# 1 K-Fold Cross-Validation

We are interested in modeling the conditional relationship between $y$ and predictors $X$ using function class $\mathcal{G}$. Suppose $X$ is from some bounded domain. Suppose $\mathcal{G}$ is a convex function class. There is some constant $G$ s.t. $\sup_{g\in\mathcal{G}}\|g\|^2 \le G < \infty$ (this envelope condition makes the proof simpler, though I think we can easily drop it with some regularity assumptions?).

We fit the function using some penalty function $P$ taken to the power $v \ge 1$. Suppose $P$ is a semi-norm, smooth, and convex.

Consider the joint optimization problem to find the best penalty parameter $\lambda$ in $\Lambda$ via $K$-fold cross validation. Let $D$ be the entire dataset of $n$ observations. For $k = 1, ..., K$, let $D_k$ represent the $k$th fold with $n_k$ observations and $D_{-k}$ denote all the folds minus the $k$th fold.

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 $D$.

We perform $K$-fold cross-validation with a small additional ridge penalty.

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

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

The model chosen via $K$-fold CV is

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

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

We show that

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

### Notation

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

# 2 Proof

### 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,

$$\|\hat g_{\hat\lambda}(\cdot|D) - g^*\|_D \le \|\hat g_{\hat\lambda}(\cdot|D) - \hat\xi_{\hat\lambda}\|_D + \|\hat\xi_{\hat\lambda} - g^*\|_D$$

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

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

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

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

The Lemmas below allow us to bound each of the summands in the inequality. Note that all the Lemmas have probability statements that condition on $\|\epsilon\|_k \leq 2\sigma$. Recall that since $\epsilon$ is sub-gaussian, Bernstein's inequality states that $\|\epsilon\|_k \leq 2\sigma$ indeed occurs with high probability:

$$Pr\left( \|\epsilon\|_k \geq 2\sigma \right) \leq \exp\left( -n_k \frac{\sigma^2}{c} \right)$$

Now to bound the first term, Lemma 5 states that with high probability

$$\sum_{k=1}^{K} \left\| g^* - \hat{g}_{\hat{\lambda}}(\cdot|D_{-k}) \right\|_k^2 \lesssim \sum_{k=1}^{K} \left\| g^* - \hat{g}_{\tilde{\lambda}}(\cdot|D_{-k}) \right\|_k^2 + \delta^2$$

To bound the last two terms, we first note that the entropy of the function class $\hat{\mathcal{G}}(D_{-k}) = \{\hat{g}_\lambda(\cdot|D_{-k}) : \lambda \in \Lambda\}$ is (for any empirical distribution $Q$)

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

So we have

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

By Lemma 3 and 4, let

$$\delta = CG \left( \left( \frac{1 + \log(v/w) + \log(4\sigma^2 + P^{v-1}(g^*) + G) + \kappa \log n}{\min_{k=1:K}\{n_k\}} \right)^{1/2} \vee 1 \right)$$

where the constant $C$ only depends on the subgaussian constants.
By Lemma 3, for some constant $c$, for all $\ell, k = 1:K$, we have

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

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

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

Combining all the probability bounds above (and Easy Lemma 1), we have with high probability

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

**Step 3: Bound $\|\hat{\xi}_{\hat{\lambda}} - g^*\|_D$**

We know that

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

We bound each term in the summation separately. For every $k$, we know

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

Again we know that $\|\epsilon\|_k \le 2\sigma$ for all folds $D_k$ with high probability.

Conditioning on $\|\epsilon\|_k \le 2\sigma$, Lemma 5 gives us that with high probability

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

As shown in Step 2, Lemma 4 and Easy Lemma 1 imply that the following bound holds for all $\ell = 1 : K$ with high probability as long as $\|\epsilon\|_k \le 2\sigma$

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

Combining these two results, we get that with high probability,

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

## 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 $h$, $\|h\|_Q \le O_p(n^p)P(h)$.

Suppose $v \ge 1$.

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

Then the entropy bound is

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

where $\kappa$ depends on $\tau_{min}, v, u, p$.

**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 $h = c\left(\hat{g}_{\lambda_0} - \hat{g}_{\lambda_0+\delta}\right)$ 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 $\lambda$, we have

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

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 $h$),

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

where the second inequality comes from our assumption that $\min_{h:P(h)=1}\|h\|^{-2} = n^u$.

**2nd multiplicand**: By definition of $\hat{g}_{\lambda_0} + \hat{m}_h(\lambda)h$,

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

Hence

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

**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 $m \to 0$, 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

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

Conbining the above bounds, we have

$$\begin{aligned}\left|\frac{\partial}{\partial\lambda}\hat{m}_h(\lambda)\right| &\lesssim n^{\tau_{min}+u}w^{-1}\left(v\left(\frac{n^{\tau_{min}}}{2}\|\epsilon\|_T^2+P^v(g^*)\right)^{(v-1)/v}+wn^pG\right)\\ &\leq n^{\kappa}w^{-1}v\left(\|\epsilon\|_T^{2(v-1)/v}+P^{v-1}(g^*)+G\right)\end{aligned}$$

for some constant $\kappa$ that depends on $\tau_{min}, u, p, v$.
By the mean value theorem, there is some $\alpha \in [0,1]$ s.t

$$\begin{aligned}\|\hat{g}_\lambda(\cdot|D_{-k})-\hat{g}_{\lambda+\delta}(\cdot|D_{-k})\|_Q &= \hat{m}_h(\lambda)\|h\|_Q\\ &\leq n^p\delta\left|\frac{\partial}{\partial\lambda}\hat{m}_h(\lambda+\alpha\delta)\right|\end{aligned}$$

Combining the inequalities above, we get a bound on the covering number

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

and the entropy

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

### 2.1.2 Lemma 1

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

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

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

**Proof**

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

The second inequality follows by convexity of $P^v$ 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 $\lambda$. Suppose $P^v$ is convex.

Then

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

(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

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

Rearranging terms, we obain

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

Since this is true for any $t$, we have that

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

Thus

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

### 2.1.4 Lemma 3

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

Suppose $\epsilon_X$ are independent sub-gaussian RVs with constants $K$ and $\sigma$ (corresponding to $X$). Same fot $\epsilon_T$. Again, $\epsilon_T$ and $\epsilon_X$ 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 $\|\epsilon\|_T \le \sigma$:

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

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

Then there exists some $C$ s.t. for all $\delta$ s.t.

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

we have for some constant $c$

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

#### Proof

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

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

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 6. For all

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

there is some constant $c$ s.t.

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

### 2.1.5 Lemma 4

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

Suppose $\epsilon_X$ are independent sub-gaussian RVs with constants $K$ and $\sigma$ (corresponding to $X$). Same fot $\epsilon_T$. Again, $\epsilon_T$ and $\epsilon_X$ 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 $\|\epsilon\|_T \le \sigma$:

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

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

Then there exists some $C$ s.t. for all $\delta$ s.t.

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

we have

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

**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 $\|\epsilon\|_T \le \sigma$, we have

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

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

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

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

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

and similarly

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

### 2.1.6 Lemma 5

Let $\hat{\lambda}$ be chosen by CV. Let $\tilde{\lambda}$ be the oracle. Suppose $\sup_{g \in \mathcal{G}} \|g\|_D \le G$. Then with high probability,

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

**Proof**

The basic inequality gives us

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

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

$$\delta \ge C \left( \frac{1 + \log 4\sigma^2 - \log w + \kappa \log n}{\min_{k=1:K}\{n_k\}} \right)^{1/2}$$

we have that for every $k$,

$$Pr\left( \frac{\left| \left( \epsilon, g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k}) \right)_k \right|}{\|g_{\hat{\lambda}}(\cdot|D_{-k}) - g_{\tilde{\lambda}}(\cdot|D_{-k})\|_k} \ge \delta \wedge \|\epsilon\|_k \le 2\sigma \right) \le c \exp\left( -n_k \frac{\delta^2}{c^2} \right)$$

Hence for constants $c_k$ (which depend on $n_k/n$),

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

Also, by Cauchy-Schwarz, we have

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

Hence the result follows.

### 2.1.7 Lemma 6

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

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

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

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

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

Then there exists some $C$ dependent only on $K,\sigma$ s.t. for all

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

we have

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

**Proof**

Proof closely follows Lemma 3.2 from Vandegeer.

For a given set of RVs $\epsilon$, let $\{f_j^s(\cdot|\epsilon)\}_{j=1}^{N_s}$ be the $2^{-s}R$-covering set of $\mathcal{F}(T,\epsilon)$ where $N_s=N_s(2^{-s}R,\mathcal{F}(T,\epsilon),\|\cdot\|_X)\le\exp\left(\psi(2^{-s}R)\right)$. Let $S=\min\{s:2^{-s}R\le\delta/2\sigma\}$. 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 $\eta_s$ s.t. $\sum_{s=1}^S\eta_s\le 1$.

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

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

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 $\epsilon$ has a continuous probability measure, I think this holds).

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

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

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

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

Now choose $\eta_s$ 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

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

then

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

**Proof**

We know

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

hence

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

# 3 Corollaries

## 3.1 Convergence Rate Equivalence between the original regression problem and the perturbed ridge problem

Consider any $\lambda$.

Suppose the original regression problem is

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

and the new perturbed ridge problem is

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

Let $g^*$ be the true minimizer

$$g^* = \arg\min_{g\in\mathcal{G}} E\left[\|y - g\|^2\right]$$

Suppose there are constants $K_0, K_1 > 0$ s.t.

$$\frac{w}{2}\|g^*\|^2 \leq K_0 P^v(g) + K_1$$

Then the rate of convergence of

$$\left\|\hat{f} - g^*\right\|_D \lesssim \|\hat{g} - g^*\|_D$$

That is, the optimal rate of convergence determined by the oracle $\tilde{\lambda}$ is preserved under the new perturbed ridge regression problem.

**Proof**

By definition,

$$\begin{aligned} \frac{1}{2}\|y - \hat{f}\|_D^2 + \lambda\left(P^v(\hat{f}) + \frac{w}{2}\|\hat{f}\|^2\right) &\leq \frac{1}{2}\|y - g^*\|_D^2 + \lambda\left(P^v(g^*) + \frac{w}{2}\|g^*\|^2\right) \\ &\leq \frac{1}{2}\|y - g^*\|_D^2 + \lambda(1 + K_0)P^v(g^*) + \lambda K_1 \end{aligned}$$

Therefore

$$\frac{1}{2}\|y - \hat{f}\|_D^2 + \lambda P^v(\hat{f}) \leq \frac{1}{2}\|y - g^*\|_D^2 + \lambda(1 + K_0)P^v(g^*) + \lambda K_1$$

Notice that this inequality is very similar to the inequality from the original regression problem

$$\frac{1}{2}\|y - \hat{f}\|_D^2 + \lambda P^v(\hat{f}) \leq \frac{1}{2}\|y - g^*\|_D^2 + \lambda P^v(g^*)$$

Therefore the arguments to bound the convergence rate of $\|\hat{g} - g^*\|_D$ should give the same convergence rate for $\left\|\hat{f} - g^*\right\|_D$. (Example: refer to Thrm 10.2 in Vandegeer)

## 3.2 Regression problems with smooth-almost-everywhere penalty functions

If the regularization functions contain smooth-almost-everywhere penalty functions, we still have the same convergence rate. To prove this, we need an additional assumption similar to Condition 1 in the hillclimbing paper. However since our function classes may be infinite-dimensional, we need to generalize the definitions/condition:

Let $L_T(g, \lambda)$ be the penalized training criterion

$$L_T(g, \lambda) = \frac{1}{2}\|y - g\|_n^2 + \lambda\left(P^v(g) + \frac{w}{2}\|g\|^2\right)$$

**Definition 1:** The differentiable space of $L_T$ at $g$ are the functions along which the directional derivative of $L_T$ exists:

$$\Omega^{L_T}(g) = \left\{h : \lim_{r \to 0} \frac{L_T(g + rh) + L_T(g)}{r} \text{ exists}\right\}$$

**Definition 2:** $S_{\lambda_0}$ is a local optimality space of $L_T$ at $\lambda_0$ if there is a neighborhood containing $\lambda_0$ s.t. for all $\lambda \in W$

$$\arg\min_{g \in \mathcal{G}} L_T(g, \lambda) = \arg\min_{g \in S_{\lambda_0}} L_T(g, \lambda)$$

**Assumption A:** For almost every $\lambda$, the differentiable space $\Omega^{L_T}(\hat{g}(\lambda))$ contains the local optimality space $S_{\lambda_0}$.

Now consider the function class

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

We have all the same assumptions as in Lemma 0, except that $P^v$ can be smooth almost everywhere. Also, we have assumed "Assumption A" above. Then the same entropy bound holds:

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

where $\kappa$ depends on $\tau_{min}, v, u, p$.

**Proof**

Using a generalized verion of Theorem 1 in the hillclimbing paper (we should probably show this), we have that $L_T(\hat{g}(\lambda), \lambda)$ is continuously differentiable with respect to $\lambda$ for almost every $\lambda$. Let $S$ be the set of knots at which $\frac{\partial}{\partial\lambda} L_T(\hat{g}(\lambda), \lambda)|_{\lambda=s}$ does not exist.

We are interested in bounding $\|\hat{g}_\lambda - \hat{g}_{\lambda+\delta}\|_Q$. Suppose knots $s_1 < s_2 < ... < s_L$ (where $L$ might equal infinity) are in between between $\lambda$ and $\lambda + \delta$.

As done in Lemma 0, we can bound

$$\|\hat{g}_{s_i} - \hat{g}_{s_{i+1}}\|_Q \le (s_{i+1} - s_i) n^\kappa w^{-1} v \left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right)$$

for some constant $\kappa$ (for details, go to Lemma 0).

Then

$$\begin{aligned}
& \|\hat{g}_\lambda - \hat{g}_{\lambda+\delta}\|_Q \\
\le\ & \|g_\lambda - g_{s_1}\|_Q + \|g_{\lambda+\delta} - g_{s_L}\|_Q + \sum_{i=1}^{L-1} \|g_{s_i} - g_{s_{i+1}}\|_Q \\
\le\ & n^\kappa w^{-1} v \left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right) \left(s_1 - \lambda + \lambda + \delta - s_L + \sum_{i=1}^{L-1} (s_{i+1} - s_i)\right) \\
=\ & \delta n^\kappa w^{-1} v \left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right)
\end{aligned}$$

Therefore we have the same entropy bound as before.

## 3.3 Multiple Regularization Parameters

Suppose we are fitting model parameters $\theta$ but we'd like to apply a separate penalty to $J$ partitions of the model parameters: $P_j(\theta_j)$ for $j = 1, ..., J$. Let $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J)$ be their corresponding penalty parameters. We suppose that $\Lambda$ is the box $[\lambda_{min}, \lambda_{max}]^J$ where $\lambda_{min} = O_p(n^{-\tau_{min}})$ and $\lambda_{max} = O_p(1)$.

That is, the training criterion is now

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

For simplicity, suppose the problem is smooth. (One can probably use the same arguments to extend this to the case where the problem is smooth almost everywhere.)

We show that the entropy is off by a constant from Lemma 0:

$$H\left(d, \hat{\mathcal{G}}(T, \epsilon_T), \|\cdot\|_Q\right) \le J\left(\log\frac{1}{d} + \kappa \log n + \log\left(\frac{v}{wJ}\left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right)\right)\right)$$

Hence the convergence rate is essentially increased by a factor of $J^{1/2}$:

$$\|\hat{g}_{\hat{\lambda}}(\cdot|D) - g^*\|_D \lesssim \sqrt{\sum_{k=1}^K \left\|g^* - \hat{g}_{\tilde{\lambda}}(\cdot|D_{-k})\right\|_k^2} + G\left(\frac{J\left(1 + \kappa \log n + \log\left(\frac{v}{wJ}\left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right)\right)\right)}{\min_{k=1:K}\{n_k\}}\right)^{1/2}$$

**Proof for the entropy of the function class**

We first determine $\nabla_{\lambda_j}\hat{m}_h(\boldsymbol{\lambda})$ by implicit differentiation. The equations get bulky, but the logic is straightforward.

Use the same directional derivative technique in Lemma 0. Since $\hat{m}_h(\boldsymbol{\lambda})$ is a local minima,

$$-\langle h, y-(\hat{g}_{\lambda_0}+mh)\rangle_T + \sum_{j=1}^{J} \lambda_j \left(\frac{\partial}{\partial m} P_j^v(\hat{g}_{\lambda_0}+mh) + w\langle h, \hat{g}_{\lambda_0}+mh\rangle\right)\Bigg|_{m=\hat{m}_h(\boldsymbol{\lambda})} = 0$$

By implicit differentiation wrt $\lambda_\ell$ for $\ell = 1:J$, we have

$$\nabla_{\lambda_\ell}\hat{m}_h(\boldsymbol{\lambda}) = -\left(\|h\|_T^2 + \sum_{j=1}^{J} \lambda_j \frac{\partial^2}{\partial m^2} P_j^v\left(\hat{g}_{\lambda_0}+mh\right) + w\|h\|^2 \sum_{j=1}^{J}\lambda_j\right)^{-1} \left(\frac{\partial}{\partial m} P_\ell^v(\hat{g}_{\lambda_0}+mh) + w\langle h, \hat{g}_{\lambda_0}+mh\rangle\right)\Bigg|_{m=\hat{m}_h(}$$

Now we can bound $\|\hat{g}(\boldsymbol{\lambda}) - \hat{g}(\boldsymbol{\lambda}+\boldsymbol{\delta})\|$. Define $h = c\left(\hat{g}(\boldsymbol{\lambda}) - \hat{g}(\boldsymbol{\lambda}+\boldsymbol{\delta})\right)$ where $c$ is some scaling constant s.t. $P(h) = 1$. Following the same logic given in Lemma 0, we get

$$|\nabla_{\lambda_\ell}\hat{m}_h(\boldsymbol{\lambda})| \le w^{-1}J^{-1}n^{\tau_{min}+v}v\left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right)$$

Therefore, applying the mean value theorem, we get that for any vector $\boldsymbol{\delta}$ there is some $\alpha \in [0,1]$ s.t.

$$\begin{aligned}
\|\hat{g}(\boldsymbol{\lambda}) - \hat{g}(\boldsymbol{\lambda}+\boldsymbol{\delta})\|_Q &= \|\hat{m}_h(\boldsymbol{\lambda}+\boldsymbol{\delta})h\|_Q \\
&\le n^p\hat{m}_h(\boldsymbol{\lambda}+\boldsymbol{\delta}) \\
&= n^p\left|\langle \boldsymbol{\delta}, \nabla_{\boldsymbol{\lambda}}\hat{m}_h(\boldsymbol{\lambda}+\alpha\boldsymbol{\delta})\rangle\right| \\
&\le \|\boldsymbol{\delta}\|w^{-1}J^{-1}n^{\kappa}v\left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right)
\end{aligned}$$

for some constant $\kappa$ dependent on $v, p, \tau_{min}$.

Hence we can form a $d$-covering set for $\hat{\mathcal{G}}(T, \epsilon_T)$ by covering the box $\Lambda = [\lambda_{min}, \lambda_{max}]^J$ with balls of radius $dwJn^{-\kappa}\left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right)^{-1}$. Hence the $u-$covering number is

$$N\left(d, \hat{\mathcal{G}}(T,\epsilon_T), \|\cdot\|_Q\right) = d^{-J}\left(\frac{n^{\kappa}v}{wJ}\left(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G\right)\right)^J$$

Hence

$$H\left(d, \hat{\mathcal{G}}(T,\epsilon_T), \|\cdot\|_Q\right) \le J\log\frac{1}{d} + J\kappa\log n + J\log\left(\frac{v}{wJ}(\|\epsilon\|_T^{2(v-1)/v} + P^{v-1}(g^*) + G)\right)$$