Oracle inequalities for validation set procedures with applications to penalized regression

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

November 19, 2016

Abstract

Many regression models are constructed based on a given set of hyperparameters. The optimal hyperparameters that minimize the model's generalization error is this the right term for the error? are unknown so they are often estimated using validation set approaches. In this paper, we establish finite-sample oracle inequalities for cases where the fitted models are smoothly parameterized by the hyperparameters. For the training/validation split framework, we establish a sharp oracle inequality on the model error, with additional near-parametric terms. Our main application is penalized regression problems with multiple penalty parameters. We show that the fitted models are indeed Lipschitz in the penalty parameters and, by our oracle inequality, we show that tuning penalty parameters only adds a near-parametric-rate error term. Hence adding multiple penalty parameters does not drastically increase the degree of overfitting.

Keywords: Hyperparameter selection, Cross-validation, Regularization, Regression

*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 by a true model g^* in some function class $\mathcal{G}$ plus random error ϵ with expectation zero, as follows

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

The function class $\mathcal{G}$ may be parametric (i.e. linear functions) or nonparametric (i.e. twice differentiable functions). Our goal is to estimate the true model g^*.

Many model-estimation procedures can be formulated as selecting a model from $\mathcal{G}$ given training data T and hyperparameters $\boldsymbol{\lambda}$. For example, in penalized regression problems, the fitted model can be expressed as the minimizer of the penalized training criterion

$$\hat{g}(\boldsymbol{\lambda}|T) = \arg\min_{g \in \mathcal{G}} \sum_{(x_i, y_i) \in T} (y_i - g(x_i))^2 + \sum_{j=1}^{J} \lambda_j P_j(g) \tag{2}$$

where P_j are penalty functions and λ_j are penalty parameters. As suggested by the notation in (2), the penalty parameters are the hyperparameters in this model-estimation procedure.

In a set of hyperparameters Λ, there is some oracle hyperparameter $\tilde{\boldsymbol{\lambda}}$ that minimizes the difference between the fitted model and the true model. Usually $\tilde{\boldsymbol{\lambda}}$ is unknown so it is estimated using training/validation split or cross-validation. The basic idea is to fit models on a random partition of the observed data and evaluate its error on the remaining data. We then select hyperparameters $\hat{\boldsymbol{\lambda}}$ that minimize the error on this validation set. For a more complete review of cross-validation, refer to Arlot (Arlot et al. 2010).

The performance of such validation set procedures is typically characterized by an oracle inequality that bounds the error of the selected model. Using 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-based approach to handle 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.

In this paper, we are interested in bounding the error of model-estimation procedures where the estimate perturbs smoothly with the hyperparameters. For $\hat{\boldsymbol{\lambda}}$ selected via a training/validation split or cross-validation, we establish finite-sample oracle inequalities of the form

$$\left\|g^* - \hat{g}\left(\hat{\boldsymbol{\lambda}}, T\right)\right\|^2 \le (1+a) \underbrace{\inf_{\lambda \in \Lambda} \left\|g^* - \hat{g}\left(\boldsymbol{\lambda}, T\right)\right\|^2}_{\text{Oracle error}} + \text{error} \tag{3}$$

for some norm $\|\cdot\|$ and $a \ge 0$. In the training/validation split setting, we consider the L2 norm over validation observations and establish a sharp oracle inequality, i.e. $a = 0$. In the cross-validation setting, we consider the functional L2 norm and establish an upper bound with $a > 0$. In both oracle inequalities, the error term shrinks at roughly a parametric rate. So for semi- and non-parametric regression problems, this term is generally dominated by the oracle error. Hence in many cases, we can actually grow the number of hyperparameters without affecting the the asymptotic convergence rate.

The main application in this paper is penalized regression models of the form (2). Our guiding question is whether having multiple penalty parameters drastically increases the degree of overfitting, a concern raised in Bengio (2000). We will focus on additive models, and other examples are provided in the Appendix. We first show that the fitted model is smoothly parameterized by the penalty parameters so our oracle inequalities apply. We show that each additional penalty parameter only adds a near-parametric error term. Hence adding more penalty parameters for nonparametric and semi-parametric penalized regression problems has negligible effects on the model error. That is, we find that the risk of overfitting is low.

During our literature search, we found little related work regarding oracle inequalities for validation set procedures for penalized regression problems with multiple penalties. The most related work were specific to ridge regression and lasso (Golub et al. 1979, Chetverikov & Liao 2016, Chatterjee & Jafarov 2015). A potential reason is that tuning multiple penalty parameters was computationally difficult, historically; so 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 (Bengio 2000, Foo et al. 2008, Snoek et al. 2012).

Section 2 provides bounds on the model estimation error for training/validation split and cross-validation. Section 3 applies these results to penalized regression models, with a focus on additive models. (The bulk will be spent on showing that fitted models are indeed Lipschitz in the penalty parameters.) Section 4 provides a simulation study to support our theoretical results. Section 5 discusses our findings and potential future work. Proofs are in Section 6.

2 Main Result

Let $D^{(n)}$ denote a dataset with n samples from the model (1) where ϵ are independent random variables with expectation zero. Suppose $\epsilon_1, ..., \epsilon_n$ are uniformly sub-Gaussian with parameter $b > 0$; i.e. $\max_{i=1,...,n} \mathbb{E} e^{t\epsilon_i} \le e^{b^2 t^2/2}$ for all $t \in \mathbb{R}$.

Our goal is to estimate the true model g^* given some training data. Mathematically, we can formulate our model-estimation procedure as an operator $\hat{g}^{(m)}(\cdot|D^{(m)})$ that maps a hyperparameter vector $\boldsymbol{\lambda}$ from some set $\Lambda \subseteq \mathbb{R}^J$ to a function in $\mathcal{G}$. For this paper, we consider model-estimation procedures that are well-behaved. In this section, we focus on those that are Lipschitz.

Definition 1. *Let $\mathcal{F}$ be a function class. Let $\Lambda \subseteq \mathbb{R}^J$. The operator $\hat{f} : \Lambda \mapsto \mathcal{F}$ is C-Lipschitz in $\boldsymbol{\lambda}$ with respect to norm $\|\cdot\|$ over Λ if*

$$\left\| \hat{f}(\boldsymbol{\lambda}) - \hat{f}(\boldsymbol{\lambda}') \right\| \le C \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2 \quad \forall \boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda \tag{4}$$

We hypothesize that many model-estimation procedures are Lipschitz in their hyperparameters. We show penalized regression problems satisfy this condition in Section 3. For more general model-estimation procedures, we provide oracle inequalities in Theorems 3 and 4.

Now suppose we have performed our model-estimation procedure to obtain a set of possible models indexed by Λ. To determine the final model, we must perform a model-selection procedure. Below, we present oracle inequalities when the models are selected by training/validation split and cross-validation.

2.1 Training/Validation Split

In the training/validation split framework, the dataset $D^{(n)}$ is randomly partitioned into a training set T and validation set V of sizes n_T and n_V, respectively. The selected hyperparameter $\hat{\boldsymbol{\lambda}}$ is the minimizer of the validation loss

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \frac{1}{2}\left\| y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D_T^{(n_T)}) \right\|_V^2 \tag{5}$$

where $\|h\|_V = \frac{1}{n_V}\sum_{i\in V} h^2(x_i)$ for any function h. We now give a sharp finite-sample oracle inequality with respect to the norm $\|\cdot\|_V$. The result is a special case of Theorem 3.

Theorem 1. *Let $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$ where $0 < \lambda_{\min} < \lambda_{\max}$. Suppose independent random variables $\epsilon_1, ...\epsilon_n$ are uniformly sub-Gaussian with parameter b. Suppose there are constants $\sigma, C_\Lambda > 0$ such that for any dataset $D^{(n_T)}$ with $\|\boldsymbol{\epsilon}\|_{D^{(n_T)}} \le \sigma$, $\hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)})$ is C_Λ-Lipschitz with respect to $\|\cdot\|_V$ over Λ. Suppose $nC_\Lambda\lambda_{\max} \ge 1$.*

Let

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\lambda\in\Lambda} \left\| g^* - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) \right\|_V^2 \tag{6}$$

Then there is a universal constant c_0 and a constant $c > 0$ only depending on b such that for all δ satisfying

$$\delta^2 \ge c\left(\frac{J(c_0 + \log(nC_\Lambda\lambda_{\max}))}{n_V} \vee \sqrt{\frac{J(c_0 + \log(nC_\Lambda\lambda_{\max}))}{n_V}} \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V \right) \tag{7}$$

we have

$$\begin{aligned} Pr\left(\left\| g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\|_V^2 - \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V^2 \ge \delta^2 \right) \quad \le \quad & c\exp\left(- \frac{n_V\delta^4}{c^2 \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V^2} \right) \\ & + c\exp\left(- \frac{n_V\delta^2}{c^2} \right) \\ & + c\exp\left(- \frac{n_T\sigma^2}{c^2} \right) \end{aligned}$$

The result is perhaps easiest to interpret in asymptotic order notation, as given in the following corollary. (Is there more to explain at all?!)

Corollary 1. *Under the assumptions given in Theorem 1, we have*

$$\left\| g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}) \right\|_V^2 \leq \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}) \right\|_V^2 \tag{8}$$

$$+ O_p\left(\frac{J(c_0 + \log(nC_\Lambda \lambda_{\max}))}{n_V} \right) \tag{9}$$

$$+ O_p\left(\sqrt{\frac{J(c_0 + \log(nC_\Lambda \lambda_{\max}))}{n_V} \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}) \right\|_V^2} \right) \tag{10}$$

We see that the validation loss of the selected model is upper bounded by the oracle error and two remainder terms: a near-parametric term in (9) and a geometric mean of the oracle error in (10). The appearance of a near-parametric term makes intuitive sense. We are trying to estimate the oracle penalty parameters using the validation set, which roughly corresponds to solving a parametric regression problem. The reason we refer to (9) as near-parametric is that the convergence rate of a J-dimensional parametric regression problem is usually $(J/n)^{1/2}$ but (9) has a $\log n$ term in the numerator. The $\log n$ term was introduced when we allowed the range of Λ to grow with the sample size.

However, the geometric mean in (10) suggests that treating the problem of tuning hyperparameters as a parametric regression problem is an oversimplification. The issue is that the model class $\mathcal{G}(T)$ does not contain the true model g^*. The bias term

$$\left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}) \right\|_V^2 \tag{11}$$

not only specifies the minimum validation loss achievable, but it also contributes to the convergence rate.

If the oracle error converges at a sub-parametric rate, the oracle error will dominate asymptotically and the two remainder terms will be negligible. In these settings, we can actually allow the number of hyperparameters J to grow with the sample size. The maximum rate J can grow without affecting the asymptotic convergence rate is proportional to

$$\frac{n_V}{\alpha_n} \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}) \right\|_V^2 \tag{12}$$

2.2 Cross-Validation

In this section, we give an oracle inequality for K-fold cross-validation. Previously, the oracle inequality was with respect to the L2 norm over the validation covariates. We are now interested in the generalization error

$$\|g - g^*\|^2 = \int |g(x) - g^*(x)|^2 \, dx \tag{13}$$

The result in this section is an application of the oracle inequality in Lecué et al. (2012).

The problem setup for K-fold cross-validation is as follows. Let dataset $D^{(n)}$ be randomly partitioned into K sets, which we assume to have equal size for simplicity. Partition k will be denoted $D_k^{(n_V)}$ and its complement will be denoted $D_{-k}^{(n_T)} = D \setminus D_k^{(n_V)}$. We perform our model-selection procedure over $D_{-k}^{(n_T)}$ for $k = 1, ..., K$ and select the hyperparameter that minimizes the average validation loss

$$\hat{\boldsymbol{\lambda}} \quad = \quad \arg\min_{\boldsymbol{\lambda} \in \Lambda} \frac{1}{2K} \sum_{k=1}^{K} \left\| y - \hat{g}(\boldsymbol{\lambda} | D_{-k}^{(n_T)}) \right\|_{D_k^{(n_V)}}^2 \tag{14}$$

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

$$\bar{g}\left(\hat{\boldsymbol{\lambda}} \middle| D^{(n)}\right) = \frac{1}{K} \sum_{k=1}^{K} \hat{g}\left(\hat{\boldsymbol{\lambda}} \middle| D_{-k}^{(n_T)}\right) \tag{15}$$

The following theorem bounds the generalization error of (15). The more general oracle inequality is in Lecué et al. (2012), which is reproduced in Theorem 4.

Theorem 2. *Suppose the dataset can be partitioned into K equal-sized sets, where $K \geq 2$. Let $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$. Suppose independent random variables $\epsilon_1, ...\epsilon_n$ have expectation zero and are uniformly sub-Gaussian with parameter b. Suppose there is a $G \geq 2$ such that $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq G$.*

Suppose there is a constant $C_\Lambda > 0$ such that for any dataset $D^{(n_T)}$ with $\|\boldsymbol{\epsilon}\|_{D^{(n_T)}} \leq \sigma$, $\hat{g}(\boldsymbol{\lambda}|D^{(n_T)})$ is C_Λ-Lipschitz with respect to $\|\cdot\|_\infty$ over Λ.

Then there is are absolute constants $c_1, c_2 > 0$ such that for all $a > 0$,

$$E_{D^{(n)}} \left\| \bar{g}(\hat{\boldsymbol{\lambda}} | D^{(n)}) - g^* \right\|^2 \quad \leq \quad (1 + a) \min_{\boldsymbol{\lambda} \in \Lambda} E_{D^{(n_T)}} \left\| \hat{g}(\boldsymbol{\lambda} | D^{(n_T)}) - g^* \right\|^2 \tag{16}$$

$$+ c_1 \frac{(1+a)^2}{a} \frac{J}{n_V} \left(G \log(G C_\Lambda \lambda_{max}) \log n + c_2 \right) \tag{17}$$

As we can see, Theorems 1 and 2 are quite similar. The upper bounds in both theorems depend on the oracle error and a near-parametric term. The asymptotic convergence rate of the selected model is determined by whichever term dominates. For both the training/validation split framework and cross-validation, we find that tuning hyperparameters is a relatively "cheap" problem to solve. If the oracle error is sub-parametric, the cost of tuning hyperparameters is negligible asymptotically.

There are also some important differences between Theorems 1 and 2. The Lipschitz condition for Theorem 2 requires the Lipschitz condition to hold with respect to $\|\cdot\|_\infty$, instead of the weaker condition with respect to $\|\cdot\|_V$. Also, we no longer have a sharp oracle inequality. The oracle rate is scaled by a constant $1+a$ where $a>0$. These differences occur since we are trying to characterize the general behavior of the selected model based on just the validation loss.

Finally, since the theorems in this section are finite-sample results, one could try to minimize the upper bound by increasing the number of hyperparameters or changing the ratio between the training and validation set sizes. Determining the optimal number of hyperparameters will unfortunately require knowing characteristics about the error variables.

3 Penalized regression models

Theorems 1 and 2 require the fitted functions $\hat{g}(\cdot|\boldsymbol{\lambda})$ to be Lipschitz when the norm of the error terms is bounded. As an example, we show that additive models are C-Lipschitz in the penalty parameters. We will start from the simple example of parametric models fitted with smooth penalty functions, then consider nonsmooth penalty functions, and finally generalize the results to nonparametric additive models.

Recall that in many cases, we will want the range of Λ to grow at some polynomial rate in n. The convergence rates given in Lemmas **??** and **??** hold if the Lipschitz constant is polynomial in n. The following results indeed show that the fitted models are Cn^κ-Lipschitz for some $\kappa>0$.

Finally, we note that additive models are not the only problems where the estimators are smoothly parameterized by the penalty functions. In the Appendix, we show that regression problems where we fit a single model $g(\cdot|\boldsymbol{\theta})$ with multiple, individually-scaled penalties $P_j(\boldsymbol{\theta})$ satisfies (**??**).

3.1 Parametric additive models

Here we consider parametric additive models of the form

$$g(\cdot|\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(J)}) = \sum_{j=1}^{J} g_j(\cdot|\boldsymbol{\theta}^{(j)}) \tag{18}$$

where $\boldsymbol{\theta}^{(j)} \in \mathbb{R}^{p_j}$ and $p = \sum_{j=1}^{J} p_j$. For simplicity, let $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(J)}\right)^\top$. Let $\boldsymbol{\theta}^*$ be the true model parameter. The number of dimensions p_j is allowed to grow with n, as commonly done in sieve estimation. We will suppose that the functions g_j are Lipschitz in $\boldsymbol{\theta}$ with respect to $\|\cdot\|_\infty$.

We consider training criteria of the form

$$L_T\left(y, \boldsymbol{\theta}|\boldsymbol{\lambda}\right) := \frac{1}{2}\left\|y - g(X|\boldsymbol{\theta})\right\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \tag{19}$$

We show that the fitted models are indeed Lipschitz in the penalty parameters with respect to $\|\cdot\|_\infty$, which satisfies the condition in both Theorems 1 and 2.

3.1.1 Parametric regression with smooth penalties

We first suppose the penalty functions are all smooth. In the following section, we will generalize the results to include certain nonsmooth penalty functions. The following lemma states that the fitted models are Lipschitz in the penalty parameter vector.

Lemma 1. *Let*

$$\hat{\boldsymbol{\theta}}^{(j)}\left(\boldsymbol{\lambda}\right) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} L_T\left(y, \boldsymbol{\theta}|\boldsymbol{\lambda}\right) \tag{20}$$

where L_T is defined in (19)

Suppose that $g_j(\cdot|\boldsymbol{\theta}^{(j)})$ are L-Lipschitz in $\boldsymbol{\theta}^{(j)}$ with respect to $\|\cdot\|_\infty$ for all $j = 1, .., J$.

Suppose $P_j(\boldsymbol{\theta})$ and $g_j(\cdot|\boldsymbol{\theta})$ are twice-differentiable and convex with respect to $\boldsymbol{\theta}^{(j)}$ for all $j = 1, .., J$. Suppose $L_T\left(y, \boldsymbol{\theta}|\boldsymbol{\lambda}\right)$ is twice-differentiable and convex with respect to $\boldsymbol{\theta}$.

Suppose there is a $m > 0$ such that the Hessian of the penalized training criterion at the minimizer satisfies

$$\nabla_\theta^2 L_T(y, \boldsymbol{\theta}|\boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq mI \tag{21}$$

Let $\lambda_{\max} > \lambda_{\min} > 0$. Let

$$C_{\theta^*,\Lambda} = \frac{1}{2}\|y - g(\cdot|\boldsymbol{\theta}^*)\|_T^2 + \lambda_{max}\sum_{j=1}^{J} P_j(\boldsymbol{\theta}^{(j),*}) \tag{22}$$

For any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda := [\lambda_{\min}, \lambda_{\max}]^J$, we have

$$\left\| g\left(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)})\right) - g\left(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)})\right)\right\|_\infty \leq \frac{L^2 J^2 \sqrt{2C_{\theta^*,\Lambda}}}{m\lambda_{min}} \left\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\right\| \tag{23}$$

Notice that the result requires that the training criterion is strongly convex at its minimizer. If this is not true, one can add augment the penalty function $P_j(\boldsymbol{\theta}^{(j)})$ with a ridge penalty $\|\boldsymbol{\theta}^{(j)}\|_2^2$ so that the training criterion becomes

$$\frac{1}{2}\|y - g(X|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J}\lambda_j\left(P_j(\boldsymbol{\theta}^{(j)}) + \frac{w}{2}\|\boldsymbol{\theta}^{(j)}\|_2^2\right) \tag{24}$$

The proofs for all the examples follow a similar recipe. We determine the gradient of the fitted model with respect to the penalty parameter vector by implicitly differentiating the KKT conditions. We can then bound the norm of the gradient to get the Lipschitz constant.

For illustration, we present the proof for Lemma 1 in the case where there is only one penalty parameter. The case with multiple penalty parameters is given in Section 6.

Proof of Lemma 1. fix me if we still want this here

By the KKT conditions, we have

$$\langle y - g(\boldsymbol{\theta}), \nabla_{\boldsymbol{\theta}} g(\boldsymbol{\theta})\rangle_T + \lambda\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) + \lambda w\boldsymbol{\theta} = \mathbf{0}$$

Its implicit derivative...

finish proof □

3.1.2 Parametric regression with non-smooth 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, such as the lasso (CITE) and group lasso (CITE), the fitted functions 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 2. *The differentiable space of a real-valued function f at $\boldsymbol{\theta}$ is*

$$\Omega^f(\boldsymbol{\theta}) = \left\{ \boldsymbol{\beta} \middle| \lim_{\epsilon \to 0} \frac{f(\boldsymbol{\theta} + \epsilon\boldsymbol{\beta}) - f(\boldsymbol{\theta})}{\epsilon} \text{ exists} \right\} \tag{25}$$

Definition 3. *S is a local optimality space for a convex function $f(\cdot, \boldsymbol{\lambda})$ over the W if for every $\boldsymbol{\lambda} \in W$,*

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in S} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{26}$$

We can now characterize a set $\Lambda_{smooth} \subseteq \Lambda$ over which the fitted functions are well-behaved. Λ_{smooth} must satisfy the following conditions:

Condition 1. *For every $\boldsymbol{\lambda} \in \Lambda_{smooth}$, there exists a ball $B(\boldsymbol{\lambda})$ with nonzero radius centered at $\boldsymbol{\lambda}$ such that*

- *For all $\boldsymbol{\lambda}' \in B(\boldsymbol{\lambda})$, the training criterion $L_T(\cdot, \cdot)$ is twice differentiable along directions in $\Omega^{L_T(\cdot,\cdot)}\left(\hat{\boldsymbol{\theta}}_\lambda\right)$.*
- *The differentiable space $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}(\boldsymbol{\theta})$ at $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ is a local optimality space for $L_T(\cdot, \boldsymbol{\lambda})$ over $B(\boldsymbol{\lambda})$.*

Condition 2. *For every $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$, let the line segment between the two points be denoted*

$$\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) = \left\{ \alpha\boldsymbol{\lambda}^{(1)} + (1-\alpha)\boldsymbol{\lambda}^{(2)} : \alpha \in [0,1] \right\}$$

Suppose the intersection $\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \cap \Lambda_{smooth}^C$ is countable.

In lasso and group lasso problems, it is hypothesized that almost every penalty parameter satisfies these properties. (CITE?) Equipped with these conditions, we can characterize the smoothness of the fitted functions when the penalties are non-smooth. In fact the Lipschitz constant is exactly the same as that in Lemma 1.

Lemma 2. *Define $\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda})$ as in (20).*

Suppose $g_j(\cdot|\boldsymbol{\theta}^{(j)})$ are L-Lipschitz in $\boldsymbol{\theta}^{(j)}$ with respect to $\|\cdot\|_\infty$ for all $j = 1, .., J$.

Suppose $P_j(\boldsymbol{\theta}^{(j)})$ and $g_j(\cdot|\boldsymbol{\theta}^{(j)})$ are convex with respect to $\boldsymbol{\theta}^{(j)}$ for all $j = 1,..,J$ and $L_T(y, \boldsymbol{\theta}|\boldsymbol{\lambda})$ is convex with respect to $\boldsymbol{\theta}$.

Let U_λ be an orthonormal matrix with columns forming a basis for the differentiable space of $L_T(\cdot|\boldsymbol{\lambda})$ at $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$. Suppose there is a $m > 0$ such that the Hessian of the penalized training criterion with respect to the differentiable space at the minimizer satisfies

$$ {}_{U_\lambda}\nabla^2_\theta L_T(y, \boldsymbol{\theta}|\boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq mI \tag{27}$$

Suppose $\Lambda_{smooth} \subseteq \Lambda \coloneqq [\lambda_{\min}, \lambda_{\max}]^J$ satisfies Conditions 1 and 2.

Then any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$ satisfies (23).

3.2 Nonparametric additive models

We now generalize the results to nonparametric additive models. We consider estimators of the form

$$\{\hat{g}_j(\cdot|\boldsymbol{\lambda})\}_{j=1}^J = \arg\min_{g\in\mathcal{G}} \left\|\boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j)\right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j) \tag{28}$$

where P_j are now penalty functionals. The following lemma states that the fitted functions are Lipschitz with respect to $\|\cdot\|_D$, which satisfies the Lipschitz condition in Theorem 1.

Lemma 3. *Let $\mathcal{G}$ be a convex function class. $\hat{g}_j(\cdot|\boldsymbol{\lambda})$ is defined in 28.*

Suppose the penalty functions P_j are twice Gateaux differentiable and convex over $\mathcal{G}$. Suppose there is a $m > 0$ such that the training criterion has a twice Gateaux derivative at $\hat{g}_j(\cdot|\boldsymbol{\lambda})$ for all $j = 1, ..., J$ satisfies

$$\left\langle D^2\left(\left\|\boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j)\right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j)\right) \circ h, h\right\rangle \geq m \quad \forall h \in \mathcal{G}, \|h\|_D = 1 \tag{29}$$

Let $\lambda_{\max} > \lambda_{\min} > 0$. Let

$$C_{\theta^*,\Lambda} = \frac{1}{2}\left\|y - \sum_{j=1}^J g_j^*(\cdot|\boldsymbol{\lambda})\right\|_T^2 + \lambda_{max}\sum_{j=1}^J P_j(g_j^*) \tag{30}$$

For any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda \coloneqq [\lambda_{\min}, \lambda_{\max}]^J$, we have

$$\left\|\sum_{j=1}^J \hat{g}_j\left(\cdot|\boldsymbol{\lambda}^{(1)}\right) - \hat{g}_j\left(\cdot|\boldsymbol{\lambda}^{(2)}\right)\right\|_D \leq \frac{J}{m\lambda_{min}}\sqrt{2C_{\theta^*,\Lambda}\frac{n}{n_T}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)}\left\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\right\| \tag{31}$$

4 Simulations

We now provide a simulation study for the prediction error bound given in Theorem 1. The penalty parameters are chosen by a training/validation split. We show that the error of the selected model converges to that of the oracle model at the near-parametric rate.

Observations were generated from the model

$$y = \exp(x_1) + x_2^2 + \sigma\epsilon \tag{32}$$

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 $(-1,1)$.

We fit a smoothing splines using the Sobolev penalty (De Boor et al. 1978, Wahba 1990, Green & Silverman 1994). The training criterion was

$$\|y - f_1(x_1) - f_2(x_2)\|_T^2 + \lambda_1 \int_0^6 (f_1^{(2)}(x))^2 dx + \lambda_2 \int_0^6 (f_2^{(2)}(x))^2 dx \tag{33}$$

The training set contained 100 samples and models were fitted with 10 knots. A grid search was performed over the penalty parameter values $\{10^{-9+0.05i} : i = 0, ..., 140\}$. We tested 36 validation set sizes $n_V = \lfloor 20 * 2^i \rfloor$ for equally log-spaced intervals from $i = 0$ to $i = 7$. A total of 20 simulations were run for each validation set size.

Figure 4 plots the difference of between the model loss and the oracle loss

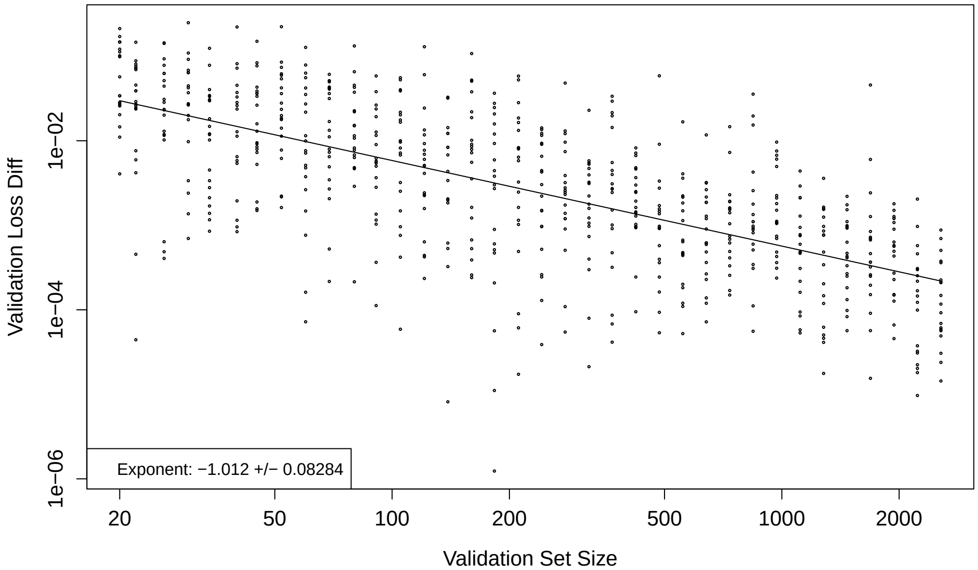
$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2$$

as the validation set size increases. The difference of the validation losses drops at a rate of about n^{-1}. This rate is in fact faster than that in Theorem 1 since the geometric term seems to play no role. We conjecture that there may be additional regularity conditions that allow the geometric term to be completely discarded.

5 Discussion

In this paper, we have established oracle inequalities for penalty parameter selection using a training/validation split framework or k-fold cross-validation. The results address the concern in Bengio (2000) regarding "the amount of overfitting that can be brought when too many

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



hyperparameters are optimized." Our results show that this should not be a major concern. In a non-parametric setting or parametric setting where p grows with n, the oracle error is the dominating term in the upper bound. At worst, the tuning penalty parameter problem contributes an error that is on the same order as the oracle error, say in a parametric setting where p is fixed.

There is recent interest in combining regularization methods, but seems to be an artificial restriction to two or three penalty parameters. The area of penalized regression methods with tens or hundreds of penalty parameters remains largely unexplored. Our results suggest that this direction of research could be fruitful. As shown in Feng and Simon (TBD), un-pooling the penalty parameters in a sparse group lasso model is surprisingly effective.

One major caveat to our results is that we have assumed that the penalty parameters can be tuned such that the validation loss is minimized. However it is difficult to find the global minimizer since the validation loss is not convex in the penalty parameters. Optimization methods need to be developed to effectively solve the bilevel optimization problems in (**??**). In addition, it would be worthwhile to understand the performance of models that are only local minimizers of the validation loss.

Finally, there are still many open questions to explore. Our results assume that the fitted models are smoothly parameterized with respect to the penalty parameters and we provide a number of examples that satisfy these conditions. There are probably many more examples of regression problems that satisfy the smoothness condition and the smoothness condition itself can probably be generalized. In addition, it would be interesting to bound the distance between the selected and oracle penalty parameters

$$\left\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\right\|_2 \tag{34}$$

Such a result would perhaps give a more intuitive understanding of penalty parameter selection methods.

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 4. *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:*

$$H(u, \mathcal{G}, \|\cdot\|) = \log N(u, \mathcal{G}, \|\cdot\|) \tag{35}$$

Theorem 3.

Theorem 4. *This is reproduced from the Mitchell paper*

Proof. Chaining and peeling. □

Proof of Theorem 1

Proof. □

Proof of Theorem 2

Proof of Lemma 1

Proof of Lemma 2

Proof of Lemma 3

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.

Buldygin, V. V. & Kozachenko, Y. V. (1980), 'Sub-gaussian random variables', *Ukrainian Mathematical Journal* **32**(6), 483–489.
URL: *http://dx.doi.org/10.1007/BF01087176*

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.

Stromberg, K. (1994), *Probability For Analysts*, Chapman & Hall/CRC Probability Series, Taylor & Francis.
URL: *https://books.google.com/books?id=gQaz79fv6QUC*

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.