Training/Validation Split Theorem

November 1, 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

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

The document is organized as follows

1. Theorem 3 proves the convergence of the selected model to the best model under general entropy conditions for fixed training data.
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 $\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$
5. We extend Theorem 3 to make a statement on the convergence of the selected model to the best model where the training data are also random variables.

1 Theorem 3

Let $r > 0$. Suppose that if $\|\epsilon\|_T \leq 2\sigma$, then $\mathcal{G}(T)$ satisfies the entropy condition

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi_T(R)$$

for all $R > r$.
Furthermore, suppose that

$$\frac{\psi_T(u)}{u^2}$$

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 $\delta > r$ with

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

we have

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

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\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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|\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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|\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\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\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\epsilon\|_T\leq 2\sigma\right)
\end{aligned}
$$

Hence

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

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

If $\|\epsilon\|_T \leq 2\sigma$, then we have the entropy condition that

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi_T(R)$$

As long as $\delta > 0$ satisfies

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

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

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

for some $c > 0$.
We also bound the summation using Vandegeer Corollary 8.3. Recall that if $\|\epsilon\|_T \leq 2\sigma$, then we have the entropy condition that

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi_T(R)$$

So we will apply Vandegeer Corollary 8.3 to bound the empirical process assuming that $\|\epsilon\|_T \leq 2\sigma$. First we check the condition for Corollary 8.3 is satisfied. In particular, we need to show that for all $s = 0, 1, 2, ...$

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

where $a > 0$ is a constant that only depends on the sub-gaussian errors.
This is true since we chose δ such that

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

and we assumed that $\psi_T(u)/u^2$ is nonincreasing for all u.
So Corollary 8.3 states that for all $s = 0, 1, ...$

$$Pr\left(\sup_{\lambda,\lambda':\|\hat{g}(\cdot|\lambda)-\hat{g}(\cdot|\lambda')\|_V\leq 2^{s+3/2}\delta} \left\langle \epsilon, \hat{g}(\cdot|\boldsymbol{\lambda}) - \hat{g}(\cdot|\boldsymbol{\lambda}')\right\rangle_V \geq 2^{2s-1}\delta^2 \wedge \|\epsilon\|_V \leq 2\sigma \wedge \|\epsilon\|_T \leq 2\sigma\right) \leq \exp\left(-n_V \frac{2^{4s-2}\delta^4}{4C^2 2^{2s+3}\delta^2}\right)$$

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

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

there exists some constant c such that

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

2 Theorem 1

Let $\Lambda = [\lambda_{min}, \lambda_{max}]^J$.
Suppose that if $\|\epsilon\|_T \leq 2\sigma$, there is a constant $C > 0$ such that for all $\boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda$

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

Then for any $\delta > n_V^{-1} \wedge n_T^{-1}$ such that

$$\delta^2 \geq \frac{\alpha^2}{n_V} \vee \alpha\frac{1}{\sqrt{n_V}}\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$$

where

$$\alpha = a\left[\left[\log\frac{1}{C_J} + J(1 + \log 4 + \log 8C) + J\log(\lambda_{max} - \lambda_{min}) + J\log(n_V \vee n_T)\right]^{1/2}\right] \vee 1$$

and C_J is a universal constant, we have

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

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.

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

By Lemma param_covering_cube (See Appendix), we know that

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

Hence

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

Then

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

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

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

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 $R \geq n_V^{-1} \wedge n_T^{-1}$

$$\int_0^R H^{1/2}(u, \mathcal{G}(T), \|\cdot\|_V)du \leq \psi_T(R) = R\left[\log\frac{1}{C_J} + J(1 + \log 4 + \log 8C) + J\log\left(\lambda_{max} - \lambda_{min}\right) + J\log\left(n_V \vee n_T\right)\right]^{1/2}$$

Notice we've replaced the last term $\log\frac{1}{R}$ with $\log\left(n_V \vee n_T\right)$, 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.

2. Apply Theorem 3

Now we apply Theorem 3 to the choice of $\delta > n_V^{-1} \wedge n_T^{-1}$ that satisfies

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

We will solve this by splitting into cases. Let

$$\alpha = a\left[\left[\log\frac{1}{C_J} + J(1 + \log 4 + \log 8C) + J\log\left(\lambda_{max} - \lambda_{min}\right) + J\log\left(n_V \vee n_T\right)\right]^{1/2}\right] \vee 1$$

Case 1: Suppose that

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

In this case, we must have

$$\delta \geq \frac{\alpha}{\sqrt{n_V}}$$

Case 2:

Suppose that

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

In this case, we must have

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

Putting these two inequalities together, we find that $\delta > n_V^{-1} \wedge n_T^{-1}$ must satisfy

$$\delta^2 \geq \frac{\alpha^2}{n_V} \vee \alpha\frac{1}{\sqrt{n_V}}\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$$

3 Lemma 1 with λ changing with n

We will express this using asymptotic notation.

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

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

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

Then

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

4 Understanding the behavior of the oracle error

All of the oracle inequalities that we have derived use the oracle error over the validation observations, $\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$, as an upper bound. Intuitively, one would think that the oracle error $\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V$ is small, but we did not prove this earlier. We are interested in showing that this quantity is small if $\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|$ is small. To do this, we use Theorem 2.1 in Vandegeer (from the paper On the uniform convergence...). Note that $\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|$ is valid to use in the upper bound since Mitchell's paper also uses this value.

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:

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

We note that this function class is dependent on the training data but independent of validation set.

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)

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

where C_1 is a constant given in the theorem.

Furthermore, suppose we know that with high probability, for any training dataset, there is an oracle penalty parameter vector $\tilde{\boldsymbol{\lambda}}$ such that the optimal convergence rate $O_p\left(n_T^{-2\omega}\right)$. That is, suppose there are constants $W, \omega > 0$ only dependent on the model class $\mathcal{G}$ such that

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

where $p(n_T)$ is some small probability tending to zero as $n_T \to \infty$.

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

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

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

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

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

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

we have

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

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.

Now we can extend **Theorem 3** further.

5 Extended Theorem 3

Let

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

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

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

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

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

and $t_1 > 0$ such that

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

where $C_1 > 0$ is a constant given in Vandegeer Theorem 2.1.
Then for all $\delta > n_V^{-1} \wedge n_T^{-1}$ such that

$$\sqrt{n_V}\delta^2 \geq \alpha\left(\psi_T(\delta) \vee \delta \vee \psi_T(\tilde{\delta}_t) \vee \tilde{\delta}_t\right)$$

we have

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

Proof

We break up the probability of interest into the following components

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

$$+Pr\left(\|\epsilon\|_T \geq 2\sigma\right)$$
$$+Pr\left(\|\epsilon\|_V \geq 2\sigma\right)$$

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

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

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\ & Pr\left(\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\right\|_V^2 \geq \tilde{\delta}_t^2\right) \\
\leq\ & 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\ & 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\ & 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\ & 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\ & 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\ & 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

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

and $t_1 > 0$ such that

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

Finally, we bound the last two terms using Bernstein's inequality since ϵ are independent sub-gaussian random variables

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

and

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

for some constant $c > 0$.

6 Appendix

Lemma Vandegeer (Based on Vandegeer Corollary 8.3)

(This lemma is directly out of Vandegeer's Empirical Process book.)

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

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

Lemma param_covering_cube

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

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

Lemma psuedo-basic inequality

Suppose

$$\left\|y - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2 \leq \left\|y - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2$$

Suppose for all samples i in dataset V, we have

$$y_i = g^*(x_i) + \epsilon_i$$

then

$$2\left\langle g^* - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}), \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\rangle_V + \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2 \leq 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V$$

Proof

From the definition, we have that

$$\begin{aligned}\left\|y-\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2 &= \left\|y-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})+\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2 \\ &= \left\|y-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2+\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2+2\left\langle y-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}),\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\rangle_V \\ &\leq \left\|y-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V^2\end{aligned}$$

Then

$$\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2+2\left\langle \epsilon+g^*-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}),\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\rangle_V \leq 0$$

Rearranging, we get

$$\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\|_V^2+2\left\langle g^*-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}),\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})\right\rangle_V \leq 2\left\langle \epsilon,\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V$$

7 Old Theorem 3

Suppose that if $\|\epsilon\|_T \leq 2\sigma$, then $\mathcal{G}(T)$ satisfies the entropy condition

$$\int_0^R H^{1/2}(u,\mathcal{G}(T),\|\cdot\|_V)du \leq \psi_T(R)$$

Furthermore, suppose that

$$\frac{\psi_T\left(a+u\right)}{u^2}$$

is nonincreasing wrt to u for all $u,a>0$ such that $a+u>\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V$.
Then there is some constant C (only dependent on the characteristics of the sub-gassian errors) such that for all $\delta>0$ such that

$$\sqrt{n_V}\delta^2 \geq 2C\left[\psi_T\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2\delta\right)\vee\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2\delta\right)\right]$$

then

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

for a constant $c>0$.

Proof

The basic inequality gives us

$$\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V^2 \leq \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V^2 + 2\left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V$$

Note that since $\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \leq \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V$, then

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

By a peeling argument, we have

$$\begin{aligned}
Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \geq \delta\right) &= \sum_{s=0}^{\infty} Pr\left(2^s\delta \leq \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \leq 2^{s+1}\delta\right) \\
&\leq \sum_{s=0}^{\infty} Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V \geq 2^s\delta \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2^{s+1}\delta\right) \\
&= \sum_{s=0}^{\infty} Pr\left(\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\right\|_V - \left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V\right)^2 \geq 2^{2s}\delta^2 \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2^{s+1}\delta\right) \\
&\leq \sum_{s=0}^{\infty} 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 2^{2s}\delta^2 \wedge \left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2^{s+1}\delta\right) \\
&\leq \sum_{s=0}^{\infty} Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V \leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2^{s+1}\delta} \left\langle \epsilon, \hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - \hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V \geq 2^{2s-1}\delta^2\right)
\end{aligned}$$

To apply the lemma based on vandegeer corollary 8.3 (see below), we must check all the conditions are satisfied.
We choose δ such that

$$\begin{aligned}
\frac{\sqrt{n_V}}{8} &\geq \frac{C}{4\delta^2}\left[\psi_T\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2\delta\right) \vee \left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2\delta\right)\right] \\
&\geq \frac{C}{2^{2s+2}\delta^2}\left[\psi_T\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2^{s+1}\delta\right) \vee \left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}) - g^*\right\|_V + 2^{s+1}\delta\right)\right]
\end{aligned}$$

where the second line follows from the assumption that $\psi_T(a+u)/u^2$ is nonincreasing wrt u. Hence we have satisfied the condition in corollary 8.3. So for all $s = 0, 1, ...$ since

$$\sqrt{n_V}2^{2s-1}\delta^2 \geq C\left[\psi_T\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)\vee\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)\right]$$

we have

$$Pr\left(\sup_{\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\|_V\leq 2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta}\left\langle\epsilon,\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})\right\rangle_V\geq 2^{2s-1}\delta^2\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\epsilon\|_T\leq 2\sigma\right)\leq\exp\left(-n_V\frac{2^{4s-2}\delta^4}{4C^2\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)^2}\right)$$

Hence we have

$$\begin{aligned}
Pr\left(\left\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}})-g^*\right\|_V-\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V\geq\delta\wedge\|\epsilon\|_V\leq 2\sigma\wedge\|\epsilon\|_T\leq 2\sigma\right) &\leq C\sum_{s=0}^{\infty}\exp\left(-n_V\frac{2^{4s-2}\delta^4}{4C^2\left(2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V+2^{s+1}\delta\right)^2}\right)\\
&\leq C\sum_{s=0}^{\infty}\exp\left(-n_V\frac{2^{4s-2}\delta^4}{64C^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2}\right)\vee\exp\left(-n_V\frac{2^{2s}\delta^2}{196C^2}\right)\\
&\leq c\exp\left(-\frac{n_V\delta^4}{c^2\left\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}})-g^*\right\|_V^2}\right)+c\exp\left(-\frac{n_V\delta^2}{c^2}\right)
\end{aligned}$$

for some constant c.
Hence we have found for the given δ choice, we have

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

8 Random Thoughts

What is wrong with the following logic?

$$\begin{aligned}\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2 &= \|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2 - \left(\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|_V^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|_V^2\right) \\ &\quad + \left(\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|_V^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|_V^2\right) \\ &\leq \left|\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|^2 - \|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|_V^2\right| + \left|\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|_V^2\right|\end{aligned}$$

For the first inequality, note that by definition, we have

$$\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|_V^2 \leq \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|_V^2$$

Using Theorem 2.1, we have that for all $\delta_n, t > 0$ such that

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

we have

$$Pr\left(\sup_{\lambda\in\Lambda}\left|\frac{\|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|^2 - \|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|_V^2}{\|\hat{g}(\cdot|\boldsymbol{\lambda}) - g^*\|^2}\right| \leq \delta_n\right) \geq 1 - \exp(-t)$$

Under the condition that $\sup_\lambda \left|\frac{\|\hat{g}(\cdot|\boldsymbol{\lambda})-g^*\|^2-\|\hat{g}(\cdot|\boldsymbol{\lambda})-g^*\|_V^2}{\|\hat{g}(\cdot|\boldsymbol{\lambda})-g^*\|^2}\right| \leq \delta_n$, we have

$$\begin{aligned}\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2 &\leq \delta_n\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|^2 + \delta_n\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2 \\ &= \delta_n\left(\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2\right) + 2\delta_n\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2\end{aligned}$$

So we have that

$$Pr\left(\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2 \leq \frac{2\delta_n}{1-\delta_n}\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2\right) \quad \geq \quad 1 - \exp(-t)$$

If we can choose $\delta_n \leq 1/2$, then

$$Pr\left(\|\hat{g}(\cdot|\hat{\boldsymbol{\lambda}}) - g^*\|^2 - \|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2 \leq 4\delta_n\|\hat{g}(\cdot|\tilde{\boldsymbol{\lambda}}_{gen}) - g^*\|^2\right) \quad \geq \quad 1 - \exp(-t)$$

This is a really really fast rate!