Training/Validation Split Theorem

November 4, 2016

We are interested in bounding the error of the selected model when tuning penalty parameters by a training validation split. We will concern ourselves with the error over the observed covariates in the validation set. Under sufficient entropy conditions, the error of the selected model will converge to the error of the oracle.

We will suppose that the data is generated from the model:

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

where ϵ are independent, sub-gaussian errors. The penalized regression models are

$$\hat{g}(\cdot|\boldsymbol{\lambda}) = \arg\min_{g \in \mathcal{G}} L_T(g|\boldsymbol{\lambda})$$

Let the model class after fitting on the training data be

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

The selected penalty parameters are

$$\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \|y - \hat{g}(\cdot|\boldsymbol{\lambda})\|_V^2$$

Suppose the "oracle" penalty parameters are

$$\tilde{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V$$

We will provide sharp oracle inequalities of the form

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \le \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 + \delta$$

The document is organized as follows

1. Theorem 3 proves the convergence of the selected model to the best model under general entropy conditions where the convergence rate δ is given as a random variable.
2. Theorem 1 applies Theorem 3 to the special case when the fitted functions are Lipschitz in the penalty parameters
3. Lemma 1 applies Theorem 1 to the special case where λ_{min} and λ_{max} are polynomial in the dataset size.
4. Understand the behavior of the oracle error $\min_{\lambda \in \Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V$.
5. An extension of Theorem 3 where the convergence rate δ does not have a random lower bound.

1 Theorem 3

Suppose there exists some $r > 0$ such that the entropy of $\mathcal{G}(T)$ for any training dataset T such that $\|\epsilon\|_T \le 2\sigma$ satisfies

$$\sup_{T:\|\epsilon\|_T \le 2\sigma} \int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du \le \psi_\sigma(R) \; \forall R > r$$

and

$$\frac{\psi_\sigma(u)}{u^2}$$

is nonincreasing wrt to u for all $u > r$.
Then there is some constant $a > 0$ (only dependent on the characteristics of the sub-gassian errors) such that for all $\delta > r$ satisfying

$$\sqrt{n_V}\delta^2 \ge a \left(\psi_\sigma(\delta) \vee \delta \vee \psi_\sigma \left(4 \left\| \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\|_V \right) \vee 4 \left\| \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\|_V \right)$$

we have

$$Pr\left(\left\| \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^* \right\|_V^2 - \left\| \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\|_V^2 \ge \delta^2 \right) \quad \le \quad c \exp\left(-\frac{n_V \delta^4}{c^2 \left\| \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\|_V^2} \right) + c \exp\left(-\frac{n_V \delta^2}{c^2} \right) + c \exp\left(-\frac{n_T \sigma^2}{c^2} \right) \tag{1}$$

for a constant $c > 0$.

Proof

We use the usual peeling argument

$$\begin{aligned}
& Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \geq \delta^2 \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\
= & \sum_{s=0}^{\infty} Pr\left(2^{2s}\delta^2 \leq \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \leq 2^{2s+2}\delta^2 \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\
\leq & \sum_{s=0}^{\infty} Pr\left(2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \leq 2^{2s+2}\delta^2 \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\
= & \sum_{s=0}^{\infty} Pr\left(2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2 + 2\left\langle \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}), \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\rangle_V \leq 2^{2s+2}\delta^2 \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\
\leq & \sum_{s=0}^{\infty} Pr\left(2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2 \leq 2^{2s+2}\delta^2 + 2\left|\left\langle \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}), \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\rangle_V\right| \middle| T, \|\epsilon\|_T \leq 2\sigma \right)
\end{aligned}$$

We can split each probability into the case where $2^{2s+2}\delta^2$ or $2\left|\left\langle \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}), \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\rangle_V\right|$ is bigger:

$$\begin{aligned}
& Pr\left(2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2 \leq 2^{2s+2}\delta^2 + 2\left|\left\langle \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}), \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^* \right\rangle_V\right| \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\
\leq & Pr\left(2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2 \leq 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\
& + Pr\left(2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2 \leq 2^{2s+3}\delta^2 \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\
\leq & Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V} 2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\
& + Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2^{s+3/2}\delta} 2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \middle| T, \|\epsilon\|_T \leq 2\sigma \right)
\end{aligned}$$

Hence

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \geq \delta^2 \middle| T, \|\epsilon\|_T \leq 2\sigma \right)$$

$$\begin{aligned} \leq \quad & Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V} \delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \\ & + \sum_{s=0}^{\infty} Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2^{s+3/2}\delta} 2^{2s}\delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \end{aligned}$$

We have the entropy condition that

$$\sup_{T:\|\epsilon\|_T \leq 2\sigma} \int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du \leq \psi_\sigma(R)$$

So as long as $\delta > 0$ satisfies

$$\sqrt{n_V}\delta^2 \geq a\left(\psi_\sigma(4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V) \vee 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \right)$$

we can bound the first probability using Vandegeer Corollary 8.3 as follows

$$Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V} \delta^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) \right\rangle_V \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \leq c\exp\left(-\frac{n_V\delta^4}{4c^2\left(16\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2\right)} \right)$$

for some $c > 0$.

We can also bound the summation using Vandegeer Corollary 8.3. First we check the condition for Corollary 8.3 is satisfied. In particular, we need to show that for all $s = 0, 1, 2, ...$

$$\sqrt{n_V}2^{2s+2}\delta^2 \quad \geq \quad a\left(\psi_\sigma\left(2^{s+1}\delta\right) \vee \delta\right)$$

where $a > 0$ is a constant that only depends on the sub-gaussian errors.

This is true since we chose δ such that

$$\sqrt{n_V}\delta^2 \quad \geq \quad a\left(\psi_\sigma\left(\delta\right) \vee \delta\right)$$

and we assumed that $\psi_\sigma(u)/u^2$ is nonincreasing for all u.

So Corollary 8.3 states that for all $s = 0, 1, ...$

$$Pr\left(\sup_{\lambda,\lambda':\|\hat{g}(\cdot|\lambda)-\hat{g}(\cdot|\lambda')\|_V \leq 2^{s+3/2}\delta} \left\langle \epsilon, \hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\boldsymbol{\lambda}') \right\rangle_V \geq 2^{2s-1}\delta^2 \middle| T, \|\epsilon\|_T \leq 2\sigma \right) \leq \exp\left(-n_V \frac{2^{4s-2}\delta^4}{4C^2 2^{2s+3}\delta^2} \right)$$

Putting this together, for all $\delta > r$ satisfying

$$\sqrt{n_V}\delta^2 \geq a\left(\psi_\sigma\left(\delta\right) \vee \delta \vee \psi_\sigma\left(4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right) \vee 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right)$$

there exists some constant c such that

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \geq \delta^2 \middle| T, \|\epsilon\|_T \leq 2\sigma\right) \quad \leq \quad c\exp\left(-\frac{n_V\delta^4}{c^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2}\right) + c\exp\left(-\frac{n_V\delta^2}{c^2}\right)$$

Finally, we note that

$$\begin{aligned} Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \geq \delta^2\right) \quad &\leq \quad c\exp\left(-\frac{n_V\delta^4}{c^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2}\right) + c\exp\left(-\frac{n_V\delta^2}{c^2}\right) \\ &\quad +Pr\left(\|\epsilon\|_T^2 \geq 4\sigma^2\right) \end{aligned}$$

The last term can easily be bounded using Bernstein's inequality since ϵ are independent sub-gaussian random variables

$$Pr\left(\|\epsilon\|_T^2 \geq 4\sigma^2\right) \leq c\exp\left(-\frac{n_T\sigma^2}{c^2}\right)$$

for some constant $c > 0$.

2 Theorem 1

Let $\Lambda = [\lambda_{min}, \lambda_{max}]^J$.

Suppose there is a constant $C > 0$ such that for all $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda$

$$\sup_{T:\|\epsilon\|_T \leq 2\sigma} \left\|\hat{g}(\cdot|\boldsymbol{\lambda}^{(1)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(2)})\right\|_V \leq C\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|$$

Then there is a constant $a > 0$ only depends on the characteristics of the sub-gassian errors.

For any $\delta > 0$ such that

$$\delta^2 \geq a\max\left\{\frac{\alpha_n^2}{n_V}, \frac{\alpha_n}{\sqrt{n_V}}\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V, \frac{1}{n^2}\right\}$$

where

$$\alpha_n = \sqrt{J}\left(1+\log\left(32Cn\left(\lambda_{max}-\lambda_{min}\right)\right)\right)^{1/2}$$

where C_J is a universal constant.
we have

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V^2-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2\ge\delta^2\right) \quad\le\quad c\exp\left(-\frac{n_V\delta^4}{c^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2}\right)+c\exp\left(-\frac{n_V\delta^2}{c^2}\right)+c\exp\left(-\frac{n_T\sigma^2}{c^2}\right)$$

for some constant $c>0$.

Proof

1. Determine entropy bound and properties

Under the given Lipschitz condition, a δ-cover for Λ is a $C\delta$-cover for $\mathcal{G}(T)$. We can therefore calculate a covering number for $\mathcal{G}(T)$ wrt $\|\cdot\|_V$ by using the covering number for Λ.

$$N\left(u,\mathcal{G}(T),\|\cdot\|_V\right)\le N\left(\frac{u}{C},\Lambda,\|\cdot\|_2\right)$$

By Lemma Cube Covering Number (See Appendix below), we know that

$$\begin{aligned}N\left(u,\Lambda,\|\cdot\|_2\right) &\le \frac{1}{C_J}\left(\frac{4\left(\lambda_{max}-\lambda_{min}\right)+2\frac{u}{C}}{\frac{u}{C}}\right)^J\\ &= \frac{1}{C_J}\left(\frac{4\left(\lambda_{max}-\lambda_{min}\right)C+2u}{u}\right)^J\\ &\le \frac{1}{C_J}\left(\frac{4C\left(\lambda_{max}-\lambda_{min}\right)+2u}{u}\right)^J\end{aligned}$$

Hence

$$H(u,\mathcal{G}(T),\|\cdot\|_V)\le\log\left[\frac{1}{C_J}\left(\frac{4C\left(\lambda_{max}-\lambda_{min}\right)+2u}{u}\right)^J\right]$$

Then

$$\begin{aligned}
\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du &\leq \int_0^R \left[\log\frac{1}{C_J} + J\log\left(\frac{4C\left(\lambda_{max}-\lambda_{min}\right)+2u}{u}\right)\right]^{1/2} du \\
&\leq \int_0^R \left[\log\frac{1}{C_J} + J\log 4 + J\log\left(\frac{8C\left(\lambda_{max}-\lambda_{min}\right)}{u}\right)\right]^{1/2} du \\
&= R\int_0^1 \left[\log\frac{1}{C_J} + J\log 4 + J\log\left(\frac{8C\left(\lambda_{max}-\lambda_{min}\right)}{Rv}\right)\right]^{1/2} dv \\
&\leq R\left[\int_0^1 \left(\log\frac{1}{C_J} + J\log 4 + J\log\left(\frac{8C\left(\lambda_{max}-\lambda_{min}\right)}{R}\right) + J\log\frac{1}{v}\right) dv\right]^{1/2} \\
&= R\left[\log\frac{1}{C_J} + J(1+\log 4) + J\log\left(8C\left(\lambda_{max}-\lambda_{min}\right)\right) + J\log\frac{1}{R}\right]^{1/2}
\end{aligned}$$

The second bound is crazy loose but I think it is okay. It comes from the fact that

$$\log\left(a+b\right) < \log\left(2a\right) + \log(2b)$$

The third inequality follows from concavity of the square root.

We will consider the following bound for $\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du$ for all $R \geq n^{-1}$

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi_\sigma(R) = R\left(J\left(1+\log\left(32C\left(\lambda_{max}-\lambda_{min}\right)n\right)\right)\right)^{1/2}$$

Notice we've replaced the last term $\log\frac{1}{R}$ with $\log n$, which is valid over the given range. We will see this is useful since solving for δ is hard with the $\log\frac{1}{R}$ term.

Also, for simplicity, we dropped $\log\frac{1}{C_J}$ since $C_J > 1$ for all $J = 1, 2, ...$

2. Apply Theorem 3

Now we apply Theorem 3 to determine δ such that (1) holds. Theorem 3 states that we need $\delta > n^{-1}$ to satisfy

$$\sqrt{n_V}\delta^2 \geq a\left(\psi_\sigma\left(\delta\right) \vee \delta \vee \psi_\sigma\left(4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right) \vee 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right)$$

where a is a constant only dependent on characteristics of the sub-gaussian random variables.

We will solve this by splitting into cases.

Case 1: Suppose that

$$\psi_\sigma\left(\delta\right) \vee \delta \geq \psi_\sigma\left(4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right) \vee 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$$

In this case, we must have

$$\sqrt{n_V}\delta^2 \geq a\delta\left(\left(J\left(1+\log\left(32C\left(\lambda_{max}-\lambda_{min}\right)n\right)\right)\right)^{1/2} \vee 1\right)$$

So

$$\delta \geq a\frac{1}{\sqrt{n_V}}\left(\left(J\left(1+\log\left(32C\left(\lambda_{max}-\lambda_{min}\right)n\right)\right)\right)^{1/2} \vee 1\right)$$

Case 2:
Suppose that

$$\psi_\sigma\left(\delta\right) \vee \delta \leq \psi_\sigma\left(4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right) \vee 4\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$$

In this case, we must have

$$\sqrt{n_V}\delta^2 \geq 4a\left\|\hat{g}\left(\cdot|\tilde{\boldsymbol{\lambda}}\right) - g^*\right\|_V\left(\left(J\left(1+\log\left(32C\left(\lambda_{max}-\lambda_{min}\right)n\right)\right)\right)^{1/2} \vee 1\right)$$

Putting these two inequalities together, we find that $\delta > 0$ must satisfy

$$\delta^2 \geq a\max\left(\frac{\alpha_n^2}{n_V}, \frac{\alpha_n}{\sqrt{n_V}}\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V, \frac{1}{n^2}\right)$$

where

$$\alpha_n = \left(J\left(1+\log\left(32C\left(\lambda_{max}-\lambda_{min}\right)n\right)\right)\right)^{1/2}$$

3 Lemma 1 with λ changing with n

We will express this using asymptotic notation.

Let $\Lambda = [n_T^{-t_{min}}, n_T^{t_{max}}]^J$ where $t_{min}, t_{max} > 0$.

Suppose that if $\|\epsilon\|_T \leq 2\sigma$, there are constants C, κ such that for any $u > 0$, we have for all $\lambda \in \Lambda$

$$\left\|\hat{g}(\cdot|\boldsymbol{\lambda}^{(1)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(2)})\right\|_V \leq Cn_T^\kappa\|\boldsymbol{\lambda}_1 - \boldsymbol{\lambda}_2\|$$

Then

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \quad \leq \quad \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 + O_p\left(\frac{J\left(\left(\kappa+t_{max}\right)\log n_T + \log\left(Cn\right)\right)}{n_V}\right) + O_p\left(\left[\frac{J\left(\left(\kappa+t_{max}\right)\log n_T + \log\left(Cn\right)\right)}{n_V}\right]^{1/2}\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right)$$

(Note: t_{min} doesn't appear in the formula, but this is because it appears in the κ term. κ is usually a linear combination of t_{min} and t_{max}.)

4 Understanding the behavior of the oracle error

All of the oracle inequalities that we have derived use the oracle error

$$\min_{\lambda \in \Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V$$

as an upper bound. Intuitively, one would think that the oracle error is small, but we did not prove this earlier. In Mitchell's paper (CITE in the real paper), he supposes the generalized loss is small

$$\min_{\lambda \in \Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|^2 = \min_{\lambda \in \Lambda} \int \left(\hat{g}(x|\boldsymbol{\lambda}) - g^*(x)\right)^2 dx$$

We will use the generalized loss as our start point to show that $\min_{\lambda \in \Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V$ is small if $\min_{\lambda \in \Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|^2$ is small. To do this, we use Theorem 2.1 in Vandegeer (CITE in the real paper, On the uniform convergence...).

Let

$$\tilde{\boldsymbol{\lambda}}_{gen} = \arg\min_{\lambda \in \Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|^2$$

Consider the function class composed of just the single function:

$$\mathcal{G}(T) = \left\{\hat{g}\left(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}\right) - g^*\right\}$$

Since this class contains a single function, its entropy is zero.

Suppose that

$$K_{gen} = \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|_\infty$$

Then by Theorem 2.1, for any $t > 0$, we can bound the conditional probability (training set T is given)

$$Pr\left(\frac{1}{C_1}\left|\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|^2\right| \le \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\| K_{gen}\sqrt{\frac{t}{n_V}} + \frac{K_{gen}^2 t}{n_V}\right| T\right) \ge 1 - \exp(-t)$$

where C_1 is a constant given in the theorem.

Extension

Perhaps instead we want to start with knowing that the oracle training loss is small. Then we need to go from the oracle training loss to the generalized loss to the validation loss. All these jumps require empirical process techniques. We can in fact use Theorem 2.1 to make all these jumps.

Suppose we know that with high probability, for any training dataset, there is an oracle penalty parameter vector $\tilde{\boldsymbol{\lambda}}_T$ such that

$$Pr\left(\min_{\lambda\in\Lambda}\|\hat{g}(\cdot|\boldsymbol{\lambda})-g^*\|_T^2\leq Wn_T^{-2\omega}\right)\geq 1-p(n_T)$$

where $p(n_T)$ is some small probability tending to zero as $n_T\to\infty$ and constants $W,\omega>0$ only dependent on the model class $\mathcal{G}$ Suppose $\tilde{\boldsymbol{\lambda}}_T=\arg\min_\lambda\|\hat{g}(\cdot|\boldsymbol{\lambda})-g^*\|_T$. Let

$$K_\Lambda=\sup_{\lambda\in\Lambda}\|\hat{g}(\cdot|\boldsymbol{\lambda})-g^*\|_\infty$$

Then if $\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T)-g^*\right\|_T^2\leq Wn_T^{-2\omega}$, according to Theorem 2.1 in Vandegeer , we have for all $\delta_t>0$ satisfying

$$\delta_t\geq\frac{2J_\infty\left(K_\Lambda,\mathcal{G}\right)+K\sqrt{t}}{\sqrt{n_T}}+\frac{4J_\infty^2(K_\Lambda,\mathcal{G})+K_\Lambda^2t}{n_T}$$

the following inequality holds with probability at least $1-\exp(-t)$

$$\begin{aligned}\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T)-g^*\right\|^2 &= \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T)-g^*\right\|^2-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T)-g^*\right\|_T^2+\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T)-g^*\right\|_T^2\\ &\leq \delta_t\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T)-g^*\right\|^2+Wn_T^{-2\omega}\end{aligned}$$

which is equivalent to saying that for all $0<\delta_t<1/2$ satisfying

$$\delta_t\geq\frac{2J_\infty\left(K,\mathcal{G}\right)+K_\Lambda\sqrt{t}}{\sqrt{n_T}}+\frac{4J_\infty^2(K_\Lambda,\mathcal{G})+K_\Lambda^2t}{n}$$

we have

$$Pr\left(\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T)-g^*\right\|^2\leq 2Wn_T^{-2\omega}\middle|\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T)-g^*\right\|_T^2\leq Wn_T^{-2\omega}\right)\geq 1-\exp(-t)$$

Note that for the models of interest, $J_\infty\left(K_\Lambda^2,\mathcal{G}\right)$ usually contains at most $\log n_T$ but not a polynomial term in n_T. These results can be used to extend Theorem 3 even further.

5 Extended Theorem 3

Let

$$K=\sup_{g\in\mathcal{G}}\|g-g^*\|_\infty$$

Suppose that there are constants $W, \omega > 0$ only dependent on the model class $\mathcal{G}$ such that

$$Pr\left(\min_{\lambda\in\Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_T^2 \le W n_T^{-2\omega}\right) \ge 1 - p(n_T)$$

where $p(n_T)$ tends to zero as $n_T \to \infty$.
Choose $t, \tilde{\delta}_t > 0$ such that

$$\tilde{\delta}_t^2 \ge C_1\left(\sqrt{2W} K n_T^{-\omega} n_V^{-1/2}\sqrt{t} + \frac{K^2 t}{n_V}\right) + 2W n_T^{-2\omega}$$

and $t_1 > 0$ such that

$$\frac{1}{2} \ge \frac{2J_\infty(K,\mathcal{G}) + K\sqrt{t_1}}{\sqrt{n_T}} + \frac{4J_\infty^2(K,\mathcal{G}) + K^2 t_1}{n_T}$$

where $C_1 > 0$ is a constant given in Vandegeer Theorem 2.1.
Suppose there is an $r > 0$ such that for all $R > r$, we have

$$\sup_{T:\|\epsilon\|_T \le 2\sigma} \int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V) du \le \psi_\sigma(R)$$

Then for all $\delta > r$ such that

$$\sqrt{n_V}\delta^2 \ge \alpha\left(\psi_\sigma(\delta) \vee \delta \vee \psi_\sigma(\tilde{\delta}_t) \vee \tilde{\delta}_t\right)$$

we have

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left[\min_{\lambda\in\Lambda}\|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2\right] \ge \delta^2\right) \le c\exp\left(-\frac{n_V\delta^4}{c^2\tilde{\delta}_t}\right) + c\exp\left(-\frac{n_V\delta^2}{c^2}\right) + c\exp\left(-\frac{n_T\sigma^2}{c^2}\right) + \exp(-t_1) + p(n_T) + \exp(-t)$$

Proof

We break up the probability of interest into the following components

$$\begin{aligned}
& Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left[\min_{\lambda\in\Lambda}\|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2\right] \ge \delta^2\right) \\
\le\ & Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left[\min_{\lambda\in\Lambda}\|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2\right] \ge \delta^2 \wedge \min_{\lambda\in\Lambda}\|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2 \le \tilde{\delta}^2\right) \\
& + Pr\left(\min_{\lambda\in\Lambda}\|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2 \ge \tilde{\delta}_t^2\right)
\end{aligned}$$

We bound the first probability term using Theorem 3: For our choice of δ, we have that

$$Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 - \left[\min_{\lambda\in\Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2\right] \geq \delta^2 \wedge \min_{\lambda\in\Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2 \leq \tilde{\delta}^2\right) \quad \leq \quad c\exp\left(-\frac{n_V\delta^4}{c^2\tilde{\delta}_t}\right) + c\exp\left(-\frac{n_V\delta^2}{c^2}\right) + c\exp\left(-\frac{n_T\sigma^2}{c^2}\right)$$

To bound the second probability term, we use the results just established

$$\begin{aligned}
& Pr\left(\min_{\lambda\in\Lambda} \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2 \geq \tilde{\delta}_t^2\right) \\
\leq \quad & Pr\left(\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|_V^2 \geq \tilde{\delta}_t^2\right) \\
\leq \quad & Pr\left(\left|\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|^2\right| + \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|^2 \geq \tilde{\delta}_t^2\right) \\
\leq \quad & Pr\left(\left|\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|^2\right| + \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|^2 \geq \tilde{\delta}_t^2 \wedge \frac{1}{C_1}\left|\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|^2\right| \leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\| K\sqrt{\frac{t}{n_V}} + \frac{K^2t}{n_V}\right) \\
& +Pr\left(\frac{1}{C_1}\left|\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|_V^2 - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|^2\right| \geq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\| K\sqrt{\frac{t}{n_V}} + \frac{K^2t}{n_V}\right) \\
\leq \quad & Pr\left(C_1\left(\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\| K\sqrt{\frac{t}{n_V}} + \frac{K^2t}{n_V}\right) + \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|^2 \geq \tilde{\delta}_t^2\right) \\
& +\exp(-t) \\
\leq \quad & Pr\left(C_1\left(\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\| K\sqrt{\frac{t}{n_V}} + \frac{K^2t}{n_V}\right) + \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\|^2 \geq \tilde{\delta}_t^2\right) + \exp(-t) \\
\leq \quad & Pr\left(C_1\left(\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\| K\sqrt{\frac{t}{n_V}} + \frac{K^2t}{n_V}\right) + \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\|^2 \geq \tilde{\delta}_t^2 \wedge \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\|^2 \leq 2Wn_T^{-2\omega} \wedge \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\|_T^2 \leq Wn_T^{-2\omega}\right) \\
& +Pr\left(\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\|^2 \geq 2Wn_T^{-2\omega} \wedge \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\|_T^2 \leq Wn_T^{-2\omega}\right) \\
& +Pr\left(\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_T) - g^*\right\|_T^2 \geq Wn_T^{-2\omega}\right) \\
& +\exp(-t) \\
\leq \quad & 0 + \exp(-t_1) + p(n_T) + \exp(-t)
\end{aligned}$$

The last line follows as long as we choose t and $\tilde{\delta}_t$ such that

$$\tilde{\delta}_t^2 \geq C_1\left(\sqrt{2W}n_T^{-\omega}K\sqrt{\frac{t}{n_V}} + \frac{K^2t}{n_V}\right) + 2Wn_T^{-2\omega}$$

and $t_1 > 0$ such that

$$\frac{1}{2} \geq \frac{2J_\infty(K,\mathcal{G}) + K\sqrt{t_1}}{\sqrt{n_T}} + \frac{4J_\infty^2(K,\mathcal{G}) + K^2 t_1}{n_T}$$

6 Appendix

Lemma Vandegeer (Based on Vandegeer Corollary 8.3)

(This lemma is directly out of Vandegeer's Empirical Process book. We apply this to the case where $\sigma = \infty$)

Let Q_m be the empirical distributon of m observations at covariates x_i.

Suppose ϵ are m independent sub-gaussian errors. Suppose the model class $\mathcal{F}(T)$ has elements $\sup_{f\in\mathcal{F}_n(T)} \|f\|_{Q_m} \leq R$ and satisfies

$$\psi_T(R) \geq \int_0^R H^{1/2}(u, \mathcal{F}(T), \|\cdot\|_{Q_m}) du$$

There is a dependent only on the sub-gaussian constants such that for all $\delta > 0$ such that

$$\sqrt{m}\delta \geq a(\psi_T(R) \vee R)$$

we have

$$Pr\left(\sup_{f\in\mathcal{F}(T)} \left| \frac{1}{m}\sum_{i=1}^m \epsilon_i f(x_i) \right| \geq \delta \middle| T \right) \leq C\exp\left(-\frac{m\delta^2}{4C^2R^2} \right)$$

Lemma Cube Covering Number

Suppose $\Lambda = [\lambda_{min}, \lambda_{max}]^J$. Then the δ-covering number is bounded as follows

$$N(\delta, \Lambda, \|\cdot\|_2) \leq \frac{1}{C_J}\left(\frac{4(\lambda_{max} - \lambda_{min}) + 2\delta}{\delta} \right)^J$$

where

$$C_J = \frac{\text{volume of ball of radius } \rho}{\rho^J} = \frac{\pi^{J/2}}{\Gamma(\frac{J}{2}+1)}$$

.

Proof

(Essentially the same proof as that for Lemma 2.5 in vandegeer)

Let $C = \{c_j\}_{j=1}^N \subset \Lambda$ be the largest set s.t. two distinct points c_{j_1}, c_{j_2} are at least δ apart. Then balls with radius δ centered at C cover Λ. Hence

$$N(\delta, \Lambda, \|\cdot\|_2) \leq N$$

If we instead consider the balls centered at C but with radius $\delta/4$, all of these smaller balls must be disjoint and are completely contained in the box $\Lambda_{bigger} = [\lambda_{min} - \delta/4, \lambda_{max} + \delta/4]^J$. So we know the aggregate volume of these smaller balls is less than the volume of Λ_{bigger}.

Hence

$$NC_J(\delta/4)^J \leq (\lambda_{max} - \lambda_{min} + \delta/2)^J$$