Statistica Sinica **30** (2020), 511-530
doi:https://doi.org/10.5705/ss.202017.0310

AN ANALYSIS OF THE COST OF HYPERPARAMETER SELECTION VIA SPLIT-SAMPLE VALIDATION, WITH APPLICATIONS TO PENALIZED REGRESSION

Jean Feng and Noah Simon

University of Washington

Abstract: In a regression setting, a model estimation procedure constructs a model from training data for given a set of hyperparameters. The optimal hyperparameters that minimize the generalization error of the model are usually unknown. Thus, in practice, they are often estimated using split-sample validation. However, how the generalization error of the selected model grows with the number of hyperparameters to be estimated remains an open question. To address this, we establish finite-sample oracle inequalities for selection based on a single training/test split and cross-validation. We show that if the model estimation procedures are smoothly parameterized by the hyperparameters, the error incurred from tuning the hyperparameters shrinks at a near-parametric rate. Hence for semiparametric and nonparametric model estimation procedures with a fixed number of hyperparameters, this additional error is negligible. For parametric model estimation procedures, adding a hyperparameter is roughly equivalent to adding a parameter to the model itself. In addition, we specialize these ideas for penalized regression problems with multiple penalty parameters. We establish that the fitted models are Lipschitz in the penalty parameters and, thus, our oracle inequalities apply. This result encourages the development of regularization methods with many penalty parameters.

Key words and phrases: Cross-validation, regression, regularization.

1. Introduction

Per the usual regression framework, suppose we observe response $y \in \mathbb{R}$ and predictors $\boldsymbol{x} \in \mathbb{R}^p$. Suppose y is generated by a true model g^* plus random error ϵ with mean zero, i.e. $y = g^*(\boldsymbol{x}) + \epsilon$. Our goal is to estimate g^*. Many model estimation procedures can be formulated as selecting a model from some function class $\mathcal{G}$, given training data T and J-dimensional hyperparameter vector $\boldsymbol{\lambda}$. For example, in penalized regression problems, the fitted model can be expressed as the minimizer of the penalized training criterion

$$\hat{g}(\boldsymbol{\lambda}|T) = \arg\min_{g \in \mathcal{G}} \sum_{(\boldsymbol{x}_i, y_i) \in T} (y_i - g(\boldsymbol{x}_i))^2 + \sum_{j=1}^{J} \lambda_j P_j(g), \tag{1.1}$$

where P_j are penalty functions and λ_j are penalty parameters that serve as hyperparameters of the model estimation procedure.

If Λ is a set of possible hyperparameters, the goal is to find a penalty parameter $\boldsymbol{\lambda} \in \Lambda$ that minimizes the expected generalization error $\mathbb{E}[(y - \hat{g}(\boldsymbol{\lambda}|T)(\boldsymbol{x}))^2]$. Typically one uses a sample-splitting procedure where models are trained on a random partition of the observed data and evaluated on the remaining data. One then chooses the hyperparameter $\hat{\boldsymbol{\lambda}}$ that minimize the error on this validation set. For a more complete review of cross-validation, refer to Arlot and Celisse (2010).

The performance of split-sample validation procedures is typically characterized by an oracle inequality that bounds the generalization error of the expected model selected from the validation set procedure. For Λ that are finite, oracle inequalities have been established for a single training/validation split (Györfi et al. (2002, Chap. 7)) and a general cross-validation framework (van der Laan and Dudoit (2003); van der Laan, Dudoit and Keles (2004)). To handle Λ over a continuous range, one can use entropy-based approaches (Lecué and Mitchell (2012)).

The goal of this paper is to characterize the performance of models when the hyperparameters are tuned by some split-sample validation procedure. We are particularly interested in an open question raised in Bengio (2000): what is the "amount of overfitting... when too many hyperparameters are optimized"? In addition, how many hyperparameters is "too many"? Here, we show that a large number of hyperparameters can be tuned without overfitting. In fact, if an oracle estimator converges at rate $R(n)$, then the number of hyper parameters J can grow at a rate of approximately $J = O_p(nR(n))$, up to log terms, without affecting the convergence rate. In practice, for penalized regression, this means that we can propose and tune over much more complex models than are currently often used.

To show these results, we prove that finite-sample oracle inequalities of the form

$$\mathbb{E}\left[\left(y - \hat{g}(\hat{\boldsymbol{\lambda}}|T)(\boldsymbol{x})\right)^2\right] \leq (1+a) \underbrace{\inf_{\lambda \in \Lambda} \mathbb{E}\left[(y - \hat{g}(\boldsymbol{\lambda}|T)(\boldsymbol{x}))^2\right]}_{\text{Oracle risk}} + \delta\,(J, n) \tag{1.2}$$

are satisfied with high probability for some constant $a \geq 0$ and remainder $\delta(J, n)$,

which depends on the number of tuned hyperparameters J and the number of samples n. Under the assumption that the model -estimation procedure is Lipschitz in the hyperparameters, we find that δ scales linearly in J. For parametric model estimation procedures, the additional error incurred from tuning hyperparameters is roughly $O_p(J/n)$, which is similar to the typical parametric model estimation rate $O_p(p/n)$ where the model parameters are not regularized. For semiparametric and nonparametric model estimation procedures, this error is generally dominated by the oracle risk. Therefore, we can increase the number of hyperparameters without affecting the asymptotic convergence rate.

In addition, we specialize our results to penalized regression models of the form given in (1.1). The models in our examples are Lipschitz; thus, our oracle inequalities apply. This suggests that multiple penalty parameters may improve the model estimation and that recent proposals for penalty functions (e.g., elastic net and sparse group lasso (Zou and Hastie (2003); Simon et al. (2013))) may have artificially restricted themselves to two-way combinations.

During our literature search, we found few theoretical results relating the number of hyperparameters to the generalization error of the selected model. Much of the existing work considers tuning a one-dimensional hyperparameter over a finite Λ, proving asymptotic optimality (van der Laan, Dudoit and Keles (2004)) and finite-sample oracle inequalities (van der Laan and Dudoit (2003); Györfi et al. (2002)). Others have addressed split-sample validation for specific penalized regression problems with a single penalty parameter, such as linear model selection (Li (1987); Shao (1997); Golub, Heath and Wahba (1979); Chetverikov, Liao and Chernozhukov (2016); Chatterjee and Jafarov (2015)). Only the results in Lecué and Mitchell (2012) are relevant to answering our question of interest. A potential reason for this dearth of literature is that, historically, tuning multiple hyperparameters was computationally difficult. However, many recent proposals have addressed this computational hurdle (Bengio (2000); Foo, Do and Ng (2008); Snoek, Larochelle and Adams (2012)).

Section 2 presents oracle inequalities for sample-splitting procedures to understand how the number of hyperparameters affects the model error. Section 3 applies these results to penalized regression models. Section 4 provides a simulation study to support our theoretical results. Oracle inequalities for general model estimation procedures and proofs are given in the online Supplementary Material.

2. Oracle Inequalities

Here, we establish oracle inequalities for models in which the hyperparameters are tuned using a single training/validation split and cross-validation. We examine model estimation procedures that vary smoothly in their hyperparameters, because such procedures tend to be easier to use and, therefore, are more popular.

Let $D^{(n)}$ denote a data set with n samples. Given data set training data $D^{(m)}$, let $\hat{g}^{(m)}(\boldsymbol{\lambda}|D^{(m)})$ be some model estimation procedure that maps hyperparameter $\boldsymbol{\lambda}$ to a function in $\mathcal{G}$. We assume the following Lipschitz-like assumption on the model estimation procedure. In particular, we suppose that for any $\boldsymbol{x}$, the predicted value $\hat{g}^{(m)}(\boldsymbol{\lambda}|D^{(m)})(\boldsymbol{x})$ is Lipschitz in $\boldsymbol{\lambda}$.

Assumption 1. Suppose there is a set $\mathcal{X}^{(L)} \subseteq \mathcal{X}$ such that for any $n_T \in \mathbb{N}$ and data set $D^{(n_T)}$, there is a function $C_\Lambda(\boldsymbol{x}|D^{(n_T)}) : \mathcal{X}^{(L)} \mapsto \mathbb{R}^+$ such that for any $\boldsymbol{x} \in \mathcal{X}^{(L)}$, we have for all $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda$

$$\left|\hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(1)}|D^{(n_T)})(\boldsymbol{x}) - \hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(1)}|D^{(n_T)})(\boldsymbol{x})\right| \leq C_\Lambda(\boldsymbol{x}|D^{(n_T)})\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|_2. \tag{2.1}$$

In Section 3, we provide examples of penalized regression models that satisfy this assumption.

2.1. A single training/validation split

In the training/validation split procedure, the data set $D^{(n)}$ is randomly partitioned into a training set $T = (X_T, Y_T)$ and validation set $V = (X_V, Y_V)$ with n_T and n_V observations, respectively. The selected hyperparameter $\hat{\boldsymbol{\lambda}}$ is a minimizer of the validation loss

$$\hat{\boldsymbol{\lambda}} \in \arg\min_{\boldsymbol{\lambda}\in\Lambda} \frac{1}{2}\left\|y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T)\right\|_V^2, \tag{2.2}$$

where $\|h\|_V^2 \coloneqq (1/n_V)\sum_{(x_i,y_i)\in V} h^2(x_i, y_i)$ for function h.

We now present a finite-sample oracle inequality for the single training/validation split, assuming Assumption 1 holds. Our oracle inequality is sharp, that is, $a = 0$ in (1.2), unlike in most other works (Györfi et al. (2002); Lecué and Mitchell (2012); van der Laan and Dudoit (2003)). Note that the result below is a special case of Theorem 3 in the Supplementary Material S1.1, which applies to general model estimation procedures.

Theorem 1. *Let* $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$, *where* $\Delta_\lambda = \lambda_{\max} - \lambda_{\min} \geq 0$. *Suppose random variables* ϵ_i *from the validation set* V *are independent with expectation*

zero and are uniformly sub-Gaussian with parameters b and B:

$$\max_{i:(x_i,y_i)\in V} B^2\left(\mathbb{E}e^{|\epsilon_i|^2/B^2}-1\right)\le b^2.$$

Let the oracle risk be denoted as

$$\tilde{R}(X_V|T)=\arg\min_{\lambda\in\Lambda}\left\|g^*-\hat{g}^{(n_T)}(\boldsymbol{\lambda}|T)\right\|_V^2. \tag{2.3}$$

Suppose Assumption 1 *is satisfied over the set X_V. Then, there is a constant $c>0$ depending only on b and B such that for all δ satisfying*

$$\delta^2\ge c\left(\frac{J\log(\|C_\Lambda(\cdot|T)\|_V\Delta_\Lambda n+1)}{n_V}\vee\sqrt{\frac{J\log(\|C_\Lambda(\cdot|T)\|_V\Delta_\Lambda n+1)}{n_V}\tilde{R}(X_V|T)}\right), \tag{2.4}$$

we have

$$\begin{aligned}&Pr\left(\left\|g^*-\hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2-\tilde{R}(X_V|T)\ge\delta^2\middle|T,X_V\right)\\&\le c\exp\left(-\frac{n_V\delta^4}{c^2\tilde{R}(X_V|T)}\right)+c\exp\left(-\frac{n_V\delta^2}{c^2}\right).\end{aligned} \tag{2.5}$$

Theorem 1 states that, with high probability, the excess risk (e.g., the error incurred during the hyperparameter selection process) is no more than δ^2. As seen in (2.4), δ^2 is the maximum of two terms: a near-parametric term, and the geometric mean of the near-parametric term and the oracle risk. To see this more clearly, we express Theorem 1 using asymptotic notation.

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

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

$$\le\min_{\lambda\in\Lambda}\left\|g^*-\hat{g}^{(n_T)}(\boldsymbol{\lambda}|T)\right\|_V^2 \tag{2.6}$$

$$+O_p\left(\frac{J\log(n\|C_\Lambda\|_V\Delta_\Lambda)}{n_V}\right) \tag{2.7}$$

$$+O_p\left(\sqrt{\frac{J\log(n\|C_\Lambda\|_V\Delta_\Lambda)}{n_V}\min_{\lambda\in\Lambda}\left\|g^*-\hat{g}^{(n_T)}(\boldsymbol{\lambda}|T)\right\|_V^2}\right). \tag{2.8}$$

Corollary 1 shows that the risk of the selected model is bounded by the oracle risk, the near-parameteric term (2.7), and the geometric mean of the two values (2.8). We refer to (2.7) as near-parametric because the error term in (unregularized) parametric regression models is typically $O_p(J/n)$, where J is the parameter dimension and n is the number of training samples. Analogously, (2.7)

is $O_p(J/n_V)$ modulo a $\log n$ term in the numerator. The geometric mean (2.8) can be thought of as a consequence of tuning hyperparameters over

$$\mathcal{G}(T) = \left\{ \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) : \boldsymbol{\lambda} \in \Lambda \right\}. \tag{2.9}$$

Because $\mathcal{G}(T)$ does not (or is very unlikely to) contain the true model g^*, tuning the hyperparameters using a training/validation split is analogous to tuning over a misspecified model class. The geometric mean takes into account this misspecification error.

In the semiparametric and nonparametric regression settings, the oracle error usually shrinks at a rate of $O_p(n_T^{-\omega})$, where $\omega \in (0,1)$. If the number of hyperparameters is fixed and n is large, the oracle risk will tend to dominate the upper bound. Hence, for such problems, we can actually let the number of hyperparameters grow, and the asymptotic convergence rate of the upper bound will remain unchanged as long as J grows no faster than $O_p\left((n_V n_T^{-\omega})/(\log(n\|C_\Lambda\|_V \Delta_\Lambda))\right)$.

2.2. Cross-validation

Now, we give an oracle inequality for K-fold cross-validation. Previously, the oracle inequality was with respect to the L_2-norm over the validation covariates. We give our result with respect to the functional L_2-norm. We suppose our data set is composed of independent and identically distributed (i.i.d.) observations (X, y), where X is independent of ϵ. The functional L_2-norm is defined as $\|h\|_{L_2}^2 = \int |h(x)|^2 \, d\mu(x)$.

For K-fold cross-validation, we randomly partition the data set $D^{(n)}$ into K sets, which we assume to be of equal size for simplicity. Partition k is denoted as $D_k^{(n_V)}$ and its complement is denoted as $D_{-k}^{(n_T)} = D^{(n)} \setminus D_k^{(n_V)}$. We train our model using $D_{-k}^{(n_T)}$, for $k = 1, \ldots, K$, and select the hyperparameter that minimizes the average validation loss

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \frac{1}{K} \sum_{k=1}^{K} \left\| y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D_{-k}^{(n_T)}) \right\|_{D_k^{(n_V)}}^2. \tag{2.10}$$

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

$$\bar{g}\left(D^{(n)}\right) = \frac{1}{K} \sum_{k=1}^{K} \hat{g}^{(n_T)}\left(\hat{\boldsymbol{\lambda}} \middle| D_{-k}^{(n_T)}\right). \tag{2.11}$$

To bound the generalization error of (2.11), we require the assumption in Lecué and Mitchell (2012) that controls the tail behavior of the fitted models. A classical approach to bounding the tail behavior of a random variable X is to bound its Orlicz norm $\|X\|_{L_{\psi_1}} = \inf\{C > 0 : \mathbb{E}\exp(|X|/C) - 1 \leq 1\}$ (Vaart and Wellner (1996, Chap. 2)).

Assumption 2. There exist constants $K_0, K_1 \geq 0$ and $\kappa \geq 1$ such that, for any $n_T \in \mathbb{N}$, data set $D^{(n_T)}$, and $\boldsymbol{\lambda} \in \Lambda$, we have

$$\left\| \left(y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) \right)^2 - (y - g^*)^2 \right\|_{L_{\psi_1}} \leq K_0, \tag{2.12}$$

$$\left\| \left(y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) \right)^2 - (y - g^*)^2 \right\|_{L_2} \leq K_1 \left\| g^* - \hat{g}(\boldsymbol{\lambda}|D^{(n_T)}) \right\|_{L_2}^{1/\kappa}. \tag{2.13}$$

Given the above assumption, the following oracle inequality bounds the risk of the averaged version of K-fold cross-validation. Note that this is a special case of Theorem 4 in the Supplementary Material, which extends Theorem 3.5 in Lecué and Mitchell (2012). The notation $\mathbb{E}_{D^{(m)}}$ indicates the expectation over random m-sample data sets $D^{(m)}$ drawn from the probability distribution μ.

Theorem 2. *Let $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$, where $\Delta_\Lambda = (\lambda_{\max} - \lambda_{\min}) \vee 1$. Suppose random variables ϵ_i are independent with expectation zero, satisfy $\|\epsilon\|_{L_{\psi_2}} = b < \infty$, and are independent of X. Suppose Assumption 1 holds over the set $\mathcal{X}$ and Assumption 2 holds. Suppose there exists a function $\tilde{h}$ and some $\sigma_0 > 0$ such that*

$$\tilde{h}(n_T) \geq 1 + \sum_{k=1}^{\infty} k \Pr\left(\|C_\Lambda(\cdot|D^{(n_T)})\|_{L_{\psi_2}} \geq 2^k \sigma_0 \right). \tag{2.14}$$

Then, there exists an absolute constant $c_1 > 0$ and a constant $c_{K_0,b} > 0$ such that for any $a > 0$,

$$\begin{aligned} &\mathbb{E}_{D^{(n)}}\left(\|\bar{g}(D^{(n)}) - g^*\|_{L_2}^2 \right) \\ &\leq (1+a) \inf_{\lambda \in \Lambda} \left[\mathbb{E}_{D^{(n_T)}}\left(\|\hat{g}(\boldsymbol{\lambda}|D^{(n_T)}) - g^*\|_{L_2}^2 \right) \right] \\ &\quad + c_1 \left(\frac{1+a}{a} \right)^2 \frac{J \log n_V}{n_V} K_0 \left[\log\left(\Delta_\Lambda c_{K_0,b} n \sigma_0 + 1 \right) + 1 \right] \tilde{h}(n_T). \end{aligned} \tag{2.15}$$

As in Theorem 1, the remainder term in Theorem 2 includes a near-parametric term $O_p(J/n_V)$. Thus, as before, adding hyperparameters to parametric model estimation incurs a similar cost to that of adding parameters to the parametric model itself. Adding hyperparameters to semiparametric and nonparametric regression settings is relatively "cheap" and negligible, asymptotically.

The differences between Theorems 1 and 2 highlight the trade-offs made to establish an oracle inequality involving the functional L_2-error. The biggest tradeoff is that Theorem 2 adds Assumption 2. Although we can relax Assumption 2 to hold over data sets D in some high-probability set, the difficulty lies in controlling the tail behavior of the fitted models over all Λ. For some model estimation procedures, K_0 may grow with n if $\lambda_{\min}$ shrinks too quickly with n. In this case, the remainder term may no longer shrink at a near-parametric rate. Unfortunately, requiring $\lambda_{\min}$ to shrink at an appropriate rate seems to defeat the purpose of cross-validation. Therefore, even though Theorem 2 helps us to better understand cross-validation, it is limited by this assumption. In addition, the Lipschitz assumption must hold over all $\mathcal{X}$ in Theorem 2, rather than just the observed covariates. Finally, the oracle inequality in Theorem 2 is no longer sharp, because the oracle risk is scaled by $1 + a$, for $a > 0$.

3. Penalized Regression Models

Here, we apply our results to analyze penalized regression procedures of the form given in (1.1). Penalty functions encourage particular characteristics in the fitted models (e.g., smoothness or sparsity). Furthermore, combining multiple penalty functions results in models that exhibit a combination of the desired characteristics. Although the latter practice has garnered much interest, few methods incorporate more than two penalties owing to (a) a concern that models may overfit the data when many penalty parameters needs to be tuned; and (b) computational issues in optimizing multiple penalty parameters. In this section, we evaluate the validity of concern (a) using the results of Section 2. We see that, contrary to popular wisdom, using split-sample validation to select multiple penalty parameters should not result in a drastic increase in the generalization error of the selected model.

In this section, we consider penalty parameter spaces of the form $\Lambda = [n^{-t_{\min}}, n^{t_{\max}}]^J$, for $t_{\min}, t_{\max} \geq 0$. This regime works well for two reasons: (a) our rates depend only quite weakly on $t_{\min}$ and $t_{\max}$; and (b) oracle λ-values are generally $O_p(n^{-\alpha})$ for some $\alpha \in (0, 1)$ (van de Geer (2000); van de Geer and Muro (2015); Bühlmann and van de Geer (2011)). As long as $t_{\min} > \alpha$, Λ will contain the optimal penalty parameter. We do not consider settings where $\lambda_{\min}$ shrinks faster than a polynomial rate because the fitted models in this case can be ill-behaved.

In the following sections, we conduct an in-depth study of additive models

of the form

$$g(\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(J)}) = \sum_{j=1}^{J} g_j(\boldsymbol{x}^{(j)}). \tag{3.1}$$

We first consider parametric additive models (with potentially growing numbers of parameters) fitted with smooth and nonsmooth penalties, followed by nonparametric additive models. We find that the Lipschitz function $C_\Lambda(\boldsymbol{x}|T)$ scales with $n^{O_p(t_{\min})}$. Applying Theorems 1 and 2, we find that the near-parametric term in the remainder grows only linearly in $t_{\min}$. We apply these results to various additive model estimation methods. For instance, in the generalized additive model (GAM) example, we show that under minimal assumptions, the error incurred from tuning penalty parameters is negligible with that from solving the penalized regression problem with oracle penalty parameters.

3.1. Parametric additive models

Parametric additive models with model parameters $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, \ldots, \boldsymbol{\theta}^{(J)}\right)$ have the form

$$g(\boldsymbol{\theta})(\boldsymbol{x}) = \sum_{j=1}^{J} g_j(\boldsymbol{\theta}^{(j)})(\boldsymbol{x}^{(j)}). \tag{3.2}$$

We denote the training criterion for training data T as

$$L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) \coloneqq \frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta}^{(j)}). \tag{3.3}$$

Suppose $\boldsymbol{\theta}^*$ is the unique minimizer of the expected loss $\|y - g(\boldsymbol{\theta})\|_{L_2}^2$.

3.1.1. Parametric regression with smooth penalties

We begin with the simple case in which the penalty functions are smooth. The following lemma states that the fitted models are Lipschitz in the penalty parameter vector. Given matrices A and B, $A \succeq B$ means that $A - B$ is a positive semi-definite matrix.

Lemma 1. *Let $\Lambda \coloneqq [\lambda_{\min}, \lambda_{\max}]^J$, where $\lambda_{\max} \geq \lambda_{\min} > 0$. For a fixed training data set $T \equiv D^{(n_T)}$, suppose for all $\boldsymbol{\lambda} \in \Lambda$, $L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ has a unique minimizer*

$$\left\{\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}|T)\right\}_{j=1}^{J} = \operatorname*{arg\,min}_{\boldsymbol{\theta} \in \mathbb{R}^p} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}). \tag{3.4}$$

Suppose for all $j = 1, \ldots, J$, the parametric class g_j is ℓ_j-Lipschitz in its parameters

$$\left|g_j(\boldsymbol{\theta}^{(1)})(\boldsymbol{x}^{(j)}) - g_j(\boldsymbol{\theta}^{(2)})(\boldsymbol{x}^{(j)})\right| \leq \ell_j(\boldsymbol{x}^{(j)})\|\boldsymbol{\theta}^{(1)} - \boldsymbol{\theta}^{(2)}\|_2 \quad \forall \boldsymbol{x}^{(j)} \in \mathcal{X}^{(j)}. \tag{3.5}$$

Further, suppose for all $j = 1, \ldots, J$, $P_j(\boldsymbol{\theta}^{(j)})$ and $g_j(\boldsymbol{\theta}^{(j)})(\boldsymbol{x})$ are twice-differentiable with respect to $\boldsymbol{\theta}^{(j)}$ for any fixed $\boldsymbol{x}$. Suppose there exists an $m(T) > 0$ such that the Hessian of the penalized training criterion at the minimizer satisfies

$$\nabla^2_\theta L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda}|T)} \succeq m(T)\boldsymbol{I} \quad \forall \boldsymbol{\lambda} \in \Lambda, \tag{3.6}$$

where I is a $p \times p$ identity matrix. Then, for any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda$, Assumption 1 is satisfied over the set $\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(J)}$ with function

$$C_\Lambda(\boldsymbol{x}|T) = \frac{1}{m(T)\lambda_{\min}} \sqrt{\left(\|\epsilon\|_T^2 + 2C^*_\Lambda\right)\left(\sum_{j=1}^J \|\ell_j\|_T^2 \ell_j^2(\boldsymbol{x}^{(j)})\right)}, \tag{3.7}$$

where $C^*_\Lambda = \lambda_{\max} \sum_{j=1}^J P_j(\boldsymbol{\theta}^{(j),*})$.

Note that Lemma 1 requires that the training criterion be strongly convex at its minimizer. This is satisfied in the following example involving multiple ridge penalties. If (3.6) is not satisfied by a penalized regression problem, we can consider a variant of the problem, in which the penalty functions $P_j(\boldsymbol{\theta}^{(j)})$ are replaced with penalty functions $P_j(\boldsymbol{\theta}^{(j)}) + (w/2)\|\boldsymbol{\theta}^{(j)}\|_2^2$ for a fixed $w > 0$.

Example 1 (Multiple ridge penalties). Let us consider fitting a linear model using ridge regression. If we can group covariates based on the similarity of their effects on the response, i.e. $\boldsymbol{x} = (\boldsymbol{x}^{(1)}, \ldots, \boldsymbol{x}^{(J)})$ where $\boldsymbol{x}^{(j)}$ is a vector of length p_j, we can incorporate this prior information by penalizing each group of covariates differently:

$$L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) \coloneqq \frac{1}{2}\left\|y - \sum_{j=1}^J \boldsymbol{x}^{(j)}\boldsymbol{\theta}^{(j)}\right\|_T^2 + \sum_{j=1}^J \frac{\lambda_j}{2}\|\boldsymbol{\theta}^{(j)}\|_2^2. \tag{3.8}$$

We tune the penalty parameters $\boldsymbol{\lambda}$ over the set Λ using a training/validation split with training and validation sets T and V, respectively. For all examples in this manuscript, let $\Lambda = \left[n^{-t_{\min}}, 1\right]^J$.

Via some algebra, we can derive (3.7) in Lemma 1; the details are deferred to the Supplementary Material. Substituting this result into Corollary 1, we find that the parametric term (2.7) in the remainder is on the order of

$$\frac{Jt_{\min}}{n_V} \log\left(C^*_T n \sum_{j=1}^J \left(\frac{1}{n_T}\sum_{(x_i,y_i)\in T} \|\boldsymbol{x}_i^{(j)}\|_2^2\right)\left(\frac{1}{n_V}\sum_{(x_i,y_i)\in V} \|\boldsymbol{x}_i^{(j)}\|_2^2\right)\right), \tag{3.9}$$

where $C^*_T = \|\epsilon\|_T^2 + \sum_{j=1}^J \|\boldsymbol{\theta}^{*,(j)}\|_2^2$. Thus, we have shown that if the lower bound

of Λ shrinks at the polynomial rate $n^{-t_{\min}}$, then the near-parametric term in the remainder of the oracle inequality grows only linearly in its power $t_{\min}$.

In the next example, we consider GAMs (Hastie and Tibshirani (1990)). Although GAMs are nonparametric models, it is well-known that they are equivalent to solving a finite-dimensional problem (Green and Silverman (1993), O'sullivan, Yandell and Raynor Jr (1986), Buja, Hastie and Tibshirani (1989)). By reformulating GAMs as parametric models, we can establish oracle inequalities for tuning the penalty parameters using a training/validation split. Here, we present an outline of the procedure; the details can be found in the Supplementary Material.

Example 2 (Multiple Sobolev penalties). To fit a GAM over the domain $\mathcal{X}^J$, where $\mathcal{X} \subseteq \mathbb{R}$, a typical setup is to solve

$$\underset{\alpha_0 \in \mathbb{R}, g_j}{\arg\min} \frac{1}{2} \sum_{i \in D^{(n_T)}} \left(y_i - \alpha_0 - \sum_{j=1}^{J} g_j(x_{ij}) \right)^2 + \sum_{j=1}^{J} \lambda_j \int_{\mathcal{X}} \left(g_j''(x_j) \right)^2 dx_j, \quad (3.10)$$

where the penalty function is the 2nd-order Sobolev norm. Let $\mathcal{X} = [0, 1]$ for this example. Using properties of the Sobolev penalty, (3.10) can be re-expressed as a finite-dimensional problem with matrices K_j,

$$\underset{\alpha_0, \boldsymbol{\alpha}_1, \boldsymbol{\theta}}{\arg\min} \frac{1}{2} \left\| y - \alpha_0 \mathbf{1} - \boldsymbol{x}\boldsymbol{\alpha}_1 - \sum_{j=1}^{j} K_j \boldsymbol{\theta}^{(j)} \right\|_T^2 + \frac{1}{2} \sum_{j=1}^{J} \lambda_j \boldsymbol{\theta}^{(j)\top} K_j \boldsymbol{\theta}^{(j)}. \quad (3.11)$$

Let $X_T \in \mathbb{R}^{n_T \times J}$ be the covariates $\boldsymbol{x}$ in the training data stacked together. If $X_T^\top X_T$ is invertible, we can derive the closed-form solution for (3.11). Then, we can directly calculate (3.7) in Lemma 1. Substituting this result into Corollary 1, we find that the parametric term in the remainder is on the order of

$$\frac{J t_{\min}}{n_V} \log \left(nJ \|y\|_T \left(J \left\| \left(X_T^\top X_T \right)^{-1} X_T^\top \right\|_2 + \sum_{j=1}^{J} h_j^{-2}(T) \right) \right), \quad (3.12)$$

where $\| \cdot \|_2$ is the spectral norm and $h_j(T)$ is the smallest distance between observations of the jth covariates in the training data T.

In particular, for $J = o(n^{1/2})$, the smoothing spline estimate (3.10) is shown to attain the minimax optimal rate of $O_p(Jn^{-4/5})$ if the penalty parameters shrink at the rate of $\sim n^{-4/5}$ (Sadhanala and Tibshirani (2017); Horowitz, Klemelä and Mammen (2006)). From Corollary 1, we see that the oracle error (2.6) asymptotically dominates the additional error terms incurred from tuning the penalty parameters. Moreover, as long as we choose $\lambda_{\min} \sim n^{-\alpha}$ for any

$\alpha > 4/5$, the model selected via training/validation split will also attain the minimax rate.

3.1.2. Parametric regression with nonsmooth penalties

If the penalty functions are nonsmooth, similar results do not necessarily hold. Nonetheless, we find that for many popular nonsmooth penalty functions, such as the lasso (Tibshirani (1996)) and group lasso (Yuan and Lin (2006)), the fitted functions are still smoothly parameterized by $\boldsymbol{\lambda}$ almost everywhere. To characterize such problems, we begin with the following definitions from Feng and Simon (2018):

Definition 1. *The differentiable space of function $f : \mathbb{R}^p \mapsto \mathbb{R}$ at $\boldsymbol{\theta}$ is*

$$\Omega^f(\boldsymbol{\theta}) = \left\{ \boldsymbol{\beta} \middle| \lim_{\epsilon \to 0} \frac{f(\boldsymbol{\theta} + \epsilon\boldsymbol{\beta}) - f(\boldsymbol{\theta})}{\epsilon} \text{ exists} \right\}. \tag{3.13}$$

Definition 2. *Let $f(\cdot, \cdot) : \mathbb{R}^p \times \mathbb{R}^J \mapsto \mathbb{R}$ be a function with a unique minimizer. $S \subseteq \mathbb{R}^p$ is a local optimality space of f over $W \subseteq \mathbb{R}^J$ if*

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in S} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda} \in W. \tag{3.14}$$

Using the above definitions, we can characterize the penalty parameters $\Lambda_{smooth} \subseteq \Lambda$, where the fitted functions are well-behaved.

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 $L_T(\cdot, \boldsymbol{\lambda}')$ is twice differentiable with respect to $\boldsymbol{\theta}$ at $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}'|T)$ along directions in the product space
$$\Omega^{L_T(\cdot,\boldsymbol{\lambda})}\left(\hat{\boldsymbol{\theta}}\left(\boldsymbol{\lambda}|T\right)\right) = \Omega^{P_1(\cdot)}\left(\hat{\boldsymbol{\theta}}^{(1)}(\boldsymbol{\lambda}|T)\right) \times \cdots \times \Omega^{P_J(\cdot)}\left(\hat{\boldsymbol{\theta}}^{(J)}(\boldsymbol{\lambda}|T)\right). \tag{3.15}$$
- $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}\left(\hat{\boldsymbol{\theta}}\left(\boldsymbol{\lambda}|T\right)\right)$ is a local optimality space for $L_T\left(\cdot, \boldsymbol{\lambda}\right)$ over $B(\boldsymbol{\lambda})$.

In addition, we need nearly all penalty parameters to be in Λ_{smooth}.

Condition 2. $\Lambda \setminus \Lambda_{smooth}$ has Lebesgue measure zero, i.e. $\mu(\Lambda^c_{smooth}) = 0$.

For instance, in the lasso, Λ_{smooth} is the sections of the lasso path between the knots. As the knots in the lasso-path are countable, the set outside Λ_{smooth} has measure zero.

Assuming the above conditions hold, the fitted models for nonsmooth penalty functions satisfy the same Lipschitz relation as that in Lemma 1.

Lemma 2. *Let $\Lambda \coloneqq [\lambda_{\min}, \lambda_{\max}]^J$, where $\lambda_{\max} \geq \lambda_{\min} > 0$. Suppose that for all $j = 1, \ldots, J$, g_j satisfies (3.5) over $\mathcal{X}^{(j)}$. Suppose for training data $T \equiv D^{(n_T)}$,*

the penalized loss function $L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ *has a unique minimizer* $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)$ *for every* $\boldsymbol{\lambda} \in \Lambda$. *Let* $\boldsymbol{U}_\lambda$ *be an orthonormal matrix with columns forming a basis for the differentiable space of* $L_T(\cdot, \boldsymbol{\lambda})$ *at* $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)$. *Suppose there exists a constant* $m(T) > 0$ *such that the Hessian of the penalized training criterion at the minimizer taken with respect to the directions in* $\boldsymbol{U}_\lambda$ *satisfies*

$$U_\lambda \nabla^2_\theta L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq m(T)\boldsymbol{I} \quad \forall \boldsymbol{\lambda} \in \Lambda, \tag{3.16}$$

where $\boldsymbol{I}$ is the identity matrix. *Suppose Conditions* 1 *and* 2 *are satisfied. Then, any* $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda$ *satisfies Assumption* 1 *over* $\mathcal{X}^{(1)} \times \cdots \times \mathcal{X}^{(J)}$, *with* C_Λ *defined in* (3.7).

As an example, we consider multiple elastic net penalties, where the penalty parameters are tuned using training/validation split and cross-validation.

Example 3 (Multiple elastic nets, training/validation split). Suppose we want to fit a linear model using the elastic net. If the covariates are grouped a priori, we can penalize each group differently using the following objective:

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \underset{\theta^{(j)} \in \mathbb{R}^{p_j}, j=1,\ldots,J}{\arg\min} \frac{1}{2} \left\| y - \sum_{j=1}^{J} \boldsymbol{X}^{(j)} \boldsymbol{\theta}^{(j)} \right\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(\|\boldsymbol{\theta}^{(j)}\|_1 + \frac{w}{2} \|\boldsymbol{\theta}^{(j)}\|_2^2 \right), \tag{3.17}$$

where $w > 0$ is a fixed constant. Here, we briefly sketch the process for deriving the oracle inequality when tuning the penalty parameters using a training/validation split over $\Lambda = [n^{-t_{\min}}, 1]^J$. Details are given in the Supplementary Material.

First, we check that all the conditions are satisfied. For this problem, the differentiable space is the subspace spanned by the nonzero elements in $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$. Because the elastic net solution paths are piecewise linear (Zou and Hastie (2003)), the differentiable space is also a local optimality space. Then, using a similar procedure to that described in Example 1, we find that the parametric term in the remainder of Corollary 1 is on the order of

$$\frac{J t_{\min}}{n_V} \log \left(\frac{C_T^* n}{w} \sum_{j=1}^{J} \left(\frac{1}{n_T} \sum_{(x_i, y_i) \in T} \|\boldsymbol{x}_i^{(j)}\|_2^2 \right) \left(\frac{1}{n_V} \sum_{(x_i, y_i) \in V} \|\boldsymbol{x}_i^{(j)}\|_2^2 \right) \right), \tag{3.18}$$

where $C_T^* = \|\epsilon\|_T^2 + \sum_{j=1}^{J} 2\|\boldsymbol{\theta}^{*,(j)}\|_1 + w\|\boldsymbol{\theta}^{*,(j)}\|_2^2$.

We can compare this additional error term to the risk of using an oracle penalty parameter. For the case of a single penalty parameter ($J = 1$), the convergence rate of using an oracle penalty parameter for the elastic net is on the order of $O_p(\log(p)/n)$ (Bunea (2008); Hebiri and van de Geer (2011)). If

we split the covariates into groups and tune the penalty parameters using a training/validation split, the incurred error (3.18) is on a similar order.

Example 4 (Multiple elastic nets, cross-validation). Now, we establish an oracle inequality for the averaged version of K-fold cross-validation using a similar setup to that in Lecué and Mitchell (2012). Suppose the noise ϵ is sub-Gaussian. For simplicity, suppose X is drawn uniformly from $[-1,1]^p$. In order to satisfy the assumptions in Theorem 2, our fitting procedure for $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ entails a thresholding operation similar to that in Lecué and Mitchell (2012). In particular, let the fitted parameters be denoted $\hat{\boldsymbol{\theta}}_{thres}(\boldsymbol{\lambda})$, where the ith element is

$$\hat{\theta}_{thres,i}(\boldsymbol{\lambda}) = \operatorname{sign}(\hat{\theta}_i(\boldsymbol{\lambda}))(|\hat{\theta}_i(\boldsymbol{\lambda})| \wedge K_0') \quad i = 1, \ldots, p, \tag{3.19}$$

where $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ is the solution to (3.17) and $K_0' > 0$ is some fixed constant. We then find the Lipschitz factor in Lemma 3 and upper-bound its Orlicz norm using exponential concentration inequalities. Let $\bar{\boldsymbol{\theta}}(D^{(n)})$ be the fitted parameters using the averaged version of K-fold cross-validation. By Theorem 2, there is some constant $\tilde{c} > 0$, such that for any $a > 0$,

$$\begin{aligned}\mathbb{P}_{D^{(n)}} \left\| X\left(\bar{\boldsymbol{\theta}}(D^{(n)}) - \boldsymbol{\theta}^*\right)\right\|_{L_2}^2 \leq (1+a) \inf_{\lambda \in \Lambda} \left[\mathbb{P}_{D^{(n_T)}} \left\| X\left(\bar{\boldsymbol{\theta}}(D^{(n_T)}) - \boldsymbol{\theta}^*\right)\right\|_{L_2}^2\right] \\ + \tilde{c}\left(\frac{1+a}{a}\right)^2 \frac{J \log n_V}{n_V} t_{\min} \log\left(\frac{1+a}{aw} Jpn\right).\end{aligned} \tag{3.20}$$

The above example is similar to the lasso example in Lecué and Mitchell (2012); the major difference is that we consider the case where the penalty parameters are tuned over a continuous range. We are able to do this because Lemma 2 specifies a Lipschitz relation between the fitted functions and the penalty parameters. This result is relevant when J is large and $\boldsymbol{\lambda}$ must be tuned using a continuous optimization procedure.

3.2. Nonparametric additive models

We now consider nonparametric additive models of the form

$$\{\hat{g}_j(\boldsymbol{\lambda})\}_{j=1}^J = \underset{g_j \in \mathcal{G}_j : j=1,\ldots,J}{\arg\min} L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right) \coloneqq \frac{1}{2}\left\| y - \sum_{j=1}^J g_j(x_j)\right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j), \tag{3.21}$$

where $\{P_j\}$ are penalty functionals, and $\{\mathcal{G}_j\}$ are linear spaces of univariate functions. Let $\{g_j^*\}_{j=1}^J$ be the minimizer of the generalization error

$$\left\{g_j^*\right\}_{j=1}^J = \underset{g_j \in \mathcal{G}_j : j=1,\ldots,J}{\arg\min} E \left\| y - \sum_{j=1}^J g_j^* \right\|_{L_2}^2 . \tag{3.22}$$

We obtain a similar Lipschitz relation in the nonparametric setting to those of previous settings.

Lemma 3. *Let $\lambda_{\max} > \lambda_{\min} > 0$ and $\Lambda := [\lambda_{\min}, \lambda_{\max}]^J$. Suppose the penalty functions P_j are twice Gateaux differentiable and convex over $\mathcal{G}_j$. Suppose there is an $m(T) > 0$ such that the second Gateaux derivative of the training criterion at $\{\hat{g}_j^{(n_T)}(\boldsymbol{\lambda}|T)\}$, for all $\boldsymbol{\lambda} \in \Lambda$, satisfies*

$$\left\langle D^2_{\{g_j\}} L_T \left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda} \right) \Big|_{g_j = \hat{g}_j(\boldsymbol{\lambda}|T)} \circ h_j, h_j \right\rangle \geq m(T) \quad \forall h_j \in \mathcal{G}_j, \|h_j\|_{D^{(n)}} = 1, \tag{3.23}$$

where $D^2_{\{g_j\}}$ is the second Gateaux derivative taken in directions $\{g_j\}$. Let $C^*_\Lambda = \lambda_{\max} \sum_{j=1}^J P_j(g_j^*)$. For any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda$, we have

$$\left\| \sum_{j=1}^J \hat{g}_j \left(\boldsymbol{\lambda}^{(1)} | T \right) - \hat{g}_j \left(\boldsymbol{\lambda}^{(2)} | T \right) \right\|_{D^{(n)}} \leq \frac{m(T)}{\lambda_{\min}} \sqrt{\left(\|\epsilon\|_T^2 + 2C^*_\Lambda \right) \frac{n_D}{n_T}} \left\| \boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)} \right\|_2 . \tag{3.24}$$

A simple example that satisfies (3.23) is a penalized regression model in which we fit values at each of the observed covariates and penalize this fitted value using a ridge penalty. Note that this procedure is allowed because the response y in the validation set is not used by the training procedure.

Note that because Lemma 3 verifies that Assumption 1 is satisfied over the observed covariates, it is suitable to be used in Theorem 1. However (3.24) is not a sufficiently strong statement to be used for Theorem 2.

4. Simulations

We now present a simulation study of the generalized additive model in Example 2 to show how the performance changes as the number of penalty parameters J increases. Corollary 1 suggests that there are two opposing forces that affect the error of the fitted model. On the one hand, (2.7) is linear in J; thus, increasing J can increase the error. On the other hand, (2.6) decreases for larger model spacesl; thus, increasing J may decrease the error. We isolate these

two behaviors using two simulation setups.

The data are generated as the sum of univariate functions. That is, let response $Y = \sum_{j=1}^{J} g_j^*(X_j) + \sigma\epsilon$, where ϵ are i.i.d. standard Gaussian random variables, and $\sigma > 0$ is chosen such that the signal-to-noise ratio is two. Here, X is drawn from a uniform distribution over $\mathcal{X} = [-2, 2]^J$. We fit models by minimizing (3.10). To vary the number of free penalty parameters, we constrain certain λ_j to be equal, while allowing others to be vary freely. (For instance, for a single penalty parameter, we constrain λ_j for $j = 1, \ldots, J$ to be the same value.) The penalty parameters are tuned using a training/validation split.

Simulation 1. The true function is the sum of identical sinusoids $g_j^*(x_j) = \sin(x_j)$ for $j = 1, \ldots, J$. Because the univariate functions are the same, the oracle risk should be roughly constant as we increase the number of free penalty parameters. The validation loss difference

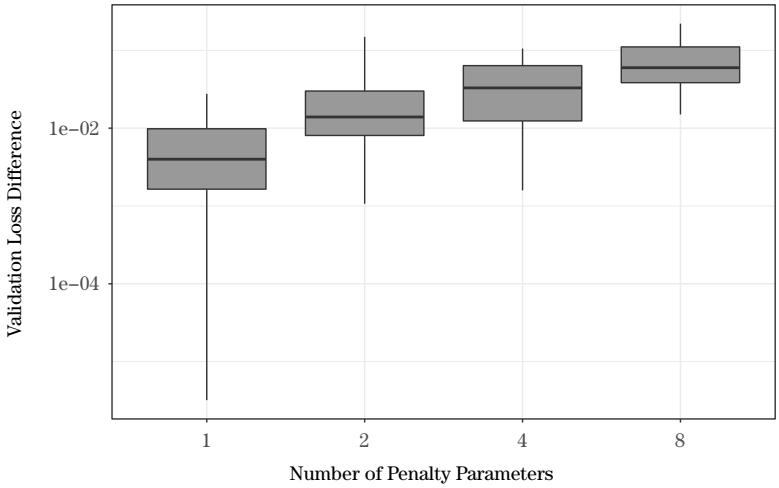
$$\left\| \sum_{j=1}^{J} \hat{g}_j^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) - g_j^* \right\|_V^2 - \min_{\lambda \in \Lambda} \left\| \sum_{j=1}^{J} \hat{g}_j^{(n_T)}(\boldsymbol{\lambda}|T) - g_j^* \right\|_V^2 \tag{4.1}$$

should grow linearly in J for this simulation setup.

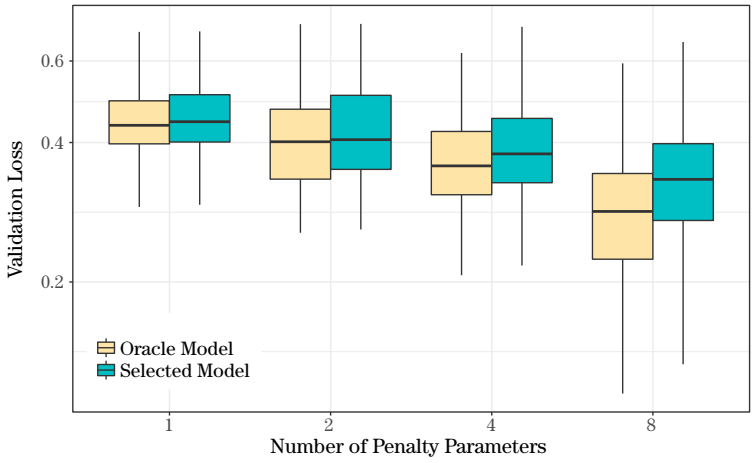
Simulation 2. The true function is the sum of sinusoids with increasing frequency $g_j^*(x_j) = \sin(x_j * 1.2^{j-4})$ for $j = 1, \ldots, J$. Because the Sobolev norms of g_j^* increase with j, we expect the penalty parameters that attain the oracle risk to be monotonically decreasing, i.e. $\lambda_1 > \cdots > \lambda_J$. As the number of penalty parameters increases, we expect the oracle risk to shrink. If the oracle risk shrinks sufficiently quickly, the performance of the selected model should improve.

For both simulations, we use $J = 8$. Each simulation was replicated 40 times with 200 training and 200 validation samples. We consider $k = 1, 2, 4, 8$ free penalty parameters by structuring the penalty parameters in a nested fashion: for each k, we constrained $\{\lambda_{8\ell/k+j}\}_{j=1,\ldots,8/k}$ to be equal for $\ell = 0, \ldots, k-1$. The penalty parameters were tuned using `nlm` in `R`, with initializations at $\{\vec{1}, 0.1 \times \vec{1}, 0.01 \times \vec{1}\}$. We did not use a grid-search, because it is computationally intractable for large numbers of penalty parameters. Multiple initializations were required, because the validation loss is not convex in the penalty parameters.

As expected, the validation loss difference increases with the number of penalty parameters in Simulation 1 (Figure 1(a)). To determine whether our oracle inequalities match the empirical results, we regressed the logarithm of the validation loss difference against the logarithm of the number of penalty parameters. We fit the model for the simulation results with at least two penalty



(a) Simulation 1: the univariate additive components are the same



(b) Simulation 2: the univariate additive components have differing levels of smoothness

Figure 1. Performance of generalized additive models as the number of free penalty parameters grows.

parameters, because the data are highly skewed for the single penalty parameter case. We estimated a slope of 1.00 (standard error 0.15), which suggests that the validation loss difference grows linearly in the number of penalty parameters. Interestingly, including the single parameter case gives us a slope of 1.45 (standard error 0.14). This suggests that our oracle inequality might not be tight for the single penalty parameter case.

For Simulation 2, the validation loss of the selected model decreases as the number of penalty parameters increases. As suggested in Figure 1(b), the validation loss of the selected model decreases because the oracle risk is decreasing at a faster rate than the rate at which the additional error (2.7) grows.

These simulation results suggest that adding more hyperparameters can improve model estimates. Having a separate penalty parameter allows GAMs to fit components with differing smoothness. However, if we know a priori that the components have the same smoothness, then it is best to use a single penalty parameter.

5. Discussion

We have characterized the generalization error of split-sample procedures that tune multiple hyperparameters. If the estimated models are Lipschitz in the hyperparameters, the generalization error of the selected model is upper bounded by a combination of the oracle risk and a near-parametric term in the number of hyperparameters. These results show that adding hyperparameters can decrease the generalization error of the selected model if the oracle risk decreases by a sufficient amount. In the semiparametric or nonparametric setting, the error incurred from tuning hyperparameters is dominated by the oracle risk asymptotically; adding hyperparameters has a negligible effect on the generalization error of the selected model. In the parametric setting, the error incurred from tuning the hyperparameters is on the same order as the oracle error. Nonetheless, one should still be careful when adding hyperparameters, even though they are not more "costly" than model parameters.

We also showed that many penalized regression examples satisfy the Lipschitz condition, which means our theoretical results apply. This implies that fitting models with multiple penalties and penalty parameters can be desirable, rather than the usual case with one or two penalty parameters.

One drawback of our theoretical results is that we have assumed that the selected hyperparameter is a global minimizer of the validation loss. Unfortunately, this is not achievable in practice since the validation loss is not convex with respect to the hyperparameters. This problem is exacerbated when there are many hyperparameters because it is computationally infeasible to perform an exhaustive grid-search. We hope to address this question in future research.

Acknowledgement

Jean Feng and Noah Simon were supported by NIH Early Independence Award grant 5DP5OD019820.

References

Arlot, S. and Celisse, A. (2010). A survey of cross-validation procedures for model selection. *Stat. Surv.* **4**, 40–79.

Bengio, Y. (2000). Gradient-based optimization of hyperparameters. *Neural Computation* **12**, 1889–1900.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data: Methods, Theory and Applications*, 1st Edition. Springer Series in Statistics. Springer-Verlag Berlin Heidelberg.

Buja, A., Hastie, T. and Tibshirani, R. (1989). Linear smoothers and additive models. *The Annals of Statistics* **17**, 453–510.

Bunea, F. (2008). Honest variable selection in linear and logistic regression models via l1 and l1+l2 penalization. *Electron. J. Stat.* **2**, 1153–1194.

Chatterjee, S. and Jafarov, J. (2015). Prediction error of cross-validated lasso. *arXiv preprint arXiv:1502.06291*.

Chetverikov, D., Liao, Z. and Chernozhukov, V. (2016). On cross-validated lasso. *arXiv preprint arXiv:1605.02214*.

Feng, J. and Simon, N. (2018). Gradient-based regularization parameter selection for problems with nonsmooth penalty functions. *J. Comput. Graph. Stat.* **27**, 426–435.

Foo, C.-S., Do, C. B. and Ng, A. Y. (2008). Efficient multiple hyperparameter learning for log-linear models. In *Advances in Neural Information Processing Systems* **20** (Edited by J. C. Platt, D. Koller, Y. Singer and S. T. Roweis), 377-384. Curran Associates, Inc.

Golub, G. H., Heath, M. and Wahba, G. (1979). Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, 215–223.

Green, P. J. and Silverman, B. W. (1993). *Nonparametric Regression and Generalized Linear Models: A Roughness Penalty Approach*. CRC Press.

Györfi, L., Kohler, M., Krzyzak, A. and Walk, H. (2002). *A Distribution-Free Theory of Nonparametric Regression*, 1st Edition. Springer Series in Statistics. Springer-Verlag New York.

Hastie, T. and Tibshirani, R. (1990). *Generalized Additive Models*, 2nd Edition. Chapman and Hall/CRC.

Hebiri, M. and van de Geer, S. (2011). The Smooth-Lasso and other l1+l2-penalized methods. *Electron. J. Stat.* **5**, 1184–1226.

Horowitz, J., Klemelä, J. and Mammen, E. (2006). Optimal estimation in additive regression models. *Bernoulli* **12**, 271–298.

Lecué, G. and Mitchell, C. (2012). Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics* **6**, 1803–1837.

Li, K.-C. (1987). Asymptotic optimality for C_p, C_L, Cross-Validation and generalized Cross-Validation: Discrete index set. *The Annals of Statistics* **15**, 958–975.

O'sullivan, F., Yandell, B. S. and Raynor Jr, W. J. (1986). Automatic smoothing of regression functions in generalized linear models. *Journal of the American Statistical Association* **81**,

96–103.

Sadhanala, V. and Tibshirani, R. J. (2017). Additive models with trend filtering. *arXiv preprint arXiv:1702.05037*.

Shao, J. (1997). An asymptotic theory for linear model selection. *Statistica Sinica* **7**, 221–242.

Simon, N., Friedman, J., Hastie, T. and Tibshirani, R. (2013). A sparse-group lasso. *Journal of Computational and Graphical Statistics* **22**, 231–245.

Snoek, J., Larochelle, H. and Adams, R. P. (2012). Practical Bayesian optimization of machine learning algorithms. In *Advances in Neural Information Processing Systems* **25** (Edited by F. Pereira, C. J. C. Burges, L. Bottou and K. Q. Weinberger), 2951-2959. Curran Associates, Inc.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **58**, 267–288.

Vaart, A. W. v. d. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes: With Applications to Statistics*. Springer Series in Statistics.

van de Geer, S. (2000). *Empirical Processes in M-Estimation*. Cambridge University Press.

van de Geer, S. and Muro, A. (2015). Penalized least squares estimation in the additive model with different smoothness for the components. *J. Statist. Plann. Inference* **162**, 43–61.

van der Laan, M. J. and 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. *U. C. Berkeley Division of Biostatistics Working Paper Series*.

van der Laan, M. J., Dudoit, S. and Keles, S. (2004). Asymptotic optimality of likelihood-based cross-validation. *Stat. Appl. Genet. Mol. Biol.* **3**, 1–23.

Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **68**, 49–67.

Zou, H. and Hastie, T. (2003). Regression shrinkage and selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology).* **67**, 301–320.

Department of Biostatistics, University of Washington, Box 357232, University of Washington, Seattle, WA 98195-7232, USA.

E-mail: jeanfeng@uw.edu

Department of Biostatistics, University of Washington, Health Sciences Building F-650, Box 357232, University of Washington, Seattle, WA 98195, USA.

E-mail: nrsimon@uw.edu

(Received July 2017; accepted May 2018)