Definitions

We find the best model for y over function class $\mathcal{G}$. Presume $g^* \in \mathcal{G}$ is the true model and

$$y = g^*(X) + \epsilon$$

Given a training set T , We define the fitted models

$$\hat{g}_\lambda = \|y - g\|_T^2 + \lambda^2 I^v(g)$$

Given a validation set T , let the CV-fitted model be

$$\hat{g}_{\hat{\lambda}} = \arg\min_\lambda \|y - \hat{g}_\lambda\|_V^2$$

We will suppose $I(g^*) > 0$.

Assumptions

Suppose we have sub-Gaussian errors ϵ for constants K and σ_0^2:

$$\max_{i=1:n} K^2 \left(E\left[\exp(|\epsilon_i|^2 K^2) - 1\right]\right) \le \sigma_0^2$$

Suppose $v > 2\alpha/(2+\alpha)$.
Suppose that the entropy of the class $\mathcal{G}'$ is

$$H\left(\delta, \mathcal{G}' = \left\{ \frac{g - g^*}{I(g) + I(g^*)} : g \in \mathcal{G}, I(g) + I(g^*) > 0 \right\}, P_n\right) \quad \le \quad \tilde{A}\delta^{-\alpha}$$

Suppose for all $\lambda \in \Lambda$, $I^v(\hat{g}_\lambda)$ is upper bounded by $\|\hat{g}_\lambda\|_n^2 = \frac{1}{n}\sum_{i=1}^n \hat{g}_\lambda(x_i)$. See Lemma 1 below for the specific assumption. This assumption includes Ridge, Lasso, Generalized Lasso, and the Group Lasso.

Result 1: Single λ, Single Penalty, cross-validation over $X_T = X_V$

For now, we will suppose $P_n = \{X_i\}_{i=1}^n$ are the same between the validation and training set.
Also, suppose the penalty normalizes the empirical norm such that:

$$\sup_{g \in \mathcal{G}} \frac{\|g - g^*\|_n}{I(g) + I(g^*)} \le R < \infty$$

Suppose for all $\lambda \in \Lambda$, $I^v(\hat{g}_\lambda)$ is upper bounded by its L_2-norm with some constant M and M_0 such that

$$I^v(\hat{g}_\lambda) \le M\|\hat{g}_\lambda\|_n^2 + M_0$$

Then

$$\|\hat{g}_{\hat{\lambda}} - g^*\|_n = O_p(n^{-1/(2+\alpha)}) \left(M^{\frac{\alpha v - 2\alpha + 2v}{v(v-2)(2+\alpha)}} R^{v/(v-2)} \vee I^{2\alpha/(2+\alpha)}(g^*)\right)$$

Proof

Let $\tilde{\lambda}$ be the optimal λ under the given assumptions, as specified by Van de geer. From the definition of $\hat{\lambda}$, we get the following basic inequality

$$\begin{aligned}\|g^* - \hat{g}_{\hat{\lambda}}\|_V^2 &\leq \|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2 + 2(\epsilon, \hat{g}_{\hat{\lambda}} - \hat{g}_{\tilde{\lambda}})_V \\ &\leq \|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2 + 2(\epsilon, \hat{g}_{\hat{\lambda}} - g^*)_V + 2(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V \\ &\leq \|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2 + 2\left|(\epsilon, \hat{g}_{\hat{\lambda}} - g^*)_V\right| + 2\left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right|\end{aligned}$$

By considering the largest term on the RHS, we have following three cases.

Case 1: $\|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2$ is the largest

Since we have assumed that the validation and training set are equal, then $\|g^* - \hat{g}_{\tilde{\lambda}}\|_V$ converges at the optimal rate $O_p(n^{-1/(2+\alpha)})$.

Case 2: $\left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right|$ is the largest

In this case, since ϵ_V is independent of $\hat{g}_{\tilde{\lambda}}$, then by Cauchy Schwarz,

$$\begin{aligned}\left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right| &\leq \|\epsilon_V\| \|g^* - \hat{g}_{\tilde{\lambda}}\|_V \\ &\leq O_p\left(n^{-1/2}\right) \|g^* - \hat{g}_{\tilde{\lambda}}\|_V\end{aligned}$$

Hence $\left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right|$ will shrink a bit faster than the optimal rate at a rate of $O_p(n^{-(\frac{1}{2+\alpha}+\frac{1}{2})})$.

Case 3: $\left|(\epsilon, g^* - \hat{g}_{\hat{\lambda}})_V\right|$ is the largest.

By the assumptions given, Vandegeer (10.6) gives us that

$$\sup_{g \in \mathcal{G}} \frac{|(\epsilon, g - g*)_n|}{\|g - g*\|_n^{1-\alpha/2} (I(g^*) + I(g))^{\alpha/2}} = O_p(n^{-1/2})$$

Hence

$$\left|(\epsilon, g^* - \hat{g}_{\hat{\lambda}})_V\right| \leq O_p(n^{-1/2}) \|\hat{g}_{\hat{\lambda}} - g*\|_n^{1-\alpha/2} \left(I(g^*) + I(\hat{g}_{\hat{\lambda}})\right)^{\alpha/2}$$

If $I(g^*) \geq I(g_{\hat{\lambda}})$, then

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_V \leq O_p(n^{-1/(2+\alpha)}) I(g^*)^{\alpha/(2+\alpha)}$$

Otherwise, we have

$$\|\hat{g}_{\hat{\lambda}} - g*\|_n^{1+\alpha/2} \leq O_p(n^{-1/2}) I(\hat{g}_{\hat{\lambda}})^{\alpha/2}$$

By Lemma 1 below, using the assumption that the penalty of $\hat{g}_\lambda$ is bounded above by its $L_2(P_n)$ norm, we have that

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_n \leq O_p(n^{-1/(2+\alpha)}) M^{\frac{\alpha v - 2\alpha + 2v}{v(v-2)(2+\alpha)}} R^{v/(v-2)}$$

Result 2: Single λ, Single Penalty, cross-validation over general X_T, X_V

Now suppose that the training and validation set are independently sampled, so the values X_i are not necessarily the same. Suppose X is bounded s.t. $|X| \leq R_X$ and the domain of $g \in \mathcal{G}$ is over $(-R_X, R_X)$.

We suppose the training and validation sets are both of size n.

Suppose the penalty normalizes the empirical norm as follows:

$$\sup_{g \in \mathcal{G}} \frac{\|g - g^*\|_T}{I(g) + I(g^*)} \leq R < \infty, \ \sup_{g \in \mathcal{G}} \frac{\|g - g^*\|_V}{I(g) + I(g^*)} \leq R < \infty$$

Suppose that

$$\sup_{g\in\mathcal{G}} \frac{\|g-g^*\|_\infty}{I(g)+I(g^*)} \quad \leq \quad K < \infty$$

Suppose for all $\lambda \in \Lambda$, $I^v(\hat{g}_\lambda)$ is upper bounded by its L_2-norm with constants M and M_0:

$$I^v(\hat{g}_\lambda) \leq M\left(\|\hat{g}_\lambda\|_T^2 + \|\hat{g}_\lambda\|_V^2\right) + M_0 = M\|\hat{g}_\lambda\|_{2n}^2 + M_0$$

Then for any $\xi > 0$,

$$\|\hat{g}_{\hat{\lambda}} - g^*\|_V = O_p(n^{-1/(2+\alpha+\xi)})I(g^*)$$

Proof: We follow the same proof structure of going thru the three cases, modifying the proofs as appropriate:

Case 1: $\|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2$ is the largest

By Lemma 2, we have

$$Pr\left(\sup_{g\in\mathcal{G}} \frac{\left|\|g^*-g\|_{P_n} - \|g^*-g\|_{P_{n''}}\right|}{I(g^*)+I(g)} \geq 6\delta\right) \leq 2\exp\left(2\tilde{A}\delta^{-\alpha} - \frac{4\delta^2 n}{K^2}\right)$$

Hence for any $\xi > 0$,

$$\frac{\left|\|g^*-\hat{g}_{\tilde{\lambda}}\|_T - \|g^*-\hat{g}_{\tilde{\lambda}}\|_V\right|}{I(g^*)+I(\hat{g}_{\tilde{\lambda}})} \leq O_p(n^{-1/(2+\alpha+\xi)})$$

Therefore

$$\begin{aligned}
\|g^*-\hat{g}_{\tilde{\lambda}}\|_V &\leq \|g^*-\hat{g}_{\tilde{\lambda}}\|_T + O_p(n^{-1/(2+\alpha+\xi)})\left(I(g^*)+I(\hat{g}_{\tilde{\lambda}})\right) \\
&\leq \|g^*-\hat{g}_{\tilde{\lambda}}\|_T + O_p(n^{-1/(2+\alpha+\xi)})I(g^*)
\end{aligned}$$

Hence we can attain a rate that is infinitely close to the optimal rate.

Case 2: $\left|(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right|$ is the largest

The same proof still holds.

Case 3: $\left|(\epsilon, g^* - \hat{g}_{\hat{\lambda}})_V\right|$ is the largest.

Again, we have by Van de geer (10.6),

$$\left|(\epsilon, g^* - \hat{g}_{\hat{\lambda}})_V\right| \leq O_p(n^{-1/2})\|\hat{g}_{\hat{\lambda}} - g*\|_V^{1-\alpha/2}(I(g^*)+I(\hat{g}_{\hat{\lambda}}))^{\alpha/2}$$

If $I(g^*) \geq I(g_{\hat{\lambda}})$ is true, then result is clearly attained.

Otherwise, we have

$$\|\hat{g}_{\hat{\lambda}} - g*\|_V^{1+\alpha/2} \leq O_p(n^{-1/2})I(\hat{g}_{\hat{\lambda}})^{\alpha/2}$$

By Lemma 1 below, since the penalty is bounded above by the $L_2(P_n)$ norm, it follows that

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_V \leq O_p(n^{-1/(2+\alpha)})M^{\frac{\alpha v - 2\alpha + 2v}{v(v-2)(2+\alpha)}}R^{v/(v-2)}$$

Result 3: Single λ, Multiple Penalties, Optimal $\tilde{\lambda}_T$ over X_T

Consider function classes $\mathcal{G}_j$ that are cones. Also, suppose we have an additive model:

$$y = \sum_{j=1}^{J} g_j^* + \epsilon$$

where $g_j^* \in \mathcal{G}_j$.

We fit the model by least squares with separate penalties for each function g_j:

$$\{\hat{g}_j\}_{j=1}^{J} = \arg\min_{g_j \in \mathcal{G}_j} \|y - \sum_{j=1}^{J} g_j\|_T^2 + \frac{\lambda^2}{J} \sum_{j=1}^{J} I_j^{v_j}(g_j)$$

Suppose $v_j \geq 1$ for all j. (This requirement on v_j stricter than Vandegeer Them 10.2)

Suppose for all j, there is some $0 < \alpha < 2$ s.t. for all $\delta > 0$,

$$H\left(\delta, \{g_j \in \mathcal{G}_j : I(g_j) \leq 1\}, \|\cdot\|_T\right) \leq A\delta^{-\alpha}$$

and that for all j

$$\sup_{g_j \in \mathcal{G}_j} \frac{\|g_j - g_j^*\|_T}{I(g_j) + I(g_j^*)} \leq R < \infty$$

If we choose λ s.t.

$$\tilde{\lambda}_T^{-1} = O_p\left(n^{1/(2+\alpha)}\right)\left(J + \sum_{j=1}^{J} I_j^{v_j}(g_j^*)\right)^{(2-\alpha)/2(2+\alpha)}$$

then

$$\|\sum_{j=1}^{J} g_j - g_j^*\|_T = O_p\left(\tilde{\lambda}_T\right) J^{\alpha/4}\left(\sum_{j=1}^{J} I_j^{v_j}(g_j^*)\right)^{1/2}$$

and

$$\sum_{j=1}^{J} I_j(\hat{g}_j) \leq J^{1/2+\alpha/4}\left(J + \sum_{j=1}^{J} I_j^{v_j}(g_j^*)\right)$$

Proof:

The basic inequality gives us:

$$\|\sum_{j=1}^{J} \hat{g}_j - g_j^*\|_T^2 + \frac{\lambda^2}{J}\sum_{j=1}^{J} I_j^{v_j}(\hat{g}_j) \leq 2\left|\left(\epsilon_T, \sum_{j=1}^{J} \hat{g}_j - g_j^*\right)\right| + \frac{\lambda^2}{J}\sum_{j=1}^{J} I_j^{v_j}\left(g_j^*\right)$$

Case 1: $\left|\left(\epsilon_T, \sum_{j=1}^{J} \hat{g}_j - g_j^*\right)\right| \leq \frac{\lambda^2}{J}\sum_{j=1}^{J} I_j^{v_j}\left(g_j^*\right)$

$$\|\sum_{j=1}^{J} \hat{g}_j - g_j^*\|_T \leq O_p(\lambda)\left(\frac{1}{J}\sum_{j=1}^{J} I_j^{v_j}\left(g_j^*\right)\right)^{1/2}$$

Case 2: $\left|\left(\epsilon_T, \sum_{j=1}^{J} \hat{g}_j - g_j^*\right)\right| \geq \frac{\lambda^2}{J}\sum_{j=1}^{J} I_j^{v_j}\left(g_j^*\right)$

By Lemma 3,

$$H\left(\delta, \left\{\frac{\sum_{j=1}^J g_j - g_j^*}{\sum_{j=1}^J I(g_j) + I(g_j^*)} : g_j \in \mathcal{G}_j, I(g_j) + I(g_j^*) > 0\right\}, \|\cdot\|_T\right) \le \tilde{A} J^{1-\alpha}\delta^{-\alpha}$$

Hence by (10.6) in Vandegeer,

$$\sup_{g_j \in \mathcal{G}_j} \frac{\left|\left(\epsilon_T, \sum_{j=1}^J g_j - g_j^*\right)\right|}{\|\sum_{j=1}^J g_j - g_j^*\|^{1-\alpha/2}\left(\sum_{j=1}^J I(g_j) + I(g_j^*)\right)^{\alpha/2}} = O_p\left(n^{-1/2}\right) J^{1-\alpha}$$

and the basic inequality becomes

$$\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T^2 + \frac{\lambda^2}{J}\sum_{j=1}^J I_j^{v_j}(\hat{g}_j) \le O_p\left(n^{-1/2}\right) J^{1-\alpha}\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T^{1-\alpha/2}\left(\sum_{j=1}^J I(\hat{g}_j) + I(g_j^*)\right)^{\alpha/2}$$

Case 2a: Suppose $\sum_{j=1}^J I(\hat{g}_j) \le \sum I(g_j^*)$.
Then

$$\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T \quad \le \quad O_p\left(n^{-1/(2+\alpha)}\right) J^{\frac{2(1-\alpha)}{\alpha+2}}\left(\sum_{j=1}^J I(g_j^*)\right)^{\alpha/(2+\alpha)}$$

Case 2b: Suppose $\sum_{j=1}^J I(\hat{g}_j) \ge \sum I(g_j^*)$.
Then

$$\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T^{1+\alpha/2} \quad \le \quad O_p\left(n^{-1/2}\right)\left(\sum_{j=1}^J I(\hat{g}_j)\right)^{\alpha/2}$$

By assuming $v_j \ge 1$, we must have $I_j(\hat{g}_j) \le I_j^{v_j}(\hat{g}_j) \vee 1$. Hence

$$\begin{aligned}\sum_{j=1}^J I_j(\hat{g}_j) &\le J + \sum_{j=1}^J I_j^{v_j}(\hat{g}_j) \\ &\le J + O_p\left(n^{-1/2}\right) J^{2-\alpha}\lambda^{-2}\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T^{1-\alpha/2}\left(\sum_{j=1}^J I(\hat{g}_j)\right)^{\alpha/2}\end{aligned}$$

Case 2ba: If $J \le O_p\left(n^{-1/2}\right) J^{2-\alpha}\lambda^{-2}\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T^{1-\alpha/2}\left(\sum_{j=1}^J I(\hat{g}_j)\right)^{\alpha/2}$, then

$$\sum_{j=1}^J I_j(\hat{g}_j) \le O_p\left(n^{-1/(2-\alpha)}\right) J^{1/2}\lambda^{-4/(2-\alpha)}\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T$$

which implies

$$\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T \le O_p\left(n^{-1/(2-\alpha)}\right) J^{\alpha/4}\lambda^{-2\alpha/(2-\alpha)}$$

and

$$\sum_{j=1}^J I_j(\hat{g}_j) \le J^{1/2+\alpha/4}\left(J + \sum_{j=1}^J I_j^{v_j}(g_j^*)\right)$$

Case 2bb: If $J \geq O_p\left(n^{-1/2}\right) J^{2-\alpha}\lambda^{-2}\|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T^{1-\alpha/2}\left(\sum_{j=1}^J I(\hat{g}_j)\right)^{\alpha/2}$, then

$$\sum_{j=1}^J I_j(\hat{g}_j) \leq J \implies \|\sum_{j=1}^J \hat{g}_j - g_j^*\|_T \leq O_p\left(n^{-1/(2+\alpha)}\right) J^{\alpha/(2+\alpha)}$$

Result 4: Single λ, Multiple Penalties, cross-validation over general X_T, X_V

Now suppose that the training and validation set are independently sampled, so the values X_i are not necessarily the same. Suppose X is bounded s.t. $|X| \leq R_X$ and the domain of $g \in \mathcal{G}$ is over $(-R_X, R_X)$.

We suppose the training and validation sets are both of size n.

Suppose the penalty normalizes the empirical norm as follows:

$$\sup_{g_j \in \mathcal{G}_j} \frac{\|g_j - g_j^*\|_T}{I(g_j) + I(g_j^*)} \leq R < \infty, \quad \sup_{g_j \in \mathcal{G}_j} \frac{\|g_j - g_j^*\|_V}{I(g_j) + I(g_j^*)} \leq R < \infty$$

We suppose the same entropy conditions as Result 3. Furthermore, suppose that

$$\sup_{g_j \in \mathcal{G}_j} \frac{\|g_j - g_j^*\|_\infty}{I(g_j) + I(g_j^*)} \quad \leq \quad K < \infty$$

Suppose there exist constants M and M_0 s.t. for all $\lambda \in \Lambda$ and all j, $I_j^{v_j}(\hat{g}_{\lambda,j})$ is upper bounded by its L_2-norm :

$$I_j^{v_j}(\hat{g}_{\lambda,j}) \leq M\left(\|\hat{g}_{\lambda,j}\|_T^2 + \|\hat{g}_{\lambda,j}\|_V^2\right) + M_0 = M\|\hat{g}_{\lambda,j}\|_{2n}^2 + M_0$$

Then for any $\xi > 0$,

$$\|\sum_{j=1}^J \hat{g}_{\hat{\lambda},j} - g_j^*\|_V = O_p(n^{-1/(2+\alpha+\xi)})\left(\sum_{j=1}^J I_j^{v_j}(g_j^*)\right)$$

Proof:

The proof is very similar to Result 2.

Case 1: $\|\sum_{j=1}^J g_j^* - \hat{g}_{\tilde{\lambda},j}\|_V^2$ is the largest

By Lemma 2, we have

$$Pr\left(\sup_{g_j \in \mathcal{G}_j} \frac{\left|\|\sum_{j=1}^J g_j^* - g_j\|_{P_n} - \|\sum_{j=1}^J g_j^* - g_j\|_{P_{n''}}\right|}{\sum_{j=1}^J I_j(g_j^*) + I_j(g_j)} \geq 6\delta\right) \leq 2\exp\left(2\tilde{A}J^{1-\alpha}\delta^{-\alpha} - \frac{4\delta^2 n}{K^2}\right)$$

Hence for any $\xi > 0$,

$$\frac{\left|\|\sum_{j=1}^J g_j^* - \hat{g}_{\tilde{\lambda},j}\|_T - \|\sum_{j=1}^J g_j^* - \hat{g}_{\tilde{\lambda},j}\|_V\right|}{\sum_{j=1}^J I_j(g_j^*) + I_j(\hat{g}_{\tilde{\lambda},j})} \leq O_p(n^{-1/(2+\alpha+\xi)})$$

Therefore

$$\begin{aligned}
\|\sum_{j=1}^J g_j^* - \hat{g}_{\tilde{\lambda},j}\|_V &\leq \|\sum_{j=1}^J g_j^* - \hat{g}_{\tilde{\lambda},j}\|_T + O_p(n^{-1/(2+\alpha+\xi)})\left(\sum_{j=1}^J I(g_j^*) + I(\hat{g}_{\tilde{\lambda},j})\right) \\
&\leq O_p(n^{-1/(2+\alpha+\xi)})J^{1/2+\alpha/4}\left(J + \sum_{j=1}^J I_j^{v_j}(g_j^*)\right)
\end{aligned}$$

Hence we can attain a rate that is infinitely close to the optimal rate.

Case 2: $\left|\left(\epsilon_V, \sum_{j=1}^J g_j^* - \hat{g}_{\tilde{\lambda},j}\right)\right|$ is the largest

Since ϵ_V is independent of $\left\{\hat{g}_{\tilde{\lambda},j}\right\}$, then this term shrinks at the rate of $O_p(n^{-1/2-1/(2+\alpha_{max})})$. (So the rate is faster than the optimal rate.)

Case 3: $\left|\left(\epsilon_V, \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j}\right)\right|$ is the largest.

Again, we have by Vandegeer (10.6),

$$\left|\left(\epsilon_V, \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j}\right)\right| \le O_p(n^{-1/2}) \| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j}\|_V^{1-\alpha/2} \left(\sum_{j=1}^J I(g_j^*) + I(\hat{g}_{\hat{\lambda},j})\right)^{\alpha/2}$$

If $\sum_{j=1}^J I(g_j^*) \ge \sum_{j=1}^J I(\hat{g}_{\hat{\lambda},j})$ is true, then result is clearly attained.
Otherwise, we have

$$\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j}\|_V^{1+\alpha/2} \le O_p(n^{-1/2}) \left(\sum_{j=1}^J I(\hat{g}_{\hat{\lambda},j})\right)^{\alpha/2}$$

By the assumption that the penalty is bounded by the $L_2(P_{2n})$ norm,

$$\begin{aligned}
\sum_{j=1}^J I(\hat{g}_{\hat{\lambda},j}) &\le \sum_{j=1}^J \left(M\|\hat{g}_{\hat{\lambda},j}\|_{2n}^2 + M_0\right)^{1/v_j} \\
&\le \sum_{j=1}^J \left(M\left(\|g_j^* - \hat{g}_{\hat{\lambda},j}\|_{2n} + \|g_j^*\|_{2n}\right)^2 + M_0\right)^{1/v_j}
\end{aligned}$$

By Lemma 1a, $\|g_j^* - \hat{g}_{\hat{\lambda},j}\|_{2n}$ is bounded and by assumption $\|g_j^*\|_{2n}$ is also bounded. Hence
Hence for some constant c dependent on $R, \|g_j^*\|_{2n}, M, j, v_j$, we have

$$\| \sum_{j=1}^J g_j^* - \hat{g}_{\hat{\lambda},j}\|_V \le O_p(n^{-1/(2+\alpha)})c$$

Result 5: Multiple λ, Multiple Penalties, Optimal λ on X_T

Consider an additive model:

$$y = \sum_{j=1}^J g_j^* + \epsilon$$

We fit the model by least squares with separate penalties and separate λ for each function g_j:

$$\{\hat{g}_j\}_{j=1}^J = \arg\min_{g_j \in \mathcal{G}_j} \|y - \sum_{j=1}^J g_j\|_T^2 + \frac{1}{J}\sum_{j=1}^J \lambda_j^2 I_j^{v_j}(g_j)$$

Suppose $v_j > \frac{2\alpha_j}{2+\alpha_j}$ for all j.
Suppose for all j, there is some $0 < \alpha_j < 2$ s.t. for all $\delta > 0$,

$$H\left(\delta, \left\{\frac{g_j - g_j^*}{I(g_j) + I(g_j^*)} : g_j \in \mathcal{G}_j, I(g_j) + I(g_j^*) > 0\right\}, \|\cdot\|_T\right) \le \frac{A}{J}\delta^{-\alpha_j}$$

and for all j,

$$\sup_{g_j \in \mathcal{G}_j} \frac{\|g_j - g_j^*\|_T}{I(g_j) + I(g_j^*)} \leq R < \infty$$

If we choose λ s.t.

$$\tilde{\lambda}_j^{-1} = ???$$

then

$$\| \sum_{j=1}^{J} g_j - g_j^* \|_T = ???$$

and

$$\sum_{j=1}^{J} I^{v_j}(\hat{g}_{\lambda,j}) = ???$$

Proof:

Lemmas

Lemma 1:

Suppose for all $\lambda \in \Lambda$, the penalty function $I^v(g_\lambda)$ is upper-bounded by $\|g_\lambda\|_n^2 = \frac{1}{n}\sum_{i=1}^n g_\lambda^2(x_i)$ with constants M_0 and M:

$$I^v(g_\lambda) \leq M\|g_\lambda\|_n^2 + M_0$$

and

$$\frac{\|g_\lambda - g^*\|_n}{I(g_\lambda) + I(g^*)} \leq R$$

Then (lemma 1a)

$$\sup_{\lambda \in \Lambda} \|g_\lambda - g^*\|_n \leq O_p(R^{v/(v-2)}) M^{1/(v-2)}$$

Furthermore, if there is some function $g^* \in \mathcal{G}$ such that

$$\|g^* - g_\lambda\|_n^{1+\alpha/2} \leq O_p(n^{-1/2}) I^{\alpha/2}(g_\lambda)$$

then for sufficiently large n, (lemma 1b)

$$\|g^* - g_\lambda\|_n \leq O_p(n^{-1/(2+\alpha)}) M^{\frac{\alpha v - 2\alpha + 2v}{v(v-2)(2+\alpha)}}$$

Proof:
First we show that $\sup_\lambda \|g_\lambda - g^*\|_n$ is bounded and does not grow with n. For any λ, we have

$$\|g_\lambda - g^*\|_n \quad \leq \quad R\left(I(g_\lambda) + I(g^*)\right)$$

Clearly if $I(g^*) \geq I(g_\lambda)$, we're done. Otherwise,

$$\|g_\lambda - g^*\|_n \leq 2RI(g_\lambda) \leq 2R\left(M\|g_\lambda\|_n^2 + M_0\right)^{1/v}$$

If $M\|g_\lambda\|_n^2 \leq M_0$, we're done. Otherwise,

$$\begin{aligned} \|g_\lambda - g^*\|_n &\leq O_p(R) M^{1/v} \|g_\lambda\|_n^{2/v} \\ &\leq O_p(R) M^{1/v} \left(\|g_\lambda - g^*\|_n + \|g^*\|_n\right)^{2/v} \end{aligned}$$

Again, if $\|g_\lambda - g^*\|_n \le \|g^*\|_n$, we're done. Otherwise,

$$\|g_\lambda - g^*\|_n \le O_p(R^{v/(v-2)})M^{1/(v-2)}$$

So we've shown that $\sup_\lambda \|g_\lambda - g^*\|_n$ is bounded (lemma 1a).
Now to prove lemma 1b, note that from the assumptions, we have

$$\|g^* - g_\lambda\|_n^{1+\alpha/2} \le O_p(n^{-1/2})\left(M\|g_\lambda\|_n^2 + M_0\right)^{\alpha/2v}$$

If $M_0 > \|g_\lambda\|_n^2$, we're done. Otherwise,

$$\begin{aligned}\|g^* - g_\lambda\|_n^{1+\alpha/2} &\le O_p(n^{-1/2})M^{\alpha/2v}\|g_\lambda\|_n^{\alpha/v} \\ &\le O_p(n^{-1/2})M^{\alpha/2v}\left(\|g_\lambda - g^*\|_n + \|g^*\|_n\right)^{\alpha/v}\end{aligned}$$

Since we showed that $\|g_\lambda - g^*\|_n$ is bounded, then

$$\|g^* - g_\lambda\|_n^{1+\alpha/2} \le O_p(n^{-1/2})M^{\alpha/2v+1/(v-2)}R^{v/(v-2)}$$

Hence

$$\|g^* - g_\lambda\|_n \le O_p(n^{-1/(2+\alpha)})M^{\frac{\alpha v-2\alpha+2v}{v(v-2)(2+\alpha)}}R^{v/(v-2)}$$

I believe we can often provide a good estimate of M for the entire class $\mathcal{G}$, which means that we can always estimate the sample size needed to ensure this case never occurs. That is, I believe we can often estimate M s.t.

$$I^v(g) \le M\|g\|_n^2 + M_0 \forall g \in \mathcal{G}$$

Lemma 2:

Let $P_{n'}$ and $P_{n''}$ be empirical distributions over $\{X_i'\}_{i=1}^n, \{X_i''\}_{i=1}^n$. Let $P_{2n} = \frac{1}{2}(P_{n'} + P_{n''})$. Suppose X is bounded s.t. $|X| < R_X$.

Let $\mathcal{G}' = \left\{\frac{g-g^*}{I(g)+I(g^*)} : g \in \mathcal{G}, I(g) + I(g^*) > 0\right\}$. Suppose g is defined over the domain over X (and zero otherwise). Suppose

$$\sup_{f\in\mathcal{G}'}\|f\|_{P_{2n}} \le R < \infty, \ \sup_{f\in\mathcal{G}'}\|f\|_\infty \le K < \infty$$

and

$$H(\delta, \mathcal{G}', P_{n'}) \le \tilde{A}\delta^{-\alpha}, \ H(\delta, \mathcal{G}', P_{n''}) \le \tilde{A}\delta^{-\alpha}$$

Then

$$Pr\left(\sup_{g\in\mathcal{G}}\frac{\left|\|g^* - g\|_{P_{n'}} - \|g^* - g\|_{P_{n''}}\right|}{I(g^*) + I(g)} \ge 6\delta\right) \le 2\exp\left(2\tilde{A}\delta^{-\alpha} - \frac{4\delta^2 n}{K^2}\right)$$

Proof: The proof is very similar to that in Pollard 1984 (page 32), so some details below are omitted.
First note that for any function f and h, we have

$$\|f\|_{P_{n'}} - \|h\|_{P_{n'}} \le \|f - h\|_{P_{n'}} \le \sqrt{2}\|f - h\|_{P_{2n}}$$

Similarly for $P_{n''}$.

Let $\{h_j\}_{j=1}^N$ be the $\sqrt{2}\delta$-cover for $\mathcal{G}'$ (where $N = N(\sqrt{2}\delta, \mathcal{G}', P_{2n})$). Let h_j be the closest function (in terms of $\|\cdot\|_{P_{2n}}$) to some $f \in \mathcal{G}'$. Then

$$\begin{aligned}
\|f\|_{P_{n'}} - \|f\|_{P_{n''}} &\leq \|f - h_j\|_{P_{n'}} + \big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| + \|f - h_j\|_{P_{n''}} \\
&\leq 4\delta + \big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big|
\end{aligned}$$

Therefore for $f = \frac{g^* - g}{I(g^*) + I(g)}$, we have

$$\begin{aligned}
Pr\left(\sup_{g \in \mathcal{G}} \frac{\big|\|g^* - g\|_{P_n} - \|g^* - g\|_{P_{n''}}\big|}{I(g^*) + I(g)} \geq 6\delta\right) &\leq Pr\left(\sup_{j \in 1:N} \big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| \geq 2\delta\right) \\
&\leq N \max_{j \in 1:N} Pr\left(\big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| \geq 2\delta\right)
\end{aligned}$$

Now note that

$$\begin{aligned}
\big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| &= \frac{\big|\|h_j\|_{P_{n'}}^2 - \|h_j\|_{P_{n''}}^2\big|}{\|h_j\|_{P_{n'}} + \|h_j\|_{P_{n''}}} \\
&\leq \frac{\big|\|h_j\|_{P_{n'}}^2 - \|h_j\|_{P_{n''}}^2\big|}{\sqrt{2}\|h_j\|_{P_{2n}}}
\end{aligned}$$

By Hoeffding's inequality,

$$\begin{aligned}
Pr\left(\big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| \geq 2\delta\right) &\leq Pr\left(\Big|\|h_j\|_{P_{n'}}^2 - \|h_j\|_{P_{n''}}^2\Big| \geq 2\sqrt{2}\delta\|h_j\|_{P_{2n}}\right) \\
&= Pr\left(\left|\sum_{i=1}^n W_i\left(h_j^2(x_i') - h_j^2(x_i'')\right)\right| \geq 2\sqrt{2}n\delta\|h_j\|_{P_{2n}}\right) \\
&\leq 2\exp\left(-\frac{16\delta^2 n^2 \|h_j\|_{P_{2n}}^2}{4\sum_{i=1}^n \left(h_j^2(x_i') - h_j^2(x_i'')\right)^2}\right)
\end{aligned}$$

Since $\|h_j\|_\infty < K$, then

$$\begin{aligned}
\sum_{i=1}^n \left(h_j^2(x_i') - h_j^2(x_i'')\right)^2 &\leq \sum_{i=1}^n h_j^4(x_i') + h_j^4(x_i'') \\
&\leq nK^2\|h_j\|_{P_{2n}}^2
\end{aligned}$$

Hence

$$Pr\left(\big|\|h_j\|_{P_{n'}} - \|h_j\|_{P_{n''}}\big| \geq 2\delta\right) \leq 2\exp\left(-\frac{4\delta^2 n}{K^2}\right)$$

Since (Pollard and Vandegeer say that)

$$N(\sqrt{2}\delta, \mathcal{G}', P_{2n}) \leq N(\delta, \mathcal{G}', P_{n''}) + N(\delta, \mathcal{G}', P_{n''})$$

then

$$Pr\left(\sup_{g \in \mathcal{G}} \frac{\big|\|g^* - g\|_{P_n} - \|g^* - g\|_{P_{n''}}\big|}{I(g^*) + I(g)} \geq 6\delta\right) \leq 2\exp\left(2\tilde{A}\delta^{-\alpha} - \frac{4\delta^2 n}{K^2}\right)$$

Using shorthand, we can write that for any $\xi > 0$,

$$\sup_{g \in \mathcal{G}} \frac{\big|\|g^* - g\|_{P_n} - \|g^* - g\|_{P_{n''}}\big|}{I(g^*) + I(g)} = O_p(n^{-1/(2+\alpha+\xi)})$$

Lemma 3:

Suppose the function classes $\mathcal{F}_j$ is a cone and $I_j : \mathcal{F}_j \mapsto [0, \infty)$ is a psuedonorm. Furthermore, suppose

$$H\left(\delta, \{f_j \in \mathcal{F}_j : I_j(f_j) \leq 1\}, \|\cdot\|_n\right) \leq A_j \delta^{-\alpha_j}$$

Then if $f_j^* \in \mathcal{F}_j$, then

$$H\left(\delta, \left\{\frac{\sum_{j=1}^J f_j - f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} : f_j \in \mathcal{F}_j, I_j(f_j) + I_j(f_j^*) > 0\right\}, \|\cdot\|_n\right) \quad \leq \quad 2\sum_{j=1}^J A_j \left(\frac{\delta}{2J}\right)^{-\alpha_j}$$

Proof: Let $\tilde{f}_j = \frac{f_j}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)}$. Then $\tilde{f}_j \in \mathcal{F}_j$ and $I_j(\tilde{f}_j) \leq 1$. Let $h_{(j)}$ be the closest function to $\tilde{f}_j$ in the δ cover of $\mathcal{F}_j$. Similarly, let $h_{(j)}^*$ be the closest function to $\tilde{f}_j^*$ in the δ cover of $\mathcal{F}_j$. Then

$$\begin{aligned}
\left\|\frac{\sum_{j=1}^J f_j - f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} - \left(\sum_{j=1}^J h_{(j)} - h_{(j)}^*\right)\right\| \quad &\leq \quad \sum_{j=1}^J \left\|\frac{f_j - f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} - \left(h_{(j)} - h_{(j)}^*\right)\right\| \\
&\leq \quad \sum_{j=1}^J \left\|\frac{f_j}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} - h_{(j)}\right\| + \left\|\frac{f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} - h_{(j)}^*\right\| \\
&\leq \quad 2J\delta
\end{aligned}$$

Hence

$$H\left(2J\delta, \left\{\frac{\sum_{j=1}^J f_j - f_j^*}{\sum_{j=1}^J I_j(f_j) + I_j(f_j^*)} : f_j \in \mathcal{F}_j, I_j(f_j) + I_j(f_j^*) > 0\right\}, \|\cdot\|_n\right) \leq 2\sum_{j=1}^J A_j \delta^{-\alpha_j}$$

Example 1: Sobelov norm (NOT DONE)

Consider the functions

$$\mathcal{G} = \left\{g : [0,1] \mapsto \mathbb{R} : \int_0^1 g^{(m)}(z)^2 dz < \infty\right\}$$

Suppose x_i are all unique. Then the Sobelov norm for the class $\{\hat{g}_\lambda \in \mathcal{G} : \lambda \in \Lambda\}$ is bounded above by its $L_2(P_n)$ norm.

$$I^2(\hat{g}_\lambda) = \int_0^1 \left(\hat{g}_\lambda^{(m)}(z)\right)^2 dz \leq 2\|\hat{g}_\lambda\|_n^2 + 4I^2(\tilde{g}) + 4\|y\|_n^2 \; \forall \lambda \in \Lambda$$

PROBLEM: as defined, it is possible that $I^2(\tilde{g})$ grows with n, which is not okay!

Proof:

Let $\tilde{g}$ satisfy $\tilde{g}(x_i) = y_i$ and have the smallest value for $\int_0^1 \left(\tilde{g}^{(m)}(z)\right)^2 dz$. This function $\tilde{g}$ should always exist.

Case 1: $\lambda \leq 1/2$

By definition of $\hat{g}_\lambda$

$$\|y - \hat{g}_\lambda\|_n^2 + \lambda^2 I^2(\hat{g}_\lambda) \leq \|y - (\tilde{g} - \lambda\hat{g}_\lambda)\|_n^2 + \lambda^2 I^2(\tilde{g} - \lambda\hat{g}_\lambda)$$

Note that

$$\begin{aligned}
I^2(\tilde{g}-\lambda\hat{g}_\lambda) &= \int_0^1 \left(\tilde{g}^{(m)}-\lambda\hat{g}_\lambda^{(m)}\right)^2 dz \\
&= 2\int_0^1 \max\left(\left|\tilde{g}^{(m)}\right|^2, \left|\lambda\hat{g}_\lambda^{(m)}\right|^2\right) dz \\
&= 2\left(\int_0^1 \left|\tilde{g}^{(m)}\right|^2 dz + \int_0^1 \left|\lambda\hat{g}_\lambda^{(m)}\right|^2 dz\right)
\end{aligned}$$

Hence

$$\lambda^2 I^2(\hat{g}_\lambda) \le \lambda^2\|\hat{g}_\lambda\|_n^2 + 2\lambda^2 I^2(\tilde{g}) + 2\lambda^4 I^2(\hat{g}_\lambda)$$

The following ineq follows, where the RHS is maximized when $\lambda = 1/2$

$$I^2(\hat{g}_\lambda) \le \frac{\lambda^2}{\lambda^2 - 2\lambda^4}\left(\|\hat{g}_\lambda\|_n^2 + 2I^2(\tilde{g})\right) \le 2\|\hat{g}_\lambda\|_n^2 + 4I^2(\tilde{g})$$

Case 2: $\lambda > 1/2$
By definition of $\hat{g}_\lambda$

$$\|y-\hat{g}_\lambda\|_n^2 + \lambda^2 I^2(\hat{g}_\lambda) \le \|y\|_n^2$$

The RHS is maximized when $\lambda = 1/2$, so

$$I^2(\hat{g}_\lambda) \le 4\|y\|_n^2$$

Hence we have an upper bound for the Sobelov norm

$$I^2(\hat{g}_\lambda) \le 2\|\hat{g}_\lambda\|_n^2 + 4I^2(\tilde{g}) + 4\|y\|_n^2$$

Appendix

A cute lemma I found but never used: Supposing that $I^v(\hat{g}_\lambda)$ is continuous in λ, then given training data T,

$$\frac{\partial}{\partial\lambda} L_T(\hat{g}_\lambda, \lambda) = 2\lambda I^v(\hat{g}_\lambda)$$

Also, L_T is convex in λ.
Proof:
By definition,

$$L_T(\hat{g}_\lambda, \lambda) = \|y-\hat{g}_\lambda\|_T^2 + \lambda^2 I^v(\hat{g}_\lambda) \le \|y-\hat{g}_{\lambda'}\|_T^2 + \lambda^2 I^v(\hat{g}_{\lambda'}) = L_T(\hat{g}_{\lambda'}, \lambda)$$

Then we can provide upper and lower bounds for $L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1)$:

$$\begin{aligned}
L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1) &\le L_T(\hat{g}_{\lambda_1}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1) \\
&= \|y-\hat{g}_{\lambda_1}\|_T^2 + \lambda_2^2 I^v(\hat{g}_{\lambda_1}) - \|y-\hat{g}_{\lambda_1}\|_T^2 - \lambda_1^2 I^v(\hat{g}_{\lambda_1}) \\
&= (\lambda_2^2 - \lambda_1^2) I^v(\hat{g}_{\lambda_1})
\end{aligned}$$

$$\begin{aligned}
L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1) &\ge L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_2}, \lambda_1) \\
&= \|y-\hat{g}_{\lambda_2}\|_T^2 + \lambda_2^2 I^v(\hat{g}_{\lambda_2}) - \|y-\hat{g}_{\lambda_2}\|_T^2 - \lambda_1^2 I^v(\hat{g}_{\lambda_2}) \\
&= (\lambda_2^2 - \lambda_1^2) I^v(\hat{g}_{\lambda_2})
\end{aligned}$$

So suppose WLOG $\lambda_2 > \lambda_1$:

$$(\lambda_2 + \lambda_1) I^v(\hat{g}_{\lambda_2}) \leq \frac{L_T(\hat{g}_{\lambda_2}, \lambda_2) - L_T(\hat{g}_{\lambda_1}, \lambda_1)}{\lambda_2 - \lambda_1} \leq (\lambda_2 + \lambda_1) I^v(\hat{g}_{\lambda_1})$$

So as $\lambda_1 \to \lambda_2 = \lambda$, we have by the sandwich theorem,

$$\frac{\partial}{\partial \lambda} L_T(\hat{g}_\lambda, \lambda) = 2\lambda I^v(\hat{g}_\lambda)$$

Furthermore, given training data T

$$\frac{\partial}{\partial \lambda} L_T(\hat{g}_\lambda, \lambda) = \frac{\partial}{\partial \lambda} \|y - \hat{g}_\lambda\|_T^2 + 2\lambda I^v(\hat{g}_\lambda) + \lambda^2 \frac{\partial}{\partial \lambda} I^v(\hat{g}_\lambda)$$

then, combining this with the lemma, we have that

$$\frac{\partial}{\partial \lambda} \|y - \hat{g}_\lambda\|_T^2 = -\lambda^2 \frac{\partial}{\partial \lambda} I^v(\hat{g}_\lambda)$$

Finally, to see that L_T is convex in λ, note that

$$\frac{\partial^2}{\partial \lambda^2} L_T(\hat{g}_\lambda, \lambda) = 2I^v(\hat{g}_\lambda) + 2\lambda v I^{v-1}(\hat{g}_\lambda) \frac{\partial}{\partial \lambda} I(\hat{g}_\lambda) > 0$$

since $\frac{\partial}{\partial \lambda} I(\hat{g}_\lambda) > 0$.

OLD

Lemma 3:

Suppose the function class $\mathcal{F}$ is bounded s.t. $\sup_{f \in \mathcal{F}} \|f\|_n \leq R < \infty$. Let

$$\tilde{\mathcal{F}} = \{\gamma f \; : \; f \in \mathcal{F}, \gamma \in (0, 1]\}$$

$$H\left(\delta(1 + R + \delta), \tilde{\mathcal{F}}, \|\cdot\|_n\right) \leq \log(1 + \lfloor \frac{1}{\delta} \rfloor) + H\left(\delta, \mathcal{F}, \|\cdot\|_n\right)$$

Proof: Let $\{h_i\}_{i=1}^N$ be the δ-cover for $\mathcal{F}$. Consider any $f \in \mathcal{F}$ and let $h_{(f)}$ be the closest function in δ-cover for $\mathcal{F}$. Choose $j \in \mathbb{Z}^+$ such that $|\gamma - \delta j| < \delta$.

$$\begin{aligned}
\|\gamma f - \delta j h_{(f)}\|_n &\leq \|\gamma f - \gamma h_{(f)}\|_n + \|\gamma h_{(f)} - \delta j h_{(f)}\|_n \\
&\leq \gamma \|f - h_{(f)}\|_n + |\gamma - \delta j| \|h_{(f)}\|_n \\
&\leq \gamma \delta + \delta \left(\|f - h_{(f)}\|_n + \|f\|_n\right) \\
&\leq \gamma \delta + \delta (\delta + R) \\
&\leq \delta (1 + R + \delta)
\end{aligned}$$

Hence we have found that the following $N(1 + \lfloor \frac{1}{\delta} \rfloor)$ functions form a $\delta(1 + R + \delta)$-cover for $\tilde{\mathcal{F}}$:

$$\{h_i\}_{i=1}^N \cup \left\{ j\delta h_i \; : \; j \in 1 : \lfloor \frac{1}{\delta} \rfloor, i \in 1 : N \right\}$$

Lemma 4:

Define function classes $\{\mathcal{F}_j\}_{j=1}^J$ and

$$\tilde{\mathcal{F}} = \left\{ \sum_{j=1}^{J} f_j \ : \ f_j \in \mathcal{F}_j \right\}$$

Then

$$H\left(J\delta, \tilde{\mathcal{F}}, \|\cdot\|_n\right) \le \sum_{j=1}^{J} H\left(\delta, \mathcal{F}_j, \|\cdot\|_n\right)$$

Proof: For every $j = 1 : J$, consider any $f_j \in \mathcal{F}_j$ and let $h_{(j)}$ be the closest function in the δ-cover for $\mathcal{F}_j$.

$$\|\sum_{j=1}^{J} f_j - \sum_{j=1}^{J} h_{(j)}\| \quad \le \quad \sum_{j=1}^{J} \|f_j - h_{(j)}\| \le J\delta$$

Hence $\exp\left(\sum_{j=1}^{J} H\left(\delta, \mathcal{F}_j, \|\cdot\|_n\right)\right)$ functions form a $J\delta$-cover for $\tilde{\mathcal{F}}$.

Lemma 5:

Suppose for all $j = 1, ..., J$, there is some $\alpha_j > 0$ and $A_j > 0$ s.t. the following entropy bound holds for all $\delta > 0$

$$H\left(\delta, \left\{ \frac{g_j - g_j^*}{I(g_j) + I(g_j^*)} : g_j \in \mathcal{G}_j, I(g_j) + I(g_j^*) > 0 \right\}, \|\cdot\|_T \right) \le A\delta^{-\alpha_j}$$

Then for sufficiently small$\delta > 0$, we have

$$H\left(\delta, \left\{ \frac{\sum_{j=1}^{J} g_j - g_j^*}{\sup_{j\in 1:J}\left(I(g_j) + I(g_j^*)\right)} : g_j \in \mathcal{G}_j, I(g_j) + I(g_j^*) > 0 \right\}, \|\cdot\|_T \right) \le 2JA\left(\frac{\delta}{2J\left(1+R\right)}\right)^{-\alpha_{max}}$$

where $\alpha_{max} = \max_{j\in 1:J} \alpha_j$.

Proof: By Lemma 3,

$$H\left(\delta(1+R+\delta), \left\{ \gamma\frac{g_j - g_j^*}{I(g_j) + I(g_j^*)} \ : \ g_j \in \mathcal{G}_j, I(g_j) + I(g_j^*) > 0, \gamma \in (0,1] \right\}, \|\cdot\|_T \right) \le \log(1+\lfloor\frac{1}{\delta}\rfloor) + A\delta^{-\alpha_j}$$

Note that

$$\frac{\sum_{j=1}^{J} g_j - g_j^*}{\sup_{j\in 1:J}\left(I(g_j) + I(g_j^*)\right)} = \sum_{j=1}^{J}\left(\frac{I(g_j) + I(g_j^*)}{\sup_{\ell\in 1:J} I(g_\ell) + I(g_\ell^*)}\right)\frac{g_j - g_j^*}{I(g_j) + I(g_j^*)}$$

By Lemma 4,

$$H\left(J\delta(1+R+\delta), \left\{ \frac{\sum_{j=1}^{J} g_j - g_j^*}{\sup_{j\in 1:J}\left(I(g_j) + I(g_j^*)\right)} \ : g_j \in \mathcal{G}_j, I(g_j) + I(g_j^*) > 0 \right\}, \|\cdot\|_T \right) \le J\log(1+\lfloor\frac{1}{\delta}\rfloor) + JA\delta^{-\alpha_j}$$

Hence for sufficiently small δ,

$$H\left(J\delta(1+R+\delta), \left\{\frac{\sum_{j=1}^{J} g_j - g_j^*}{\sup_{j\in 1:J}\left(I(g_j)+I(g_j^*)\right)} \ : g_j \in \mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\}, \|\cdot\|_T\right) \leq 2JA\delta^{-\alpha_{max}}$$

Rearranging, we get

$$\begin{aligned} H\left(\delta, \left\{\frac{\sum_{j=1}^{J} g_j - g_j^*}{\sup_{j\in 1:J}\left(I(g_j)+I(g_j^*)\right)} : g_j \in \mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\}, \|\cdot\|_T\right) &\leq 2AJ\left(\sqrt{\left(\frac{1+R}{2}\right)^2+\frac{\delta}{J}}-\frac{1+R}{2}\right)^{-\alpha_{max}} \\ &\leq 2AJ\left(\frac{\delta}{2J\,(1+R)}\right)^{-\alpha_{max}} \end{aligned}$$

(Used the fact that for $b>0$ small enough, $\sqrt{a^2+b}-a \geq \sqrt{(a+\frac{b}{4a})^2}-a = \frac{b}{4a}$)

Lemma 5b:

Suppose for all $j=1,...,J$, there is some $\alpha_j>0$ and $A_j>0$ s.t. the following entropy bound holds for all $\delta>0$

$$H\left(\delta, \left\{\frac{g_j - g_j^*}{I(g_j)+I(g_j^*)} : g_j \in \mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\}, \|\cdot\|\right) \leq A\delta^{-\alpha_j}$$

Then for sufficiently small$\delta>0$, we have

$$H\left(\delta, \left\{\frac{\sum_{j=1}^{J} g_j - g_j^*}{\sum_{j=1}^{J} I(g_j)+I(g_j^*)} : g_j \in \mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\}, \|\cdot\|\right) \leq 2JA\,(STUFF)^{-\alpha_{max}}$$

where $\alpha_{max} = \max_{j\in 1:J}\alpha_j$.

Proof: Note that

$$\frac{\sum_{j=1}^{J} g_j - g_j^*}{\sum_{j=1}^{J} I(g_j)+I(g_j^*)} = \sum_{j=1}^{J}\left(\frac{I(g_j)+I(g_j^*)}{\sum_{j=1}^{J} I(g_\ell)+I(g_\ell^*)}\right)\frac{g_j-g_j^*}{I(g_j)+I(g_j^*)}$$

So we can express

$$\left\{\frac{\sum_{j=1}^{J} g_j - g_j^*}{\sum_{j=1}^{J} I(g_j)+I(g_j^*)} : g_j \in \mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\} \subseteq \left\{\sum_{j=1}^{J}\gamma_j\frac{g_j-g_j^*}{I(g_j)+I(g_j^*)} : g_j \in \mathcal{G}_j, I(g_j)+I(g_j^*)>0, \sum_{j=1}^{J}\gamma_j = 1\right\}$$

Let $\mathcal{H}_j$ be the set of functions that form a δ-cover for $\left\{\frac{g_j-g_j^*}{I(g_j)+I(g_j^*)} : g_j \in \mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\}$. Consider the set of functions

$$\left\{\sum_{j=1}^{J}\delta k_j h_j \ : \ h_j \in \mathcal{H}_j, 1-\frac{1}{\delta} \leq \delta\textstyle\sum_{j=1}^{J} k_j \leq 1, k_j \in 1:\lfloor\frac{1}{\delta}\rfloor\right\}$$

Let $|\delta k_j - \gamma_i| < \delta/2$. Then

$$\begin{aligned}
\left\|\sum_{j=1}^{J}\gamma_j\frac{g_j-g_j^*}{I(g_j)+I(g_j^*)}-\sum_{j=1}^{J}\delta k_j h_j\right\| &\le \sum_{j=1}^{J}\left\|\gamma_j\frac{g_j-g_j^*}{I(g_j)+I(g_j^*)}-\delta k_j h_j\right\| \\
&\le \sum_{j=1}^{J}\left\|\gamma_j\frac{g_j-g_j^*}{I(g_j)+I(g_j^*)}-\gamma_i h_j\right\|+|\delta k_j-\gamma_i|\|h_j\| \\
&\le \sum_{j=1}^{J}\left(\gamma_j\delta+\frac{\delta}{2}\left(\left\|\frac{g_j-g_j^*}{I(g_j)+I(g_j^*)}-h_j\right\|+\left\|\frac{g_j-g_j^*}{I(g_j)+I(g_j^*)}\right\|\right)\right) \\
&\le \delta(1+JR+J\delta)
\end{aligned}$$

Hence these $\left(\Pi_{j=1}^{J}N_j\right)\binom{\lfloor\frac{1}{\delta}\rfloor+J-1}{J-1}$ functions form a $\delta(1+JR+J\delta)$ cover. Hence the entropy is

$$H\left(\delta(1+JR+J\delta),\left\{\frac{\sum_{j=1}^{J}g_j-g_j^*}{\sum_{j=1}^{J}I(g_j)+I(g_j^*)}\ :g_j\in\mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\},\|\cdot\|\right)\le(J-1)\log(1+J+\lfloor\frac{1}{\delta}\rfloor)+A\sum_{j=1}^{J}\delta^{-\alpha_j}$$

Note:

$$\binom{\lfloor\frac{1}{\delta}\rfloor+J-1}{J-1}\le\left(\lfloor\frac{1}{\delta}\rfloor+J-1\right)^{J-1}$$

Hence for sufficiently small δ,

$$H\left(\delta(1+JR+J\delta),\left\{\frac{\sum_{j=1}^{J}g_j-g_j^*}{\sum_{j=1}^{J}I(g_j)+I(g_j^*)}\ :g_j\in\mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\},\|\cdot\|\right)\le 2JA\delta^{-\alpha_{max}}$$

Rearranging, we get

$$\begin{aligned}
H\left(\delta,\left\{\frac{\sum_{j=1}^{J}g_j-g_j^*}{\sum_{j=1}^{J}I(g_j)+I(g_j^*)}\ :g_j\in\mathcal{G}_j, I(g_j)+I(g_j^*)>0\right\},\|\cdot\|\right) &\le 2AJ\left(\frac{-JR+1+\sqrt{(JR+1)^2+4\delta J}}{2J}\right)^{-\alpha_{max}} \\
&\le 2AJ\left(\frac{\sqrt{2}\delta J^{3/2}}{1+JR}\right)^{-\alpha_{max}}
\end{aligned}$$

(Used the fact that for $b>0$ small enough, $\sqrt{a^2+b}-a\ge\sqrt{(a+\frac{b}{4a})^2}-a=\frac{b}{4a}$)

Lemma 6:

Suppose ϵ_i are sub-gaussian errors and for the function class $\mathcal{F}$, we have that for some $0<\alpha<2$, $A'>0$, and $J>0$

$$H\left(\delta,\mathcal{F},\|\cdot\|_T\right)\le A'J^{\tau}\delta^{-\alpha}\ \forall\delta>0$$

Then for $T=2C_1CA'^{1/2}J^{\tau/2}2^{1-\alpha/2}$

$$Pr\left(\sup_{f\in\mathcal{F}}\frac{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(z_i)\right|}{\|f\|_n^{1-\alpha/2}}\ge T\right)\le c\exp(-T^2/c^2)$$

Proof: Follow proof for Lemma 8.4 in Vandegeer, but with $A = A'J^{-\alpha}$. Note that we then have $A_0 = A'^{1/2}J^{\tau/2}$. We then get

$$Pr\left(\sup_{f\in\mathcal{F}} \frac{\left|\frac{1}{\sqrt{n}}\sum_{i=1}^{n}\epsilon_i f(z_i)\right|}{\|f\|_n^{1-\alpha/2}} \geq 2C_1CA'^{1/2}J^{\tau/2}2^{1-\alpha/2}\right) \leq c\exp(-T^2/c^2)$$

Note that we can write via shorthand that

$$\sup_{f\in\mathcal{F}} \frac{\left|\frac{1}{n}\sum_{i=1}^{n}\epsilon_i f(z_i)\right|}{\|f\|_n^{1-\alpha/2}} = O_p(J^{\tau/2}n^{-1/2})$$