1 Simple model

Definitions

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

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

where ϵ are sub-Gaussian errors for constants K and σ_0^2

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

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 V , 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 the entropy of the class $\mathcal{G}'$ is

$$H\left(\delta, \mathcal{G}' = \left\{ \frac{g - g^*}{I(g) + I(g^*)} : g \in \mathcal{G}, I(g) + I(g^*) > 0 \right\}, P_T \right) \quad \leq \quad \tilde{A}\delta^{-\alpha} \tag{1}$$

Suppose $v > 2\alpha/(2+\alpha)$.

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 $X_T = X_V$

Suppose $P_T = P_V = P_n = \{X_i\}_{i=1}^n$ are the same between the validation and training set.

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

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

Then

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

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

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

(Note: Here we've assumed function I is from the optimization criterion, but that is not necessary!)

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

Suppose that the training and validation set are independently sampled, so 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 entropy bound (1) for both training set P_T and validation set P_V.

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:

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

Then

$$\|\hat{g}_{\hat{\lambda}} - g^*\|_V = O_p(n^{-1/(2+\alpha)})\left(I(g^*) \vee I^{(4a-4v+a^2v)/2a(2+a)}(g^*) \vee M^{\alpha v/(2+\alpha)}\|g^*\|_V^{2\alpha/v(2+\alpha)}\right)$$

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 $\xi > 0$,

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

Therefore

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

where

$$\tilde{\lambda}^{-1} = O_p(n^{-1/(2+\alpha)})I^{(2v-2\alpha+\alpha v)/2(2+\alpha)}(g^*)$$

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

Same proof still holds.

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

Again, by Vandegeer (10.6),

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

Case 3a: $I(g^*) \geq I(g_{\hat{\lambda}})$

The result is clearly attained.

Case 3b: $I(g^*) \leq I(g_{\hat{\lambda}})$

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

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

2 General Additive Model

Definitions

We find the best model for y over function classes $\mathcal{G}_j$. Suppose we observe:

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

where $g_j^* \in \mathcal{G}_j$ are the true functions. ϵ are sub-Gaussian errors for constants K and σ_0^2

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

Given a training set T , we fit models by least squares with multiple penalties

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

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

$$\{\hat{g}_{\hat{\lambda},j}\}_{j=1}^{J} = \arg\min_{\lambda} \|y - \sum_{j=1}^{J} \hat{g}_{\lambda,j}\|_V^2$$

Reasonable assumption:

- The entropy bound (2) in result 3 comes from the assumptions in Lemma 3. The α below is $\alpha = \max_{j=1:J}\{\alpha_j\}$, so convergence is only as fast as fitting the highest-entropy function class. The constant A must be appropriately inflated such that the entropy bound holds for all $\delta \in (0, R]$.

Special assumptions:

- I assume exponents v_k in the optimization criteria are greater than one, whereas Vandegeer Thrm 10.2 only assumes $v > 2\alpha/(2+\alpha)$. Without this assumption, I wasn't able to form inequalities between $\sum_{j=1}^{J} I_j(g_j) \leq O_p(1) + \sum_{j=1}^{J} I_j^{v_j}(g_j)$. To remove this assumption, we need something else in the denominator of the entropy bound. (Currently, I use $\sum_{j=1}^{J} I_j(g_j) + I_j(g_j^*)$).
- In Result 1 and 2, I bounded $\left|(\epsilon_V, g^* - \hat{g}_{\hat{\lambda}})\right|$ by assuming the penalty function $I^v(g)$ was upper bounded by $\|g\|_n^2$. However, that isn't enough for the case of additive penalties. I've assumed that there is some function $\tilde{I} : \left\{ \sum_{j=1}^{J} g_j : g_j \in \mathcal{G}_j \right\} \mapsto \mathbb{R}$ such $I^v(\sum_{j=1}^{J} g_j)$ is upper bounded by $\|\sum_{j=1}^{J} g_j\|_n^2$ AND it gives the same entropy bound.

Result 3: Additive Model with multiple penalties, Single oracle λ over X_T

Suppose there is some $0 < \alpha < 2$ s.t. for all $\delta \in (0, R]$,

$$H\left(\delta, \left\{ \frac{\sum_{j=1}^{J} g_j - g_j^*}{\sum_{j=1}^{J} I_j(g_j) + I_j(g_j^*)} : g_j \in \mathcal{G}_j, \sum_{j=1}^{J} I_j(g_j) + I_j(g_j^*) > 0 \right\}, \|\cdot\|_T \right) \leq A\delta^{-\alpha} \tag{2}$$

Special assumption: Suppose $v_k \geq 1$ for all k.
If λ is chosen s.t.

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

then

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

and

$$\sum_{k=1}^{K} I_k(\{\hat{g}_j\}) \leq K + \sum_{j=1}^{J} I_j(g_j^*)$$

Proof:

The basic inequality gives us:

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

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

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

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

By Vandegeer (10.6), the basic inequality becomes

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

Case 2a: $\sum_{j=1}^{J} I_j(g_j) \leq \sum_{j=1}^{J} I_j(g_j^*)$

Then

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

Case 2b: $\sum_{j=1}^{J} I_j(g_j) \geq \sum_{j=1}^{J} I_j(g_j^*)$

First note that for exponent $v \geq 1$, we must have $x \leq x^v + 1$. So by assuming $v_j \geq 1$,

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

Case 2b part a: 2nd term on the RHS in the inequality above is bigger

Then

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

which implies

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

and

$$\sum_{j=1}^{J} I_j(\hat{g}_j) \le J + \sum_{j=1}^{J} I_j^{v_j}(g_j^*)$$

Case 2b part b: 1st term on the RHS in the inequality above is bigger
Then

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

Result 4: Additive Model with multiple penalties, Single cross-validation λ over general X_T, X_V

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

In addition to the assumptions in Result 3, assume the following:

Suppose the same entropy bound (2) for both the training set P_T and validation set P_V.

Suppose the infinity norm is also bounded

$$\sup_{g_j \in \mathcal{G}_j} \frac{\|\sum_{j=1}^{J} g_j - g_j^*\|_\infty}{\sum_{j=1}^{J} I_j(g_j) + I_j(g_j^*)} \quad \le \quad K < \infty$$

Special Assumption: Suppose there is some function $\tilde{I} : \left\{\sum_{j=1}^{J} g_j : g_j \in \mathcal{G}_j\right\} \mapsto \mathbb{R}$ such

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

Furthermore, suppose there exist constants M,M_0, and $w > 2\alpha/(2+\alpha)$ s.t. for all $\lambda \in \Lambda$

$$\tilde{I}^w\left(\sum_{j=1}^{J}\hat{g}_{\lambda,j}\right) \le M\|\sum_{j=1}^{J}\hat{g}_{\lambda,j}\|_V^2 + M_0$$

Then

$$\|\sum_{j=1}^{J}\hat{g}_{\hat{\lambda},j} - g_j^*\|_V = O_p(n^{-1/(2+\alpha)})\left[\left(K + \sum_{k=1}^{K} I_k^{v_k}(\{g_j^*\})\right) \vee \left(M^{\alpha w/(2+\alpha)}\|\sum_{j=1}^{J} g_j^*\|_V^{2\alpha/w(2+\alpha)}\right)\right]$$

Proof:

The proof is very similar to Result 2.

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

By Lemma 2, we have

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

Hence

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

Therefore

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

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 $O_p(n^{-1/2-1/(2+\alpha)})$. (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.

By our **special assumption**, we can again apply Vandegeer (10.6),

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

Case 3a:$\tilde{I}(\{\hat{g}_{\hat{\lambda},j}\})\leq\tilde{I}(\{g_j^*\})$
The result is clearly attained.
Case 3b:$\tilde{I}(\{\hat{g}_{\hat{\lambda},j}\})>\tilde{I}(\{g_j^*\})$
By the assumption that $\tilde{I}^w(\sum_{j=1}^J\hat{g}_{\hat{\lambda},j})$ is bounded above by $\|\sum_{j=1}^J\hat{g}_{\hat{\lambda},j}\|_V^2$, Lemma 1 gives us

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

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:

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

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

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

Then

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

Proof:
From the assumptions, we have

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

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

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

Case 1: $\|g_\lambda - g^*\|_n \ge \|g^*\|_n$
Then

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

Note that $\sup_v -\frac{v}{2v+\alpha v-2\alpha} = -\frac{1}{2+\alpha}$, so this rate is faster than $O_p(n^{-\frac{1}{2+\alpha}})$.
Case 2: $\|g_\lambda - g^*\|_n \le \|g^*\|_n$
Then

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

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 $P_{2n} = \frac{1}{2}\left(P_{n'} + P_{n''}\right)$. 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

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

and

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

Then

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

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

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

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

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

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

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

Now note that

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

By Hoeffding's inequality,

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

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

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

Hence

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

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

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

Using shorthand, we can write

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

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

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

Hence

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