Lemma: parametric proof, smooth penalties

Let $\bar{\lambda} = \frac{1}{J}\sum_{j=1}^{J}\lambda_j$

The function class is

$$\mathcal{G}(T) = \left\{ \hat{\theta}_\lambda = \arg\min \|y - g(\cdot|\theta)\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\theta) + J\bar{\lambda}\frac{w}{2}\|\theta\|_2^2 \right\}$$

Suppose

$$\sup_{\theta\in\Theta} \|\theta\| \le G$$

Suppose one can show that for some constant K, we have

$$\frac{\partial}{\partial m} P(\theta + m\beta) \le K\|\beta\|_2$$

Suppose we have

$$\|g(\cdot|\theta_1) - g(\cdot|\theta_2)\|_\infty \le Lp^r \|\theta_1 - \theta_2\|_2$$

For any $d > 0$, for any $\lambda^{(1)}$ and $\lambda^{(2)}$ chosen such that

$$\|\lambda^{(2)} - \lambda^{(1)}\|_2 \le d\frac{wJ}{2n^{t_{min}}(K + wG)}$$

we have

$$\|\theta_{\lambda^{(1)}} - \theta_{\lambda^{(2)}}\|_2 \le d$$

so it follows that

$$\|g(\cdot|\hat{\theta}_{\lambda^{(1)}}) - g(\cdot|\hat{\theta}_{\lambda^{(2)}})\|_\infty \le Lp^r d$$

Proof

Consider any $\lambda^{(1)}, \lambda^{(2)}$ that satisfy the above conditions. Let $\beta = \theta_{\lambda^{(1)}} - \theta_{\lambda^{(2)}}$. For contradiction, suppose $\|\beta\|_2 \ge d$.

Define

$$\hat{m}_\beta(\lambda) = \arg\min_m \|y - g(\cdot|\theta_{\lambda^{(1)}} + m\beta)\|_T^2 + \sum_{j=1}^{J}\lambda_j P_j(\theta_{\lambda^{(1)}} + m\beta) + J\bar{\lambda}\frac{w}{2}\|\theta_{\lambda^{(1)}} + m\beta\|_2^2$$

We have

$$\nabla_m \|y - g(\cdot|\theta_{\lambda^{(1)}} + m\beta)\|_T^2 + \sum_{j=1}^{J}\lambda_j \nabla_m P_j(\theta_{\lambda^{(1)}} + m\beta) + \frac{w}{2}J\bar{\lambda}\nabla_m\|\theta_{\lambda^{(1)}} + m\beta\|_2^2 = 0$$

and implicit differentiation wrt λ_ℓ (assuming everything is smooth)

$$\frac{\partial}{\partial\lambda_\ell}\hat{m}_\beta(\lambda) = -\left[\nabla_m^2\|y - g(\cdot|\theta_{\lambda^{(1)}} + m\beta)\|_T^2 + \sum_{j=1}^{J}\lambda_j\nabla_m^2 P_j(\theta_{\lambda_1} + m\beta) + wJ\bar{\lambda}\|\beta\|_2^2\right]^{-1} [\nabla_m P_\ell(\theta_{\lambda_1} + m\beta) + w\langle\beta, \theta_{\lambda_1} + m\beta$$

That is,

$$\begin{aligned} \left|\frac{\partial}{\partial\lambda_\ell}\hat{m}_\beta(\lambda)\right| &\le \left(wJ\bar{\lambda}\|\beta\|_2^2\right)^{-1}(K\|\beta\|_2 + wG\|\beta\|_2) \\ &= \frac{n^{t_{min}}(K + wG)}{wJd} \end{aligned}$$

Therefore by MVT, there is some $\alpha \in (0,1)$ such that

$$\begin{aligned}
\hat{m}_\beta(\lambda^{(2)}) &= \left\langle \lambda^{(2)} - \lambda^{(1)}, \nabla_\lambda \hat{m}_\beta(\lambda) \right\rangle \Big|_{\lambda=\alpha\lambda^{(1)}+(1-\alpha)\lambda^{(2)}} \\
&\le \|\lambda^{(2)} - \lambda^{(1)}\|_2 \|\nabla_\lambda \hat{m}_\beta(\lambda)\|_{\lambda=\alpha\lambda^{(1)}+(1-\alpha)\lambda^{(2)}} \\
&\le \|\lambda^{(2)} - \lambda^{(1)}\|_2 \frac{n^{t_{min}}(K+wG)}{wJd} \\
&\le 1/2
\end{aligned}$$

But this is a contradiction since we knew that $\hat{m}_\beta(\lambda^{(2)}) = 1$.

Lemma: Parametric Regression witih Nonsmooth Penalties

Suppose the differentiable space and local optimality space assumptions

Suppose the same conditions as Lemma Parametric with smooth penalties.

For any $d > 0$, for any $\lambda^{(1)}$ and $\lambda^{(2)}$ chosen such that

$$\|\lambda^{(2)} - \lambda^{(1)}\|_2 \le d \frac{wJ}{2n^{t_{min}}(K+wG)}$$

we have

$$\|\theta_{\lambda^{(1)}} - \theta_{\lambda^{(2)}}\|_2 \le d$$

so it follows that

$$\|g(\cdot|\hat{\theta}_{\lambda^{(1)}}) - g(\cdot|\hat{\theta}_{\lambda^{(2)}})\|_\infty \le Lp^r d$$

Proof

Under the given assumptions, for almost every pair $\lambda^{(1)}, \lambda^{(2)}$, there is a line

$$\mathcal{L} = \left\{ \alpha\lambda^{(1)} + (1-\alpha)\lambda^{(2)} : \alpha \in [0,1] \right\}$$

such that there is a finite set of points $\{\ell_i\}_{i=1}^N \subset \mathcal{L}$ such that union of their differentiable space $\Omega^{L_T(\cdot,\ell_i)}(\hat{g}(\cdot|\hat{\theta}_{\ell_i}))$ satisfies

$$\mathcal{L} \subset \cup_{i=0}^{N+1} \Omega^{L_T(\cdot,\ell_i)}(\hat{g}(\cdot|\hat{\theta}_{\ell_i}))$$

where $\ell_0 = \lambda^{(1)}$ and $\ell_{N+1} = \lambda^{(2)}$ and each of the differentiable spaces above satisfy conditions 1 and 2.

Let $\{\ell_{(i)}\}_{i=0}^N \subset \mathcal{L}$ be the points such that $\ell_{(i)}$ is in the differentiable space $\Omega^{L_T(\cdot,\ell_i)}(\hat{g}(\cdot|\hat{\theta}_{\ell_i}))$ and $\Omega^{L_T(\cdot,\ell_{i+1})}(\hat{g}(\cdot|\hat{\theta}_{\ell_{i+1}}))$. That is, we choose

$$\ell_{(i)} \in \Omega^{L_T(\cdot,\ell_i)}(\hat{g}(\cdot|\hat{\theta}_{\ell_i})) \cap \Omega^{L_T(\cdot,\ell_{i+1})}(\hat{g}(\cdot|\hat{\theta}_{\ell_{i+1}}))$$

Then consider applying the smooth lemma to the following pairs of points:

$$\left(\ell_0, \ell_{(0)}\right), \left(\ell_{(0)}, \ell_1\right), ..., \left(\ell_N, \ell_{(N)}\right), \left(\ell_{(N)}, \ell_{N+1}\right)$$

By the lemma for parametric regression with smooth penalties, we get that

$$\|g(\cdot|\hat{\theta}_{\ell_i}) - g(\cdot|\hat{\theta}_{\ell_{(i)}})\|_\infty \le Lp^r \frac{n^{t_{min}}(K+wG)}{wJ\|\beta\|_2} \|\ell_i - \ell_{(i)}\|_2$$

and similarly

$$\|g(\cdot|\hat{\theta}_{\ell_{i+1}}) - g(\cdot|\hat{\theta}_{\ell_{(i)}})\|_\infty \le Lp^r \frac{n^{t_{min}}(K+wG)}{wJ\|\beta\|_2} \|\ell_{i+1} - \ell_{(i)}\|_2$$

Hence

$$\begin{aligned}
\|g(\cdot|\hat{\theta}_{\lambda^{(1)}}) - g(\cdot|\hat{\theta}_{\lambda^{(2)}})\|_\infty &\leq \sum_{i=0}^{N} \|g(\cdot|\hat{\theta}_{\ell_i}) - g(\cdot|\hat{\theta}_{\ell_{(i)}})\|_\infty + \|g(\cdot|\hat{\theta}_{\ell_{i+1}}) - g(\cdot|\hat{\theta}_{\ell_{(i)}})\|_\infty \\
&\leq Lp^r \frac{n^{t_{min}}(K + wG)}{wJ\|\beta\|_2} \left(\sum_{i=0}^{N} \|\ell_i - \ell_{(i)}\|_2 + \|\ell_{i+1} - \ell_{(i)}\|_2 \right) \\
&= Lp^r \frac{n^{t_{min}}(K + wG)}{wJ\|\beta\|_2} \|\lambda^{(1)} - \lambda^{(2)}\|_2
\end{aligned}$$

Example parametric penalties

Ridge, assuming $\sup_{\theta\in\Theta} \|\theta\| \leq G$:

$$\begin{aligned}
\frac{\partial}{\partial m}\|\theta + m\beta\|_2^2 &= \langle \theta + m\beta, \beta \rangle \\
&\leq G\|\beta\|_2
\end{aligned}$$

Lasso:

$$\begin{aligned}
\frac{\partial}{\partial m}\|\theta + m\beta\|_1 &= \langle sgn(\theta + m\beta), \beta \rangle \\
&\leq \|sgn(\theta + m\beta)\|_2 \|\beta\|_2 \\
&\leq p\|\beta\|_2
\end{aligned}$$

Generalized Lasso: let G be the maximum eigenvalue of D.

$$\begin{aligned}
\frac{\partial}{\partial m}\|D(\theta + m\beta)\|_1 &= \langle sgn(D(\theta + m\beta)), D\beta \rangle \\
&\leq \|sgn(D(\theta + m\beta))\|_2 \|D\beta\|_2 \\
&\leq pG\|\beta\|_2
\end{aligned}$$

Group Lasso:

$$\begin{aligned}
\frac{\partial}{\partial m}\|\theta + m\beta\|_2 &= \langle \frac{\theta + m\beta}{\|\theta + m\beta\|_2}, \beta \rangle \\
&\leq \|\beta\|_2
\end{aligned}$$