Oracle Inequalities for multiple penalty parameters

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

August 19, 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 are usually unknown. 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 of models chosen by a training/validation split. We find that tuning multiple penalty parameters over a continuum of values only increases the oracle error rate by a near-parametric rate. This result justifies recent work on combining regularization methods 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:

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

The penalty parameters λ_j ultimately determine the fitted model, so it is important to select them properly. Their oracle values give fast convergence rates (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 allow the convergence rates of each f_i to be as fast as in the case where the other components are known (Vandegeer additive models). However, these oracle values commonly depend on unknown values, such as the complexity of the true model and the magnitude of the noise variables.

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 can be understood using oracle inequalities on 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

We will consider the problem of tuning our penalty parameters over the J-dimensional box $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$ by minimizing the validation loss. Let D be the observed data of size n.

Suppose it is split into a training set T of size n_T and validation set V of size n_V. For a given set of data A with size $|A|$, define $\|h\|_A^2 = \frac{1}{|A|}\sum_{i \in A} h^2(x_i)$. Define $\langle h, \ell \rangle_A = \frac{1}{|A|}\sum_{i \in A} h(x_i)\ell(x_i)$. In this section, we will bound the mean squared error over the validation set $\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V$.

Let the fitted models over the range of penalty parameter values Λ be denoted

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

Let the final penalty parameter chosen by the training/validation split be denoted $\hat{\lambda}$, which by definition satisfies

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

To handle the fact that (**??**) could potentially be a dense, multi-dimensional set, we will use results from empirical process theory. In particular, we will consider the metric entropy of (**??**), a classical measure of the complexity of a function class. 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:*

$$H(u, \mathcal{G}, \|\cdot\|) = \log N(u, \mathcal{G}, \|\cdot\|) \tag{5}$$

Standard chaining and peeling arguments then give us a finite sample upper bound on the mean squared prediction error of $\hat{g}_{\hat{\lambda}}(\cdot|T)$ over the validation points.

Theorem 1. *Suppose ϵ are independent sub-Gaussian random variables. Suppose for any training dataset $T \subseteq D$, if $\|\epsilon\|_T \le 2\sigma$, then we*

$$\int_0^R H^{1/2}\left(u, \mathcal{G}(\cdot|\mathcal{T})\|\cdot\|_D\right) du \le \psi(u, n, J) \tag{6}$$

Let $\tilde{\lambda} \in \Lambda$ be the oracle set of penalty parameters. Then with high probability, we have

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V \le \|\hat{g}_{\tilde{\lambda}}(\cdot|T) - g^*\|_V + G\frac{\psi(u, n, J)}{\sqrt{n_V}} \tag{7}$$

Proof Sketch. The basic inequality gives us

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V^2 \le \|\hat{g}_{\tilde{\lambda}}(\cdot|T) - g^*\|_V^2 + 2\left|\langle \epsilon, \hat{g}_{\tilde{\boldsymbol{\lambda}}}(\cdot|T) - \hat{g}_{\hat{\boldsymbol{\lambda}}}(\cdot|T)\rangle_V\right| \tag{8}$$

By Lemma ?something?, for all

$$\delta \geq \left(\frac{\log n}{n_V}\right)^{1/2} \tag{9}$$

there is some constant c such that

$$Pr\left(\sup_{\lambda\in\Lambda} \frac{|\langle \epsilon, \hat{g}_{\boldsymbol{\lambda}}(\cdot|T) - \hat{g}_{\boldsymbol{\lambda}}(\cdot|T)\rangle_V|}{\|\hat{g}_{\boldsymbol{\lambda}}(\cdot|T) - \hat{g}_{\boldsymbol{\lambda}}(\cdot|T)\|_V} \geq \delta \wedge \|\epsilon\|_V \leq 2\sigma\right) \leq \exp\left(-n_V \frac{\delta^2}{c}\right) \tag{10}$$

Also, by Bernstein's inequality, we have that

$$Pr\left(\|\epsilon\|_V \leq 2\sigma\right) \leq \exp\left(-n_V \frac{\sigma^2}{K}\right) \tag{11}$$

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

From Theorem **??**, we see that the key to bounding the validation loss is to bound the entropy of the fitted models in (**??**). The theorem is very general, so one could conceivably apply this to various other regression problems.

Here we focus on the penalized regression setting. Bounding the fitted models from minimizing (**??**) is difficult, so we consider models that fit the training criterion with a slight perturbation. That is, we will consider the functions

$$\hat{g}_\lambda(\cdot|T) = \arg\min_{g\in\mathcal{G}} \frac{1}{2}\|y-g\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j^{v_j}(g) + \frac{w}{2}\|g\|_D^2\right) \tag{12}$$

where w is some positive constant. Of course if the existing penalties already bound the additional ridge penalty by some constant (e.g. Elastic Net), it is sufficient to set $w = 0$. As shown in the following section, the ridge penalty implies that $\hat{g}_\lambda(\cdot|T)$ evaluated over the observed covariates is smoothly parametrized by λ. Thus the entropy bound is very similar to that for parametric models, with an additional $\log n$ term

$$H(u, \mathcal{G}(T), \|\cdot\|_D) \leq \log\frac{1}{u} + \kappa \log n \tag{13}$$

for some constant κ dependent on things. The $\log n$ term results from the range of Λ increasing at some polynomial rate. The rate the oracle λ decreases in n is unknown (due to unknown constants), so one should have the lower limit of Λ decreasing at a rate that essentially guarantees that the oracle λ is in Λ.

Applying this entropy bound to Theorem **??**, we get the following corollary

Corollary 1. *Suppose in Theorem* **??** *that*

$$H(u, \mathcal{G}(T), \| \cdot \|_D) \le \log \frac{1}{u} + \kappa \log n \tag{14}$$

Then with high probability, we have

$$\|\hat{g}_{\hat{\lambda}}(\cdot|T) - g^*\|_V \le \|\hat{g}_{\tilde{\lambda}}(\cdot|T) - g^*\|_V + G\sqrt{\frac{\log n}{n_V}} \tag{15}$$

The fact that the model fit from cross-validation $\hat{g}_{\hat{\lambda}}$ converges to the oracle model $\hat{g}_{\tilde{\lambda}}$ at a parametric rate makes intuitive sense. Λ is a J-dimensional grid so tuning penalty parameters is just solving a J-dimensional optimization problem. In semi-/non-parametric settings, adding penalty parameters is therefore "cheap"; adding more penalty parameters incurs only an increase in the prediction error by a parametric rate. In fact, one could conceivably increase the number of penalty parameters at some polynomial rate in n to minimize the upper bound. This hinges on the fact that one know how the oracle rate decreases with an increasing number of penalty parameters.

Theorem **??** 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_

We recognize that these results unfortunately only apply to our perturbed regression problem with the additional ridge penalty. 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. The importance of the ridge penalty in our proof is interesting though. It seems to suggest that regularization methods with a ridge penalty are indeed more stable.

Next we derive the entropy bounds for (**??**). We will consider smooth and non-smooth penalty functions separately.

2.1.1 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 $\sup_{g\in\mathcal{G}(\mathcal{T})} \|g_\lambda\| \leq G$. Suppose $\Lambda = [n^{-\tau_{\min}}, n^{\tau_{\max}}]^J$. Then the entropy is bounded above by*

$$H(u, \mathcal{G}(T), \|\cdot\|_V) \leq J\left(2\log\frac{1}{u} + \kappa\log n + \log\frac{C}{Jw}\right) \tag{16}$$

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 with any λ_0

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

Clearly $\hat{m}_h(\delta_i) = 0$ and $\hat{m}_h(\lambda) = 1$. By the mean-value theorem, there is some $\alpha \in (\delta_i, \lambda)$ such that

$$\hat{m}_h(\lambda) = (\lambda - \delta_i)\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|_{\lambda_0=\alpha} \leq \delta(d)\left|\frac{\partial}{\partial\lambda_0}\hat{m}_h(\lambda_0)\right|_{\lambda_0=\alpha}$$

We use implicit differentiation of the KKT conditions with respect to λ_0 to get

$$\frac{\partial}{\partial \lambda_0}\hat{m}_h(\lambda_0) \;=\; -\left(\|h\|_T^2 + \lambda_0 \frac{\partial^2}{\partial m^2} P^v\left(\hat{g}_{\delta_i} + mh\right) + \lambda_0 w \|h\|_D^2\right)^{-1} \left(\frac{\partial}{\partial m} P^v(\hat{g}_{\delta_i} + mh) + w\langle h, \hat{g}_{\delta_i} + mh\rangle_D\right)\Bigg|$$

Since penalty P is convex and by the assumption that $\|h\|_D \geq d$, we have

$$\left|\|h\|_T^2 + \lambda_0 \frac{\partial^2}{\partial m^2} P^v\left(\hat{g}_{\delta_i} + mh\right) + \lambda_0 w \|h\|_D^2\right|^{-1} \;\leq\; n^{\tau_{min}} w^{-1} d^{-2}$$

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. After some algebra, we get

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

Hence the mean-value theorem tells that $\hat{m}_h(\lambda) \leq 1/2$, which is a contradiction. □

2.1.2 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 (**??**) 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 (**??**), 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*

$$\Omega^L(\boldsymbol{\eta}) = \left\{ \boldsymbol{u} \middle| \lim_{\epsilon \to 0} \frac{L(\boldsymbol{\eta} + \epsilon \boldsymbol{u}) - L(\boldsymbol{\eta})}{\epsilon} \text{ exists} \right\} \tag{17}$$

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

$$\arg\min_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in S} L(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{18}$$

It follows that as long as the local optimality space is a subset of the differentiable space, the function class in (**??**) 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 $v_j \ge 1$. Suppose $\sup_{g\in\mathcal{G}(\mathcal{T})}\|g_\lambda\| \le G$. Suppose $\Lambda = [n^{-\tau_{\min}}, n^{\tau_{\max}}]^J$. Then the entropy for non-smooth functions is bounded by*

$$H\left(u, G, \|\cdot\|_D\right) \le J\left(2\log\frac{1}{u} + \kappa\log n + stuff\right) \tag{19}$$

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

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 expected prediction error rather than the prediction error on the validation set. Toward this end, we will apply the results by Mitchell on the modified average CV procedure.

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:

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda}\frac{1}{2}\sum_{k=1}^{K}\|y - \hat{g}_\lambda(\cdot|D_{-k})\|_k^2 \tag{20}$$

$$\hat{g}(\lambda|D_{-k}) = \arg\min_{g\in\mathcal{G}}\frac{1}{2}\|y-g\|_{-k}^2 + \sum_{j=1}^{J}\lambda_j P_j^{v_j}(g) + \frac{w}{2}\|g\|^2 \tag{21}$$

Define the modified average CV model as

$$\frac{1}{K}\sum_{k=1}^{K}\hat{g}_{\hat{\lambda}}(\cdot|D_{-k}) \tag{22}$$

We need the following set of assumptions:

- tail behavior of the loss function (orlicz norm)
- margin behavior of the loss function (L2 norm)
- Errors are bounded: $\|\epsilon\|_\infty < \infty$

- Penalty is not too crazy $\|g_{\lambda_1} - g_{\lambda_2}\| \ge n^{-p} P(g_{\lambda_1} - g_{\lambda_2})$

Theorem **??** then gives a bound on the expected prediction error of the modified average CV model.

Theorem 2. *Suppose blah. With high probability, we have*

$$\left\| \frac{1}{K} \sum_{k=1}^{K} \hat{g}(\hat{\lambda}|D_{-k}) - g^* \right\|_D \le \sqrt{\sum_{k=1}^{K} \|\hat{g}(\tilde{\lambda}|D_{-k}) - g^*\|_k^2} + G \left(J \frac{\kappa \log n + otherthings}{\min_{k=1:K} n_k} \right)^{1/2} \tag{23}$$

3 Simulations

We show that the empirical process term is indeed very small.

Maybe a simulation on using lots of penalty parameters.

4 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.

5 The Proof

Lemma 3. *The oracle rate isn't changed when we add the ridge penalty*

Proof. short proof □

Proof of Theorem ??

Proof. one page □

Proof of Entropy for nonsmooth penalties

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

6 Other things