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. For $k = 1, ..., K$, let D_k represent the kth fold and D_{-k} denote all the folds minus the kth 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.

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 $(h,g)_k = \frac{1}{n_k}\sum_{i\in D_k} h(x_i)g(x_i)$ and $(h,g)_{-k}$ for the set D_{-k} and $(h,g)_D$ for the set D.

Let us define

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

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

The K-fold CV model is

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

Let the range of $\Lambda = [\lambda_{min}, \lambda_{max}]$. Both limits can grow and shrink at any polynomial rate, e.g. $\lambda_{max} = O_P(1)$ and $\lambda_{min} = O_P(n^{-\tau_{min}})$.

We show that

$$\|\hat{g}_{\hat{\lambda}}(\cdot|D) - g^*\|_D \le ? + \sum_{k=1}^K \|g^* - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k + \|\hat{g}_{\tilde{\lambda}}(\cdot|D) - g^*\|_D$$

Notation

$a \lesssim b$ means that $a \le Cb + c$ 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, the additional ridge regression penalty allows us to bound the entropy of $\hat{\mathcal{G}}_k = \{\hat{g}_\lambda(\cdot|D_{-k}) : \lambda\in\Lambda\}$. Once this is complete, we can use results similar to Vandegeer that bound the Rademacher process

$$\sum_{i=1}^n W_i \hat{g}_\lambda(x_i|D_{-k})$$

and the empirical process

$$\sum_{i=1}^n \epsilon_i \hat{g}_\lambda(x_i|D_{-k})$$

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,

$$\begin{aligned}
\|\hat{g}_{\hat{\lambda}}(\cdot|D) - g^*\|_D &\leq \|\hat{g}_{\hat{\lambda}}(\cdot|D) - \xi_{\hat{\lambda}}\|_D + \|\xi_{\hat{\lambda}} - \xi_{\tilde{\lambda}}\|_D + \|\xi_{\tilde{\lambda}} - \hat{g}_{\tilde{\lambda}}(\cdot|D)\|_D + \|\hat{g}_{\tilde{\lambda}}(\cdot|D) - g^*\|_D \\
&\leq \|\hat{g}_{\hat{\lambda}}(\cdot|D) - \xi_{\hat{\lambda}}\|_D + \|\xi_{\tilde{\lambda}} - \hat{g}_{\tilde{\lambda}}(\cdot|D)\|_D + \frac{1}{K-1}\sum_{k=1}^{K} \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_D + \|\hat{g}_{\tilde{\lambda}}(\cdot|D) - g^*\|_D
\end{aligned}$$

We bound the first two terms in step 2. We bound the third term in step 3.

Step 2:

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

$$\|\hat{g}_\lambda(\cdot|D) - \hat{\xi}_\lambda\|_D^2 \leq \frac{1}{K-1}\sum \frac{n-n_k}{n}\left(\left|\langle \epsilon, \hat{\xi}_\lambda - \hat{g}_\lambda(\cdot|D_{-k})\rangle_{-k}\right| + \left|\langle g^* - \hat{\xi}_\lambda, \hat{\xi}_\lambda - \hat{g}_\lambda(\cdot|D_{-k})\rangle_k\right|\right)$$

The first term is bounded by

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

By Lemma 4, we know that $\sup_\lambda \|\hat{g}_\lambda(\cdot|D_{-\ell}) - g^*\| \leq F\sigma$. Hence by Lemma 3, we have that for all $\delta \geq CR\sqrt{J}\left(\frac{1+\log(C/\sqrt{w})+\kappa\log n}{n-n_k}\right)^{1/2}$, we have for all $\ell = 1:K$ and $k = 1:K$,

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

The second term is bounded by

$$\begin{aligned}
&\left|\langle g^* - \hat{\xi}_\lambda, \hat{\xi}_\lambda - \hat{g}_\lambda(\cdot|D_{-k})\rangle_k\right| \\
\leq\; &\left|\left\langle \frac{1}{K-1}\sum_{\ell=1}^{K}\frac{n-n_\ell}{n}\left(g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\right), \frac{1}{K-1}\sum_{\ell=1}^{K}\frac{n-n_\ell}{n}\left(g^* - \hat{\xi}_\lambda\right) + \hat{g}_\lambda(\cdot|D_{-k}) - g^*\right\rangle_k\right| \\
\lesssim\; &\frac{1}{K-1}\sum_{\ell=1}^{K}\frac{n-n_k}{n}\|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_{-k}^2 \\
\leq\; &\frac{1}{K-1}\sum_{\ell=1}^{K}\frac{n-n_k}{n}\sum_{h\neq k}\frac{n_h}{n-n_k}\|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_h^2 \\
\lesssim\; &\sum_{\ell=1}^{K}\|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_\ell^2 + \left(\|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_h^2 - \|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_\ell^2\right)
\end{aligned}$$

We will use a symmetrization argument to bound the term in the parenthesis. Let W_i be RV s.t. $Pr(W_i = 1) = \frac{n_h}{n_h+n_k}$ and $Pr(W_i = -\frac{n_h+n_k}{n_k}) = \frac{n_k}{n_h+n_k}$ (so $EW_i = 0$). We have

$$\begin{aligned}
&Pr_{X_h,X_\ell}\left(\|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_h^2 - \|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_\ell^2 \geq \delta\right) \\
\leq\; &2Pr_{W,X_h,X_\ell}\left(\sup_\lambda \frac{1}{n_h+n_\ell}\sum_{i\in D_h\cup D_\ell} W_i\left(g^*(x_i) - \hat{g}_\lambda(x_i|D_{-\ell})\right)^2 \geq \delta/2\right)
\end{aligned}$$

where the second inequality follows from a symmetrization argument (check this!)

Since W_i are sub-gaussian, we can apply Lemma 3 again. For all $\delta \geq CR\sqrt{J}\left(\frac{1+\log(C/\sqrt{w})+\kappa\log n}{n_h+n_\ell}\right)^{1/2}$, we have for all $\ell = 1 : K$ and $h = 1 : K$, (and some constants C, R)

$$Pr_{X_h,X_\ell}\left(\sup_{\lambda}\left|\|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_h^2 - \|g^* - \hat{g}_\lambda(\cdot|D_{-\ell})\|_\ell^2\right| \geq \delta\right) \leq C\exp\left(-(n_h+n_\ell)\frac{\delta^2}{C^2R^2}\right)$$

Hence for all $\lambda \in \Lambda$, we have with high probability that

$$\|\hat{g}_\lambda(\cdot|D) - \hat{\xi}_\lambda\|_D^2 \lesssim \sum_{\ell=1}^{K}\left\|g^* - \hat{g}_{\tilde{\lambda}}(\cdot|D_{-\ell})\right\|_\ell^2 + \max_{h,\ell} CR\sqrt{J}\left(\frac{1+\log(C/\sqrt{w})+\kappa\log n}{n_h+n_\ell}\right)^{1/2}$$

Step 3:

For every k, we have

$$\begin{aligned}
& \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_D^2 \\
\leq\ & \sum_{\ell\neq k}\frac{n_\ell}{n}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_\ell^2 + \frac{n_k}{n}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2 \\
\leq\ & \sum_{\ell\neq k} 2\frac{n_k}{n}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2 + \left(\frac{n_\ell}{n}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_\ell^2 - \frac{n_k}{n}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2\right)
\end{aligned}$$

Using Lemma 3 (and the same arguments given above in Step 2), we get that with high probability,

$$\left|\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_\ell^2 - \|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2\right| \lesssim CR\sqrt{J}\left(\frac{1+\log(C/\sqrt{w})+\kappa\log n}{n_h+n_\ell}\right)$$

Also, by the definition of $\hat{\lambda}$, the basic inequality gives us that

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

If the first sum on the RHS (the empirical process term) is bigger, then from the same arguments in Step 2, we can bound with high probability that

$$\sum_{k=1}^{K}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2 \leq CR\sqrt{J}\left(\frac{1+\log(C/\sqrt{w})+\kappa\log n}{n_h+n_\ell}\right)^{1/2} (whp)$$

If the second sum on the RHS is bigger, note that by Cauchy-Schwarz

$$\begin{aligned}
\sum_{k=1}^{K}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2 &\lesssim \sum_{k=1}^{K}\left|\left\langle g^* - \hat{g}_{\tilde{\lambda}}(x|D_{-k}), \hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\right\rangle_k\right| \\
&\leq \sqrt{\left(\sum_{k=1}^{K}\|g^* - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2\right)\left(\sum_{k=1}^{K}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2\right)}
\end{aligned}$$

Hence with high probability,

$$\sum_{k=1}^{K}\|\hat{g}_{\hat{\lambda}}(x|D_{-k}) - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2 \leq \sum_{k=1}^{K}\|g^* - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k^2 + CR\sqrt{J}\left(\frac{1+\log(C/\sqrt{w})+\kappa\log n}{n_h+n_\ell}\right)^{1/2}$$

Step 4

Combining the above steps, we have shown that with high probability,

$$\|\hat{g}_{\hat{\lambda}}(\cdot|D) - g^*\|_D \leq \max_{h,\ell} CR\sqrt{J}\left(\frac{1+\log(C/\sqrt{w})+\kappa\log n}{n_h+n_\ell}\right)^{1/2} + \sum_{k=1}^{K}\|g^* - \hat{g}_{\tilde{\lambda}}(x|D_{-k})\|_k$$

Therefore CV converges at a parametric rate modulo a log term plus the optimal rates of convergence.

2.1 Lemmas

2.1.1 Lemma 0

Consider any empirical distributions Q_1 and Q_2.

Suppose the penalty function is smooth. Suppose $O_p(n^{-u}) = \min_{h:P(h)=1}\|h\|_{Q_1}^2$ and for all h,$\|h\|_{Q_1} \leq O_p(n^v)P(h)$ and $\|h\|_{Q_2} \leq O_p(n^v)P(h)$. Suppose $\lambda_{min} = O_P(n^{-\tau_{min}})$ and $\lambda_{max} = O_P(n^{\tau_{max}})$.

Consider the function class

$$\hat{\mathcal{G}} = \{\hat{g}_\lambda(\cdot|Q_1) : \lambda \in \Lambda\}$$

We have that the entropy is bounded at a near-parametric rate:

$$H\left(u, \hat{\mathcal{G}}, \|\cdot\|_{Q_2}\right) \leq \log\left(\frac{C}{u\sqrt{w}}\right) + \kappa\log n$$

Proof

To find the covering number for $\hat{\mathcal{G}}$, we bound the distance$\|\hat{g}_\lambda(\cdot|Q_1) - \hat{g}_{\lambda+\delta}(\cdot|Q_1)\|_{Q_2}$ for every $\lambda \in \Lambda$.

Consider the function $h = c\left(\hat{g}_\lambda - \hat{g}_{\lambda+\delta}\right)$ where $c > 0$ is some constant s.t. $P(h) = 1$. Consider the 1-dimensional optimization problem

$$\hat{m}(\lambda+\delta) = \arg\min_m \frac{1}{2}\|y - (\hat{g}_\lambda + mh)\|_{Q_1}^2 + (\lambda+\delta)\left(P(\hat{g}_\lambda + mh) + \frac{w}{2}\|\hat{g}_\lambda + mh\|_{Q_1}^2\right)$$

Clearly $\hat{m}_\lambda = 0$ and $\hat{m}_{\lambda+\delta} = c^{-1}$.

Taking the derivative of the criterion wrt m, we get

$$-\langle h, y - (\hat{g}_\lambda + mh)\rangle_{Q_1} + \lambda\left(\frac{\partial}{\partial m}P(\hat{g}_\lambda + mh) + w\langle h, g + mh\rangle_{Q_1}\right) = 0$$

By implicit differentiation wrt δ, we have

$$\frac{\partial}{\partial\delta}\hat{m}(\lambda+\delta) = -\left(\|h\|_T^2 + \lambda\frac{\partial^2}{\partial m^2}P\left(\hat{g}_\lambda + mh\right) + \lambda w\|h\|_{Q_1}^2\right)^{-1}\left(\frac{\partial}{\partial m}P(\hat{g}_\lambda + mh) + w\langle h, \hat{g}_\lambda + mh\rangle_{Q_1}\right)\Bigg|_{m=\hat{m}(\lambda+\delta)}$$

To bound $|\frac{\partial}{\partial\delta}\hat{m}(\lambda+\delta)|$, consider each multiplicand.

Since penalty P is convex (regardless of the direction of h), the first multiplicand is bounded by

$$\begin{aligned}\left|\|h\|_T^2 + \lambda\frac{\partial^2}{\partial m^2}P\left(\hat{g}_\lambda + mh\right) + \lambda w\|h\|_{Q_1}^2\right|^{-1} &\leq \left(\lambda w\|h\|_{Q_1}^2\right)^{-1} \\ &\leq \lambda^{-1}w^{-1}O_P(n^u)\end{aligned}$$

For the second multiplicand, note that

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

and with high probability,

$$\begin{aligned} w\langle h, \hat{g}_\lambda + mh\rangle_{Q_1} &\le w\|h\|_{Q_1}\|\hat{g}_\lambda + mh\|_{Q_1} \\ &\le O_P(n^v)w\sqrt{(\lambda w)^{-1}4\sigma^2 + w^{-1}P(g^*) + \|g^*\|^2_{Q_1}} \end{aligned}$$

where we have bounded $\|g + m_\lambda h\|_T$ using the definition

$$\frac{\lambda w}{2}\|g + m_\lambda h\|_T^2 \le \frac{1}{2}\|y - g^*\|_T^2 + \lambda P(g^*) + \frac{\lambda w}{2}\|g^*\|^2_{Q_1}$$

Hence there is a constant C that only depends on $g^*$and σ s.t.

$$\begin{aligned} \left|\frac{\partial}{\partial\delta}\hat{m}(\lambda + \delta)\right| &= \lambda^{-1}w^{-1}O_P(n^u)\left|1 + O_P(n^v)w\sqrt{(\lambda w)^{-1}4\sigma^2 + w^{-1}P(g^*) + \|g^*\|^2_{Q_1}}\right| \\ &\le \lambda_{min}^{-1}O_P(n^{u+v})\sqrt{(\lambda_{min}w)^{-1}4\sigma^2 + w^{-1}P(g^*) + \|g^*\|^2_{Q_1}} \\ &\le \frac{O_P(n^{1.5\tau_{min}+u+v})}{\sqrt{w}}C \end{aligned}$$

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

$$\begin{aligned} \|\hat{g}_\lambda(\cdot|D_{-k}) - \hat{g}_{\lambda+\delta}(\cdot|D_{-k})\|_{Q_2} &= \hat{m}(\lambda + \delta)\|h\|_{Q_2} \\ &\le n^{-v}\delta\left|\frac{\partial}{\partial u}\hat{m}(\lambda + u)\right|_{u=\alpha\delta} \\ &\le \delta\frac{C}{\sqrt{w}}O_P(n^{1.5\tau_{min}+u}) \end{aligned}$$

Therefore there is a constant κ that linearly grows with $u, \tau_{min}, \tau_{max}$ s.t. the covering number is

$$N\left(u, \hat{\mathcal{G}}, \|\cdot\|_{Q_2}\right) \le \frac{C}{u\sqrt{w}}O_p(n^\kappa)$$

so the entropy is

$$H\left(u, \hat{\mathcal{G}}, \|\cdot\|_{Q_2}\right) \le \log\left(\frac{C}{u\sqrt{w}}\right) + \kappa\log n$$

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

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

(This is a version of the beginning of the proof for Thrm 1 in Chetverikov, Chaterjee probably does the same thing.)

Proof

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

The second inequality follows by convexity of P and $\|\cdot\|^2$.

Now note that

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

2.1.3 Lemma 2

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

Then

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

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

Proof

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

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

Rearranging terms, we obain

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

Since this is true for any t, we have that

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

Thus

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

2.1.4 Lemma 3

Suppose ϵ are independent sub-gaussian RV with constants K and σ.

Suppose X are n random (or fixed) covariate values.

Suppose for any empirical distribution Q, the (random) function class $\mathcal{F}(X,\epsilon)$ has its entropy uniformly bounded

$$H\left(u, \mathcal{F}(X,\epsilon), \|\cdot\|_Q\right) \leq \psi(u) = J\left(\log\left(\frac{C}{u\sqrt{w}}\right) + \kappa \log n\right)$$

for positive constants J, C, w, κ.

Suppose $\sup_{f\in\mathcal{F}(X,\epsilon)} \|f\|_Q \leq R$.

Then there exists some C s.t. for all δ s.t. $R \geq \delta/\sigma$ and

$$\delta \geq CR\sqrt{J}\left(\frac{1 + \log(C/\sqrt{w}) + \kappa \log n}{|Q|}\right)^{1/2}$$

we have

$$Pr\left(\sup_{f\in\mathcal{F}(X,\epsilon), \|f\|_Q \leq R} |\langle \epsilon, f\rangle_Q| \geq \delta \wedge \|\epsilon\|_Q \leq \sigma\right) \leq C\exp\left(-|Q|\frac{\delta^2}{C^2R^2}\right)$$

Proof

$$\langle \epsilon, \hat{g}_\lambda(\cdot|D_{-\ell}) - g^*\rangle_{-k}$$

We apply Lemma 10 in Vandegeer to determine the value δ s.t. δ bounds the empirical process term with high probability.

For $R \geq \delta/\sqrt{2}\sigma$,

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

Apply Lemma 10 to $\delta > 0$ s.t.

$$\delta \geq CR\sqrt{J}\left(\frac{1 + \log(C/\sqrt{w}) + \kappa \log n}{|Q|}\right)^{1/2}$$

2.1.5 Lemma 4

Suppose we are working within a restricted domain, so there is some constant R s.t. $\|g^*\|_\infty \le R$.

Suppose ϵ is sub-gaussian with constants K, σ.

Then for some constant C that depends on λ_{max} and g^*, we have

$$Pr\left(\sup_\lambda \|\hat{g}_\lambda(\cdot|D_{-k}) - g^*\|_{-k} \ge 2\sigma + C\right) \le \exp\left(-(n-n_k)\frac{\sigma^2}{12K^2}\right)$$

Proof

By triangle inequality

$$\|g^* - \hat{g}_\lambda(\cdot|D_{-k})\|_{-k} \le \|y - \hat{g}_\lambda(\cdot|D_{-k})\|_{-k} + \|y - g^*\|_{-k}$$

By definition of $\hat{g}_\lambda$, we have

$$\begin{aligned}
\|y - \hat{g}_\lambda(\cdot|D_{-k})\|_{-k}^2 &\le \|y - g^*\|_{-k}^2 + \lambda\left(P(g^*) + \frac{w}{2}\|g^*\|_{-k}^2\right) \\
&\le \|y - g^*\|_{-k}^2 + \lambda_{max} P(g^*) + \frac{w}{2} R
\end{aligned}$$

Since ϵ is sub-gaussian, then by Bernstein's inequality, we have

$$Pr\left(\|\epsilon\|_{-k}^2 \ge 2\sigma^2\right) \le \exp\left(-(n-n_k)\frac{\sigma^2}{12K^2}\right)$$

2.1.6 Lemma 10

Suppose ϵ are n independent sub-gaussian RVs with constants K and σ.

Let X be n covariate values (potentially randomly drawn).

Suppose that we have function classes $\mathcal{F}(X,\epsilon)$ dependent on the sub-gaussian RV with entropy $H(\delta, \mathcal{F}(X,\epsilon), \|\cdot\|_X)$. Suppose there is a universal bound

$$H(u, \mathcal{F}(X,\epsilon), \|\cdot\|_X) \le \psi(u)$$

Suppose $\sup_{f\in\mathcal{F}(X,\epsilon)} \|f\|_X \le R$ (with high probability).

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

$$\delta \ge \int_0^R \psi^{1/2}(u)du$$

we have

$$Pr_\epsilon\left(\sup_{f_\theta\in\mathcal{F}(X,\epsilon)} |\langle \epsilon, f_\theta\rangle_X| \ge \delta \wedge \|\epsilon\|_X \le \sigma\right) \le C\exp\left(-|X|\frac{\delta^2}{C^2R^2}\right)$$

Proof

Proof closely follows Lemma 3.2 from Vandegeer.

Let $\{f_j^s(\cdot|\epsilon)\}_{j=1}^{N_s}$ be the $2^{-s}R$-covering set of $\mathcal{F}(X,\epsilon)$ where $N_s = N_s(2^{-s}R, \mathcal{F}(X,\epsilon), X) \le \exp(\psi(2^{-s}R))$. Let $S = \min\{s : 2^{-s}R \le \delta/2\sigma\}$

Let $f_\theta^s(\cdot|\epsilon)$ be the closest element to f_θ in the $2^{-s}R$-covering set. If $\|\epsilon\|_X \le \sigma$, then

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

Therefore it suffices to bound

$$Pr_\epsilon\left(\sup_{j=1:N_S}\left|\langle\epsilon, f_j^s(\cdot|\epsilon)\rangle_X\right| \geq \delta/2 \wedge \|\epsilon\|_X \leq \sigma\right)$$

Let's chain! Let $f_\theta^S(\cdot|\epsilon) = \sum_{s=1}^S f_\theta^s(\cdot|\epsilon) - f_\theta^{s-1}(\cdot|\epsilon)$. Note that

$$\begin{aligned}\|f_\theta^s(\cdot|\epsilon) - f_\theta^{s-1}(\cdot|\epsilon)\|_X &\leq \|f_\theta^s(\cdot|\epsilon) - f_\theta\|_X + \|f_\theta - f_\theta^{s-1}(\cdot|\epsilon)\|_X \\ &\leq 3(2^{-s}R)\end{aligned}$$

Then for some positive numbers s.t. $\sum_{s=1}^S \eta_s \leq 1$, we have

$$\begin{aligned}& Pr_\epsilon\left(\sup_{j=1:N_s}\left|\langle\epsilon, \sum_{s=1}^S f_\theta^s(\cdot|\epsilon) - f_\theta^{s-1}(\cdot|\epsilon)\rangle_X\right| \geq \delta/2\right) \\ \leq & \sum_{s=1}^S Pr_\epsilon\left(\sup_{j=1:N_s}\left|\langle\epsilon, f_\theta^s(\cdot|\epsilon) - f_\theta^{s-1}(\cdot|\epsilon)\rangle_X\right| \geq \delta/2\eta_s\right) \\ \leq & \sum_{s=1}^S \exp\left(2\psi(2^{-s}R) - C\frac{n(\delta/2)^2\eta_s^2}{9(2^{-2s}R^2)}\right)\end{aligned}$$

Choose η_s as Vandegeer does. Then after a lot of algebraic massaging, we get that for some constants C_1, C_2

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

2.2 OLD

2.2.1 Lemma 3

Consider the function class $\mathcal{F}$ with entropy bound

$$H\left(u, \mathcal{F}, \|\cdot\|_Q\right) \leq J\left(\log\left(\frac{C}{u\sqrt{w}}\right) + \kappa\log n\right)$$

We will suppose that $\sup_{f\in\mathcal{F}}\|f\|_Q \leq F\sigma$. (check this!!!)
Suppose ϵ are independent sub-gaussian RV with constants K and σ.
We get that there exists some C s.t. for all δ s.t. $R \geq \delta/\sigma$ and

$$\delta \geq CR\sqrt{J}\left(\frac{1+\log(C/\sqrt{w}) + \kappa\log n}{|Q|}\right)^{1/2}$$

we have

$$Pr\left(\sup_{f_1:\|f_1\|_Q\leq R}|\langle\epsilon, f_1\rangle_Q| \geq \delta \wedge \|\epsilon\|_Q \leq 2\sigma\right) \leq C\exp\left(-|Q|\frac{\delta^2}{C^2R^2}\right)$$

Proof

Using Lemma $2\frac{3}{4}$, we bound the empirical process term by a standard chaining argument (basically a copy of Thrm 9.1 in Vandegeer).

Let $S = \min\{s \in \{0,1,...\} : 2^s > F\sigma\}$. For

$$\delta \geq 16C\left(\frac{1+\log(C/\sqrt{w})+\kappa\log n}{|Q|}\right)^{1/2}$$

we have

$$\begin{aligned}
& Pr\left(\sup_{f_1:\|f_1\|_Q\leq F}|\langle\epsilon, f_1\rangle_Q| \geq \delta^2 \wedge \|\epsilon\|_Q \leq 2\sigma\right) \\
\leq\ & \sum_{s=0}^{S} Pr\left(\sup_{f_1:\|f_1\|_Q\leq 2^{s+1}\delta}|\langle\epsilon, f_1\rangle_Q| \geq 2^{2s-1}\delta^2 \wedge \|\epsilon\|_Q \leq 2\sigma\right) \\
\leq\ & \sum_{s=0}^{S} C\exp\left(-|Q|\frac{2^{4s-2}\delta^4}{4C^2 2^{2s+2}\delta^2}\right) \\
\leq\ & C\exp\left(-|Q|\frac{\delta^2}{c^2}\right)
\end{aligned}$$

Note that Lemma $2\frac{3}{4}$ can be applied since for $s = 0, ..., S$,

$$\sqrt{|Q|}2^{2s+2}\delta^2 \geq 16C2^{s+1}\delta\left(1+\log(C/\sqrt{w})+\kappa\log n\right)^{1/2}$$