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 18, 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 increasing the number of penalty parameters in penalized regression problems is likely to improve model error.

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 possible 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

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

for some norm $\|\cdot\|$ and $a \geq 0$. 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.

Our 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 point raised in Bengio (2000). Our main example within penalized regression models will be additive models; other examples are provided in the Appendix. We first show that our oracle inequalities apply since the fitted model is smoothly parameterized by the penalty parameters. We then show that the additional penalty parameters incur a roughly a near-parametric error term. Hence for nonparametric and semi-parametric penalized regression problems, adding more penalty parameters results in a negligible increase in the model error. These results suggest that regularization methods that combine many penalty functions are effective model-estimation procedures.

We were unable to find oracle inequalities for validation set procedures for penalized regression problems with multiple penalties. The only related oracle inequalities we found are specific to ridge regression and lasso (Golub et al. 1979, Chetverikov & Liao 2016, Chatterjee & Jafarov 2015). A potential reason is that, historically, tuning multiple penalty parameters was computationally difficult; 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

Suppose we have observations $D = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ generated from the model

$$y_i = g^*(\boldsymbol{x}_i) + \epsilon_i \quad i = 1, ..., n \tag{4}$$

where g^* is from our model class $\mathcal{G}$ and ϵ_i are independent random variables with expectation zero. We suppose $\epsilon_1, ..., \epsilon_n$ are uniformly sub-Gaussian random variables 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 $t \in \mathbb{R}$.

2.1 Training/Validation Split

Can we do this without discussing penalized regression? We have some operator τ (pick something cooler) that maps from training data to models (give formal defs of those!) Suppose dataset D 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 (**??**) can be written as

$$\hat{g}(\cdot|\boldsymbol{\lambda}) = \arg\min_{g\in\mathcal{G}} \frac{1}{2}\|y - g\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g) \tag{5}$$

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

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \frac{1}{2}\|y - \hat{g}(\cdot|\boldsymbol{\lambda})\|_V^2 \tag{6}$$

We are interested in comparing its performance to the oracle penalty parameters $\tilde{\boldsymbol{\lambda}}$, which minimize the model error Should there be an expectation there?

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda}\in\Lambda} \frac{1}{2}\|g^* - \hat{g}(\cdot|\boldsymbol{\lambda})\|_V^2 \tag{7}$$

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

$$\mathcal{G}(T) = \{\hat{g}(\cdot|\boldsymbol{\lambda}) : \lambda \in \Lambda\} \tag{8}$$

This should go later! It distracts here From the definition of $\hat{\boldsymbol{\lambda}}$, we can bound the difference of the validation losses

$$\begin{aligned}
\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 &\leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V & (9)\\
&\leq \sup_{g\in\mathcal{G}(T)} 2\left\langle \epsilon, g - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V & (10)
\end{aligned}$$

where $\langle h, \ell\rangle_V = \frac{1}{n_V}\sum_{i\in V} h(x_i)\ell(x_i)$. Our goal is to bound this empirical process term with high probability.

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 (ha! if we have space). For this paper, we mostly concern ourselves with Lipschitz functions.

Definition 1. *A function $f(\cdot|\boldsymbol{\lambda})$ is C-Lipschitz with respect to norm $\|\cdot\|$ over Λ if*

$$\|f(\cdot|\boldsymbol{\lambda}_1) - f(\cdot|\boldsymbol{\lambda}_2)\| \leq C\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2 \quad \forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda \tag{11}$$

Function classes that satisfy (11) have low metric entropy. Hence their empirical process terms are small with high probability.

We are interested in bounding the metric entropy of $\mathcal{G}(T)$ to bound (10). $\mathcal{G}(T)$ is clearly a smaller class than $\mathcal{G}$ that is parametric in $\boldsymbol{\lambda}$, but more work needs to be done to show that the functions are Lipschitz in the penalty parameters ? can't parse this sentence. In Section 3, we present penalized regression problems for additive models as one such example.

We now present a sharp oracle inequality for the penalty parameters selected by a training/validation split.

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 is a constant*

$\sigma > 0$ such that for any dataset with $\|\boldsymbol{\epsilon}\|_T \le \sigma$, $\hat{g}(\cdot|\boldsymbol{\lambda})$ is C-Lipschitz with respect to $\|\cdot\|_V$ over Λ.

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

$$\delta^2 \ge c\left(\frac{\alpha_n^2}{n_V} \vee \frac{\alpha_n}{\sqrt{n_V}}\left\|\hat{g}\left(\cdot|\tilde{\boldsymbol{\lambda}}\right) - g^*\right\|_V\right) \tag{12}$$

where

$$\alpha_n = \sqrt{J\left(1 + \log\left(32Cn(\lambda_{\max} - \lambda_{\min})\right)\right)} \vee 1 \tag{13}$$

we have

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

can we upperbound α_n by something less crazy like $C\sqrt{J\log\left(32Cn(\lambda_{\max} - \lambda_{\min})\right)}$

We need a lot more exposition about what the heck the above means!!! The result is a special case of Theorem 3, which applies to general function classes that are not Lipschitz. For example, in other penalized regression examples, we will find that the fitted functions satisfy

$$\|f(\cdot|\boldsymbol{\lambda}_1) - f(\cdot|\boldsymbol{\lambda}_2)\| \le C\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|_2^2 \quad \forall \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda \tag{14}$$

Theorem 3 shows that the convergence rate in this case is very similar to that in Theorem 1.

We see in Theorem 1 that the choice of $\lambda_{\min}$ and $\lambda_{\max}$ are important contributors to the convergence rate. Ideally we would want to choose $\Lambda = \mathbb{R}_+^J$, but $\hat{g}(\cdot|\boldsymbol{\lambda})$ can be very ill-behaved under such general conditions. Therefore Λ is usually chosen to be just large enough so that it contains

$$\tilde{\boldsymbol{\lambda}}_{\mathbb{R}_+} = \arg\min_{\boldsymbol{\lambda} \in \mathbb{R}_+^J} \|g^* - \hat{g}(\cdot|\boldsymbol{\lambda})\|_V^2 \tag{15}$$

As shown in van de Geer (2000), $\tilde{\boldsymbol{\lambda}}_{\mathbb{R}_+}$ shrinks at polynomial rate $O_p(n_T^{-\omega})$ for some $\omega > 0$, so the lower limit of Λ just needs to shrink at a faster polynomial rate.

We now apply Theorem 1 to this special case where $\lambda_{\min} = n_T^{-t_{\min}}$ and $\lambda_{\max} = n^{t_{\max}}$ for $0 < t_{\min} < t_{\max}$. In the examples in Section 3, the Lipschitz constant for $\hat{g}(\cdot|\boldsymbol{\lambda})$ turns out to be proportional to $\lambda_{\min}^{-1}$. Hence we also allow for $\hat{g}(\cdot|\boldsymbol{\lambda})$ to be Cn^{κ}-Lipschitz for some $\kappa \geq 0$. For ease of interpretation, we present the results in asymptotic notation this time:

Lemma 1. *Let $\Lambda = [n_V^{-t_{min}}, n_V^{t_{max}}]^J$ where $0 < t_{\min} < t_{\max}$. Suppose for $\sigma > 0$ such that for any dataset with $\|\boldsymbol{\epsilon}\|_T \leq \sigma$, $\hat{g}(\cdot|\boldsymbol{\lambda})$ is Cn^{κ}-Lipschitz with respect to $\|\cdot\|_V$ over Λ. Then*

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \tag{16}$$

$$+O_p\left(\frac{J\alpha_n}{n_V}\right) \tag{17}$$

$$+O_p\left(\sqrt{\frac{J\alpha_n}{n_V}\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2}\right) \tag{18}$$

where

$$\alpha_n = (t_{\max} + \kappa)\log n_T + \log(Cn)$$

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 (17) and a geometric mean of the oracle error in (18). sThe 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 (17) as near-parametric is that the convergence rate of a J-dimensional parametric regression problem is usually $(J/n)^{1/2}$ but (17) 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 (18) suggests that treating the problem of tuning penalty parameters 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

$$\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \tag{19}$$

not only specifies the minimum validation loss achievable, but it also appears in the convergence rate.

Lemma 1 shows that 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 penalty parameters J to grow with the sample size without affecting the asymptotic convergence rate. The maximum rate J can grow without affecting the asymptotic convergence rate is of the form

$$O_p\left(\frac{n_V}{\alpha_n}\left\|g^* - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2\right) \tag{20}$$

Of course if the oracle error converges at a parametric rate, we need to keep the number of penalty parameters fixed.

2.2 Cross-Validation

In this section, we give an oracle inequality for K-fold cross-validation. Instead of bounding the model error over the observed covariates, we will bound the generalization error, which is the squared $L2$-norm of the difference:

$$\|g - g^*\|^2 = \int |g(x) - g^*(x)|^2 \, dx \tag{21}$$

Toward this end, we will apply the oracle inequality in Lecué et al. (2012). what does this mean?

We will generalize the notation from the previous section. Let a dataset with n samples be denoted $D^{(n)}$. The fitted model given any training data $D^{(n)}$ is denoted

$$\hat{g}_{D^{(n)}}(\cdot|\boldsymbol{\lambda}) = \arg\min_{g\in\mathcal{G}} \frac{1}{2}\|y - g\|^2_{D^{(n)}} + \sum_{j=1}^{J} \lambda_j P_j(g) \tag{22}$$

For K-fold cross-validation, the problem setup is as follows. As before, let $D^{(n)}$ be the entire dataset. For simplicity, suppose the dataset can be partitioned into K sets of equal size n_V. Let $n_T = n - n_V$. Then partition k will be denoted $D_k^{(n_V)}$ and its complement will be denoted $D_{-k}^{(n_T)} = D \setminus D_k^{(n_V)}$. The selected penalty parameter vector is

$$\hat{\boldsymbol{\lambda}} \;=\; \arg\min_{\boldsymbol{\lambda}\in\Lambda} \frac{1}{2K} \sum_{k=1}^{K} \left\| y - \hat{g}_{D^{n_T}_{-k}}(\cdot|\boldsymbol{\lambda}) \right\|^2_{D_k} \tag{23}$$

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 following "averaged version of cross-validation" instead

$$\bar{g}_{D^n}(x) = \frac{1}{K}\sum_{k=1}^{K} \hat{g}_{D^{(n_T)}_{-k}}(x|\hat{\boldsymbol{\lambda}}) \tag{24}$$

The following theorem bounds the generalization error of $\bar{g}_{D^n}$. We note that the generalization error of the the oracle inequality is no longer sharp; the oracle rate is scaled by a constant $1+a$ for any $a>0$. This is a consequence of trying to characterize the behavior of the selected model based on its validation error. One could try to shrink a towards zero, but the additional error term grows as a decreases.

Theorem 2. *Let $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$. Suppose independent random variables $\epsilon_1, ...\epsilon_n$ are uniformly sub-Gaussian with parameter b. Suppose $\sup_{g\in\mathcal{G}} \|g\|_\infty \le G$.*

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

Then there is an absolute constant $c>0$ such that for all $a>0$,

$$E_{D^{(n)}} \left\|\bar{g}_{D^n} - g^*\right\|^2 \le (1+a) \min_{\boldsymbol{\lambda}\in\Lambda} E_{D^{(n_T)}} \left\|\hat{g}_{D^{(n_T)}}(\cdot|\boldsymbol{\lambda}) - g^*\right\|^2 \tag{25}$$

$$+c\frac{(1+a)^2}{a}\frac{J}{n_V}\left(C_\Lambda + \frac{1}{2}\log n_V + 4GC_\Lambda \log n_V\right) \tag{26}$$

where

$$C_\Lambda = 1 + \log\left(128GC(\lambda_{max} - \lambda_{min})\right) \tag{27}$$

Notice that Theorem 2 requires a stronger Lipschitz condition than that in Theorem 1. Here we require that the fitted functions are Lipschitz with respect to $\|\cdot\|_\infty$.

Now we apply Theorem 2 to the special case where Λ grows at a polynomial rate with the sample size. We will use the same assumptions as we did in Lemma 1.

Lemma 2. *Suppose $\sup_{g\in\mathcal{G}} \|g\|_\infty \le G$. Suppose $\Lambda = [n_T^{-t_{\min}}, n_T^{t_{\max}}]^J$.*

Suppose that if $\|\epsilon\|_T \le \sigma$, there are constants C, κ such that for any dataset $D^{(n_T)}$, $\hat{g}_{D^{(n_T)}}(\cdot|\boldsymbol{\lambda})$ is Cn^κ-Lipschitz with respect to $\|\cdot\|_\infty$ over Λ.

Then for any $a>0$, there are positive constants c_a and c_G only dependent on a and G, respectively, such that

$$E_{D^{(n)}} \left\|\bar{g}_{D^{(n)}} - g^*\right\|^2 \le (1+a) \min_{\boldsymbol{\lambda}\in\Lambda} E_{D^{(n_T)}} \left\|\hat{g}_{D^{(n_T)}}(\cdot|\boldsymbol{\lambda}) - g^*\right\|^2 \tag{28}$$

$$+c_a\frac{J}{n_V}\left(\left(c_G \log n_V + 1\right)\left((\kappa + t_{max})\log n_T + 1\right) + c_G\right) \tag{29}$$

(We have simplified the constants c and a in this expression for readibility. Refer to the original theorem for the actual constants)

Lemmas 2 and 1 are quite similar in that the upper bounds are functions of 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 penalty parameters is a relatively "cheap" problem to solve. If the oracle error is sub-parametric, the cost of tuning penalty parameters is negligible asymptotically.

The theorems and lemmas given in this section are all finite-sample results. One could try to minimize the upper bound by increasing the number of penalty parameters or changing the ratio between the training and validation set sizes. Determining the optimal number of penalty parameters will unfortunately require knowing characteristics about the error variables ϵ. (Perhaps you can use cross-validation to determine the number of penalty parameters to use. Ha! How meta!)

3 Examples: Additive 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 1 and 2 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 (14).

3.1 Parametric additive models

Here we consider parametric additive models of the form

$$g(\cdot|\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(J)}) = \sum_{j=1}^{J} g_j(\cdot|\boldsymbol{\theta}^{(j)}) \tag{30}$$

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

$$L_T\left(y, \boldsymbol{\theta}|\boldsymbol{\lambda}\right) \coloneqq \frac{1}{2}\left\|y - g(X|\boldsymbol{\theta})\right\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \tag{31}$$

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 3. *Let*

$$\hat{\boldsymbol{\theta}}^{(j)}\left(\boldsymbol{\lambda}\right) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^p} L_T\left(y, \boldsymbol{\theta}|\boldsymbol{\lambda}\right) \tag{32}$$

where L_T is defined in (31)

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

$$\nabla_\theta^2 L_T(y, \boldsymbol{\theta}|\boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq mI \tag{33}$$

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

$$C_{\theta^*,\Lambda} = \frac{1}{2}\|y - g(\cdot|\boldsymbol{\theta}^*)\|_T^2 + \lambda_{max}\sum_{j=1}^{J} P_j(\boldsymbol{\theta}^{(j),*}) \tag{34}$$

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

$$\left\| g\left(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)})\right) - g\left(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)})\right)\right\|_\infty \le \frac{L^2 J^2\sqrt{2C_{\theta^*,\Lambda}}}{m\lambda_{min}}\left\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\right\| \tag{35}$$

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

$$\frac{1}{2}\|y - g(X|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J}\lambda_j\left(P_j(\boldsymbol{\theta}^{(j)}) + \frac{w}{2}\|\boldsymbol{\theta}^{(j)}\|_2^2\right) \tag{36}$$

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 3 in the case where there is only one penalty parameter. The case with multiple penalty parameters is given in Section 6.

Proof of Lemma 3. 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*

$$\Omega^f(\boldsymbol{\theta}) = \left\{ \boldsymbol{\beta} \middle| \lim_{\epsilon \to 0} \frac{f(\boldsymbol{\theta} + \epsilon \boldsymbol{\beta}) - f(\boldsymbol{\theta})}{\epsilon} \text{ exists} \right\} \tag{37}$$

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

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in S} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{38}$$

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 $\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \cap \Lambda^C_{smooth}$ 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 3.

Lemma 4. *Define $\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda})$ as in (32).*

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

$$ {}_{U_\lambda}\nabla^2_\theta L_T(y, \boldsymbol{\theta}|\boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq mI \tag{39}$$

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

Then any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$ satisfies (35).

3.2 Nonparametric additive models

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

$$\{\hat{g}_j(\cdot|\boldsymbol{\lambda})\}_{j=1}^J = \arg\min_{g\in\mathcal{G}} \left\| \boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j) \right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j) \tag{40}$$

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 5. *Let $\mathcal{G}$ be a convex function class. $\hat{g}_j(\cdot|\boldsymbol{\lambda})$ is defined in 40.*

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

$$\left\langle D^2 \left(\left\| \boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j) \right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j) \right) \circ h, h \right\rangle \geq m \quad \forall h \in \mathcal{G}, \|h\|_D = 1 \tag{41}$$

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

$$C_{\theta^*,\Lambda} = \frac{1}{2} \left\| y - \sum_{j=1}^J g_j^*(\cdot|\boldsymbol{\lambda}) \right\|_T^2 + \lambda_{max} \sum_{j=1}^J P_j(g_j^*) \tag{42}$$

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

$$\left\| \sum_{j=1}^J \hat{g}_j\left(\cdot|\boldsymbol{\lambda}^{(1)}\right) - \hat{g}_j\left(\cdot|\boldsymbol{\lambda}^{(2)}\right) \right\|_D \leq \frac{J}{m\lambda_{min}} \sqrt{2C_{\theta^*,\Lambda} \frac{n}{n_T} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)} \left\| \boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)} \right\| \tag{43}$$

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

$$y = \exp(x_1) + x_2^2 + \sigma\epsilon \tag{44}$$

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

$$\|y - f_1(x_1) - f_2(x_2)\|_T^2 + \lambda_1 \int_0^6 (f_1^{(2)}(x))^2 dx + \lambda_2 \int_0^6 (f_2^{(2)}(x))^2 dx \tag{45}$$

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

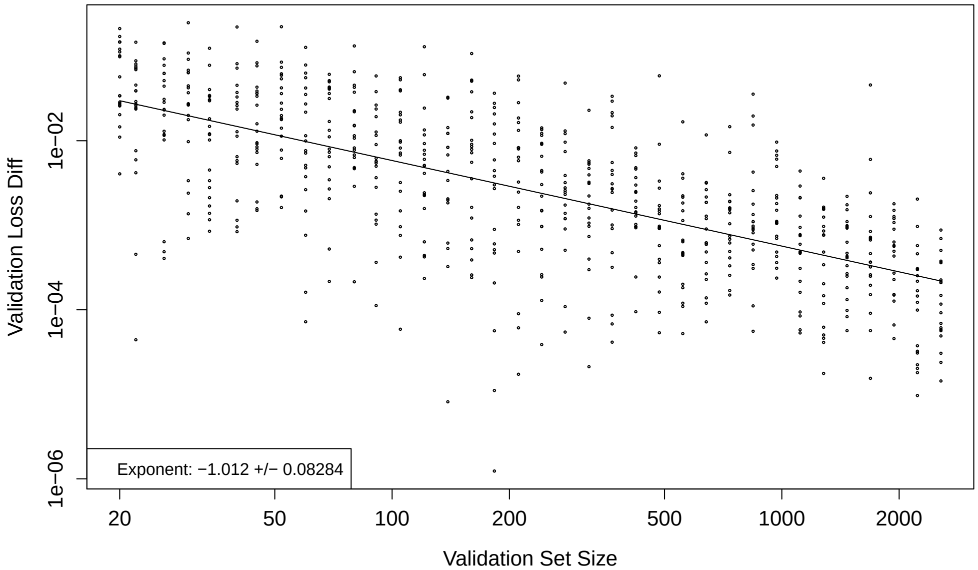
$$\left\| \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^* \right\|_V^2 - \left\| \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\|_V^2$$

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

$$\left\|\hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}}\right\|_2 \tag{46}$$

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:*

$$H(u, \mathcal{G}, \|\cdot\|) = \log N(u, \mathcal{G}, \|\cdot\|) \tag{47}$$

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*

$$\int_0^R H^{1/2}\left(u, \mathcal{G}(\cdot|\mathcal{T})\| \cdot \|_V\right) du \leq \psi(n, J, \sigma) \tag{48}$$

Then for all $\delta > 0$ such that

$$\sqrt{n_V}\delta^2 \geq c\left[\psi_T\left(2\left\|\hat{g}_{\tilde{\lambda}} - g^*\right\|_V + 2\delta\right) \vee \left(2\left\|\hat{g}_{\tilde{\lambda}} - g^*\right\|_V + 2\delta\right)\right] \tag{49}$$

Then with high probability, we have

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V \leq \min_{\lambda\in\Lambda} \|\hat{g}_\lambda(\cdot|T) - g^*\|_V + \delta \tag{50}$$

Proof. Chaining and peeling. ∎

Proof of Theorem 1

Proof. □

Proof of Theorem 2

Proof of Lemma 3

Proof of Lemma 4

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.

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.