Definitions

We presume g^* is true model and

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

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

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

We will be minimizing $\arg\min_{g \in \mathcal{G}} \|y - g\|_T^2 + \lambda^2 I^v(g)$ to obtain fitted models $\hat{g}_\lambda$.
We will also restrict $\lambda > \lambda_{min} = O_p(n^{-t})$ for a t we will specify later.

Goal:

Bound

$$Pr\left(\|\hat{g}_{\hat{\lambda}} - g^*\|_V \geq \delta \right) \leq ???$$

Proof

Consider the class

$$\mathcal{G}' = \left\{ \frac{g - g^*}{I(g) + I(g^*)} : g \in \mathcal{G}, I(g) + I(g^*) > 0 \right\}$$

Suppose this class is bounded and its entropy is for $\alpha \in (0, 2)$

$$H\left(\delta, \mathcal{G}', Q_n\right) \leq A\delta^{-\alpha} \forall \delta > 0, n \geq 1$$

Alo note that the class

$$\mathcal{G}\text{"} = \left\{ \left(\frac{g - g^*}{I(g) + I(g^*)} \right)^2 : g \in \mathcal{G}, I(g) + I(g^*) > 0 \right\}$$

must also be bounded. For some other constant $\tilde{A}$, we have that its entropy is bounded above by (proof in the mini appendix below)

$$H\left(\delta, \mathcal{G}\text{"}, Q_n\right) \leq \tilde{A}\delta^{-\alpha} \forall \delta > 0, n \geq 1$$

Concentration inequality 1:

By Lemma 8.4, since ϵ is sub-gaussian and we've assumed that $\mathcal{G}'$ is bounded $\left(\sup_{g' \in \mathcal{G}'} \|g'\|_n \leq R\right)$ then for some constant c depending on A,α, R,K,σ_0, we have for all $\delta\sqrt{n} \geq c$

$$Pr\left(\sup_{g \in \mathcal{G}} \frac{|(\epsilon, g - g^*)_n|}{\|g - g^*\|^{1-\alpha/2} \left(I(g) + I(g^*)\right)^{\alpha/2}} > \delta \right) \leq c \exp\left(-\frac{\delta^2 n}{c^2} \right)$$

Concentration inequality 2:

Now consider two sets of samples $\{X_i\}_{i=1}^n, \{X_i'\}_{i=1}^n$. We are interested in the concentration inequality for

$$\frac{\left| \|g - g^*\|_n^2 - \|g - g^*\|_{n'}^2 \right|}{\left(I(g) + I(g^*)\right)^2}$$

where $\|g - g^*\|_{n'}^2 = \sum_{i=1}^n (g - g^*)^2(X_i')$.

Using the Rademacher sequence $\{W_i\}_{i=1}^n$, we know that

$$\begin{aligned} Pr\left(\sup_{g\in\mathcal{G}} \frac{\left|\|g-g^*\|_n^2 - \|g-g^*\|_{n'}^2\right|}{\left(I(g)+I(g^*)\right)^2} > \delta\right) &= Pr\left(\sup_{g\in\mathcal{G}} \frac{\left|\frac{1}{n}\sum_{i=1}^n W_i\left((g-g^*)^2(X_i) - (g-g^*)^2(X_i')\right)\right|}{\left(I(g)+I(g^*)\right)^2} > \delta\right) \\ &\leq 2Pr\left(\sup_{g\in\mathcal{G}} \frac{\left|\frac{1}{n}\sum W_i(g-g^*)^2(X_i)\right|}{\left(I(g)+I(g^*)\right)^2} > \delta/2\right) \end{aligned}$$

By Lemma 3.2, since the Rademacher sequence is sub-Gaussian and we've assumed that $\mathcal{G}$" is bounded $\left(\sup_{g"\in\mathcal{G}"}\|g"\|_n \leq R^2\right)$, then there exists constants C, A_0 s.t.

$$\delta\sqrt{n} \geq A_0\delta^{1-\alpha/2} \geq C\left(\int_0^\delta H^{1/2}(u,\mathcal{G}",Q_n)du \vee R^2\right)$$

That is, for all

$$\delta \geq A_0^{2/\alpha} n^{-1/\alpha}$$

there is some constant c depending only on A_0 and α

$$Pr\left(\sup_{g\in\mathcal{G}} \frac{\left|\frac{1}{n}\sum W_i(g-g^*)^2(X_i)\right|}{\left(I(g)+I(g^*)\right)^2} > \delta\right) \leq c\exp\left(-\frac{n\delta^2}{c^2R^2}\right)$$

That is,

$$Pr\left(\sup_{g\in\mathcal{G}} \frac{\left|\|g-g^*\|_n^2 - \|g-g^*\|_{n'}^2\right|}{\left(I(g)+I(g^*)\right)^2} > \delta\right) \leq \frac{c}{2}\exp\left(-\frac{n\delta^2}{4c^2R^2}\right)$$

Construct our high probability set $\mathcal{T}$

Let $\delta = o_p(n^{-1/2})$. Consider the set

$$\begin{aligned} \mathcal{T} &= \left\{\{X_i\}_{i=1}^n, \{X_i'\}_{i=1}^{n'} \text{where the conditions (1),(2),(3) hold}\right\} \\ (1)\quad &\sup_g \frac{\left|\|g-g^*\|_n^2 - \|g-g^*\|_{n'}^2\right|}{\left(I(g)+I(g^*)\right)^2} \leq \delta \\ (2)\quad &\sup_g \frac{|(\epsilon, g-g^*)_{n'}|}{\|g-g^*\|_{n'}^{1-\alpha/2}\left(I(g)+I(g^*)\right)^{\alpha/2}} \leq \delta \\ (3)\quad &\sup_g \frac{|(\epsilon, g-g^*)_{n}|}{\|g-g^*\|_{n}^{1-\alpha/2}\left(I(g)+I(g^*)\right)^{\alpha/2}} \leq \delta \end{aligned}$$

This set occurs with high probability on the order of $Pr(\mathcal{T}) = c\exp\left(-O_p(1)\frac{\delta^2 n}{c^2}\right)$ as shown by the concentration inequalities given above. Hence we can now suppose our training and validation set come from $\mathcal{T}$.

Define the following:

- $\hat{g}_\lambda \equiv \arg\min_{g\in\mathcal{G}} \|y-g\|_T^2 + \lambda^2 I^v(g)$ as the minimizer of the penalized loss on the training set.
- $\hat{\lambda} \equiv \arg\min_{\lambda\in\Lambda} \|y-\hat{g}_\lambda\|_V^2$ as the minimizer of the loss on the validation set (but constrained to minimizers of the training set).
- $\tilde{\lambda}$ as the penalty parameter that attains the asymptotically optimal convergence rate. By Theorem 10.2, assuming $I(g^*) > 0$ and $v > \frac{2\alpha}{2+\alpha}$, we have chosen $\tilde{\lambda}$ to satisfy

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

Show $\hat{g}_{\hat{\lambda}}$ behaves well on $\mathcal{T}$

By definition, we have

$$\|y - \hat{g}_{\hat{\lambda}}\|_V^2 \leq \|y - \hat{g}_{\tilde{\lambda}}\|_V^2$$

By adding and subtracting g^* in the squared norms, we have

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

Case 1: $\|g^* - \hat{g}_{\tilde{\lambda}}\|_V^2$ is the largest term on the RHS
On the set $\mathcal{T}$, we have

$$\left|\|g^* - \hat{g}_{\hat{\lambda}}\|_V^2 - \|g^* - \hat{g}_{\tilde{\lambda}}\|_T^2\right| \leq \delta\left(I(\hat{g}_{\tilde{\lambda}}) + I(g^*)\right)^2$$

Since $\|\hat{g}_{\tilde{\lambda}} - g^*\|_T = O_p(\tilde{\lambda})I^{v/2}(g^*)$, then

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

Since we also know the order of λ^*, we have

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

Here, we are looking at a convergence rate of

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_V = O_p(n^{-1/4})I(g^*)$$

or

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

Case 2: $\left|2(\epsilon, g^* - \hat{g}_{\tilde{\lambda}})_V\right|$ is the largest term on the RHS
On set $\mathcal{T}$, we have

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

Hence

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

Here we are looking at a convergence rate of

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_V = O_p(n^{(\alpha-3)/4})I^{2-\alpha/2}(g^*)$$

Case 3: $\left|2(\epsilon, \hat{g}_{\hat{\lambda}} - g^*)_V\right|$ is the largest term on the RHS

On set $\mathcal{T}$, we have

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

So

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_V^2 \leq 6\delta \left\|\hat{g}_{\hat{\lambda}} - g^*\right\|_V^{1-\alpha/2} \left(I(\hat{g}_{\hat{\lambda}}) + I(g^*)\right)^{\alpha/2}$$

Dividing both sides, we get

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

This is tricky since $I(\hat{g}_{\hat{\lambda}})$ is unknown.

Let's add the assumption that $\lambda_{min} = O_p(n^{-t})$ for some $t > 0$. We must make sure that $\lambda_{min} \leq \tilde{\lambda}$. Let's consider these two cases:

Case 3a: $I(\hat{g}_{\hat{\lambda}}) \geq I(\hat{g}_{\tilde{\lambda}})$

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

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

which implies that

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

Case 3aa: If $\|y - \hat{g}_{\tilde{\lambda}}\|_T^2 \leq \hat{\lambda}^2 I^v(\hat{g}_{\tilde{\lambda}})$, then $I^v(\hat{g}_{\hat{\lambda}}) \leq 2I^v(\hat{g}_{\tilde{\lambda}})$. Refer to Case 3b below to see that

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

Case 3ab: If $\|y - \hat{g}_{\tilde{\lambda}}\|_T^2 \geq \hat{\lambda}^2 I^v(\hat{g}_{\tilde{\lambda}})$, then $\hat{\lambda}^2 I^v(\hat{g}_{\hat{\lambda}}) \leq 2\|y - \hat{g}_{\tilde{\lambda}}\|_T^2$. Since we know that $\|y - \hat{g}_{\tilde{\lambda}}\|_T = O_p(\tilde{\lambda}) I^{v/2}(g^*)$. Hence we find that

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

Plugging the above into the inequality

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

we get

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

Since $\delta = O(n^{-1/2})$, then we just need t s.t.

$$(2t - \frac{2}{2+\alpha})\frac{\alpha}{(2+\alpha)v} - \frac{1}{2+\alpha} < 0$$

in order to have $\|g^* - \hat{g}_{\hat{\lambda}}\|_V \to 0$ as $n \to \infty$. Rearranging, we get that we must choose t s.t.

$$t < \frac{v}{2\alpha} + \frac{1}{2+\alpha}$$

We check that indeed, we can choose $\lambda_{min} \leq \tilde{\lambda}$ since $\tilde{\lambda} = O_p(n^{-1/(2+\alpha)}) I^{(2v-2\alpha+v\alpha)/(2(2+\alpha))}(g^*)$.

Let $\theta \in (0, 1)$. Then reparameterizing $t = \theta\frac{v}{2\alpha} + \frac{1}{2+\alpha}$, we get the following convergence rate:

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

So if we choose θ close to 1, then we get super slow convergence rate. If we choose θ close to 0, then we are essentially requiring $\lambda_{min} = \tilde{\lambda}$, which gives us a great convergence rate but the value λ_{min} will be hard to determine. If we choose $\theta = 1/2$, then we get a convergence rate about $O_p(n^{-1/(4+2\alpha)})$. There's a tradeoff between choosing λ_{min} as small as possible and the convergence rate.

Case 3b: $I(\hat{g}_{\hat{\lambda}}) \leq I(\hat{g}_{\tilde{\lambda}})$

We're all happy in this case since we know that $I(g_{\tilde{\lambda}}) = O_p(1)I(g^*)$:

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

So we have convergence rates of either

$$\|g^* - \hat{g}_{\hat{\lambda}}\|_V = O_p(n^{-\frac{\theta}{2+\alpha}}) I^{\frac{\alpha}{2+\alpha}\frac{2\alpha}{v(2+\alpha)}}(g^*)$$

or

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

Summary

From the three cases, we've found that $\|g^* - \hat{g}_{\hat{\lambda}}\|_V$ converges the slowest in case 3ab, for $\theta \in (0, 1)$:

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

Mini Appendix

Lemma

Define function classes $\mathcal{G}' = \{f\}$ and $\mathcal{G}" = \{f^2\}$ and let Q_n be an empirical measure. Suppose $\|f\|_{Q_n}^2 < R < \infty \forall f \in \mathcal{G}'$. Then for some constant K, we have

$$H\left(\delta K, \mathcal{G}", Q_n\right) \leq H\left(\delta, \mathcal{G}', Q_n\right)$$

Proof

Let the δ-cover set for $\mathcal{G}'$ be $\{f_1, ..., f_N\}$. Consider any function $f \in \mathcal{G}'$. WLOG, suppose

$$\frac{1}{n}\sum (f - f_1)^2(x_i) \leq \delta$$

Note that

$$\begin{aligned} \sum \left|f^2 - f_1^2\right|(x_i) &= \sum \left|(f - f_1)(f + f_1)\right|(x_i) \\ &\leq \sqrt{\left(\sum (f - f_1)^2(x_i)\right)\left(\sum (f + f_1)^2(x_i)\right)} \\ &\leq n\sqrt{\delta K} \end{aligned}$$

Hence

$$\sum \left|f^2 - f_1^2\right|^2(x_i) \leq \left(\sum \left|f^2 - f_1^2\right|(x_i)\right)^2 \leq n^2 \delta K$$

That is,

$$\|f^2 - f_1^2\|_{Q_n} \leq \delta K$$