The effect of adding a small ridge penalty

November 11, 2016

We consider the case of p-dimensional parametric models. Let the original training criterion be denoted

$$L_T(\boldsymbol{\theta}|\boldsymbol{\lambda}) = \frac{1}{2}\|y - f(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta})$$

Let the minimizer to the perturbed training criterion be denoted for any $w \ge 0$,

$$\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w) = \arg\min_{\theta \in \mathbb{R}^p} L_T(\boldsymbol{\theta}|\boldsymbol{\lambda}) + \sum_{j=1}^{J} \lambda_j \frac{w}{2}\|\boldsymbol{\theta}\|^2$$

We show that adding a small ridge penalty scaled by some constant w does not change the fitted model by very much.

This document is organized as follows

1. We quantify the effect of the ridge penalty for w that are small enough. The proof uses the implicit function theorem and the mean value theorem. We assume that the original training criterion is locally strongly convex around its minimizer.

2. We extend the result to parametric models where the training criterion has nonsmooth penalties.

1 Result

Let

$$D(w, \boldsymbol{\theta}) = \nabla_\theta \left[L_T(\boldsymbol{\theta}|\boldsymbol{\lambda}) + \sum_{j=1}^{J} \lambda_j \frac{w}{2}\|\boldsymbol{\theta}\|^2 \right]$$

and suppose $D(w, \boldsymbol{\theta})$ is continuously differentiable in a neighborhood Θ_0 containing $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)$.

Suppose that there is an $m > 0$ such that

$$\nabla_\theta^2 L_T(\boldsymbol{\theta})\big|_{\theta=\hat{\theta}_\lambda(0)} \succeq mI$$

There exists a $W > 0$ such that for all $w \in [0, W)$

$$\left\|\hat{\boldsymbol{\theta}}_\lambda(0) - \hat{\boldsymbol{\theta}}_\lambda(w)\right\| \quad \le \quad \frac{w}{m}\left(\sum_{j=1}^{J} \lambda_j\right) \|\hat{\boldsymbol{\theta}}_\lambda(0)\|$$

Proof

By the implicit function, since $D(w, \boldsymbol{\theta})$ is continuously differentiable in a neighborhood Θ_0 containing $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)$ and $\nabla_\theta D(w, \boldsymbol{\theta}) = \nabla_\theta^2 L_T(\boldsymbol{\theta}|\boldsymbol{\lambda})$ is nonsingular at $\hat{\boldsymbol{\theta}}_\lambda(0)$, $\hat{\boldsymbol{\theta}}_\lambda(w)$ is continuously differentiable over $[0, W)$ for some $W > 0$. Furthermore, the implicit function theorem states that for all $w \in [0, W)$

$$\begin{aligned}
\nabla_w \hat{\boldsymbol{\theta}}_\lambda(w) &= -\left(\nabla_\theta^2 L_T(\boldsymbol{\theta})|_{\theta=\hat{\theta}_\lambda(0)}\right)^{-1} \nabla_w D\left(w, \hat{\boldsymbol{\theta}}_\lambda(w)\right) \\
&= -\left(\nabla_\theta^2 L_T(\boldsymbol{\theta})|_{\theta=\hat{\theta}_\lambda(0)}\right)^{-1} \left(\sum_{j=1}^J \lambda_j\right) \hat{\boldsymbol{\theta}}_\lambda(w)
\end{aligned}$$

Since $\nabla_\theta^2 L_T(\boldsymbol{\theta})|_{\theta=\hat{\theta}_\lambda(0)} \succeq mI$, then for all $w \in [0, W)$

$$\left\|\nabla_w \hat{\boldsymbol{\theta}}_\lambda(w)\right\| \le m^{-1} \left(\sum_{j=1}^J \lambda_j\right) \|\hat{\boldsymbol{\theta}}_\lambda(w)\|$$

We bound $\|\hat{\boldsymbol{\theta}}_\lambda(w)\|$ using the definitions of $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)$ and $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)$:

$$L_T(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)) + \sum_{j=1}^J \lambda_j \frac{w}{2} \|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w)\|^2 \le L_T(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)) + \sum_{j=1}^J \lambda_j \frac{w}{2} \|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)\|^2$$

and

$$L_T(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(0)) \le L_T(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}(w))$$

Adding these two inequalities, we get that for all $w \in [0, W)$

$$\|\hat{\boldsymbol{\theta}}_\lambda(w)\|^2 \le \|\hat{\boldsymbol{\theta}}_\lambda(0)\|^2$$

By the Mean Value Inequality, for all $w \in [0, W)$, there is a $w' \in (0, w)$ such that

$$\begin{aligned}
\left\|\hat{\boldsymbol{\theta}}_\lambda(0) - \hat{\boldsymbol{\theta}}_\lambda(w)\right\| &\le w \left\|\nabla_w \hat{\boldsymbol{\theta}}_\lambda(w)\Big|_{w=w'}\right\| \\
&\le \frac{w}{m} \left(\sum_{j=1}^J \lambda_j\right) \|\hat{\boldsymbol{\theta}}_\lambda(w')\| \\
&\le \frac{w}{m} \left(\sum_{j=1}^J \lambda_j\right) \|\hat{\boldsymbol{\theta}}_\lambda(0)\|
\end{aligned}$$

2 Nonsmooth Case

Let the differentiable space at $\hat{\boldsymbol{\theta}}_0$ be defined as

$$\Omega_\lambda = \left\{ \boldsymbol{\eta} | \lim_{\epsilon \to 0} \frac{L_T\left(\hat{\boldsymbol{\theta}}_\lambda(0) + \epsilon \boldsymbol{\eta} | \boldsymbol{\lambda}\right) - L_T\left(\hat{\boldsymbol{\theta}}_\lambda(0) | \boldsymbol{\lambda}\right)}{\epsilon} \text{ exists} \right\}$$

Let U_λ be an orthonormal basis of Ω.

Suppose that for all $w < W'$, we have that

$$\min_{\theta \in \mathbb{R}^p} L_T(\boldsymbol{\theta}|\boldsymbol{\lambda}) + \sum_{j=1}^J \lambda_j \frac{w}{2} \|\boldsymbol{\theta}\|^2 = \min_{\beta \in \mathbb{R}^q} L_T(U_\lambda \boldsymbol{\beta}|\boldsymbol{\lambda}) + \sum_{j=1}^J \lambda_j \frac{w}{2} \|U_\lambda \boldsymbol{\beta}\|^2$$

Let

$$\hat{\boldsymbol{\beta}}_\lambda(w) = \arg\min_{\beta \in \mathbb{R}^q} L_T(U_\lambda \boldsymbol{\beta} | \boldsymbol{\lambda}) + \sum_{j=1}^J \lambda_j \frac{w}{2} \|U_\lambda \boldsymbol{\beta}\|^2$$

Suppose ${}_U\nabla^2_\beta L_T(U\boldsymbol{\beta}|\boldsymbol{\lambda})$ is exists and is continuous in a neighborhood of $\hat{\boldsymbol{\beta}}_\lambda(w)$. Furthermore suppose there is a $m > 0$ such that

$${}_U\nabla^2_\beta L_T(U\boldsymbol{\beta}|\boldsymbol{\lambda})\Big|_{\beta=\hat{\boldsymbol{\beta}}_\lambda(w)} \succeq mI$$

Then there is a $W > 0$ such that for all $w \in [0, W)$, we have

$$\left\|\hat{\boldsymbol{\theta}}_\lambda(0) - \hat{\boldsymbol{\theta}}_\lambda(w)\right\|_2 \quad \le \quad \frac{w}{m}\left(\sum_{j=1}^J \lambda_j\right) \|\hat{\boldsymbol{\theta}}_\lambda(0)\|_2$$

Proof

By the result in Section 1, we know that

$$\left\|\hat{\boldsymbol{\beta}}_\lambda(0) - \hat{\boldsymbol{\beta}}_\lambda(w)\right\|_2 \quad \le \quad \frac{w}{m}\left(\sum_{j=1}^J \lambda_j\right) \|\hat{\boldsymbol{\beta}}_\lambda(0)\|_2$$

Since $\hat{\boldsymbol{\theta}}_w = U_\lambda \hat{\boldsymbol{\beta}}_w$ and U is an orthonormal matrix, the result follows.