1 Cross-validation Proof

The problem

Let θ^* be the true model parameters and suppose we observe

$$y = g_{\theta^*} + \epsilon$$

where ϵ are the sub-gaussian random errors. Let T be the training set and V be the validation set. Let the total number of observations be n.

Let the loss function for our regression problem be least squares. We denote it as $\|h\|_T^2 = \frac{1}{|T|}\sum_{i=1}^{|T|} h(x_i)$ and similarly for $\|h\|_V^2$.

Consider the joint optimization problem on a training/validation split to find the best regularization parameter λ in Λ:

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \frac{1}{2}\|y - g_{\hat{\theta}(\lambda)}\|_V^2$$

$$\hat{\theta}(\lambda) = \arg\min_{\theta} \frac{1}{2}\|y - g_\theta\|_T^2 + \lambda\left(P(\theta) + \frac{w}{2}\|\theta\|_2^2\right)$$

Here $w > 0$ is some fixed parameter. (One should choose this to be on the order of 1e-15 or smaller. In practice, we can probably just drop it entirely.)

Let the range of Λ be from $[\lambda_{min}, \lambda_{max}]$. Both limits can grow and shrink at any polynomial rate, e.g. $\lambda_{max} = O_P(n^{\tau_{max}})$ and $\lambda_{min} = O_P(n^{-\tau_{min}})$.

Suppose there is an optimal $\tilde{\lambda}$ s.t. $\|g_{\hat{\theta}(\lambda)} - g^*\|$ is some optimal rate $O_p(n^{-r})$ where $r \geq 1/2$.

We show that $\hat{\lambda}$will converge to $\tilde{\lambda}$ at an almost-parametric rate or just the optimal rate $O_p(n^{-r})$.

The almost-parametric rate is

$$\|g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})}\|_V \leq O_P(1)\left(\frac{1 - \log w + \kappa\log n}{|V|}\right)^{1/2}$$

where κ is a constant that depends on known constants, including τ_{max} and $\tau_{\min}$.

Other Assumptions

- Suppose that penalty $P(\theta)$ is smooth and convex wrt θ – its 1st and 2nd derivatives wrt θ are defined. Suppose $\|\nabla_\theta P(\theta)|_{\hat{\theta}(\lambda)}\| \leq O_p(n^{k_1})$.
- Suppose g_θ is a smooth convex functions wrt θ – its 1st and 2nd derivatives wrt θ are defined.
- Suppose g_θ is Lipschitz with constant M s.t.

$$\|g_{\theta_1} - g_{\theta_2}\|_n \leq M\left\|\theta_1 - \theta_2\right\|_2$$

Proof

Step 1: Find the entropy of the model class $\mathcal{G}_\lambda$

First we show that the entropy $H(u, \mathcal{G}_\lambda, \|\cdot\|_V)$ of the class

$$\mathcal{G}_\lambda = \{\theta_\lambda : \lambda \in \Lambda\}$$

is bounded at a near-parametric rate:

$$H\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \leq \log\left(\frac{M}{uw}\right) + \kappa\log n$$

By the mean value theorem, for any $\delta > 0$, there is some $\alpha \in [0,1]$ s.t.

$$\|\hat{\theta}(\ell) - \hat{\theta}(\ell + \delta)\| = \delta \left\| \nabla_\lambda \hat{\theta}(\lambda)|_{\lambda = \ell + \alpha\delta} \right\|$$

Since the problem is smooth, we can apply implicit differentiation to get the derivative

$$\nabla_\lambda \hat{\theta}(\lambda) = -H(\lambda)^{-1} \left(\nabla_\theta P(\theta) + w\theta \right)$$

where the Hessian matrix is

$$H(\lambda) = \nabla_\theta^2 \left(\|y - g_\theta\|_T^2 + \lambda P(\theta) \right) + \lambda w I$$

Under the assumption that g_θ and $P(\theta)$ are both smooth and convex wrt θ which means the minimum eigenvalue of $H(\lambda)$ is at least $\lambda w = O_p(n^{-\tau_{min}})w$.

From the asumptions, we have that

$$\begin{aligned} \left\| \nabla_\lambda \hat{\theta}(\lambda) \right\| &\leq \left\| H(\lambda)^{-1} \left(\nabla_\theta P(\theta) + w\hat{\theta}(\lambda) \right) \right\| \\ &\leq O_p(n^{\tau_{min}}) w^{-1} \left(O_p(n^{k_1}) + w\|\hat{\theta}(\lambda)\| \right) \end{aligned}$$

Note that $\|\hat{\theta}(\lambda)\|$ is easily bounded by the solution to the ridge regression problem:

$$\hat{\theta}_r(\lambda) = \arg\min \frac{1}{2} \|y - g_\theta\|_T^2 + \lambda w \|\theta\|_2^2$$

The solution $\hat{\theta}_r(\lambda)$ has norm at most $w^{-1} O_P(n^{\tau_{\min}})$. It is straightforward to show that

$$\|\hat{\theta}(\lambda)\| \leq \|\hat{\theta}_r(\lambda)\|$$

Therefore

$$\left\| \nabla_\lambda \hat{\theta}(\lambda) \right\| = w^{-1} O_p(n^{\max(2\tau_{min}, \tau_{min} + k_1)})$$

Now we can know bound the distance between $g_{\hat{\theta}(\ell)}$ and $g_{\hat{\theta}(\ell+\delta)}$

$$\begin{aligned} \|g_{\hat{\theta}(\ell)} - g_{\hat{\theta}(\ell+\delta)}\|_V &\leq M \left\| \hat{\theta}(\ell) - \hat{\theta}(\ell + \delta) \right\|_2 \\ &= M\delta w^{-1} O_p(n^{\max(2\tau_{min}, \tau_{min} + k_1)}) \end{aligned}$$

Then the covering number is on the order of n^κ where κ grows linearly in τ_{min}, τ_{max} and k_1.

$$N\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \leq \frac{L}{uw} O_p(n^\kappa) \implies H\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \leq \log\left(\frac{M}{uw}\right) + \kappa \log n$$

Step 2: Find the basic inequality

By definition,

$$\|y - g_{\hat{\theta}(\hat{\lambda})}\|_V^2 \leq \|y - g_{\hat{\theta}(\tilde{\lambda})}\|_V^2$$

Rearranging, we get the basic inequality

$$\|g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})}\|_V^2 \leq 2 \left| \left(\epsilon, g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})} \right) \right|_V + \left| \left(g^* - g_{\hat{\theta}(\tilde{\lambda})}, g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})} \right) \right|_V$$

Clearly if the second term dominates, Cauchy Schwarz (and a symmetrization argument) give us that $\|g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})}\|$ converges that the same rate as $\|g^* - g_{\hat{\theta}(\tilde{\lambda})}\|$.

In the next step we bound the empirical process term.

Step 3: Bound the empirical process with high probability

We apply Corollary 8.3 in Vandegeer to determine the value δ s.t. δ bounds the empirical process term with high probability.

Let

$$\tilde{\mathcal{G}_\lambda} = \left\{ \frac{g_{\hat{\theta}(\lambda)} - g_{\hat{\theta}(\tilde{\lambda})}}{\|g_{\hat{\theta}(\lambda)} - g_{\hat{\theta}(\tilde{\lambda})}\|_V} : \lambda \in \Lambda \right\}$$

Consequently,

$$\begin{aligned} \int_0^1 H^{1/2}\left(u, \tilde{\mathcal{G}_\lambda}, \|\cdot\|_V\right) du &= \int_0^1 \left(\log\left(\frac{1}{uw}\right) + \kappa \log n \right)^{1/2} du \\ &\le \left(\int_0^1 \log\left(\frac{1}{uw}\right) + \kappa \log n du \right)^{1/2} \\ &\le (1 - \log w + \kappa \log n)^{1/2} \end{aligned}$$

By Corollary 8.3, if we choose

$$\delta \ge \left(\frac{1 - \log w + \kappa \log n}{|V|} \right)^{1/2}$$

then

$$Pr\left(\frac{\left|\left(\epsilon, g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})}\right)\right|_V}{\|g_{\hat{\theta}(\lambda)} - g_{\hat{\theta}(\tilde{\lambda})}\|_V} \ge \delta \right) \le \exp\left(-|V| \frac{\delta^2}{C^2} \right)$$

Step 4: Win

Returning to the basic inequality in Step 2, we find that in the case that the empirical process term is bigger, we have

$$\begin{aligned} \|g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})}\|_V &\le 4 \left| \frac{(\epsilon, g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})})_V}{\|g_{\hat{\theta}(\lambda)} - g_{\hat{\theta}(\tilde{\lambda})}\|_V} \right| \\ &\le 4 \left(\frac{1 - \log w + \kappa \log n}{|V|} \right)^{1/2} \end{aligned}$$

2 Lemmas

2.1 Convergence Rate Equivalence between the original regression problem and the perturbed ridge problem

Suppose the original regression problem is

$$\hat{\beta}(\lambda) = \arg\min_{\beta \in \Theta} \frac{1}{2} \|y - g_\beta\|_T^2 + \lambda P(\beta)$$

and the new perturbed ridge problem is

$$\hat{\theta}(\lambda) = \arg\min_{\theta \in \Theta} \frac{1}{2} \|y - g_\theta\|_T^2 + \lambda \left(P(\theta) + \frac{w}{2} \|\theta\|_2^2 \right)$$

Let θ^* be the true model parameters. Suppose there are constants $K_0, K_1 > 0$ s.t.

$$\frac{w}{2} \|\theta^*\|_2^2 \le K_0 P(\theta^*) + K_1$$

Then the rate of convergence of $\left\| g_{\hat{\beta}(\lambda)} - g^* \right\|$ is the same as $\left\| g_{\hat{\theta}(\lambda)} - g^* \right\|$ modulo some constant.

Proof

For ease of notation, we will write $\hat{\theta} = \hat{\theta}(\lambda)$ and $\hat{\beta} = \hat{\beta}(\lambda)$.

By definition,

$$\begin{aligned} \frac{1}{2}\|y - g_{\hat{\theta}}\|^2 + \tilde{\lambda}\left(P(\hat{\theta}) + \frac{w}{2}\|\hat{\theta}\|^2\right) &\leq \frac{1}{2}\|y - g_{\theta^*}\|^2 + \tilde{\lambda}\left(P(\theta^*) + \frac{w}{2}\|\theta^*\|^2\right) \\ &\leq \frac{1}{2}\|y - g_{\theta^*}\|^2 + \tilde{\lambda}(1+K_0)P(\theta^*) + \tilde{\lambda}K_1 \end{aligned}$$

Therefore

$$\frac{1}{2}\|y - g_{\hat{\theta}}\|^2 + \tilde{\lambda}P(\hat{\theta}) \leq \frac{1}{2}\|y - g_{\theta^*}\|^2 + \tilde{\lambda}(1+K_0)P(\theta^*) + \tilde{\lambda}K_1$$

Notice that this inequality is very similar to the inequality from the original regression problem

$$\frac{1}{2}\|y - g_{\hat{\beta}}\|^2 + \tilde{\lambda}P(\hat{\beta}) \leq \frac{1}{2}\|y - g_{\theta^*}\|^2 + \tilde{\lambda}P(\theta^*)$$

Therefore the same arguments to prove the convergence rate of $\left\|g_{\hat{\beta}(\lambda)} - g^*\right\|$ give the same convergence rate for $\left\|g_{\hat{\theta}(\lambda)} - g^*\right\|$. (Example: refer to Thrm 10.2 in Vandegeer)

2.2 Regression problems with smooth-almost-everywhere penalty functions

If the regularization functions contain smooth-almost-everywhere penalty functions, we still have the same convergence rate. This requires the additional assumption that the local optimality space is the same as the differentiable space (refer to condition 1 in the hillclimbing paper).

A small modification to Step 1 of the proof shows that the entropy of the class $\mathcal{G}_\lambda$ is still the same. Hence the rest of the proof remains unchanged.

Proof

Step 1 for Smooth-almost-everywhere functions: Find the entropy of the model class $\mathcal{G}_\lambda$

Let S be the set of knots at which $\nabla_\lambda \theta(\lambda)|_{\lambda=s}$ does not exist. Since the regression problem is smooth almost everywhere, S should have measure zero.

First we apply the mean value theorem to two points ℓ and $\ell + \delta$ that have no knots in between. (Think about what happens when λ is multi-dim?) That is, for any $\delta > 0$, there is some $\alpha \in [0, 1]$ s.t.

$$\|\hat{\theta}(\ell) - \hat{\theta}(\ell + \delta)\| = \delta \left\| \nabla_\lambda \hat{\theta}(\lambda)|_{\lambda = \ell + \alpha\delta} \right\|$$

Apply the same assumptions from the hillclimbing paper regarding the differentiable space and the local optimality space. We can then reformulate the solutions in terms of the differentiable space. Suppose U forms an orthonormal basis for the differentiable space. Hence we can rewrite θ in terms of β s.t. $\theta = U\beta$. The derivative of $\hat{\beta}(\lambda)$ can be calculated since the locally equivalent regression problem is now smooth:

$$\nabla_\lambda \hat{\beta}(\lambda) = H(\lambda)^{-1} \left({}_U\nabla P(U\hat{\beta}) + wU\hat{\beta} \right)$$

where the Hessian matrix is

$$H(\lambda) =_U \nabla^2 \left(\|y - g_{U\hat{\beta}}\|_T^2 + \lambda P(U\hat{\beta}) \right) + \lambda w I$$

Under the assumption that g_θ and $P(\theta)$ are both smooth and convex wrt θ, the minimum eigenvalue of $H(\lambda)$ is at least $\lambda w = O_p(n^{-\tau_{min}})w$. Hence

$$\left\|\nabla_\lambda \hat{\beta}(\lambda)\right\| \quad \leq \quad O_p(n^{\tau_{min}}) w^{-1} \left(O_p(n^{k_1}) + w\|\hat{\beta}(\lambda)\| \right)$$

By the same argument as above, $\|\hat{\beta}(\lambda)\|$ is easily bounded by the solution to the analogous ridge regression problem:

$$\hat{\beta}_r(\lambda) = \arg\min \frac{1}{2}\|y - g_{U\beta}\|_T^2 + \lambda w \|U\beta\|_2^2$$

Therefore

$$\left\|\nabla_\lambda \hat{\beta}(\lambda)\right\| = w^{-1} O_p(n^{\max(2\tau_{min}, \tau_{min}+k_1)})$$

Now we can bound the distance between $g_{\hat{\theta}(\ell)}$ and $g_{\hat{\theta}(\ell+\delta)}$

$$\begin{aligned} \|g_{\hat{\theta}(\ell)} - g_{\hat{\theta}(\ell+\delta)}\|_V &\leq M \left\| U\hat{\beta}(\ell) - U\hat{\beta}(\ell+\delta) \right\|_2 \\ &= M\delta w^{-1} O_p(n^{\max(2\tau_{min}, \tau_{min}+k_1)}) \end{aligned}$$

Next consider if there are knots between points ℓ and $\ell+\delta$. We can recover the same upper bound by chaining. That is, suppose there are some countable number of knots λ_i for $i = 1, ..., s$ (where s can equal ∞) between ℓ and $\ell+\delta$, with distances

$$\delta_0 = \lambda_1 - \ell, \delta_i = \lambda_{i+1} - \lambda_i, \delta_s = \ell + \delta - \lambda_s$$

Then

$$\begin{aligned} \|g_{\hat{\theta}(\ell)} - g_{\hat{\theta}(\ell+\delta)}\|_V &\leq \|g_{\hat{\theta}(\ell)} - g_{\hat{\theta}(\lambda_1)}\|_V + \|g_{\hat{\theta}(\lambda_s)} - g_{\hat{\theta}(\ell+\delta)}\|_V + \sum_{i=1}^{s} \|g_{\hat{\theta}(\lambda_i)} - g_{\hat{\theta}(\lambda_{i+1})}\|_V \\ &= M w^{-1} O_p(n^{\max(2\tau_{min}, \tau_{min}+k_1)}) \sum_{i=0}^{s} \delta_i \\ &= M\delta w^{-1} O_p(n^{\max(2\tau_{min}, \tau_{min}+k_1)}) \end{aligned}$$

Then the covering number is on the order of n^κ where κ grows linearly in τ_{min}, τ_{max} and k_1 and is independent of the number of knots.

$$N\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \leq \frac{L}{uw} O_p(n^\kappa) \implies H\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \leq \log\left(\frac{M}{uw}\right) + \kappa \log n$$

2.3 K-fold Cross-Validation

The same convergence rate holds for K-fold cross-validation. For $k = 1, ..., K$, let D_k represent the kth fold and D_{-k} denote all the folds minus the kth fold. For a given λ, train over D_{-k} and then validate over D_k.

Let $\|h\|_k^2 = \frac{1}{|D_k|}\sum_{i\in D_k} h(x_i)^2$ and similarly for $\|h\|_{-k}^2$ for the set D_{-k}. Let $(h,g)_k = \frac{1}{|D_k|}\sum_{i\in D_k} h(x_i)g(x_i)$ and $(h,g)_{-k}$ for the set D_{-k}.

The joint optimization problem for k-fold cross-validation is

$$\hat{\lambda} = \arg\min_{\lambda\in\Lambda} \frac{1}{2}\sum_{k=1}^{K} \|y - g_{\hat{\theta}_k(\lambda)}\|_k^2$$

$$\hat{\theta}_k(\lambda) = \arg\min_{\theta \in \Theta} \frac{1}{2} \|y - g_\theta\|_{-k}^2 + \lambda \left(P(\theta) + \frac{w}{2} \|\theta\|_2^2 \right)$$

We show that the convergence rate of $\sqrt{\sum_{k=1}^K \|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k^2}$ is either nearly-parametric or is the same as the optimal convergence rate of $\sqrt{\sum_{k=1}^K \|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\theta^*}\|_k^2}$. The nearly-parametric rate is

$$O_P(1) \left(1 - 2 \log w + \kappa \log n\right)^{1/2} \left(\sum_{k=1}^K |D_k|^{-1/2} \right)$$

Proof

Step 1: Basic inequality crunching

By definition,

$$\sum_{k=1}^K \|y - g_{\hat{\theta}_k(\hat{\lambda})}\|_k^2 \le \sum_{k=1}^K \|y - g_{\hat{\theta}_k(\tilde{\lambda})}\|_k^2$$

Rearranging as usual, we get

$$\begin{aligned}
\sum_{k=1}^K \|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k^2 &\le 2 \sum_{k=1}^K \left(y - g_{\hat{\theta}_k(\tilde{\lambda})}, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right)_k \\
&\le 2 \left| \sum_{k=1}^K \left(\epsilon, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right)_k \right| + 2 \left| \sum_{k=1}^K \left(g^* - g_{\hat{\theta}_k(\tilde{\lambda})}, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right)_k \right|
\end{aligned}$$

Case 1: Suppose $\left| \sum_{k=1}^K \left(g^* - g_{\hat{\theta}_k(\tilde{\lambda})}, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right)_k \right|$ is bigger

$$\begin{aligned}
\sum_{k=1}^K \|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k^2 &\le 4 \left| \sum_{k=1}^K \left(g^* - g_{\hat{\theta}_k(\tilde{\lambda})}, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right)_k \right| \\
&\le 4 \sqrt{ \left(\sum_{k=1}^K \left\| g^* - g_{\hat{\theta}_k(\tilde{\lambda})} \right\|_k^2 \right) \left(\sum_{k=1}^K \left\| g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right\|_k^2 \right) }
\end{aligned}$$

where the second inequality follows from Cauchy-Schwarz. Then

$$\sum_{k=1}^K \left\| g^* - g_{\hat{\theta}_k(\tilde{\lambda})} \right\|_k^2 \le O_p(1) \sum_{k=1}^K \left\| g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right\|_k^2$$

Hence in this case, we recover whatever the convergence rate is for the optimal $\tilde{\lambda}$.

Case 2: Suppose $\left| \sum_{k=1}^K \left(\epsilon, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right)_k \right|$ is bigger

$$\sum_{k=1}^K \|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k^2 \le 4 \left| \sum_{k=1}^K \left(\epsilon, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})} \right)_k \right|$$

Obviously

$$
\begin{aligned}
\sqrt{\sum_{k=1}^{K} \|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k^2} &\leq 4 \left| \frac{\sum_{k=1}^{K} \left(\epsilon, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})}\right)_k}{\sqrt{\sum_{k=1}^{K} \|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k^2}} \right| \\
&\leq 4 \sum_{k=1}^{K} \frac{\left| \left(\epsilon, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})}\right)_k \right|}{\|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k}
\end{aligned}
$$

Consider model classes $\mathcal{G}_{\lambda,k} = \left\{ g_{\hat{\theta_k}(\lambda)} : \lambda \in \Lambda \right\}$ for $k = 1, ..., K$. Apply Proof Step 1 to each $\mathcal{G}_{\lambda,k}$. The entropy bound for each class is

$$
H\left(u, \mathcal{G}_{\lambda}, \|\cdot\|_k\right) \leq \log\left(\frac{M}{uw}\right) + \kappa \log n
$$

Then for every k,

$$
\frac{\left| \left(\epsilon, g_{\hat{\theta}_k(\hat{\lambda})} - g_{\hat{\theta}_k(\tilde{\lambda})}\right)_k \right|}{\|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k} = O_P(1) \left(\frac{1 - 2\log w + \kappa \log n}{|D_k|}\right)^{1/2}
$$

Therefore

$$
\sqrt{\sum_{k=1}^{K} \|g_{\hat{\theta}_k(\tilde{\lambda})} - g_{\hat{\theta}_k(\hat{\lambda})}\|_k^2} \quad \leq \quad O_P(1) \left(1 - 2\log w + \kappa \log n\right)^{1/2} \left(\sum_{k=1}^{K} |D_k|^{-1/2}\right)
$$

2.4 Multiple Regularization Parameters

Suppose we are fitting model parameters θ but we'd like to apply a separate penalty to J partitions of the model parameters: $P_j(\theta_j)$ for $j = 1, ..., J$. Let $\boldsymbol{\lambda} = (\lambda_1, ..., \lambda_J)$ be their corresponding penalty parameters. We suppose that Λ is the box $[\lambda_{min}, \lambda_{max}]^J$ where $\lambda_{min} = O_p(n^{-\tau_{min}})$ and $\lambda_{max} = O_p(n^{\tau_{max}})$.

For simplicity, suppose the problem is smooth and that we tune $\boldsymbol{\lambda}$ over a training/validation split. (One can probably use the same arguments as above to extend this to the case when the problem is smooth almost everywhere and K-fold cross validation.)

Consider the joint optimization problem

$$
\hat{\boldsymbol{\lambda}} = \arg\min_{\boldsymbol{\lambda} \in \Lambda} \frac{1}{2} \|y - g_{\hat{\theta}(\boldsymbol{\lambda})}\|_V^2
$$

$$
\hat{\theta}(\boldsymbol{\lambda}) = \arg\min_{\theta} \frac{1}{2} \|y - g_{\theta}\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j(\theta_j) + \frac{w}{2}\|\theta_j\|_2^2\right)
$$

The convergence rate of $\|g_{\hat{\theta}(\hat{\boldsymbol{\lambda}})} - g_{\hat{\theta}(\tilde{\boldsymbol{\lambda}})}\|_V$ is either nearly-parametric or the same convergence rate as $\|g_{\theta^*} - g_{\hat{\theta}(\tilde{\boldsymbol{\lambda}})}\|_V$. The nearly-parametric rate is

$$
\|g_{\hat{\theta}(\hat{\boldsymbol{\lambda}})} - g_{\hat{\theta}(\tilde{\boldsymbol{\lambda}})}\|_V \quad = \quad O_P(1) \left(\frac{J(1 - \log w + \kappa \log n)}{|V|}\right)^{1/2}
$$

Note that we do indeed have to pay a price for tuning over J dimensions.

Proof

Step 0: Find $\nabla_{\lambda_j}\hat{\theta}(\boldsymbol{\lambda})$ via implicit differentiation

We can again use implicit differentiation to get $\nabla_{\lambda_j}\hat{\theta}(\boldsymbol{\lambda})$. The equations get bulky, but the logic is straightforward.

Since $\hat{\theta}(\boldsymbol{\lambda})$ is a local minima,

$$\nabla_\theta \left(\|y - g_{\hat{\theta}}\|_T^2 + \sum_{j=1}^J \lambda_j \left(P(\hat{\theta}_j) + \frac{w}{2}\|\hat{\theta}_j\|_2^2 \right) \right) = 0$$

which simplifies to

$$\nabla_\theta \|y - g_{\hat{\theta}}\|_T^2 + \sum_{j=1}^J \lambda_j \left(\nabla_\theta P(\hat{\theta}_j) + w\hat{\theta}_j \right) = 0$$

Implicit differentiation wrt λ_ℓ (for $\ell = 1, ..., J$) gives us

$$\nabla_{\lambda_\ell}\hat{\theta}(\boldsymbol{\lambda}) = \left(\nabla_\theta^2 \|y - g_{\hat{\theta}(\boldsymbol{\lambda})}\|_T^2 + \sum_{j=1}^J \lambda_j \nabla_\theta^2 P(\theta_j) + diag\left(\lambda_j w I_j\right) \right)^{-1} \left(\nabla_\theta P(\theta_\ell) + w \begin{bmatrix} 0 \\ \theta_\ell \\ 0 \end{bmatrix} \right)$$

where I_j are appropriately-sized identity matrices.

Step 1: Find the entropy of the model class $\mathcal{G}_{\boldsymbol{\lambda}}$

First we show that the entropy $H(u, \mathcal{G}_{\boldsymbol{\lambda}}, \|\cdot\|_V)$ of the class

$$\mathcal{G}_{\boldsymbol{\lambda}} = \{\theta_{\boldsymbol{\lambda}} : \boldsymbol{\lambda} \in \Lambda\}$$

is bounded at a near-parametric rate:

$$H\left(u, \mathcal{G}_\lambda, \|\cdot\|_V\right) \le ??$$

By the mean value theorem, for any vector $\boldsymbol{\delta}$ there is some $\alpha \in [0,1]$ s.t.

$$\|\hat{\theta}(\boldsymbol{\ell}) - \hat{\theta}(\boldsymbol{\ell} + \boldsymbol{\delta})\| = \|\boldsymbol{\delta}\| \left\| \nabla_{\boldsymbol{\lambda}}\hat{\theta}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda} = \boldsymbol{\ell} + \alpha\boldsymbol{\delta}} \right\|$$

Since the problem is smooth, we can apply implicit differentiation to get the derivative

$$\nabla_{\boldsymbol{\lambda}}\hat{\theta}(\boldsymbol{\lambda}) = -H(\boldsymbol{\lambda})^{-1}\left(\nabla_\theta P(\theta) + w\theta\right)$$

where the Hessian matrix was derived above

$$H(\boldsymbol{\lambda}) = \nabla_\theta^2 \|y - g_{\hat{\theta}(\boldsymbol{\lambda})}\|_T^2 + \sum_{j=1}^J \lambda_j \nabla_\theta^2 P(\theta_j) + diag\left(\lambda_j w I_j\right)$$

The minimum eigenvalue of $H(\boldsymbol{\lambda})$ is at least $wO_p(n^{-\tau_{min}})$.

From the asumptions, we have that for every $\ell = 1, ..., J$

$$\begin{aligned} \left\| \nabla_{\lambda_\ell}\hat{\theta}(\boldsymbol{\lambda}) \right\| &\le \left\| H(\boldsymbol{\lambda})^{-1}\left(\nabla_\theta P(\theta) + w\hat{\theta}(\boldsymbol{\lambda}) \right) \right\| \\ &\le O_p(n^{\tau_{min}}) w^{-1} \left(O_p(n^{k_1}) + w\|\hat{\theta}(\boldsymbol{\lambda})\| \right) \end{aligned}$$

Again we can bound $\|\hat{\theta}(\boldsymbol{\lambda})\|$ by the solution to the ridge regression problem to get that

$$\|\hat{\theta}(\boldsymbol{\lambda})\| \le w^{-1} O_P(n^{\tau_{\min}})$$

Therefore

$$\left\|\nabla_{\lambda_j}\hat{\theta}(\boldsymbol{\lambda})\right\| = w^{-1}O_p(n^{\max(2\tau_{min}, \tau_{min}+k_1)})$$

Now we can bound the distance between $g_{\hat{\theta}(\boldsymbol{\ell})}$ and $g_{\hat{\theta}(\boldsymbol{\ell}+\boldsymbol{\delta})}$

$$\begin{aligned}
\|g_{\hat{\theta}(\boldsymbol{\ell})} - g_{\hat{\theta}(\boldsymbol{\ell}+\boldsymbol{\delta})}\|_V &\le M\left\|\hat{\theta}(\boldsymbol{\ell}) - \hat{\theta}(\boldsymbol{\ell}+\boldsymbol{\delta})\right\|_2 \\
&= M\left\|\sum_{j=1}^J \delta_j \nabla_{\lambda_j}\hat{\theta}(\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\boldsymbol{\ell}+\alpha\boldsymbol{\delta}}\right\| \\
&\le M\sum_{j=1}^J |\delta_j| \left\|\nabla_{\lambda_j}\hat{\theta}(\boldsymbol{\lambda})\Big|_{\boldsymbol{\lambda}=\boldsymbol{\ell}+\alpha\boldsymbol{\delta}}\right\| \\
&\le M\|\boldsymbol{\delta}\| w^{-1}O_p(n^{\max(2\tau_{min}, \tau_{min}+k_1)})
\end{aligned}$$

Hence we can form a u-covering set for $\mathcal{G}_{\boldsymbol{\lambda}} = \left\{g_{\hat{\theta}(\boldsymbol{\lambda})} : \boldsymbol{\lambda} \in \Lambda\right\}$ by covering the box $\Lambda = [\lambda_{min}, \lambda_{max}]^J$ with balls of radius $d = uwO_p(n^{-\max(2\tau_{min}, \tau_{min}+k_1)})/M$.

So the total number of balls needed is on the order of

$$\left(\frac{M}{uw}O_p(n^\kappa)\right)^J$$

where κ grows linearly in τ_{min}, τ_{max} and k_1.

$$N\left(u, \mathcal{G}_{\boldsymbol{\lambda}}, \|\cdot\|_V\right) \le \left(\frac{M}{uw}O_p(n^\kappa)\right)^J \implies H\left(u, \mathcal{G}_{\boldsymbol{\lambda}}, \|\cdot\|_V\right) \le J\left(\log\left(\frac{M}{uw}\right) + \kappa\log n\right)$$

Step 2: Exactly the same

Step 3: Bound the empirical process with high probability

We apply Corollary 8.3 in Vandegeer to determine the value δ s.t. δ bounds the empirical process term with high probability.

From Step 1, we have that

$$\int_0^1 H^{1/2}\left(u, \tilde{\mathcal{G}_\lambda}, \|\cdot\|_V\right)du \quad \le \quad \left(J\left(1 - \log w + \kappa\log n\right)\right)^{1/2}$$

By Corollary 8.3, if we choose

$$\delta \ge \left(\frac{J(1-\log w + \kappa\log n)}{|V|}\right)^{1/2}$$

then

$$Pr\left(\frac{\left|\left(\epsilon, g_{\hat{\theta}(\hat{\lambda})} - g_{\hat{\theta}(\tilde{\lambda})}\right)\right|_V}{\|g_{\hat{\theta}(\lambda)} - g_{\hat{\theta}(\tilde{\lambda})}\|_V} \ge \delta\right) \le \exp\left(-|V|\frac{\delta^2}{C^2}\right)$$

Step 4: Same with an additional term J

3 Examples

Ridge Regression

$$\hat{\theta}(\lambda) = \arg\min_{\theta} \frac{1}{2}\|y - X\theta\|_T^2 + \frac{\lambda}{2}\|\theta\|^2$$

The Hessian is

$$X^T X + \lambda I$$

so its minimum eigenvalue is λ.

The derivative of the penalty wrt θ is

$$\nabla_\theta P(\theta) = \theta$$

Since $\theta = (X^T X + \lambda I)^{-1} X^T y$, the derivative has bounded norm $\|\nabla_\theta P(\theta)\| = O_P(n^{\tau_{min}})$.

All other assumptions are obviously satisfied. Note that in this problem, we can drop the tiny additional ridge penalty entirely.

Smoothing Splines with a Sobolev Penalty

The optimization problem can be formulated as

$$\hat{\theta}(\lambda) = \arg\min_{\theta} \frac{1}{2}\|y - \theta\|_T^2 + \frac{\lambda}{2}\theta^T K\theta$$

where $K = N^{-T}\Omega N$, N is the normalized B-splines evaluated at the input points, and Ω has entries $\Omega_{ij} = \int b_i"(x) b_j"(x) dx$.

Assume the input points i/n for $i = 1 : n$ and we fit y with cubic B-splines. Then $K = DC^{-1}Dn^3$ where D is the (universal, non-data-dependent) second-order discrete diference operator and C is a tridiagonal matrix with diagonal elements equal to $2/3$ and off-diagonal elements equal to $1/6$.

The Hessian is

$$I + \lambda K$$

so its minimum eigenvalue is 1.

The derivative of the penalty wrt θ is

$$\nabla_\theta P(\theta) = K\theta$$

The maximum eigenvalue of K is on the order of $O_p(n^{-3})$. Clearly$\|\theta\|$ can be bounded by $\|y\|$ modulo a constant. So $\|\nabla_\theta P(\theta)\| = O_P(n^3)$. Also, $\nabla_\theta^2 P(\theta) = K$ so its norm is shrinking at a rate of $O_P(n^{-3})$.

All other assumptions are obviously satisfied. Note that in this problem, we can drop the tiny additional ridge penalty entirely.

Lasso (with a tiny ridge)

$$\hat{\theta}(\lambda) = \arg\min_{\theta} \frac{1}{2}\|y - X\theta\|_T^2 + \lambda\left(\|\theta\|_1 + \frac{w}{2}\|\theta\|_2^2\right)$$

With modifications to the proof above, we can probably show that the proof carries through for penalties that are smooth almost everywhere. The main tricky part is dealing with the fact that the Hessian with respect to the differentiable space changes in size for different values of λ.

Anyhow, by a hand-wavy argument, we have that the Hessian matrix with respect to the differentiable space S_λ is

$$X_{S_\lambda}^T X_{S_\lambda} + \lambda w I_{S_\lambda}$$

so its minimum eigenvalue is $w\lambda$.

The lasso penalty clearly satisfies our assumptions since its derivative $\nabla_\theta\|\theta\|_1 = sgn(\theta)$ has bounded norm $\|\nabla_\theta\|\theta\|_1\| \le p$.

All other assumptions are satisfied.