

1 K-Fold Cross-Validation

Consider the joint optimization problem to find the best regularization parameter λ in Λ via K -fold cross validation. Let D be the entire dataset (so both y, X). For $k = 1, \dots, K$, let D_k represent the k th fold and D_{-k} denote all the folds minus the k th fold. For a given λ , train over D_{-k} and then validate over D_k . Let the number of observations for each fold be n_k and let the total number of observations be n .

Suppose the penalty function P is a semi-norm, smooth, and convex. Suppose the penalty function is taken to the power $v \geq 1$.

We will presume that the domain of the function class is bounded.

Suppose that there is some G s.t. $\sup_{g \in \mathcal{G}} \|g\|^2 \leq G$.

Let $\|h\|^2 = \int h(x)d\mu(x)$. Let $\|h\|_k^2 = \frac{1}{n_k} \sum_{i \in D_k} h(x_i)^2$ and similarly for $\|h\|_{-k}^2$ for the set D_{-k} and $\|h\|_D^2$ for the set D . Let $\langle h, g \rangle_k = \frac{1}{n_k} \sum_{i \in D_k} h(x_i)g(x_i)$ and $\langle h, g \rangle_{-k}$ for the set D_{-k} and $\langle h, g \rangle_D$ for the set D .

Let us define

$$\hat{\lambda} = \arg \min_{\lambda \in \Lambda} \frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D_{-k})\|_k^2$$

$$\hat{g}(\lambda|D_{-k}) = \arg \min_{g \in \mathcal{G}} \frac{1}{2} \|y - g\|_{-k}^2 + \lambda \left(P^v(g) + \frac{w}{2} \|g\|^2 \right)$$

The K -fold CV model is

$$\hat{g}(\lambda|D) = \arg \min_{g \in \mathcal{G}} \frac{1}{2} \|y - g\|_D^2 + \lambda \left(P^v(g) + \frac{w}{2} \|g\|^2 \right)$$

Let the range of Λ be from $\lambda_{min} = O_P(n^{-\tau_{min}})$ to $\lambda_{max} = O_P(1)$.

We show that

$$\|\hat{g}_{\hat{\lambda}}(\cdot|D) - g^*\|_D \lesssim \sqrt{\sum_{k=1}^K \|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_k^2} + \left(\frac{1 + \log(4\sigma^2) + \kappa \log n}{\min_{k=1:K} \{n_k\}} \right)^{1/2}$$

Notation

$a \lesssim b$ means that $a \leq Cb + c$ where $C > 0, c$ are constants independent of n .

2 Proof

The proof is based on two main ideas.

First, we bound the error of the retrained K -fold CV model by a convex combination of the K trained models from each fold.

Second, we bound the entropy of $\hat{\mathcal{G}}(D_{-k}) = \{\hat{g}_\lambda(\cdot|D_{-k}) : \lambda \in \Lambda\}$, which in turn allows us to bound empirical and Rademacher processes.

Step 1:

Define the convex combination

$$\hat{\xi}_\lambda(x) = \frac{1}{K-1} \sum_{k=1}^K \frac{n - n_k}{n} \hat{g}_\lambda(x|D_{-k})$$

By the triangle inequality,

$$\|\hat{g}_{\hat{\lambda}}(\cdot|D) - g^*\|_D \leq \|\hat{g}_{\hat{\lambda}}(\cdot|D) - \hat{\xi}_{\hat{\lambda}}\|_D + \|\hat{\xi}_{\hat{\lambda}} - g^*\|_D$$

We bound the first and second summands in steps 2 and 3, respectively.

Step 2: Bound $\|\hat{g}_{\hat{\lambda}}(\cdot|D) - \hat{\xi}_{\hat{\lambda}}\|_D$

Adding the two inequalities from Lemma 1 and 2, we have

$$\begin{aligned}
& \|\hat{g}_{\hat{\lambda}}(\cdot|D) - \hat{\xi}_{\hat{\lambda}}\|_D^2 \\
& \leq \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \left(\left| \langle \epsilon, \hat{\xi}_{\hat{\lambda}} - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k}) \rangle_{-k} \right| + \left| \langle g^* - \hat{\xi}_{\hat{\lambda}}, \hat{\xi}_{\hat{\lambda}} - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k}) \rangle_{-k} \right| \right) \\
& \lesssim \sum_{k=1}^K \|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_k^2 + \left(\sum_{\ell=1}^K \left| \langle \epsilon, \hat{g}_{\hat{\lambda}}(\cdot|D_{-k}) - g^* \rangle_{-\ell} \right| + \|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_{\ell}^2 - \|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_k^2 \right)
\end{aligned}$$

Our goal is to bound the empirical process and the rademacher process terms using Lemma 3 and 4.

By Lemma 0, if $\|\epsilon\|_k \leq 2\sigma$ for all folds D_k , then for any empirical distribution Q ,

$$H\left(d, \hat{\mathcal{G}}(D_{-k}), \|\cdot\|_Q\right) \lesssim \psi(u) = \log\left(\frac{1}{wd}\right) + \kappa \log n + \log(4\sigma^2)$$

So we have

$$\begin{aligned}
\int_0^R \psi^{1/2}(u) du &= \int_0^R \left(\log\left(\frac{1}{uw}\right) + \kappa \log n + \log(4\sigma^2) \right)^{1/2} du \\
&\lesssim R \left(\int_0^1 \log\left(\frac{1}{u}\right) - \log w + \kappa \log n + \log(4\sigma^2) du \right)^{1/2} \\
&\leq R (1 + \log(4\sigma^2) - \log w + \kappa \log n)^{1/2}
\end{aligned}$$

Let

$$\delta_{max} = CR \left(\left(\frac{1 + \log(4\sigma^2) + \kappa \log n}{\min_{k=1:K} \{n_k\}} \right)^{1/2} \vee 1 \right)$$

By Lemma 3, for some constant c , for all $\ell, k = 1 : K$ (and $\ell \neq k$), we have

$$Pr \left(\sup_{\lambda \in \Lambda} \frac{|\langle \epsilon, \hat{g}_{\hat{\lambda}}(\cdot|D_{-k}) - g^* \rangle_{-\ell}|}{\|\hat{g}_{\hat{\lambda}}(\cdot|D_{-k}) - g^*\|_{-\ell}} \geq \delta_{max} \wedge \|\epsilon\|_{-k} \leq 2\sigma \wedge \|\epsilon\|_{-\ell} \leq 2\sigma \right) \leq c \exp \left(-(n - n_k) \frac{\delta_{max}^2}{c^2 R^2} \right)$$

By Lemma 4, for some constant c , we have for all $\ell, k = 1 : K$ (and $\ell \neq k$), we have

$$Pr \left(\sup_{\lambda \in \Lambda} \frac{\|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_{\ell}^2 - \|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_k^2}{\|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_{k \cup \ell}} \geq \delta_{max} \wedge \|\epsilon\|_{-k} \leq 2\sigma \right) \leq c \exp \left(-n_{\ell} \frac{\delta_{max}^2}{c^2 R^2} \right) + c \exp \left(-n_k \frac{\delta_{max}^2}{c^2 R^2} \right)$$

Hence with high probability, we have

$$\|\hat{g}_{\hat{\lambda}}(\cdot|D) - \hat{\xi}_{\hat{\lambda}}\|_D^2 \lesssim \sum_{k=1}^K \|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_k^2 + \delta_{max} \sum_{\ell=1}^K \|\hat{g}_{\hat{\lambda}}(\cdot|D_{-k}) - g^*\|_{-\ell}$$

Combining Easy Lemma 1 and Lemma 5, we get

$$\|\hat{g}_{\hat{\lambda}}(\cdot|D) - \hat{\xi}_{\hat{\lambda}}\|_D^2 \lesssim \sum_{k=1}^K \|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_k^2 + \delta_{max} \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_k + \delta_{max}^2$$

Step 3:

For every k , we have

$$\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_D^2 \lesssim \sum_{\ell=1}^k \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_k^2 + (\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_\ell^2 - \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_k^2)$$

Lemma 5 directly bounds $\sum_{\ell=1}^k \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\hat{\lambda}}(x|D_{-k})\|_k^2$.

By Lemma 4 and Easy Lemma 1, we have

$$|\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_\ell^2 - \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_k^2| \leq \delta_{max} (\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_k + \delta_{max})$$

Hence

$$\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_D^2 \lesssim \sum_{k=1}^K \|g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k})\|_k^2 + \delta_{max} \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - g^*\|_k + \delta_{max}^2$$

2.1 Lemmas

2.1.1 Lemma 0

Consider any empirical distributions T and Q .

Consider the function class

$$\hat{\mathcal{G}}(T, \epsilon_T) = \left\{ \hat{g}_\lambda(\cdot|T, \epsilon_T) = \arg \min_{g \in \mathcal{G}} \frac{1}{2} \|y - g\|_T^2 + \lambda \left(P^v(g) + \frac{w}{2} \|g\|^2 \right) : \lambda \in \Lambda \right\}$$

Suppose the penalty function P is a semi-norm, smooth, and convex. Suppose $\min_{h \in \mathcal{G}: P(h)=1} \|h\|^2 = O_p(n^{-u})$ and for all h , $\|h\|_Q \leq O_p(n^p)P(h)$.

Suppose $v \geq 1$.

Suppose $\lambda_{min} = O_p(n^{-\tau_{min}})$ and $\lambda_{max} = O_p(1)$.

Then the entropy bound is

$$H(d, \hat{\mathcal{G}}(T, \epsilon_T), \|\cdot\|_Q) \lesssim \log \left(\frac{1}{dw} \right) + \kappa \log n + \log \|\epsilon\|_T^2$$

where κ depends on τ_{min}, v, u, p .

Proof

To find the covering number for $\hat{\mathcal{G}}$, we bound the distance $\|\hat{g}_{\lambda_0}(\cdot|T) - \hat{g}_{\lambda_0+\delta}(\cdot|T)\|_Q$ for every $\lambda_0 \in \Lambda$.

Consider the function $h = c(\hat{g}_{\lambda_0} - \hat{g}_{\lambda_0+\delta})$ where $c > 0$ is some constant s.t. $P(h) = 1$. (We'll assume that $\|\hat{g}_{\lambda_0} - \hat{g}_{\lambda_0+\delta}\|_Q > 0$, since we'll be done otherwise.) Consider the 1-dimensional optimization problem

$$\hat{m}_h(\lambda) = \arg \min_m \frac{1}{2} \|y - (\hat{g}_{\lambda_0} + mh)\|_T^2 + \lambda \left(P^v(\hat{g}_{\lambda_0} + mh) + \frac{w}{2} \|\hat{g}_{\lambda_0} + mh\|^2 \right)$$

Taking the derivative of the criterion wrt m , we get

$$-\langle h, y - (\hat{g}_{\lambda_0} + mh) \rangle_T + \lambda \left(\frac{\partial}{\partial m} P^v(\hat{g}_{\lambda_0} + mh) + w \langle h, \hat{g}_{\lambda_0} + mh \rangle \right) \Big|_{m=\hat{m}_h(\lambda)} = 0$$

By implicit differentiation wrt λ , we have

$$\frac{\partial}{\partial \lambda} \hat{m}_h(\lambda) = - \left(\|h\|_T^2 + \lambda \frac{\partial^2}{\partial m^2} P^v(\hat{g}_{\lambda_0} + mh) + \lambda w \|h\|^2 \right)^{-1} \left(\frac{\partial}{\partial m} P^v(\hat{g}_{\lambda_0} + mh) + w \langle h, \hat{g}_{\lambda_0} + mh \rangle \right) \Big|_{m=\hat{m}_\lambda(\delta)}$$

To bound $|\frac{\partial}{\partial \lambda} \hat{m}_h(\lambda)|$, we bound each multiplicand.

1st multiplicand: Since penalty P is convex (regardless of the direction of h),

$$\begin{aligned} \left| \|h\|_T^2 + \lambda \frac{\partial^2}{\partial m^2} P^v(\hat{g}_{\lambda_0} + mh) + \lambda w \|h\|^2 \right|^{-1} &\leq \lambda w \|h\|^{-2} \\ &\lesssim \lambda^{-1} w^{-1} n^u \end{aligned}$$

2nd multiplicand: By definition of \hat{g}_{λ_0} ,

$$\lambda_0 P^v(\hat{g}_{\lambda_0} + mh) \leq \frac{1}{2} \|y - g^*\|_T^2 + \lambda_0 P^v(g^*)$$

Hence

$$\begin{aligned} P^{v-1}(\hat{g}_{\lambda_0} + mh) &\leq \left(\frac{1}{2\lambda_0} \|\epsilon\|_T^2 + P^v(g^*) \right)^{(v-1)/v} \\ &\lesssim (n^{\tau_{min}} \|\epsilon\|_T^2)^{(v-1)/v} \end{aligned}$$

3rd multiplicand: Note that since P is a semi-norm, then

$$|P(\hat{g}_\lambda + mh) - P(\hat{g}_\lambda)| \leq mP(h)$$

Therefore as we take $m \rightarrow 0$, we have

$$\left| \frac{\partial}{\partial m} P(\hat{g}_\lambda + mh) \right| \leq P(h) = 1$$

Also by Cauchy Schwarz and the assumption that $\sup_{g \in \mathcal{G}} \|g\| \leq G$, we have

$$|\langle h, \hat{g}_{\lambda_0} + mh \rangle| \leq \|h\| \|\hat{g}_{\lambda_0} + mh\| \leq n^p G$$

Combining the above bounds, we have

$$\begin{aligned} \left| \frac{\partial}{\partial \delta} \hat{m}_\lambda(\delta) \right| &\lesssim n^{\tau_{min}+u} w^{-1} \left(v (n^{\tau_{min}} \|\epsilon\|_T^2)^{(v-1)/v} + w n^p G \right) \\ &\leq n^{\tau_{min}(2v-1)/v+u} w^{-1} v \|\epsilon\|_T^{2(v-1)/v} + n^{\tau_{min}+u+p} G \end{aligned}$$

By the mean value theorem, there is some $\alpha \in [0, 1]$ s.t

$$\begin{aligned} \|\hat{g}_\lambda(\cdot|D_{-k}) - \hat{g}_{\lambda+\delta}(\cdot|D_{-k})\|_Q &= \hat{m}_\lambda(\delta) \|h\|_Q \\ &\leq n^p \delta \left| \frac{\partial}{\partial a} \hat{m}_\lambda(a) \right|_{a=\alpha\delta} \\ &\lesssim \delta \left(n^{\tau_{min}(2v-1)/v+u+p} w^{-1} v \|\epsilon\|_T^{2(v-1)/v} + n^{\tau_{min}+u+2p} G \right) \end{aligned}$$

Therefore the covering number is for some constant κ that depends on τ_{min}, u, p, v

$$N \left(d, \hat{\mathcal{G}}(T, \epsilon_T), \|\cdot\|_Q \right) \lesssim \frac{1}{d} n^\kappa w^{-1} v G \|\epsilon\|_T^{2(v-1)/v}$$

so the entropy is

$$H \left(d, \hat{\mathcal{G}}(T, \epsilon_T), \|\cdot\|_Q \right) \lesssim \log \left(\frac{1}{dw} \right) + \kappa \log n + \log \|\epsilon\|_T^2$$

2.1.2 Lemma 1

Define the convex combination $\hat{\xi}_\lambda(x) = \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \hat{g}_\lambda(x|D_{-k})$. Then

$$\frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D)\|_D^2 + \hat{\lambda} P(\hat{g}_\lambda(\cdot|D)) \geq \frac{1}{2} \|y - \hat{\xi}_\lambda\|_D^2 + \hat{\lambda} P(\hat{\xi}_\lambda) + \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \langle y - \hat{\xi}_\lambda, \hat{\xi}_\lambda - \hat{g}_\lambda(\cdot|D_{-k}) \rangle_{-k}$$

(This is Thrm 1 in Chetverikov, Chaterjee probably does the same thing.)

Proof

$$\begin{aligned} & \frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D)\|_D^2 + \hat{\lambda} P(\hat{g}_\lambda(\cdot|D)) \\ &= \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \left(\frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D)\|_{-k}^2 + \hat{\lambda} P(\hat{g}_\lambda(\cdot|D)) \right) \\ &\geq \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \left(\frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D_{-k})\|_{-k}^2 + \hat{\lambda} P(\hat{g}_\lambda(\cdot|D_{-k})) \right) \\ &\geq \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \left(\frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D_{-k})\|_{-k}^2 \right) + \hat{\lambda} P(\hat{\xi}_\lambda) \end{aligned}$$

The second inequality follows by convexity of P .

Now note that

$$\begin{aligned} \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D_{-k})\|_{-k}^2 &= \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \frac{1}{2} \|y - \hat{\xi}_\lambda + \hat{\xi}_\lambda - \hat{g}_\lambda(\cdot|D_{-k})\|_{-k}^2 \\ &\geq \frac{1}{2} \|y - \hat{\xi}_\lambda\|_D^2 + \frac{1}{K-1} \sum_{k=1}^K \frac{n-n_k}{n} \langle y - \hat{\xi}_\lambda, \hat{\xi}_\lambda - \hat{g}_\lambda(\cdot|D_{-k}) \rangle_{-k} \end{aligned}$$

2.1.3 Lemma 2

Consider any $\xi \in \mathcal{G}$ and $\lambda \in \Lambda$. Suppose P is convex.

Then

$$\frac{1}{2} \|\hat{g}_\lambda(\cdot|D) - \xi\|_D^2 \leq \frac{1}{2} \|y - \xi\|_D^2 - \frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D)\|_D^2 + \lambda P(\xi) - \lambda P(\hat{g}_\lambda(\cdot|D))$$

(This is Lemma 10 in Chetverikov, which is based on Chaterjee.)

Proof

Since P is convex, then for $t \in (0, 1)$, we have

$$\begin{aligned} & \frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D)\|_D^2 + \lambda P(\hat{g}_\lambda(\cdot|D)) \\ &\leq \frac{1}{2} \|y - (t\xi + (1-t)\hat{g}_\lambda(\cdot|D))\|_D^2 + \lambda P(t\xi + (1-t)\hat{g}_\lambda(\cdot|D)) \\ &\leq \frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D)\|_D^2 + t \langle y - \hat{g}_\lambda(\cdot|D), \hat{g}_\lambda(\cdot|D) - \xi \rangle_D + t^2 \|\xi - \hat{g}_\lambda\|_D^2 + \lambda (tP(\xi) + (1-t)P(\hat{g}_\lambda(\cdot|D))) \\ &\leq \frac{1}{2} \|y - \hat{g}_\lambda(\cdot|D)\|_D^2 + t \langle y - \hat{g}_\lambda(\cdot|D), \hat{g}_\lambda(\cdot|D) - \xi \rangle_D + \frac{t^2}{2} \|\xi - \hat{g}_\lambda\|_D^2 + \lambda (tP(\xi) + (1-t)P(\hat{g}_\lambda(\cdot|D))) \end{aligned}$$

Rearranging terms, we obtain

$$\lambda(P(\hat{g}_\lambda(\cdot|D)) - P(\xi)) \leq \langle y - \hat{g}_\lambda(\cdot|D), \hat{g}_\lambda(\cdot|D) - \xi \rangle_D + \frac{t}{2} \|\xi - \hat{g}_\lambda\|_D^2$$

Since this is true for any t , we have that

$$\lambda(P(\hat{g}_\lambda(\cdot|D)) - P(\xi)) \leq \langle y - \hat{g}_\lambda(\cdot|D), \hat{g}_\lambda(\cdot|D) - \xi \rangle_D$$

Thus

$$\begin{aligned} \frac{1}{2} \|\hat{g}_\lambda(\cdot|D) - \xi\|_D^2 &= \frac{1}{2} \|\hat{g}_\lambda(\cdot|D) - y + y - \xi\|_D^2 \\ &= \frac{1}{2} \|\hat{g}_\lambda(\cdot|D) - y\|_D^2 + \frac{1}{2} \|y - \xi\|_D^2 - \langle \hat{g}_\lambda(\cdot|D) - y, \xi - y \rangle_D \\ &= -\frac{1}{2} \|\hat{g}_\lambda(\cdot|D) - y\|_D^2 + \frac{1}{2} \|y - \xi\|_D^2 - \langle \hat{g}_\lambda(\cdot|D) - y, \xi - \hat{g}_\lambda(\cdot|D) \rangle_D \\ &\leq -\frac{1}{2} \|\hat{g}_\lambda(\cdot|D) - y\|_D^2 + \frac{1}{2} \|y - \xi\|_D^2 - \lambda(P(\hat{g}_\lambda(\cdot|D)) - P(\xi)) \end{aligned}$$

2.1.4 Lemma 3

Suppose X, T are random (or fixed) covariate values. X and T might overlap.

Suppose ϵ_X are independent sub-gaussian RVs with constants K and σ (corresponding to X). Same for ϵ_T . Again, ϵ_T and ϵ_X might have overlapping samples.

Suppose the (random) function class $\mathcal{F}(T, \epsilon_T)$ has its entropy uniformly bounded by $\psi(\cdot)$, as long as $\|\epsilon\|_T \leq \sigma$:

$$H(u, \mathcal{F}(T, \epsilon_T), \|\cdot\|_X) \leq \psi(u)$$

Suppose $\sup_{f \in \mathcal{F}(T, \epsilon_T)} \|f\|_X \leq R$.

Then there exists some C s.t. for all δ s.t.

$$\sqrt{|X|}\delta \geq C \left(\int_0^R \psi^{1/2}(u) du \vee 1 \right)$$

we have for some constant c

$$Pr_\epsilon \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle \epsilon, f \rangle_X|}{\|f\|_X} \geq \delta \wedge \|\epsilon\|_X \leq \sigma \wedge \|\epsilon\|_T \leq \sigma \right) \leq C \exp \left(-|X| \frac{\delta^2}{c^2} \right)$$

Proof

We use the peeling device. Let $S = \min\{s \in 0, 1, \dots : 2^s \delta > R\}$. Conditional on $\|\epsilon\|_X \leq \sigma$ and $\|\epsilon\|_T \leq \sigma$, we have

$$\begin{aligned} &Pr_\epsilon \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle \epsilon, f \rangle_X|}{\|f\|_X} \geq \delta \right) \\ &= \int 1 \left[\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle \epsilon, f \rangle_X|}{\|f\|_X} \geq \delta \right] dF(\epsilon) \\ &= \int \sum_{s=0}^S 1 \left[\sup_{f \in \mathcal{F}(T, \epsilon_T): 2^s \delta \leq \|f\|_X \leq 2^{s+1} \delta} \frac{|\langle \epsilon, f \rangle_X|}{\|f\|_X} \geq \delta \right] dF(\epsilon) \end{aligned}$$

$$\begin{aligned}
&= \sum_{s=0}^S Pr \left(\sup_{f \in \mathcal{F}(T, \epsilon_T): 2^s \delta \leq \|f\|_X \leq 2^{s+1} \delta} \frac{|\langle \epsilon, f \rangle_X|}{\|f\|_X} \geq \delta \right) \\
&\leq \sum_{s=0}^S Pr \left(\sup_{f \in \mathcal{F}(T, \epsilon_T): \|f\|_X \leq 2^{s+1} \delta} |\langle \epsilon, f \rangle_X| \geq 2^s \delta^2 \right)
\end{aligned}$$

In the last equality, we swap the order of integration and summation. This is allowed under the assumption that the identity functions are measurable, which should be okay.

To bound the summation, apply Lemma 10. For all

$$\sqrt{|X|} \delta \geq C \left(\int_0^R \psi^{1/2}(u) du \vee 1 \right)$$

there is some constant c s.t.

$$\begin{aligned}
Pr_\epsilon \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle \epsilon, f(\cdot|\epsilon) \rangle_X|}{\|f(\cdot|\epsilon)\|_X} \geq \delta \wedge \|\epsilon\|_X \leq \sigma \wedge \|\epsilon\|_T \leq \sigma \right) &\leq \sum_{s=0}^S C \exp \left(-|X| \frac{2^{2s} \delta^4}{C^2 2^{2s+2} \delta^2} \right) \\
&\leq c \exp \left(-|X| \frac{\delta^2}{c^2} \right)
\end{aligned}$$

2.1.5 Lemma 4

Suppose X, Z, T are random (or fixed) covariate values. X, Z, T might overlap.

Suppose ϵ_X are independent sub-gaussian RVs with constants K and σ (corresponding to X). Same for ϵ_T . Again, ϵ_T and ϵ_X might have overlapping samples.

Suppose the (random) function class $\mathcal{F}(T, \epsilon_T)$ has its entropy uniformly bounded by $\psi(\cdot)$, as long as $\|\epsilon\|_T \leq \sigma$:

$$H(u, \mathcal{F}(T, \epsilon_T), \|\cdot\|_{X \cup Z}) \leq \psi(u)$$

Suppose $\sup_{f \in \mathcal{F}(T, \epsilon_T)} \|f\|_X \leq R$.

Then there exists some C s.t. for all δ s.t.

$$(\min\{|X|, |Z|\})^{1/2} \delta \geq C \left(\int_0^R \psi^{1/2}(u) du \vee 1 \right)$$

we have

$$Pr \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\|f\|_X^2 - \|f\|_Z^2|}{\|f\|_{X \cup Z}} \geq \delta \wedge \|\epsilon\|_T \leq \sigma \right) \leq c \exp \left(-|X| \frac{\delta^2}{c^2 R^2} \right) + c \exp \left(-|X| \frac{\delta^2}{c^2 R^2} \right)$$

Proof

We use a symmetrization argument. Let W_i be Rademacher-like RV s.t. $Pr(W_i = 1) = \frac{|T|}{|T|+|X|}$ and $Pr(W_i = -\frac{|T|+|X|}{|T|}) = \frac{|X|}{|T|+|X|}$ (so $EW_i = 0$). Note that W_i are sub-gaussian by Hoeffding's inequality. Conditional on $\|\epsilon\|_T \leq \sigma$, we have

$$\begin{aligned}
&Pr_{X,T} \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\|f\|_X^2 - \|f\|_Z^2|}{\|f\|_{X \cup Z}} \geq \delta \right) \\
&= Pr_{W,X,T} \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle W, f^2 \rangle_X + \langle W, f^2 \rangle_T|}{\|f\|_{X \cup Z}} \geq \delta \right)
\end{aligned}$$

$$\leq Pr_{W,X,T} \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle W, f^2 \rangle_X|}{\|f\|_X} \geq \frac{\delta}{2} \frac{|X|}{|T| + |X|} \right) + Pr_{W,X,T} \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle W, f^2 \rangle_Z|}{\|f\|_Z} \geq \frac{\delta}{2} \frac{|T|}{|T| + |X|} \right)$$

Therefore we apply Lemma 3. (Note that the RVs W determine the model class $\hat{\mathcal{G}}(D_{-\ell})$, but Lemma 3 allows for this.)

Then there is a constant C such that for all

$$(\min\{|X|, |Z|\})^{1/2} \delta \geq C \left(\int_0^R \psi^{1/2}(u) du \vee 1 \right)$$

we have for some constant c (depends on the ratio $|X|/|Z|$),

$$Pr_{W,X,T} \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle W, f^2 \rangle_X|}{\|f\|_X} \geq \frac{\delta}{2} \frac{|X|}{|T| + |X|} \wedge \|\epsilon\|_T \leq \sigma \right) \leq c \exp \left(-|X| \frac{\delta^2}{c^2 R^2} \right)$$

and similarly

$$Pr_{W,X,T} \left(\sup_{f \in \mathcal{F}(T, \epsilon_T)} \frac{|\langle W, f^2 \rangle_Z|}{\|f\|_Z} \geq \frac{\delta}{2} \frac{|T|}{|T| + |X|} \wedge \|\epsilon\|_T \leq \sigma \right) \leq c \exp \left(-|Z| \frac{\delta^2}{c^2 R^2} \right)$$

2.1.6 Lemma 5

Let $\hat{\lambda}$ be chosen by CV. Let $\tilde{\lambda}$ be the oracle. Then with high probability,

$$\sqrt{\sum_{k=1}^K \|g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k})\|_k^2} \lesssim \left(\frac{1 + \log 4\sigma^2 - \log w + \kappa \log n}{n} \right)^{1/2} + \sqrt{\sum_{k=1}^K \|g_{\tilde{\lambda}}(\cdot|D_{-k}) - g^*\|_k^2}$$

Proof

The basic inequality gives us

$$\begin{aligned} & \sum_{k=1}^K \|g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k})\|_k^2 \\ & \leq 2 \left| \sum_{k=1}^K (\epsilon, g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k}))_k \right| + 2 \left| \sum_{k=1}^K (g^* - g_{\tilde{\lambda}}(\cdot|D_{-k}), g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k}))_k \right| \end{aligned}$$

We bound the empirical process term by a standard peeling argument (omitted). One can show that for all

$$\delta \geq C \left(\frac{1 + \log 4\sigma^2 - \log w + \kappa \log n}{n} \right)^{1/2}$$

we have for constants c_k (which depend on n_k/n),

$$\begin{aligned} & Pr \left(\frac{\left| \sum_{k=1}^K (\epsilon, g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k}))_k \right|}{\sqrt{\sum_{k=1}^K \|g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k})\|_k^2}} \geq \delta \wedge \|\epsilon\|_D \leq 2\sigma \right) \\ & \leq \sum_{k=1}^K Pr \left(\frac{\left| (\epsilon, g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k}))_k \right|}{\|g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k})\|_k} \geq \frac{\delta c_k}{k} \wedge \|\epsilon\|_D \leq 2\sigma \right) \\ & \leq c \exp \left(- \min_{k=1:K} \{n_k\} \frac{\delta^2}{c^2} \right) \end{aligned}$$

Also, by Cauchy-Schwarz, we have

$$\frac{\left| \sum_{k=1}^K (g^* - g_{\tilde{\lambda}}(\cdot|D_{-k}), g_{\tilde{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k}))_k \right|}{\sqrt{\sum_{k=1}^K \|g_{\tilde{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k})\|_k^2}} \leq \sqrt{\sum_{k=1}^K \|g^* - g_{\tilde{\lambda}}(\cdot|D_{-k})\|_k^2}$$

Hence the result follows.

2.1.7 Lemma 10

Let X be n covariate values (potentially randomly drawn). It is also possible for T and X to contain overlapping samples.

Suppose ϵ_X is a set of n independent sub-gaussian RVs with constants K and σ (corresponding to X). Suppose ϵ_T are also independent sub-gaussian RVs with constants K and σ (corresponding to T). Again, it is possible for ϵ_T and ϵ_X can have overlapping samples.

Suppose $\sup_{f \in \mathcal{F}(T, \epsilon_T)} \|f\|_X \leq R$.

Suppose we have (random) function classes $\mathcal{F}(T, \epsilon_T)$ with entropy $H(\delta, \mathcal{F}(T, \epsilon_T), \|\cdot\|_X)$. Suppose that there is a universal bound on the entropy if $\|\epsilon_T\| \leq \sigma$:

$$H(u, \mathcal{F}(T, \epsilon_T), \|\cdot\|_X) \leq \psi(u)$$

Then there exists some C dependent only on K, σ s.t. for all

$$\sqrt{n}\delta \geq C \left(\int_0^R \psi^{1/2}(u) du \vee R \right)$$

we have

$$Pr_{\epsilon} \left(\sup_{f_{\theta}(\cdot|\epsilon) \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_{\theta}(\cdot|\epsilon) \rangle_X| \geq \delta \wedge \|\epsilon\|_X \leq \sigma \wedge \|\epsilon\|_T \leq \sigma \right) \leq C \exp \left(-n \frac{\delta^2}{C^2 R^2} \right)$$

Proof

Proof closely follows Lemma 3.2 from Vandegeer.

For a given set of RVs ϵ , let $\{f_j^s(\cdot|\epsilon)\}_{j=1}^{N_s}$ be the $2^{-s}R$ -covering set of $\mathcal{F}(T, \epsilon)$ where $N_s = N_s(2^{-s}R, \mathcal{F}(T, \epsilon), \|\cdot\|_X) \leq \exp(\psi(2^{-s}R))$. Let $S = \min\{s : 2^{-s}R \leq \delta/2\sigma\}$. We can write $f_{\theta}^S(\cdot|\epsilon) = \sum_{s=1}^S f_{\theta}^s(\cdot|\epsilon) - f_{\theta}^{s-1}(\cdot|\epsilon)$ where $f_{\theta}^0(\cdot|\epsilon) = 0$.

Also, consider a set of constants η_s s.t. $\sum_{s=1}^S \eta_s \leq 1$.

Then as long as $\|\epsilon\|_X \leq \sigma$ and $\|\epsilon\|_T \leq \sigma$, we have

$$\begin{aligned} & Pr_{\epsilon} \left(\sup_{\theta \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_{\theta}(\cdot|\epsilon) \rangle_X| \geq \delta \right) \\ &= \int 1 \left[\sup_{\theta \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_{\theta}(\cdot|\epsilon) \rangle_X| \geq \delta/2 \right] dF(\epsilon) \\ &= \int 1 \left[\sup_{\theta \in \mathcal{F}(T, \epsilon_T)} \left| \left\langle \epsilon, f_{\theta}(\cdot|\epsilon) - f_{\theta}^S(\cdot|\epsilon) + \sum_{s=1}^S f_{\theta}^s(\cdot|\epsilon) - f_{\theta}^{s-1}(\cdot|\epsilon) \right\rangle_X \right| \geq \delta/2 \right] dF(\epsilon) \\ &\leq \int 1 \left[\sup_{\theta \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_{\theta} - f_{\theta}^S(\cdot|\epsilon) \rangle_X| \geq \delta/2 \right] + \sum_{s=1}^S \int 1 \left[\sup_{\theta \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_{\theta}^s(\cdot|\epsilon) - f_{\theta}^{s-1}(\cdot|\epsilon) \rangle_X| \geq \delta\eta_s/2 \right] dF(\epsilon) \\ &= Pr_{\epsilon} \left(\sup_{\theta \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_{\theta} - f_{\theta}^S(\cdot|\epsilon) \rangle_X| \geq \delta/2 \right) + \sum_{s=1}^S Pr_{\epsilon} \left(\sup_{\theta \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_{\theta}^s(\cdot|\epsilon) - f_{\theta}^{s-1}(\cdot|\epsilon) \rangle_X| \geq \delta\eta_s/2 \right) \end{aligned}$$

In the last equality, we swap the order of the summation and integration, which is allowed under the assumption that identity functions are measurable (I think this is the measurability assumption required here) (If ϵ has a continuous probability measure, I think this holds).

We know that the first summand is zero since by Cauchy-Schwarz,

$$\begin{aligned} |\langle \epsilon, f_\theta - f_\theta^S(\cdot|\epsilon) \rangle_X| &\leq \sigma \|f_\theta - f_\theta^S(\cdot|\epsilon)\|_X \\ &\leq \delta/2 \end{aligned}$$

Also, for any ϵ , we must have $\|f_\theta^s(\cdot|\epsilon) - f_\theta^{s-1}(\cdot|\epsilon)\| \leq 3(2^{-s}R)$. Furthermore, since ϵ is sub-gaussian,

$$Pr_\epsilon \left(\sup_{\theta \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_\theta^s(\cdot|\epsilon) - f_\theta^{s-1}(\cdot|\epsilon) \rangle_X| \geq \delta \eta_s/2 \right) \leq \exp \left(2\psi(2^{-s}R) - C \frac{n(\delta/2)^2 \eta_s^2}{9(2^{-2s}R^2)} \right)$$

Now choose η_s as Vandegeer does in Lemma 3.2. After a lot of algebraic massaging, we get that for some constants C_1, C_2

$$Pr_\epsilon \left(\sup_{f_\theta \in \mathcal{F}(T, \epsilon_T)} |\langle \epsilon, f_\theta \rangle_X| \geq \delta \wedge \|\epsilon\|_X \leq \sigma \wedge \|\epsilon\|_T \leq \sigma \right) \leq C_1 \exp \left(-n \frac{\delta^2}{C_2^2 R^2} \right)$$

2.2 Easy algebra notes

2.2.1 Easy Lemma 1

Suppose

$$\frac{|\|f\|_X^2 - \|f\|_Z^2|}{\|f\|_{X \cup Z}} \leq \delta$$

then

$$|\|f\|_X - \|f\|_Z| \leq \delta$$

Proof

We know

$$\begin{aligned} \sqrt{\|f\|_X^2 + \|f\|_Z^2} |\|f\|_X - \|f\|_Z| &\leq |\|f\|_X + \|f\|_Z| |\|f\|_X - \|f\|_Z| \\ &= |\|f\|_X^2 - \|f\|_Z^2| \\ &\leq \delta \sqrt{\frac{|X|}{|X| + |Z|} \|f\|_X^2 + \frac{|Z|}{|X| + |Z|} \|f\|_Z^2} \\ &\leq \delta \sqrt{\|f\|_X^2 + \|f\|_Z^2} \end{aligned}$$

hence

$$|\|f\|_X - \|f\|_Z| \leq \delta$$