Oracle inequalities for tuning hyper-parameters 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

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

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 $\mathrm{E}\left[\epsilon\right] = 0$, 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

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

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 and the true model:

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

$\tilde{\boldsymbol{\lambda}}$ 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 et al. 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) (and others): what is the "amount of overfitting... when too many hyper-parameters are optimized"? To do this, we establish finite-sample oracle inequalities of the form

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

for some norm $\| \cdot \|$ and constant $a \geq 0$; where $\delta(J, n)$ 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 parameters 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 did find any theoretical results addressing the relationship between the number of hyper-parameters and generalization error of the selected model. Most oracle inequalities only consider tuning a one-dimensional hyper-parameter over a finite Λ (Van Der Laan & Dudoit 2003, van der Laan et al. 2004, Györfi et al. 2006). Others have addressed split-sample validation for specific penalized regression problems, but they solely consider problems with 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 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. 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 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 (from Sections 2.1, 2.2) show

4

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 in such problems without asymptotically increasing the generalization error from the oracle 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

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

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 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 et al. 2012, Van Der Laan & Dudoit 2003). Note that the result below is a special case of Theorem 3 in Section 6.1, 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

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

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

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

5

we have

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

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*

$$\left\|g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 \leq \left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2 \tag{8}$$
$$+ O_p\left(\frac{J\log(nC_\Lambda\lambda_{\max})}{n_V}\right) \tag{9}$$
$$+ O_p\left(\sqrt{\frac{J\log(nC_\Lambda\lambda_{\max})}{n_V}\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2}\right) \tag{10}$$

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.

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 small increases in the model error/degree of overfitting. In fact, if the oracle error rate is $O_p(n_T^{-\omega})$, the number of hyper-parameters J can grow at the rate

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

6

without affecting the asymptotic convergence rate. Note that for parametric regression problems, this will not be the case.

The appearance of the parametric term (9) 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

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

$\mathcal{G}(T)$ 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. Now, we give our result with respect to the generalization error

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

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

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

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

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

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 4 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* $a > 0$,

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

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

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 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 when we are interested in characterizing the generalization error instead.

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: a) our rates depend only quite weakly on $t_{\min}$ and $t_{\max}$; and 2) 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 will 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 3.1.1, 3.1.2, and 3.2. That is, there exist constants $C, \kappa \geq 0$ such that

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

In our examples, the Lipschitz constant is inversely proportional to $\lambda_{\min}$, so κ grows linearly

in $t_{\min}$. Assuming that (19) and (20) hold, we get the following oracle inequalities when tuning penalty parameters via the 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* $C, \kappa > 0$.

1. *Single 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* $a > 0$,

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

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

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 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 those 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

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

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 exactly Lipschitz, one would need the more general oracle inequalities from Theorems 3 and 4 instead).

3.1 Parametric additive models

Parametric additive models have the form

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

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

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

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

$$\boldsymbol{\theta}^* = \arg\min_{\theta} E_T \left[\|y - g(\boldsymbol{\theta})\|^2 \right] \tag{29}$$

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*

$$\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}|T) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) \tag{30}$$

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

$$\nabla_\theta^2 L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\theta = \hat{\theta}(\boldsymbol{\lambda}|T)} \succeq mI \tag{31}$$

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

$$C_{\theta^*, \Lambda} = \frac{1}{2} \|y - g(\boldsymbol{\theta}^*)\|_T^2 + \lambda_{max} \sum_{j=1}^{J} P_j(\boldsymbol{\theta}^{(j),*}) \tag{32}$$

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

$$\left\| g\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}|T)\right) - g\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)}|T)\right) \right\|_\infty \leq \frac{L^2 J^2 \sqrt{2C_{\theta^*, \Lambda}}}{m\lambda_{min}} \left\| \boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)} \right\| \tag{33}$$

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

$$\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j(\boldsymbol{\theta}^{(j)}) + \frac{w}{2} \|\boldsymbol{\theta}^{(j)}\|_2^2 \right) \tag{34}$$

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*

$$\Omega^f(\boldsymbol{\theta}) = \left\{ \boldsymbol{\beta} \Big| \lim_{\epsilon \to 0} \frac{f(\boldsymbol{\theta} + \epsilon\boldsymbol{\beta}) - f(\boldsymbol{\theta})}{\epsilon} \ exists \right\} \tag{35}$$

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*

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^p} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) = \arg\min_{\boldsymbol{\theta} \in S} f(\boldsymbol{\theta}, \boldsymbol{\lambda}) \quad \forall \boldsymbol{\lambda} \in W \tag{36}$$

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 $\Omega^{L_T(\cdot,\cdot)}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T), \boldsymbol{\lambda}\right)$.*

- *$\Omega^{L_T(\cdot,\boldsymbol{\lambda})}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)\right)$ is a local optimality space for $L_T(\cdot, \boldsymbol{\lambda})$ 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.

As an example, the solution to the lasso follows a piecewise linear path, so Λ_{smooth}^C contains a countable number of points. 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 (30).*
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 $L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ 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 exists $m > 0$ such that the Hessian of the penalized training criterion with respect to the differentiable space at the minimizer satisfies

$$_{U_\lambda}\nabla_\theta^2 L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq mI \tag{37}$$

Suppose $\Lambda_{smooth} \subseteq \Lambda := [\lambda_{\min}, \lambda_{\max}]^J$ satisfies Conditions 1 and 2.
Then any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$ satisfies (33).

3.2 Nonparametric additive models

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

$$
\{\hat{g}_j(\boldsymbol{\lambda})\}_{j=1}^J \;\; = \;\; \underset{g_j \in \mathcal{G}_j : j=1,\ldots,J}{\arg\min} \; L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right) \tag{38}
$$

$$
\text{where} \quad L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right) \;\; = \;\; \frac{1}{2}\left\| \boldsymbol{y} - \sum_{j=1}^J g_j(\boldsymbol{x}_j) \right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j) \tag{39}
$$

where P_j are now penalty functionals. For some problems, eg. smoothing splines, we can show that the solution to (38) 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 $\{g_j^*\}_{j=1}^J$ be the minimizer of the generalization error

$$
\{g_j^*\}_{j=1}^J = \underset{g_j \in \mathcal{G}_j : j=1,\ldots,J}{\arg\min} \; E_T\left[\left\| y - \sum_{j=1}^J g_j^* \right\|_T^2 \right] \tag{40}
$$

Lemma 3. *Suppose $\mathcal{G}_1, \ldots, \mathcal{G}_J$ be convex 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 $j = 1, \ldots, J$, the training criterion has a second Gateaux derivative with respect to g_j at $\hat{g}_j^{(n_T)}(\boldsymbol{\lambda}|T)$ satisfies

$$
\left\langle D_{g_j}^2 L_T\left(\{g_j\}_{j=1}^J, \boldsymbol{\lambda}\right)\bigg|_{g_j=\hat{g}_j(\boldsymbol{\lambda}|T)} \circ h_j, h_j \right\rangle \geq m \quad \forall h_j \in \mathcal{G}_j, \|h_j\|_{D^{(n)}} = 1 \tag{41}
$$

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

$$
C_\Lambda^* = \frac{1}{2}\left\| y - \sum_{j=1}^J g_j^* \right\|_T^2 + \lambda_{max}\sum_{j=1}^J P_j(g_j^*) \tag{42}
$$

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

$$
\left\| \sum_{j=1}^J \hat{g}_j\left(\boldsymbol{\lambda}^{(1)}|T\right) - \hat{g}_j\left(\boldsymbol{\lambda}^{(2)}|T\right) \right\|_{D^{(n)}} \leq \frac{J}{m\lambda_{min}}\sqrt{2C_\Lambda^*\frac{n}{n_T}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)}\left\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\right\| \tag{43}
$$

4 Simulations

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

Observations were generated from the model

$$y = \exp(x_1) + x_2^2 + \sigma\epsilon \tag{44}$$

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 an additive smoothing spline model using the Sobolev penalty (De Boor et al. 1978, Wahba 1990, Green & Silverman 1994). The fitted models were of the form

$$\left\{\hat{g}_j^{(n_T)}(\boldsymbol{\lambda}|T)\right\}_{j=1}^2 = \underset{g_1,g_2}{\arg\min} \|y - g_1(x_1) - g_2(x_2)\|_T^2 + \lambda_1 \int_{-1}^{1} (g_1^{(2)}(x))^2 dx + \lambda_2 \int_{-1}^{1} (g_2^{(2)}(x))^2 dx \tag{45}$$

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 $n_V = \lfloor 20 * 2^i \rceil$ for equally log-spaced intervals from $i = 0$ to $i = 7$. A total of 20 simulations were run for each validation set size.

Figure 1 plots the difference of between the model loss and the oracle loss

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

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 the bound given 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 considered model-estimation procedures that construct models with hyperparameters tuned via split-sample validation. Our goal was to address the following open

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

Exponent: −1.012 +/− 0.08284
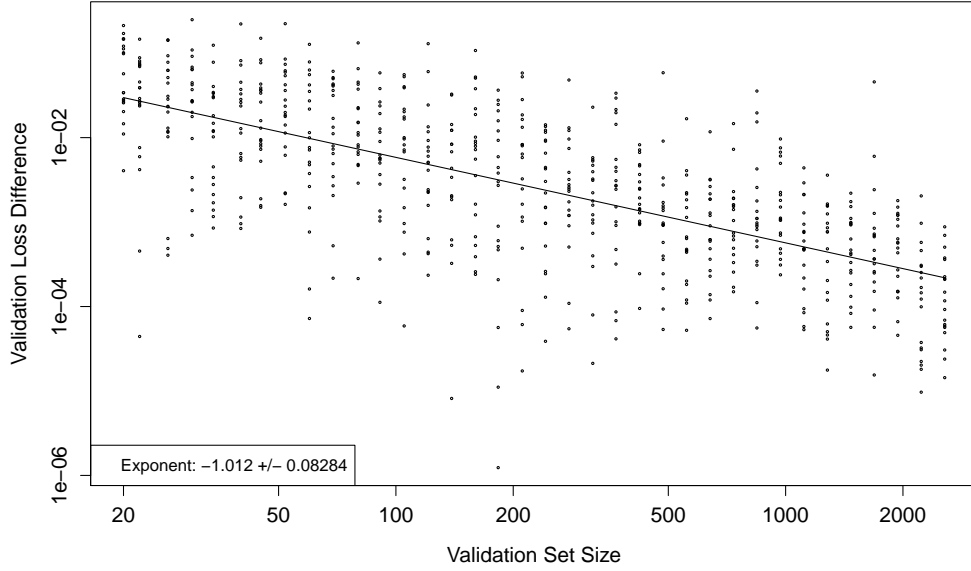
question: How does the generalization error of a model grow with the number of hyper-parameters to be estimated. To do so, we established finite-sample oracle inequalities for selection based on a single training/validation split and based on K-fold cross-validation. We showed that if the model-estimation 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. For example, Feng and Simon (TBD) fit an un-pooled version of sparse group lasso model with a hundred penalty parameters

16

and significantly decreased the model's generalization error.

We note that our results only assume that the selected hyper-parameter has a smaller validation loss compared to the oracle error. In particular, the oracle inequalities do not require the selected hyper-parameter to be the global minimizer. These oracle inequalities may explain why some have noted that for many model-estimation procedures, local minimizers seem to perform fairly well (Kunapuli et al. 2008).

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.

6   Proofs

This section presents proofs to the previous theorems and lemmas. In addition, it presents generalizations of Theorems 1 and 2.

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   A single 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 "basic inequalities", since they are derived directly from the definition. In this "basic inequality" we see that 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*

$$\left\| g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\|_V^2 - \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V^2 \leq \left\langle \epsilon, \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\rangle_V \qquad (46)$$

Proof. By definition,

$$\left\| y - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\|_V^2 \leq \left\| y - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V^2 \qquad (47)$$

$\square$

We are therefore interested in bounding the empirical process term in (46). A common approach is to use a measure of complexity of the function class. For a single training/validation split, where we treat the training set as fixed, we only need to consider the complexity of the fitted models from the model-selection procedure

$$\mathcal{G}(T) = \left\{ \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) : \boldsymbol{\lambda} \in \Lambda \right\} \tag{48}$$

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:*

$$H(u, \mathcal{F}, \|\cdot\|) = \log N(u, \mathcal{F}, \|\cdot\|) \tag{49}$$

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. *Suppose we have dataset $D^{(m)}$ where $\epsilon_1, ..., \epsilon_m$ are independent random variables with mean zero and uniformly sub-gaussian with parameter $b > 0$. Suppose the model class $\mathcal{F}$ has elements $\sup_{f \in \mathcal{F}} \|f\|_{D^{(m)}} \le R$ and satisfies*

$$\psi(R) \ge \int_0^R H^{1/2}(u, \mathcal{F}, \|\cdot\|_{D^{(m)}}) du$$

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

$$\sqrt{m}\delta \ge a(\psi(R) \vee R)$$

we have

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

We are now ready to prove the oracle inequality. It uses a standard peeling argument.

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*

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du \leq \psi(R) \quad \forall R > r \tag{50}$$

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

$$\sqrt{n_V}\delta^2 \geq c\left(\psi(\delta) \vee \delta \vee \psi\left(4\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2\right) \vee 4\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2\right) \tag{51}$$

we have

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

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

Proof. We will use the simplified notation $\hat{g}(\hat{\boldsymbol{\lambda}}) := \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)$ and $\hat{g}(\tilde{\boldsymbol{\lambda}}) := \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)$. In addition, the following probabilities are all conditional on X_V and T but we leave them out for readability.

$$Pr\left(\left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \geq \delta^2\right) \tag{52}$$

$$= \sum_{s=0}^{\infty} Pr\left(2^{2s}\delta^2 \leq \left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \leq 2^{2s+2}\delta^2\right) \tag{53}$$

$$\leq Pr\left(2^{2s}\delta^2 \leq 2\left\langle\epsilon, \hat{g}(\hat{\boldsymbol{\lambda}}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V\right. \tag{54}$$

$$\left.\wedge \left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V^2 \leq 2^{2s+2}\delta^2 + 2\left|\left\langle\hat{g}(\tilde{\boldsymbol{\lambda}}) - \hat{g}(\hat{\boldsymbol{\lambda}}), \hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\rangle_V\right|\right) \tag{55}$$

where we applied the basic inequality in the last inequality. Each summand in (54) can be bounded by splitting the event into the cases where either $2^{2s+2}\delta^2$ 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 larger. Splitting up the probability and applying Cauchy Schwarz gives us the following

bound for (52)

$$Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V\leq 4\left\|\hat{g}(\tilde{\boldsymbol{\lambda}})-g^*\right\|_V} 2\left\langle\epsilon,\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V\geq\delta^2\right)\tag{56}$$

$$+\sum_{s=0}^{\infty}Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V\leq 2^{s+3/2}\delta} 2\left\langle\epsilon,\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V\geq 2^{2s}\delta^2\right)\tag{57}$$

We can bound both (56) and (57) using Lemma 5. For our choice of δ in (51), there is some constant $a > 0$ dependent only on b such that (56) is bounded above by

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

In addition, our choice of δ from (51) 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

$$Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V\leq 2^{s+3/2}\delta} 2\left\langle\epsilon,\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V\geq 2^{2s}\delta^2\right)\leq a\exp\left(-n_V\frac{2^{4s-2}\delta^4}{4a^22^{2s+3}\delta^2}\right)$$

Putting this all together, there is a constant c such that (52) is bounded above by

$$c\exp\left(-\frac{n_V\delta^4}{4c^2\left(16\left\|\hat{g}(\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2\right)}\right)+c\exp\left(-\frac{n_V\delta^2}{c^2}\right)\tag{58}$$

$\square$

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

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*

$$H\left(u,\mathcal{G}(T),\|\cdot\|\right)\leq J\log\left(\frac{4C\left(\lambda_{max}-\lambda_{min}\right)+2u}{u}\right)\tag{59}$$

20

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

$$N\left(u, \Lambda, \|\cdot\|_2\right) \leq \left(\frac{4\left(\lambda_{max} - \lambda_{min}\right) + 2u}{u}\right)^J \tag{60}$$

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

$$
\begin{aligned}
N\left(u, \mathcal{G}(T), \|\cdot\|_V\right) &\leq N\left(\frac{u}{C}, \Lambda, \|\cdot\|_2\right) \tag{61} \\
&\leq \left(\frac{4\left(\lambda_{max} - \lambda_{min}\right) + 2u/C}{u/C}\right)^J \tag{62}
\end{aligned}
$$

$\square$

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

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi(R) := 2R\left(J\log(C_\Lambda\lambda_{\max}n)\right)^{1/2} \tag{63}$$

Applying Theorem 3, we get our desired result. $\square$

6.2 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*

$$\psi(d) \geq \int_0^{\sqrt{d}} H^{1/2}(u, \mathcal{Q}_d, \|\cdot\|_2)du + \frac{\log n_V}{\sqrt{n_V}} \int_0^{2G} H(u, \mathcal{Q}_d, \|\cdot\|_\phi)du \quad \forall d > d_{\min} \tag{64}$$

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*

$$E_{D^{(n)}} \left\| \bar{g}(\hat{\boldsymbol{\lambda}}|D^{(n)}) - g^* \right\|^2 \leq (1+a) \inf_{\boldsymbol{\lambda} \in \Lambda} \left[E_{D^{(n_T)}} \left\| \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - g^* \right\|^2 \right] \tag{65}$$

$$+ \frac{ac}{q} \left[\psi^* \left(\frac{2q^r(1+a)}{a\sqrt{n_V}} \right) \vee d_{\min} \right] \tag{66}$$

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

Theorem 2

Proof. First we determine ψ in (64). 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

$$H(u, \mathcal{Q}_d(T), \|\cdot\|_\infty) \leq J \log \left[\frac{16GC\lambda_{\max} + 2u}{u} \right] \tag{67}$$

Hence we can let

$$\psi(d) := \sqrt{d}K_{n,1} + \frac{K_{n,2}}{\sqrt{n_V}} \tag{68}$$

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 (64) for all $d > n_V^{-1}$.

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

$$\psi^*(z) = \frac{z^2 K_{n,1}^2}{4} + \frac{zK_{n,2}}{\sqrt{n_V}} \tag{69}$$

Plugging in (69) into Theorem 4 gives us the result in Theorem 2. $\qquad\square$

6.3   Penalized regression for additive models

We now show that penalized regression problems for additive models satisfies the Lipschitz condition.

Lemma 1

Proof. We will use the notation $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \coloneqq \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)$. By the gradient optimality conditions, we have for all $j = 1 : J$

$$\nabla_{\theta^{(j)}} \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right] \Bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\lambda)} = 0 \tag{70}$$

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

$$\nabla_\lambda \left\{ \nabla_{\theta^{(j)}} \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right] \Bigg|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\} = 0 \tag{71}$$

Define the following matrices

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

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

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

From the product rule and chain rule, we can then write the system of equations in (71) as

$$\left(\nabla_\lambda \hat{\boldsymbol{\theta}}_1(\boldsymbol{\lambda}) \quad \nabla_\lambda \hat{\boldsymbol{\theta}}_2(\boldsymbol{\lambda}) \quad ... \quad \nabla_\lambda \hat{\boldsymbol{\theta}}_p(\boldsymbol{\lambda}) \right) = -M^\top (S + D)^{-1} \tag{72}$$

We now bound each column in M. From the (70) and Cauchy Schwarz, we have

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

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

$$\frac{1}{2} \left\| y - g(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right\|_T^2 + \sum_{j=1}^J \lambda_j P_j \left(\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}) \right) \leq \frac{1}{2} \|y - g(\boldsymbol{\theta}^*)\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta}^{(j),*}) \tag{73}$$

$$\leq C_{\theta^*, \Lambda} \tag{74}$$

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

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

Now we bound the norm of $\nabla_\lambda \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$. From (72), 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

$$\left\| \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}') \right\|_2 \leq \frac{L J^{3/2} \sqrt{2C_{\theta^*,\Lambda}}}{m \lambda_{min}} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2 \tag{75}$$

Finally the result (33) clearly follows since $g_j(\boldsymbol{\theta})$ are Lipschitz in $\boldsymbol{\theta}$ with respect to $\| \cdot \|_\infty$. $\quad \square$

Lemma 2

Proof. Consider any $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda_{smooth}$.

From Condition 1, every point $\boldsymbol{\lambda} \in \Lambda_{smooth}$ is the center of a ball $B(\boldsymbol{\lambda})$ with nonzero radius where the differentiable space within $B(\boldsymbol{\lambda})$ is constant. Hence by Condition 2, it can be shown that there must exist a countable set of points $\bar{\boldsymbol{\ell}} \subset \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ where $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \bar{\boldsymbol{\ell}}$ such that the union of their differentiable neighborhoods cover $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ entirely:

$$\mathcal{L}\left(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \right) \subseteq \cup_{\ell \in \bar{\boldsymbol{\ell}}} B(\boldsymbol{\ell})$$

Consider the intersections of boundaries of the differentiable neighborhoods with the line segment:

$$P = \left(\cup_{\ell \in \bar{\boldsymbol{\ell}}} \mathrm{Bd} B\left(\boldsymbol{\ell} \right) \right) \cap \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$$

Every point $p \in P$ can be expressed as $\alpha_p \boldsymbol{\lambda^{(1)}} + (1 - \alpha_p) \boldsymbol{\lambda^{(2)}}$ for some $\alpha_p \in [0, 1]$. So points in P can ordered by increasing α_p to get the sequence $\boldsymbol{p}^{(1)}, \boldsymbol{p}^{(2)}, ...$ By Condition 1, the differentiable space of the training criterion is also constant over $\mathcal{L}\left(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)} \right)$ since each of these sub-segments are a subset of $B(\boldsymbol{\ell})$ for some $\boldsymbol{\ell} \in \bar{\boldsymbol{\ell}}$.

Let the differentiable space over the interior of line segment $\mathcal{L}\left(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)}\right)$ be denoted Ω_i. By Condition 1, the differentiable space is also a local optimality space. Let $U^{(i)}$ be an orthonormal basis of Ω_i. For each i, we can express $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)$ for all $\boldsymbol{\lambda} \in \operatorname{Int}\left\{\mathcal{L}\left(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)}\right)\right\}$ as

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T) = U^{(i)}\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}|T)$$

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}|T) = \arg\min_{\beta} L_T(U^{(i)}\boldsymbol{\beta}, \boldsymbol{\lambda})$$

Apply Lemma 1 with $\Lambda = \mathcal{L}\left(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)}\right)$, and modify the proofs to take directional derivatives along the columns of $U^{(i)}$ instead. Then there is a constant constant $c > 0$ independent of i such that for all $i = 1, 2...$, such that

$$\left\|\hat{\boldsymbol{\beta}}(\boldsymbol{p}^{(i)}|T) - \hat{\boldsymbol{\beta}}(\boldsymbol{p}^{(i)}|T)\right\|_2 \leq c\|\boldsymbol{p}^{(i)} - \boldsymbol{p}^{(i+1)}\|_2$$

Finally, we can sum these inequalities. By the triangle inequality,

$$\begin{aligned}
\left\|\hat{\theta}(\boldsymbol{\lambda}^{(1)}|T) - \hat{\theta}(\boldsymbol{\lambda}^{(2)}|T)\right\|_2 &\leq \sum_{i=1}^{\infty} \|\hat{\boldsymbol{\theta}}_{p^{(i)}} - \hat{\boldsymbol{\theta}}_{p^{(i+1)}}\|_2 \\
&\leq \sum_{i=1}^{\infty} c\|\boldsymbol{p}^{(i)} - \boldsymbol{p}^{(i+1)}\|_2 \\
&= c\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|_2
\end{aligned}$$

$\square$

Lemma 3

Proof. Let $H_0 = \left\{j : \left\|\hat{g}_j(\boldsymbol{\lambda}^{(2)}|T) - \hat{g}_j(\boldsymbol{\lambda}^{(1)}|T)\right\|_{D^{(n)}} \neq 0 \;\forall j = 1, ..., J\right\}$. For all $j \in H_0$, let

$$h_j = \frac{\hat{g}_j(\boldsymbol{\lambda}^{(2)}|T) - \hat{g}_j(\boldsymbol{\lambda}^{(1)}|T)}{\left\|\hat{g}_j(\boldsymbol{\lambda}^{(2)}|T) - \hat{g}_j(\boldsymbol{\lambda}^{(1)}|T)\right\|_{D^{(n)}}}$$

For notational convenience, let $\hat{g}_{1,j} = \hat{g}_j(\boldsymbol{\lambda}^{(1)}|T)$. Consider the optimization problem

$$\hat{\boldsymbol{m}}(\boldsymbol{\lambda}) = \{\hat{m}_j(\boldsymbol{\lambda})\}_{j \in H_0} = \arg\min_{m_j \in \mathbb{R}: j \in H_0} \frac{1}{2}\left\|y - \sum_{j=1}^{J}(\hat{g}_{1,j} + m_j h_j)\right\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j\left(\hat{g}_{1,j} + m_j h_j\right) \quad (76)$$

By the gradient optimality conditions, we have

$$\nabla_m \left[\frac{1}{2}\|y - \sum_{j=1}^{J}(\hat{g}_{1,j} + m_j h_j)\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\hat{g}_{1,j} + m_j h_j)\right]\Bigg|_{m=\hat{m}(\lambda)} = 0 \quad (77)$$

25

Implicit differentiation with respect to $\boldsymbol{\lambda}$ gives us

$$\nabla_\lambda \nabla_m \left[\frac{1}{2} \|y - \sum_{j=1}^J (\hat{g}_{1,j} + m_j h_j)\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\hat{g}_{1,j} + m_j h_j) \right] \Bigg|_{m=\hat{m}(\lambda)} = 0 \qquad (78)$$

Define the following matrices

$$S : S_{ij} = \langle h_j, h_j \rangle_T$$

$$D_1 = diag \left(\left\{ \lambda_j \frac{\partial^2}{\partial m_j^2} P_j(\hat{g}_{1,j} + m_j h_j) \Big|_{m=\hat{m}(\lambda)} \right\}_{j=1}^J \right)$$

$$D_2 = diag \left(\left\{ \frac{\partial}{\partial m_j} P_j(\hat{g}_{1,j} + m_j h_j) \Big|_{m=\hat{m}(\lambda)} \right\}_{j=1}^J \right)$$

$$M : \text{ column } M_j = \nabla_\lambda \hat{m}_j(\lambda) \quad \forall j = 1, ..., J$$

The system of equations from (78) can be written as $M = D_2 (S + D_1)^{-1}$.

Now we bound each element in D_2. From (77) and Cauchy Schwarz, we have for all $k = 1, ..., J$

$$\left| \frac{\partial}{\partial m_k} P_k(\hat{g}_{1,k} + m_k h_k) \right|_{m=\hat{m}(\lambda)} \leq \frac{1}{\lambda_{min}} \left\| y - \sum_{j=1}^J (\hat{g}_{1,j} + \hat{m}_j(\boldsymbol{\lambda}) h_j) \right\|_T \|h_k\|_T \qquad (79)$$

We note that $\|h_k\|_T \leq \sqrt{\frac{n_D}{n_T}}$. Also, by definition of $\hat{m}(\boldsymbol{\lambda})$,

$$\frac{1}{2} \left\| y - \sum_{j=1}^J (\hat{g}_{1,j} + \hat{m}_j(\boldsymbol{\lambda}) h_j) \right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\hat{g}_{1,j}) \leq \frac{1}{2} \left\| y - \sum_{j=1}^J \hat{g}_{1,j} \right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\hat{g}_{1,j})$$

$$\leq C_\Lambda^* + J\lambda_{max} \max_{j=1:J} P_j(\hat{g}_{1,j})$$

Hence every diagonal element in D_2 is bounded above by

$$\frac{1}{\lambda_{min}} \sqrt{2 C_\Lambda^* \frac{n_D}{n_T} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}} \right)} \qquad (80)$$

Now we can bound the norm of the gradient $\hat{\boldsymbol{m}}_k(\boldsymbol{\lambda})$ for $k \in H_0$

$$\|\nabla_\lambda \hat{m}_k(\boldsymbol{\lambda})\| = \left\| D_2 (S + D_1)^{-1} e_k \right\| \qquad (81)$$

$$\leq \frac{1}{\lambda_{min}} \sqrt{2 C_\Lambda^* \frac{n_D}{n_T} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}} \right)} \left\| (S + D_1)^{-1} e_k \right\| \qquad (82)$$

$$\qquad (83)$$

By the assumption in (41), $e_k^\top (S + D_1) e_k \geq m$.

By the mean value theorem and Cauchy Schwarz, we have

$$\left| \hat{m}_j(\boldsymbol{\lambda}^{(2)}) - \hat{m}_j(\boldsymbol{\lambda}^{(1)}) \right| \leq \left\| \boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)} \right\| \frac{1}{d\lambda_{min}} \sqrt{2 C_\Lambda^* \frac{n_D}{n_T} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}} \right)} \qquad (84)$$

Since $\left| \hat{m}_j(\boldsymbol{\lambda}^{(2)}) - \hat{m}_j(\boldsymbol{\lambda}^{(1)}) \right| = \left\| \hat{g}_j(\boldsymbol{\lambda}^{(2)}|T) - \hat{g}_j(\boldsymbol{\lambda}^{(1)}|T) \right\|_{D^{(n)}}$, the result in (43) follows.

$\square$

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.

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. 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.