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

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 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

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

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

Other Assumptions

- Suppose that penalty P is smooth and convex – its m-th derivative $\nabla_\theta^{(m)} P(\theta)$ is defined. Suppose $\|\nabla_\theta P(\theta)|_{\hat{\theta}(\lambda)}\| \leq O_p(n^{k_1})$. For $m \geq 2$, suppose the derivative is bounded $\nabla_\theta^{(m)} P(\theta) \leq K_m$.
- Suppose that the m-th derivative $\nabla_\theta^{(m)} g_\theta$ exists. Suppose $\|\nabla_\theta g_{\theta=\hat{\theta}(\lambda)}\| \leq O_p(n^{g_1})$. Suppose the same things as above for the mth derivatives of g_θ wrt θ.
- Suppose g_θ is Lipschitz with constant L s.t.

$$\|g_{\theta_1} - g_{\theta_2}\|_n \leq L\,\|\theta_1 - \theta_2\|_2$$

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:

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

Using a Taylor expansion, we know that for some $\delta > 0$, then

$$\|\hat{\theta}(\ell) - \hat{\theta}(\ell + \delta)\| = \delta \left\| \nabla_\lambda \hat{\theta}(\lambda)|_{\lambda=\ell} \right\| + O_P(\delta^2)$$

Since the problem is smooth, we can find the second derivative via implicit differentiation:

$$\nabla_\lambda \hat{\theta}(\lambda) = \left[\nabla^2_\theta L_T(y, g_\theta)\right]^{-1} (\nabla_\theta P(\theta) + w\theta)$$

Note that the Hessian has the form

$$\nabla^2_\theta L_T(y, g_\theta) = \nabla^2_\theta \left(\|y - g_\theta\|^2 + \lambda P(\theta)\right) + \lambda w I$$

which means its minimum eigenvalue is at least $\lambda w = O_p(n^{-\tau_{min}})w$.

From the asumptions, we have that

$$\begin{aligned} \left\| \nabla_\lambda \hat{\theta}(\lambda) \right\| &\leq \left\| \left[\nabla^2_\theta L_T(y, g_\theta)\right]^{-1} \left(\nabla_\theta P(\theta) + w\hat{\theta}(\lambda)\right) \right\| \\ &\leq O_p(n^{\tau_{min}}) w^{-1} \left(O_p(n^{k_1}) + \|\hat{\theta}(\lambda)\|\right) \end{aligned}$$

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

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

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

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

Therefore

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

Using the Taylor's expansion from above, then we can know bound the distance between $g_{\hat{\theta}(\ell)}$ and $g_{\hat{\theta}(\ell+\delta)}$

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

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

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

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,

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

By Corollary 8.3, if we choose

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

then

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

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

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

2 Lemmas

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.

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

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)

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, $\nabla^2_\theta P(\theta) = K$ 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)

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

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

$$X^T_{S_\lambda} X_{S_\lambda} + \lambda w I_{S_\lambda}$$

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 $\|\nabla_\theta\|\theta\|_1\| \leq p$.

All other assumptions are satisfied.