteness), and their CDFs are $\widehat{F}_n$ and F_n respectively, then we write $\widehat{P}_n \xrightarrow{\mathrm{d}} P_n$ as $n \to \infty$ to mean that $|\widehat{F}_n(z) - F_n(z)| \to 0$ for every z that is a continuity point of F_n for all n large enough.

We remark that if we make the stronger assumption that P_λ converges weakly to a continuous distribution, then Theorem 1 can be strengthened from pointwise to uniform convergence in the following sense: in place of (16), we have $\sup_{z\in\mathbb{R}} |\widehat{F}_\lambda^{\mathrm{gcv}}(z) - F_\lambda(z)| \to 0$, where F_λ and $\widehat{F}_\lambda^{\mathrm{gcv}}$ are the distribution functions associated with P_λ and $\widehat{P}_\lambda^{\mathrm{gcv}}$, respectively, and the same result holds for LOOCV as well. This follows from standard arguments (e.g., Chapter 3 of Durrett, 2019), and we omit the details.

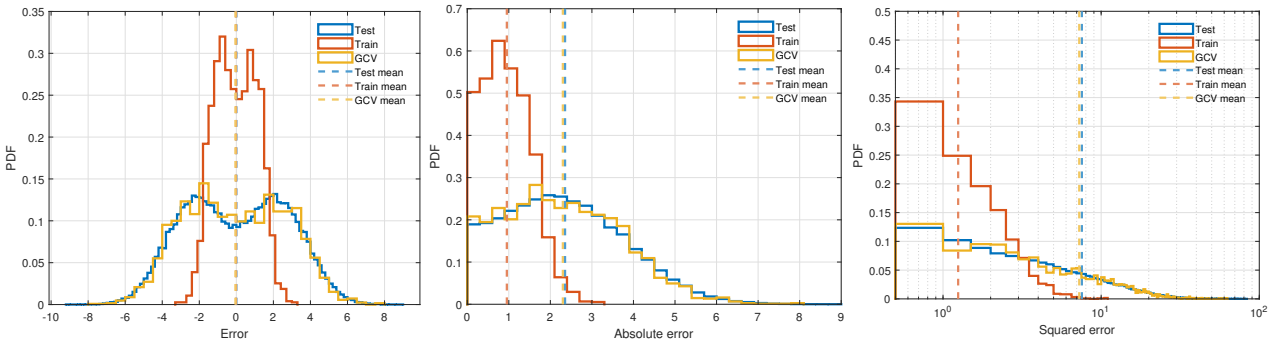
An extension (resembling the continuous mapping theorem) of Theorem 1 is given next.

Corollary 2. *Let $h : \mathbb{R} \to \mathbb{R}$ be a continuous function, and H_λ denote the distribution of the transformed error $h(e_\lambda)$ conditional on the training data. Let $\widehat{H}_\lambda^{\mathrm{gcv}}$ and $\widehat{H}_\lambda^{\mathrm{loo}}$ denote the empirical distributions as in (11)–(14), but where the point mass in each summand is evaluated at h of its argument. Then, under Assumptions 1 and 2, for $\lambda > \lambda_{\min}$,*

$$\widehat{H}_\lambda^{\mathrm{gcv}} \xrightarrow{\mathrm{d}} H_\lambda \quad \text{and} \quad \widehat{H}_\lambda^{\mathrm{loo}} \xrightarrow{\mathrm{d}} H_\lambda, \tag{17}$$

almost surely as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

Some remarks on the above results are in order. The assumptions required on the distributions of response



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

Figure 2: An example with $n = 2500$, $p = 5000$. We generated each x_i according to a Bernoulli distribution, and y_i by adding Bernoulli noise to a nonlinear (quadratic) function of x_i. The ridge tuning 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) tracks the true distribution (blue) closely. Empirical results for LOOCV are given in the supplement.

and features are very weak. Notably, we do not require that the response comes from a well-specified model. Further, the distributions of the response and feature components could be arbitrary so long as they satisfy the moment bounds. As an illustration, we consider examples with binary features and noise in Figure 2. Finally, since $\lambda_{\min} < 0$, the results cover the case of the min-norm interpolator (except when $\gamma = 1$).

We next provide some intuition as to why the above results are true. Consider the special case of an underlying linear model $y_0 = x_0^\top \beta_0 + \varepsilon_0$, where $\beta_0 \in \mathbb{R}^p$ is deterministic unknown parameter vector and ε_0 is independent of x_0. In this case, the out-of-sample prediction error simplifies to $e_\lambda = x_0^\top(\beta_0 - \widehat{\beta}_\lambda) + \varepsilon_0$, and

$$P_\lambda = \mathcal{L}\big(x_0^\top(\beta_0 - \widehat{\beta}_\lambda)\big) \star \mathcal{L}(\varepsilon_0),$$

where $\star$ denotes convolution. Further assuming that the features x_0 are Gaussian, as is the noise ε_0, with mean zero and variance σ^2, this law will be Gaussian with mean zero and variance $\|\beta_0 - \widehat{\beta}_\lambda\|_\Sigma^2 + \sigma^2$, where $\|a\|_\Sigma^2 = a^\top \Sigma a$. The variance here is the same as the mean squared prediction error of $\widehat{\beta}_\lambda$. As LOOCV and GCV (in their usual forms (8) and (9)) track this variance term, Theorem 1 can be viewed as establishing asymptotic normality of the empirical distributions of LOOCV and GCV errors, in this special case.

However, Theorem 1 is considerably more general and applies even when $\mathcal{L}(x_0^\top(\beta_0 - \widehat{\beta}_\lambda))$ does not have an analytically known asymptotic limit (and to reiterate, applies even when $\mathbb{E}[y_0 \mid x_0]$ is not linear in x_0). In fact, Theorem 1 is itself a consequence of an more general result on the convergence of certain functionals of the error distribution, which is covered next.

4 FUNCTIONAL ESTIMATION

Now we derive convergence theory on the estimation of linear functionals (5) of the out-of-sample prediction error distribution. In addition to serving as the main ingredient for proving Theorem 1, it forms a building block for establishing convergence results that apply to certain nonlinear functionals of the error distribution, discussed in the next section.

4.1 Pointwise Convergence

We impose the following assumption on the error function t in (5).

Assumption 3 (Growth rate for the error function)**.** There are constants $a, b, c > 0$ such that $|t(z)| \leq az^2 + b|z| + c$ for any $z \in \mathbb{R}$.

The quadratic growth condition on the error function t in Assumption 3 is tied to the moment conditions in Assumptions 1 and 2. In particular, both assumptions together let us bound $\mathbb{E}[|t(e_\lambda)|^{2+\xi}]$, where $\xi > 0$. One can thus relax the requirement on the growth rate by assuming higher moments in Assumptions 1 and 2.

Henceforth, let T_λ denote the linear functional in (5) corresponding to an error function t, and let $\widehat{T}_\lambda^{\text{gcv}}, \widehat{T}_\lambda^{\text{loo}}$ denote the associated plug-in estimators in (15). Next we give the first functional convergence result.

Theorem 3 (Linear functional estimation)**.** *Suppose Assumptions 1 and 2 hold, and the function t is continuous and satisfies Assumption 3. Then, for $\lambda > \lambda_{\min}$,*

$$\widehat{T}_\lambda^{\text{gcv}} - T_\lambda \to 0 \quad \textit{and} \quad \widehat{T}_\lambda^{\text{loo}} - T_\lambda \to 0, \tag{18}$$

almost surely as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

Several remarks on the above result follow. As before, the allowed range of tuning parameter values includes the min-norm estimator, since $\lambda_{\min} < 0$ (except when $\gamma = 1$). Moreover, the convergence result in (18) holds almost surely (with respect to the training data X, y). This is stronger than many previous results for CV that hold either in probability or expectation over the training data. Lastly, the error function t can be any arbitrary continuous, subquadratic function. In particular, it does *not* need to be bounded (which, by the Portmanteau theorem, would be equivalent to the weak convergence result in Theorem 1).

A special case of the last result was recently given in Patil et al. (2021) for squared error, $t(e) = e^2$, who assume a much more restricted setting of a well-specified linear model. The current result greatly extends this last one, by allowing for general error functions as well as nonlinear models. The proofs in Patil et al. (2021) exploit the bias-variance decomposition that accompanies squared error, analyze the asymptotic behavior of GCV first, and then tie this to LOOCV. Our approach in this paper is completely different (as it must be, due to the general lack of bias-variance decompositions for non-squared error functions). Below we highlight key steps involved in the proof of Theorem 3.

Proof overview. Our strategy is to study LOOCV first, and then connect it to GCV. It helps to introduce an intermediate quantity:

$$\widetilde{T}_\lambda = \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \mid X_{-i}, y_{-i}\big], \tag{19}$$

where we use X_{-i} and y_{-i} for the feature matrix and response vector with the i^{th} row and element removed, respectively, and $\widehat{\beta}_{-i,\lambda}$ for the ridge estimator trained on X_{-i} and y_{-i}. One can interpret (19) as the average of the functionals of the leave-one-out estimators $\widehat{\beta}_{-i,\lambda}$, $i = 1, \ldots, n$. The result then follows from establishing that: (i) $T_\lambda - \widetilde{T}_\lambda \xrightarrow{\text{a.s.}} 0$, (ii) $\widetilde{T}_\lambda - \widehat{T}_\lambda^{\text{loo}} \xrightarrow{\text{a.s.}} 0$, and (iii) $\widehat{T}_\lambda^{\text{loo}} - \widehat{T}_\lambda^{\text{gcv}} \xrightarrow{\text{a.s.}} 0$. In step (i), we use the modulus of continuity of a suitably truncated error function and the stability of the ridge regression estimator. Step (ii) is based on identifying a martingale difference sequence and applying the Burkholder concentration inequality. In step (iii), we use a key lemma from Patil et al. (2021) on the asymptotic equivalence of certain functionals of sample covariance matrices. The full proof is deferred to the supplement (as with all others in this paper).

4.2 Uniform Convergence

The result in Theorem 3, which is pointwise in λ, can be made uniform in λ under a stronger assumption on the error function t.

Assumption 4 (Growth rate for the derivative of the error function)**.** There are constants $g, h > 0$ such that $|t'(z)| \le g|z| + h$ for any $z \in \mathbb{R}$.

Theorem 4 (Linear functional estimation, uniform in λ)**.** *Assume the conditions of Theorem 3, and that t is differentiable and satisfies Assumption 4. Then, for any compact $\Lambda \subseteq (\lambda_{\min}, \infty)$,*

$$\sup_{\lambda\in\Lambda} \big|\widehat{T}_\lambda^{\text{gcv}} - T_\lambda\big| \to 0 \quad \textit{and} \quad \sup_{\lambda\in\Lambda} \big|\widehat{T}_\lambda^{\text{loo}} - T_\lambda\big| \to 0, \tag{20}$$

almost surely as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.

We remark that it is not essential that the error function t be differentiable. We can prove a similar result assuming that the error function t is Lipschitz continuous. We assume a global Lipschitz error function t to simplify the proof, but it should be possible to further relax this to a locally Lipschitz assumption, where we have control over the average Lipschitz constant. We do not pursue this in the current paper.

Theorem 5 (Linear functional estimation, uniform in λ, nonsmooth t)**.** *Assume the conditions of Theorem 3, and that t is Lipschitz continuous. Then, for any compact $\Lambda \subseteq (\lambda_{\min}, \infty)$, the same result as in (20) holds, almost surely as $n, p \to \infty$ and $p/n \to \gamma \in (0, \infty)$.*

Such uniform convergence will come in handy in the applications discussed next.

5 OTHER APPLICATIONS

The main application of Theorem 3 discussed thus far is the weak convergence in Theorem 1. Several other applications are possible, as detailed in this section.

5.1 Variational Functional Estimation

We consider estimation of certain nonlinear functionals that can be represented in variational form as minimizers of parametrized linear functionals over a sufficiently "nice" family of error functions. The main idea behind such an approach is to exploit uniform convergence of the plug-in estimators over the family.

Let $\mathcal{T}_\mathcal{V} = \{t(\cdot, v) : \mathbb{R} \to \mathbb{R} : v \in \mathcal{V}\}$ denote a family of functions indexed by a set $\mathcal{V} \subseteq \mathbb{R}$. Corresponding to each error function $t(\cdot, v)$ in $\mathcal{T}_\mathcal{V}$, let $T_\lambda(v)$ denote the linear functional (5) associated with $\widehat{\beta}_\lambda$. A variational error functional, denoted by V_λ, is defined as

$$V_\lambda = \arg\min_{v\in\mathcal{V}} T_\lambda(v). \tag{21}$$

This is assumed to be unique.[2] Meanwhile, denoting by $\widehat{T}_\lambda^{\text{gcv}}(v)$ and $\widehat{T}_\lambda^{\text{loo}}(v)$ the plug-in estimators (15) associated with the error function $t(\cdot, v)$, for $v \in \mathcal{V}$, we

[2]This is done for simplicity, so we do not have to appeal

can then define:

$$\widehat{V}_\lambda^{\mathrm{gcv}} \in \operatorname*{arg\,min}_{v\in\mathcal{V}} \widehat{T}_\lambda^{\mathrm{gcv}}(v), \tag{22}$$

$$\widehat{V}_\lambda^{\mathrm{loo}} \in \operatorname*{arg\,min}_{v\in\mathcal{V}} \widehat{T}_\lambda^{\mathrm{loo}}(v). \tag{23}$$

Note that we do not assume that these are unique (as is reflected by the element notation above). Our main result in the variational setting is as follows.

Theorem 6 (Variational functional estimation)**.** *Suppose Assumptions 1 and 2 hold. Let $\mathcal{T}_\mathcal{V}$ be a pointwise equicontinuous family of functions, where $\mathcal{V}$ is compact, and each $t(\cdot, v)$ satisfies Assumption 3. For $\lambda > \lambda_{\min}$,*

$$\widehat{V}_\lambda^{\mathrm{gcv}} - V_\lambda \to 0 \quad \textit{and} \quad \widehat{V}_\lambda^{\mathrm{loo}} - V_\lambda \to 0, \tag{24}$$

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

The proof of Theorem 6 builds on the previous results. We apply Theorem 3 on $t(\cdot, v)$ to establish the convergence of $\widehat{T}_\lambda^{\mathrm{gcv}}(v)$ to $T_\lambda(v)$ for each $v \in \mathcal{V}$. The pointwise equicontinuity of functions in $\mathcal{T}_\mathcal{V}$ leads to stochastic equicontinuity of $\widehat{T}_\lambda^{\mathrm{gcv}}(v) - T_\lambda(v)$, which then provides GCV part of (24). Similar arguments hold for LOOCV.

5.2 Quantile Estimation

To illustrate the use of Theorem 6, we consider estimating quantiles of the out-of-sample prediction error distribution. For $\tau \in (0, 1)$, let $Q_\lambda(\tau)$ denote the level-τ conditional quantile (6), assumed unique for simplicity. While this is a nonlinear functional of P_λ, we will exploit the fact that (6) can expressed in an equivalent variational form (Koenker and Bassett Jr., 1978):

$$Q_\lambda(\tau) = \operatorname*{arg\,min}_{u\in\mathcal{U}} \mathbb{E}\big[t_\tau\big(y_0 - x_0^\top \widehat{\beta}_\lambda - u\big) \mid X, y\big], \tag{25}$$

where $t_\tau(u) = u(\tau - \mathbb{I}(u < 0))$, sometimes called the pinball or tilted ℓ_1 loss. If $\mathcal{U}$ is any set containing the true quantile, we can recognize $Q_\lambda(\tau)$ as being in the form (21), for the family $\mathcal{T}_\mathcal{U} = \{t_\tau(\cdot, u) : u \in \mathcal{U}\}$. We can then define plug-in estimators $\widehat{Q}_\lambda^{\mathrm{gcv}}(\tau)$ and $\widehat{Q}_\lambda^{\mathrm{loo}}(\tau)$ as in (22) and (23), or to be fully explicit:

$$\widehat{Q}_\lambda^{\mathrm{gcv}}(\tau) \in \operatorname*{arg\,min}_{u\in\mathcal{U}} \frac{1}{n}\sum_{i=1}^n t_\tau\bigg(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \frac{\mathrm{tr}[L_\lambda]}{n}} - u\bigg), \tag{26}$$

$$\widehat{Q}_\lambda^{\mathrm{loo}}(\tau) \in \operatorname*{arg\,min}_{u\in\mathcal{U}} \frac{1}{n}\sum_{i=1}^n t_\tau\bigg(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} - u\bigg), \tag{27}$$

with suitable adaptations based on (13), (14) if $\lambda = 0$. These are essentially just the sample quantiles of GCV and LOOCV residuals, up to discretization issues (the sample quantiles not being unique for integral τn).

to set-theoretic notation for convergence of minimizers in the statements that follow. More general formulations that do not assume uniqueness, via variational analysis, should be possible.

Corollary 7 (Quantile estimation)**.** *Suppose Assumptions 1 and 2 hold. Given $\tau \in (0, 1)$, assume the level-τ quantile $Q_\lambda(\tau)$ of P_λ is unique, and assume $\mathcal{U}$ in (26), (27) is any compact set that contains the true quantile. For any $\lambda > \lambda_{\min}$,*

$$\widehat{Q}_\lambda^{\mathrm{gcv}}(\tau) - Q_\lambda(\tau) \to 0 \quad \textit{and} \quad \widehat{Q}_\lambda^{\mathrm{loo}}(\tau) - Q_\lambda(\tau) \to 0, \tag{28}$$

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

Thanks to the general result in Theorem 6, the proof of (28) reduces to verifying the pointwise equicontinuity of the family of pinball loss functions.

Estimating quantiles gives us a way to construct prediction intervals for the out-of-sample response y_0, of the form:

$$\mathcal{I}_\lambda^{\mathrm{gcv}} = \big[x_0^\top \widehat{\beta}_\lambda - \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_l),\, x_0^\top \widehat{\beta}_\lambda + \widehat{Q}_\lambda^{\mathrm{gcv}}(\tau_u)\big], \tag{29}$$

$$\mathcal{I}_\lambda^{\mathrm{loo}} = \big[x_0^\top \widehat{\beta}_\lambda - \widehat{Q}_\lambda^{\mathrm{loo}}(\tau_l),\, x_0^\top \widehat{\beta}_\lambda + \widehat{Q}_\lambda^{\mathrm{loo}}(\tau_u)\big], \tag{30}$$

where $\tau_l < \tau_u$ are appropriate lower and upper quantile levels chosen to provide the desired coverage. These intervals have asymptotically exact coverage conditional on the training set, as a consequence of Corollary 7. See Figure 3 for empirical results.

5.3 Regularization Tuning

One important application of convergence results that are uniform in λ, for given functionals, is that we can tune the amount of regularization according to those functionals, and uniformity will imply that any minimizer of the plug-in estimator converges to a minimizer of the population functional. A typical strategy is to tune by minimizing the mean squared GCV or LOOCV error; but we can also tune via more robust measures such as absolute error, Huber error, or the length of the prediction intervals.

The next corollary certifies that the the level of regularization tuned by using the plug-in GCV and LOOCV estimators is almost surely optimal for a wide range of error functions.

Corollary 8 (Convergence of tuned errors)**.** *Suppose Assumptions 1 and 2 hold. Suppose the error function t satisfies Assumption 3, and furthermore, it is either differentiable and satisfies Assumption 4, or else it is Lipschitz. Let $\Lambda \subseteq (\lambda_{\min}, \infty)$ be compact, and let $\lambda^\star$ be a minimizer of T_λ over Λ. Similarly, let $\widehat{\lambda}^{\mathrm{gcv}}$ and $\widehat{\lambda}^{\mathrm{loo}}$ denote minimizers of $\widehat{T}_\lambda^{\mathrm{gcv}}$ and $\widehat{T}_\lambda^{\mathrm{loo}}$ over Λ, respectively. Then,*

$$T_{\widehat{\lambda}^{\mathrm{gcv}}} - T_{\lambda^\star} \to 0 \quad \textit{and} \quad T_{\widehat{\lambda}^{\mathrm{loo}}} - T_{\lambda^\star} \to 0, \tag{31}$$

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

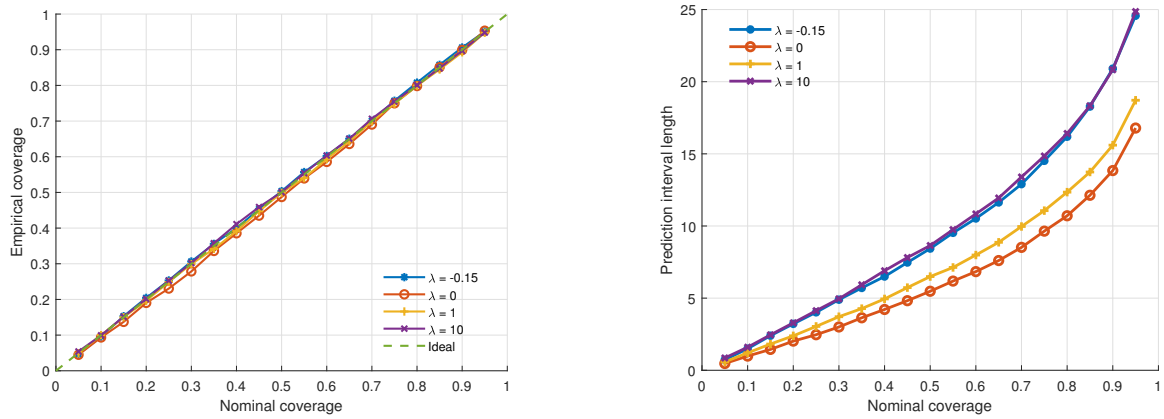


Figure 3: Illustration of empirical coverage and length of GCV prediction intervals (29) against nominal coverage, where $n = 2500$, $p = 5000$. The data model has a latent structure with autoregressive feature covariance and true signal aligned with the principal eigenvector, similar to that in Kobak et al. (2020) (the supplement gives details), who investigated the empirical optimality of the min-norm interpolator. Here 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).

6 DISCUSSION

In this paper, we investigate the distribution of errors arising from both generalized and leave-one-out cross-validation in the context of ridge regression. We show that these distributions converge to the out-of-sample prediction error distribution, under generic conditions. A core result in our work is on consistent estimation of linear functionals of the error distribution, yielding wide implications, including an extension to estimating certain nonlinear functionals which has applications in conditional predictive inference.

Amazingly (and surprisingly, even to us), these results continue to hold in an high-dimensional setting when $p > n$. LOOCV for ridge regression takes on a special form, based on the beautiful "shortcut" relation:

$$y_i - x_i^\top \widehat{\beta}_{-i,\lambda} = \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}} \approx \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}.$$

When $p > n$ and $\lambda = 0$, the numerator and denominator in both fractions here are zero. However, as $\lambda \to 0$ the numerator and denominator (in each fraction) tend to zero at exactly the same rate, allowing us to "cancel" the dependence on λ infinitesimally, leading to:

$$y_i - x_i^\top \widehat{\beta}_{-i,0} = \frac{[(XX^\top)^\dagger y]_i}{[(XX^\top)^\dagger]_{ii}} \approx \frac{[(XX^\top)^\dagger y]_i}{\mathrm{tr}[(XX^\top)^\dagger]/n}.$$

This fact was first derived in Hastie et al. (2019), and it is key for our results.

The most immediate next direction is to study kernel ridge regression, which yields a similar "shortcut" formula (Hastie, 2020) where $XX^\top$ gets replaced by the kernel gram matrix. For other predictive models that do not yield exact leave-one-out formulae (in terms of training errors), examining to what degree similar results hold true is an interesting direction for future study. This is especially interesting for "benign" interpolators, now an active area of research, which decompose into a "simple" component useful for prediction and a "spiky" component that interpolates the training data (Bartlett et al., 2021). As interpolators gain a central role in modern machine learning, adapting CV methods to work seamlessly with them is becoming of foundational importance. This current paper serves as a step in that direction.

Acknowledgements

We thank Arun Kumar Kuchibhotla and Yuting Wei for helpful discussions. We also thank the anonymous reviewers for their comments that improved the presentation of this paper. PP and RJT were supported by ONR grant N00014-20-1-2787.

References

David M. Allen. The relationship between variable selection and data agumentation and a method for prediction. *Technometrics*, 16(1):125–127, 1974.

Sylvain Arlot and Alain Celisse. A survey of cross-validation procedures for model selection. *Statistics surveys*, 4:40–79, 2010.

Morgane Austern and Wenda Zhou. Asymptotics of cross-validation. *arXiv preprint arXiv:2001.11111*, 2020.

Zhi-Dong Bai and Jack W. Silverstein. No eigenvalues outside the support of the limiting spectral distribu-

tion 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.

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

Peter L. Bartlett, Andrea Montanari, and Alexander Rakhlin. Deep learning: a statistical viewpoint. *arXiv preprint arXiv:2103.09177*, 2021.

Pierre Bayle, Alexandre Bayle, Lucas Janson, and Lester Mackey. Cross-validation confidence intervals for test error. *arXiv preprint arXiv:2007.12671*, 2020.

Ahmad Beirami, Meisam Razaviyayn, Shahin Shahrampour, and Vahid Tarokh. On optimal generalizability in parametric learning. *arXiv preprint arXiv:1711.05323*, 2017.

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

Olivier Bousquet and André Elisseeff. Stability and generalization. *The Journal of Machine Learning Research*, 2:499–526, 2002.

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

Alain Celisse and Benjamin Guedj. Stability revisited: new generalisation bounds for the leave-one-out. *arXiv preprint arXiv:1608.06412*, 2016.

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.

Rick Durrett. *Probability: theory and examples*, volume 49. Cambridge university press, 2019.

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

Seymour Geisser. The predictive sample reuse method with applications. *Journal of the American statistical Association*, 70(350):320–328, 1975.

Ryan Giordano, William Stephenson, Runjing Liu, Michael Jordan, and Tamara Broderick. A swiss army infinitesimal jackknife. In *The 22nd International Conference on Artificial Intelligence and Statistics*, pages 1139–1147. PMLR, 2019.

Gene H. Golub, Michael Heath, and Grace Wahba. Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics*, 21(2):215–223, 1979.

László Györfi, Michael Kohler, Adam Krzyzak, and Harro Walk. *A Distribution-free Theory of Nonparametric Regression*. Springer Science & Business Media, 2006.

Trevor Hastie. Ridge regularization: An essential concept in data science. *Technometrics*, 62(4):426–433, 2020.

Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning*. Springer Series in Statistics, 2009. Second edition.

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

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, 12(1):55–67, 1970a.

Arthur E. Hoerl and Robert W. Kennard. Ridge regression: applications to nonorthogonal problems. *Technometrics*, 12(1):69–82, 1970b.

Satyen Kale, Ravi Kumar, and Sergei Vassilvitskii. Cross-validation and mean-square stability. In *In Proceedings of the Second Symposium on Innovations in Computer Science*. Citeseer, 2011.

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.

Roger Koenker and Gilbert Bassett Jr. Regression quantiles. *Econometrica: Journal of the Econometric Society*, pages 33–50, 1978.

Ravi Kumar, Daniel Lokshtanov, Sergei Vassilvitskii, and Andrea Vattani. Near-optimal bounds for cross-validation via loss stability. In *International Conference on Machine Learning*, pages 27–35. PMLR, 2013.

Ker-Chau Li. From Stein's unbiased risk estimates to the method of generalized cross validation. *The Annals of Statistics*, pages 1352–1377, 1985.

Ker-Chau Li. Asymptotic optimality of C_L and generalized cross-validation in ridge regression with appli-

cation to spline smoothing. *The Annals of Statistics*, 14(3):1101–1112, 1986.

Ker-Chau Li. Asymptotic optimality for C_p, C_L, cross-validation and generalized cross-validation: Discrete index set. *The Annals of Statistics*, 15(3):958–975, 1987.

Tengyuan Liang and Alexander Rakhlin. Just interpolate: Kernel "ridgeless" regression can generalize. *The Annals of Statistics*, 48(3):1329–1347, 2020.

Yong Liu, Shali Jiang, and Shizhong Liao. Efficient approximation of cross-validation for kernel methods using bouligand influence function. In *International Conference on Machine Learning*, pages 324–332. PMLR, 2014.

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.

Rosa J. Meijer and Jelle J. Goeman. Efficient approximate k-fold and leave-one-out cross-validation for ridge regression. *Biometrical Journal*, 55(2):141–155, 2013.

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

Tomoyuki Obuchi and Yoshiyuki Kabashima. Cross validation in LASSO and its acceleration. *Journal of Statistical Mechanics: Theory and Experiment*, 2016.

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.

Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the out-of-sample prediction error via approximate leave-one-out cross-validation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 82(4):965–996, 2020.

Kamiar Rahnama Rad, Wenda Zhou, and Arian Maleki. Error bounds in estimating the out-of-sample prediction error using leave-one-out cross validation in high-dimensions. In *International Conference on Artificial Intelligence and Statistics*, pages 4067–4077. PMLR, 2020.

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

Lukas Steinberger and Hannes Leeb. Leave-one-out prediction intervals in linear regression models with many variables. *arXiv preprint arXiv:1602.05801*, 2016.

Lukas Steinberger and Hannes Leeb. Conditional predictive inference for high-dimensional stable algorithms. *arXiv preprint arXiv:1809.01412*, 2018.

William Stephenson and Tamara Broderick. Approximate cross-validation in high dimensions with guarantees. In *International Conference on Artificial Intelligence and Statistics*, pages 2424–2434. PMLR, 2020.

Mervyn Stone. Cross-validatory choice and assessment of statistical predictions. *Journal of the Royal Statistical Society: Series B*, 36(2):111–133, 1974.

Mervyn Stone. Asymptotics for and against cross-validation. *Biometrika*, 64(1):29–35, 1977.

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

Andrei N. Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady Akademii Nauk SSSR*, volume 151, pages 501–504, 1963.

Andrey N. Tikhonov. On the stability of inverse problems. In *Doklady Akademii Nauk SSSR*, volume 39, pages 195–198, 1943.

Grace Wahba. Spline bases, regularization, and generalized cross-validation for solving approximation problems with large quantities of noisy data. *Approximation Theory III*, 1980.

Grace Wahba. *Spline Models for Observational Data*. SIAM, 1990.

Shuaiwen Wang, Wenda Zhou, Arian Maleki, Haihao Lu, and Vahab Mirrokni. Approximate leave-one-out for high-dimensional non-differentiable learning problems. *arXiv preprint arXiv:1810.02716*, 2018.

Ashia Wilson, Maximilian Kasy, and Lester Mackey. Approximate cross-validation: Guarantees for model assessment and selection. In *International Conference on Artificial Intelligence and Statistics*, pages 4530–4540. PMLR, 2020.

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

Ji Xu, Arian Maleki, Kamiar Rahnama Rad, and Daniel Hsu. Consistent risk estimation in moderately high-dimensional linear regression. *IEEE Transactions on Information Theory*, 67(9):5997–6030, 2021.

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

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." The content of the supplement is organized as follows. In Appendices A to C, 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 Appendix D contains proof of Theorem 1, while Appendix E contains proofs related to Theorem 6, along with further theoretical results related to quantile estimation. Additional numerical results and experimental details are provided in Appendix F. Finally, Appendix G 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.

A 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: $\widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, and use it to show the first part: $\widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$.

- To prove $\widehat{T}_\lambda^{\mathrm{loo}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, we introduce an intermediate quantity $\widetilde{T}_\lambda$ as in (19) and break the difference

$$T_\lambda - \widehat{T}_\lambda^{\mathrm{loo}} = (T_\lambda - \widetilde{T}_\lambda) + (\widetilde{T}_\lambda - \widehat{T}_\lambda^{\mathrm{loo}}). \tag{32}$$

  We will show that both terms in the decomposition (32) almost surely vanish. Appendix A.1 shows the convergence for the first term, while Appendix A.2 shows the convergence for the second term.

- To prove $\widehat{T}_\lambda^{\mathrm{gcv}} - T_\lambda \xrightarrow{\text{a.s.}} 0$, we similarly break the difference

$$T_\lambda - \widehat{T}_\lambda^{\mathrm{gcv}} = (T_\lambda - \widehat{T}_\lambda^{\mathrm{loo}}) + (\widehat{T}_\lambda^{\mathrm{loo}} - \widehat{T}_\lambda^{\mathrm{gcv}}). \tag{33}$$

  We have already dealt with the first term in the decomposition (33) in (32). We show the second term almost surely goes to zero in Appendix A.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 Appendix A.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 $\overline{\omega}_t(r) \leq 2\omega_t(r)$ for $r \geq 0$. We will make use of these properties below.

A.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

$$\left|\frac{1}{n}\sum_{i=1}^n \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) - t(y_0 - x_0^\top \widehat{\beta}_{-i,\lambda}) \mid X, y\big]\right| \xrightarrow{\text{a.s.}} 0. \tag{34}$$

Using the modulus of continuity of t and its least concave majorant, we first bound the summands in (34) 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 (34) as

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

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

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \;\Big|\; X, y\Big] \xrightarrow{\text{a.s.}} 0, \tag{35}$$

which along with the continuity of the modulus that vanishes at 0 shows (34), leading to the desired conclusion that $T_\lambda - \widetilde{T}_\lambda \xrightarrow{\text{a.s.}} 0$.

Towards showing (35), first note that under Assumption 1, we can bound the summands for each $i = 1, \dots, n$ as

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

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 (35) by

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}\Big[\big|x_0^\top(\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda})\big| \mid X, y\Big] \le \sqrt{r_{\max}}\left(\frac{1}{n}\sum_{i=1}^n \big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2\right). \tag{36}$$

We show in Lemma 10 that the term in the parenthesis on the right-hand side of (36) almost surely goes to zero under Assumptions 1 and 2, proving (35) and completing the proof.

A.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:

$$\begin{aligned}
\big|\widetilde{T}_\lambda - \widehat{T}_\lambda^{\text{loo}}\big| &= \left|\widetilde{T}_\lambda - \frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) + \frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widehat{T}_\lambda^{\text{loo}}\right| \\
&\le \left|\widetilde{T}_\lambda - \frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda})\right| + \left|\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widehat{T}_\lambda^{\text{loo}}\right|.
\end{aligned} \tag{37}$$

In the sequel, we will show that each of two pieces in (37) vanishes almost surely under proportional asymptotics.

For the second piece in (37), using the modulus of t and its concave majorant, we can bound the difference as

$$\begin{aligned}
\left|\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widetilde{T}_\lambda^{\text{loo}}\right| &= \left|\frac{1}{n}\sum_{i=1}^n t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \frac{1}{n}\sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right| \\
&\le \frac{1}{n}\sum_{i=1}^n \left|t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right| \\
&\le \frac{1}{n}\sum_{i=1}^n \overline{\omega}_t\left(\left|y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\le \overline{\omega}_t\left(\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right) \\
&\le 2\omega\left(\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_{-i,\lambda} - \frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right|\right),
\end{aligned} \tag{38}$$

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

$$\left| \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,0}) - \widetilde{T}_\lambda^{\text{loo}} \right| \leq 2\omega \left(\frac{1}{n} \sum_{i=1}^{n} \left| y_i - x_i^\top \widehat{\beta}_{-i,0} - \frac{[(XX^\top/n)^\dagger]_i}{[(XX^\top/n)^\dagger]_{ii}} \right| \right). \tag{39}$$

The argument of ω in either cases of (38) and (39) goes to 0 almost surely, and thus the continuity of ω provides the desired convergence of the second piece in (37) It is worth mentioning that the only reason we need to worry about (38) and (39) 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 (37). Note that we can express

$$\begin{aligned} \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \widetilde{T}_\lambda &= \frac{1}{n} \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \frac{1}{n} \sum_{i=1}^{n} \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\big|\; X_{-i}, y_{-i} \big] \\ &= \frac{1}{n} \sum_{i=1}^{n} \Big\{ t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\big|\; X_{-i}, y_{-i} \big] \Big\}. \end{aligned} \tag{40}$$

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

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

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 16 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

$$\begin{aligned} &\mathbb{E}\left[\left| \sum_{i=1}^{n} t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\big|\; X_{-i}, y_{-i} \big] \right|^q \right] \\ &\leq C\mathbb{E}\left[\left\{ \sum_{i=1}^{n} \mathbb{E}\left[\left| t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\big|\; X_{-i}, y_{-i} \big] \right|^2 \;\Big|\; \mathcal{F}_{i-1} \right] \right\}^{q/2} \right] \\ &\quad + C\mathbb{E}\left[\sum_{i=1}^{n} \left| t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) - \mathbb{E}\big[t(y_i - x_i^\top \widehat{\beta}_{-i,\lambda}) \;\big|\; X_{-i}, y_{-i} \big] \right|^q \right] \end{aligned} \tag{41}$$

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

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

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 9, 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 11 that implies $\|\widehat{\beta}_{-i,\lambda}\|_2 \leq C$ almost surely for n large enough under Assumptions 1 and 2, we have

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

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

Therefore, from (41) we can bound q-th moment of normalized sum (40) 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 22 provides the desired convergence for the first piece in (37). This concludes the proof.

A.3 LOOCV Estimator to GCV Estimator

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

$$\begin{aligned}\left|\widehat{T}_\lambda^{\mathrm{gcv}} - \widehat{T}_\lambda^{\mathrm{loo}}\right| &= \left|\frac{1}{n}\sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) - \frac{1}{n}\sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right| \\ &\leq \frac{1}{n}\sum_{i=1}^n \left|t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)\right|. \end{aligned} \tag{42}$$

We will show below that the right-hand side of the expression (42) 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 (42) 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

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

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

$$\frac{1}{n}\sum_{i=1}^n \left|y_i - x_i^\top \widehat{\beta}_\lambda\right| \left|\frac{1}{1 - \mathrm{tr}[L_\lambda]/n} - \frac{1}{1 - [L_\lambda]_{ii}}\right| \xrightarrow{\text{a.s.}} 0. \tag{43}$$

To build towards proving (43), 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 $i = 1, \dots, n$, by adding and subtracting $1 + \operatorname{tr}\left[(X^\top X/n + \lambda I)^\dagger \Sigma\right]/n$, and $\operatorname{tr}\left[(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger \Sigma\right]/n$, we decompose the difference

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

This lets us decompose

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

Finally, we verify that each of the term in the decomposition almost surely vanishes. Using the $\lambda \neq 0$ case of Lemma 19, we have for the first term

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

For the second term, following the proof of Lemma 19, for $i = 1, \dots, n$ we can bound

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

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

$$(X^\top X/n + \lambda I)^\dagger - (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger = -\frac{(X_{-i}^\top X_{-i}/n + \lambda I)^\dagger x_i x_i^\top /n (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger}{1 + x_i^\top (X_{-i}^\top X_{-i}/n + \lambda I)^\dagger x_i}. \tag{44}$$

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 $x_i^\top\left(X_{-i}^\top X_{-i}/n+\lambda I\right)^\dagger x_i$ 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 18, the above quantity is of order $O(n/n^q)$. Choosing $q>2$ and applying Lemma 22 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 Appendix A.4.) This reduces to showing

$$\frac{1}{n}\sum_{i=1}^n\left|[(XX^\top/n)^\dagger y]_i\right|\left|\frac{1}{\mathrm{tr}[(XX^\top/n)^\dagger]/n}-\frac{1}{[(XX^\top/n)^\dagger]_{ii}}\right|\xrightarrow{\text{a.s.}}0. \tag{45}$$

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 Appendix A.4. Finally, analogous to the argument used to bound d, we can now use the case of $\lambda=0$ equivalence in Lemma 19 for the difference vector in the modified errors of (45). This takes care of both the cases and concludes the proof.

A.4 Truncation Arguments

We established the converges in Appendices A.1 to A.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 Appendices A.1 to A.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:

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

where the third line uses the Cauchy-Schwarz inequality, the fourth line uses Lemmas 9 and 11 with $q = 2$, the fifth line uses Chebychev's inequality, and the last line again uses Lemmas 9 and 11 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:

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

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 Appendix A.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

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

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

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

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

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

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

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:

$$\begin{aligned}y_i - x_i^\top \widehat{\beta}_\lambda &= y_i - x_i^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n \\ &= y_i - [X^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n]_i \\ &= [y - X^\top (X^\top X/n + \lambda I)^\dagger X^\top y/n]_i \\ &= [(I - X^\top (X^\top X/n + \lambda I)^\dagger X/n)y]_i \\ &= \lambda[(XX^\top/n + \lambda I)^\dagger y]_i \end{aligned} \tag{46}$$

Similarly, we rewrite the denominator of GCV using

$$\begin{aligned}1 - \text{tr}[L_\lambda]/n &= 1 - \text{tr}[X(XX^\top/n + \lambda I)^\dagger X^\top]/n \\ &= \text{tr}[I - X(XX^\top/n + \lambda I)^\dagger X^\top]/n \\ &= \lambda\,\text{tr}[(XX^\top/n + \lambda I)^\dagger]/n. \end{aligned} \tag{47}$$

This lets us rewrite the invidual GCV reweighted errors as

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

Thus, we can now bound

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

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 11. This finishes the argument.

A.5 Auxiliary Lemmas

In this section, we gather supporting lemmas used in the proofs in Appendices A.1 to A.3, along with their proofs.

Lemma 9 (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 17 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 10 (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

$$\frac{1}{n}\sum_{i=1}^{n}\big\|\widehat{\beta}_\lambda - \widehat{\beta}_{-i,\lambda}\big\|_2 \le \frac{1}{n}\sum_{i=1}^{n}\big\|(X_{-i}^\top X_{-i}/n+\lambda I)^\dagger\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\}\big\| + \frac{1}{n}\sum_{i=1}^{n}\big\|\{(X^\top X/n+\lambda I)^\dagger - (X_{-i}^\top X_{-i}/n+\lambda I)^\dagger\}X^\top y/n\big\|. \tag{48}$$

We will see below that each of the two terms on the right-hand side of (48) almost surely goes to zero providing the desired convergence. Note that for each $i = 1, \dots, n$, we can bound

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

where the second line follows from the fact that $\big\|(X_{-i}^T X_{-i}/n+\lambda I)^\dagger\big\|_{\text{op}}$ is almost surely bounded for n large enough (as explained in the proof of Lemma 11), and last line uses triangle inequality. Now writing $x_i = \Sigma^{1/2} z_i$, note that for each $i = 1, \dots, n$,

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

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

$$\begin{aligned}\frac{1}{n}\sum_{i=1}^{n}\big\|(X_{-i}^\top X_{-i}+\lambda I)^\dagger\{X^\top y/n - X_{-i}^\top y_{-i}/(n-1)\}\big\| &\le \frac{C}{\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n}|y_i| + \frac{C}{(n-1)\sqrt{n}}\frac{1}{n}\sum_{i=i}^{n}\sum_{j\neq i}|y_j| \\ &\le \frac{C}{\sqrt{n}}\frac{(2n-1)}{(n-1)n}\sum_{i=1}^{n}|y_i| \\ &\le \frac{C}{\sqrt{n}} \to 0.\end{aligned} \tag{49}$$

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 (44), we can similarly show that the second term goes to zero almost surely. Hence, we have that (48) almost surely goes to zero. This completes the proof.

□

Lemma 11 (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

$$\begin{aligned}\big\|\widehat{\beta}_\lambda\big\|_2 &= \big\|(X^\top X/n+\lambda I)^\dagger X^\top y/n\big\|_2 \\ &\le \big\|(X^\top X/n+\lambda I)^\dagger X^\top/\sqrt{n}\big\|_{\text{op}}\|y\|_2/\sqrt{n} \\ &\le \big\|(X^\top X/n+\lambda I)^\dagger\big\|_{\text{op}}\big\|X^\top/\sqrt{n}\big\|_{\text{op}}\|y\|_2/\sqrt{n}.\end{aligned} \tag{50}$$

Now for $\lambda \in (\lambda_{\min}, \infty)$, the first two terms in the product (50) 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. □

B PROOFS RELATED TO Theorem 4

To show almost sure uniform convergence (in λ), we will appeal to Lemma 20. 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 $\widehat{T}_\lambda^{\text{gcv}}$ and their derivates are bounded over Λ, implying strong stochastic equicontinuity of the family of functions $\{T_\lambda - \widehat{T}_\lambda^{\text{gcv}}\}_{\lambda\in\Lambda}$. Analogous analysis holds for $\{T_\lambda - \widehat{T}_\lambda^{\text{loo}}\}_{\lambda\in\Lambda}$, 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 9 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

$$T_\lambda = \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \le \big(C_1 + C_2 \|\widehat{\beta}_\lambda\|_2\big)^2, \tag{51}$$

for some positive constants C_1 and C_2. Now following Lemma 11, 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 $\widehat{T}_\lambda^{\text{gcv}}$. We next consider $\widehat{T}_\lambda^{\text{gcv}}$. Using the alternate representation (46), for error function t satisfying Assumption 3, for some positive constants C, C_1, C_2, we can bound

$$\begin{aligned}
\widehat{T}_\lambda^{\text{gcv}} &= \frac{1}{n}\sum_{i=1}^n t\left(\frac{[(XX^\top/n + \lambda I)^\dagger y]_i}{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n}\right) \\
&\le \frac{C_2}{n}\sum_{i=1}^n \frac{\big\{[(XX^\top/n + \lambda I)^\dagger y]_i\big\}^2}{\big\{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n\big\}^2} + \frac{C_1}{n}\sum_{i=1}^n \frac{\big|[(XX^\top/n + \lambda I)^\dagger y]_i\big|}{\big|\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n\big|} + C \\
&\le \frac{C_2}{n}\sum_{i=1}^n \big\{[(XX^\top/n + \lambda I)^\dagger y]_i\big\}^2 + \frac{C_1}{n}\sum_{i=1}^n \big|[(XX^\top/n + \lambda I)^\dagger y]_i\big| + C.
\end{aligned} \tag{52}$$

The last inequality above follows by noting that the map $\lambda \mapsto \operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n$ is non-increasing over $[\underline{\lambda}, \overline{\lambda}]$, so $\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n$ is lower bounded by $\operatorname{tr}[(XX^\top/n + \underline{\lambda} I)^\dagger/n]$. Since $\lambda_{\min} < \underline{\lambda}$, we then have that $\{\operatorname{tr}[(XX^\top/n + \lambda I)^\dagger]/n\}^{-1}$ is upper bounded by $(\lambda_{\min} + \underline{\lambda})^{-1}$. Now, observe that for the first term in (52):

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

Similarly, note that for the second term in (52):

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

Since $\|(XX^\top/n + \lambda I)^\dagger\|_{\text{op}}$ 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 $\widehat{T}_\lambda^{\text{gcv}}$ 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

$$T_\lambda' = \mathbb{E}\big[t'(y_0 - x_0^\top \widehat{\beta}_\lambda)\; x_0^\top \widehat{\beta}_\lambda' \mid X, y\big] \le \mathbb{E}\big[\{t'(y_0 - x_0^\top \widehat{\beta}_\lambda)\}^2 \mid X, y\big]^{1/2} \cdot \mathbb{E}\big[(\widehat{\beta}_\lambda')^\top x_0 x_0^\top \widehat{\beta}_\lambda' \mid X, y\big] \le C\sqrt{r_{\max}}\|\widehat{\beta}_\lambda'\|_2. \tag{53}$$

In the above chain, the first inequality follows from Cauchy-Schwarz inequality, and the second inequality follows from the bounding of T_λ per (51) above (because under Assumption 3, t' is bounded above by a linear function), and the fact that $\|\Sigma\|_{\text{op}} \le r_{\max}$. Applying Lemma 12 on the last term of (53), we thus conclude that the derivative of T_λ is almost surely uniformly bounded over $\lambda \in \Lambda$, as desired.

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} t'\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right) \\
&\leq \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\{t'\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2}\sqrt{\frac{1}{n}\sum_{i=1}^{n}\left\{\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2} \qquad (54)\\
&\leq C\sqrt{\sum_{i=1}^{n}\left\{\frac{d}{d\lambda}\left(\frac{[(XX^\top/n+\lambda I)^\dagger y]_i}{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n}\right)\right\}^2} \qquad (55)
\end{aligned}$$

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 $\widehat{T}_\lambda^{\mathrm{gcv}}$ per (52) above shows that the first term of (54) is almost surely bounded. Applying Lemma 13, we can now upper bound the final term of (55). This leads the derivative of $\widehat{T}_\lambda^{\mathrm{gcv}}$ to be almost surely bounded over $\lambda \in \Lambda$ and concludes the proof.

Lemma 12 (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 11. □

Lemma 13 (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*

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

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

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

Note that $\{\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger]/n\}^{-1}$ 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 $\mathrm{tr}[(XX^\top/n+\lambda I)^\dagger(XX^\top/n+\lambda I)^\dagger]/n$ 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. □

C PROOFS RELATED TO Theorem 5

The proof is similar to that of proof of Theorem 4. We will again use Lemma 20. 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 $\{\widehat{T}_\lambda^{\mathrm{gcv}}\}_{\lambda\in\Lambda}$, assuming that the loss function t is Lipschitz continuous. This will show that $\{T_\lambda - \widehat{T}_\lambda^{\mathrm{gcv}}\}_{\lambda\in\Lambda}$ is almost surely Lipschitz continuous from which the theorem follows. A similar analysis holds for $\{T_\lambda - \widehat{T}_\lambda^{\mathrm{loo}}\}_{\lambda\in\Lambda}$.

Lemma 14 (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:

$$\begin{aligned}
\left\|\widehat{\beta}_{\lambda_1}-\widehat{\beta}_{\lambda_2}\right\|_2 &= \left\|X(XX^\top/n+\lambda_1)^\dagger y/n - X(XX^\top/n+\lambda_2)^\dagger y/n\right\|_2 \\
&\le \left\|X/\sqrt{n}\right\|_{\mathrm{op}}\left\|(XX^\top/n+\lambda_1)-(XX^\top/n+\lambda_2)\right\|_{\mathrm{op}}\|y\|_2/\sqrt{n}. \qquad (56)
\end{aligned}$$

As argued before, both the first and the last term in the product (56) 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

$$\begin{aligned}
\left\|(XX^\top/n+\lambda_1 I)^\dagger-(XX^\top/n+\lambda_2 I)^\dagger\right\|_{\mathrm{op}} &\le \left\|(XX^\top/n+\lambda I)^\dagger(XX^\top/n+\lambda I)^\dagger\right\|_{\mathrm{op}}|\lambda_1-\lambda_2| \\
&\le \left\|(XX^\top/n+\lambda I)^\dagger\right\|_{\mathrm{op}}\left\|(XX^\top/n+\lambda I)^\dagger\right\|_{\mathrm{op}}|\lambda_1-\lambda_2| \\
&\le C\,|\lambda_1-\lambda_2|, \qquad (57)
\end{aligned}$$

where the last inequality follows because $\lambda \geq \underline{\lambda} > \lambda_{\min}$ as explained in the proof of Lemma 11. This concludes the proof. □

Lemma 15 (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 (46) for the numerator and (47) 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

$$\begin{aligned}
&\widehat{T}_{\lambda_1}^{\mathrm{gcv}} - \widehat{T}_{\lambda_2}^{\mathrm{gcv}} \qquad (58)\\
&= \frac{1}{n}\sum_{i=1}^{n} t\left(\frac{\left[(XX^\top/n + \lambda_1 I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n + \lambda_1 I)^\dagger\right]/n}\right) - t\left(\frac{\left[(XX^\top/n + \lambda_2 I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n + \lambda_2 I)\right]/n}\right)\\
&\leq \frac{1}{n}\sum_{i=1}^{n} L\left|\frac{\left[(XX^\top/n + \lambda_1 I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n + \lambda_1 I)^\dagger\right]/n} - \frac{\left[(XX^\top/n + \lambda_2 I)^\dagger y\right]_i}{\mathrm{tr}\left[(XX^\top/n + \lambda_2 I)^\dagger\right]/n}\right|\\
&\leq L\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n + \lambda_1 I)^\dagger\right]/n} - \frac{1}{\mathrm{tr}\left[(XX^\top/n + \lambda_2 I)^\dagger\right]/n}\right| \frac{1}{n}\sum_{i=1}^{n}\left|\left[(XX^\top/n + \lambda_1 I)^\dagger y\right]_i - \left[(XX^\top/n + \lambda_2 I)^\dagger y\right]_i\right|\\
&\leq L\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n + \lambda_1 I)^\dagger\right]/n} - \frac{1}{\mathrm{tr}\left[(XX^\top/n + \lambda_2 I)^\dagger\right]/n}\right| \frac{1}{n}\sum_{i=1}^{n}\left|\left[\left\{(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\right\} y\right]_i\right|\\
&\leq L\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n + \lambda_1 I)^\dagger\right]/n} - \frac{1}{\mathrm{tr}\left[(XX^\top/n + \lambda_2 I)^\dagger\right]/n}\right| \frac{1}{n}\left\|\left\{(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\right\} y\right\|_1 \qquad (59)
\end{aligned}$$

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 (59) using

$$\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n + \lambda_1 I)^\dagger\right]/n} - \frac{1}{\mathrm{tr}\left[(XX^\top/n + \lambda_2 I)^\dagger\right]/n}\right| \leq 2\left|\frac{1}{\mathrm{tr}\left[(XX^\top/n + \underline{\lambda} I)^\dagger\right]/n}\right|. \qquad (60)$$

For bounding the second term of (59), note that

$$\begin{aligned}
\left\|\left\{(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\right\} y\right\|_1 / n &\leq \left\|\left\{(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\right\} y\right\|_2 / \sqrt{n}\\
&\leq \left\|(XX^\top/n + \lambda_1 I)^\dagger - (XX^\top/n + \lambda_2 I)^\dagger\right\|_{\mathrm{op}} \|y\|_2 / \sqrt{n}\\
&\leq C\,|\lambda_1 - \lambda_2|\,, \qquad (61)
\end{aligned}$$

where we used the bound from (57), 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 (60) and (61) into (59) then finishes the proof. □

D 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

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

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

$$\widehat{F}_\lambda^{\mathrm{gcv}}(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} \leq z\right\} \leq \frac{1}{n}\sum_{i=1}^n t_{z,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right). \tag{62}$$

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

$$\frac{1}{n}\sum_{i=1}^n t_{z,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) - \mathbb{E}\big[t_{z,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \xrightarrow{\text{a.s.}} 0. \tag{63}$$

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

$$\mathbb{E}\big[t_{z,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \leq \mathbb{E}\big[\mathbb{I}\{y_0 - x_0^\top \widehat{\beta}_\lambda \leq z + \epsilon\} \mid X, y\big] = \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \leq z + \epsilon \mid X, y\big]. \tag{64}$$

Thus, combining (62) to (64), we get that almost surely

$$\limsup_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \leq \limsup_{n\to\infty} \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \leq z + \epsilon \mid X, y\big] = \limsup_{n\to\infty} F_\lambda(z + \epsilon). \tag{65}$$

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

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

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

$$\widehat{F}_\lambda^{\mathrm{gcv}}(z) = \frac{1}{n}\sum_{i=1}^n \mathbb{I}\left\{\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n} \leq z\right\} \geq \frac{1}{n}\sum_{i=1}^n t_{z-\epsilon,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right). \tag{66}$$

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

$$\frac{1}{n}\sum_{i=1}^n t_{z-\epsilon,\epsilon}\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}\right) - \mathbb{E}\big[t_{z-\epsilon,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \xrightarrow{\text{a.s.}} 0. \tag{67}$$

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

$$\mathbb{E}\big[t_{z-\epsilon,\epsilon}(y_0 - x_0^\top \widehat{\beta}_\lambda) \mid X, y\big] \geq \mathbb{E}\big[\mathbb{I}\{y_0 - x_0^\top \widehat{\beta}_\lambda \leq z - \epsilon\} \mid X, y\big] = \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \leq z - \epsilon\big]. \tag{68}$$

Combining (66) to (68), we have almost surely,

$$\liminf_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \geq \liminf_{n\to\infty} \mathbb{P}\big[y_0 - x_0^\top \widehat{\beta}_\lambda \leq z - \epsilon\big] = \liminf_{n\to\infty} F_\lambda(z - \epsilon). \tag{69}$$

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

Combining (65) and (69), we conclude that almost surely $\limsup_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) - \liminf_{n\to\infty} \widehat{F}_\lambda^{\mathrm{gcv}}(z) \to 0$, and $\widehat{F}_\lambda^{\mathrm{gcv}}(z) - F(z) \to 0$, completing the proof.

E PROOFS RELATED TO Theorem 6

E.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 21, 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 $\delta > 0$,

$$\begin{aligned}
&\sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \big| T_\lambda(v_1) - T_\lambda(v_2) \big| \\
&= \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \Big| \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) \mid X, y \big] - \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \mid X, y \big] \Big| \\
&= \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \Big| \mathbb{E}\big[t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \mid X, y \big] \Big| \\
&\le \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \mathbb{E}\Big[\big| t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \big| \;\Big|\; X, y \Big] \\
&\le \mathbb{E}\left[\sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \big| t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_1) - t(y_0 - x_0^\top \widehat{\beta}_\lambda, v_2) \big| \;\Big|\; X, y \right],
\end{aligned} \tag{70}$$

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 $\delta > 0$,

$$\begin{aligned}
&\sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \big| \widehat{T}_\lambda^{\mathrm{gcv}}(v_1) - \widehat{T}_\lambda^{\mathrm{gcv}}(v_2) \big| \\
&= \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \left| \frac{1}{n} \sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1 \right) - \frac{1}{n} \sum_{i=1}^n t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2 \right) \right| \\
&\le \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \frac{1}{n} \sum_{i=1}^n \left| t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1 \right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2 \right) \right| \\
&\le \frac{1}{n} \sum_{i=1}^n \sup_{|v_1 - v_2| \le \delta,\, v_1, v_2 \in \mathcal{V}} \left| t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_1 \right) - t\left(\frac{y_i - x_i^\top \widehat{\beta}_\lambda}{1 - \mathrm{tr}[L_\lambda]/n}, v_2 \right) \right|.
\end{aligned} \tag{71}$$

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, (70) and (71) 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 21 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.

E.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.

F 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

$$x_i = \Sigma^{1/2} z_i, \tag{72}$$

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

$$y_i = \beta_0^\top x_i + \big(x_i^\top A x_i - \mathrm{tr}[A\Sigma]\big)/p + \varepsilon_i, \tag{73}$$

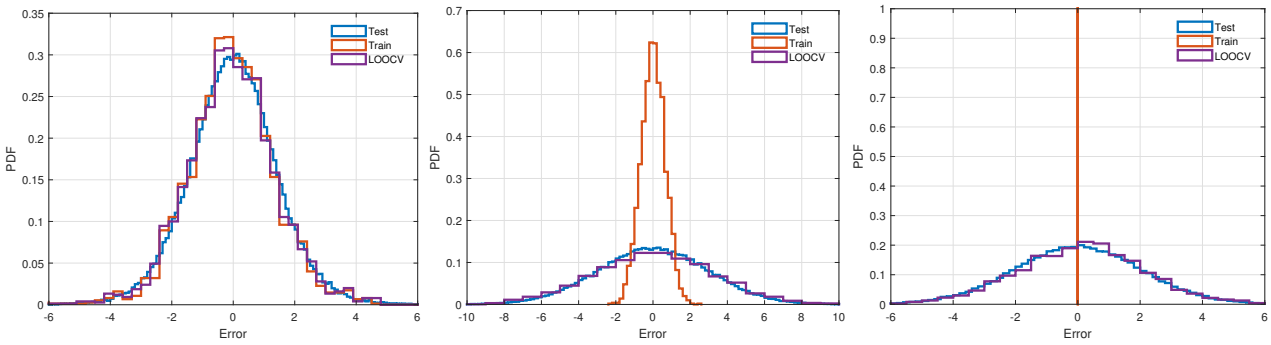
where $\beta_0 \in \mathbb{R}^p$ is a fixed signal vector, $A \in \mathbb{R}^{p\times p}$ is a fixed matrix, and $\varepsilon_i \in \mathbb{R}$ is a random noise variable. Note that we have subtracted the mean from the squared nonlinear component and scaled it to keep the variance of the nonlinear component 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.

F.1 Distribution Estimation

As promised in the paper, we first present illustrations with LOOCV reweighted errors for Figures 1 and 2 in Figures 4 and 5, respectively.

Note that both in Figures 1 and 2 in the paper, as well as Figures 4 and 5, 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 6, 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 4: 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 (72) 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 (73). 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.

F.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 7 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 8 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.

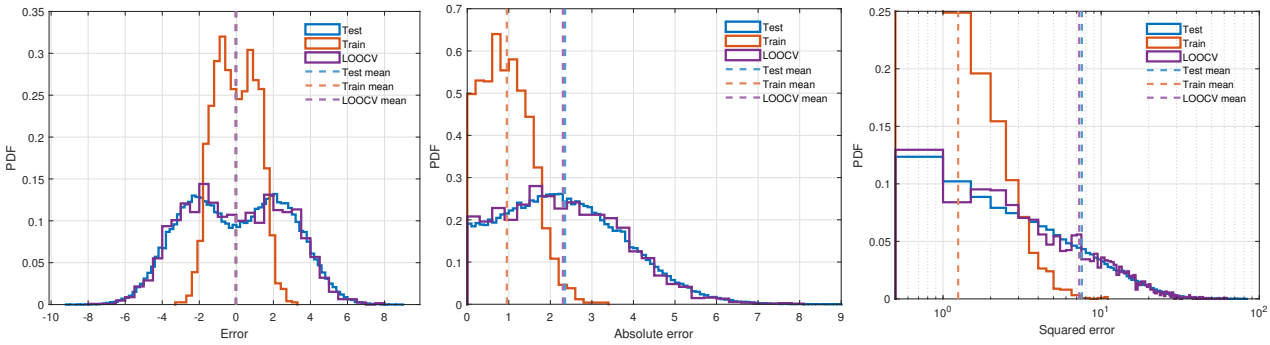
G SUPPLEMENTARY RESULTS

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).

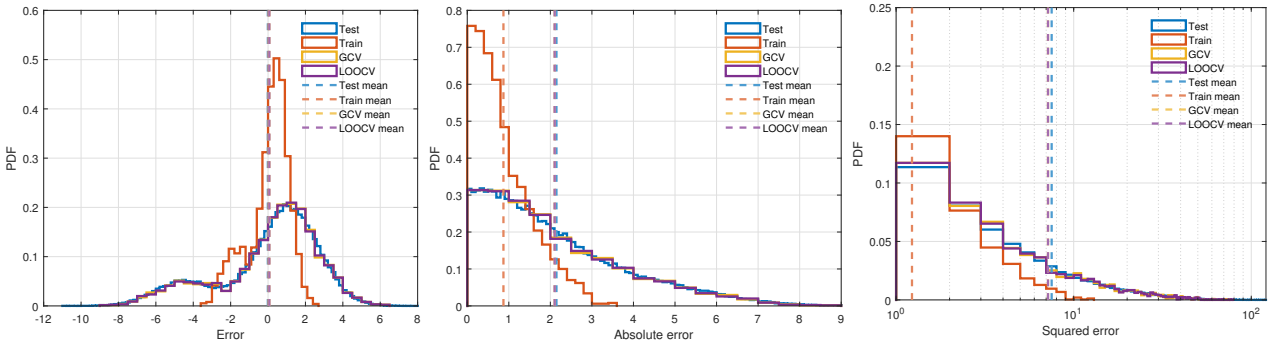
Lemma 16 (Burkholder's inequality)**.** *Let $\{Z_k\}$ be a martingale difference sequence with respect to the increasing σ-field $\{\mathcal{F}_k\}$. Then, for $q \geq 2$,*

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



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

Figure 5: An example with $n = 2500$, $p = 5000$. We generated each x_i according to (72) with identity covariance with the components of z_i sampled from a symmetric Bernoulli distribution, and each response y_i is generated according to (73). 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 6: An example with $n = 2500$, $p = 5000$. We generated each x_i according to (72) with identity covariance and components of z_i sampled from a Gaussian distribution, and each response y_i according to (73) 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 17 (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 18 (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 $q \ge 1$,*

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

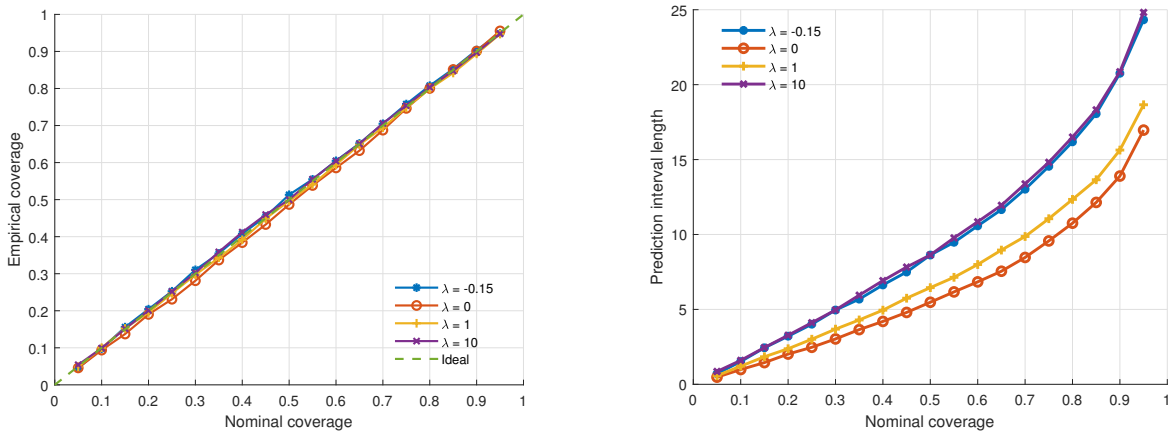


Figure 7: Illustration of empirical coverage and length of LOOCV prediction intervals constructed using (30) against nominal coverage, where $n = 2500$, $p = 5000$. We generated features x_i according to (72) with autoregressive covariance structure (with $\rho = 0.25$) and t-distributed components of z_i with 5 degrees of freedom. The responses y_i are generated according to (73) 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 7 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 19 (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 20 (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 20 is the following statement.

Lemma 21 (Convergence of minimizers)**.** *Assume the setting of Lemma 20. 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 22 (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 $n \to \infty$, $S_n \xrightarrow{\text{a.s.}} 0$.*

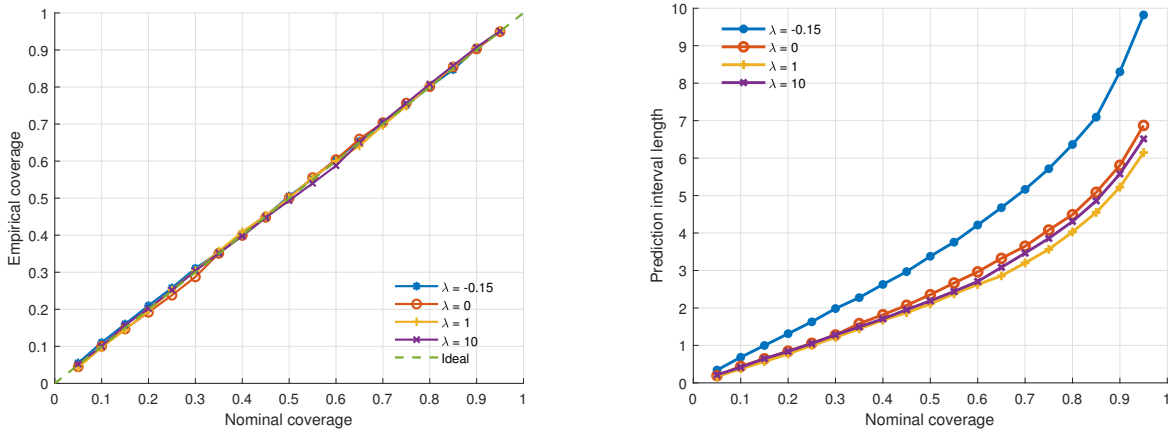


Figure 8: Illustration of empirical coverage and length of LOOCV prediction intervals (30) against nominal coverage, where $n = 2500$, $p = 5000$. The features x_i are generated according to (72) with identity covariance and components of z_i having Gaussian distribution. The responses y_i are generated according to (73) 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.