Oracle inequalities for tuning hyperparameters via validation sets with applications to penalized regression

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

November 22, 2016

Abstract

In the regression setting, model-estimation procedures construct models given training data and some hyperparameters. The optimal hyperparameters that minimize the model error are unknown so they are often estimated using validation set approaches. Up to now, there is an open question regarding how the model error grows with the number of hyperparameters. To answer this question, we establish finite-sample oracle inequalities for training/validation split framework and cross-validation. If the model-estimation procedures are smoothly parameterized by the hyperparameters, the error incurred from tuning hyperparameters shrinks at roughly a parametric rate. Hence for semi- and non-parametric model-estimation procedures, this additional error is negligible and for parametric model-estimation procedures, adding a hyperparameter is roughly equivalent to adding a parameter to the model itself. Our main application is penalized regression problems with multiple penalty parameters. We establish the fitted models are Lipschitz in the penalty parameters, so a similar relationship holds between model error and the number of penalty parameters. This result encourages development of regularization methods with many penalty parameters.

Keywords: Hyperparameter selection, Cross-validation, Regularization, Regression

*Jean Feng was supported by NIH grants ???. Noah Simon was supported by NIH grant DP5OD019820. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

1 Introduction

Per the usual regression framework, we observe response $y \in \mathbb{R}$ and predictors $\boldsymbol{x} \in \mathbb{R}^p$. Suppose y is generated by a true model g^* plus random error ϵ with expectation zero, 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 hyperparameter vector $\boldsymbol{\lambda}$. For example, in penalized regression problems, the fitted model can be expressed as the minimizer of the penalized training criterion

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

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

Given a set of possible hyperparameters Λ, there is some oracle hyperparameter $\tilde{\boldsymbol{\lambda}}$ in Λ that minimizes the difference between the fitted model and the true model. Usually $\tilde{\boldsymbol{\lambda}}$ is unknown so it is estimated using 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 hyperparameters $\hat{\boldsymbol{\lambda}}$ 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 validation set 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 training/validation framework 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 et al. 2012).

The goal of this paper is to characterize the performance of models when the hyperparameters must be tuned by some validation set procedure. We are particularly interested in an open question raised in Bengio (2000) (and possibly by others): what is the "amount

of overfitting... when too many hyperparameters are optimized"? To do this, we establish finite-sample oracle inequalities of the form

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

for some norm $\|\cdot\|$ and constant $a \geq 0$. Under the assumption that the model-estimation procedure is smoothly parameterized by the hyperparameters, we find that the error from tuning hyperparameters shrinks at roughly a parametric rate. For parametric model-estimation procedures, the additional error of adding a hyperparameter 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 hyperparameters can actually grow without affecting the asymptotic convergence rate.

The main application in this paper is 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 parameters only add a near-parametric error term, which has a negligible effect on the model error 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-component combinations. Adding more penalties may lead to better models.

We did not find any theoretical results addressing the relationship betwen the number of hyperparameters and model error. Most oracle inequalities only consider one-dimensional hyperparameters (Van Der Laan & Dudoit 2003, van der Laan et al. 2004, Györfi et al. 2006). Within the context of penalized regression problems, the oracle inequalities are also only for the case of a single penalty parameter (Golub et al. 1979, Chetverikov & Liao 2016, Chatterjee & Jafarov 2015). Only Lecué et al. (2012) has a result that is relevant to answering our question of interest. We will use his framework to address the question of cross-validation for multiple hyperparameters. A potential reason for this dearth of literature is that, historically, tuning multiple hyperparameters was 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 hyperparameters. These results answer our question regarding 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. Section 5 discusses our findings and potential future work. Section 6 presents oracle inequalities for general model-estimation procedures and proofs for all the results.

2 Main Result

In this section, we establish oracle inequalities for the model error from tuning hyperparameters by 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 hyperparameter of dimension J and training data of size m to output a fitted model from some model class $\mathcal{G}$. This can be formulated as an operator $\hat{g}^{(m)}(\cdot|D^{(m)})$ that maps a hyperparameter 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 show that the additional error from tuning multiple hyperparameters for such procedures shrinks at roughly a parametric rate. Hence for semi- or non-parametric model-estimation procedures, the error from tuning multiple hyperparameters is very small.

2.1 Training/Validation Split

In the training/validation split framework, 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 hyperparameter $\hat{\boldsymbol{\lambda}}$ is the minimizer of the validation loss

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

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 training/validation split framework assuming the model-estimation procedure is Lipschitz. The oracle inequality is sharp, i.e. $a = 0$ in (3), unlike most other work (Györfi et al. 2006, Lecué et al. 2012, Van Der Laan & Dudoit 2003). The reason for this difference is that the model error is taken with respect to the norm $\|\cdot\|_V$. Note that the result below is a special case of Theorem 3, which applies to general model-estimation procedures.

Theorem 1. *Let $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$ where $0 < \lambda_{\min} < \lambda_{\max}$. Suppose independent random variables $\epsilon_1, ...\epsilon_n$ have expectation zero and are uniformly sub-Gaussian with parameter $b > 0$:*

$$\max_{i=1,...,n} \mathbb{E} e^{t\epsilon_i} \leq e^{b^2 t^2 / 2} \quad \forall t \in \mathbb{R}$$

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

Let

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

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

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

we have

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

Theorem 1 states that as the number of validation samples grows, the difference between the selected model error and the oracle model error shrinks at the rate of δ^2 with high probability. δ^2 can be thought of as the error incurred during the hyperparameter 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*

$$\left\|g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 \leq \left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2 \tag{8}$$

$$+O_p\left(\frac{J\log(nC_\Lambda\lambda_{\max})}{n_V}\right) \tag{9}$$

$$+O_p\left(\sqrt{\frac{J\log(nC_\Lambda\lambda_{\max})}{n_V}\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2}\right) \tag{10}$$

Hence the error of the selected model is bounded by the error of the oracle model, the near-parameteric term (9), and the geometric mean of the two values (10). We refer to (9) 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, (9) is roughly $O_p(J/n_V)$ modulo a $\log n$ term in the numerator. ($\log n$ grows at a sub-polynomial rate, so it changes the convergence rate by a very small amount.)

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

$$\frac{n_V n^\omega}{\log(nC_\Lambda\lambda_{\max})} \tag{11}$$

without affecting the asymptotic convergence rate. Note that for parametric regression problems, this will not be the case. Adding hyperparameters incurs a similar cost as adding parameters to the model itself.

The appearance of the parametric term (9) suggests that we can interpret the problem of tuning hyperparameters 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 due to model misspecification. Recall that we perform training/validation split over the model class

$$\mathcal{G}(T) = \left\{\hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) : \boldsymbol{\lambda} \in \Lambda\right\} \tag{12}$$

$\mathcal{G}(T)$ is unlikely to contain the true model g^* and is biased by

$$\min_{\lambda \in \Lambda} \left\|g^* - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T)\right\|_V^2 \tag{13}$$

This bias term contributes to the convergence rate in the geometric mean (10).

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 L2 norm over the validation covariates. We are now interested in the generalization error

$$\|g - g^*\|^2 = \int |g(x) - g^*(x)|^2 \, dx \tag{14}$$

We will follow the framework in Lecué et al. (2012).

The problem setup for K-fold cross-validation 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 $D_{-k}^{(n_T)} = D \setminus D_k^{(n_V)}$. We perform our model-selection procedure over $D_{-k}^{(n_T)}$ for $k = 1, ..., K$ and select the hyperparameter that minimizes the average validation loss

$$\hat{\boldsymbol{\lambda}} \;=\; \arg\min_{\boldsymbol{\lambda} \in \Lambda} \frac{1}{2K} \sum_{k=1}^{K} \left\| y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D_{-k}^{(n_T)}) \right\|_{D_k^{(n_V)}}^2 \tag{15}$$

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 et al. 2012). We consider the "averaged version of cross-validation" instead

$$\bar{g}\left(\hat{\boldsymbol{\lambda}} \middle| D^{(n)}\right) = \frac{1}{K} \sum_{k=1}^{K} \hat{g}^{(n_T)}\left(\hat{\boldsymbol{\lambda}} \middle| D_{-k}^{(n_T)}\right) \tag{16}$$

The following theorem bounds the generalization error of (16). It is an application of Theorem 3.5 in Lecué et al. (2012), which is reproduced in Theorem **??** for convenience.

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

Consider datasets of size n *where* n *is divisible by* K, *for* $K \geq 2$. *Suppose random variables* ϵ_i *are independent with expectation zero and are bounded* $\|\epsilon_i\|_\infty \leq \sigma$. *Suppose there is a constant* $C_\Lambda > 0$ *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 are absolute constants $c_1, c_2 > 0$ *such that for all* $a > 0$,

$$E_{D^{(n)}} \left\| \bar{g}(\hat{\boldsymbol{\lambda}}|D^{(n)}) - g^* \right\|^2 \leq (1+a) \min_{\boldsymbol{\lambda} \in \Lambda} E_{D^{(n_T)}} \left\| \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - g^* \right\|^2 \tag{17}$$

$$+ c_1 \frac{(1+a)^2}{a} \frac{J}{n_V} \left(G \log(G C_\Lambda \lambda_{max}) \log n + c_2 \right) \tag{18}$$

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 hyperparameters incurs a similar cost as the model-estimation procedure itself. In semi- and non-parametric regression settings, tuning hyperparameters 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 when we are interested in characterizing the generalization error instead.

Finally, since the theorems in this section are finite-sample results, one could try to minimize the upper bound by increasing the number of hyperparameters or changing the ratio between the training and validation set sizes. Unfortunately, optimizing the upper bound in these oracle inequalities require knowing characteristics about the error variables. Instead one may need to rely on heuristic approaches (or even another layer of cross-validation).

3 Penalized regression models

The main application of this paper is penalized regression models 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 no popular methods use more than two penalties since there is a concern that models may overfit the data when there are many penalty parameters. In this section, we answer this open question by applying the oracle inequalities from Section 2.

Recall that in our framework, the model-estimation procedure is to find the minimizer of the penalized training criterion, where the penalty parameters serve as the hyperparameters. Models are fit for penalty parameters over some range 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.

We are particularly interested in the case where the limits of Λ are polynomial in the sample size. The reason is that the oracle penalty parameters over $\mathbb{R}_+^J$ are usually polynomial in n (van de Geer 2000, van de Geer & Muro 2014, Bühlmann & Van De Geer 2011). In order to ensure Λ contains the optimal penalty parameters, we consider

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

for sufficiently large $t_{\min}, t_{\max} \geq 0$. (Ideally we would set $\Lambda = \mathbb{R}_+^J$, but the penalized regression can be ill-behaved for very small penalty parameters.)

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

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

In our examples, the Lipschitz constant is inversely proportional to $\lambda_{\min}$, so κ is grows linearly in $t_{\min}$. Assuming that (19) and (20) hold, we get the following oracle inequalities when tuning penalty parameters via 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 $C, \kappa > 0$.*

1. *Training/validation split: Suppose conditions in Theorem 1 hold with $C_\Lambda = Cn^\kappa$. Then*

$$\left\|g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 \leq \left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2 \tag{21}$$

$$+O_p\left(\frac{J(1+\kappa+t_{\max})\log n}{n_V}\right) \tag{22}$$

$$+O_p\left(\sqrt{\frac{J(1+\kappa+t_{\max})\log n}{n_V}\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2}\right) \tag{23}$$

2. *Averaged version of K-fold cross-validation: Suppose conditions in Theorem 2 hold with* $C_\Lambda = Cn^\kappa$. *Then for sufficiently large* n, *there is an absolute constant* $c_1 > 0$ *such that for all* $a > 0$,

$$E_{D^{(n)}} \left\| \bar{g}(\hat{\boldsymbol{\lambda}}|D^{(n)}) - g^* \right\|^2 \leq (1+a) \min_{\boldsymbol{\lambda}\in\Lambda} E_{D^{(n_T)}} \left\| \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - g^* \right\|^2 \quad (24)$$

$$+c_1 \frac{(1+a)^2}{a} \frac{J(t_{\max}+\kappa)(\log n)^2}{n_V} \quad (25)$$

We get very similar results to those in Corollary 1 and Theorem 2. We still have the same near-parametric term and geometric mean in the oracle inequality for 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 model error. In semi- and non-parametric penalized regression problems, the error from tuning penalty parameters is negligible compared the error from solving the penalized regression problem itself. In parametric regression problems, the error from tuning penalty parameters is comparable to that from solving the penalized regression problem.

It remains to show that the fitted models are Lipschitz in the penalty parameters. We will do an in-depth study of additive models, which have the form

$$g(x_1, ..., x_J) = \sum_{j=1}^{J} g_j(x_j) \quad (26)$$

We first consider parametric additive models fitted with smooth and non-smooth penalties and then nonparametric additive models. Results for more general penalized regression problems are included in Section 6.

3.1 Parametric additive models

Parametric additive models have the form

$$g(\cdot|\boldsymbol{\theta}^{(1)}, ..., \boldsymbol{\theta}^{(J)}) = \sum_{j=1}^{J} g_j(\cdot|\boldsymbol{\theta}^{(j)}) \quad (27)$$

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

$$L_T\left(\boldsymbol{\theta}, \boldsymbol{\lambda}\right) \coloneqq \frac{1}{2}\left\|y - g(\boldsymbol{\theta})\right\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \tag{28}$$

Suppose the true model has parameters $\boldsymbol{\theta}^*$.

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*

$$\hat{\boldsymbol{\theta}}^{(j)}\left(\boldsymbol{\lambda}|T\right) = \underset{\boldsymbol{\theta} \in \mathbb{R}^p}{\arg\min}\, L_T\left(\boldsymbol{\theta}, \boldsymbol{\lambda}\right) \tag{29}$$

Suppose that $g_j(\boldsymbol{\theta}^{(j)})$ are L-Lipschitz in $\boldsymbol{\theta}^{(j)}$ with respect to $\|\cdot\|_\infty$ for all $j = 1, .., J$.

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 $j = 1, .., J$. Suppose $L_T\left(\boldsymbol{\theta}, \boldsymbol{\lambda}\right)$ is twice-differentiable and convex with respect to $\boldsymbol{\theta}$.

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

$$\nabla_\theta^2 L_T\left(\boldsymbol{\theta}, \boldsymbol{\lambda}\right)\big|_{\theta = \hat{\theta}(\boldsymbol{\lambda}|T)} \succeq mI \tag{30}$$

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

$$C_{\theta^*, \Lambda} = \frac{1}{2}\left\|y - g(\boldsymbol{\theta}^*)\right\|_T^2 + \lambda_{max} \sum_{j=1}^{J} P_j(\boldsymbol{\theta}^{(j),*}) \tag{31}$$

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

$$\left\| g\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda^{(1)}}|T)\right) - g\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda^{(2)}}|T)\right)\right\|_\infty \leq \frac{L^2 J^2 \sqrt{2C_{\theta^*, \Lambda}}}{m\lambda_{min}} \left\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\right\| \tag{32}$$

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

$$\frac{1}{2}\left\|y - g(\boldsymbol{\theta})\right\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j(\boldsymbol{\theta}^{(j)}) + \frac{w}{2}\|\boldsymbol{\theta}^{(j)}\|_2^2\right) \tag{33}$$

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 use the approach in Feng & Simon (TBD- CITE?). We begin with the following definitions:

Definition 2. *The differentiable space of a real-valued function f at $\boldsymbol{\theta}$ is*

$$\Omega^f(\boldsymbol{\theta}) = \left\{ \boldsymbol{\beta} \middle| \lim_{\epsilon \to 0} \frac{f(\boldsymbol{\theta} + \epsilon\boldsymbol{\beta}) - f(\boldsymbol{\theta})}{\epsilon} \text{ exists} \right\} \tag{34}$$

Definition 3. *S is a local optimality space for convex function $f(\cdot, \boldsymbol{\lambda}) : \mathbb{R}^p \times \mathbb{R}^J \mapsto \mathbb{R}$ over $W \subseteq \mathbb{R}^J$ if*

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in S} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda} \in W \tag{35}$$

We can now characterize a set $\Lambda_{smooth} \subseteq \Lambda$ over which the fitted functions are well-behaved. Λ_{smooth} must satisfy 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 $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \cap \Lambda^C_{smooth}$ is countable.

In lasso and group lasso problems, it is hypothesized that almost every penalty parameter satisfies these properties. (CITE?) 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* $\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}|T)$ *as in* (29).

Suppose $g_j(\boldsymbol{\theta}^{(j)})$ *is L-Lipschitz in* $\boldsymbol{\theta}^{(j)}$ *with respect to* $\|\cdot\|_\infty$ *for all* $j = 1, .., J$.

Suppose $P_j(\boldsymbol{\theta}^{(j)})$ *and* $g_j(x|\boldsymbol{\theta}^{(j)})$ *are convex with respect to* $\boldsymbol{\theta}^{(j)}$ *for any fixed* x *and all* $j = 1, .., J$. *Suppose* $L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ *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 is a* $m > 0$ *such that the Hessian of the penalized training criterion with respect to the differentiable space at the minimizer satisfies*

$$U_\lambda \nabla^2_\theta L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq mI \tag{36}$$

Suppose $\Lambda_{smooth} \subseteq \Lambda \coloneqq [\lambda_{\min}, \lambda_{\max}]^J$ *satisfies Conditions 1 and 2.*

Then any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$ *satisfies* (32).

3.2 Nonparametric additive models

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

$$\{\hat{g}_j(\boldsymbol{\lambda})\}_{j=1}^J = \underset{g_j \in \mathcal{G}_j : j=1,\ldots,J}{\arg\min} L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right) \tag{37}$$

$$\text{where} \quad L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right) = \frac{1}{2}\left\|\boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j)\right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j) \tag{38}$$

where P_j are now penalty functionals. The following lemma states that the fitted functions are Lipschitz with respect to $\|\cdot\|_{D^{(n)}}$. Let the true model be $\left\{g_j^*\right\}_{j=1}^J$.

Lemma 3. *Suppose* $\mathcal{G}_1, ..., \mathcal{G}_J$ *are convex univariate function classes.*

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 training criterion has a twice Gateaux derivative with respect to* g_j *at* $\hat{g}_j^{(n_T)}(\boldsymbol{\lambda}|T)$ *satisfies*

$$\left\langle D^2_{g_j} L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right)\Big|_{g_j=\hat{g}_j^{(n_T)}(\boldsymbol{\lambda}|T)} \circ h_j, h_j \right\rangle \geq m \quad \forall h_j \in \mathcal{G}_j, \|h_j\|_{D^{(n)}} = 1 \tag{39}$$

where $D^2_{g_j}$ *is the second Gateuax derivative with respect to* g_j.

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

$$C_{\theta^*,\Lambda} = \frac{1}{2}\left\| y - \sum_{j=1}^{J} g_j^* \right\|_T^2 + \lambda_{max} \sum_{j=1}^{J} P_j(g_j^*) \tag{40}$$

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

$$\left\| \sum_{j=1}^{J} \hat{g}_j^{(n_T)}\left(\boldsymbol{\lambda}^{(1)}|T\right) - \hat{g}_j^{(n_T)}\left(\boldsymbol{\lambda}^{(2)}|T\right) \right\|_{D^{(n)}} \leq \frac{J}{m\lambda_{min}} \sqrt{2C_{\theta^*,\Lambda}\frac{n}{n_T}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)} \left\| \boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)} \right\| \tag{41}$$

4 Simulations

We now provide a simulation study for the prediction error bound given in Theorem 1. The penalty parameters are chosen by a training/validation split. We show that the error of the selected model converges to that of the oracle model at the near-parametric rate.

Observations were generated from the model

$$y = \exp(x_1) + x_2^2 + \sigma\epsilon \tag{42}$$

where $\epsilon \sim N(0,1)$ and σ scaled the error term such that the signal to noise ratio was 2. The covariates x_1 and x_2 were uniformly distributed over the interval $(-1, 1)$.

We fit a smoothing splines using the Sobolev penalty (De Boor et al. 1978, Wahba 1990, Green & Silverman 1994). The fitted models were of the form

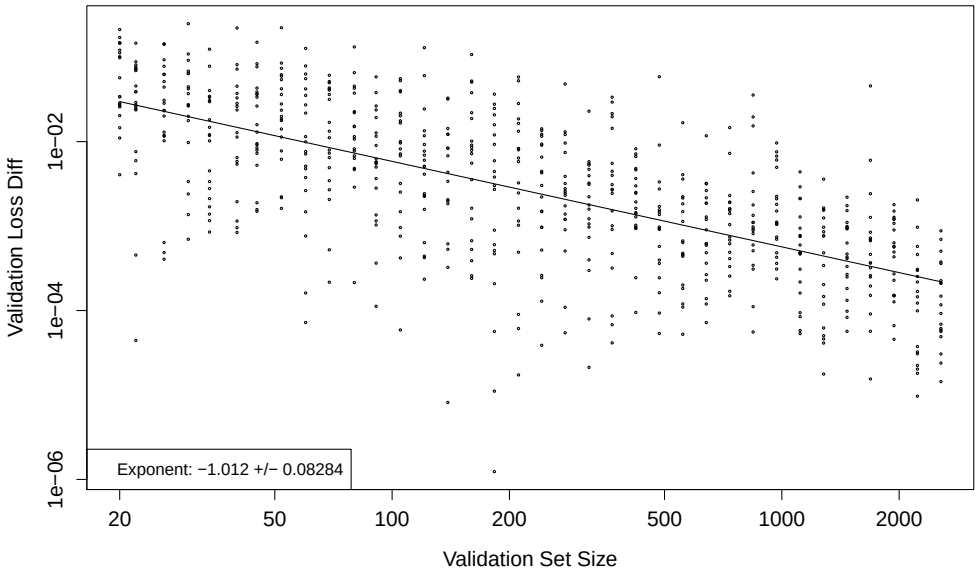
$$\left\{\hat{g}_j^{(n_T)}(\boldsymbol{\lambda}|T)\right\}_{j=1}^{2} = \underset{g_1,g_2}{\arg\min} \|y - g_1(x_1) - g_2(x_2)\|_T^2 + \lambda_1 \int_{-1}^{1} (g_1^{(2)}(x))^2 dx + \lambda_2 \int_{-1}^{1} (g_2^{(2)}(x))^2 dx \tag{43}$$

The training set contained 100 samples and models were fitted with 10 knots. A grid search was performed over the penalty parameter values $\{10^{-9+0.05i} : i = 0, ..., 140\}$. We tested 36 validation set sizes $n_V = \lfloor 20 * 2^i \rfloor$ for equally log-spaced intervals from $i = 0$ to $i = 7$. A total of 20 simulations were run for each validation set size.

Figure 4 plots the difference of between the model loss and the oracle loss

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

Figure 1: Validation loss difference between oracle and selected model as validation set size grows



as the validation set size increases. The difference of the validation losses drops at a rate of about n^{-1}. This rate is in fact faster than that in Theorem 1; the geometric mean in the oracle inequality seems to play no role in the convergence rate. We conjecture that smoothing splines may satisfy additional regularity conditions such that the geometric mean may be discarded.

5 Discussion

In this paper, we established oracle inequalities for hyperparameter selection using a training/validation split framework or K-fold cross-validation. The results address the open question regarding model error when many hyperparameters need to be tuned. If the model-estimation procedure is smoothly parameterized by the hyperparameters, then in a non-parametric setting or parametric setting where p grows with n, the oracle error is the dominating term in the upper bound. In the parametric setting, the tuning penalty parameter problem contributes an error that is on the same order as the oracle error.

We then applied our results to penalized regression problems. We showed that the fitted models are Lipschitz in the penalty parameters and get the same relationship between penalty

parameters and model error. This suggests that the we should push past the artificial barrier of two penalty parameters and consider combining tens or even hundreds of penalty parameters. For example, Feng and Simon (TBD) fit an un-pooled version of sparse group lasso model with a hundred penalty parameters and significantly decreased the model's generalization error.

A major caveat to our results is that we have assumed that it is possible to find the global minimizer of the validation loss over the penalty parameter space. Unfortunately, this is currently computationally intractable since the validation loss is not convex in the penalty parameters. Better optimization methods need to be developed to solve this problem. Current methods only guarantee finding a local minimizer. More theoretical results are necessary to understand how robust the models are to optimization error.

An interesting direction for future work is to understand the behavior of hyperparameter selection in the hyperparameter space, in contrast to our approach of characterizing model error. For example, bounds on the distance between the selected and oracle penalty parameters

$$\left\| \hat{\boldsymbol{\lambda}} - \tilde{\boldsymbol{\lambda}} \right\|_2 \tag{44}$$

can perhaps lend to a more intuitive understanding of hyperparameter selection methods.

6 The Proof

For functions f and g and a dataset $D^{(m)}$ with m samples, we denote the inner product of f and g at covariates D as $\langle f, g \rangle_D = \frac{1}{m} \sum_{x_i \in D} f(x_i) g(x_i)$.

6.1 Proof for training/validation split

Theorem 1 is a special case of Theorem 3, which applies to general model-estimation procedures. The proof is based on the inequality below. Inequalities of this form are often called a "basic inequality", since it is derived directly from the definition and the quantity of interest, the difference in the error of the selected model and the oracle model, is bounded by an empirical process term.

Lemma 4. *Basic inequality*

$$\left\|g^* - \hat{g}^{n_T}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 - \left\|g^* - \hat{g}^{n_T}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2 \leq \left\langle \epsilon, \hat{g}^{n_T}(\tilde{\boldsymbol{\lambda}}|T) - \hat{g}^{n_T}(\hat{\boldsymbol{\lambda}}|T)\right\rangle_V \tag{45}$$

Proof. By definition,

$$\left\|y - \hat{g}^{n_T}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 \leq \left\|y - \hat{g}^{n_T}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2 \tag{46}$$

□

We are therefore interested in bounding the empirical process term in (45). A common approach is to use a measure of complexity of the function class. In the training/validation split framework where we treat the training set as fixed, we only need to consider the complexity of the fitted models from the model-selection procedure

$$\mathcal{G}(T) = \left\{\hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) : \boldsymbol{\lambda} \in \Lambda\right\} \tag{47}$$

This model class can be considerably less complex compared to the original function class $\mathcal{G}$, such as the special case in Theorem 1 where we suppose $\mathcal{G}(T)$ is Lipschitz. For this proof, we will use metric entropy as a measure of model class complexity. We recall its definition below.

Definition 4. *Let $\mathcal{F}$ be a function class. Let the covering number $N(u, \mathcal{F}, \|\cdot\|)$ be the smallest set of u-covers of $\mathcal{F}$ with respect to the norm $\|\cdot\|$. The metric entropy of $\mathcal{F}$ is defined as the log of the covering number:*

$$H(u, \mathcal{F}, \|\cdot\|) = \log N(u, \mathcal{F}, \|\cdot\|) \tag{48}$$

We will bound the empirical process term using the following Lemma, which is a simplification of Corollary 8.3 in van de Geer (2000).

Lemma 5. *adapt to our new notation Let Q_m be the empirical distributon of m observations at covariates x_i.*

Suppose ϵ are m independent sub-gaussian errors with parameter $b > 0$. Suppose the model class $\mathcal{F}(T)$ has elements $\sup_{f \in \mathcal{F}_n(T)} \|f\|_{Q_m} \leq R$ and satisfies

$$\psi_T(R) \geq \int_0^R H^{1/2}(u, \mathcal{F}(T), \|\cdot\|_{Q_m}) du$$

There is a constant $a > 0$ dependent only on b such that for all $\delta > 0$ satisfying

$$\sqrt{m}\delta \geq a(\psi_T(R) \vee R)$$

we have

$$Pr\left(\sup_{f\in\mathcal{F}(T)}\left|\frac{1}{m}\sum_{i=1}^{m}\epsilon_i f(x_i)\right| \geq \delta \middle| T\right) \leq a\exp\left(-\frac{m\delta^2}{4a^2R^2}\right)$$

We are now ready to prove the oracle inequality. It uses a standard peeling argument. For readability, we will use the simplified notation $\hat{g}(\hat{\boldsymbol{\lambda}})$ and $\hat{g}(\tilde{\boldsymbol{\lambda}})$.

Theorem 3. *Let $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$ where $0 < \lambda_{\min} < \lambda_{\max}$. Suppose independent random variables $\epsilon_1, ...\epsilon_n$ have expectation zero and are uniformly sub-Gaussian with parameter $b > 0$. Suppose there is a function $\psi : \mathbb{R} \mapsto \mathbb{R}$ and constant $r > 0$ such that*

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi(R) \quad \forall R > r \tag{49}$$

Also, suppose $\psi(u)/u^2$ is non-increasing in u for all $u > r$. Let $\tilde{\boldsymbol{\lambda}}$ be defined as in (6).

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

$$\sqrt{n_V}\delta^2 \geq c\left(\psi(\delta) \vee \delta \vee \psi\left(4\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2\right) \vee 4\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2\right) \tag{50}$$

we have

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

Proof. The following probabilities are all conditional on X_V and T. We leave them out for readability.

$$Pr\left(\left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \geq \delta^2\right) \tag{51}$$

$$= \sum_{s=0}^{\infty} Pr\left(2^{2s}\delta^2 \leq \left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \leq 2^{2s+2}\delta^2\right) \tag{52}$$

$$\leq Pr\left(2^{2s}\delta^2 \leq 2\left\langle\epsilon, \hat{g}(\hat{\boldsymbol{\lambda}}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V \wedge \left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V^2 \leq 2^{2s+2}\delta^2 + 2\left|\left\langle\hat{g}(\tilde{\boldsymbol{\lambda}}) - \hat{g}(\hat{\boldsymbol{\lambda}}), \hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\rangle_V\right|\right) \tag{53}$$

where we applied the basic inequality in the last inequality. Each summand in (53) can be bounded by splitting the event into the cases where $2^{2s+2}\delta^2$ is bigger or $2\left|\left\langle \hat{g}(\tilde{\boldsymbol{\lambda}}) - \hat{g}(\hat{\boldsymbol{\lambda}}), \hat{g}(\tilde{\boldsymbol{\lambda}}) - g^* \right\rangle_V\right|$ is bigger. Splitting up the probability and applying Cauchy Schwarz gives us the following bound for (51)

$$Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\cdot|\boldsymbol{\lambda})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V} 2\left\langle \epsilon, \hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V \geq \delta^2\right) \tag{54}$$

$$+\sum_{s=0}^{\infty} Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\cdot|\boldsymbol{\lambda})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2^{s+3/2}\delta} 2\left\langle \epsilon, \hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V \geq 2^{2s}\delta^2\right) \tag{55}$$

We can bound both (54) and (55) using Lemma 5. For our choice of δ in (50), there is some constant $a > 0$ dependent only on b such that (54) is bounded above by

$$a\exp\left(-\frac{n_V\delta^4}{4a^2\left(16\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2\right)}\right)$$

In addition, our choice of δ from (50) and our assumption that $\psi(u)/u^2$ is non-increasing implies that the condition in Lemma 5 is satisfied for all $s = 0, 1, ..., \infty$ simultaneously. Hence for all $s = 0, 1, ..., \infty$, we have

$$Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\cdot|\boldsymbol{\lambda})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2^{s+3/2}\delta} 2\left\langle \epsilon, \hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V \geq 2^{2s}\delta^2\right) \leq a\exp\left(-n_V\frac{2^{4s-2}\delta^4}{4a^2 2^{2s+3}\delta^2}\right)$$

Putting this all together, there is a constant c such that (51) is bounded above by

$$c\exp\left(-\frac{n_V\delta^4}{4c^2\left(16\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2\right)}\right) + c\exp\left(-\frac{n_V\delta^2}{c^2}\right) \tag{56}$$

□

We can apply Theorem 3 to get Theorem 1. Before proceeding, we determining the entropy of $\mathcal{G}(T)$ when the functions are Lipschitz in the hyperparameters.

Lemma 6. *Let $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$. Suppose $\mathcal{G}(T)$ is C-Lipschitz in $\boldsymbol{\lambda}$ with respect to some norm $\|\cdot\|$. Then the entropy of $\mathcal{G}(T)$ with respect to $\|\cdot\|$ is*

$$H\left(u, \mathcal{G}(T), \|\cdot\|\right) \leq J\log\left(\frac{4C\left(\lambda_{max} - \lambda_{min}\right) + 2u}{u}\right) \tag{57}$$

Proof. Using a slight variation of the proof for Lemma 2.5 in van de Geer (2000), we can show

$$N\left(u, \Lambda, \|\cdot\|_2\right) \leq \left(\frac{4\left(\lambda_{max} - \lambda_{min}\right) + 2u}{u}\right)^J \tag{58}$$

Under the Lipschitz assumption, a δ-cover for Λ is a $C\delta$-cover for $\mathcal{G}(T)$. The covering number for $\mathcal{G}(T)$ wrt $\|\cdot\|_V$ is bounded by the covering number for Λ as follows

$$\begin{aligned} N\left(u, \mathcal{G}(T), \|\cdot\|_V\right) &\leq N\left(\frac{u}{C}, \Lambda, \|\cdot\|_2\right) & (59) \\ &\leq \left(\frac{4\left(\lambda_{max} - \lambda_{min}\right) + 2u/C}{u/C}\right)^J & (60) \end{aligned}$$

□

Now we ready to prove Theorem 1.

Proof. By Lemma 6, we have

$$\begin{aligned} \int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du &= \int_0^R \left(J \log\left(\frac{4C_\Lambda\left(\lambda_{max} - \lambda_{min}\right) + 2u}{u}\right)\right)^{1/2} du \\ &\leq J^{1/2} \int_0^R \left[\log 4 + \log\left(\frac{8C_\Lambda\left(\lambda_{max} - \lambda_{min}\right)}{u}\right)\right]^{1/2} du \\ &\leq RJ^{1/2}\left[\int_0^1 \left(\log 4 + \log\left(\frac{8C_\Lambda\left(\lambda_{max} - \lambda_{min}\right)}{R}\right) + J\log\frac{1}{v}\right) dv\right]^{1/2} \\ &= R\left[J\left(1 + \log\left(32C_\Lambda\left(\lambda_{max} - \lambda_{min}\right)\right) + \log\frac{1}{R}\right)\right]^{1/2} \end{aligned}$$

where the second inequality follows from a change of variables and the concavity of the square root function. If we restrict $R > n^{-1}$ and $C_\Lambda \geq 32e/(n\lambda_{\max})$, then

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du \leq \psi(R) \coloneqq 2R\left(J\log(C_\Lambda \lambda_{\max} n)\right)^{1/2} \tag{61}$$

Applying Theorem 3, we get our desired result.

□

6.2 Proof for cross-validation

We now present the proof for Theorem 2, which is an application of Theorem 3.5 in Lecué et al. (2012). We reproduce Theorem 3.5 below for convenience. The theorem calculates entropy with respect to the Orlicz norm $\|\cdot\|_\phi = \inf\{C > 0 : E[\exp(|f|/C) - 1] \leq 1\}$.

Theorem 4. *Let* $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$. *Let* $\mathcal{Q} = \{\|g^* - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T)\|_2^2 : \boldsymbol{\lambda} \in \Lambda\}$. *Suppose there is* $G > 0$ *such that* $\sup_{g \in \mathcal{G}} \|g\|_\infty \leq G$.

Suppose there is a $d_{\min}$ *and a strictly increasing function* ψ *such that* ψ^{-1} *is strictly convex, the convex conjugate* ψ^* *of* ψ^{-1} *increases,* $\psi^*(\infty) = \infty$ *and there exists* $r \geq 1$ *such that* $\psi^*(x)/x^r$ *is decreasing in* x *and*

$$\psi(d) \geq \int_0^{\sqrt{d}} H^{1/2}(u, \mathcal{Q}_d, \|\cdot\|_2)du + \frac{\log n_V}{\sqrt{n_V}} \int_0^{2G} H(u, \mathcal{Q}_d, \|\cdot\|_\phi)du \quad \forall d > d_{\min} \tag{62}$$

where $\mathcal{Q}_d = \{Q \in \mathcal{Q} : \|Q\|_2 \leq \sqrt{d}\}$.

Suppose that the model-estimation procedure is exchangeable (i.e. any ordering of the same training data produces the same fitted model).

Then for every $a > 0$ *and* $q > 1$, *the following inequality holds*

$$E_{D^{(n)}} \left\| \bar{g}(\hat{\boldsymbol{\lambda}}|D^{(n)}) - g^* \right\|^2 \leq (1+a) \inf_{\boldsymbol{\lambda} \in \Lambda} \left[E_{D^{(n_T)}} \left\| \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - g^* \right\|^2 \right] \tag{63}$$

$$+ \frac{ac}{q} \left[\psi^* \left(\frac{2q^r(1+a)}{a\sqrt{n_V}} \right) \vee d_{\min} \right] \tag{64}$$

In order to apply Theorem 4, we need to determine the entropy of the loss functions with respect to the Orlicz norm.

Proof. First we determine ψ in (62). Note that because $\|\cdot\|_\phi \leq 2\|\cdot\|_\infty$ and $\|\cdot\|_2 \leq \|\cdot\|_\infty$, then both $H(2u, \mathcal{Q}_d(T), \|\cdot\|_\phi)$ and $H(u, \mathcal{Q}_d(T), \|\cdot\|_2)$ are bounded by $H(u, \mathcal{Q}_d(T), \|\cdot\|_\infty)$. By Lemma 6, we know

$$H(u, \mathcal{Q}_d(T), \|\cdot\|_\infty) \leq J \log \left[\frac{16GC\lambda_{\max} + 2u}{u} \right] \tag{65}$$

Hence we can let

$$\psi(d) \coloneqq \sqrt{d}K_{n,1} + \frac{K_{n,2}}{\sqrt{n_V}} \tag{66}$$

for $K_{n,1} = \left[J \left(1 + \log \left(128\sqrt{n_V} GC\lambda_{max} \right) \right) \right]^{1/2}$ and $K_{n,2} = \log n_V 2GJ \left(1 + \log \left(128GC\lambda_{max} \right) \right)$. $\psi(d)$ is a valid upper bound in (62) for all $d > n_V^{-1}$.

We can show that the convex conjugate of ψ^{-1} is

$$\psi^*(z) = \frac{z^2 K_{n,1}^2}{4} + \frac{zK_{n,2}}{\sqrt{n_V}} \tag{67}$$

Plugging in (67) into Theorem 4 gives us the result in Theorem 2. □

6.3 Proof for Lispchitz conditions

Proof of Lemma 1

Proof. By the gradient optimality conditions, we have for all $j = 1 : J$

$$\nabla_{\theta^{(j)}} \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} = 0 \tag{68}$$

Now we implicitly differentiate with respect to $\boldsymbol{\lambda}$

$$\nabla_{\lambda} \left\{ \nabla_{\theta^{(j)}} \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\} = 0 \tag{69}$$

Define the following matrices

$$S = \nabla_{\theta} \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda|T)}$$

$$D = diag \left(\left\{ \nabla^2_{\boldsymbol{\theta}^{(j)}} \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right\}_{j=1}^{J} \right) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda|T)}$$

$$M : \text{column } M_j = \nabla_{\theta} P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda|T)}$$

From the product rule and chain rule, we can then write the system of equations in (69) as

$$\left(\begin{array}{cccc} \nabla_{\lambda} \hat{\boldsymbol{\theta}}_1(\boldsymbol{\lambda}) & \nabla_{\lambda} \hat{\boldsymbol{\theta}}_2(\boldsymbol{\lambda}) & ... & \nabla_{\lambda} \hat{\boldsymbol{\theta}}_p(\boldsymbol{\lambda}) \end{array} \right) = -M^{\top} (S + D)^{-1} \tag{70}$$

We now bound each column in M. From the (68) and Cauchy Schwarz, we have

$$\left\| \nabla_{\theta^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\| \leq \frac{1}{\lambda_{min} \sqrt{n_T}} \left\| y - g(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right\|_T \sqrt{\sum_{i=1}^{n_T} \left\| \nabla_{\theta^{(j)}} g_j(x_i|\boldsymbol{\theta}^{(j)}) \right\|_2^2}$$

The norm of the gradients of g_j can be bounded since g_j is Lipschitz. Also, by definition of $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$, we have

$$\begin{aligned} \frac{1}{2} \left\| y - g(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j \left(\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}) \right) &\leq \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta}^*)\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta}^{(j),*}) \quad (71) \\ &\leq C_{\theta^*,\Lambda} \quad (72) \end{aligned}$$

Hence for all $j = 1, ..., J$

$$\left\| \nabla_{\theta^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\| \leq \frac{L}{\lambda_{min}} \sqrt{2 C_{\theta^*,\Lambda}}$$

Now we bound the norm of $\nabla_\lambda \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$. From (70), we have for all $j = 1, ..., J$

$$\begin{aligned}
\|\nabla_\lambda \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})\| &= \|M^\top (S+D)^{-1} e_k\| \\
&\leq \sum_{j=1}^{J} \left\| \nabla_{\theta^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\|_2 \left\|(S+D)^{-1}\right\|_2 \\
&\leq J\left(\frac{L}{\lambda_{min}}\sqrt{2C_{\theta^*,\Lambda}}\right)\frac{1}{m}
\end{aligned}$$

where the last line follows from the fact that $(S+D)^{-}1 \succeq m^{-1}I$. Since the norm of the gradient is bounded, $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ must be Lipschitz

$$\left\|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}')\right\|_2 \leq \frac{LJ^{3/2}\sqrt{2C_{\theta^*,\Lambda}}}{m\lambda_{min}}\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2 \tag{73}$$

Finally the result (32) clearly follows since $g_j(\boldsymbol{\theta})$ are Lipschitz in $\boldsymbol{\theta}$ with respect to $\|\cdot\|_\infty$. □

Proof of Lemma 2

Proof of Lemma 3

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.

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.

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

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.

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.