An analysis of the cost of hyper-parameter selection via split-sample validation, with applications to penalized regression

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

July 1, 2017

Abstract

In the regression setting, given a set of hyper-parameters, a model-estimation procedure constructs a model from training data. The optimal hyper-parameters that minimize generalization error of the model are usually unknown. In practice they are often estimated using split-sample validation. Up to now, there is an open question regarding how the generalization error of the selected model grows with the number of hyper-parameters to be estimated. To answer this question, we establish finite-sample oracle inequalities for selection based on a single training/test split and based on cross-validation. We show that if the model-estimation procedures are smoothly parameterized by the hyper-parameters, the error incurred from tuning hyper-parameters shrinks at nearly a parametric rate. Hence for semi- and non-parametric model-estimation procedures with a fixed number of hyper-parameters, this additional error is negligible. For parametric model-estimation procedures, adding a hyper-parameter 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 development of regularization methods with many penalty parameters.

Keywords: Cross-validation, Oracle inequalities, Regression, Regularization

*Jean Feng was supported by NIH grants DP5OD019820 and T32CA206089.
†Noah Simon was supported by NIH grant DP5OD019820.

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 $\mathrm{E}[\epsilon] = 0$, as follows

$$y = g^*(\boldsymbol{x}) + \epsilon \tag{1}$$

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 hyper-parameter vector $\boldsymbol{\lambda}$. For example, in penalized regression problems, the fitted model can be expressed as the minimizer of the penalized training criterion

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

where P_j are penalty functions and λ_j are penalty parameters. As suggested by the notation in (2), the penalty parameters are the hyper-parameters in this model-estimation procedure.

Given a set of possible hyper-parameters Λ, for a given training dataset T and norm $\|\cdot\|$, there is some oracle hyper-parameter $\tilde{\boldsymbol{\lambda}} \in \Lambda$ that minimizes the difference between the fitted model $\hat{g}(\boldsymbol{\lambda}|T)$ and the true model:

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\lambda \in \Lambda} \left\| g^* - \hat{g}(\boldsymbol{\lambda}|T) \right\|^2$$

$\tilde{\boldsymbol{\lambda}}$ is unknown and often estimated using a single training/validation split or cross-validation. The basic idea is to fit models on a random partition of the observed data and evaluate their error on the remaining data. The final hyper-parameters $\hat{\boldsymbol{\lambda}}$ are the minimizer of the error on this validation set. For a more complete review of cross-validation, refer to Arlot et al. (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. (2006) and a general cross-validation framework (Van Der Laan & Dudoit 2003, van der Laan et al. 2004). To handle continuous Λ, one can use entropy-based approaches (Lecué & Mitchell 2012).

The goal of this paper is to characterize the performance of models when the hyper-parameters must be 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 hyper-parameters are optimized"? To do this, we show that finite-sample oracle inequalities of the form

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

are satisfied with high probability, where $\|\cdot\|$ is some norm; a is some nonnegative constant; and $\delta(J,n)$ is a function of the number of parameters to tune and the number of validation samples n. Under the assumption that the model-estimation procedure is smoothly parameterized by the hyper-parameters, we find that the contribution to δ from tuning J hyper-parameters scales linearly in J. For parametric model-estimation procedures, the additional error from adding a hyper-parameter is roughly equivalent to adding a parameter to the model itself. For semi- and non-parametric model-estimation procedures, this error is generally dominated by the oracle error and the number of hyper-parameters can actually grow without affecting the asymptotic convergence rate.

In this paper, we also specialize these results to penalized regression models of the form (2). We show that the fitted model is indeed smoothly parameterized by the penalty parameters so our oracle inequalities apply. Again, we find that additional penalty pa-

rameters only add a near-parametric error term, which has a negligible effect in semi- and non-parametric settings. This result suggests that the recent interest in combining penalty functions (e.g. elastic net and sparse group lasso (Zou & Hastie 2003, Simon et al. 2013)) may have artificially restricted themselves to two-way combinations. Adding more penalties may lead to better models.

During our literature search, we found few theoretical results addressing the relationship between the number of hyper-parameters and generalization error of the selected model. Much of the previous work only considered tuning a one-dimensional hyper-parameter over a finite Λ, proving asymptotic optimality (van der Laan et al. 2004) and finite-sample oracle inequalities (Van Der Laan & Dudoit 2003, Györfi et al. 2006). 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 et al. 1979, Chetverikov & Liao 2016, Chatterjee & Jafarov 2015). Only Lecué & Mitchell (2012) has a result that is relevant to answering our question of interest by using techniques from empirical process theory. A potential reason for this dearth of literature is that, historically, tuning multiple hyper-parameters has been computationally difficult. However, there have been many proposals recently for overcoming this computational hurdle (Bengio 2000, Foo et al. 2008, Snoek et al. 2012).

Section 2 presents oracle inequalities for model-estimation procedures that are smoothly parameterized by the hyper-parameters. These results answer our question regarding how the number of hyper-parameters affects the model error. Section 3 applies these results to penalized regression models. Section 4 provides a simulation study to support our theoretical results. Section 5 discusses our findings and potential future work. Oracle inequalities for general model-estimation procedures and proofs for all the results are given in the Appendix.

2 Oracle Inequalities

In this section, we establish oracle inequalities for models where the hyper-parameters are tuned by a single training/validation split and cross-validation. We first introduce some notation and formalize the model-estimation procedure.

Let $D^{(n)}$ denote a dataset with n samples from the model (1). The model-estimation procedure accepts some hyper-parameter of dimension J and training data of size n_T to output a fitted model from some model class $\mathcal{G}$. This can be formulated as an operator $\hat{g}^{(n_T)}(\cdot|D^{(n_T)})$ that maps a hyper-parameter vector $\boldsymbol{\lambda}$ from some set $\Lambda \subseteq \mathbb{R}^J$ to a function in $\mathcal{G}$.

In this section, we focus on model-estimation procedures that are Lipschitz.

Definition 1. *Let $\mathcal{F}$ be a function class. Let $\Lambda \subseteq \mathbb{R}^J$. The operator $\hat{f} : \Lambda \mapsto \mathcal{F}$ is C-Lipschitz in $\boldsymbol{\lambda}$ with respect to norm $\|\cdot\|$ over Λ if*

$$\left\| \hat{f}(\boldsymbol{\lambda}) - \hat{f}(\boldsymbol{\lambda}') \right\| \leq C\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2 \quad \forall \boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda \tag{4}$$

We hypothesize that many model-estimation procedures satisfy this Lipschitz assumption since it ensures that the procedure is well-behaved. Section 3 shows that penalized regression models indeed satisfy this assumption. The following results in Sections 2.1 and 2.2 show that the contribution to the error from tuning multiple hyper-parameters for such procedures is roughly parametric. Hence for semi- or non-parametric model-estimation procedures, the error from tuning a fixed number of hyper-parameters is negligible. In fact, we specify a bound on the rate at which the number of hyper-parameters can grow without asymptotically increasing the generalization error of the selected model.

2.1 A single Training/Validation Split

For a single training/validation split, the dataset $D^{(n)}$ 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 hyper-parameter $\hat{\boldsymbol{\lambda}}$ is the minimizer of the validation loss

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \frac{1}{2} \left\| y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) \right\|_V^2 \tag{5}$$

where $\|h\|_V = \frac{1}{n_V} \sum_{i \in V} h^2(x_i)$ for any function h.

We now present a finite-sample oracle inequality for the single training/validation split assuming the model-estimation procedure is Lipschitz. Our oracle inequality is sharp, i.e. $a = 0$ in (3), unlike most other work (Györfi et al. 2006, Lecué & Mitchell 2012, Van Der Laan & Dudoit 2003). Note that the result below is a special case of Theorem 3 in Appendix A.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 independent random variables $\epsilon_1, ...\epsilon_n$ have expectation zero and are uniformly sub-Gaussian with parameters b and B:*

$$\max_{i=1,...,n} B^2 \left(\mathbb{E} e^{|\epsilon_i|^2/B^2} - 1 \right) \leq b^2$$

Suppose there is a constant $C_\Lambda \geq 32e/(n\Delta_\Lambda)$ such that $\hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)})$ is C_Λ-Lipschitz with respect to $\|\cdot\|_V$ over Λ.

Let

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\lambda \in \Lambda} \left\| g^* - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) \right\|_V^2 \tag{6}$$

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

$$\delta^2 \geq c \left(\frac{J \log(n C_\Lambda \Delta_\Lambda)}{n_V} \vee \sqrt{\frac{J \log(n C_\Lambda \Delta_\Lambda)}{n_V} \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V^2} \right) \tag{7}$$

we have

$$Pr\left(\left\|g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 - \left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2 \geq \delta^2 \middle| T, X_V\right) \tag{8}$$

$$\leq c \exp\left(-\frac{n_V \delta^4}{c^2 \left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2}\right) \tag{9}$$

$$+ c \exp\left(-\frac{n_V \delta^2}{c^2}\right) \tag{10}$$

Theorem 1 states that with high probability, the difference between the true model and the selected model is bounded above by the difference between the true model and the oracle model plus δ^2. Hence δ^2 can be thought of as the error incurred during the hyper-parameter selection process. As seen in (7), it is the maximum of two terms: a near-parametric term and a geometric mean of the near-parametric term and the oracle error. To see this more clearly, we express Theorem 1 using asymptotic notation.

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

$$\left\|g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 \leq \left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2 \tag{11}$$

$$+ O_p\left(\frac{J \log(n C_\Lambda \Delta_\Lambda)}{n_V}\right) \tag{12}$$

$$+ O_p\left(\sqrt{\frac{J \log(n C_\Lambda \Delta_\Lambda)}{n_V} \left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2}\right) \tag{13}$$

Hence the error of the selected model is bounded by the error of the oracle model, the near-parameteric term (12), and the geometric mean of the two values (13). We refer to (12) as near-parametric because the error term in parametric regression models are usually $O_p(J/n)$, where J is the parameter dimension and n is the number of training samples. Analogously, (12) is roughly $O_p(J/n_V)$ modulo a $\log n$ term in the numerator.

In the semi- and non-parametric regression settings, the oracle error usually shrinks at a rate of $n^{-\omega}$ where $\omega \in (0, 1)$, which means that for large n, the oracle error will tend to dominate both the error terms. Therefore increasing the number of hyper-parameters for such problems only results in a small increase in the model error. In fact, if the oracle error rate is $O_p(n_T^{-\omega})$, the number of hyper-parameters J can grow at the rate

$$\frac{n_V n_T^{-\omega}}{\log(nC_\Lambda \Delta_\Lambda)} \tag{14}$$

without affecting the asymptotic convergence rate.

The appearance of the parametric term (12) suggests that we can interpret the problem of tuning hyper-parameters as a parametric regression problem over a J-dimensional parameter space where the validation data is the training data. However, this interpretation is an oversimplification. Recall that we perform the training/validation split over the model class

$$\mathcal{G}(T) = \left\{ \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) : \boldsymbol{\lambda} \in \Lambda \right\} \tag{15}$$

Since $\mathcal{G}(T)$ is unlikely to contain the true model g^*, we should instead consider the problem of tuning hyper-parameters as a misspecified parametric regression problem. The minimum bias of $\mathcal{G}(T)$ contributes to the convergence rate via the geometric mean (13).

2.2 Cross-Validation

In this section, 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. Now we give our result with respect to functional L_2 norm

$$\|g - g^*\|^2 = \int |g(x) - g^*(x)|^2 \, dx \tag{16}$$

For K-fold cross-validation our setup is as follows. Let dataset $D^{(n)}$ be randomly partitioned into K sets, which we assume to have equal size for simplicity. Partition k will be denoted $D_k^{(n_V)}$ and its complement will be denoted $D_{-k}^{(n_T)} = D^{(n)} \setminus D_k^{(n_V)}$. We perform our model-estimation procedure given $D_{-k}^{(n_T)}$ for $k = 1, ..., K$ and select the hyper-parameter that minimizes the average validation loss

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \frac{1}{2K} \sum_{k=1}^{K} \left\| y - \hat{g}^{(n_T)}(\boldsymbol{\lambda} | D_{-k}^{(n_T)}) \right\|_{D_k^{(n_V)}}^2 \tag{17}$$

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

$$\bar{g}\left(\hat{\boldsymbol{\lambda}} \middle| D^{(n)}\right) = \frac{1}{K} \sum_{k=1}^{K} \hat{g}^{(n_T)}\left(\hat{\boldsymbol{\lambda}} \middle| D_{-k}^{(n_T)}\right) \tag{18}$$

The following theorem bounds the generalization error of (18). It is an application of Theorem 3.5 in Lecué & Mitchell (2012), which is reproduced in Theorem 4 in the Appendix A.2 for convenience.

Theorem 2. *Let* $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$ *where* $\Delta_\Lambda = \lambda_{\max} - \lambda_{\min} \geq 0$. *Suppose there is a* $G \geq 2$ *such that* $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq G$.

Consider the setting of averaged version of K*-fold cross-validation where* $K \geq 2$ *and the datasets are of size* n *where* n *is divisible by* K. *Let* $n_V = n/K = n_V$ *and* $n_T = n - n_V$. *Suppose random variables* ϵ_i *are independent with expectation zero and are bounded* $\|\epsilon_i\|_\infty \leq \sigma$. *Suppose there is a constant* $C_\Lambda > e^5/G\Delta_\Lambda$ *such that for any dataset* $D^{(n_T)}$, $\hat{g}(\boldsymbol{\lambda}|D^{(n_T)})$ *is* C_Λ*-Lipschitz with respect to* $\|\cdot\|_\infty$ *over* Λ.

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

$$E_{D^{(n)}} \left\| \bar{g}(\hat{\boldsymbol{\lambda}}|D^{(n)}) - g^* \right\|^2 \leq (1+a) \min_{\boldsymbol{\lambda} \in \Lambda} E_{D^{(n_T)}} \left\| \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - g^* \right\|^2 + c \frac{(1+a)^2}{a} \frac{JG \log(GC_\Lambda \Delta_\Lambda) \log n_V}{n_V}$$

As we can see, Theorems 1 and 2 are quite similar. The upper bounds in both theorems depend on the oracle error and a near-parametric term. For parametric model-estimation procedures, tuning hyper-parameters incurs a similar cost as the model-estimation procedure itself. In semi- and non-parametric regression settings, tuning hyper-parameters is a relatively "cheap" and incurs an error that is negligible asymptotically.

There are also some notable differences between Theorems 1 and 2. The Lipschitz condition in Theorem 2 is required to hold with respect to $\|\cdot\|_\infty$, which is stricter than that in Theorem 1. Also, we no longer have a sharp oracle inequality since the oracle error is scaled by $1 + a$ where $a > 0$. These differences occur because we are interested in characterizing the functional L_2 error instead of the L_2 error over the observed covariates.

3 Penalized regression models

Penalized regression is a class of model-generating procedures where hyper-parameters are of particular interest. In this manuscript, we consider penalized regression procedures of the form (2). Penalty functions are used to control model complexity and induce desired characteristics (e.g. smoothness or sparsity). When multiple penalty functions are used, the resulting model exhibits a combination of the desired characteristics. Hence there has been recent interest in combining penalty functions, such as the elastic net or sparse group lasso. However few popular methods use more than two penalties. This has been due to a) the concern that models may overfit the data when selection of many penalty parameters

is required; 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 overfitting (even if many parameters need to be tuned). For computational concerns (b), we refer the reader to recent papers on hyper-parameter estimation (Bengio 2000, Foo et al. 2008, Snoek et al. 2012).

Recall that in our framework, the model-estimation procedure takes in penalty parameters (i.e. hyper-parameters) and then finds the minimizer of the penalized training criterion (2). Models are fit for penalty parameters within some set Λ. As long as we can show that the fitted models are Lipschitz in the penalty parameters over Λ, we can apply Theorems 1 and 2.

Our results do not allow us to consider split-sample validation over all $\lambda \in \mathbb{R}_+^J$: generally as $\lambda_{min} \to 0$, for any finite n, our fitted models become very widely behaved. Instead we restrict ourselves to

$$\Lambda = [n^{-t_{\min}}, n^{t_{\max}}]^J \tag{19}$$

for sufficiently large $t_{\min}, t_{\max} \geq 0$. This regime works well for two reasons: one, our rates depend only quite weakly on $t_{\min}$ and $t_{\max}$; and two, generally oracle λ-values are $\sim n^{-\alpha}$ for some $\alpha \in (0, 1)$ (van de Geer 2000, van de Geer & Muro 2014, Bühlmann & Van De Geer 2011). So long as $t_{\min} > \alpha$, we ensure that Λ contains the optimal penalty parameters over all of $\mathbb{R}_+^J$.

If Λ is of the form (19), we find that the Lipschitz constant for many penalized regression examples are polynomial in n, such as in Sections 3.1 and 3.2. That is, there exist constants $C, \kappa \geq 0$ such that

$$\left\|\hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) - \hat{g}^{(n_T)}(\boldsymbol{\lambda}'|T)\right\| \leq C n^{\kappa} \left\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\right\| \quad \forall \boldsymbol{\lambda}, \boldsymbol{\lambda}' \in \Lambda \tag{20}$$

In our examples, the Lipschitz constant is inversely proportional to $\lambda_{\min}$, so κ grows linearly in $t_{\min}$. Assuming that (19) and (20) hold, we get the following oracle inequalities when penalty parameters are tuned via a single training/validation split and cross-validation.

Corollary 2. *Suppose* $\lambda_{\min} = n^{-t_{\min}}$ *and* $\lambda_{\max} = n^{t_{\max}}$ *for* $t_{\min}, t_{\max} \geq 0$. *Let* $\kappa > 0$.

1. *Single training/validation split: Suppose conditions in Theorem 1 hold with* $C_\Lambda = O_p(n^\kappa)$. *Then*

$$\begin{aligned}\left\|g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 \leq \left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2 \\ + O_p\left(\frac{J(1+\kappa+t_{\max})\log n}{n_V}\right) \\ + O_p\left(\sqrt{\frac{J(1+\kappa+t_{\max})\log n}{n_V}\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2}\right)\end{aligned}$$

2. *Averaged version of* K*-fold cross-validation: Suppose conditions in Theorem 2 hold with* $C_\Lambda = O_p(n^\kappa)$. *For sufficiently large* n, *there is an absolute constant* $c > 0$ *such that for all* $a > 0$,

$$\begin{aligned}E_{D^{(n)}}\left\|\bar{g}(\hat{\boldsymbol{\lambda}}|D^{(n)}) - g^*\right\|^2 \quad \leq \quad & (1+a)\min_{\boldsymbol{\lambda}\in\Lambda} E_{D^{(n_T)}}\left\|\hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - g^*\right\|^2 \\ & + c\frac{(1+a)^2}{a}\frac{JG(t_{\max}+\kappa)(\log n)^2}{n_V}\end{aligned}$$

These results are very similar to those in Corollary 1 and Theorem 2. We still have the same near-parametric term and geometric mean in the oracle inequality for the single training/validation split and a near-parametric term in the oracle inequality for cross-validation. The near-parametric terms have the familiar $\log n$ in the numerator because the error terms are proportional to the log of the Lipschitz constant and $\lambda_{\max}$.

Therefore we reach the same conclusion for the relationship between penalty parameters and the generalization error of the selected model. In high-dimensional and non-parametric penalized regression problems satisfying (20), the error from tuning penalty parameters is negligible compared to the error from solving the penalized regression problem with the oracle penalty parameters.

It remains to characterize penalized problems wherein the fitted models are Lipschitz in the penalty parameters. In this manuscript we will do an in-depth study of additive models, which have the form

$$g(x_1, ..., x_J) = \sum_{j=1}^{J} g_j(x_j) \tag{21}$$

though we believe that these results hold much more generally. We first consider parametric additive models (with potentially growing numbers of parameters) fitted with smooth and non-smooth penalties and then nonparametric additive models. Results for more general penalized regression may be derived in a similar manner (note that if the fitted models are not Lipschitz, one would need the general oracle inequalities from Theorems 3 and 4 instead).

3.1 Parametric additive models

Parametric additive models have the form

$$g(\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(J)}) = \sum_{j=1}^{J} g_j(\boldsymbol{\theta}^{(j)}) \tag{22}$$

where $\boldsymbol{\theta}^{(j)} \in \mathbb{R}^{p_j}$ and $p = \sum_{j=1}^{J} p_j$. The number of dimensions p_j is allowed to grow with n, as commonly done in sieve estimation. For simplicity, let the full parameter vector be denoted $\boldsymbol{\theta} = \left(\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(J)}\right)^{\top}$. Then we can write the training criterion for training data T

as

$$L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) \coloneqq \frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \tag{23}$$

Suppose $\boldsymbol{\theta}^*$ is the minimizer of the expected generalization error

$$\boldsymbol{\theta}^* = \underset{\theta \in \mathbb{R}^p}{\arg\min}\, E_T\left[\|y - g(\boldsymbol{\theta})\|^2\right] \tag{24}$$

3.1.1 Parametric regression with smooth penalties

We begin with the simple case where the penalty functions are smooth. The following lemma states that the fitted models are Lipschitz in the penalty parameter vector.

Lemma 1. *Let*

$$\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}|T) = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\min}\, L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{25}$$

Suppose that $g_j(\boldsymbol{\theta}^{(j)})$ are L-Lipschitz in $\boldsymbol{\theta}^{(j)}$ with respect to $\|\cdot\|_\infty$ for all $j = 1, .., J$. Further suppose $P_j(\boldsymbol{\theta}^{(j)})$ and $g_j(x|\boldsymbol{\theta}^{(j)})$ are twice-differentiable and convex with respect to $\boldsymbol{\theta}^{(j)}$ for any fixed x and all $j = 1, .., J$. Also suppose $L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ is twice-differentiable and convex with respect to $\boldsymbol{\theta}$.

Suppose there exists $m > 0$ such that the Hessian of the penalized training criterion at the minimizer satisfies

$$\nabla_\theta^2 L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\theta = \hat{\theta}(\boldsymbol{\lambda}|T)} \succeq mI \tag{26}$$

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

$$C_{\theta^*, \Lambda} = \frac{1}{2} \|y - g(\boldsymbol{\theta}^*)\|_T^2 + \lambda_{max} \sum_{j=1}^{J} P_j(\boldsymbol{\theta}^{(j),*}) \tag{27}$$

Then, for any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda \coloneqq [\lambda_{\min}, \lambda_{\max}]^J$, we have

$$\left\| g\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda^{(1)}}|T)\right) - g\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda^{(2)}}|T)\right) \right\|_\infty \leq \frac{L^2 J^2 \sqrt{2C_{\theta^*, \Lambda}}}{m\lambda_{min}} \left\| \boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)} \right\|_2 \tag{28}$$

Notice that the result assumes that the training criterion is strongly convex at its minimizer. If this is not true, one can augment the penalty function $P_j(\boldsymbol{\theta}^{(j)})$ with a ridge penalty $\|\boldsymbol{\theta}^{(j)}\|_2^2$ so that the training criterion becomes

$$\frac{1}{2}\|y - g(\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j(\boldsymbol{\theta}^{(j)}) + \frac{w}{2}\|\boldsymbol{\theta}^{(j)}\|_2^2 \right) \tag{29}$$

3.1.2 Parametric regression with non-smooth penalties

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

Definition 2. *The differentiable space of function $f : \mathbb{R}^p \mapsto \mathbb{R}$ at $\boldsymbol{\theta}$ is*

$$\Omega^f(\boldsymbol{\theta}) = \left\{ \boldsymbol{\beta} \middle| \lim_{\epsilon \to 0} \frac{f(\boldsymbol{\theta} + \epsilon\boldsymbol{\beta}) - f(\boldsymbol{\theta})}{\epsilon} \text{ exists} \right\} \tag{30}$$

Definition 3. *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*

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in S} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda} \in W \tag{31}$$

We are interested in the set $\Lambda_{smooth} \subseteq \Lambda$ in which the fitted functions are well-behaved. Consider Λ_{smooth} that satisfies the following conditions:

Condition 1. *For every $\boldsymbol{\lambda} \in \Lambda_{smooth}$, there exists a ball $B(\boldsymbol{\lambda})$ with nonzero radius centered at $\boldsymbol{\lambda}$ such that*

- *For all* $\boldsymbol{\lambda}' \in B(\boldsymbol{\lambda})$*, the training criterion* L_T *is twice differentiable at* $(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}'|T), \boldsymbol{\lambda}')$ *along directions in* $\Omega^{L_T(\cdot,\cdot)}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T), \boldsymbol{\lambda}\right)$.
- $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}\left(\hat{\boldsymbol{\theta}}\left(\boldsymbol{\lambda}|T\right)\right)$ *is a local optimality space for* $L_T\left(\cdot, \boldsymbol{\lambda}\right)$ *over* $B(\boldsymbol{\lambda})$.

Condition 2. *For every* $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda_{smooth}$*, let the line segment between the two points be denoted*

$$\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) = \left\{\alpha\boldsymbol{\lambda^{(1)}} + (1-\alpha)\boldsymbol{\lambda^{(2)}} : \alpha \in [0,1]\right\}$$

Suppose the intersection $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \cap \Lambda_{smooth}^{C}$ *is countable.*

For example, the lasso satisfies these conditions since its solution follows a piecewise linear path (Efron et al. 2004, Tibshirani et al. 2011). We believe other L_p-norm penalties, like the group lasso, also satisfy these conditions.

Equipped with these conditions, we can characterize the smoothness of the fitted functions when the penalties are non-smooth. In fact the Lipschitz constant is exactly the same as that in Lemma 1.

Lemma 2. *Define* $\hat{\boldsymbol{\theta}}^{(j)}\left(\boldsymbol{\lambda}|T\right)$ *as in* (25).
Suppose $g_j(\boldsymbol{\theta}^{(j)})$ *is L-Lipschitz in* $\boldsymbol{\theta}^{(j)}$ *with respect to* $\|\cdot\|_\infty$ *for all* $j = 1, .., J$.
Further suppose that $P_j(\boldsymbol{\theta}^{(j)})$ *and* $g_j(x|\boldsymbol{\theta}^{(j)})$ *are convex with respect to* $\boldsymbol{\theta}^{(j)}$ *for any fixed* x *and all* $j = 1, .., J$. *Also suppose that* $L_T\left(\boldsymbol{\theta}, \boldsymbol{\lambda}\right)$ *is convex with respect to* $\boldsymbol{\theta}$.

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

$$_{U_\lambda}\nabla^2_\theta L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq mI \tag{32}$$

Suppose $\Lambda_{smooth} \subseteq \Lambda$ *satisfies Conditions 1 and 2. Then any* $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$ *satisfies* (28).

3.2 Nonparametric additive models

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

$$\{\hat{g}_j(\boldsymbol{\lambda})\}_{j=1}^J = \underset{g_j \in \mathcal{G}_j : j=1,\ldots,J}{\arg\min} L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right) \tag{33}$$

$$\text{where} \quad L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right) = \frac{1}{2}\left\|\boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j)\right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j) \tag{34}$$

where P_j are now penalty functionals. For some problems, e.g. smoothing splines, we can show that the solution to (33) lies in a parametric family (of growing dimension); and use the results of Sections 3.1.1, and 3.1.2 to give our Lipschitz bounds. However, this is not always possible. Nevertheless we can still obtain similar results in the non-parametric problem. The following lemma states that the fitted functions are Lipschitz with respect to $\|\cdot\|_{D^{(n)}}$. Let $\left\{g_j^*\right\}_{j=1}^J$ be the minimizer of the generalization error

$$\left\{g_j^*\right\}_{j=1}^J = \underset{g_j \in \mathcal{G}_j : j=1,\ldots,J}{\arg\min} E_T\left[\left\|y - \sum_{j=1}^J g_j^*\right\|^2\right] \tag{35}$$

Lemma 3. *Suppose $\mathcal{G}_1, ..., \mathcal{G}_J$ be linear spaces of univariate functions.*

Suppose the penalty functions P_j are twice Gateaux differentiable and convex over $\mathcal{G}_j$. Suppose there is a $m > 0$ such that for all $j = 1, ..., J$, the second Gateaux derivative of the training criterion at $\hat{g}_j^{(n_T)}(\boldsymbol{\lambda}|T)$ satisfies

$$\left\langle D_{g_j}^2 L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right)\Big|_{g_j = \hat{g}_j(\boldsymbol{\lambda}|T)} \circ h_j, h_j \right\rangle \geq m \quad \forall h_j \in \mathcal{G}_j, \|h_j\|_{D^{(n)}} = 1 \tag{36}$$

where $D_{g_j}^2$ is the second Gateaux derivative where both derivatives are taken in the direction of g_j.

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

$$C^*_\Lambda = \frac{1}{2}\left\| y - \sum_{j=1}^{J} g_j^* \right\|_T^2 + \lambda_{max} \sum_{j=1}^{J} P_j(g_j^*) \tag{37}$$

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

$$\left\| \sum_{j=1}^{J} \hat{g}_j\left(\boldsymbol{\lambda}^{(1)}|T\right) - \hat{g}_j\left(\boldsymbol{\lambda}^{(2)}|T\right) \right\|_{D^{(n)}} \leq \frac{J}{m\lambda_{min}} \sqrt{2C^*_\Lambda \frac{n}{n_T}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)} \left\| \boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)} \right\| \tag{38}$$

4 Simulations

We now present a simulation study of nonparametric additive models to understand how the performance changes as the number of penalty parameters increases. Corollary 2 states that increasing the number of penalty parameters J affects the oracle inequalities in two ways: the oracle error may decrease while the remaining terms on the right hand side increase linearly in J. We use two simulations to isolate out these two behaviors.

The data is generated as the sum of univariate functions

$$Y = \sum_{j=1}^{p} g_j(X_j) + \sigma\epsilon, \tag{39}$$

where ϵ are iid standard Gaussian random variables and $\sigma > 0$ was chosen such that the signal to noise ratio was two. X was drawn from a uniform distribution over $\mathcal{X} = [-2, 2]^p$. We fit nonparametric additive models by minimizing the training loss penalized with the Sobolev norm

$$\min_{g_1,\ldots,g_p} \left\| y - \sum_{j=1}^{p} g_j(x_j) \right\|_T^2 + \sum_{j=1}^{p} \lambda_j \int_{\mathcal{X}} \left(g_j''(x)\right)^2 dx \tag{40}$$

where λ_j are the penalty parameters. To vary the number of penalty parameters, we constrain certain λ_j to be equal while allowing others to be completely free. (For instance,

if we wanted a single penalty parameter, λ_j for $j = 1, ..., p$ are all set to the same value.) The penalty parameters are tuned using a training/validation split and we select the one with the minimum error on the validation set.

Simulation 1: The response is the sum of identical univariate sinusoid functions where

$$g_j(x_j) = \sin(x_j) \quad \text{for } j = 1, ..., p. \tag{41}$$

Since the univariate functions are all the same in (41), the oracle penalty parameters $\tilde{\lambda}_j$ should be roughly the same for $j = 1, ..., p$ and the oracle error should stay relatively constant, even if we increase the number of penalty parameters. So for this first simulation, we should be able to observe that the difference in validation error between the selected model and the oracle model grows linearly in J.

Simulation 2: The response is the sum of sinusoid functions with increasing frequency where

$$g_j(x_j) = \sin(x_j * 1.2^{j-4}) \quad \text{for } j = 1, ..., p. \tag{42}$$

Since the Sobolev norms of g_j increase with j, we expect the oracle penalty parameters to be monotonically decreasing, $\tilde{\lambda}_1 > ... > \tilde{\lambda}_p$. As we increase the number of penalty parameters, we expect the upper bound in the oracle inequality to shrink if the decrease in the oracle error term outweighs changes in the other terms.

For both simulations, we used $p = 8$ features and fit the models using 200 training and 200 validation samples. We considered 1, 2, 4, and 8 free penalty parameters, where the penalty parameter structure was nested. That is, when considering k free penalty parameters, we constrained $\lambda_{8\ell/k+1}, ..., \lambda_{8(\ell+1)/k}$ to be equal, for $\ell = 1, ..., k$. Each simulation setting was replicated ten times.

The penalty parameters were tuned using the nonlinear minimization algorithm `nlm` in `R` (CITE) with initializations at $\{\vec{1}, 0.1 \times \vec{1}, 0.01 \times \vec{1}\}$. Multiple initializations were required

since minimizing the validation loss with respect to the penalty parameters is a non-convex problem. We did not use the usual method of grid-search since it is computationally intractable when there are more than three penalty parameters.

The results for Simulation 1 are shown in Figure 1(a). We display the validation loss difference on the y-axis, which is defined as the difference between the validation loss of the selected model and the oracle model

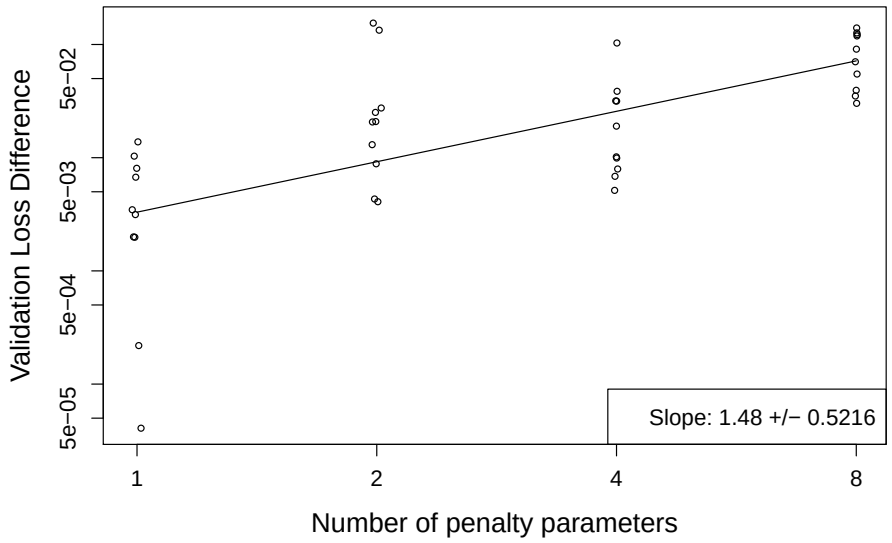
$$\left\| \sum_{j=1}^{2} \hat{g}_j^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) - g_j^* \right\|_V^2 - \left\| \sum_{j=1}^{2} \hat{g}_j^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) - g_j^* \right\|_V^2 .$$

We performed linear regression with the logarithm of the validation loss difference as the response and the logarithm of the number of penalty parameters as the covariate. (We discarded the two outliers.) Since the slope of the fitted model is close to one, our simulation results aligns with our theoretical results: the validation loss difference indeed seems to grow linearly in the number of penalty parameters J.

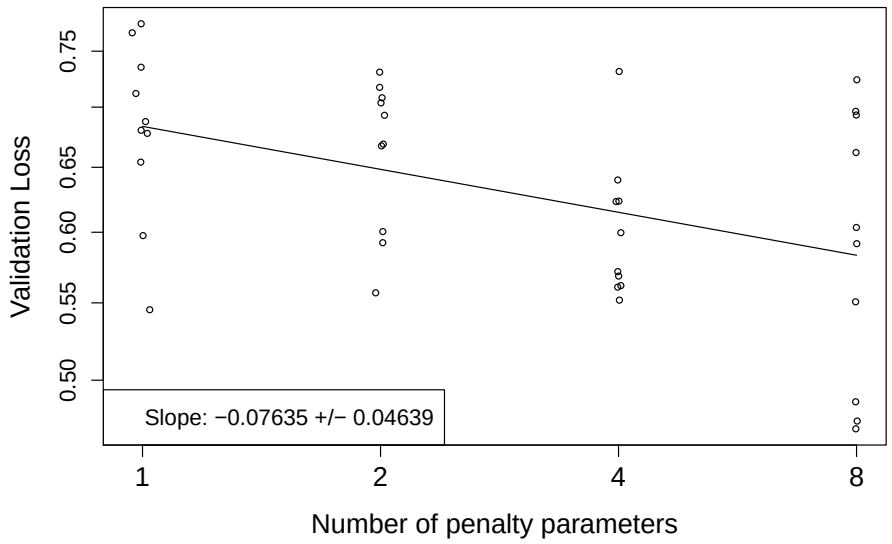
The results for Simulation 2 are shown in Figure 1(b). We see that the validation loss of the selected model decreases as the number of penalty parameters increases. Therefore, if our response is the sum of heterogeneous functions, using multiple penalty parameters can help improve the performance of the model. On the other hand, if the response is the sum of homogeneous functions, allowing multiple penalty parameters will not help since it increases the variance of our model without changing the bias.

5 Discussion

In this paper, we considered model-estimation procedures where the hyper-parameters were tuned via split-sample validation. Our goal was to address the following open question: how does the generalization error of a model grow with the number of hyper-parameters to



(a) Simulation 1: sum of identical sinusoids



(b) Simulation 2: sum of sinusoids with increasing frequency

Figure 1: Performance of the selected model on the validation set as the number of penalty parameters grows.

be estimated? To do so, we established finite-sample oracle inequalities for selection based on a single training/validation split and based on K-fold cross-validation. We showed that if the model-estimation procedure is smoothly parameterized by the hyper-parameters, the generalization error of the model is bounded by a combination of the oracle error, a near-parametric term, and the geometric mean of the two. In a semi- or non-parametric setting, the error incurred from tuning hyper-parameters is asymptotically negligible compared to the oracle error. In the parametric setting, tuning hyper-parameters contributes an error that is roughly on the same order as the oracle error.

We then specialized our results to penalized regression problems with multiple penalty parameters. We showed that in many cases the fitted models are Lipschitz in the penalty parameters and thus the same relationship between the number of penalty parameters and model error holds. This result suggests that the recent efforts to combine penalty functions should push past the artificial barrier of two-way combinations and consider regularization methods with tens or even hundreds of penalty parameters.

Our results may also address the phenomenon that for many model-estimation procedures, the local minimizers seem to perform fairly well (Kunapuli et al. 2008). The oracle inequalities do not need the assumption that the hyper-parameters are global minimizers of the validation loss. The only criteria necessary for the proofs to carry through is that the selected hyper-parameter vector has a smaller validation loss compared to the oracle validation loss.

Finally, an interesting direction of work would be to understand the behavior of hyper-parameter selection in the hyper-parameter space. Bounds on the distance between the selected and oracle hyper-parameters may lend a more intuitive understanding of hyper-parameter selection methods.

References

Arlot, S., Celisse, A. et al. (2010), 'A survey of cross-validation procedures for model selection', *Statistics surveys* **4**, 40–79.

Bengio, Y. (2000), 'Gradient-based optimization of hyperparameters', *Neural computation* **12**(8), 1889–1900.

Bühlmann, P. & Van De Geer, S. (2011), *Statistics for high-dimensional data: methods, theory and applications*, Springer Science & Business Media.

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

Chetverikov, D. & Liao, Z. (2016), 'On cross-validated lasso', *arXiv preprint arXiv:1605.02214* .

De Boor, C., De Boor, C., Mathématicien, E.-U., De Boor, C. & De Boor, C. (1978), *A practical guide to splines*, Vol. 27, Springer-Verlag New York.

Efron, B., Hastie, T., Johnstone, I., Tibshirani, R. et al. (2004), 'Least angle regression', *The Annals of statistics* **32**(2), 407–499.

Foo, C.-s., Do, C. B. & Ng, A. Y. (2008), Efficient multiple hyperparameter learning for log-linear models, *in* 'Advances in neural information processing systems', pp. 377–384.

Golub, G. H., Heath, M. & Wahba, G. (1979), 'Generalized cross-validation as a method for choosing a good ridge parameter', *Technometrics* **21**(2), 215–223.

Green, P. & Silverman, B. (1994), 'Nonparametric regression and generalized linear models, vol. 58 of', *Monographs on Statistics and Applied Probability* .

Györfi, L., Kohler, M., Krzyzak, A. & Walk, H. (2006), *A distribution-free theory of nonparametric regression*, Springer Science & Business Media.

Kunapuli, G., Bennett, K., Hu, J. & Pang, J.-S. (2008), 'Bilevel model selection for support vector machines', *Manuscript, Department of Mathematical Sciences, Rensselaer Polytechnic Institute (March 2007)* .

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

Li, K.-C. (1987), 'Asymptotic optimality for cp, cl, cross-validation and generalized cross-validation: discrete index set', *The Annals of Statistics* pp. 958–975.

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

Simon, N., Friedman, J., Hastie, T. & Tibshirani, R. (2013), 'A sparse-group lasso', *Journal of Computational and Graphical Statistics* **22**(2), 231–245.

Snoek, J., Larochelle, H. & Adams, R. P. (2012), Practical bayesian optimization of machine learning algorithms, *in* 'Advances in neural information processing systems', pp. 2951–2959.

Tibshirani, R. (1996), 'Regression shrinkage and selection via the lasso', *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 267–288.

Tibshirani, R. J., Taylor, J. E., Candes, E. J. & Hastie, T. (2011), *The solution path of the generalized lasso*, Stanford University.

van de Geer, S. (2000), 'Empirical processes in m-estimation (cambridge series in statistical and probabilistic mathematics)'.

van de Geer, S. & Muro, A. (2014), 'The additive model with different smoothness for the components', *arXiv preprint arXiv:1405.6584* .

Van Der Laan, M. J. & Dudoit, S. (2003), 'Unified cross-validation methodology for selection among estimators and a general cross-validated adaptive epsilon-net estimator: Finite sample oracle inequalities and examples'.

van der Laan, M. J., Dudoit, S. & Keles, S. (2004), 'Asymptotic optimality of likelihood-based cross-validation', *Statistical Applications in Genetics and Molecular Biology* **3**(1), 1–23.

Wahba, G. (1990), *Spline models for observational data*, Vol. 59, Siam.

Yuan, M. & Lin, Y. (2006), 'Model selection and estimation in regression with grouped variables', *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **68**(1), 49–67.

Zou, H. & Hastie, T. (2003), 'Regression shrinkage and selection via the elastic net', *Journal of the Royal Statistical Society: Series B. v67* pp. 301–320.