Oracle Inequalities for multiple penalty parameters

Jean Feng*
Department of Biostatistics, University of Washington
and
Noah Simon
Department of Biostatistics, University of Washington

October 17, 2016

Abstract

In penalized regression problems, the choice of penalty parameters is important since they ultimately determine the fitted model. The penalty parameters that minimize the generalization error are generally unknown and must be estimated. In this paper, we establish finite-sample oracle inequalities for models selected by a validation set approach. Our upper bounds on the model error depend on the oracle error and a near-parametric term. Therefore in settings where the oracle error shrinks at a sub-parametric rate, the number of penalty parameters can grow with the sample size without affecting the asymptotic convergence rate. Our oracle inequalities hold for penalized regression problems where the fitted models are smoothly parameterized by the penalty parameters. We show that this smoothness condition is satisfied by adding a ridge penalty to the training criterion.

Keywords: Regression, Cross-validation, Regularization

*Jean Feng was supported by NIH grants ???. Noah Simon was supported by NIH grant DP5OD019820. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1 Introduction

Per the usual regression framework, we observe response $y \in \mathbb{R}$ and predictors $\boldsymbol{x} \in \mathbb{R}^p$. Suppose y is generated from the model g^* from model class $\mathcal{G}$

$$y = g^*(\boldsymbol{x}) + \epsilon \tag{1}$$

where ϵ_i are random errors. Our goal is to find the best model in $\mathcal{G}$ to model y given $\boldsymbol{x}$.

In high-dimensional ($p \gg n$) or ill-posed problems, the ordinary least squares estimate performs poorly as it overfits to the training data. A common solution is to add regularization, or penalization, to control model complexity and induce desired structure. The penalized least squares estimate minimizes a criterion of the form

$$\hat{g}(\cdot|\boldsymbol{\lambda}) = \arg\min_{g \in \mathcal{G}} \frac{1}{n} \sum_{i=1}^{n} \left(y_i - g(x_i)\right)^2 + \sum_{j=1}^{J} \lambda_j P_j(g) \tag{2}$$

where P_j are the penalty functions and λ_j are the penalty parameters.

Selecting the penalty parameters is an important task since they ultimately determine the fitted model. Their oracle values balance the residual least squares and the penalty terms to ensure fast convergence rates (van de Geer 2000). For example, when fitting an additive model $f(\boldsymbol{x}) = \sum_{j=1}^{J} f_j(x_j)$ with a roughness penalty for each component, the penalty parameters should be inversely proportional to the penalties of the true model (van de Geer & Muro 2014). When fitting a linear model using the lasso, the penalty parameter should be on the order $\sigma(\log p/n)^{1/2}$ where σ^2 is the variance of the error terms (Bühlmann & Van De Geer 2011).

The obvious problem is that the oracle penalty parameters depend on unknown values. Thus penalty parameters are usually tuned via a training/validation split or cross-validation. The basic idea is to train a model on a random partition of the data and evaluate its error on the remaining data. The penalty parameters that minimize the error on this validation set are then selected. For a more complete review of cross-validation, refer to Arlot (Arlot et al. 2010).

The performance of cross-validation-like procedures is typically characterized by an oracle inequality that bounds the error of the selected model. In a general cross-validation framework, Van Der Laan & Dudoit (2003), van der Laan et al. (2004) provides finite sample oracle

inequalities assuming that cross-validation is performed over a finite model class and Lecué et al. (2012) uses an entropy approach to bound the error for cross-validated models from potentially infinite model classes. In the regression setting, Györfi et al. (2006) provides a finite sample inequality for training/validation split for least squares and Wegkamp (2003) proves an oracle inequality for a penalized least squares holdout procedure. There are also bounds for cross-validated models from ridge regression and lasso (Golub et al. 1979, Chetverikov & Liao 2016, Chatterjee & Jafarov 2015), though the proofs usually rely on the linearity of the model class and are therefore hard to generalize.

Despite the wealth of literature on cross-validation, there is very little work on characterizing the prediction error when the regularization method has multiple penalty parameters. A potential reason is that tuning multiple penalty parameters is computationally difficult; most regularization methods only have one or two tuning parameters (e.g. the Elastic Net and Sparse Group Lasso (Zou & Hastie 2003, Simon et al. 2013)). This computational hurdle has been addressed recently by using continuous optimization methods. For many penalized regression problems, the gradient of the validation loss with respect to the penalty parameters can be calculated using an implicit differentiation trick (Bengio 2000, Foo et al. 2008). Thus a gradient descent procedure can be used to tune the penalty parameters. For more general "hyperparameter selection" problems, one can use a gradient-free approach such as Bayesian optimization Snoek et al. (2012) or Nelder-Mead (CITE).

This paper provides a finite-sample upper bound on the prediction error when multiple penalty parameters are tuned via a training/validation split or cross-validation. We establish an upper bound on the model error that depends on the oracle error and a near-parametric term. Roughly speaking, the additional price for not knowing the oracle penalty parameters is a parametric term. For semi- and non-parametric regression problems, this term is generally much smaller than the oracle error and do not affect the asymptotic convergence rate. In fact, in these cases, the number of penalty parameters can grow with the sample size. Our oracle inequalities depend on the assumption that the fitted models are smoothly parameterized by the penalty parameters. We will show that this smoothness assumption can be easily satisfied if we add a ridge penalty to the training criterion.

Section 2 provides bounds on the prediction error for a training/validation framework

and cross-validation. Section 3 gives examples of penalized regression problems where the fitted models are smoothly parameterized by the penalty parameters. Section 4 provides a simulation study to support our theoretical results. Section 5 discusses our findings and potential future work. Section 6 contains the proofs and other technical details.

2 Main Result

2.1 Training/Validation Split

Given the total observed dataset D of size n, suppose it is randomly split into a training set T of size n_T and validation set V of size n_V. For a function h, define $\|h\|_V^2 = \frac{1}{n_V}\sum_{i\in V} h^2(x_i)$ and similarly for T. Using this notation, the fitted model defined in (2) can be written as

$$\hat{g}(\cdot|\boldsymbol{\lambda}) = \arg\min_{g\in\mathcal{G}} \frac{1}{2}\|y-g\|_T^2 + \sum_{j=1}^{J}\lambda_j P_j(g) \tag{3}$$

In the training/validation framework, we minimize the validation error by tuning over the range of possible penalty parameters values Λ. The selected penalty parameter can be expressed as

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \frac{1}{2}\|y-\hat{g}(\cdot|\boldsymbol{\lambda})\|_V^2 \tag{4}$$

We are interested in comparing its performance to the oracle penalty parameters $\tilde{\boldsymbol{\lambda}}$, which minimizes the model error as follows

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \frac{1}{2}\|g^*-\hat{g}(\cdot|\boldsymbol{\lambda})\|_V^2 \tag{5}$$

We will establish a sharp oracle inequality for the model over the observed covariates in the validation set. Our bound is based on the basic inequality (van de Geer 2000). Let the set of fitted models be denoted

$$\mathcal{G}(T) = \{\hat{g}(\cdot|\boldsymbol{\lambda}) : \lambda\in\Lambda\} \tag{6}$$

From the definition of $\hat{\boldsymbol{\lambda}}$, we have

$$\begin{aligned}
\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 &\leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 + 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V & (7)\\
&\leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 + \sup_{g\in\mathcal{G}(T)} 2\left\langle \epsilon, g - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V & (8)
\end{aligned}$$

where $\langle h, \ell \rangle_V = \frac{1}{n_V} \sum_{i \in V} h(x_i)\ell(x_i)$. The first term in the upper bound is the best error achievable in the model class $\mathcal{G}(T)$. In this paper, the model class is pre-defined so this term can be treated as fixed. Our primary interest is bounding the second term, which is the supremum of empirical processes.

The supremum of empirical processes can be bounded using the complexity of the model class $\mathcal{G}(T)$. Complexity can be measured in a number of ways; we will use metric entropy in this paper. A more thorough review of empirical process theory is presented in Section 6. In this section, we only concern ourselves with smoothly-parameterized functions. More formally, a functions is C-smoothly-parameterized by $\boldsymbol{\lambda}$ over Λ if

$$\|f(\cdot|\boldsymbol{\lambda}_1) - f(\cdot|\boldsymbol{\lambda}_2)\| \leq C\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda \tag{9}$$

Function classes that are C-smoothly parameterized have low metric entropy. Hence their empirical process terms are small with high probability.

In the penalized regression setting, we are interested in bounding the metric entropy of $\mathcal{G}(T)$. It is clear that the fitted functions $\hat{g}(\cdot|\boldsymbol{\lambda})$ are parameterized by $\boldsymbol{\lambda}$, but more work is required to show that they are *smoothly*-parameterized. We defer that discussion to Section 3, where we provide examples of penalized regression problems that satisfy 9.

We now present a sharp oracle inequality for the penalty parameters selected by a training/validation split. The result is a special case of Theorem 3.

Theorem 1. *(give the theorem with λ_{min} and λ_{max}) Suppose $\Lambda = [n_V^{-t_{min}}, n_V^{t_{max}}]^J$.*

Suppose that if $\|\epsilon\|_T \leq 2\sigma$, there are constants C, κ such that for any $u > 0$, we have for all $\lambda \in \Lambda$ (give the theorem with C rather than Cn^κ)

$$\left\|\hat{g}(\cdot|\boldsymbol{\lambda}^{(1)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(2)})\right\|_V \leq Cn^\kappa \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|$$

Then there are universal constants $c_1, c_2 > 0$ and constants $c_3, c_4 > 0$ such that

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \geq \delta \wedge \|\epsilon\|_V \leq 2\sigma\right)$$

$$\leq c_3 \exp\left(-\frac{n_V \delta^4}{c_3^2 \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2}\right) + c_3 \exp\left(-\frac{n_V \delta^2}{c_3^2}\right) + Pr\left(\|\epsilon\|_T \leq 2\sigma\right)$$

where

$$\delta = c_1 \left(\frac{J\left(1 + t_{max} + \kappa\right) \log n_V + c_4}{n_V} \right)^{1/2} + c_2 \sqrt{\left(\frac{J\left(1 + t_{max} + \kappa\right) \log n_V + c_4}{n_V} \right)^{1/2} \left\| \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\|_V}$$

Technically we would like to tune the penalty parameters over the entire range $\Lambda = \mathbb{R}_+^J$, but $\hat{g}(\cdot|\boldsymbol{\lambda})$ can be very ill-behaved under such general conditions. A close approximation is if we ensure that Λ includes the penalty parameter

$$\tilde{\boldsymbol{\lambda}}_{\mathbb{R}_+} = \underset{\boldsymbol{\lambda} \in \mathbb{R}_+^J}{\arg\min} \left\| g^* - \hat{g}(\cdot|\boldsymbol{\lambda}) \right\|_V^2 \tag{10}$$

As shown in Lemma ? (Or cite van de geer), $\tilde{\boldsymbol{\lambda}}_{\mathbb{R}_+}$ shrinks at some polynomial rate $O_p(n^{-\omega})$, so the lower limit of Λ just needs to shrink at a faster polynomial rate $O_p(n^{-t_{\min}})$ where $t_{\min} > \omega$. Suppose $\Lambda = [n^{-t_{min}}, n^{t_{max}}]^J$. Under this condition, we find that the fitted models $\hat{g}(\cdot|\boldsymbol{\lambda})$ for all our examples in Section 3 are $Cn^{-\kappa}$-smoothly parameterized by $\boldsymbol{\lambda}$. We can now apply Theorem 1 to this special case. For ease of interpretation, we present the results in asymptotic notation this time:

Lemma 1. $\Lambda = [n_V^{-t_{min}}, n_V^{t_{max}}]^J$ *Suppose that if* $\|\epsilon\|_T \leq 2\sigma$*, there are constants* C, κ *such that for any* $u > 0$*, we have for all* $\lambda \in \Lambda$ *(give the theorem with* C *rather than* Cn^κ*)*

$$\left\| \hat{g}(\cdot|\boldsymbol{\lambda}^{(1)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(2)}) \right\|_V \leq Cn^\kappa \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|$$

Then

$$\left\| \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^* \right\|_V \leq \quad \left\| \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\|_V \tag{11}$$

$$+ O_p \left(\frac{J\left(1 + t_{max} + \kappa\right) \log n_V}{n_V} \right)^{1/2} \tag{12}$$

$$+ O_p \left(\frac{J\left(1 + t_{max} + \kappa\right) \log n_V}{n_V} \right)^{1/4} \left\| \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\|_V^{1/2} \tag{13}$$

Therefore the model error is upper bounded by the sum of three terms: the oracle error, the near-parametric term in 12, and the geometric mean of the two. We call 12 near-parametric since the convergence rate for a J-dimensional parametric model is usually $(J/n)^{1/2}$ and 12 has a $\log n$ term in the numerator. The $\log n$ term appears because the range of the penalty parameters grows at a polynomial rate in the number of samples. Since $\log n$ grows very slowly, its contribution is nearly negligible.

In the semi- or non-parametric regression setting, the oracle error is likely to dominate in finite samples and will certainly dominate asymptotically.

2. There is an interesting log n term. it appears because the range of Lambda grows with n_V The $\log n_V$ terms are the result of increasing the range of Λ at a polynomial rate. (Not true). Increasing the range of Λ in important to ensure fast convergence rates. (True... but is this important to put here)

3. An interesting geometric mean has appeared. We analyze if this is truly necessary in the simuations... though i still have no idea if it is.

4. Mention no one else has a sharp oracle inequality. Possibly caus they are dealing with generalization error rather than the validation error.

5. The oracle error is small.

2.2 Cross-Validation

In practice, K-fold cross-validation is a far more common procedure than a training/validation split. In addition, we will now bound the generalization error rather than the prediction error over the validation observations. Toward this end, we will apply the oracle inequality in Lecué et al. (2012) to the problem of penalized regression.

The problem setup for K-fold CV is as follows. Let the K partitions for $k = 1, ..., K$ be denoted D_k (with size n_k) and the entire set minus the D_k will be denoted D_{-k}. Consider the joint optimization problem for K-fold CV:

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\lambda \in \Lambda} \frac{1}{2K} \sum_{k=1}^{K} \|y - \hat{g}_k(\cdot|\boldsymbol{\lambda})\|^2_{D_k} \tag{14}$$

$$\hat{g}_k(\cdot|\boldsymbol{\lambda}) = \arg\min_{g \in \mathcal{G}} \frac{1}{2}\|y - g\|^2_{D_{-k}} + \sum_{j=1}^{J} \lambda_j P_j(g) \tag{15}$$

In traditional cross-validation, the final model is retrained on all the data with $\hat{\lambda}$. However, bounding its generalization error requires additional regularity assumptions (Lecué et al. 2012). We consider the following "averaged version of cross-validation" instead

$$\hat{g}_{ACV}(\cdot) = \frac{1}{K} \sum_{k=1}^{K} \hat{g}_k(\cdot|\hat{\boldsymbol{\lambda}}) \tag{16}$$

The following theorem bounds the generalization error of $\hat{g}_{ACV}$.

Theorem 2. *Suppose the errors have expectation zero and* $\|\epsilon\|_\infty < \infty$.

Suppose $\sup_{g\in\mathcal{G}} \|g\|_\infty \leq G$.

Suppose there are constants C, κ *such that*

$$\|\hat{g}(\cdot|\boldsymbol{\lambda}_1) - \hat{g}(\cdot|\boldsymbol{\lambda}_2)\|_\infty \leq \frac{1}{C} n^{-\kappa} \|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \tag{17}$$

Suppose that $\Lambda = [n^{-t_{\min}}, n^{t_{\max}}]^J$.

With high probability, we have for any $a > 0$,

$$E_D \|\hat{g}_{ACV} - g^*\|^2 \leq (1+a) \min_{k\in 1:K, \lambda\in\Lambda} E_D \|\hat{g}_k(\cdot|\boldsymbol{\lambda}) - g^*\|^2 + c_a \max_{k=1:K} \frac{\log^2(n)}{n_k} \tag{18}$$

where c_a *is given in Mitchell.*

Theorem 2 is a stronger result than Theorem 1, but one is required to show that $\hat{g}_\lambda$ is smoothly parameterized by λ over the entire domain, not just the validation points.

2.2.1 Example

We present a nonparametric additive model as an example.

Consider the training criterion with Sobolev penalties on each of the components

$$\hat{f}(\cdot|\boldsymbol{\lambda}), \hat{g}(\cdot|\boldsymbol{\lambda}) = \arg\min_{f,g\in\mathcal{G}} \|y - (f+g)\|_n^2 + \lambda_1 \int_0^1 |f^{(s)}(x)|^2 dx + \lambda_2 \int_0^1 |g^{(t)}(x)|^2 dx \tag{19}$$

where $\mathcal{G}$ are functions defined over the domain [0,1]. It can be shown that the estimate is a spline. According to van de Geer & Muro (2014), the oracle convergence rate is

$$\|\hat{f}(\cdot|\tilde{\boldsymbol{\lambda}}) - f^*\| = O_p(n^{-\frac{s}{2s+1}}), \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\| = O_p(n^{-\frac{t}{2t+1}}) \tag{20}$$

However, one would need to know the Sobolev penalties of f^* and g^* in order to determine the oracle penalty parameters. If the penalty parameters are chosen using the training/validation framework, Theorem 1 gives us

$$\begin{aligned} \|\hat{f}(\cdot|\hat{\boldsymbol{\lambda}}) + \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - (f^* + g^*)\|_V &= O_p(n_T^{-\frac{s}{2s+1}}) + O_p(n_T^{-\frac{t}{2t+1}}) + c_1 \left(\frac{J(\log n_V + c_2)}{n_V}\right)^{1/2} \\ &\quad + \sqrt{c_1 \left(\frac{J(\log n_V + c_2)}{n_V}\right)^{1/2} \left(O_p(n_T^{-\frac{s}{2s+1}}) + O_p(n_T^{-\frac{t}{2t+1}})\right)} \end{aligned}$$

If the penalty parameters are chosen instead using a K-fold cross-validation framework, Theorem 2 states that the averaged version of cross-validation has a generalization error that is bounded as follows

$$E_D\|\hat{f}_{ACV} + \hat{g}_{ACV} - (f^* + g^*)\|^2 \quad = \quad (1+a)E_D\left(O_p(n_k^{-\frac{s}{2s+1}}) + O_p(n_k^{-\frac{t}{2t+1}})\right) + c_a \max_{k=1:K} \frac{\log^2(n)}{n_k}$$

2.2.2 Implications

Theorem 1 and 2 imply that $\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})$ has a semi-parametric convergence rate: the nonparametric (or potentially parametric) convergence rate of the oracle and the parametric convergence rate of the cross-validated model to the oracle. As long as the number of penalty parameters is finite, it does not affect the convergence rate of the model asymptotically.

We can minimize the prediction error by balancing the two terms in the upper bound. The inequalities suggest that there are many approaches. One could change increase the training to validation ratio as the sample size grows. Alternatively, one could increase the number of penalties and penalty parameters with the number of samples. Of course, finding the true minimizer of the upper bound will require knowing properties about the true model, so this would have to be done in some heuristic manner.

3 Smoothness of $\hat{g}(\cdot|\boldsymbol{\lambda})$ in $\boldsymbol{\lambda}$

We now show that $\hat{g}(\cdot|\boldsymbol{\lambda})$ is smoothly parametrized by λ. Theorem 1 requires this smoothness assumption to hold over the validation observations whereas Theorem 2 requires this to hold over the entire domain. Smoothness over the validation set is generally easier to show. We will prove it for nonparametric additive models with smooth penalties and certain nonsmooth penalties. Smoothness over the entire domain is harder to show, so we consider two specific examples: parametric regression problems (where p can grow with n) and smoothing splines.

Throughout, we will presume that $\mathcal{G}$ is a convex function class.

3.1 Smoothness over the Validation Set

We will show that $\hat{g}(\cdot|\boldsymbol{\lambda})$ varies smoothly with respect to λ over the observed covariates, which will directly imply smoothness over the validation set. Suppose we are in the additive

model setting. The fitted models minimize the training criterion

$$\{\hat{g}_j(\cdot|\boldsymbol{\lambda})\}_{j=1}^J = \arg\min_{g\in\mathcal{G}} \|\boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j)\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j) \tag{21}$$

We will not directly consider functions that minimize (21). Instead we will characterize the function class that minimize the perturbed training criterion

$$\{\hat{g}_j(\cdot|\boldsymbol{\lambda})\}_{j=1}^J = \arg\min_{g\in\mathcal{G}} \|\boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j)\|_T^2 + \sum_{j=1}^J \lambda_j \left(P_j(g_j) + \frac{w}{2}\|g_j\|_D^2\right) \tag{22}$$

where $w > 0$ is some constant. The additional ridge penalty will be used in the proofs to ensure that the model is "well-conditioned" over the observed covariates.

In practice, one can choose w to be small enough such that the fitted models from (22) are indistinguishable from those in (21). Lemma 6 proves that the convergence rate of the oracle penalty parameters is preserved for small enough w.

3.1.1 Smooth Penalties

We will first consider the simple case where the penalties P_j are twice-differentiable everywhere. The following lemma shows that fitted function values vary smoothly in the penalty parameters.

Lemma 2. *We suppose the penalty functions P_j are convex and twice-differentiable. Suppose that $\sup_{g\in\mathcal{G}} \|g\|_D \leq G$. Suppose $\lambda_j \geq \lambda_{\min}$ for all j.*

For all $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda$, we have

$$\left\|\sum_{j=1}^J \hat{g}_j(\cdot|\boldsymbol{\lambda}^{(1)}) - \hat{g}_j(\cdot|\boldsymbol{\lambda}^{(2)})\right\|_D \leq \frac{2J}{w\lambda_{\min}}\left(\frac{n}{n_T\lambda_{\min}}\left(2G + \|\epsilon\|_T\right) + wG + G\right)\left\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\right\|$$

All of the proofs on the smoothness of $\hat{g}(\cdot|\boldsymbol{\lambda})$ follow the same recipe. The first step is to consider the optimization problem (21) restricted to models on the line

$$\left\{\hat{g}(\cdot|\boldsymbol{\lambda}^{(1)}) + m\left(\hat{g}(\cdot|\boldsymbol{\lambda}^{(2)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})\right) : m \in [0,1]\right\} \tag{23}$$

By implicit differentiation of the KKT conditions, we can then determine how the fitted models change with respect to the penalty parameters. Finally, the difference $\|\hat{g}(\cdot|\boldsymbol{\lambda}^{(1)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(2)})\|$

is bounded using the mean value theorem. Depending on the situation, the result can be proven directly or by contradiction.

For illustration, we present the proof for Lemma 2 in the case where there is only one penalty parameter. The case with multiple penalty parameters is given in Section 6.

Proof of Lemma 2. Let $h = \hat{g}(\cdot|\lambda^{(1)}) - \hat{g}(\cdot|\lambda^{(2)})$. Suppose for contradiction that $\|h\|_D > d$. Consider the one-dimensional optimization problem

$$\hat{m}(\lambda) = \arg\min_m \frac{1}{2}\|y - (g + mh)\|_T^2 + \lambda \left(P(g + mh) + \frac{w}{2}\|g + mh\|_D^2 \right)$$

Now by the KKT conditions, we have

$$\langle y - (g + mh), h \rangle_T + \lambda \frac{\partial}{\partial m} P(g + mh) + \lambda w \langle h, g + mh \rangle_D = 0$$

It's implicit derivative with respect to λ is

$$\frac{\partial \hat{m}(\lambda)}{\partial \lambda} = \left(\|h\|_T^2 + \lambda \frac{\partial^2}{\partial m^2} P(g + mh) + \lambda w \|h\|_D^2 \right)^{-1} \left(\frac{\partial}{\partial m} P(g + mh) + w \langle h, g + mh \rangle_D \right) \tag{24}$$

From the KKT conditions, we can show

$$\left| \frac{\partial}{\partial m} P(g + mh) \right| \leq \left(\frac{n}{n_T \lambda_{\min}} (2G + \|\epsilon\|_T) + wG + G \right) \|h\|_D$$

Hence

$$\left| \frac{\partial}{\partial \lambda} \hat{m}(\lambda) \right| \leq \left(\frac{n}{n_T} n^{\tau_{min}} (2G + \|\epsilon\|_T) + wG + G \right) n^{\tau_{\min}} w^{-1} \|h\|_D^{-1}$$

By the MVT, there is some $\alpha \in (\lambda^{(1)}, \lambda^{(2)})$ such that

$$\begin{aligned} \left| \hat{m}(\lambda^{(2)}) - \hat{m}(\lambda^{(1)}) \right| &= \left| \left(\lambda^{(2)} - \lambda^{(1)} \right) \left. \frac{\partial \hat{m}(\lambda)}{\partial \lambda} \right|_{\lambda = \alpha} \right| \\ &\leq |\lambda^{(2)} - \lambda^{(1)}| \left(\frac{n}{n_T \lambda_{\min}} (2G + \|\epsilon\|_T) + wG + G \right) \frac{1}{wd\lambda_{\min}} \\ &= 1/2 \end{aligned}$$

But this is a contradiction since we know that $\hat{m}(\lambda^{(2)}) = 1$ and $\hat{m}_{\tilde{k}}(\lambda^{(1)}) = 0$. □

3.1.2 Nonsmooth penalties

If the regression problem contains non-smooth penalty functions, similar results do not necessarily hold. Nonetheless, we find that for many popular non-smooth penalty functions, the functions $\hat{g}(\cdot|\boldsymbol{\lambda})$ are still smoothly parameterized by $\boldsymbol{\lambda}$ almost everywhere. To characterize such problems, we use the approach in Feng & Simon (TBD- CITE?). We begin with the following definitions:

Definition 1. *The differentiable space of a real-valued function L at $g \in \mathcal{G}$ is the set of functions*

$$\Omega^L(g) = \left\{ h \in \mathcal{G} \middle| \lim_{\epsilon \to 0} \frac{L(g + \epsilon h) - L(g)}{\epsilon} \text{ exists} \right\} \tag{25}$$

Definition 2. *S is a local optimality space for a convex function $L(\cdot, \boldsymbol{\lambda}_0)$ if there exists a neighborhood W containing $\boldsymbol{\lambda}_0$ such that for every $\boldsymbol{\lambda} \in W$,*

$$\arg\min_{g \in \mathcal{G}} L(g, \boldsymbol{\lambda}) = \arg\min_{g \in S} L(g, \boldsymbol{\lambda}) \tag{26}$$

Let the training criterion be denoted

$$L_T(g, \lambda) \quad = \quad \arg\min_{g \in \mathcal{G}} \left\| \boldsymbol{y} - \sum_{j=1}^{J} g_j(\boldsymbol{x}_j) \right\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j(g_j) + \frac{w}{2} \|g_j\|_D^2 \right)$$

We will need following conditions to hold for almost every $\boldsymbol{\lambda}$:

Condition 1. *The differentiable space $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}))$ is a local optimality space for $L_T(\cdot, \boldsymbol{\lambda})$.*

Condition 2. *$L_T(\cdot, \boldsymbol{\lambda})$ is twice continuously differentiable along directions in $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}))$.*

Nonsmooth penalties that satisfy Conditions 1 and 2 include ? (I'm not sure which nonparametric penalties actually satisfy this.)

Equipped with the conditions above, we can characterize the smoothness of the fitted functions when the penalties are nonsmooth. In fact the result is exactly the same as Lemma 2. The proof is given in Section 6.

Lemma 3. *Suppose that $\sup_{g \in \mathcal{G}} \|g\|_D \leq G$. Suppose $\lambda_j \geq \lambda_{\min}$ for all j. Suppose the penalty functions are convex. Suppose Conditions 1 and 2 hold for almost every $\boldsymbol{\lambda}$.*

For any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda$, we have

$$\left\| \sum_{j=1}^{J} \hat{g}_j(\cdot|\boldsymbol{\lambda}^{(1)}) - \hat{g}_j(\cdot|\boldsymbol{\lambda}^{(2)}) \right\|_D \leq \frac{2J}{w\lambda_{\min}} \left(\frac{n}{n_T \lambda_{\min}} (2G + \|\epsilon\|_T) + wG + G \right) \left\| \boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)} \right\|$$

3.2 Smoothness over the entire domain

To show that the fitted functions $\hat{g}(\cdot|\boldsymbol{\lambda})$ vary smoothly with respect to $\boldsymbol{\lambda}$ over the entire domain, we will need additional assumptions. In Section 3.1, we controlled the difference between the fitted values at the validation points by adding a ridge penalty. Unfortunately this trick does not allow us to control the sup norm between the fitted functions.

Instead we will just consider specific regression problems. In the parametric regression setting, smoothness over the entire domain is easy to control as long as the derivative of the penalty is controlled by the 2-norm of the model parameters. In the smoothing spline problem, we rely on special properties of the roughness penalty.

3.2.1 Parametric Regression

Consider the parametric regression setting where the model parameters have dimension p. We allow p to grow with the number of samples n, as is common in sieve estimation. Again, we consider the perturbed regression problem:

$$\hat{\theta}(\lambda) = \underset{\theta \in \Theta}{\arg\min} \|y - g(X|\theta)\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j^{v_j}(\theta) + \frac{w}{2}\|\theta\|_2^2 \right) \tag{27}$$

where the additional ridge penalty is now over the model parameters rather than the fitted values.

We can show smoothness over the entire domain under the following conditions

Condition 3. *There exists some constant K such that $\frac{\partial}{\partial m} P(\theta + m\beta) \leq K\|\beta\|_2$*

Condition 4. *There exist constants L, r such that the functions are Lp^r-Lipschitz in the model parameters:*

$$\|g(\cdot|\theta^{(1)}) - g(\cdot|\theta^{(2)})\|_\infty \leq Lp^r \|\theta^{(1)} - \theta^{(2)}\|_2 \tag{28}$$

It is easy to show that Condition 3 is satisfied by many popular parametric penalties, such as the ridge penalty $\|\cdot\|_2^2$, lasso $\|\cdot\|_1$, and group lasso $\|\cdot\|_2$. (Proofs are given in Section 6, if you insist.) Condition 4 requires that the Lipschitz constant grows at a polynomial rate in the number of features. Many models satisfy this condition, assuming they are parameterized appropriately. For example, Condition 4 is satisfied by linear regression when the covariates

are bounded. With these assumptions, we can show that the fitted values are smooth with respect to the penalty parameters over the entire domain.

Lemma 4. *Suppose*

$$\sup_{\theta \in \Theta} \|\theta\| \leq G$$

Suppose Condition 3 and 4 are satisfied. For any $\lambda^{(1)}, \lambda^{(2)} \in \Lambda$*, we have*

$$\|g(\cdot|\hat{\theta}_{\lambda^{(1)}}) - g(\cdot|\hat{\theta}_{\lambda^{(2)}})\|_\infty \leq \frac{2Lp^r\left(K + wG\right)}{wJ\lambda_{\min}} \left\|\lambda^{(2)} - \lambda^{(1)}\right\|_2$$

3.2.2 Smoothing Splines with a Sobolev Penalty

Finally, we consider the additive regression model of fitting a smoothing spline using the Sobolev penalty (De Boor et al. 1978, Wahba 1990, Green & Silverman 1994). For a given set of penalty parameters, the smoothing spline estimate is

$$\{\hat{g}_j(\cdot|\boldsymbol{\lambda})\}_{j=1}^J = \arg\min_{g_j \in \mathcal{G}} \frac{1}{2} \left\| y - \sum_{j=1}^J g_j \right\|_T^2 + \sum_{j=1}^J \lambda_j \int (g_j^{(r_j)}(x))^2 dx \tag{29}$$

where $g^{(r)}$ is the derivative of g of order $r \geq 2$. Unlike the previous sections, we will not need an additional ridge penalty to control the model class.

Due to the special property of the Sobolev penalty given in Lemma 7, we can prove a stronger statement compared to the previous Lemmas 2, 3, and 4. The following lemma shows that the fitted models are Lipschitz in the penalty parameters.

Lemma 5. *Suppose* $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq G$*. Suppose* $\lambda_j \geq \lambda_{\min}$ *for all* j*. Then*

$$\left\| \sum_{j=1}^J \hat{g}_j(\cdot|\lambda^{(1)}) - \hat{g}_j(\cdot|\lambda^{(2)}) \right\|_\infty \leq \left\|\lambda^{(1)} - \lambda^{(2)}\right\| \frac{G}{\lambda_{\min}} \sqrt{\frac{1}{2\lambda_{\min}} \|\epsilon\|_T^2 + \frac{\lambda_{\max}}{\lambda_{\min}} \sum_{j=1}^J P\left(g_j^*\right)} \tag{30}$$

4 Simulations

We now provide a simulation study for the prediction error bound given in Theorem 3. The penalty parameters are chosen by a training/validation split. We show that the error of the select model converges to that of the oracle model at the expected $(\log(n_V)/n_V)^{1/2}$ rate.

Observations were generated from the model

$$y = sin(x_1) + \frac{1}{2} sin(2x_2 + 1) + \sigma\epsilon \tag{31}$$

where $\epsilon \sim N(0, 1)$ and σ scaled the error term such that the signal to noise ratio was 2. The covariates x_1 and x_2 were uniformly distributed over the interval $(0, 6)$. Smoothing splines were fit with a Sobolev penalty

$$\hat{g}_{1,\lambda}, \hat{g}_{2,\lambda} = \underset{g_1, g_2}{\arg\min} \|y - f_1(x_1) - f_2(x_2)\|_T^2 + \int_0^6 (f_1^{(2)}(x))^2 dx + \int_0^6 (f_2^{(2)}(x))^2 dx \tag{32}$$

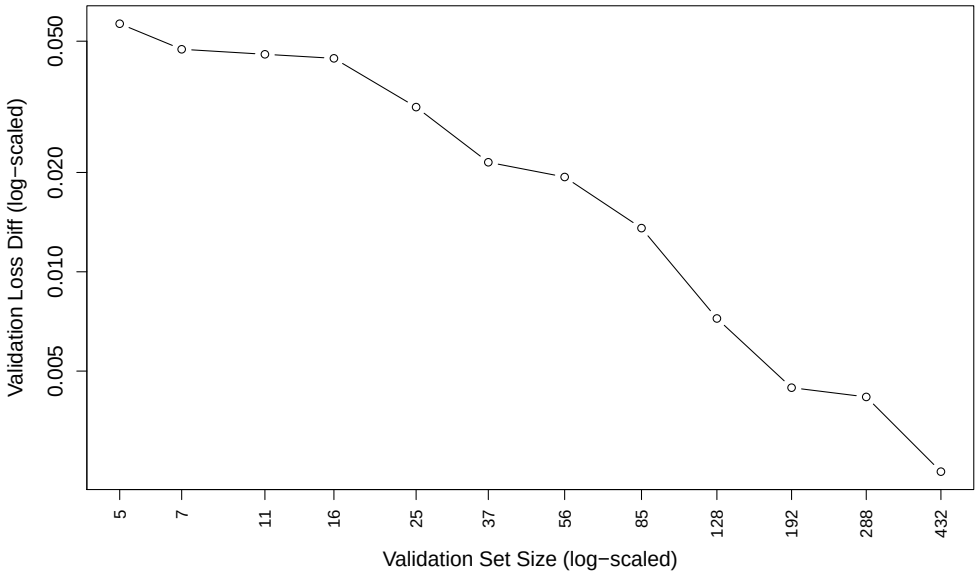
The training set contained 100 samples and models were fitted with 10 knots. A grid search was performed over the penalty parameter values $\{10^{-6+0.2i} : i = 0, ..., 25\}$. We tested validation set sizes $n_V = 20, 30, ..., 80$. The oracle penalty parameters were chosen by minimizing the difference between the fitted model and the true model over a separate test set of 800 samples. A total of 30 simulations were run for each validation set size.

Figure 4 plots the validation loss $\|\hat{g}_\lambda - g^*\|_V$ of the model tuned using a validation set versus the model fit using the oracle penalty parameters. As the validation set increases, the error of the tuned model converges towards the oracle model as expected. In addition we compare the observed difference between the validation losses for the two models and the expected convergence rate of $O_p\left(n_V^{-1/4}\right)$. (Note that all other factors that influence the convergence rate are constant since we only vary the validation set size.) The plot shows that theory closely matches the empirical evidence.

5 Discussion

In this paper, we have shown that the difference in prediction error of the model chosen by cross-validation and the oracle model decreases at a near-parametric rate if the fitted models are smoothly parameterized in terms of the penalty parameters. For many penalized regression problems, we find that this is indeed the case. Our results show that adding penalty parameters does not drastically increase the model complexity. This supports recent efforts to combine regularization methods and "un-pool" regularization parameters. Furthermore, since our result holds for a search over a dense set of penalty parameters, our prediction error

Figure 1: Validation loss difference between oracle and selected model as validation set size grows



bounds apply to cross-validation over a continuum of values, as done in hyper-parameter optimization methods.

The main caveat is that our theorems bound the prediction error of the global minimizer of the validation set. However this is hard to achieve practically since the validation loss is not convex in the penalty parameters. More investigation needs to be done to bound the prediction error of fitted models that are local minima.

A different approach we could have taken in this paper is to bound the distance between the estimated and oracle penalty parameters

$$\left\| \hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}} \right\|_2 \tag{33}$$

instead of the fitted values. Bounding 33 is not obvious from the definitions of $\hat{\boldsymbol{\lambda}}$ and would probably require more regularity assumptions on the model class. However, it could provide a more intuitive understanding of the behavior of cross-validation-like procedures.

6 The Proof

In this paper, we will measure the the complexity of $\mathcal{G}(T)$ by its metric entropy. Let us recall its definition here:

Definition 3. *Let the covering number $N(u, \mathcal{G}, \|\cdot\|)$ be the smallest set of u-covers of $\mathcal{G}$ with respect to the norm $\|\cdot\|$. The metric entropy of $\mathcal{G}$ is defined as the log of the covering number:*

$$H(u, \mathcal{G}, \|\cdot\|) = \log N(u, \mathcal{G}, \|\cdot\|) \tag{34}$$

Theorem 3. *Let ϵ be independent sub-Gaussian random variables. Suppose that $\sup_{g\in\mathcal{G}} \|g\|_\infty \leq G < \infty$. Suppose for any training dataset $T \subseteq D$ with $\|\epsilon\|_T \leq 2\sigma$, we have*

$$\int_0^R H^{1/2}\left(u, \mathcal{G}(\cdot|\mathcal{T})\|\cdot\|_V\right) du \leq \psi(n, J, \sigma) \tag{35}$$

Then for all $\delta > 0$ such that

$$\sqrt{n_V}\delta^2 \geq c\left[\psi_T\left(2\left\|\hat{g}_{\tilde{\lambda}} - g^*\right\|_V + 2\delta\right) \vee \left(2\left\|\hat{g}_{\tilde{\lambda}} - g^*\right\|_V + 2\delta\right)\right] \tag{36}$$

Then with high probability, we have

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V \leq \min_{\lambda\in\Lambda} \|\hat{g}_\lambda(\cdot|T) - g^*\|_V + \delta \tag{37}$$

Proof. Chaining and peeling. □

Proof of Theorem 1

Proof. □

Proof of Theorem 2

Lemma 6. *The oracle rate isn't changed when we add the ridge penalty*

Proof of Lemma 2

Proof of Lemma 3

Proof of Lemma 4

Lemma 7. *Sobolev penalty has nice properties*

Proof of Lemma 5

References

Arlot, S., Celisse, A. et al. (2010), 'A survey of cross-validation procedures for model selection', *Statistics surveys* **4**, 40–79.

Bengio, Y. (2000), 'Gradient-based optimization of hyperparameters', *Neural computation* **12**(8), 1889–1900.

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

Chatterjee, S. & Jafarov, J. (2015), 'Prediction error of cross-validated lasso', *arXiv preprint arXiv:1502.06291* .

Chetverikov, D. & Liao, Z. (2016), 'On cross-validated lasso', *arXiv preprint arXiv:1605.02214* .

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C. & De Boor, C. (1978), *A practical guide to splines*, Vol. 27, Springer-Verlag New York.

Foo, C.-s., Do, C. B. & Ng, A. Y. (2008), Efficient multiple hyperparameter learning for log-linear models, *in* 'Advances in neural information processing systems', pp. 377–384.

Golub, G. H., Heath, M. & Wahba, G. (1979), 'Generalized cross-validation as a method for choosing a good ridge parameter', *Technometrics* **21**(2), 215–223.

Green, P. & Silverman, B. (1994), 'Nonparametric regression and generalized linear models, vol. 58 of', *Monographs on Statistics and Applied Probability* .

Györfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2006), *A distribution-free theory of nonparametric regression*, Springer Science & Business Media.

Lecué, G., Mitchell, C. et al. (2012), 'Oracle inequalities for cross-validation type procedures', *Electronic Journal of Statistics* **6**, 1803–1837.

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2013), 'A sparse-group lasso', *Journal of Computational and Graphical Statistics* **22**(2), 231–245.

Snoek, J., Larochelle, H. & Adams, R. P. (2012), Practical bayesian optimization of machine learning algorithms, *in* 'Advances in neural information processing systems', pp. 2951–2959.

van de Geer, S. (2000), 'Empirical processes in m-estimation (cambridge series in statistical and probabilistic mathematics)'.

van de Geer, S. & Muro, A. (2014), 'The additive model with different smoothness for the components', *arXiv preprint arXiv:1405.6584* .

Van Der Laan, M. J. & Dudoit, S. (2003), 'Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples'.

van der Laan, M. J., Dudoit, S. & Keles, S. (2004), 'Asymptotic optimality of likelihood-based cross-validation', *Statistical Applications in Genetics and Molecular Biology* **3**(1), 1–23.

Wahba, G. (1990), *Spline models for observational data*, Vol. 59, Siam.

Wegkamp, M. (2003), 'Model selection in nonparametric regression', *Annals of Statistics* pp. 252–273.

Zou, H. & Hastie, T. (2003), 'Regression shrinkage and selection via the elastic net', *Journal of the Royal Statistical Society: Series B. v67* pp. 301–320.