1 Appendix

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 $\langle f, g \rangle_D = \frac{1}{m} \sum_{x_i \in D} f(x_i) g(x_i)$.

1.1 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*

$$\left\| g^* - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\|_V^2 - \left\| g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) \right\|_V^2 \leq \left\langle \epsilon, \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T) - \hat{g}^{(n_T)}(\hat{\boldsymbol{\lambda}}|T) \right\rangle_V \tag{1}$$

Proof. By definition,

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

□

We are therefore interested in bounding the empirical process term in (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

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

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:*

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

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

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

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

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

we have

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

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

Theorem 3. *Consider a set of hyper-parameters Λ. Suppose independent random variables $\epsilon_1, ...\epsilon_n$ have expectation zero and are uniformly sub-Gaussian with parameter b and B. Suppose there is a function $\psi : \mathbb{R} \mapsto \mathbb{R}$ and constant $r > 0$ such that*

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi(R) \quad \forall R > r \tag{5}$$

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

$$\sqrt{n_V}\delta^2 \geq c\left(\psi(\delta) \vee \delta \vee \psi\left(4\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2\right) \vee 4\left\|g^* - \hat{g}^{(n_T)}(\tilde{\boldsymbol{\lambda}}|T)\right\|_V^2\right) \tag{6}$$

we have

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

Proof. 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.

$$Pr\left(\left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \geq \delta^2\right) \tag{7}$$

$$= \sum_{s=0}^{\infty} Pr\left(2^{2s}\delta^2 \leq \left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \leq 2^{2s+2}\delta^2\right) \tag{8}$$

$$\leq \sum_{s=0}^{\infty} Pr\left(2^{2s}\delta^2 \leq 2\left\langle\epsilon, \hat{g}(\hat{\boldsymbol{\lambda}}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V\right. \tag{9}$$

$$\left. \wedge \left\|\hat{g}(\hat{\boldsymbol{\lambda}}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V^2 \leq 2^{2s+2}\delta^2 + 2\left|\left\langle\hat{g}(\tilde{\boldsymbol{\lambda}}) - \hat{g}(\hat{\boldsymbol{\lambda}}), \hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\rangle_V\right|\right) \tag{10}$$

where we applied the basic inequality in the last inequality. Each summand in (9) 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 (7)

$$Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V \le 4\left\|\hat{g}(\tilde{\boldsymbol{\lambda}})-g^*\right\|_V} 2\left\langle \epsilon, \hat{g}(\boldsymbol{\lambda}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V \ge \delta^2\right) \tag{11}$$

$$+\sum_{s=0}^{\infty} Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V \le 2^{s+3/2}\delta} 2\left\langle \epsilon, \hat{g}(\boldsymbol{\lambda}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V \ge 2^{2s}\delta^2\right) \tag{12}$$

We can bound both (11) and (12) using Lemma 5. For our choice of δ in (6), there is some constant $a > 0$ dependent only on b such that (11) 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 (6) 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

$$Pr\left(\sup_{\boldsymbol{\lambda}\in\Lambda:\left\|\hat{g}(\boldsymbol{\lambda})-\hat{g}(\tilde{\boldsymbol{\lambda}})\right\|_V \le 2^{s+3/2}\delta} 2\left\langle \epsilon, \hat{g}(\boldsymbol{\lambda}) - \hat{g}(\tilde{\boldsymbol{\lambda}})\right\rangle_V \ge 2^{2s}\delta^2\right) \tag{13}$$

$$\le a\exp\left(-n_V\frac{2^{4s-2}\delta^4}{4a^2 2^{2s+3}\delta^2}\right) \tag{14}$$

Putting this all together, there is a constant c such that (7) is bounded above by

$$c\exp\left(-\frac{n_V\delta^4}{4c^2\left(16\left\|\hat{g}(\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2\right)}\right) + c\exp\left(-\frac{n_V\delta^2}{c^2}\right) \tag{15}$$

□

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 C-Lipschitz in $\boldsymbol{\lambda}$ with respect to some norm $\|\cdot\|$. Then the entropy of $\mathcal{G}(T)$ with respect to $\|\cdot\|$ is*

$$H\left(u, \mathcal{G}(T), \|\cdot\|\right) \le J\log\left(\frac{4C\left(\lambda_{max} - \lambda_{min}\right) + 2u}{u}\right) \tag{16}$$

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

$$N\left(u, \Lambda, \|\cdot\|_2\right) \leq \left(\frac{4\left(\lambda_{max} - \lambda_{min}\right) + 2u}{u}\right)^J \tag{17}$$

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

$$\begin{aligned} N\left(u, \mathcal{G}(T), \|\cdot\|_V\right) &\leq N\left(\frac{u}{C}, \Lambda, \|\cdot\|_2\right) && (18)\\ &\leq \left(\frac{4\left(\lambda_{max} - \lambda_{min}\right) + 2u/C}{u/C}\right)^J && (19)\end{aligned}$$

□

Theorem 1

Proof. By Lemma 6, we have

$$\begin{aligned} &\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du && (20)\\ &= \int_0^R \left(J \log\left(\frac{4C_\Lambda\Delta_\Lambda + 2u}{u}\right)\right)^{1/2} du && (21)\\ &\leq J^{1/2}\int_0^R \left[\log 4 + \log\left(\frac{8C_\Lambda\Delta_\Lambda}{u}\right)\right]^{1/2} du && (22)\\ &\leq RJ^{1/2}\left[\int_0^1 \left(\log 4 + \log\left(\frac{8C_\Lambda\Delta_\Lambda}{R}\right) + J\log\frac{1}{v}\right) dv\right]^{1/2} && (23)\\ &= R\left[J\left(1 + \log\left(32C_\Lambda\Delta_\Lambda\right) + \log\frac{1}{R}\right)\right]^{1/2} && (24)\end{aligned}$$

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 $C_\Lambda \geq 32e/(n\Delta_\Lambda)$, then

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi(R) \coloneqq 2R\left(J\log(C_\Lambda\Delta_\Lambda n)\right)^{1/2} \tag{25}$$

Applying Theorem 3, we get our desired result. □

1.2 Cross-validation

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

Theorem 4. *Consider a set of hyper-parameters Λ. 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 $d_{\min} \geq 0$ 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

$$\psi(d) \geq \int_0^{\sqrt{d}} H^{1/2}(u, \mathcal{Q}_d, \|\cdot\|_2)du + \frac{\log n_V}{\sqrt{n_V}} \int_0^{2G} H(u, \mathcal{Q}_d, \|\cdot\|_\phi)du \quad \forall d > d_{\min} \tag{26}$$

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 there is an absolute constant $c > 0$ such that for every $a > 0$ and $q > 1$, the following inequality holds

$$E_{D^{(n)}} \left\| \bar{g}(\hat{\boldsymbol{\lambda}}|D^{(n)}) - g^* \right\|^2 \leq (1+a) \inf_{\boldsymbol{\lambda}\in\Lambda} \left[E_{D^{(n_T)}} \left\| \hat{g}^{(n_T)}(\boldsymbol{\lambda}|D^{(n_T)}) - g^* \right\|^2 \right] \tag{27}$$

$$+ \frac{ac}{q} \left[\psi^* \left(\frac{2q^r(1+a)}{a\sqrt{n_V}} \right) \vee d_{\min} \right] \tag{28}$$

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 (26). 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

$$H(u, \mathcal{Q}_d(T), \|\cdot\|_\infty) \leq J \log \left[\frac{16GC\Delta_\Lambda + 2u}{u} \right] \tag{29}$$

Hence we can let

$$\psi(d) \coloneqq \sqrt{d}K_{n,1} + \frac{K_{n,2}}{\sqrt{n_V}} \tag{30}$$

for

$$K_{n,1} = \left[J \left(1 + \log \left(128\sqrt{n_V}GC\Delta_\Lambda \right) \right) \right]^{1/2}$$

and

$$K_{n,2} = \log n_V 2GJ \left(1 + \log \left(128GC\Delta_\Lambda \right) \right)$$

$\psi(d)$ is a valid upper bound in (26) for all $d > n_V^{-1}$.

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

$$\psi^*(z) = \frac{z^2 K_{n,1}^2}{4} + \frac{zK_{n,2}}{\sqrt{n_V}} \tag{31}$$

Plugging in (31) into Theorem 4 gives us the result in Theorem 2. □

1.3 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 $j = 1 : J$

$$\nabla_{\theta^{(j)}} \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} = 0 \tag{32}$$

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

$$\nabla_{\lambda} \left\{ \nabla_{\theta^{(j)}} \left[\frac{1}{2} \|y - g(\boldsymbol{\theta})\|_T^2 + \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\} = 0 \tag{33}$$

Define the following matrices

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

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

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

From the product rule and chain rule, we can then write the system of equations in (33) as

$$\left(\begin{array}{cccc} \nabla_{\lambda} \hat{\boldsymbol{\theta}}_1(\boldsymbol{\lambda}) & \nabla_{\lambda} \hat{\boldsymbol{\theta}}_2(\boldsymbol{\lambda}) & \ldots & \nabla_{\lambda} \hat{\boldsymbol{\theta}}_p(\boldsymbol{\lambda}) \end{array} \right) = -M^{\top} (S + D)^{-1} \tag{34}$$

We now bound each column in M. From the (32) 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

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

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

$$\left\| \nabla_{\theta^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\| \leq \frac{L}{\lambda_{min}} \sqrt{2C_{\theta^*,\Lambda}} \tag{35}$$

Now we bound the norm of $\nabla_\lambda \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$. From (34), we have for all $j = 1, ..., J$

$$\|\nabla_\lambda \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})\| = \|M^\top (S + D)^{-1} e_k\| \tag{36}$$

$$\leq \sum_{j=1}^{J} \left\| \nabla_{\theta^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\lambda)} \right\|_2 \left\| (S + D)^{-1} \right\|_2 \tag{37}$$

$$\leq J \left(\frac{L}{\lambda_{min}} \sqrt{2C_{\theta^*,\Lambda}} \right) \frac{1}{m} \tag{38}$$

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

$$\left\| \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}') \right\|_2 \leq \frac{LJ^{3/2}\sqrt{2C_{\theta^*,\Lambda}}}{m\lambda_{min}} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|_2 \tag{39}$$

Finally the result (28) clearly follows since $g_j(\boldsymbol{\theta})$ are Lipschitz in $\boldsymbol{\theta}$ with respect to $\|\cdot\|_\infty$. □

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 $\boldsymbol{\lambda} \in \mathrm{Int}\left\{\mathcal{L}\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)\right\}$ 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,

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

□

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

$$\hat{\boldsymbol{m}}(\boldsymbol{\lambda}) = \{\hat{m}_j(\boldsymbol{\lambda})\}_{j\in H_0} = \underset{m_j\in\mathbb{R}:j\in H_0}{\arg\min} \frac{1}{2}\left\|y - \sum_{j=1}^{J}\left(\hat{g}_{1,j} + m_j h_j\right)\right\|_T^2 + \sum_{j=1}^{J}\lambda_j P_j\left(\hat{g}_{1,j} + m_j h_j\right) \quad (40)$$

By the gradient optimality conditions, we have

$$\nabla_m\left[\frac{1}{2}\|y - \sum_{j=1}^{J}\left(\hat{g}_{1,j} + m_j h_j\right)\|_T^2 + \sum_{j=1}^{J}\lambda_j P_j(\hat{g}_{1,j} + m_j h_j)\right]\Bigg|_{m=\hat{m}(\lambda)} = 0 \quad (41)$$

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

$$\nabla_\lambda\nabla_m\left[\frac{1}{2}\|y - \sum_{j=1}^{J}\left(\hat{g}_{1,j} + m_j h_j\right)\|_T^2 + \sum_{j=1}^{J}\lambda_j P_j(\hat{g}_{1,j} + m_j h_j)\right]\Bigg|_{m=\hat{m}(\lambda)} = 0 \quad (42)$$

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 (42) can be written as $M = D_2 \left(S + D_1\right)^{-1}$.

Now we bound each element in D_2. From (41) and Cauchy Schwarz, we have for all $k = 1, ..., J$

$$\left|\frac{\partial}{\partial m_k} P_k(\hat{g}_{1,k} + m_k h_k)\right|_{m=\hat{m}(\lambda)} \leq \frac{1}{\lambda_{min}} \left\| y - \sum_{j=1}^{J} (\hat{g}_{1,j} + \hat{m}_j(\boldsymbol{\lambda}) h_j) \right\|_T \|h_k\|_T \tag{43}$$

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

$$\begin{aligned}
&\frac{1}{2} \left\| y - \sum_{j=1}^{J} (\hat{g}_{1,j} + \hat{m}_j(\boldsymbol{\lambda}) h_j) \right\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\hat{g}_{1,j}) \\
&\leq \frac{1}{2} \left\| y - \sum_{j=1}^{J} \hat{g}_{1,j} \right\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\hat{g}_{1,j}) \\
&\leq C_\Lambda^* + J\lambda_{max} \max_{j=1:J} P_j(\hat{g}_{1,j})
\end{aligned}$$

Hence every diagonal element in D_2 is bounded above by

$$\frac{1}{\lambda_{min}} \sqrt{2C_\Lambda^* \frac{n_D}{n_T} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)} \tag{44}$$

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

$$\begin{aligned}
\|\nabla_\lambda \hat{m}_k(\boldsymbol{\lambda})\| &= \left\| D_2 \left(S + D_1\right)^{-1} e_k \right\| && (45) \\
&\leq \frac{1}{\lambda_{min}} \sqrt{2C_\Lambda^* \frac{n_D}{n_T} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)} \left\| \left(S + D_1\right)^{-1} e_k \right\| && (46) \\
& && (47)
\end{aligned}$$

By the assumption in (36), $e_k^\top \left(S + D_1\right) e_k \geq m$.

By the mean value theorem and Cauchy Schwarz, we have

$$\left|\hat{m}_j(\boldsymbol{\lambda}^{(2)}) - \hat{m}_j(\boldsymbol{\lambda}^{(1)})\right| \leq \left\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\right\| \frac{1}{d\lambda_{min}} \sqrt{2C_\Lambda^* \frac{n_D}{n_T} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)} \tag{48}$$

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 (38) follows. □

References

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

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