Oracle Inequalities for multiple penalty parameters

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

August 24, 2016

Abstract

In penalized least squares problems, penalty parameters determine the tradeoff between minimizing the residual sum of squares and the model complexity. The oracle set of penalty parameter values that minimize the generalization error are usually estimated by evaluating the models on a separate validation set or by cross-validation. We show that in many problems, the difference between the generalization error of the selected model and the oracle converges at a near-parametric rate. The key idea to show that the fitted models are smoothly parameterized by the penalty parameters. This finding justifies recent work on combining penalty functions using separate penalty parameters.

Keywords: Regression, Cross-validation, Regularization

[*] Jean Feng was supported by NIH grants DP5OD019820 and T32CA206089. 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 and predictors $\boldsymbol{x} \in \mathbb{R}^p$. Suppose y is generated from the true model g^* from model class $\mathcal{G}$

$$y_i = g^*(\boldsymbol{x}) + \epsilon_i \tag{1}$$

where ϵ_i are random errors. In high-dimensional ($p >> n$) 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

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

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-book, Wahba-smoothing spline paper, and others?). For example, when fitting an additive model $f = \sum f_i$ with roughness penalties, the penalty parameters should be inversely proportional to the penalties of the true model (cite Vandegeer additive models). Then the convergence rates of each f_i is as fast as in the case where the other components are known. In a high-dimensional lasso problem, the penalty parameter should be on the order $\sigma(\log p/n)^{1/2}$ where σ^2 is the variance of the error terms.

The obvious problem is that the oracle penalty parameters depend on unknown values. Instead, the 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. One then searches for the penalty parameters that yield the lowest validation error.

The performance of cross-validation-like procedures is characterized by bounding the prediction error. Typically the upper bound is composed of two terms: the error of the oracle plus a complexity term. In a general CV framework, Van Der Laan (2003, 2004) provides finite sample oracle inequalities assuming that CV is performed over a finite model class and Mitchell () uses an entropy approach to bound CV for potentially infinite model classes. In the regression setting, Gyorfi (2002) 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, Heath and Wahba, Chetverikov, and Chaterjee), 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 so most regularization methods only have one or two tuning parameters (e.g. Elastic Net, Sparse Group Lasso, etc.). However, recent efforts have used continuous optimization methods to make this "hyperparameter selection" problem computationally tractable. A popular gradient-free approach is to use Bayesian optimization (Snoek). For more specialized problems, e.g. when the hyperparameters are exactly the penalty parameters, the gradient of the validation loss with respect to the penalty parameters can be calculated by implicit differentiation and the parameters can be tune by gradient descent (Bengio, Foo). Another potential reason is that there seems to be a widely held belief that having multiple penalty functions drastically increases model complexity and leads to overfitting. (CITE SOMETHING or say that our JASA referees thought it was a dumb idea).

Our paper provides a finite sample upper bound on the prediction error when tuning multiple penalty parameters via a training/validation split and cross-validation. The upper bound is composed of the error of the oracle and an empirical process term that converges at a near-parametric rate. Our main contribution is proving that the fitted function in many penalized regression problems vary smoothly in the penalty parameters. By reducing the model class to a parametric model class, we can show that the empirical process term converges to faster than the error of the oracle. The proofs use results from empirical process theory and an implicit differentiation trick.

Section 2 provides bounds on the prediction error for a training/validation framework and cross-validation. Section 3 proves that for many penalized regression problems, the fitted models are smoothly parameterized by the penalty parameters. Section 4 provides simulation studies to support the theory. Section 5 discusses the results in the paper. Section 6 contains the full proofs for the main results and additional lemmas.

2 Main Result

2.1 Training/Validation Split

Consider the training/validation split framework. Given the total observed dataset D of size n, suppose it is split into a training set T of size n_T and validation set V of size n_V. Define $\|h\|_V^2 = \frac{1}{n_V}\sum_{i\in A} h^2(x_i)$ and similarly for T. Let the fitted models over the range of penalty parameter values Λ be denoted

$$\mathcal{G}(T) = \{\hat{g}_{\boldsymbol{\lambda}}(\cdot|T) : \lambda \in \Lambda\} \tag{3}$$

The final penalty parameter chosen by the training/validation split is

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

We are interested in bounding $\left\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\right\|_V$, the error between the fitted model and the true model at the observed covariates in the validation set.

The bound is based on the basic inequality (cite?). From the definition of $\hat{\lambda}$, we have

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V^2 \leq \|\hat{g}_{\tilde{\lambda}}(\cdot|T) - g^*\|_V^2 + 2\left|\langle\epsilon, \hat{g}_{\tilde{\boldsymbol{\lambda}}}(\cdot|T) - \hat{g}_{\hat{\boldsymbol{\lambda}}}(\cdot|T)\rangle_V\right| \tag{5}$$

where $\langle h, \ell\rangle_A = \frac{1}{|A|}\sum_{i\in A} h(x_i)\ell(x_i)$. The second term on the right hand is the empirical process term. Bounding this will rely on results from empirical process theory.

Empirical process results state that when the complexity of the class $\mathcal{G}(T)$ is small, the empirical process term will be small with high probability. In this paper, we will measure the the complexity of $\mathcal{G}(T)$ by its metric entropy. Let us recall its definition here:

Definition 1. *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:*

$$H(u, \mathcal{G}, \|\cdot\|) = \log N(u, \mathcal{G}, \|\cdot\|) \tag{6}$$

The following theorem gives a finite-sample upper bound on the error of the fitted model $\hat{g}_{\hat{\lambda}}(\cdot|T)$ over the observed points in the validation set. The proof leverages standard chaining and peeling arguments.

Theorem 1. *Let ϵ be independent sub-Gaussian random variables. Suppose that $\sup_{g \in \mathcal{G}} \|g\|_\infty \le G < \infty$. Suppose for any training dataset $T \subseteq D$ with $\|\epsilon\|_T \le 2\sigma$, we have*

$$\int_0^R H^{1/2}\left(u, \mathcal{G}(\cdot|\mathcal{T})\| \cdot \|_V\right) du \le \psi(n, J, \sigma) \tag{7}$$

Then for all $\delta > 0$ such that

$$\sqrt{n_V}\delta^2 \ge c\left[\psi_T\left(2\left\|\hat{g}_{\tilde{\lambda}} - g^*\right\|_V + 2\delta\right) \vee \left(2\left\|\hat{g}_{\tilde{\lambda}} - g^*\right\|_V + 2\delta\right)\right] \tag{8}$$

Then with high probability, we have

$$\left\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\right\|_V \le \min_{\lambda \in \Lambda} \|\hat{g}_\lambda(\cdot|T) - g^*\|_V + \delta \tag{9}$$

In the penalized regression setting, each function $\hat{g}_\lambda$ in $\mathcal{G}(T)$ directly maps to a set of penalty parameters, so one would expect that the covering number of $\mathcal{G}(T)$ and Λ to be related. In Section 3, we show that $\hat{g}_\lambda$ is smoothly parameterized by λ in many penalized regression problems. Corollary 1 uses this insight to build a d-cover set of $\mathcal{G}(T)$ from a $\delta(d)$-cover set of Λ. Applying Theorem 1, we then get a bound on the prediction error of the penalized least squares estimate. Note that the complexity term in the upper bound contains a $\log n$ term. This is the result of allowing the range of Λ to increase at a polynomial rate.

Corollary 1. *Suppose that $\sup_{g \in \mathcal{G}(\cdot|\mathcal{T})} \|g\|_\infty \le G < \infty$. Suppose that $\Lambda = [n^{-t_{\min}}, n^{t_{\max}}]^J$.*

Suppose that if $\|\epsilon\|_T \le 2\sigma$, there is some constant C, κ such that for any $u > 0$, we have

$$\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\| \le Cn^\kappa u^2 \implies \|\hat{g}_{\boldsymbol{\lambda}_1} - \hat{g}_{\boldsymbol{\lambda}_2}\|_V \le u \tag{10}$$

Then with high probability, we have

$$\left\|\hat{g}_{\hat{\lambda}} - g^*\right\|_V \le \|\hat{g}_{\tilde{\lambda}} - g^*\|_V + \frac{c_1\left(J(\log n_V + c_2)\right)^{1/2}}{\sqrt{n_V}} + \sqrt{c\left(J(\log n_V + c_2)\right)^{1/2}\left\|\hat{g}_{\tilde{\lambda}} - g^*\right\|_V n_V^{-1/2}} \tag{11}$$

Proof. By Lemma param_covering_cube, we have

$$H(u, \mathcal{G}(T), \| \cdot \|_V) \le \log \frac{1}{C_J} + J \log \left(\frac{2n^{t_{max} - \kappa} + 2Cu^2}{Cu^2} \right)$$

Let $R_1 = R \wedge \sqrt{n^{t_{max}-\kappa}/C}$.

Then after immense algebraic massaging, we get

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \| \cdot \|_V) du \le R \left(\left[\log \frac{1}{C_J} + J(2 + \log 4) + J \log \left(\frac{4n^{t_{max}-\kappa}}{C} \right) \right]^{1/2} + \sqrt{2J \log \frac{1}{R} \vee 0} \right) \tag{12}$$

We note since $\delta > \frac{1}{n_V}$ (modulo a constant?), it suffices to choose δ such that

$$\sqrt{n_V} \delta^2 \ge c \left(\| \hat{g}_{\tilde{\lambda}} - g^* \|_V + \delta \right) \left(\left[\log \frac{1}{C_J} + J(2 + \log 4) + J \log \left(\frac{4n^{t_{max}-\kappa}}{C} \right) \right]^{1/2} + \sqrt{2J \log n_V} \right)$$

Let

$$K = c \left(\left[\log \frac{1}{C_J} + J(2 + \log 4) + J \log \left(\frac{4n^{t_{max}-\kappa}}{C} \right) \right]^{1/2} + \sqrt{2J \log n_V} \right)$$

and

$$\omega = \| \hat{g}_{\tilde{\lambda}} - g^* \|_V$$

The quadratic formula gives us that

$$\delta \ge \frac{K + \sqrt{K^2 + 4K\omega\sqrt{n_V}}}{2\sqrt{n_V}}$$

□

2.2 Cross-Validation

In practice, K-fold cross-validation is a far more common procedure than a training/validation split. Furthermore, one is usually interested in bounding the generalization error rather than the prediction error on the validation set. Toward this end, we will apply the oracle inequality in Mitchell (CITE) 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:

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \frac{1}{2} \sum_{k=1}^{K} \|y - \hat{g}_{\lambda}(\cdot|D_{-k})\|_k^2 \tag{13}$$

$$\hat{g}(\lambda|D_{-k}) = \arg\min_{g \in \mathcal{G}} \frac{1}{2}\|y - g\|_{-k}^2 + \sum_{j=1}^{J} \lambda_j P_j^{v_j}(g) + \frac{w}{2}\|g\|^2 \tag{14}$$

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 (CITE mitchell). Instead, we will bound the generalization error of a model from the "averaged version of cross-validation":

$$\frac{1}{K}\sum_{k=1}^{K} \hat{g}_{\hat{\boldsymbol{\lambda}}}(\cdot|D_{-k}) \tag{15}$$

The following theorem bounds the generalization error of the model from the averaged version of cross-validation. For any function h, we use the notation $\|h\|^2 = \int h^2(x)d\mu(x)$.

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

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

Suppose there is a constant C such that

$$\|\hat{g}_{\lambda_1} - \hat{g}_{\lambda_2}\|_\infty \leq \|\lambda_1 - \lambda_2\| C n^\kappa \tag{16}$$

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

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

$$E_D \left\| \frac{1}{K}\sum_{k=1}^{K} \hat{g}(\hat{\lambda}|D_{-k}) - g^* \right\|^2 \leq (1+a) \min_{k \in 1:K, \lambda \in \Lambda} E_D \|\hat{g}(\lambda|D_{-k}) - g^*\|^2 + c_a \max_{k=1:K} \frac{\log^2(n)}{n_k} \tag{17}$$

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

2.2.1 Implications

Theorem 2 and Corollary 1 imply that $\hat{g}_{\hat{\lambda}}$ is indeed a semi-parametric model. Its convergence rate can be separated into the convergence rate of the oracle to the truth and the parametric convergence rate of the cross-validated model to the oracle. One could try to minimize the upper bound by balancing the two terms, though it would require knowledge that is usually unknown. Nonetheless, adding more penalty parameters is “cheap.” It is very possible that adding more penalties or un-pooling penalties could actually increase the convergence rate. For example, in the additive model setting, there is usually a single penalty parameter, but this could be replaced by an un-pooled version:

$$\lambda \sum_{j=1}^{J} P_j^{v_j}(g_j) \rightarrow \sum_{j=1}^{J} \lambda_j P_j^{v_j}(g_j) \tag{18}$$

Of course, there is a limit to the number of penalty parameters one can add. For example, if the number of penalty parameters grows with n, the cross-validated model no longer converges to the oracle at a near-parametric rate.

Theorem 1 also provides guidance on choosing the optimal ratio between the training and validation sets. As the sample size increases, the ratio between the training and validation sets should change. For example, consider the nonparametric setting with the oracle convergence $n^-1/4$. With 100 training samples, one would want about 70 samples in the training set. With 1000 training samples, one would want about 850 samples in the training set. _Insert plot_

3 Smoothness of $\hat{g}_\lambda$ in λ

We now show that $\hat{g}_\lambda$ is smoothly parametrized by λ. Corollary 1 requires this smoothness assumption to hold over the validation observations whereas Theorem 2 requires this to hold over the entire domain. Under varying assumptions, we are able to satisfy these conditions. Below, we will begin with the general case of a nonparametric regression problem with smooth penalties. Then we will consider the more specific case of a parametric regression problem. Finally, we consider the example of smoothing splines fitted with the Sobolev penalty.

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

3.1 The Implicit Differentiation trick

All the proofs rely on an implicit differentiation trick, so we will highlight it here. For any function $h \in \mathcal{G}$ and any $\lambda, \delta \in \Lambda$, consider the one-dimensional optimization problem

$$\hat{m}_h(\lambda) = \arg\min_{m\in\mathbb{R}} \frac{1}{2} \|y - (\hat{g}_\delta + mh)\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j^{v_j}(\hat{g}_\delta + mh) \tag{19}$$

Suppose the penalty functions P_j are twice-differentiable everywhere.

From the KKT conditions, we have

$$\langle h, y - (\hat{g}_\delta + mh)\rangle + \sum_{j=1}^{J} \lambda_j \frac{\partial}{\partial m} P_j^{v_j}(\hat{g}_\delta + mh)\Big|_{m=\hat{m}_h(\lambda)} = 0 \tag{20}$$

Implicit differentiation of (20) with respect to λ_ℓ for $\ell = 1, ..., J$ gives us

$$\frac{\partial}{\partial \lambda_\ell}\hat{m}_h(\lambda) = -\left(\|h\|_T^2 + \sum_{j=1}^{J} \lambda_j \frac{\partial^2}{\partial m^2} P_j^{v_j}(\hat{g}_\delta + mh)\right)^{-1} \frac{\partial}{\partial m} P_\ell^{v_\ell}(\hat{g}_\delta + mh)\Big|_{m=\hat{m}_h(\lambda)} \tag{21}$$

A key step in all the proofs is to bound the absolute value of (21).

3.2 Regression problems with Smooth Penalties

We will first consider regression problems with smooth penalties and then those with nonsmooth penalties. Suppose the penalties P_j are norms that are twice differentiable everywhere.

Instead of considering models fitted as in (2), we perturb the training criterion with an additional ridge penalty

$$\hat{g}(\lambda) = \arg\min_{g\in\mathcal{G}} \|y - g(X)\|_n^2 + \sum_{j=1}^{J} \lambda_j \left(P_j^{v_j}(g) + \frac{w}{2}\|g\|_V^2\right) \tag{22}$$

where w is some constant. Clearly this is unnecessary if there is already some penalty $P_j^{v_j}(g) = C\|g\|_V^2$ for some constant C (such as in the elastic net). Otherwise, one can choose w to be small enough to maintain the oracle convergence rate of the original problem, as shown in Lemma 5 (and in practice, w can be chosen to be so small that the fitted models are indistinguishable). The additional ridge penalty is useful in our proof as it ensures the model space is"well-conditioned".

Lemma 1. *Suppose the penalty functions P_j are smooth norms and that $v_j \ge 1$. Suppose $\sup_{g\in\mathcal{G}} \|g\| \le G$. Suppose $\Lambda = [n^{-\tau_{\min}}, n^{\tau_{\max}}]^J$. Then for all $d > 0$, if $\lambda_1, \lambda_2 \in \Lambda$ satisfy*

$$\|\lambda_1 - \lambda_2\| \le d^2 w \left(C n^c v \left(\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_V^2 + G \right)\right)^{-1} whereisJ \tag{23}$$

then

$$\|\hat{g}_{\lambda_1} - \hat{g}_{\lambda_2}\|_V \le d \tag{24}$$

Proof. We present the proof here in the case where there is only one penalty parameter. It readily extends into the case for J penalty parameters.

Consider any λ_1, λ_2 satisfying (31). Let $h = \hat{g}_{\lambda_1}(\cdot|T) - \hat{g}_{\lambda_2}(\cdot|T)$. Suppose $\|h\|_V > d$ for contradiction.

Consider the one-dimensional problem as done in Section 3.1

$$\hat{m}_h(\lambda) = \arg\min_m \frac{1}{2}\|y - (\hat{g}_{\lambda_1} + mh)\|_T^2 + \lambda \left(P^v(\hat{g}_{\lambda_1} + mh) + \frac{w}{2}\|\hat{g}_{\lambda_1} + mh\|_V^2 \right)$$

By our assumptions that $\|h\|_V \ge d$ and P is convex, we have

$$\left|\frac{\partial}{\partial\lambda}\hat{m}_h(\lambda)\right| \le \frac{n^{\tau_{\min}}}{wd^2}\left|\frac{\partial}{\partial m}P^v(\hat{g}_{\lambda_1} + mh) + w\langle h, \hat{g}_{\lambda_1} + mh\rangle_D\right|_{m=\hat{m}_\lambda(\lambda_0)} \tag{25}$$

The second term can be bounded by the definitions of $\hat{m}_h(\lambda_0)$ and $\hat{g}_{\delta_i}$ and the fact that P is a semi-norm:

$$\begin{aligned} \left|\frac{\partial}{\partial m}P(g + mh)\right| &\le P(h) \\ P(h) &\le P(\hat{g}_{\lambda_1}) + P(\hat{g}_{\lambda_2}) \\ P(\hat{g}_\lambda) &\le \frac{1}{2\lambda}\|\epsilon\|_T^2 + P(g^*) + \frac{w}{2}\|g^*\|_V^2 \forall\lambda \in \Lambda \end{aligned}$$

Combining these facts, we get that

$$\left|\frac{\partial}{\partial\lambda}\hat{m}_h(\lambda)\right| \le Cd^{-2}n^c w^{-1} v \left(\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_V^2 + G \right)$$

By the mean-value theorem, there is some $\alpha \in (\lambda_1, \lambda_2)$ such that

$$\begin{aligned} |\hat{m}_h(\lambda_2) - \hat{m}_h(\lambda_1)| &= (\lambda_2 - \lambda_1)\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|_{\lambda_0=\alpha} && (26) \\ &\le 1/2 && (27) \end{aligned}$$

However clearly $\hat{m}_h(\lambda_1) = 0$ and $\hat{m}_h(\lambda_2) = 1$, so there is a contradiction. □

3.3 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 like the lasso and the group lasso, the functions $\hat{g}_\lambda(\cdot|T)$ are still smoothly parameterized by λ almost everywhere. To characterize such problems, we generalize the approach used in Feng (CITE). We begin with the following definitions:

Definition 2. *The differentiable space of a real-valued function L at $g \in \mathcal{G}$ is the set of functions*

$$\Omega^L(g) = \left\{ h \in \mathcal{G} \middle| \lim_{\epsilon \to 0} \frac{L(g+\epsilon h) - L(g)}{\epsilon} \text{ exists} \right\} \tag{28}$$

Definition 3. *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 $\boldsymbol{\lambda} \in W$,*

$$\arg\min_{g \in \mathcal{G}} L(g, \boldsymbol{\lambda}) = \arg\min_{g \in S} L(g, \boldsymbol{\lambda}) \tag{29}$$

Suppose the training criterion has the form

$$L_T(g, \lambda) \quad = \quad \frac{1}{2}\|y - g\|_T^2 + \sum_{j=1}^J \lambda_j \left(P^{v_j}(g) + \frac{w}{2}\|g\|_V^2 \right)$$

To be able to take an implicit derivative, 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 $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}))$.*

Consequently, we get the following smoothness result.

Lemma 2. *Let $\lambda_0 \in \Lambda$. Suppose Conditions 1 and 2 hold for almost every $\boldsymbol{\lambda}$.*

Furthermore, suppose one can show that

$$\tag{30}$$

Then for all $d > 0$, if $\lambda_1, \lambda_2 \in \Lambda$ satisfy

$$\|\lambda_1 - \lambda_2\| \le d^2 w \left(C n^c v \left(\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_V^2 + G \right) \right)^{-1} whereisJ \tag{31}$$

then

$$\|\hat{g}_{\lambda_1} - \hat{g}_{\lambda_2}\|_V \le d \tag{32}$$

The proof is given in Section 6. Nonsmooth penalties that satisfy the conditions in Lemma 2 include the lasso and sparse group lasso.

3.4 Smoothness over the full domain

??

In Section 3.2, proofs proceeded by contradiction. Here we will proceed using a direct proof. We consider only specific problems since ?

3.4.1 Parametric Regression

We will now consider the parametric regression setting where the model parameters have dimension p. Again, we will perturb the original penalization problem with an additional ridge penalty.

$$\hat{\theta}(\lambda) = \underset{\theta \in \Theta}{\arg\min} \|y - g(X)\|_n^2 + \sum_{j=1}^{J} \lambda_j \left(P_j^{v_j}(\theta) + \frac{w}{2}\|\theta\|_2^2 \right) \tag{33}$$

Define the function class as $\mathcal{G}(T) = \{g_{\hat{\theta}(\lambda)} : \lambda \in \Lambda\}$.

Lemma 3. *Suppose*

$$\|\sup_{\theta \in \Theta} \theta\|_2 \le R$$

and the penalty functions P_j are norms that are smooth and

$$P(\beta) \le c \forall \|\beta\|_2 \le 1$$

Suppose $v_j \ge 1$. Suppose $g_\theta(x)$ is Lp^r-lipschitz in θ

$$|g_{\theta_1}(x) - g_{\theta_2}(x)| \le Lp^r \|\theta_1 - \theta_2\|_2$$

Suppose $\Lambda = [n^{-\tau_{\min}}, n^{\tau_{\max}}]^J$. Then the entropy is bounded above by

$$H\left(u, \mathcal{G}(T), \|\cdot\|_D\right) \le J \left(2\log\frac{1}{u} + \kappa \log n + r \log p + stuff \right) \tag{34}$$

Proof. The proof here is only for one penalty parameter, but it generalizes to the multi-parameter case.

Consider any $\beta = c_0 \left(\hat{\theta}_{\lambda_0} - \hat{\theta}_\lambda\right)$ where c is s.t. $\|\beta\|_2 \le 1$. Consider the optimization problem

$$\hat{m}_\beta(\lambda) = \arg\min_m \frac{1}{2}\|y - g_{\hat{\theta}_\lambda + m\beta}\|_T^2 + \lambda_0 \left(P^v(\hat{\theta}_\lambda + m\beta) + \frac{w}{2}\|\hat{\theta}_\lambda + m\beta\|_2^2\right)$$

By implicit differentiation of the KKT conditions, we get

$$\begin{aligned} \left|\frac{\partial}{\partial\lambda}\hat{m}_\beta(\lambda)\right| &\le \frac{n^{\tau_{min}}}{w}\left|\frac{\partial}{\partial m}P^v(\hat{\theta}_\lambda + m\beta) + w\langle\hat{\theta}_\lambda + m\beta, \beta\rangle\right|_{m=\hat{m}_\lambda(\lambda)} \\ &\le \frac{n^{\tau_{min}}}{w}\left(v\left(n^\kappa C\right)^{v-1} c + wR\right) \end{aligned}$$

where $C = O_p(1)\left(\|\epsilon\|_T^2 + P(\theta^*) + w\|\theta^*\|_2^2\right)$

By the assumption that g_θ is Lp^r-lipschitz in θ, we have

$$\begin{aligned} \|g_{\theta_\lambda} - g_{\theta_{\lambda_0}}\|_\infty &\le Lp^r \hat{m}_\beta(\lambda)\|\beta\|_2 \\ &= Lp^r|\lambda_0 - \lambda|\left|\frac{\partial}{\partial\lambda}\hat{m}_\beta(\lambda)\right|_{\alpha\in[\lambda,\lambda_0]} \\ &\le |\lambda_0 - \lambda|\frac{n^{\tau_{min}}L}{w}p^r\left(v\left(n^{\tau_{min}}C\right)^{v-1} c + wR\right) \end{aligned}$$

Hence

$$N\left(u, \hat{\mathcal{G}}(T), \|\cdot\|_\infty\right) \le n^\kappa p^r \frac{L}{w}\left(v\left(n^{\tau_{min}}C\right)^{v-1} c + wR\right)$$

□

An analogous lemma holds for nonsmooth penalties P_j that satisfy the assumptions given in **??**.

3.4.2 Smoothing Splines with a Sobolev Penalty

Finally, we consider the classic nonparametric problem of fitting a smoothing spline using a Sobolev penalty. The function class of interest here is

$$\hat{\mathcal{G}}(T) = \left\{\hat{g}_\lambda \equiv \sum_{j=1}^J \hat{g}_\lambda(\cdot|T) : \hat{g}_\lambda(\cdot|T) = \arg\min_{g_j\in\mathcal{G}} \frac{1}{2}\|y - \sum_{j=1}^J g_j\|_T^2 + \sum_{j=1}^J \lambda_j \int (g_j^{(m)}(x))^2 dx, \lambda \in \Lambda\right\}$$

Lemma 4. *Suppose* $\sup_{g\in\mathcal{G}}\|g\|_\infty \le R$. *Suppose* $\Lambda = [n^{-\tau_{\min}}, n^{\tau_{\max}}]^J$. *Then the entropy is bounded above by*

$$H\left(u, \mathcal{G}(T), \|\cdot\|_D\right) \le J\left(2\log\frac{1}{u} + \kappa\log n + stuff\right) \tag{35}$$

Proof. First, note the following properties of the Sobolev norm. For any function h, we have

$$\left|\frac{\partial}{\partial m}P(g+mh)\right| = \left|2\int(g^{(m)}(x)+mh^{(m)}(x))h^{(m)}(x)dx\right| \leq 2\sqrt{P(g+mh)P(h)}$$

and

$$\frac{\partial^2}{\partial m^2}P(g+mh) = 2\int(h^{(m)}(x))^2dx = 2P(h)$$

Consider the function $h = c(g_\lambda - g_\delta)$ where c is some constant such that $P(h) = 1$ (Note that $P(h) = 0$ if and only if $g_\lambda \equiv g_\delta$).

Define the following one-dimensional optimization problem

$$\hat{m}_h(\lambda_0) = \arg\min_m \frac{1}{2}\|y - (\hat{g}_\delta + mh)\|_T^2 + \lambda_0 P(\hat{g}_\delta + mh)$$

Implicit differentiation of the KKT conditions, we get

$$\begin{aligned} \left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right| &\leq n^{\tau_{min}}\sqrt{P(g+mh)/P(h)} \\ &\leq n^{\tau_{\min}}\sqrt{\frac{n^{\tau_{\min}}}{2}\|\epsilon\|_T^2 + P(g^*)} \end{aligned}$$

where $\sqrt{P(g+mh)}$ is bounded using the same logic as in Lemma **??**.

By the mean value theorem, there is some $\alpha \in (\delta, \lambda)$ such that

$$\begin{aligned} \|g_\lambda - g_\delta\|_\infty &= \|\hat{m}_h(\lambda)h\|_\infty \\ &\leq |\lambda-\delta|R\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|_{\lambda_0=\alpha} \\ &\leq |\lambda-\delta|Rn^{\tau_{\min}}\sqrt{\frac{n^{\tau_{\min}}}{2}\|\epsilon\|_T^2 + P(g^*)} \end{aligned}$$

Hence

$$N\left(u, \hat{\mathcal{G}}(T), \|\cdot\|_\infty\right) \leq Rn^{\tau_{\max}-\tau_{\min}}\sqrt{\frac{n^{\tau_{\min}}}{2}\|\epsilon\|_T^2 + P(g^*)}$$

□

4 Simulations

In this section, we provide empirical evidence that supports the oracle inequalities we have found.

In this (first?) simulation, we show that the model chosen by a training/validation split framework converges to the oracle model at the $(\log(n)/n)^{1/2}$ rate. We generated observations from the model

$$y = sin(x_1) + sin(4x_2 + 1) + \sigma\epsilon \tag{36}$$

where $\epsilon \sim U(-1, 1)$ 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

$$\hat{g}_{1,\lambda}, \hat{g}_{2,\lambda} = \arg\min_{g_1,g_2} \|y - f_1(x_1) - f_2(x_2)\|_T^2 + \int_0^6 (f_1^{(2)}(x))^2 dx + \int_0^6 (f_2^{(2)}(x))^2 dx \tag{37}$$

The training set contained 30 samples. Penalty parameters were tuned using validation set sizes $n_V = 5, 10, ..., 30$. The oracle penalty parameters were chosen by minimizing over a separate test set of 400 samples. A total of 25 simulations were run for each validation set size.

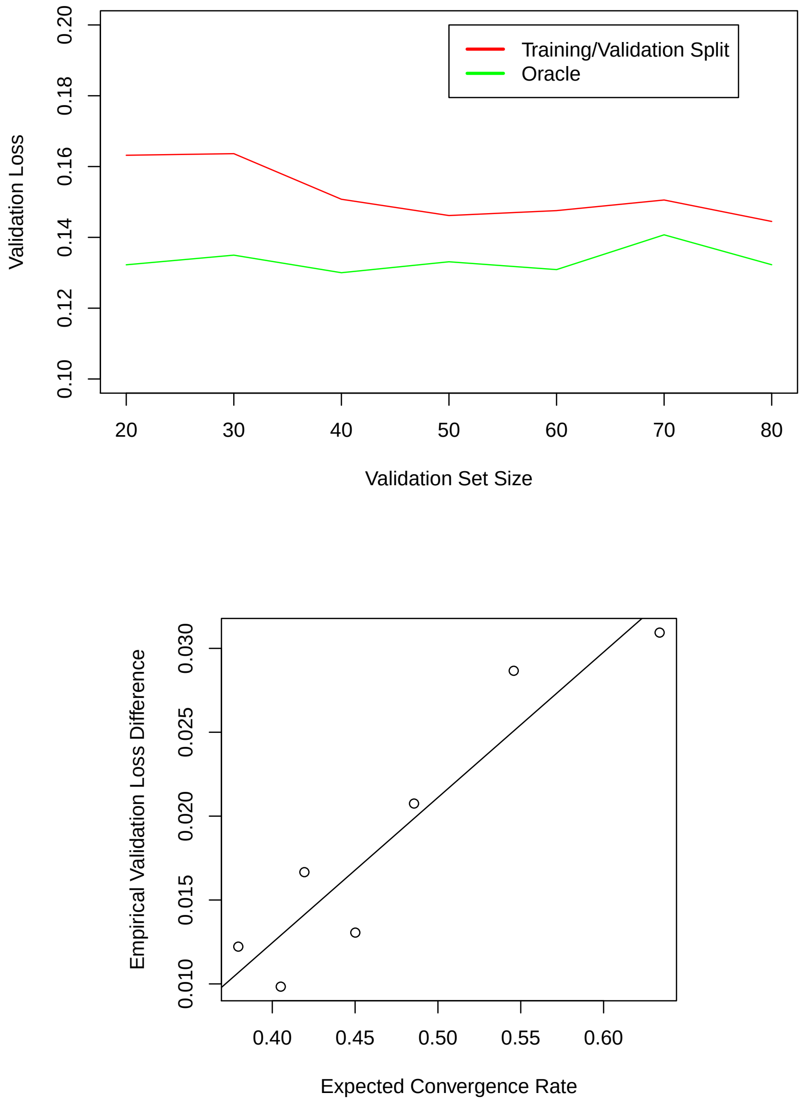
Figure 4 plots the validation loss $\|\hat{g}_\lambda - g^*\|_V$ of the model tuned using a validation set versus the model fit using the oracle penalty parameters. As the validation set increases, the error of the tuned model converges towards the oracle model as expected. In addition we compare the observed difference between the validation losses for the two models and the expected convergence rate of $(\log(n)/n)^{1/2}$. The plot shows that theory closely matches the empirical evidence.

Maybe a simulation on using lots of penalty parameters.

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. Contrary to popular opinion, adding penalty parameters does not drastically increase the model complexity. This finding supports recent efforts to combine regularization methods and "un-pool" regularization parameters. Since the fitted models are smoothly parameterized in terms of the penalty parameters, cross-validation over a continuum of penalty parameters does not increase the model complexity either.

Figure 1: Empirical vs. Theory



The main caveat is that we have proven results for a perturbed penalized regression problem, rather than the original. Determining the entropy of fitted models from the original penalized regression is still an open question.

Our theorems assume that the global minimizer has been found over the penalty parameter set, but 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 are local minima.

6 The Proof

Lemma 5. *The oracle rate isn't changed when we add the ridge penalty*

Proof. short proof □

Proof of Theorem 1

Proof. one page □

Proof of Entropy for nonsmooth penalties

Proof. one page, including the implicit function theorem. □