Oracle Inequalities for multiple penalty parameters

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

August 20, 2016

Abstract

In high-dimensional or non-parametric problems, regularization is used to control model complexity. Each penalty function is scaled by a penalty parameter that must be tuned. The oracle penalty parameters guarantee fast convergence rates but they depend on unknown constants. Therefore one usually tunes the penalty parameters by evaluating the fitted models on a validation set. In this paper, we provide finite sample oracle inequalities on the prediction error for models in which multiple penalty parameters are tuned over a continuum of values. We find that the fitted functions are smoothly parameterized by the penalty parameters. Hence the difference in prediction error between the chosen model and the oracle model decreases at a near-parametric rate. This result justifies recent work on having multiple penalties with separate penalty parameters and tuning penalty parameters using continuous optimization methods instead of relying on a pre-defined finite-sized grid of values.

Keywords: ...?

*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_i and p predictors x_i. Suppose y_i is generated from the true model g^* from model class $\mathcal{G}$

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

where ϵ_i are random errors. Penalized regression methods are important in high-dimensional ($p >> n$) or ill-posed problems as they control model complexity and induce desired structure. Here we will consider least squares regression, in which the model is estimated by minimizing a criterion of the form:

$$\hat{g}(\lambda) = \underset{g \in \mathcal{G}}{\arg\min} \|y - g(X)\|_n^2 + \sum_{j=1}^{J} \lambda_j P_j^{v_j}(g) \tag{2}$$

The penalty parameters λ_j ultimately determine the fitted model, so it is important to select them properly. In many cases, one can work out their oracle values to ensure fast convergence rates with high probability (Van de geer-book, Wahba-smoothing spline paper, and others?). For example, in the case of an additive model $f = \sum f_i$, the oracle set of penalty parameters are inversely proportional to the penalties of the true model (taken to some power). If the model is fit using the oracle penalty parameters, the convergence rates of each f_i is as fast as in the case where the other components are known (Vandegeer additive models). However, these oracle values commonly depend on unknown values.

Given the oracle penalty parameter values are unknown, one usually tunes the penalty parameters 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 chooses the penalty parameters with the lowest validation error. When $J \leq 2$, a simple grid search over the penalty parameters is used; when J is much larger, one must use continuous optimization methods. The machine learning literature addresses this "hyperparameter selection" problem using continuous optimization methods such as Bayesian optimization and gradient descent (Bengio, Foo, Feng, MacLaurin, Snoek).

The performance of cross-validation-like procedures is characterized by bounding the prediction error. Typically these inequalities provide an upper bound 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 (our inequality bound has faster convergence I think?). 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 very difficult computationally. Hence the most popular regularization methods only have at most two tuning parameters (e.g. Elastic Net, Sparse Group Lasso, etc.). Also, there is 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. The upper bound is composed of the error of the oracle and an empirical process term that converges at a near-parametric rate. In semi- and non-parametric problems, the error of the oracle term dominates, so the prediction error could be minimized by cross-validation over more penalty parameters. The proof takes a general approach of bounding the empirical process term using entropy methods (sara's book).

Section 1 provides the theorem. Section 2 provides simulation studies. Section 3 is a discussion. Section 4 provides the proof.

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 $\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V$, 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

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

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 $\mathcal{G}(T)$ 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 $\sup_{g\in\mathcal{G}(\cdot|\mathcal{T})} \|g\|_\infty \leq \infty$. Suppose for any training dataset $T \subseteq D$ with $\|\epsilon\|_T \leq 2\sigma$, we have*

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

Then with high probability, we have

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V \leq \min_{\lambda\in\Lambda} \|\hat{g}_\lambda(\cdot|T) - g^*\|_V + G\frac{\psi(u, n, J)}{\sqrt{n_V}} \tag{8}$$

Proof Sketch. Define $\langle h, \ell \rangle_A = \frac{1}{|A|} \sum_{i \in A} h(x_i)\ell(x_i)$. The basic inequality gives us

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V^2 \leq \|\hat{g}_{\tilde{\lambda}}(\cdot|T) - g^*\|_V^2 + 2\left|\langle \epsilon, \hat{g}_{\tilde{\boldsymbol{\lambda}}}(\cdot|T) - \hat{g}_{\hat{\boldsymbol{\lambda}}}(\cdot|T)\rangle_V\right| \tag{9}$$

By Lemma ?something?, for all

$$\delta \geq \left(\frac{\log n}{n_V}\right)^{1/2} \tag{10}$$

there is some constant c such that

$$Pr\left(\sup_{\lambda \in \Lambda} \frac{|\langle \epsilon, \hat{g}_{\boldsymbol{\lambda}}(\cdot|T) - \hat{g}_{\boldsymbol{\lambda}}(\cdot|T)\rangle_V|}{\|\hat{g}_{\boldsymbol{\lambda}}(\cdot|T) - \hat{g}_{\boldsymbol{\lambda}}(\cdot|T)\|_V} \geq \delta \wedge \|\epsilon\|_V \leq 2\sigma\right) \leq \exp\left(-n_V \frac{\delta^2}{c}\right) \tag{11}$$

Also, by Bernstein's inequality, we have that

$$Pr\left(\|\epsilon\|_V \leq 2\sigma\right) \leq \exp\left(-n_V \frac{\sigma^2}{K}\right) \tag{12}$$

Therefore the result in (21) holds with high probability. □

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 **??**, we will show that $\hat{g}_\lambda$ is smoothly parameterized by λ in many penalized regression problems. That is, we will show that for any $d > 0$, there is some $\delta > 0$ such that for all

$$\|\lambda_1 - \lambda_2\| \leq \delta \implies \|\hat{g}_{\lambda_1} - \hat{g}_{\lambda_2}\|_V \leq d \tag{13}$$

This implies that the covering number of $\mathcal{G}(T)$ has the form:

$$H(u, \mathcal{G}(T), \|\cdot\|_V) \leq \log \frac{1}{u} + \log \frac{\lambda_{\max}}{\lambda_{\min}} + C \tag{14}$$

The first term on the right hand size is the usual metric entropy of a parametric family. The second term is the price we pay for searching over a large space.

Applying Theorem 1 to the setting of penalized regression, we have the following corollary

Corollary 1. *Suppose that* $\sup_{g \in \mathcal{G}(\cdot|\mathcal{T})} \|g\|_\infty \leq \infty$. *If for any* $d > 0$, *there is some* $\delta = O_p(d^2)$ *such that for all*

$$\|\lambda_1 - \lambda_2\| \leq \delta \implies \|\hat{g}_{\lambda_1} - \hat{g}_{\lambda_2}\|_V \leq d \tag{15}$$

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

Then with high probability, we have

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V \leq \min_{\lambda \in \Lambda} \|\hat{g}_\lambda(\cdot|T) - g^*\|_V + G\sqrt{\frac{\kappa \log n + C + 2}{n_V}} \tag{16}$$

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:

$$\hat{\lambda} = \arg\min_{\lambda \in \Lambda} \frac{1}{2} \sum_{k=1}^{K} \|y - \hat{g}_\lambda(\cdot|D_{-k})\|_k^2 \tag{17}$$

$$\hat{g}(\lambda|D_{-k}) = \arg\min_{g \in \mathcal{G}} \frac{1}{2}\|y - g\|_{-k}^2 + \sum_{j=1}^{J} \lambda_j P_j^{v_j}(g) + \frac{w}{2}\|g\|^2 \tag{18}$$

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":

$$\frac{1}{K} \sum_{k=1}^{K} \hat{g}_{\hat{\boldsymbol{\lambda}}}(\cdot|D_{-k}) \tag{19}$$

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 for any $d > 0$, there is some $\delta = O_p(d^2)$ such that for all

$$\|\lambda_1 - \lambda_2\| \leq \delta \implies \|\hat{g}_{\lambda_1} - \hat{g}_{\lambda_2}\|_\infty \leq d \tag{20}$$

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

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

$$\left\| \frac{1}{K} \sum_{k=1}^{K} \hat{g}(\hat{\lambda}|D_{-k}) - g^* \right\|^2 \leq (1 + a) \min_{k \in 1:K, \lambda \in \Lambda} \|\hat{g}(\lambda|D_{-k}) - g^*\|^2 + c_a \max_{k=1:K} \frac{\log^2(n_k)}{n_k} \tag{21}$$

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:

$$\lambda \sum_{j=1}^{J} P_j^{v_j}(g_j) \to \sum_{j=1}^{J} \lambda_j P_j^{v_j}(g_j) \tag{22}$$

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 Covering number/Entropy of the fitted models

The results in Section **??** hinge on bounding the metric entropy of the function class $\mathcal{G}(T)$. We approach this by showing 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. The former can be shown for a general set of penalized regression problems; we will consider problems with smooth penalties and then those with nonsmooth penalties. The latter is harder to show and so we will consider two

specific examples: parametric regression problems (where p can grow with n) and smoothing splines.

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

3.1 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 λ, consider the one-dimensional optimization problem

$$\hat{m}_h(\lambda) = \arg\min_m \frac{1}{2}\|y - (\hat{g}_\delta + mh)\|_T^2 + \sum_{j=1}^J \lambda_j P_j^{v_j}(\hat{g}_\delta + mh) \tag{23}$$

Suppose the penalty functions P_j are twice-differentiable everywhere. Then the KKT conditions states that

$$\langle h, y - (\hat{g}_\delta + mh)\rangle + \sum_{j=1}^J \lambda_j \frac{\partial}{\partial m} P_j^{v_j}(\hat{g}_\delta + mh) = 0 \tag{24}$$

Implicit differentiation of (24) with respect to λ_ℓ for $\ell = 1, ..., J$ gives us

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

The primary challenge is bounding the first term.

3.2 Bounds on the Metric Entropy over the Validation Set

Bounding the entropy of $\mathcal{G}(T)$ directly is difficult at this level of generality. Instead, we will consider the function class when the training criterion is slightly perturbed:

$$\hat{g}(\lambda) = \arg\min_{g \in \mathcal{G}} \|y - g(X)\|_n^2 + \sum_{j=1}^J \lambda_j \left(P_j^{v_j}(g) + \frac{w}{2}\|g\|_V^2\right) \tag{25}$$

Under certain regularity assumptions, one could probably show that the addition of w only modifies the fitted model slightly. In practice, one could certainly choose w sufficiently small such that the model fit is not different from when $w = 0$. In Lemma ?something?, we show that the additional ridge penalty does not affect the oracle convergence rate.

Nonetheless, the importance of the ridge penalty in our proof is interesting. Adding the ridge penalty allows us to characterize the model class and thereby increases the stability of its estimates.

We use a proof by contradiction to bound the metric entropy with respect to $\|\cdot\|_V$.

3.3 Smooth Norms

Suppose the penalties P_j are semi-norms that are differentiable everywhere. The entropy is bounded using an implicit differentiation trick.

Lemma 1. *Suppose the penalty functions P_j are smooth norms and that $v_j \geq 1$. Suppose $\sup_{g\in\mathcal{G}} \|g\| \leq G$. Suppose $\Lambda = [n^{-\tau_{\min}}, n^{\tau_{\max}}]^J$. Then the entropy is bounded above by*

$$H(u, \mathcal{G}(T), \|\cdot\|_V) \leq J\left(2\log\frac{1}{u} + \kappa\log n + \log\frac{C}{Jw}\right) \tag{26}$$

where

$$C = \sqrt{2}\left(2v_{\max}(1+J)c + wc^{1/v_{min}}G\right)$$

and

$$c = \frac{1}{2}\|\epsilon\|_T^2 + n^{\tau_{max}}\sum_{j=1}^{J}\left(P_j^{v_j}(g^*) + \frac{w}{2}\|g^*\|_D^2\right)$$

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.

Let

$$\delta(d) = \left(Cd^{-2}n^c w^{-1} v\left(\|\epsilon\|_T^2 + P^v(g^*) + \frac{w}{2}\|g^*\|_D^2 + G\right)\right)^{-1}$$

We will show that the following set $\Omega_{\delta(d)}$ forms a d-cover set for $\hat{\mathcal{G}}(T, \epsilon_T)$:

$$\Omega_{\delta(d)} = \left\{\hat{g}_{\delta_i}(\cdot|T) : \delta_i = i\delta(d) + \lambda_{min} \text{ for } i = 0, ..., \left\lceil\frac{\lambda_{max} - \lambda_{min}}{\delta(d)}\right\rceil\right\}$$

Consider any $\lambda \in [\lambda_{min}, \lambda_{max}]$ and suppose $\delta_i < \lambda < \delta_{i+1}$. Let $h = \hat{g}_{\delta_i}(\cdot|T) - \hat{g}_\lambda(\cdot|T)$. Suppose $\|h\|_D > d$ for contradiction.

Consider the one-dimensional problem as done in Section 3.1

$$\hat{m}_h(\lambda_0) = \arg\min_m \frac{1}{2}\|y - (\hat{g}_{\delta_i} + mh)\|_T^2 + \lambda_0\left(P^v(\hat{g}_{\delta_i} + mh) + \frac{w}{2}\|\hat{g}_{\delta_i} + mh\|_D^2\right)$$

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

$$\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right| \leq \frac{n^{\tau_{\min}}}{wd^2}\left|\frac{\partial}{\partial m}P^v(\hat{g}_{\delta_i}+mh)+w\langle h,\hat{g}_{\delta_i}+mh\rangle_D\right|_{m=\hat{m}_\lambda(\lambda_0)} \tag{27}$$

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:

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

Combining these facts, we get that

$$\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right| \leq Cd^{-2}n^c w^{-1}v\left(\|\epsilon\|_T^2+P^v(g^*)+\frac{w}{2}\|g^*\|_D^2+G\right)$$

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

$$\begin{aligned}|\hat{m}_h(\lambda)-\hat{m}_h(\delta_i)| &= (\lambda-\delta_i)\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|_{\lambda_0=\alpha} &(28)\\ &\leq \delta(d)\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|_{\lambda_0=\alpha} &(29)\\ &\leq 1/2 &(30)\end{aligned}$$

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

3.4 Nonsmooth penalties

If the regression problem contains non-smooth penalty functions, similar results do not necessarily hold. The key problem is that the entropy of the function class defined in (3) may not well-controlled. 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. Hence their entropy is actually the same as that in (14), modulo some constant.

To characterize such problems, we need the following definitions:

Definition 2. *The differentiable space of a real-valued function L at $\boldsymbol{\eta}$ in its domain is the set such that*

$$\Omega^L(\boldsymbol{\eta})=\left\{\boldsymbol{u}\middle|\lim_{\epsilon\to0}\frac{L(\boldsymbol{\eta}+\epsilon\boldsymbol{u})-L(\boldsymbol{\eta})}{\epsilon}\ \textit{exists}\right\} \tag{31}$$

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 $\boldsymbol{\lambda} \in W$,*

$$\arg\min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in S} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{32}$$

It follows that as long as the local optimality space is a subset of the differentiable space, the function class in (3) satisfies the following entropy bound.

Lemma 2. *For almost every $\boldsymbol{\lambda}$, the differentiable space $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}))$ is a local optimality space for $L_T(\cdot, \boldsymbol{\lambda})$. Suppose the penalty functions P_j are semi-norms that are smooth almost everywhere and that $v_j \geq 1$. Suppose $\sup_{g \in \mathcal{G}(\mathcal{T})} \|g_\lambda\| \leq G$. Suppose $\Lambda = [n^{-\tau_{\min}}, n^{\tau_{\max}}]^J$. Then the entropy for non-smooth functions is bounded by*

$$H(u, G, \|\cdot\|_D) \leq J\left(2\log\frac{1}{u} + \kappa \log n + stuff\right) \tag{33}$$

The proof is requires using the implicit function theorem to show that $\nabla_\lambda L$ exists. The proof is given in Section 6.

3.5 Entropy Bounds over the full domain

3.5.1 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.

$$\hat{\theta}(\lambda) = \arg\min_{\theta \in \Theta} \|y - g(X)\|_n^2 + \sum_{j=1}^{J} \lambda_j \left(P_j^{v_j}(\theta) + \frac{w}{2}\|\theta\|_2^2\right) \tag{34}$$

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

Lemma 3. *Suppose*

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

and the penalty functions P_j are norms that are smooth and

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

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

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

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

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

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. $\|\beta\|_2 \leq 1$. 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

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

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

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

Hence

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

□

An analogous lemma holds for nonsmooth penalties P_j that satisfy the assumptions given in 2.

3.5.2 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

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

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

$$H\left(u, \mathcal{G}(T), \|\cdot\|_D\right) \leq J\left(2\log\frac{1}{u} + \kappa \log n + r\log p + stuff\right) \tag{36}$$

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

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

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 26.

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

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

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

$$y = sin(x_1) + sin(4x_2 + 1) + \sigma\epsilon \tag{37}$$

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

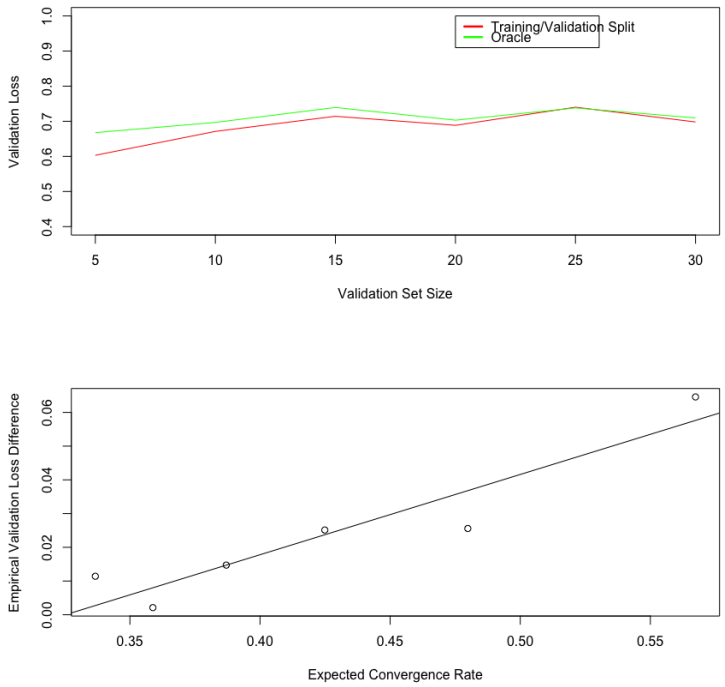
$$\hat{g}_{1,\lambda}, \hat{g}_{2,\lambda} = \underset{g_1,g_2}{\arg\min} \|y - f_1(x_1) - f_2(x_2)\|_T^2 + \int_0^6 (f_1^{(2)}(x))^2 dx + \int_0^6 (f_2^{(2)}(x))^2 dx \tag{38}$$

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.

Figure 1: Empirical vs. Theory



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.

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. □

7 Other things