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

Pratik Patil **Yuting Wei** **Alessandro Rinaldo** **Ryan J. Tibshirani**
Carnegie Mellon University

Abstract

We examine generalized and leave-one-out cross-validation for ridge regression in a proportional asymptotic framework where the dimension of the feature space grows proportionally with the number of observations. Given i.i.d. samples from a linear model with an arbitrary feature covariance and a signal vector that is bounded in ℓ_2 norm, we show that generalized cross-validation for ridge regression converges almost surely to the expected out-of-sample prediction error, uniformly over a range of ridge regularization parameters that includes zero (and even negative values). We prove the analogous result for leave-one-out cross-validation. As a consequence, we show that ridge tuning via minimization of generalized or leave-one-out cross-validation asymptotically almost surely delivers the optimal level of regularization for predictive accuracy, whether it be positive, negative, or zero.

1 INTRODUCTION

Fitting high-dimensional statistical models typically requires some form of regularization, both for computational and statistical reasons. For optimization-based models, this can be achieved by adding to the data fitting objective function a tunable regularization term. The optimal level of regularization usually depends on unknown characteristics of the data generating distribution. In practice, one performs regularization tuning based on the observed data. Proper calibration of regularization can significantly affect the performance of the fitted model, and consequently proper data-dependent tuning is one of the core tasks in statistical learning.

Proceedings of the 24[th] International Conference on Artificial Intelligence and Statistics (AISTATS) 2021, San Diego, California, USA. PMLR: Volume 130.

Cross-validation (e.g., Allen, 1974; Stone, 1974; Geisser, 1975) is a widely used method for regularization tuning. While it has many variants, the most common variant is arguably k-fold cross-validation (e.g., Hastie et al., 2009; Györfi et al., 2006). Here we split the data into k "folds", leave out the first fold for model fitting so that we can use it to assess the out-of-sample performance of the fitted model, then we leave out the second fold, and so on. By aggregating the errors made across the k folds, we produce a final estimate of the expected out-of-sample error profile as a function of regularization level, and select the regularization level by minimizing of the cross-validated error profile.

While a typical choice of k is 5 or 10, such a choice of can suffer from high bias in high-dimensional problems. Setting $k = n$, the number of observations, leads to a variant called leave-one-out cross-validation (LOOCV). This alleviates the bias issues but it is computationally expensive in general, requiring n model fits. Despite recent important advances in the theoretical study of LOOCV and its various approximations in high dimensions (including Kale et al., 2011; Kumar et al., 2013; Meijer and Goeman, 2013; Obuchi and Kabashima, 2016; Miolane and Montanari, 2018; Wang et al., 2018; Xu et al., 2019; Stephenson and Broderick, 2020; Wilson et al., 2020; Celentano et al., 2020), the theoretical understanding of these methods, especially statistical properties of the tuned estimators under general distributional assumptions, is still incomplete.

In this paper, we focus on ridge regression (Hoerl and Kennard, 1970), a widely-used estimator in statistics that entails fitting linear regression with ℓ_2 regularization. We consider two commonly used cross-validation procedures, LOOCV and an approximation to LOOCV called generalized cross-validation (GCV) (Golub et al., 1979; Wahba, 1980, 1990). For ridge regression, both procedures can be computed efficiently—in a manner that requires no model refitting whatsoever—and are popular choices in practice. Our main goal is to investigate the theoretical behavior of ridge regression when tuned using one of these cross-validation methods.

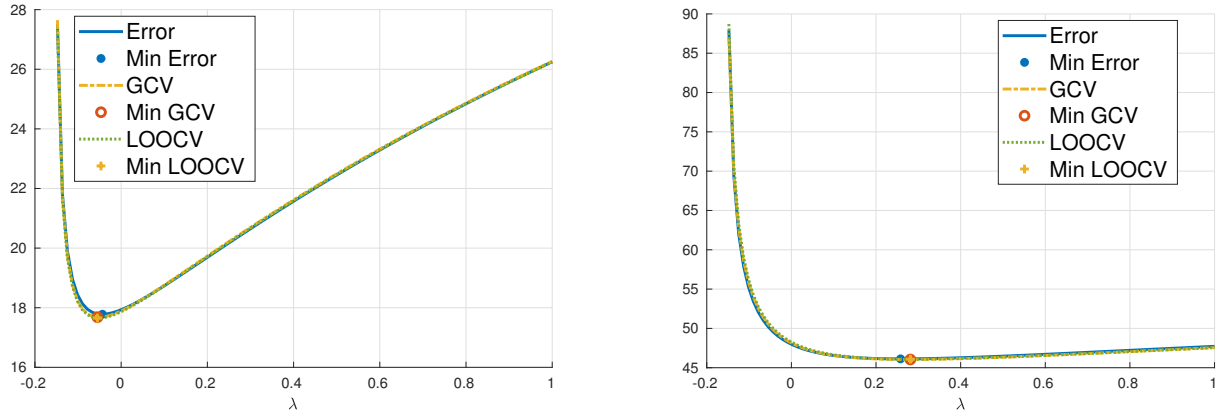

Figure 1: Comparison of the GCV and LOOCV estimates of the expected out-of-sample prediction error for ridge regression as a function of the regularization parameter λ. We consider an overparametrized regime where the number of observations is $n = 6000$ and the number of features is $p = 12000$. The features are random with a ρ-autoregressive covariance Σ (such that $\Sigma_{ij} = \rho^{|i-j|}$ for all i, j) with $\rho = 0.25$. The response is generated from a linear model with a nonrandom signal vector β_0. In the left figure, the signal is aligned with the eigenvector corresponding to the largest eigenvalue of Σ, while in the right figure, the signal is aligned with the eigenvector corresponding to the smallest eigenvalue. The effective signal-to-noise ratio is set to $\beta_0^T \Sigma \beta_0 = 60$ to illustrate that, in the overparametrized regime, the optimal regularization could be negative or positive depending on how the signal aligns with the covariance eigenstructure. Note that in both the cases, the GCV and LOOCV curves track the prediction error over the whole range of λ very closely. The optimal regularization is recovered very well by the GCV and LOOCV estimates in both cases.

For our theoretical analysis, we adopt a proportional asymptotic framework in which the number of features grows linearly with the number of observations (that is, their ratio converges to a constant). We show that both the GCV and LOOCV error curves, as functions of the ridge regularization parameter, converge uniformly almost surely to the expected out-of-sample prediction error curve. Our results hold under weaker assumptions on the data generating distribution compared to others in the literature thus far, and provide a rigorous theoretical justification for the use of both GCV and LOOCV for regularization tuning for ridge regression in high dimensions. Below we summarize our main contributions, and illustrate key points with a numerical example in Figure 1.

1. GCV pointwise convergence. Given n i.i.d. samples from a standard linear model $y = x^T \beta_0 + \varepsilon$, where x is p-dimensional feature such that $x = \Sigma^{1/2} z$ for a covariance matrix Σ, and z having i.i.d. entries, we establish the equivalence of the GCV estimator and the expected out-of-sample prediction error for ridge regression, under proportional asymptotics (p/n converging to a constant). This result holds for an arbitrary sequence of covariance matrices Σ with eigenvalues bounded away from zero and infinity, and an arbitrary sequence of signal vectors β_0 with bounded ℓ_2 norm.

2. GCV uniform convergence. Moreover, we show that this GCV convergence holds uniformly over compact intervals of the regularization parameter λ that include zero and negative regularization.

3. LOOCV convergences. We establish the analogous properties (pointwise and uniform convergence) for the LOOCV estimator by relating it to GCV.

4. Optimal tuning. As a direct consequence of uniform convergence, we demonstrate that the level of regularization chosen based on either of the GCV or LOOCV estimators almost surely delivers a limiting prediction accuracy that an oracle with full knowledge of the out-of-sample prediction error curve would achieve. Thus, in this sense, both methods are asymptotically optimal for tuning the prediction error of ridge regression.

2 RELATED WORK

Ridge Error Analysis. The predictive performance of ridge regression has been studied comprehensively in various settings, both asymptotic and non-asymptotic; see, e.g., Hsu et al. (2012); Karoui (2013); Dicker (2016); Dobriban and Wager (2018). More recently, there has been a surge of interest in understanding its prediction error driven by the successes of interpolating models in high dimensions; e.g., Hastie et al. (2019); Mei and Montanari (2019); Wu and Xu (2020); Richards et al. (2020); Tsigler and Bartlett (2020). Interestingly, Wu

and Xu (2020); Richards et al. (2020) study the nature of optimal regularization and provide conditions on the feature covariance and signal structure that result in a positive or negative level of optimal regularization.

Ridge Cross-Validation. In the low-dimensional setting, the consistency of LOOCV and GCV for ridge regression error estimation and regularization tuning has been established in Stone (1974, 1977); Craven and Wahba (1979); Li (1985, 1986, 1987); Dudoit and van der Laan (2005), among others. More recently, statistical and computational aspects of cross-validation for regularized estimators in high dimensions have also been thoroughly studied; see, e.g., Beirami et al. (2017); Rad and Maleki (2018); Wang et al. (2018); Xu et al. (2019); Rad et al. (2020); Austern and Zhou (2020).

Most similar to our work in this paper is probably the result of Hastie et al. (2019) on the asymptotic optimality of LOOCV and GCV tuning for ridge regression in high dimensions. These authors also adopt a proportional asymptotic model, but use stronger assumptions on the data generating distribution: they assume $\Sigma = I$ (independent features) and that the signal β_0 is drawn from a spherical prior (taking a Bayesian view). Under these conditions, the optimal level of regularization is always positive. We significantly generalize the scope of this analysis by allowing for *arbitrary* Σ and *nonrandom* β_0, in which case the optimal regularization level can be positive, negative, or zero.

Our Work. We highlight the main contributions of our paper below.

Analyzing differences. We do not seek to characterize the limiting risk (we will use the terms risk and prediction error interchangeably), but instead, we analyze the limiting differences between the LOOCV and GCV estimators and the risk, and show that these differences tend to zero. As such, we are able to work in a general regime where it may not even be possible to precisely characterize the limiting risk in the first place.

Conditional statements. Our theory is all conditional on the training data $\{(x_i, y_i)\}_{i=1}^n$ (results hold almost surely with respect to the draws from the training distribution). Most other papers provide cross-validation results that hold in an integrated sense over the training data. Our conditional setup allows for stronger statements about tuning based on the observed data rather than in an average sense.

Direct analysis of GCV. Most previous papers rely on the stability of some estimator in question to establish the properties of LOOCV, while we directly tackle the explicit forms of prediction error and GCV, and derive a crucial empirical equivalence lemma to first tie the risk to GCV, and then GCV to LOOCV.

Uniform convergence. To analyze the cross-validation-tuned risks, we establish uniform convergence results, by leveraging the explicit form of the ridge estimator. This aspect has not been focused on in previous cross-validation work, except Hastie et al. (2019).

Proof technique. To reiterate what was mentioned earlier, in comparison to Hastie et al. (2019) (who take $\Sigma = I$ and β_0 drawn from a prior), we allow Σ and β_0 to be essentially arbitrary, only requiring Σ to have bounded eigenvalues and β_0 to have bounded ℓ_2 norm. While the flavor of final results is similar to those in Hastie et al. (2019), the proof techniques are different. We isolate the individual equivalences for the bias- and variance-like components in the GCV and LOOCV estimators, which helps shed light into the structure underlying the overall combined equivalence. Further, we derive (and rely extensively on) an equivalence that relates certain functionals involving the sample covariance $\widehat{\Sigma}$ and population covariance Σ, in a proportional asymptotic setup. This is in a sense much simpler than the approach taken in Hastie et al. (2019), which relies on equating certain limiting formulae that arise from studying GCV, LOOCV, and ridge risk (equating such formulae involves difficult and unintuitive manipulations with Stieltjes transforms).

Result utility. Recently, it has been observed that models with very small or even zero regularization can generalize well in certain overparametrized settings (e.g., Zhang et al., 2017; Belkin et al., 2019). This is also the case with ridge regression where the optimal level of regularization can be zero or even negative (Kobak et al., 2020; Richards et al., 2020; Wu and Xu, 2020). Certain nontrivial interactions between the properties of the signal and feature distributions is what leads to these recent surprises. Our framework automatically accommodates these cases and affirms that that GCV and LOOCV can indeed pick risk-optimal interpolators when they need to.

3 PROBLEM SETUP

We consider the standard regression setting in which we observe n i.i.d. pairs $\{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathbb{R}^p$ is the i^{th} feature vector and $y_i \in \mathbb{R}$ is the corresponding response variable. In matrix notation, we denote by $X \in \mathbb{R}^{n \times p}$ the feature matrix whose i^{th} row is x_i^T and by $y \in \mathbb{R}^n$ the response vector whose i^{th} entry is y_i.

Extended Ridge Regression. For a regularization parameter $\lambda > 0$, the ridge regression estimate $\widehat{\beta}_\lambda \in \mathbb{R}^p$ based on features X and response y can be formulated

as the solution to the convex optimization problem

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

The can be explicitly written as

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

where $I_p \in \mathbb{R}^{p\times p}$ is the identity matrix. To allow for an extended range of λ (including $\lambda = 0$), we simply define the extended ridge regression estimate as

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

Here A^+ denotes the Moore-Penrose pseudoinverse of a matrix A. Note this definition allows for any $\lambda \in \mathbb{R}$. For $\lambda > 0$, there is no difference between (1) and the usual definition of ridge (second to last display). For $\lambda = 0$, we can see that (1) reduces to the least squares solution that lies in the row space of X, and hence has minimum ℓ_2 norm among all least squares solutions. Of particular interest is when $\text{rank}(X) = n < p$: then it reduces to the least squares solution that interpolates the data ($X\widehat{\beta}_\lambda = y$), and has minimum ℓ_2 norm among all such interpolators.

Prediction Error. The expected out-of-sample prediction error (or risk) of the ridge model $\widehat{\beta}_\lambda$ is defined as

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

Here the expectation is taken with respect to the distribution of a new test pair (x_0, y_0) sampled from the same distribution as the training data $\{(x_i, y_i)\}_{i=1}^n$, and independent of the training data. The prediction error is a random variable (it is conditional on—and thus a function of—X, y) that quantifies how well a given fitted ridge model $\widehat{\beta}_\lambda$ performs in the task of predicting the response.

The prediction error as a function of the regularization parameter λ yields an error curve that we denote by

$$\text{err}(\lambda) = \text{Err}(\widehat{\beta}_\lambda).$$

As far as we are concerned in this paper, the optimal regularization parameter is defined as the value that minimizes the risk curve $\text{err}(\lambda)$. This is the value of λ that an oracle with knowledge of the risk curve would pick. We seek to construct a faithful estimate of the risk curve $\text{err}(\lambda)$ based on the available data X and y, uniformly over λ, in order to select the regularization level that leads to prediction error close to that of the oracle prediction error. To do so, we will consider LOOCV and GCV whose definitions we recall next.

LOOCV and GCV. The LOOCV estimate for the risk of a given ridge model $\widehat{\beta}_\lambda$ is defined as

$$\text{loo}(\lambda) = \frac{1}{n}\sum_{i=1}^n \Big(y_i - x_i^T\widehat{\beta}_{-i,\lambda}\Big)^2,$$

where $\widehat{\beta}_{-i,\lambda} = (X_{-i}^T X_{-i}/n + \lambda I_p)^+ X_{-i}^T y_{-i}/n$ denotes the ridge estimate with the i^{th} observation pair (x_i, y_i) excluded from the training set. Computing the LOOCV estimate with this definition requires (re)fitting ridge model n times. Recall that ridge regression is a linear smoother, $X\widehat{\beta}_\lambda = L_\lambda y$, where the smoothing matrix $L_\lambda \in \mathbb{R}^{n\times n}$ is

$$L_\lambda = X(X^T X/n + \lambda I_p)^+ X^T/n, \tag{3}$$

Fortunately, there is a so-called shortcut formula for the LOOCV estimate (see, e.g., Chapter 7 of Hastie et al., 2009):

$$\text{loo}(\lambda) = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - [L_\lambda]_{ii}}\right)^2, \tag{4}$$

where $[L_\lambda]_{ii}$ denotes the i^{th} diagonal element of L_λ.

The GCV estimate is a further convenient approximation to the LOOCV shortcut formula (4) given by

$$\text{gcv}(\lambda) = \frac{1}{n}\sum_{i=1}^n \left(\frac{y_i - x_i^T\widehat{\beta}_\lambda}{1 - \text{tr}[L_\lambda]/n}\right)^2, \tag{5}$$

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

Caution needs to be taken when the smoothing matrix L_λ reduces to the identity matrix I_n, or in other words, ridge regression is an interpolator, with $X\widehat{\beta}_\lambda = y$. This happens when $\lambda = 0$ and X has rank n. In this case, both the numerators and denominators of $\text{loo}(\lambda)$ and $\text{gcv}(\lambda)$ are 0, however, we can define the corresponding LOOCV and GCV estimates as their respective limits as $\lambda \to 0$; see Hastie et al. (2019) for details.

Goal of This Paper. Our main goal is to analyze the differences between the cross-validation estimators of risk and the risk itself, $\text{loo}(\lambda) - \text{err}(\lambda)$ and $\text{gcv}(\lambda) - \text{err}(\lambda)$. Let $\lambda_I^\star$ denote the optimal oracle ride tuning parameter that minimizes $\text{err}(\lambda)$ over an interval $I \subseteq \mathbb{R}$,

$$\lambda_I^\star = \underset{\lambda\in I}{\arg\min}\ \text{err}(\lambda).$$

(If there are multiple minimizers, simply let $\lambda_I^\star$ denote one of them.) Similarly, let $\widehat{\lambda}_I^{\text{gcv}}$ and $\widehat{\lambda}_I^{\text{loo}}$ be the corresponding tuning parameters that minimize GCV and LOOCV over $\lambda \in I$. We seek to compare the prediction errors of the models tuned using GCV and LOOCV, $\text{Err}(\widehat{\beta}_{\widehat{\lambda}_I^{\text{gcv}}})$ and $\text{Err}(\widehat{\beta}_{\widehat{\lambda}_I^{\text{loo}}})$, against the prediction error under oracle tuning, $\text{Err}(\widehat{\beta}_{\lambda_I^\star})$.

4 MAIN RESULTS

In this section, we state and discuss our main results. We first list the required assumptions in Section 4.1. In Section 4.2, we state the limiting equivalence between the GCV estimator and prediction risk, followed by the limiting equivalence between the LOOCV and GCV estimators in Section 4.3.

4.1 Assumptions

We begin by stating the assumptions we impose on the structure of response and feature distributions.

Assumption 1 (Response distribution)**.** There exists a signal vector $\beta_0 \in \mathbb{R}^p$ such that $y = X\beta_0 + \varepsilon$, where the noise vector $\varepsilon = (\varepsilon_1, \ldots, \varepsilon_n) \in \mathbb{R}^n$ is independent of X, and its components are i.i.d. with mean 0, variance σ^2, and finite $4+\eta$ moment for some $\eta > 0$.

Assumption 2 (Feature distribution)**.** The feature vectors (rows of X) can be decomposed as $x = \Sigma^{1/2} z$, where $\Sigma \in \mathbb{R}^{p\times p}$ is a deterministic positive definite matrix, and $z \in \mathbb{R}^p$ is a random vector whose components are i.i.d. with mean zero 0, variance 1, and finite $4+\eta$ moment for some $\eta > 0$.

We consider a proportional asymptotic framework in which the number of features p grows with the number of observations n in such a way that their ratio p/n approaches a constant $\gamma \in (0, \infty)$. Accordingly, in our asymptotic analysis, we must deal with a sequence of feature covariance matrices Σ and signal vectors β_0. (For ease of readability, we do not make the dependence of these quantities and many others on p explicit in our notation.) We make the following assumptions on the eigenvalues of Σ and the signal energy.

Assumption 3 (Extreme eigenvalues of Σ)**.** The maximum and minimum eigenvalues of Σ are upper and lower bounded by constants $r_{\max} < \infty$ and $r_{\min} > 0$, respectively, independent of p.

The lower bound $r_{\min}$ on the minimal eigenvalue of Σ is that it will determine, asymptotically, the smallest possible value of the regularization parameter for which our results hold. We denote it by $\lambda_{\min} = -(\sqrt{\gamma} - 1)^2 r_{\min}$.

Assumption 4 (Signal energy)**.** The signal energy $\|\beta_0\|_2^2$ is upper bounded by a constant $\tau < \infty$ independent of p.

We note that it should be possible to relax the assumptions on the maximum and minimum eigenvalues of Σ, to allow a certain fraction of eigenvalues to diverge and others to accumulate near zero. We leave such an extension to future work.

4.2 GCV Versus Prediction Error

We are ready to state our first result comparing the GCV estimator to prediction error of ridge regression.

Theorem 4.1 (GCV equals prediction error in limit)**.** *Under Assumptions 1 to 4, for every $\lambda \in (\lambda_{\min}, \infty)$, it holds that*

$$\mathrm{gcv}(\lambda) - \mathrm{err}(\lambda) \xrightarrow{a.s.} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Furthermore, the convergence is uniform in λ over compact subintervals $I \subseteq (\lambda_{\min}, \infty)$; consequently, for any such interval I,

$$\mathrm{Err}(\widehat{\beta}_{\widehat{\lambda}_I^{\mathrm{gcv}}}) - \mathrm{Err}(\widehat{\beta}_{\lambda_I^\star}) \xrightarrow{a.s.} 0,$$

where $\widehat{\lambda}_I^{\mathrm{gcv}}$ and $\lambda_I^\star$ are the corresponding optimal GCV and prediction error tuning parameters, respectively.

We note that in this and in all the other asymptotic statements in the paper, the almost sure qualification refers to the randomness in both X and y.

Range of λ. The lower limit $\lambda_{\min}$ in Theorem 4.1 is used to ensure that the resulting smoothing matrix L_λ stays positive semidefinite; this is simply a function of the behavior of the minimum non-zero eigenvalue of the sample covariance matrix $\widehat{\Sigma}$ (see Bai and Silverstein, 1998).

Note that this range of λ allows for potentially negative regularization (when $\gamma \neq 1$), including zero; the latter case, in particular, results in the least squares interpolator when $p > n$. The fact that GCV works in this case is interesting because both the numerator and denominator in the expression (5) for $\mathrm{gcv}(\lambda)$ are 0—implying the particular form of the ridge estimator somehow preserves the information about the predictive performance in the GCV limit even when the training error is 0.

The statement in Theorem 4.1 does not cover the behavior of GCV at the endpoints $\lambda = \lambda_{\min}$ and $\lambda \to \infty$. In fact, it is easy to check that the limiting behavior of GCV and prediction error matches at these endpoints as well. In particular, under the same assumptions as the theorem, if $r_{\min}$ is the limit inferior of minimum eigenvalues of the Σ sequence, then indeed both

$$\mathrm{gcv}(\lambda_{\min}) \to \infty \quad \text{and} \quad \mathrm{err}(\lambda_{\min}) \to \infty$$

as $n, p \to \infty$ with $p/n \to \gamma$. Similarly, both

$$\mathrm{gcv}(\lambda) \to c^2 \quad \text{and} \quad \mathrm{err}(\lambda) \to c^2$$

as $\lambda \to \infty$ and $n, p \to \infty$ with $p/n \to \gamma$, where $c^2 = \mathbb{E}[y_0^2]$ is the prediction error of the null estimator. In this regard, the pointwise equivalence between GCV and prediction error extends to the entire range of λ.

4.3 LOOCV Versus GCV

As a byproduct of our analysis, we establish a limiting equivalence between the LOOCV and GCV estimators. This implies a limiting equivalence between LOOCV and prediction error.

Theorem 4.2 (LOOCV equals GCV in limit)**.** *If the components of the response vector $y \in \mathbb{R}^n$ have mean zero and finite second moment, and Assumptions 2 to 3 hold, then for every $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{loo}(\lambda) - \mathrm{gcv}(\lambda) \xrightarrow{a.s.} 0$$

as $n, p \to \infty$ with $p/n \to \gamma \in (0, \infty)$. Furthermore, the convergence is uniform in λ over compact subintervals $I \subseteq (\lambda_{\min}, \infty)$.

It is worth pointing out that, compared to Theorem 4.1, the last guarantee only requires that the response variables have a finite second moment. In particular, it does not postulate a linear model. So the equivalence between the GCV and LOOCV estimators holds even when the model is misspecified.

In general, the analysis of LOOCV is challenging because of complex dependencies between its summands. Fortunately, for ridge regression, the equivalent shortcut expression given in (4) for the LOOCV estimate simplifies such dependence. Unlike GCV in (5), which weights training errors by $1 - \mathrm{tr}[L_\lambda]/n$, the shortcut expression for LOOCV weights the i^{th} training error by $1 - [L_\lambda]_{ii}$. Theorem 4.1 effectively shows that this different reweighting does not affect the limiting behavior, providing a way to directly tie GCV to LOOCV.

An important consequence of the last theorem is the following.

Corollary 4.3 (LOOCV equals prediction error in limit)**.** *Under the assumptions as Theorem 4.1, the same results hold but for LOOCV in place of GCV.*

(The same remarks about the range of λ that were made following the GCV theorem also apply here.)

In light of this corollary, we conclude that both the GCV and the LOOCV estimators are uniformly close to the true risk in the limit. Thus regularization tuning using either method will be asymptotically optimal for ridge regression.

5 PROOF OUTLINES

In this section, we outline the main ideas behind the proofs of Theorem 4.1 and Theorem 4.2. The complete proofs are provided in the supplement.

5.1 GCV Versus Prediction Error

The proof of Theorem 4.2 involves two steps. In the first step, we decompose both the prediction error and the GCV estimator into asymptotic bias- and variance-like components as summarized in Lemma 5.1 and Lemma 5.2. In the second step, we establish limiting equivalences for both the bias and variance components as summarized in Lemma 5.3 and Lemma 5.4. The key reason why the limiting bias-variance equivalences hold is a certain property obeyed by the denominator of GCV as elucidated in Lemma 5.5.

Prediction Error Decomposition. We begin with a familiar asymptotic bias-variance decomposition for the prediction risk. For convenience, let $\widehat{\Sigma} = X^T X/n$ denote the sample covariance matrix. Also, define bias- and variance-like components as follows:

$$\mathrm{err}_b(\lambda) = \beta_0^T \big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\Sigma\big(I_p - \widehat{\Sigma}(\widehat{\Sigma}+\lambda I_p)^+\big)\beta_0,$$

$$\mathrm{err}_v(\lambda) = \frac{\varepsilon^T}{\sqrt{n}}\left(\frac{X(\widehat{\Sigma}+\lambda I_p)^+\Sigma(\widehat{\Sigma}+\lambda I_p)^+X^T}{n}\right)\frac{\varepsilon}{\sqrt{n}} + \sigma^2.$$

The decomposition of the prediction error can now be summarized as follows.

Lemma 5.1 (Error bias-variance decomposition)**.** *Under Assumptions 1 to 4, for every $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{err}(\lambda) - \mathrm{err}_b(\lambda) - \mathrm{err}_v(\lambda) \xrightarrow{a.s.} 0$$

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

GCV Decomposition. We decompose GCV into terms that mimic the bias- and variance-like terms in the decomposition for the risk. For $\lambda \neq 0$, define GCV bias- and variance-like components as follows:

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

$$\mathrm{gcv}_v(\lambda) = \frac{\varepsilon^T}{\sqrt{n}}\left(I_n - \frac{X(\widehat{\Sigma}+\lambda I_p)^+X^T}{n}\right)^2\frac{\varepsilon}{\sqrt{n}}.$$

Additionally, write the GCV denominator as:

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

When $\lambda = 0$, the corresponding quantities after taking the limit $\lambda \to 0$ the take the form:

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

$$\mathrm{gcv}_v(0) = \frac{\varepsilon^T}{\sqrt{n}}\big(\widehat{\Sigma}^+\big)^2\frac{\varepsilon}{\sqrt{n}},$$

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

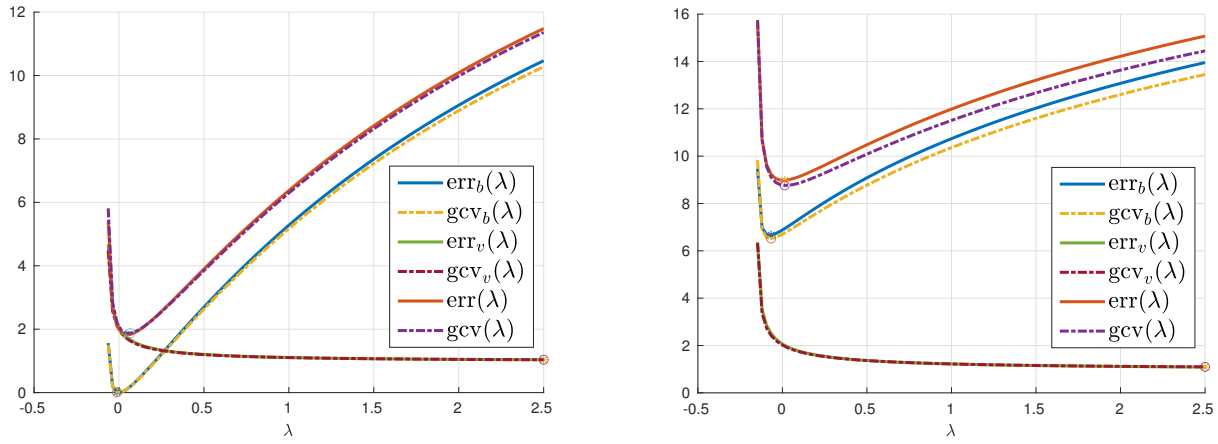

Figure 2: Comparison of the bias and variance decompositions of the GCV estimate and the prediction error. Similar to Figure 1, the features are random from a ρ-autoregressive covariance matrix Σ with $\rho = 0.25$. The response is generated from a linear model where the signal is nonrandom and aligned with the principle eigenvector of Σ. The effective signal-to-noise ratio is $\beta_0^T \Sigma \beta_0 = 25$. The left figure illustrates an underparametrized regime (with $n = 6000$ and $p = 3000$ such that $\gamma = 0.5$) while the right illustrates an overparametrized regime (with $n = 6000$ and $p = 12000$ such that $\gamma = 2$). In both cases, the bias-variance-like components of the GCV risk estimate track the bias-variance components in the prediction risk over the entire range of λ very well. In the underparametrized regime, the bias of the prediction risk is 0 at $\lambda = 0$ and increases on either sides when $\lambda \neq 0$, while the variance always decreases as λ increases (from the most negative allowed λ), resulting in a positive optimal regularization. On the other hand, in the overparametrized regime, the bias is no longer minimized at $\lambda = 0$, but at a negative λ, while the variance is again a decreasing function of λ. Since the bias dominates the total prediction risk, it results in negative optimal regularization.

(We remark that the limiting expressions for the bias- and variance-like components and the denominator for the $\lambda = 0$ case can alternately be written in terms of the gram matrix XX^T/n. The representation in terms of the sample covariance matrix $\widehat{\Sigma}$ is for consistency with the $\lambda \neq 0$ case.)

Next we establish the decomposition of GCV into bias- and variance-like quantities.

Lemma 5.2 (GCV bias-variance decomposition)**.** *Under Assumptions 1 to 4, for every $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{gcv}(\lambda) - \frac{\mathrm{gcv}_b(\lambda) + \mathrm{gcv}_v(\lambda)}{\mathrm{gcv}_d(\lambda)} \xrightarrow{a.s.} 0$$

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

Bias-Variance Equivalences. The two bias terms $\mathrm{err}_b(\lambda)$ and $\mathrm{gcv}_b(\lambda)$ differ in the sense that the latter has the unknown Σ replaced by its natural plug-in estimator $\widehat{\Sigma}$ and a rescaling by the denominator $\mathrm{gcv}_d(\lambda)$. The difference between the variance terms is analogous, albeit slightly more involved. For both, the denominator adjustment, which can be thought of as a correction for optimism in the training error by an number of effective degrees of freedom, turns out to be critical. Indeed, it is only through this normalization that $\mathrm{gcv}_b(\lambda)$ and $\mathrm{gcv}_b(\lambda)$ become consistent estimators of their population counterparts, as summarized next and illustrated in Figure 2.

Lemma 5.3 (Bias equivalence)**.** *Under Assumptions 2 to 4, for $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{err}_b(\lambda) - \frac{\mathrm{gcv}_b(\lambda)}{\mathrm{gcv}_d(\lambda)} \xrightarrow{a.s.} 0$$

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

Lemma 5.4 (Variance equivalence)**.** *Under Assumptions 1 to 3, for $\lambda \in (\lambda_{\min}, \infty)$,*

$$\mathrm{err}_v(\lambda) - \frac{\mathrm{gcv}_v(\lambda)}{\mathrm{gcv}_d(\lambda)} \xrightarrow{a.s.} 0$$

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

Basic GCV Equivalence. At the heart of why the rescaling in the GCV bias and variance-like terms yields consistency is a certain asymptotic equivalence of random matrices as summarized below.

Lemma 5.5 (Basic GCV equivalence)**.** *Under Assumption 2 and Assumption 3, for any sequence of matrices $B_p \in \mathbb{R}^{p \times p}$ that are bounded in trace norm (indepen-*

dent of p), and for $\lambda \in (\lambda_{min}, \infty) \setminus \{0\}$*, it holds that*

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

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

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

Finally, to prove uniform convergence in λ, we show that both the prediction risk $\operatorname{err}(\lambda)$ and the GCV estimator $\operatorname{gcv}(\lambda)$, and their derivatives, as functions of λ, are uniformly bounded over compact subintervals of $(\lambda_{\min}, \infty)$. This yields equicontinuity of the family of functions $\lambda \to \operatorname{err}(\lambda)$ and $\lambda \to \operatorname{gcv}(\lambda)$ almost surely and the result then follows from an application of the Arzela-Ascoli theorem. The uniform convergence subsequently leads to the convergence of the tuned errors.

5.2 LOOCV Versus GCV

There are two steps involved in establishing the limiting equivalence between LOOCV and GCV. The first is to show that the LOOCV estimator in the limit is equal to a scalar corrected factor of the training error. The second is that the correction happens to match with the factor that appears in the GCV estimator in the limit. The following lemma provides the LOOCV limit.

Lemma 5.6 (LOOCV limit as rescaled train error)**.** *If the components of the response* $y \in \mathbb{R}^n$ *have mean zero and finite second moment, and Assumptions 2 to 3 hold, then for every* $\lambda \in (\lambda_{\min}, \infty)$*,*

$$\operatorname{loo}(\lambda) - \Big(1+\operatorname{tr}[(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma]/n\Big)^2 \frac{1}{n}\sum_{i=1}^{n}(y_i - x_i^T\widehat{\beta}_\lambda)^2 \xrightarrow{a.s.} 0$$

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

The limiting equivalence then follows by tying the scale factor in the GCV estimator to the scale factor in the limiting LOOCV using an instantiation of Lemma 5.5.

6 DISCUSSION

In this paper, we established uniform consistency of the GCV and LOOCV estimators for ridge regression prediction error under a proportional asymptotic framework. At a high level, the key reason why the limiting equivalences hold is a certain asymptotic equivalence of random matrices, where on one side we have a quantity that involves both the feature covariance Σ and the sample covariance $\widehat{\Sigma}$, while on the other side, we have a quantity that only involves $\widehat{\Sigma}$, appropriately normalized. That is,

$$(\widehat{\Sigma}+\lambda I_p)^{+}\Sigma \asymp \frac{(\widehat{\Sigma}+\lambda I_p)^{+}\widehat{\Sigma}}{1-\operatorname{tr}[(\widehat{\Sigma}+\lambda I_p)^{+}\widehat{\Sigma}]/n}$$

where for two sequences of matrices A_p and B_p, $A_p \asymp B_p$ is used to mean that $\lim_{n\to\infty} \operatorname{tr}[C_pA_p] - \operatorname{tr}[C_pB_p] = 0$ almost surely for any sequence of deterministic matrices C_p of bounded trace norm.

A similar notion of equivalence has appeared in the random matrix theory literature (e.g., Serdobolskii, 1983; Silverstein and Choi, 1995; Hachem et al., 2007; Ledoit and Peche, 2011; Rubio and Mestre, 2011; Couillet and Debbah, 2011), and recently, has been utilized and developed further in Dobriban and Sheng (2018, 2020). Our work takes a slightly differently approach in that, instead of expressing the resolvents in terms of limits of unknown population quantities (which has been called a *deterministic equivalence*), we relate two sets of resolvents, neither of which needs to have a computable asymptotic limit in the first place.

For statistical applications, we believe this could have broad utility because it allows to tie potentially interesting out-of-sample quantities to purely data-dependent quantities. For example, it should be possible to asymptotically equate more general functionals involving Σ and $\widehat{\Sigma}$ in terms of $\widehat{\Sigma}$ alone. Exploring such connections for both a wider class of statistical problems and for metrics other than the expected out-of-sample error is a future direction.

Beyond asymptotics, it is also of interest to carry out a finite sample analysis that explicitly reveals how the interaction between the signal vector and the feature covariance affects rates of convergence. This may, for example, facilitate constructions of confidence intervals for the tuned parameters. It may also reveal that GCV and LOOCV—though consistent across a very broad set of problem settings, as demonstrated in this paper—can struggle in terms of their speed of convergence for certain problems, like (say) when the optimal regularization parameter is around zero. Finally, the assumptions on the feature and response distribution should be able to be relaxed; pursuing minimal assumptions that allow for equivalences is of general interest.

Acknowledgements

We thank Arun Kumar Kuchibhotla and Matey Nekov for helpful discussions. We also thank the anonymous reviewers whose comments helped improve 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.

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 distribution of large-dimensional sample covariance matrices. *The Annals of Probability*, 26(1):316–345, 1998.

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, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116 (32):15849–15854, 2019.

Michael Celentano, Andrea Montanari, and Yuting Wei. The lasso with general Gaussian designs with applications to hypothesis testing. *arXiv preprint arXiv:2007.13716*, 2020.

Romain Couillet and Merouane Debbah. *Random Matrix Methods for Wireless Communications*. Cambridge University Press, 2011.

Peter Craven and Grace Wahba. Estimating the correct degree of smoothing by the method of generalized cross-validation. *Numerische Mathematik*, 31:377–403, 1979.

Lee H. Dicker. Ridge regression and asymptotic minimax estimation over spheres of growing dimension. *Bernoulli*, 22(1):1–37, 2016.

Edgar Dobriban and Yue Sheng. Distributed linear regression by averaging. *arXiv preprint arXiv:1810.00412*, 2018.

Edgar Dobriban and Yue Sheng. Wonder: Weighted one-shot distributed ridge regression in high dimensions. *Journal of Machine Learning Research*, 21(66): 1–52, 2020.

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

Sandrine Dudoit and Mark J. van der Laan. Asymptotics of cross-validated risk estimation in estimator selection and performance assessment. *Statistical methodology*, 2(2):131–154, 2005.

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

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.

Walid Hachem, Philippe Loubaton, and Jamal Najim. Deterministic equivalents for certain functionals of large random matrices. *The Annals of Applied Probability*, 17(3):875–930, 2007.

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, 1970.

Daniel Hsu, Sham M. Kakade, and Tong Zhang. Random design analysis of ridge regression. In *Conference on Learning Theory*, 2012.

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*, 2011.

Noureddine El Karoui. Asymptotic behavior of unregularized and ridge-regularized high-dimensional robust regression estimators: rigorous results. *arXiv preprint arXiv:1311.2445*, 2013.

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.

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

Olivier Ledoit and Sandrine Peche. Eigenvectors of some large sample covariance matrix ensembles. *Probability Theory and Related Fields*, 151(1–2):233–264, 2011.

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.

Song Mei and Andrea Montanari. The generalization error of random features regression: precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 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.

Léo Miolane and Andrea Montanari. The distribution of the lasso: Uniform control over sparse balls and adaptive parameter tuning. *arXiv preprint arXiv:1811.01212*, 2018.

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

Kamiar Rahnama Rad and Arian Maleki. A scalable estimate of the extra-sample prediction error via approximate leave-one-out. *arXiv preprint arXiv:1801.10243*, 2018.

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. *arXiv preprint arXiv:2003.01770*, 2020.

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

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

A. V. Serdobolskii. On minimal error probability in discriminant analysis. *Reports of the Academy of Sciences of the USSR*, 270:1066–1070, 1983.

Jack W. Silverstein and Sang-Il Choi. Analysis of the limiting spectral distribution of large dimensional random matrices. *Journal of Multivariate Analysis*, 54(2):295–309, 1995.

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

Alexander Tsigler and Peter L. Bartlett. Benign overfitting in ridge regression. *arXiv preprint arXiv:2009.14286*, 2020.

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, Haihao Lu, Arian Maleki, and Vahab Mirrokni. Approximate leave-one-out for fast parameter tuning in high dimensions. *arXiv preprint arXiv:1807.02694*, 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*, 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, and Kamiar Rahnama Rad. Consistent risk estimation in high-dimensional linear regression. *arXiv preprint arXiv:1902.01753*, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. In *International Conference on Learning Representations*, 2017.