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

Jean Feng and Noah Simon

Department of Biostatistics, University of Washington

Supplementary Material

S1 Appendix

We will use the following notation: for functions f and g and a dataset D with m samples, we denote the inner product of f and g at covariates D as $\langle f, g \rangle_D = \frac{1}{m}\sum_{(x_i,y_i)\in D} f(x_i, y_i)g(x_i, y_i)$.

S1.1 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 so-called "basic inequality" below.

Lemma 4. *For any* $\tilde{\boldsymbol{\lambda}} \in \tilde{\Lambda}$, *we have*

$$\left\| g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\|_V^2 - \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V^2 \le 2 \left\langle \epsilon, \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\rangle_V \quad \text{(S1.1)}$$

Proof. The desired result can be attained by rearranging the definition of $\hat{\boldsymbol{\lambda}}$

$$\left\| y - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\|_V^2 \le \min_{\tilde{\boldsymbol{\lambda}}\in\tilde{\Lambda}} \left\| y - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V^2 . \quad \text{(S1.2)}$$

□

We are therefore interested in bounding the empirical process term in (S1.1). 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

$$\mathcal{G}(T) = \left\{ \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T) : \boldsymbol{\lambda} \in \Lambda \right\}. \tag{S1.3}$$

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:

$$H(u, \mathcal{F}, \|\cdot\|) = \log N(u, \mathcal{F}, \|\cdot\|). \tag{S1.4}$$

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 $D^{(m)} = \{x_1, ..., x_m\}$ are fixed and $\epsilon_1, ..., \epsilon_m$ are independent random variables with mean zero and uniformly sub-gaussian with parameters b and B. Suppose the model class $\mathcal{F}$ satisfies $\sup_{f\in\mathcal{F}} \|f\|_{D^{(m)}} \le R$ and*

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

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

$$\sqrt{m}\delta \geq a(\mathcal{J}(R) \vee R),$$

we have

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

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

Theorem 3. *Consider a set of hyper-parameters Λ. Let training data T be fixed, as well as the covariates of the validation set X_V. Let the oracle risk be denoted*

$$\tilde{R}(X_V|T) = \arg\min_{\lambda\in\Lambda}\left\|g^* - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|T)\right\|_V^2. \tag{S1.5}$$

Suppose independent random variables ϵ_i for validation set V have expectation zero and are uniformly sub-Gaussian with parameter b and B. Suppose there is a function $\mathcal{J}(\cdot|T) : \mathbb{R} \mapsto \mathbb{R}$ and constant $r > 0$ such that

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \mathcal{J}(R|T) \quad \forall R > r \tag{S1.6}$$

Also, suppose $\mathcal{J}(u|T)/u^2$ is non-increasing in u for all $u > r$.

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

$$\sqrt{n_V}\delta^2 \geq c\left(\mathcal{J}(\delta|T) \vee \delta \vee \mathcal{J}\left(\tilde{R}(X_V|T)\middle|T\right) \vee 4\tilde{R}(X_V|T)\right), \tag{S1.7}$$

we have

$$Pr\left(\left\|g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T)\right\|_V^2 - \tilde{R}(X_V|T) \ge \delta^2 \middle| T, X_V\right) \le c\exp\left(-\frac{n_V\delta^4}{c^2\tilde{R}(X_V|T)}\right) + c\exp\left(-\frac{n_V\delta^2}{c^2}\right). \tag{S1.8}$$

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

$$\begin{aligned}
&\Pr\left(\left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \tilde{R}(X_V|T) \ge \delta^2\right) && \text{(S1.9)}\\
&= \sum_{s=0}^{\infty} \Pr\left(2^{2s}\delta^2 \le \left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \tilde{R}(X_V|T) \le 2^{2s+2}\delta^2\right) && \text{(S1.10)}\\
&\le \sum_{s=0}^{\infty} \Pr\left(2^{2s}\delta^2 \le 2\left\langle \epsilon, \hat{g}(\hat{\boldsymbol{\lambda}}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V\right. && \text{(S1.11)}\\
&\qquad \left.\wedge \left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V^2 \le 2^{2s+2}\delta^2 + 2\left|\left\langle \hat{g}(\tilde{\boldsymbol{\lambda}}) - \hat{g}(\hat{\boldsymbol{\lambda}}), \hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\rangle_V\right|\right), && \text{(S1.12)}
\end{aligned}$$

where we applied the basic inequality (S1.1) in the last line. Each summand in (S1.11) 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 (S1.9)

$$\begin{aligned}
&Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda: \left\|\hat{g}(\boldsymbol{\lambda}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V \le 4\left\|\hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V} 2\left\langle \epsilon, \hat{g}(\boldsymbol{\lambda}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V \ge \delta^2\right) && \text{(S1.13)}\\
&+ \sum_{s=0}^{\infty} Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda: \left\|\hat{g}(\boldsymbol{\lambda}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V \le 2^{s+3/2}\delta} 2\left\langle \epsilon, \hat{g}(\boldsymbol{\lambda}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V \ge 2^{2s}\delta^2\right). && \text{(S1.14)}
\end{aligned}$$

We can bound both (S1.13) and (S1.14) using Lemma 5. For our choice of δ in (S1.7), there is some constant $a > 0$ dependent only on b such that (S1.13) 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 (S1.7) 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

$$Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V \le 2^{s+3/2}\delta} 2\left\langle \epsilon, \hat{g}(\boldsymbol{\lambda}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V \ge 2^{2s}\delta^2\right) \le a \exp\left(-n_V \frac{2^{4s-2}\delta^4}{4a^2 2^{2s+3}\delta^2}\right). \tag{S1.15}$$

Putting this all together, we have that there is a constant c such that (S1.9) is bounded above by

$$c \exp\left(-\frac{n_V \delta^4}{c^2 \tilde{R}(X_V|T)}\right) + c \exp\left(-\frac{n_V \delta^2}{c^2}\right). \tag{S1.16}$$

□

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 $\Lambda = [\lambda_{\min}, \lambda_{\max}]^J$ where $\lambda_{\min} \le \lambda_{\max}$. Suppose $\mathcal{G}(T)$ is Lipschitz with function $C(\cdot|T)$ over $\boldsymbol{\lambda}$. Then the entropy of $\mathcal{G}(T)$ with respect to $\|\cdot\|$ is*

$$H\left(u, \mathcal{G}(T), \|\cdot\|\right) \le J \log\left(\frac{4\|C(\cdot|T)\|\left(\lambda_{max} - \lambda_{min}\right) + 2u}{u}\right). \tag{S1.17}$$

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

$$N\left(u, \Lambda, \|\cdot\|_2\right) \le \left(\frac{4\left(\lambda_{max} - \lambda_{min}\right) + 2u}{u}\right)^J. \tag{S1.18}$$

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

$$\begin{aligned} N\left(u, \mathcal{G}(T), \|\cdot\|\right) &\le N\left(\frac{u}{\|C(\cdot|T)\|}, \Lambda, \|\cdot\|_2\right) && \text{(S1.19)} \\ &\le \left(\frac{4\left(\lambda_{max} - \lambda_{min}\right) + 2u/\|C(\cdot|T)\|}{u/\|C(\cdot|T)\|}\right)^J. && \text{(S1.20)} \end{aligned}$$

$\square$

Proof for Theorem 1

Proof. By Lemma 6, we have

$$\begin{aligned} \int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du &= \int_0^R \left(J \log\left(\frac{4\|C_\Lambda\|_V \Delta_\Lambda + 2u}{u}\right)\right)^{1/2} du && \text{(S1.21)} \\ &\le J^{1/2} \int_0^R \left[\log\left(\frac{4\|C_\Lambda(\cdot|T)\|_V \Delta_\Lambda + 2R}{u}\right)\right]^{1/2} du && \text{(S1.22)} \\ &= J^{1/2} R \int_0^1 \left[\log\left(\frac{4\|C_\Lambda(\cdot|T)\|_V \Delta_\Lambda + 2R}{vR}\right)\right]^{1/2} dv && \text{(S1.23)} \\ &\le J^{1/2} R \int_0^1 \log^{1/2}\left(\frac{4\|C_\Lambda(\cdot|T)\|_V \Delta_\Lambda + 2R}{R}\right) + \log^{1/2}(1/v) dv && \text{(S1.24)} \\ &< J^{1/2} R \left(\log^{1/2}\left(\frac{4\|C_\Lambda(\cdot|T)\|_V \Delta_\Lambda + 2R}{R}\right) + 1\right). && \text{(S1.25)} \end{aligned}$$

If we restrict $R > n^{-1}$, then for an absolute constant c, we have

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du \leq \mathcal{J}(R) \coloneqq cR \left(J \log(\|C_\Lambda(\cdot|T)\|_V \Delta_\Lambda n + 1) \right)^{1/2}. \quad \text{(S1.26)}$$

Applying Theorem 3, we get our desired result. □

S1.2 Cross-validation

In order to obtain an oracle inequality for averaged version of cross-validation, we need to extend Theorem 3.5 in Lecué and Mitchell [2012]. Let the class of fitted functions for given training data T be denoted

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

In Lecué and Mitchell [2012], they assume that there is a function $\mathcal{J}$ that uniformly bounds the size of the class $\mathcal{G}(T)$ for any training data T. However the complexity of $\mathcal{G}(T)$ depends on training data – for instance, if there is a lot of noise in the training data, the size of $\mathcal{G}(T)$ can be very high. In our extension, we allow the function $\mathcal{J}$ to depend on the training data.

Throughout this section, we use Talagrand's gamma function [Talagrand, 2005] to characterize the size of a function class. We present it below as it will be used later on.

Definition 5. For metric space (T, d) and $\alpha \geq 0$, define

$$\gamma_\alpha(T, d) = \inf \sup_{t \in T} \sum_{s=0}^{\infty} 2^{s/\alpha} d(t, T_s)$$

where the infimum is taken over all sequences $\{T_s : s \in \mathbb{N}, T_s \subseteq T, |T_s| \le 2^{2^s}\}$. (Here, $|A|$ denotes the cardinality of the set A.)

We begin with some notation. Suppose we have a measurable space $(\mathcal{Z}, \mathcal{T})$ where we observe $Z = (X, y)$ random variables with values in $\mathcal{Z}$. Let $\mathcal{G}$ is a class of measurable functions from $\mathcal{Z} \mapsto \mathbb{R}$; the model-estimation procedure selects functions from the class $\mathcal{G}$. In contrast to the main manuscript, we will consider a very general setting. In particular, the noise $\epsilon = y - E[y|X = x]$ is not necessarily independent of X. In addition, we consider a general loss function $Q : \mathcal{Z} \times \mathcal{G} \mapsto \mathbb{R}$ (rather than solely the least squares loss). Define the risk function $R(g)$ as the expected loss $\mathbb{E}Q(Z, g)$ and suppose the risk function is convex. Let $\bar{g}^{(n)}(D^{(n)})$ denote the averaged version of cross-validation and g^* denote the minimizer of the risk function over $\mathcal{G}$.

In this more general setting, we require a more general version of Assumption 2:

Assumption 3. There exist constants $K_0, K_1 \ge 0$ and $\kappa \ge 1$ such that for any $m \in \mathbb{N}$ and any dataset $D^{(m)}$,

$$\left\|Q(\cdot, \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - Q(\cdot, g^*)\right\|_{L_{\psi_1}} \le K_0 \tag{S1.27}$$

$$\left\|Q(\cdot, \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - Q(\cdot, g^*)\right\|_{L_2} \le K_1 \left(R(\hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)})) - R(g^*)\right)^{1/2\kappa}. \tag{S1.28}$$

Our theorem relies on the basic inequality established in Lemma 3.1 in Lecué and Mitchell [2012]. We reproduce it here for convenience. From henceforth, $c_i > 0$ denotes absolute constants, that may not necessarily be the same if they share the same subscript.

Lemma 7. *For any constant $a > 0$, we have the following inequality*

$$\mathbb{E}_{D^{(n)}}\left(R(\bar{g}^{(n)}(D^{(n)})) - R(g^*)\right) \le (1+a) \inf_{\lambda \in \Lambda} \left[\mathbb{E}_{D^{(n_V)}} R(\hat{g}^{(n_V)}(\boldsymbol{\lambda}|D^{(n_V)})) - R(g^*)\right] + \mathbb{E}_{D^{(n)}} \sup_{\lambda \in \Lambda} \left[(P - (1+a)P_{n_V})\left(Q(\cdot, \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)})) - Q(\cdot, g^*)\right)\right] \tag{S1.29}$$

where $P_{n_V} = 1/n_V \sum_{i=n_T+1}^{n} \delta_{Z_i}$ is the empirical probability measure on $\{Z_{n_T+1}, ..., Z_n\}$.

We need to bound the supremum of the second term on the right hand side, which is a shifted empirical process term. Lemma 3.4 in Lecué and Mitchell [2012] already bounds the shifted empirical process term. However to extend their result to our purposes, we restate it to clarify the conditional dependencies. This allows us to introduce two new functions h and J_δ that will be used later on.

Lemma 8. *Let $\mathcal{Q}(D^{(m)}) \equiv \left\{Q(\lambda|D^{(m)}) : \lambda \in \Lambda\right\}$ and $\mathcal{Q} \equiv \cup_{m\in\mathbb{N}} \cup_{D^{(m)}} \mathcal{Q}(D^{(m)})$. Suppose there exists $C_1 > 0$ and an increasing function $G(\cdot)$ such that $\forall Q \in \mathcal{Q}$,*

$$\|Q(Z)\|_{L_2} \le G\left(\mathbb{E}Q(Z)\right).$$

Let $n_T, n_V \in \mathbb{N}$. Suppose there exists a function h that maps training data $D^{(n_T)}$ to $\mathbb{R}^+$, a function $J_\delta : \mathbb{R}^+ \mapsto \mathbb{R}^+$ indexed by $\delta > 0$, and a constant $w_{\min} > 0$ such that for any dataset $D^{(n_T)}$ and any $w \ge w_{\min}$,

$$h(D^{(n_T)}) \le \delta \implies \frac{\log n_V}{\sqrt{n_V}} \gamma_1\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \|\cdot\|_{L_{\psi_1}}\right) + \gamma_2\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \|\cdot\|_{L_2}\right) \le J_\delta(w) \tag{S1.30}$$

where $\mathcal{Q}_w^{L_2}(D^{(n_T)}) \equiv \left\{Q \in \mathcal{Q}(D^{(n_T)}) : \|Q(Z)\|_{L_2} \le G(w)\right\}$.

Then there exists absolute constants $L, c > 0$ *such that for all* $w \ge w_{\min}$ *and all* $u \ge 1$,

$$\Pr\left(\sup_{Q \in \mathcal{Q}(D^{(n_T)}): PQ \le w} \left((P - P_{n_V}) Q \right)_+ \le uL \frac{J_\delta(w)}{\sqrt{n_V}} \middle| h\left(D^{(n_T)}\right) \le \delta \right) \ge 1 - L \exp(-cu). \tag{S1.31}$$

Now that we have established a concentration inequality for the function class $\{Q \in \mathcal{Q}(D^{(n_T)}) : PQ \le w\}$, we need to aggregate the results to establish a concentration inequality for the function class $\mathcal{Q}(D^{(n_T)})$. Again, we use Lemma 3.2 in Lecué and Mitchell [2012] but restate it using our new functions h and J_δ.

Lemma 9. *Let* $a > 0$*. Let* $\mathcal{Q}(D^{(m)}) \equiv \left\{Q(\lambda | D^{(m)}) : \lambda \in \Lambda\right\}$ *be a set of measurable functions. For all* $m \in \mathbb{N}$ *and any dataset* $D^{(m)}$*, suppose* $\mathbb{E}Q(Z) \ge 0$ *for all* $Q \in \mathcal{Q}\left(D^{(m)}\right)$.

Suppose for any $n_T, n_V \in \mathbb{N}$ *and dataset* $D^{(n_T)}$ *there exists some absolute constant* $L, c > 0$ *such that for all* $w \ge w_{\min}$ *and for all* $u \ge 1$,

$$\Pr\left(\sup_{Q \in \mathcal{Q}(D^{(n_T)}): PQ \le w} \left((P - P_{n_V}) Q \right)_+ \le uL \frac{J_\delta(w)}{\sqrt{n_V}} \middle| h\left(D^{(n_T)}\right) \le \delta \right) \ge 1 - L \exp(-cu).$$

For any $\delta > 0$*, suppose* J_δ *is strictly increasing and its inverse is strictly convex. Let* ψ_δ *be the convex conjugate of* J_δ^{-1}*, e.g.* $\psi_\delta(u) = \sup_{v>0} uv - J_\delta^{-1}(v)$ *for all* $u > 0$*. Assume there is a* $r \ge 1$ *such that* $x > 0 \mapsto \psi_\delta(x)/x^r$ *decreases. For all* $q > 1$ *and*

$u \geq 1$, define

$$\tilde{\psi}_{q,\delta}(u) = \psi_\delta \left(\frac{2q^{r+1}(1+a)u}{a\sqrt{n_V}} \right) \vee w_{\min}.$$

Then there exists a constant L_1 that only depends on L such that for every $u \geq 1$,

$$\Pr\left(\sup_{Q \in \mathcal{Q}(D^{(n_T)})} \left((P - (1+a)P_{n_V}) Q \right)_+ \leq \frac{a\tilde{\psi}_{q,\delta}(u/q)}{q} \middle| h\left(D^{(n_T)}\right) \leq \delta \right) \geq 1 - L_1 \exp(-cu).$$

Moreover, assume that $\psi_\delta(x)$ is an increasing function in x such that $\psi_\delta(\infty) = \infty$. Then there exists a constant c_1 that depends only on L and c such that

$$\mathbb{E}\left[\sup_{Q \in \mathcal{Q}(D^{(n_T)})} \left((P - (1+a)P_{n_V}) Q \right)_+ \middle| h\left(D^{(n_T)}\right) \leq \delta \right] \leq \frac{ac_1\tilde{\psi}_{q,\delta}(1/q)}{q}. \tag{S1.32}$$

Finally, we are ready to bound the expectation of the shifted empirical process term in (S1.29). We accomplish this via a simple chaining argument; we omit its proof as this is a standard application of the chaining argument.

Lemma 10. *Consider any $a > 0$. Suppose there exists a constant c_1 such that for any $n_T, n_V \in \mathbb{N}$, $\delta > 0$, and $q > 1$, (S1.32) holds. Then for any $\sigma > 0$, we have*

$$\mathbb{E}\left[\sup_{Q \in \mathcal{Q}(D^{(n_T)})} \left((P - (1+a)P_{n_V}) Q \right)_+ \right] \leq \frac{ac_1}{q} \left(\tilde{\psi}_{q,2\sigma}(1/q) + \sum_{k=1}^{\infty} \Pr\left(h\left(D^{(n_T)}\right) \geq 2^k \sigma \right) \tilde{\psi}_{q,2^k\sigma}(1/q) \right).$$

Putting Lemmas 7 and 10 together, we have the following result.

Theorem 4. *Consider a set of hyper-parameters Λ. Consider a loss function $Q : (\mathcal{Z}, \mathcal{G}) \mapsto \mathbb{R}$ with convex risk function $R : \mathcal{G} \mapsto \mathbb{R}$. Let*

$$\mathcal{Q} = \{ Q(\cdot, \hat{g}^{(n_T)}(\boldsymbol{\lambda} | D^{(n_T)}) - Q(\cdot, g^*) : \boldsymbol{\lambda} \in \Lambda \}.$$

Suppose Assumption 3 holds. Suppose there is an $w_{\min} > 0$ *and functions* $h : \mathcal{Z}^{(n_T)} \mapsto \mathbb{R}$ *and* $\mathcal{J}_\delta : \mathbb{R} \mapsto \mathbb{R}$ *such that for all* $w \geq w_{\min}$,

$$h(D^{(n_T)}) \leq \delta \implies \frac{\log n_V}{\sqrt{n_V}} \gamma_1 \left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \| \cdot \|_{L_{\psi_1}} \right) + \gamma_2 \left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \| \cdot \|_{L_2} \right) \leq \mathcal{J}_\delta(w) \tag{S1.33}$$

where $\mathcal{Q}_w = \{Q \in \mathcal{Q} : \|Q\|_{L_2} \leq w^{1/2\kappa}\}$*. Moreover, suppose that for all* $\delta > 0$*,* J_δ *is a strictly increasing function and* $\mathcal{J}_\delta^{-1}(\epsilon)$ *is strictly convex. Let the convex conjugate of* $\mathcal{J}_\delta^{-1}$ *be denoted* ψ_δ*. Suppose* $\psi_\delta(x)$ *increases in* x*,* $\psi_\delta(\infty) = \infty$*, and there exists* $r \geq 1$ *such that* $\psi_\delta(x)/x^r$ *decreases.*

Consider any $\sigma > 0$*. Then there is a constant* $c > 0$ *such that for every* $a > 0$ *and* $q > 1$*, the following inequality holds*

$$\begin{aligned} \mathbb{E}_{D^{(n)}} \left(R\left(\bar{g}(D^{(n)})\right) - R(g^*) \right) \leq (1+a) \inf_{\boldsymbol{\lambda} \in \Lambda} \mathbb{E}_{D^{(n_T)}} \left(R\left(\bar{g}(\hat{\boldsymbol{\lambda}} | D^{(n)}) \right) - R(g^*) \right) \\ + \frac{ac}{q} \left(\tilde{\psi}_{q,2\sigma}(1/q) + \sum_{k=1}^{\infty} \Pr\left(h\left(D^{(n_T)}\right) \geq 2^k \sigma \right) \tilde{\psi}_{q,2^k\sigma}(1/q) \right). \end{aligned} \tag{S1.34}$$

where $\tilde{\psi}_{q,\delta}(u) = \psi_\delta \left(\frac{2q^{r+1}(1+a)u}{a\sqrt{n_V}} \right) \vee w_{\min}$ *for all* $u > 0$*.*

Of course, this theorem is only useful if we can show that $h(D^{(n_T)})$ is bounded with high probability. For instance, in an example in the main manuscript, we show that $h(D^{(n_T)})$ has sub-exponential tails; so the latter term in (S1.34) is well-controlled.

We now apply Theorem 4 to prove Theorem 2. Recall that Theorem 2 concerns the squared error loss $Q((x, y), g) = (y - g(x))^2$ and only considers model-estimation methods where the estimated functions are Lipschitz in the hyper-parameters. First we need the following lemma that describes the relationship between Lipschitz functions

Lemma 11. *Suppose the same conditions as Theorem 4. Suppose Assumptions 1 and 2 hold. Also suppose that* $\|\epsilon\|_{L_{\psi_2}} = b < \infty$. *Define* $\mathcal{Q}_w^{L_2} = \{g^* - \hat{g}(\boldsymbol{\lambda}|D^{(n_T)}) : P(g^* - \hat{g}(\boldsymbol{\lambda}|D^{(n_T)}))^2 < w\}$ *for* $w > 0$. *Then there is an absolute constant* $c_0 > 0$ *such that*

$$N\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), u, \|\cdot\|_{L_2}\right) \le N\left(\Lambda, \frac{u}{c_0\left(b+\sqrt{w}\right)\|C_\Lambda(x|D^{(n_T)})\|_{L_2}}, \|\cdot\|_2\right). \quad \text{(S1.35)}$$

then we also have

$$N\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), u, \|\cdot\|_{L_{\psi_1}}\right) \le N\left(\Lambda, \frac{u}{c_{K_0,b}\|C_\Lambda(x|D^{(n_T)})\|_{L_{\psi_2}}}, \|\cdot\|_2\right) \quad \text{(S1.36)}$$

for a constant $c_{K_0,b} > 0$ *that only depends on* K_0 *and* b.

Proof. Let us first consider a general norm $\|\cdot\|$ such that for any random variables X, Y, we have $\|XY\| \le \|X\|_*\|Y\|_*$. Then for all $\boldsymbol{\lambda} \in \Lambda$ such that $P(g^* - \hat{g}(\boldsymbol{\lambda}|D^{n_T}))^2 \le$

w, we have

$$\left\| Q(\cdot, \hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(1)}|D^{(n_T)})(x)) - Q(\cdot, \hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(2)}|D^{(n_T)})(x)) \right\| \tag{S1.37}$$

$$= \left\| \left(y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(1)}|D^{(n_T)})(x) \right)^2 - \left(y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(2)}|D^{(n_T)})(x) \right)^2 \right\| \tag{S1.38}$$

$$\left(\hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(2)}|D^{(n_T)})(x) - \hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(1)}|D^{(n_T)})(x) \right)^2 \| \tag{S1.39}$$

$$\begin{aligned} &\leq \left\| 2\epsilon + g^*(x) - \hat{g}(\boldsymbol{\lambda}^{(1)}|D^{(n_T)})(x) + g^*(x) - \hat{g}(\boldsymbol{\lambda}^{(2)}|D^{(n_T)})(x) \right\|_* \\ &\quad \times \left\| \hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(2)}|D^{(n_T)})(x) - \hat{g}^{(n_T)}(\boldsymbol{\lambda}^{(1)}|D^{(n_T)})(x) \right\|_* \end{aligned} \tag{S1.40}$$

$$\leq \left(2\|\epsilon\|_* + 2 \sup_{\lambda \in \Lambda : P(g^* - \hat{g}(\boldsymbol{\lambda}|D^{n_T}))^2 \leq w} \left\| g^*(x) - \hat{g}(\boldsymbol{\lambda}^{(1)}|D^{(n_T)})(x) \right\|_* \right) \left\| C_\Lambda(x|D^{(n_T)}) \right\|_* \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2 \tag{S1.41}$$

For $\|\cdot\| = \|\cdot\|_{L_2}$, the L_2 norm is its own dual norm so (S1.41) reduces to

$$c_0 \left(b + \sqrt{w} \right) \|C_\Lambda(x|D^{(n_T)})\|_{L_2} \|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|_2$$

for an absolute constant $c_0 > 0$.

For $\|\cdot\| = \|\cdot\|_{L_{\psi_1}}$, the dual of the L_{ψ_1} norm is L_{ψ_2}. Thus applying Assumption 2 and the fact that $\|\epsilon\|_{L_{\psi_2}} = b < \infty$, (S1.41) reduces to

$$2 \left(b + K_0 \right) \|C_\Lambda(x|D^{(n_T)})\|_{L_{\psi_2}} \|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|_2.$$

$\square$

Talagrand's gamma function of a class T can be bounded by Dudley's integral

$$\gamma_\alpha(T, D) \leq c \int_0^{\mathrm{Diam}(T,d)} \left(\log N(T, \epsilon, d) \right)^{1/\alpha} d\epsilon \tag{S1.42}$$

[Talagrand, 2005]. Combining the above bound with Lemma 11 gives the following lemma.

Lemma 12. *Suppose Assumptions 1 and 2 hold. Suppose* $\|\epsilon\|_{L_{\psi_2}} = b < \infty$. *Define* $\mathcal{Q}_w^{L_2}$ *as before. For* Λ, *let* $\Delta_\Lambda = (\lambda_{\max} - \lambda_{\min}) \vee 1$. *Let* $w > 0$. *Let* $\mathcal{Q}_w^{L_2}(D^{(n_T)})$ *be defined as before.*

Then there exist absolute constants $c_0, c_1 > 0$ *and a constant* $c_{K_0,b} > 0$ *such that*

$$\gamma_2\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \|\cdot\|_{L_2}\right) \leq c_0\sqrt{wJ}\left[\sqrt{\log\left(\left(\frac{b}{\sqrt{w}}+1\right)\Delta_\Lambda\|C_\Lambda(x|D^{(n_T)})\|_{L_2}+1\right)}+1\right] \tag{S1.43}$$

$$\gamma_1\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \|\cdot\|_{L_{\psi_1}}\right) \leq c_1 J K_0\left[\log\left(\Delta_\Lambda\|C_\Lambda(x|D^{(n_T)})\|_{L_{\psi_2}} c_{K_0,b}+1\right)+1\right]. \tag{S1.44}$$

Proof. By definition of $\mathcal{Q}_w^{L_2}$, we have $\mathrm{Diam}\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \|\cdot\|_{L_2}\right) = 2\sqrt{w}$. Using Lemma 11 and (S1.42), we have

$$\gamma_2\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \|\cdot\|_{L_2}\right) \leq c\int_0^{2\sqrt{w}}\sqrt{\log N\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), u, \|\cdot\|_{L_2}\right)}du \tag{S1.45}$$

$$\leq c\int_0^{2\sqrt{w}}\sqrt{\log N\left(\Lambda, \frac{u}{c_0\left(b+\sqrt{w}\right)\|C_\Lambda(x|D^{(n_T)})\|_{L_2}}, \|\cdot\|_2\right)}du \tag{S1.46}$$

$$\leq c\int_0^{2\sqrt{w}}\sqrt{J\log\left(\frac{4c_0\Delta_\Lambda\left(b+\sqrt{w}\right)\|C_\Lambda(x|D^{(n_T)})\|_{L_2}+2u}{u}\right)}du \tag{S1.47}$$

$$\leq 2c\sqrt{wJ}\left[\sqrt{\log\left(\frac{4c_0\Delta_\Lambda\left(b+\sqrt{w}\right)\|C_\Lambda(x|D^{(n_T)})\|_{L_2}+4\sqrt{w}}{2\sqrt{w}}\right)}+\frac{\sqrt{\pi}}{2}\right] \tag{S1.48}$$

Using very similar logic, we now bound the γ_1 function. First we bound the diameter of $\mathcal{Q}_w^{L_2}$ with respect to the norm $\|\cdot\|_{L_{\psi_1}}$:

$$\mathrm{Diam}(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \|\cdot\|_{L_{\psi_1}}) \le 2 \sup_{\boldsymbol{\lambda}\in\Lambda} \left\| \left(y - \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)})\right)^2 - (y - g^*(x))^2 \right\|_{L_{\psi_1}} \le c_1 K_0. \tag{S1.49}$$

Thus

$$\gamma_1\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), \|\cdot\|_{L_{\psi_1}}\right) \le c\int_0^{c_1K_0} \log N\left(\mathcal{Q}_w^{L_2}(D^{(n_T)}), u, \|\cdot\|_{L_{\psi_1}}\right) du \tag{S1.50}$$

$$\le c_2 J K_0 \left[\log\left(\frac{4\Delta_\Lambda c_{K_0,b}\|C_\Lambda(x|D^{(n_T)})\|_{L_{\psi_2}} + 2c_1K_0}{c_1K_0}\right) + 1\right] \tag{S1.51}$$

□

To apply Theorem 4, we need to define h and J_δ so that (S1.33) is satisfied. Based on the lemma above, we see that it suffices to let

$$h(D^{(n_T)}) \coloneqq \|C_\Lambda(x|D^{(n_T)})\|_{L_{\psi_2}} \tag{S1.52}$$

and

$$\mathcal{J}_\delta(w) = c_1\frac{\log n_V}{\sqrt{n_V}} JK_0\left[\log\left(\Delta_\Lambda \delta c_{K_0,b} + 1\right) + 1\right] + c_3\sqrt{Jw}\left[\sqrt{\log\left(\Delta_\Lambda b\delta n + 1\right)} + 1\right]. \tag{S1.53}$$

Finally using the results above, we can prove Theorem 2.

Proof for Theorem 2. We now apply Theorem 4 to our Lipschitz case. From (S1.49), we find that Assumption 3 is satisfied. We have defined h and J_δ so that (S1.33) is satisfied for all $w \ge 1/n$. Moreover, $\mathcal{J}_\delta(w)$ is strictly increasing and concave in

w. This implies that $\mathcal{J}_\delta^{-1}$ is strictly convex. Via algebra, we find that the convex conjugate of $\mathcal{J}_\delta^{-1}$ is

$$\psi_\delta(u) = c_1 u \frac{\log n_V}{\sqrt{n_V}} J K_0 \left[\log\left(\Delta_\Lambda \delta c_{K_0,b} + 1\right) + 1\right] + u^2 c_4 J \left[\sqrt{\log\left(\Delta_\Lambda b \delta n + 1\right)} + 1\right]^2. \tag{S1.54}$$

Now let us determine $\tilde{\psi}_{q,\delta}(1/q)$ as $q \to 1$. We have

$$\lim_{q\to 1} \tilde{\psi}_{q,\delta}(1/q) = \psi_\delta\left(\frac{2(1+a)}{a}\frac{1}{\sqrt{n_V}}\right) \vee \frac{1}{n_V} \tag{S1.55}$$

$$\leq c_5 \left(\frac{1+a}{a}\right)^2 \frac{J \log n_V}{n_V} K_0 \left[\log\left(\Delta_\Lambda \delta c_{K_0,b} n + 1\right) + 1\right]. \tag{S1.56}$$

So the summation in (S1.34) reduces to

$$\lim_{q\to 1}\left(\tilde{\psi}_{q,2\sigma_0}(1/q) + \sum_{k=1}^{\infty} \Pr\left(h\left(D^{(n_T)}\right) \geq 2^k \sigma\right) \tilde{\psi}_{q,2^k\sigma_0}(1/q)\right) \tag{S1.57}$$

$$\leq c_6 \left(\frac{1+a}{a}\right)^2 \frac{J \log n_V}{n_V} K_0 \left[\log\left(\Delta_\Lambda c_{K_0,b} n \sigma_0 + 1\right) + 1\right] \left(1 + \sum_{k=1}^{\infty} k \Pr\left(\|C_\Lambda(x|D^{(n_T)})\|_{L_{\psi_2}} \geq 2^k \sigma_0\right)\right) \tag{S1.58}$$

$$\leq c_6 \left(\frac{1+a}{a}\right)^2 \frac{J \log n_V}{n_V} K_0 \left[\log\left(\Delta_\Lambda c_{K_0,b} n \sigma_0 + 1\right) + 1\right] \tilde{h}(n_T). \tag{S1.59}$$

Taking $q \to 1$ in (S1.34) and plugging in (S1.59) to Theorem 4, we get our desired result. $\square$

S1.3 Penalized regression for additive models

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

Proof for Lemma 1

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

$$\nabla_\theta \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right] \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} = 0. \tag{S1.60}$$

After implicitly differentiating with respect to $\boldsymbol{\lambda}$, we have

$$\nabla_\lambda \left\{ \nabla_\theta \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right] \Bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\} = 0. \tag{S1.61}$$

From the product rule and chain rule, we can then write the system of equations in (S1.61) as

$$\nabla_\lambda \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} = - \left(\nabla_\theta^2 L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}) \big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right)^{-1} \operatorname{diag} \left\{ \nabla_{\theta^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\}_{j=1:J}. \tag{S1.62}$$

We can bound the norm of the second term in (S1.62) by rearranging (S1.60) and using the Cauchy-Schwarz inequality:

$$\left\| \nabla_{\theta^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\|_2 \le \frac{1}{\lambda_{min}} \left\| y - g(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right\|_T \left\| \left\| \nabla_{\theta^{(j)}} g_j(x|\boldsymbol{\theta}^{(j)}) \right\|_2 \right\|_T.$$

Since g_j is Lipschitz by assumption, then

$$\left\| \nabla_{\theta^{(j)}} g_j(x|\boldsymbol{\theta}^{(j)}) \right\|_2 \le \ell_j(x). \tag{S1.63}$$

Also, by the definition of $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$, we have

$$\frac{1}{2} \left\| y - g(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right\|_T^2 \le \frac{1}{2} \|\epsilon\|_T^2 + C_\Lambda^*. \tag{S1.64}$$

Hence

$$\left\| \nabla_\theta P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\|_2 \leq \frac{\|\ell_j\|_T}{\lambda_{min}} \sqrt{\|\epsilon\|_T^2 + 2C_\Lambda^*}. \tag{S1.65}$$

Plugging in the results from above and using the assumption that the Hessian of the objective function has a minimum eigenvalue of $m(T)$, we have for all

$$\nabla_{\lambda_k} \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} = \mathbf{0} \text{ if } j \neq k \tag{S1.66}$$

$$\left\| \nabla_{\lambda_j} \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\|_2 = \left\| \nabla_{\lambda_j} \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\|_2 \tag{S1.67}$$

$$\leq \frac{1}{m(T)} \frac{\|\ell_j\|_T}{\lambda_{min}} \sqrt{\|\epsilon\|_T^2 + 2C_\Lambda^*}. \tag{S1.68}$$

Since the norm of the gradient is bounded, $\hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda})$ must be Lipschitz:

$$\left\| \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}^{(1)}) - \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}^{(2)}) \right\|_2 \leq \frac{1}{m(T)} \frac{\|\ell_j\|_T}{\lambda_{min}} \sqrt{\|\epsilon\|_T^2 + 2C_\Lambda^*} \left| \lambda_j^{(1)} - \lambda_j^{(2)} \right|. \tag{S1.69}$$

Finally we combine the above results to get

$$\left| g\left(x \middle| \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}) \right) - g\left(x \middle| \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)}) \right) \right| \tag{S1.70}$$

$$\leq \sum_{j=1}^J \left| g_j\left(x \middle| \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}) \right) - g_j\left(x \middle| \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)}) \right) \right| \tag{S1.71}$$

$$\leq \sum_{j=1}^J \ell_j(x_j) \left\| \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}^{(1)}) - \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}^{(2)}) \right\|_2 \tag{S1.72}$$

$$\leq \sum_{j=1}^J \ell_j(x_j) \frac{1}{m(T)} \frac{\|\ell_j\|_T}{\lambda_{min}} \sqrt{\|\epsilon\|_T^2 + 2C_\Lambda^*} \left| \lambda_j^{(1)} - \lambda_j^{(2)} \right| \tag{S1.73}$$

$$\leq \frac{1}{m(T)\lambda_{min}} \sqrt{\left(\|\epsilon\|_T^2 + 2C_\Lambda^* \right) \left(\sum_{j=1}^J \|\ell_j\|_T^2 \ell_j^2(x_j) \right)} \left\| \boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)} \right\|_2 \tag{S1.74}$$

□

Proof for Lemma 2

Before proving Lemma 2, we need to introduce some notation. Let $\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})$ be the line segment connecting $\boldsymbol{\lambda}^{(1)}$ and $\boldsymbol{\lambda}^{(2)}$. Let $\mu_1(z)$ be the 1-dimensional Lebesgue measure in the direction of z (so if z is a continuous line segment, $\mu_1(z) = \|z\|_2$; if z is composed of multiple line segments z_i, then $\mu(z) = \sum \mu(z_i)$).

Before proving the Lipschitz property over all of Λ, we show that the fitted function is Lipschitz over Λ_{smooth}. For convenience, define $\Lambda_{smooth}^c \coloneqq \Lambda \setminus \Lambda_{smooth}$.

Lemma 13. *Suppose that $g_j(\boldsymbol{\theta})(x)$ satisfies the Lipschitz condition in Lemma 1. Let $T \equiv D^{(n_T)}$ be a fixed set of training data. Suppose the penalized loss function $L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ has a unique minimizer $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)$ for every $\boldsymbol{\lambda} \in \Lambda$. Let $\boldsymbol{U}_\lambda$ be an orthonormal matrix with columns forming a basis for the differentiable space of $L_T(\cdot, \boldsymbol{\lambda})$ at $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)$. Suppose there exists a constant $m(T) > 0$ such that the Hessian of the penalized training criterion at the minimizer taken with respect to the directions in $\boldsymbol{U}_\lambda$ satisfies*

$$ {}_{U_\lambda}\nabla_\theta^2 L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})\big|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \succeq m(T)\boldsymbol{I} \quad \forall \boldsymbol{\lambda} \in \Lambda \tag{S1.75} $$

where $\boldsymbol{I}$ is the identity matrix. *Suppose Condition 1 is satisfied by some $\Lambda_{smooth} \subseteq \Lambda$. Define*

$$ \Lambda_{ext} = \left\{ (\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) : \boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda, \mu_1\left(\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \cap \Lambda_{smooth}^c \right) > 0 \right\}. \tag{S1.76} $$

Then any $(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \in \Lambda_{ext}^c$ satisfies (4.24).

Proof. 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.

Now consider any $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda_{ext}$. By (S1.76), there must exist a countable set of points $\cup_{i=1}^{\infty} \boldsymbol{\ell}^{(i)} \subset \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ where $\cup_{i=1}^{\infty} \boldsymbol{\ell}^{(i)} \subset \Lambda_{smooth}$, $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \cup_{i=1}^{\infty} \boldsymbol{\ell}^{(i)}$, and the union of their differentiable neighborhoods cover $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ entirely:

$$\mathcal{L}\left(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}\right) \subseteq \cup_{i=1}^{\infty} B\left(\boldsymbol{\ell}^{(i)}\right).$$

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

$$P = \cup_{i=1}^{\infty} \left[bd\left(B\left(\boldsymbol{\ell}^{(i)}\right)\right) \cap \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \right]. \tag{S1.77}$$

Every point $p \in P$ can be expressed as $\alpha_p \boldsymbol{\lambda^{(1)}} + (1-\alpha_p)\boldsymbol{\lambda^{(2)}}$ for some $\alpha_p \in [0,1]$. We can order the points in P 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 constant over $\mathcal{L}\left(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)}\right)$ since each of these sub-segments are contained in some $B(\boldsymbol{\ell}^{(i)})$ for $i \in \mathbb{N}$. Moreover, the differentiable space over the interior of line segment $\mathcal{L}\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$ can be decomposed as the product of differentiable spaces, which we denote as

$$\Omega_i^{(1)} \times ... \times \Omega_i^{(J)}. \tag{S1.78}$$

By Condition 1, (S1.78) is also a local optimality space. Let $U^{(i,j)}$ be an orthonormal basis of $\Omega_i^{(j)}$ for $j = 1, ..., J$. For each i, we can express $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)$ for all $\boldsymbol{\lambda} \in \text{Int}\left\{\mathcal{L}\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)\right\}$ as

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

$$\hat{\boldsymbol{\beta}}(\boldsymbol{\lambda}|T) = \begin{pmatrix} \hat{\boldsymbol{\beta}}^{(1)}(\boldsymbol{\lambda}|T) & ... & \hat{\boldsymbol{\beta}}^{(J)}(\boldsymbol{\lambda}|T) \end{pmatrix} = \arg\min_{\beta} L_T\left(\{U^{(i,j)}\boldsymbol{\beta}^{(j)}\}_{j=1}^{J}, \boldsymbol{\lambda}\right).$$

We can show that the fitted parameters satisfy the Lipschitz condition (S1.69) over $\Lambda = \mathcal{L}\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$ by using a similar proof as in Lemma 1. The only difference is that the proofs starts with taking directional derivatives along the columns of $U^{(i)} = (U^{(i,1)}...U^{(i,J)})$ to establish the KKT conditions. Then for all j and i, we have

$$\left\| \hat{\boldsymbol{\beta}}^{(j)}(\boldsymbol{p}^{(i)}|T) - \hat{\boldsymbol{\beta}}^{(j)}(\boldsymbol{p}^{(i)}|T) \right\|_2 \leq \frac{1}{m(T)} \frac{\|\ell_j\|_T}{\lambda_{min}} \sqrt{\|\epsilon\|_T^2 + 2C_\Lambda^*} \left| p_j^{(i)} - p_j^{(i+1)} \right|. \quad \text{(S1.79)}$$

We can sum these inequalities by the triangle inequality:

$$\begin{aligned}
\left\| \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}^{(1)}|T) - \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}^{(2)}|T) \right\|_2 &\leq \sum_{i=1}^{\infty} \left\| \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{p}^{(i)}|T) - \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{p}^{(i+1)}|T) \right\|_2 \\
&\leq \frac{1}{m(T)} \frac{\|\ell_j\|_T}{\lambda_{min}} \sqrt{\|\epsilon\|_T^2 + 2C_\Lambda^*} \sum_{i=1}^{\infty} \left| p_j^{(i)} - p_j^{(i+1)} \right| \\
&= \frac{1}{m(T)} \frac{\|\ell_j\|_T}{\lambda_{min}} \sqrt{\|\epsilon\|_T^2 + 2C_\Lambda^*} \left| \lambda_j^{(1)} - \lambda_j^{(2)} \right|.
\end{aligned}$$

Finally, using the fact that g_j is ℓ_j-Lipschitz, we have by the triangle inequality and Cauchy Schwarz that

$$C_\Lambda(\boldsymbol{x}|T) = \frac{\sqrt{\|\epsilon\|_T^2 + 2C_\Lambda^*}}{m(T)\lambda_{min}} \sqrt{\sum_{j=1}^{J} \|\ell_j\|_T^2 \ell_j^2(x_j)}. \quad \text{(S1.80)}$$

□

In order to extend the result in Lemma 13 to all of Λ, we need to show that Λ_{ext} is a set with measure zero.

Lemma 14. *Suppose Condition 2. Then* $\mu_{2J}(\Lambda_{ext}) = 0$ *where* μ_{2J} *is the Lebesgue measure in* $\mathbb{R}^{2J}$ *and* Λ_{ext} *was defined in* (S1.76).

Proof. Suppose for contradiction that $\mu_{2J}(\Lambda_{ext}) > 0$. If this is the case, then there exists a ball $B_r\left(\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right)\right)$ contained in Λ_{ext} with nonzero radius $r > 0$ centered at $\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right)$ where $\boldsymbol{\lambda}^{(1)} \neq \boldsymbol{\lambda}^{(2)}$ and

$$\mu_1\left(\mathcal{L}\left(\boldsymbol{\lambda}^{'}, \boldsymbol{\lambda}^{''}\right) \cap \Lambda_{smooth}^c\right) > 0 \quad \forall \left(\boldsymbol{\lambda}^{'}, \boldsymbol{\lambda}^{''}\right) \in B_r\left(\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right)\right). \tag{S1.81}$$

Suppose that $\mu_1\left(\mathcal{L}\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right) \cap \Lambda_{smooth}^c\right) = \delta > 0$. We claim that for a sufficiently small radius r', we also have

$$\mu_1\left(\mathcal{L}\left(\boldsymbol{\lambda}^{'}, \boldsymbol{\lambda}^{''}\right) \cap \Lambda_{smooth}^c\right) > \delta/2 > 0 \quad \forall \left(\boldsymbol{\lambda}^{'}, \boldsymbol{\lambda}^{''}\right) \in B_{r'}\left(\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right)\right). \tag{S1.82}$$

To see why this claim is true, let us define a monotonically decreasing sequence $\{r_i\}$ where $r_i > 0$ for all $i \in \mathbb{N}$ and $\lim_{i\to\infty} r_i = 0$. By the monotone convergence theorem,

$$\lim_{i\to\infty} \inf_{\left(\boldsymbol{\lambda}', \boldsymbol{\lambda}''\right) \in B_{r_i}\left(\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right)\right)} \mu_1\left(\mathcal{L}\left(\boldsymbol{\lambda}^{'}, \boldsymbol{\lambda}^{''}\right) \cap \Lambda_{smooth}^c\right) = \mu_1\left(\mathcal{L}\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right) \cap \Lambda_{smooth}^c\right) = \delta > 0. \tag{S1.83}$$

By the definition of limits, there is some sufficiently large i' such that for $r' := r_{i'} > 0$, we have

$$\inf_{\left(\boldsymbol{\lambda}', \boldsymbol{\lambda}''\right) \in B_{r'}\left(\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right)\right)} \mu_1\left(\mathcal{L}\left(\boldsymbol{\lambda}^{'}, \boldsymbol{\lambda}^{''}\right) \cap \Lambda_{smooth}^c\right) > \delta/2. \tag{S1.84}$$

Given our ball is non-empty, there exist points $\left(\boldsymbol{\lambda}^{(3)}, \boldsymbol{\lambda}^{(4)}\right), \left(\boldsymbol{\lambda}^{(5)}, \boldsymbol{\lambda}^{(6)}\right) \in B_{r'}\left(\left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\right)\right)$ where

$$\lambda_j^{(3)} > \lambda_j^{(5)}, \lambda_j^{(4)} > \lambda_j^{(6)} \quad \forall j = 1, .., J. \tag{S1.85}$$

For any $\alpha \in (0, 1)$, the line

$$\mathcal{L}_\alpha = \mathcal{L}\left(\alpha\boldsymbol{\lambda}^{(3)} + (1-\alpha)\boldsymbol{\lambda}^{(5)}, \alpha\boldsymbol{\lambda}^{(4)} + (1-\alpha)\boldsymbol{\lambda}^{(6)}\right) \tag{S1.86}$$

has

$$\mu_1 \left(\mathcal{L}_\alpha \cap \Lambda^c_{smooth} \right) > \delta/2. \tag{S1.87}$$

As the lines $\mathcal{L}_\alpha$ do not intersect for $\alpha \in (0,1)$, then

$$\mu \left(\cup_{\alpha \in [0,1]} \left(\mathcal{L}_\alpha \cap \Lambda^c_{smooth} \right) \right) = \int_0^1 \mu_1 \left(\mathcal{L}_\alpha \cap \Lambda^c_{smooth} \right) d\alpha > \delta/2 \tag{S1.88}$$

Thus

$$\mu \left(\Lambda^c_{smooth} \right) \geq \mu \left(\cup_{\alpha \in [0,1]} \left(\mathcal{L}_\alpha \cap \Lambda^c_{smooth} \right) \right) > \delta/2. \tag{S1.89}$$

However this is a contradiction of our assumption that $\mu \left(\Lambda^c_{smooth} \right) = 0$. □

Finally, combining Lemmas 13 and 14, we can show that the Lipschitz condition is satisfied over all of Λ.

Proof for Lemma 2. Since we already showed Lemma 13, it suffices to show that the Lipschitz condition is satisfied for any $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{ext}$. Lemma 14 states that $\mu_{2J}(\Lambda_{ext}) = 0$, which means that there exists a sequence $\left\{ \left(\boldsymbol{\lambda}^{(1,i)}, \boldsymbol{\lambda}^{(2,i)} \right) \right\}_{i=1}^\infty \subseteq \Lambda^c_{ext}$ such that $\lim_{i \to \infty} \left(\boldsymbol{\lambda}^{(1,i)}, \boldsymbol{\lambda}^{(2,i)} \right) = \left(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \right)$. As L_T is continuous and we have assumed that there exists a unique minimizer of $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ for all $\boldsymbol{\lambda} \in \Lambda$, then $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ is continuous in $\boldsymbol{\lambda}$ over all Λ. As $g(\boldsymbol{\theta})(x)$ is also continuous in $\boldsymbol{\theta}$, then for any

$\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda$, we have

$$\left| g(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}|T)(\boldsymbol{x}) - g(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)}|T)(\boldsymbol{x}) \right| = \lim_{i\to\infty} \left| g(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1,i)}|T))(\boldsymbol{x}) - g(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2,i)}|T))(\boldsymbol{x}) \right| \tag{S1.90}$$

$$\leq \lim_{i\to\infty} C_\Lambda(\boldsymbol{x}|T)\|\boldsymbol{\lambda}^{(1,i)} - \boldsymbol{\lambda}^{(2,i)}\|_2 \tag{S1.91}$$

$$= C_\Lambda(\boldsymbol{x}|T)\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|_2 \tag{S1.92}$$

where $C_\Lambda(\boldsymbol{x}|T)$ is defined in (S1.80). □

Proof for 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

$$\hat{\boldsymbol{m}}(\boldsymbol{\lambda}) = \{\hat{m}_j(\boldsymbol{\lambda})\}_{j\in H_0} = \underset{m_j\in\mathbb{R}:j\in H_0}{\arg\min} \frac{1}{2} \left\| y - \sum_{j=1}^J (\hat{g}_{1,j} + m_j h_j) \right\|_T^2 + \sum_{j=1}^J \lambda_j P_j (\hat{g}_{1,j} + m_j h_j). \tag{S1.93}$$

By the gradient optimality conditions, we have

$$\nabla_m \left[\frac{1}{2} \|y - \sum_{j=1}^J (\hat{g}_{1,j} + m_j h_j) \|_T^2 + \sum_{j=1}^J \lambda_j P_j(\hat{g}_{1,j} + m_j h_j) \right] \Bigg|_{m=\hat{m}(\lambda)} = 0. \tag{S1.94}$$

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

$$\nabla_\lambda \nabla_m \left[\frac{1}{2} \|y - \sum_{j=1}^J (\hat{g}_{1,j} + m_j h_j) \|_T^2 + \sum_{j=1}^J \lambda_j P_j(\hat{g}_{1,j} + m_j h_j) \right] \Bigg|_{m=\hat{m}(\lambda)} = 0. \tag{S1.95}$$

From the product rule and chain rule, we can write the system of equations from (S1.95) as

$$\nabla_\lambda \hat{\boldsymbol{m}}(\boldsymbol{\lambda}) = -\left(\nabla_m^2 L_T(\boldsymbol{m},\boldsymbol{\lambda})\right)^{-1} \operatorname{diag}\left\{ \frac{\partial}{\partial m_j} P_j(\hat{g}_{1,j} + m_j h_j)\Big|_{m=\hat{m}(\lambda)} \right\}_{j=1}^{J} \tag{S1.96}$$

where $L_T(\boldsymbol{m},\boldsymbol{\lambda})$ is the loss in (S1.94).

We now bound the second term in (S1.96). From (S1.94) and Cauchy Schwarz, we have for all $k = 1, ..., J$

$$\left|\frac{\partial}{\partial m_k} P_k(\hat{g}_{1,k} + m_k h_k)\right|_{m=\hat{m}(\lambda)} \le \frac{1}{\lambda_{min}} \left\| y - \sum_{j=1}^{J} \left(\hat{g}_{1,j} + \hat{m}_j(\boldsymbol{\lambda}) h_j\right) \right\|_T \|h_k\|_T. \tag{S1.97}$$

From the definition of h_k, we know that $\|h_k\|_T \le \sqrt{\frac{n_D}{n_T}}$. By definition of $\hat{m}(\boldsymbol{\lambda})$ and $\hat{g}_1$, we also have

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

Hence

$$\left|\frac{\partial}{\partial m_k} P_k(\hat{g}_{1,k} + m_k h_k)\right|_{m=\hat{m}(\lambda)} \le \frac{1}{\lambda_{min}} \sqrt{\left(\|\epsilon\|_T^2 + 2C_\Lambda^*\right) \frac{n_D}{n_T}}. \tag{S1.98}$$

By (4.40), we know $\nabla_m^2 L_T(\boldsymbol{m},\boldsymbol{\lambda}) \succeq m(T) I$. So for all k,

$$\|\nabla_\lambda \hat{m}_k(\boldsymbol{\lambda})\|_2 \le \frac{m(T)}{\lambda_{min}} \sqrt{\left(\|\epsilon\|_T^2 + 2C_\Lambda^*\right) \frac{n_D}{n_T}} \tag{S1.99}$$

By the mean value inequality and Cauchy Schwarz, we have

$$\left|\hat{m}_k(\boldsymbol{\lambda}^{(2)}) - \hat{m}_k(\boldsymbol{\lambda}^{(1)})\right| \le \frac{m(T)}{\lambda_{min}} \sqrt{\left(\|\epsilon\|_T^2 + 2C_\Lambda^*\right) \frac{n_D}{n_T}}. \tag{S1.100}$$

By construction, $\left|\hat{m}_k(\boldsymbol{\lambda}^{(2)}) - \hat{m}_k(\boldsymbol{\lambda}^{(1)})\right| = \left\|\hat{g}_k(\boldsymbol{\lambda}^{(2)}|T) - \hat{g}_k(\boldsymbol{\lambda}^{(1)}|T)\right\|_{D^{(n)}}$. So we obtain our desired result in (4.41). $\square$

S1.4 Examples: detailed derivations

Example 1 (Multiple ridge penalties) Here we present the details for deriving (4.24) for Example 1. The additive components $g_j(\boldsymbol{\theta}^{(j)})(\boldsymbol{x}^{(j)})$ are linear functions that are ℓ_j-Lipschitz where $\ell_j(\boldsymbol{x}^{(j)}) = \|\boldsymbol{x}^{(j)}\|_2$. Then by Lemma 1, the fitted function $g(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}))(\boldsymbol{x})$ satisfy Assumption 1 over $\mathbb{R}^p$ with

$$C_\Lambda\left(\boldsymbol{x}|T\right) = n^{2t_{\min}}\sqrt{C_T^*\left(\sum_{j=1}^{J}\|\boldsymbol{x}^{(j)}\|_2^2\left(\frac{1}{n_T}\sum_{(x_i,y_i)\in T}\|\boldsymbol{x}_i^{(j)}\|_2^2\right)\right)} \tag{S1.101}$$

where C_T^* is defined in Example 1 of the main manuscript.

Example 2 (Multiple sobolev penalties) here we present the details for deriving (4.24) for Example 2 Since the solution to (4.27) must be the sum of natural cubic splines [Buja et al., 1989], we can parameterize the space using a Reproducing Kernel Hilbert Space with inner product

$$\langle f, g\rangle = \int_0^1 f^{''}(x)g^{''}(x)dx \tag{S1.102}$$

and the reproducing kernel

$$R(s,t) = st(s\wedge t) + \frac{s+t}{2}(s\wedge t)^2 + \frac{1}{3}(s\wedge t)^3 \tag{S1.103}$$

[Heckman et al., 2012]. Then one can instead solve for (4.27) over the functions g of the form

$$g(x_1, ..., x_J) = \alpha_0 + \sum_{j=1}^{J} g_j(x_j) \tag{S1.104}$$

where the functions g_j are split into a linear component and an orthogonal non-linear component

$$g_j(x_j) = \alpha_{1j} x_j + \sum_{i=1}^{n_T} \theta_{ij} R(x_{ij}, x_j). \tag{S1.105}$$

For notational simplicity, we will also denote $\vec{R}(x|D)_{ij} = R(x_{ij}, x_j)$. We will also write

$$g_{j,\perp}(x_j) = \sum_{i=1}^{n_T} \theta_{ij} R(x_{ij}, x_j). \tag{S1.106}$$

Using this finite-dimensional representation, we find that

$$\int_0^1 \left(g_j^{''}(x) \right)^2 dx = \sum_{u=1}^{n_T} \sum_{v=1}^{n_T} \theta_{uj} \theta v j R(x_{uj}, x_{vj}) = \theta_j^\top K_j \theta_j \tag{S1.107}$$

where the matrix K_j has elements $K_{j,(u,v)} = R(x_{uj}, x_{vj})$. Since any g_j with non-zero $\boldsymbol{\theta}_j$ will have a positive Sobolev penalty, then the matrix K_j must be positive definite. Using the formulation above, we re-express (4.27) as the finite-dimensional problem

$$\hat{\alpha_0}(\boldsymbol{\lambda}), \hat{\boldsymbol{\alpha}}_1(\boldsymbol{\lambda}), \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \underset{\alpha_0, \boldsymbol{\alpha}_1, \boldsymbol{\theta}}{\arg\min} \frac{1}{2} \left\| \boldsymbol{y}_T - \alpha_0 \mathbf{1} - X_T \boldsymbol{\alpha}_1 - K\boldsymbol{\theta} \right\|_2^2 + \frac{1}{2} \boldsymbol{\theta}^\top \operatorname{diag}\left(\{ \lambda_j K_j \} \right) \boldsymbol{\theta}. \tag{S1.108}$$

where $K = (K_1 ... K_J)$. In order to make the fitted functions $\hat{g}_j$ identifiable, we add the usual constraint that $\sum_{i=1}^{n_T} g_j(x_{ij}) = 0$ for all j. We also assume that $X_T^\top X_T$ is nonsingular to ensure that there is a unique $\hat{\alpha}_1$.

The KKT conditions then gives us

$$\hat{\alpha}_0 = \frac{1}{n_T} \sum_{(x_i, y_i) \in T} y_i \tag{S1.109}$$

$$\hat{\boldsymbol{\alpha}}_1(\boldsymbol{\lambda}) = (X_T^\top X_T)^{-1} X_T^\top (\boldsymbol{y}_T - \hat{\alpha}_0 \mathbf{1} - K\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \tag{S1.110}$$

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \mathrm{diag}(K_j^{-1/2}) \left(K^{(1/2)\top} P_{X_T}^\top K^{(1/2)} + \mathrm{diag}(\lambda_j I) \right)^{-1} K^{(1/2)\top} P_{X_T}^\top (I - \frac{1}{n} \mathbf{1}\mathbf{1}^\top) \boldsymbol{y}_T \tag{S1.111}$$

where $K^{(1/2)} = (K_1^{1/2} ... K_J^{1/2})$, I is the $n_T \times n_T$ identity matrix, and $P_{X_T}^\top = I - X_T (X_T^\top X_T)^{-1} X_T^\top$.

To apply Theorem 1, we need to characterize how $\hat{g}(\boldsymbol{\lambda})(\cdot)$ varies with $\boldsymbol{\lambda}$. Since we have the closed form solution to (S1.111), we use it to directly bound the Lipschitz factor $C_\Lambda(\boldsymbol{x}|D^{(n_T)})$. From Green and Silverman [1993], we know that the value of the cubic $\hat{g}_j$ on the interval $[t_L, t_R]$ can be defined using its values and second derivatives at the ends of the interval. Let $h = t_R - t_L$. Then the value of the cubic

$$\begin{aligned} \hat{g}_{j,\perp}(x_j) = \hat{\alpha_{1j}} x_j &+ \frac{(x_j - t_L)\hat{g}_{j,\perp}(t_R) + (t_R - t)\hat{g}_{j,\perp}(t_L)}{h} \\ &- \frac{1}{6}(x_j - t_L)(t_R - x_j) \left\{ \left(1 + \frac{x_j - t_L}{h} \right) \hat{g}''_{j,\perp}(t_R^+) \left(1 + \frac{t_R - x_j}{h} \right) \hat{g}''_{j,\perp}(t_L^+) \right\}. \end{aligned} \tag{S1.112}$$

Let $\hat{\boldsymbol{\gamma}}_j$ be the vector of second derivatives of $\hat{g}''_{j,\perp}$ for observations in the training data. Since the fitted functions $\hat{g}_{j,\perp}$ must be natural cubic splines, $\hat{\boldsymbol{\gamma}}_j$ and $\hat{\boldsymbol{\theta}}_j$ have a linear relationship:

$$\hat{\boldsymbol{\gamma}}_j = R_j^{-1} Q_j^\top K_j \hat{\boldsymbol{\theta}}_j \tag{S1.113}$$

where the matrix R_j is a banded diagonally dominant matrix and Q_j is a banded negative-semi-definite matrix that depend on the covariates x_j in the training data. For the definitions of R_j and Q_j, refer to Green and Silverman [1993]. Let $h_j(D^{(n_T)})$ be the smallest distance between observations of the jth covariates in the training data T. Then using the Gershgorin circle theorem [Gershgorin, 1931], one can show that all the eigenvalues of R_j are larger than $\frac{1}{3}h_j(D^{(n_T)})$ and all the eigenvalues of Q_j have magnitudes no greater than $4/h_j(D^{(n_T)})$. Thus using (S1.112) and (S1.113), we have that

$$\|\nabla_\lambda \hat{g}_{j,\perp}(\boldsymbol{\lambda})(x_j)\|_2 \le \frac{c}{h_j(D^{(n_T)})^2} \left\|\nabla_\lambda K_j \hat{\boldsymbol{\theta}}_j(\boldsymbol{\lambda})\right\|_2 \tag{S1.114}$$

for some absolute constant $c > 0$. To bound the second term on the right hand side, we know from (S1.111) that

$$\nabla_{\lambda_\ell} K_j \hat{\boldsymbol{\theta}}_j(\boldsymbol{\lambda}) \tag{S1.115}$$

$$= \begin{bmatrix} 0 & .. & 0 & K_j^{1/2} & 0 & .. & 0 \end{bmatrix} \left(K^{(1/2),\top} P_{X_T}^\top K^{(1/2)} + \mathrm{diag}\{\lambda_j I\}_{j=1:J}\right)^{-2} K^{(1/2),\top} P_{X_T}^\top (I - \frac{1}{n}\mathbf{1}\mathbf{1}^\top)\boldsymbol{y}_T \tag{S1.116}$$

if $\ell = j$. Otherwise $\nabla_{\lambda_\ell} K_j \hat{\boldsymbol{\theta}}_j(\boldsymbol{\lambda}) = 0$. Thus

$$\left\|\nabla_{\lambda_\ell} K_j \hat{\boldsymbol{\theta}}_j(\boldsymbol{\lambda})\right\|_2 \le \lambda_{\min}^{-2} \|\boldsymbol{y}_T\|_2 \sqrt{\|K_j\|_2 \sum_{j'=1}^{J} \|K_{j'}\|_2^2} \tag{S1.117}$$

The eigenvalues of K_j are bounded above by the largest row sum, which is no more than $2n_T$ (assuming all training covariates are between 0 and 1). Putting the results above together, we have

$$\|\nabla_\lambda \hat{g}_{j,\perp}(\boldsymbol{\lambda})(x_j)\|_2 \leq \frac{c\sqrt{J}n_T}{h_j(D^{(n_T)})^2\lambda_{\min}^2}\|\boldsymbol{y}_T\|_2. \tag{S1.118}$$

Also, we have from (S1.110) that

$$\begin{aligned}
\|\nabla_\lambda \hat{\boldsymbol{\alpha}}_1(\boldsymbol{\lambda})\|_2 &= \left\| \left(X_T^\top X_T\right)^{-1} X_T^\top \nabla_{\lambda_j} K \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) \right\|_2 && \text{(S1.119)}\\
&= \left\| \left(X_T^\top X_T\right)^{-1} X_T^\top \nabla_{\lambda_j} K_j \hat{\boldsymbol{\theta}}_j(\boldsymbol{\lambda}) \right\|_2 && \text{(S1.120)}\\
&\leq \left\| \left(X_T^\top X_T\right)^{-1} X_T^\top \right\|_2 \lambda_{\min}^{-2}\|\boldsymbol{y}_T\|_2 n_T\sqrt{J} && \text{(S1.121)}
\end{aligned}$$

Finally we can conclude that

$$\begin{aligned}
\left\|\hat{g}_j(\boldsymbol{\lambda}^{(1)})(x_j) - \hat{g}_j(\boldsymbol{\lambda}^{(2)})(x_j)\right\|_2 \leq &\left(|x_j| \left\| \left(X_T^\top X_T\right)^{-1} X_T^\top \right\|_2 + \frac{c}{h_j(D^{(n_T)})^2}\right) \\
&\times \sqrt{J}n_T\lambda_{\min}^{-2}\|\boldsymbol{y}_T\|_2\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|_2
\end{aligned} \tag{S1.122}$$

By triangle inequality, we get the Lipschitz factor for the fitted model $\hat{g}$ by summing up (S1.122) for $j = 1, .., J$. We find that the Lipschitz factor in (4.24) is

$$C_\Lambda(\boldsymbol{x}|T) = \left(J\left\| \left(X_T^\top X_T\right)^{-1} X_T^\top \right\|_2 + \sum_{j=1}^{J} \frac{c}{h_j(T)^2}\right)\sqrt{J}n^{2t_{\min}+1}\|\boldsymbol{y}\|_T. \tag{S1.123}$$

Example 3 (Multiple elastic nets, training-validation split) Here we check that all the conditions for Lemma 2 are satisfied.

First we check Condition 1. Since the absolute value function $|\cdot|$ is twice-continuously differentiable everywhere except at zero, the directional derivatives of $||\boldsymbol{\theta}^{(j)}||_1$ at $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ only exist along directions spanned by the columns of $\boldsymbol{I}_{I^{(j)}(\boldsymbol{\lambda})}$. Thus the penalized training loss $L_T(\cdot, \boldsymbol{\lambda})$ is twice differentiable with respect to the directions in

$$\Omega^{L_T(\cdot,\lambda)}(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}|T)) = \text{span}(I_{I^{(1)}(\boldsymbol{\lambda})}) \times ... \times \text{span}(I_{I^{(J)}(\boldsymbol{\lambda})}). \tag{S1.124}$$

Moreover, the elastic net solution paths are piecewise linear [Zou and Hastie, 2003]. This implies that the nonzero indices of the elastic net estimates stay locally constant for almost every $\boldsymbol{\lambda}$; so (S1.124) is also a local optimality space for $L_T(\cdot, \boldsymbol{\lambda})$. In addition, this implies that Condition 2 is satisfied.

We also check that the Hessian of the penalized training loss has a minimum eigenvalue bounded away from zero. Consider the following orthogonal basis of (S1.124) at $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$: $U(\boldsymbol{\lambda}) = \{U^{(j)}(\boldsymbol{\lambda})\}_{j=1}^{J}$ where

$$U^{(j)} = \begin{pmatrix} \mathbf{0} \\ I_{I^{(j)}(\boldsymbol{\lambda})} \\ \mathbf{0} \end{pmatrix} \quad \forall j = 1, ..., J. \tag{S1.125}$$

The Hessian matrix of $L_T(\cdot, \boldsymbol{\lambda})$ with respect to directions $U(\boldsymbol{\lambda})$ is

$$\boldsymbol{U}(\boldsymbol{\lambda})^\top \boldsymbol{X}_T^\top \boldsymbol{X}_T \boldsymbol{U}(\boldsymbol{\lambda}) + \lambda_1 w \boldsymbol{I} \tag{S1.126}$$

where $\boldsymbol{X}_T = (\boldsymbol{X}^{(1)} ... \boldsymbol{X}^{(J)})$ and $\boldsymbol{I}$ is the identity matrix with length equal to the number of nonzero elements in $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$. Since the first summand is positive semi-definite and $\lambda_1 > \lambda_{\min}$, (S1.126) has a minimum eigenvalue of $\lambda_{\min} w$.

Example 4 (Multiple elastic nets, cross-validation) Here we present details for establishing an oracle inequality when multiple elastic net penalties are tuned via the averaged version of K-fold cross-validation. First we check the conditions in

Theorem 2 are satisfied. In the problem setup, X is a log-concave vector and $\sup_{\|a\|_\infty=1} \left\|X^\top a\right\|_{L_{\psi_2}} < c_R < \infty$ for some constant c_R. Using a similar procedure as Lecué and Mitchell [2012], we can then show that (3.14) and (3.15) in Assumption 2 are satisfied with $K_0 \coloneqq (\|\boldsymbol{\theta}^*\|_\infty + K_0')c_R$.

Next we find the Lipschitz factor. We can upper bound the Lipschitz factor of the thresholded model with the Lipschitz factor of the un-thresholded model. So Assumption 1 is satisfied over $\mathbb{R}^p$ with

$$C_\Lambda(\boldsymbol{x}|D^{(n_T)}) = \frac{n^{2t_{\min}}}{w} R^2 \sqrt{Jp\left(\|\epsilon\|^2_{D^{(n_T)}} + \sum_{j=1}^{J} 2\|\boldsymbol{\theta}^{*,(j)}\|_1 + w\|\boldsymbol{\theta}^{*,(j)}\|_2^2\right)}. \quad \text{(S1.127)}$$

Finally, to apply Theorem 2, we must find a bound for (3.16). Let $\sigma_0 = O_p(n^{4t_{\min}} R^4 Jp/w^2)$. Using the fact that $\left\|C_\Lambda(\cdot|D^{(n_T)})\right\|^2_{L_{\psi_2}}$ is a linear function of $\|\epsilon\|^2_{D^{(n_T)}}$, which is a sub-exponential random variable, we have that

$$\sum_{k=1}^{\infty} k \Pr\left(\|C_\Lambda(\cdot|D^{(n_T)})\|_{L_{\psi_2}} \geq 2^k\sigma_0\right) \leq \sum_{k=1}^{\infty} k \Pr\left(\|\epsilon\|^2_{D^{(n_T)}} \geq 2^{2k}\right) \leq c_1 \exp\left(-\frac{c_0 n_T}{\|\epsilon\|^2_{L_{\psi_2}}}\right) \quad \text{(S1.128)}$$

for constants $c_0, c_1 > 0$. Plugging in this bound to Theorem 2 gives us our desired result.

Bibliography

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

Guillaume Lecué and Charles Mitchell. Oracle inequalities for cross-validation type procedures. *Electronic Journal of Statistics*, 6:1803–1837, 2012.

Michel Talagrand. *The generic chaining: upper and lower bounds of stochastic processes*. Springer Monographs in Mathematics, 2005.

Andreas Buja, Trevor Hastie, and Robert Tibshirani. Linear smoothers and additive models. *The Annals of Statistics*, pages 453–510, 1989.

Nancy Heckman et al. The theory and application of penalized methods or reproducing kernel hilbert spaces made easy. *Statistics Surveys*, 6:113–141, 2012.

Peter J Green and Bernard W Silverman. *Nonparametric regression and generalized linear models: a roughness penalty approach*. CRC Press, 1993.

Semyon Aranovich Gershgorin. Uber die abgrenzung der eigenwerte einer matrix. *Bulletin de l'Académie des Sciences de l'URSS. Classe des sciences mathématiques et na*, (6):749–754, 1931.

Hui Zou and Trevor Hastie. Regression shrinkage and selection via the elastic net. *Journal of the Royal Statistical Society: Series B. v67*, pages 301–320, 2003.