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.

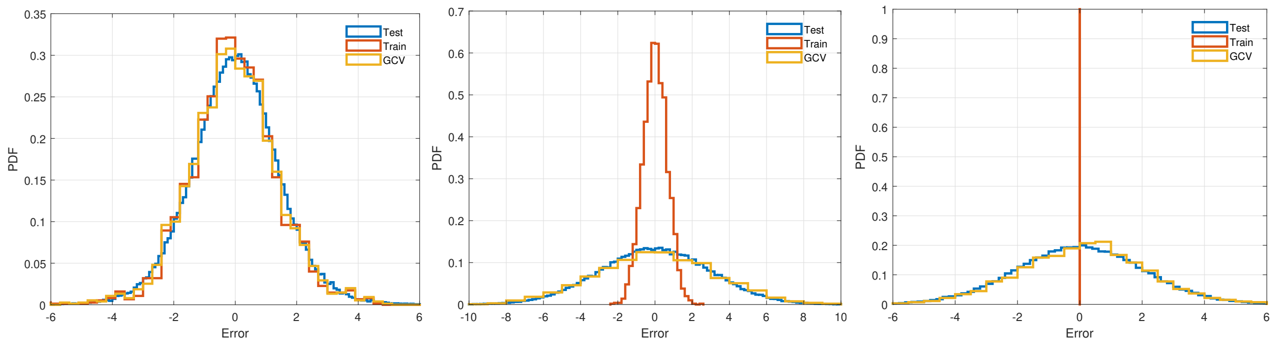
1 INTRODUCTION

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 $i = 1, \dots, n$, 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),

$i = 1, \dots, n$, 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, $i = 1, \dots, n$) 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

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

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

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

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

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

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

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

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

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

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

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

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 $\lambda > \lambda_{\min}$,*

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

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 $\widehat{P}_n \xrightarrow{\text{d}} P_n$ 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 $\sup_{z\in\mathbb{R}} |\widehat{F}_\lambda^{\text{gcv}}(z) - F_\lambda(z)| \to 0$, where F_λ and $\widehat{F}_\lambda^{\text{gcv}}$ are the distribution functions associated with P_λ and $\widehat{P}_\lambda^{\text{gcv}}$, respectively. The analogous result holds for LOOCV as well. This follows from standard arguments (e.g., Chapter 3 of Durrett, 2019), and we omit the details.

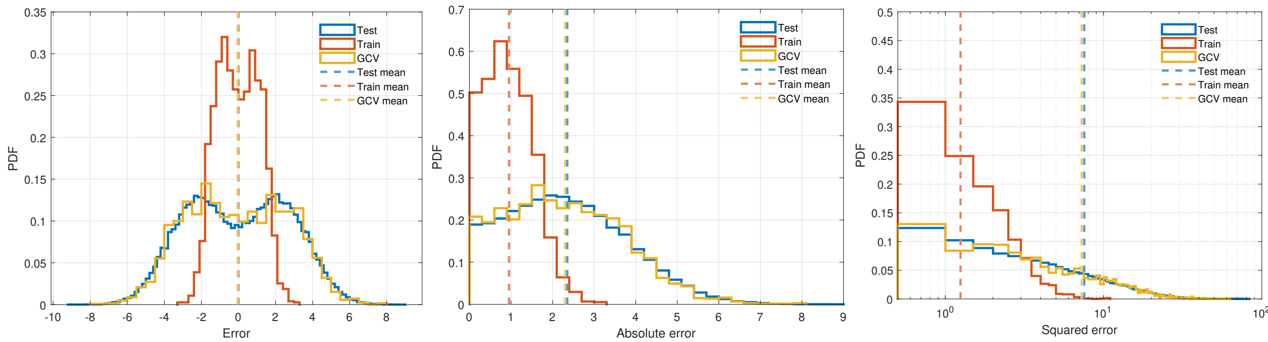
An extension (resembling the continuous mapping theorem) of Theorem 1 is given next.

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 $\widehat{H}_\lambda^{\text{gcv}}$ and $\widehat{H}_\lambda^{\text{loo}}$ 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 $\lambda > \lambda_{\min}$,*

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

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 a 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 $|t(z)| \le az^2 + b|z| + c$ 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 $\widehat{T}_\lambda^{\mathrm{gcv}}, \widehat{T}_\lambda^{\mathrm{loo}}$ 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 $\lambda > \lambda_{\min}$,*

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

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:

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

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) $\widetilde{T}_\lambda - \widehat{T}^{\text{loo}}_\lambda \xrightarrow{\text{a.s.}} 0$, and (iii) $\widehat{T}^{\text{loo}}_\lambda - \widehat{T}^{\text{gcv}}_\lambda \xrightarrow{\text{a.s.}} 0$. 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 $\Lambda \subseteq (\lambda_{\min}, \infty)$,*

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

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

$$V_\lambda = \operatorname*{arg\,min}_{v\in\mathcal{V}} T_\lambda(v). \qquad (21)$$

This is assumed to be unique.[2] Meanwhile, denoting by $\widehat{T}^{\text{gcv}}_\lambda(v)$ and $\widehat{T}^{\text{loo}}_\lambda(v)$ 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 $\lambda > \lambda_{\min}$,*

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

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:

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

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

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 $\lambda > \lambda_{\min}$,*

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

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,*

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

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

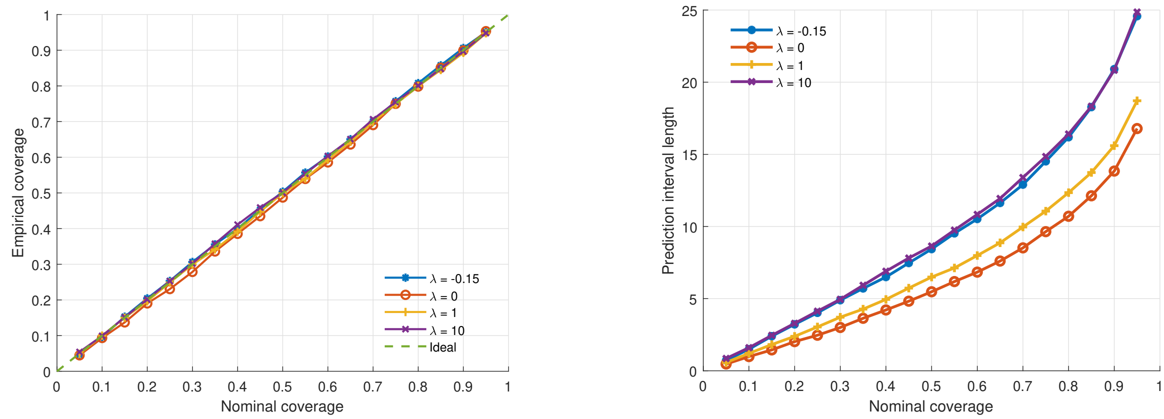


Figure 3: Illustration of empirical coverage and length of GCV prediction intervals (29) against nominal coverage, where $n = 2500$, $p = 5000$. The data model has a latent structure with autoregressive feature covariance and true signal aligned with the principal eigenvector, similar to that in Kobak 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.

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.

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

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

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 application 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.

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.

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.

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.