Lemma: parametric proof, smooth penalties

Suppose we observe training samples from the model

$$y = g(x|\boldsymbol{\theta}^*) + \epsilon$$

Suppose we are fitting parametric functions $g(\cdot|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$.
The function class of interest are the minimizers of the penalized least squares criterion:

$$\mathcal{G}(T) = \left\{ \hat{\boldsymbol{\theta}}_\lambda = \arg\min_{\theta \in \mathbb{R}^p} \frac{1}{2}\|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^J \lambda_j \left(P_j(\boldsymbol{\theta}) + \frac{w}{2}\|\boldsymbol{\theta}\|_2^2 \right) : \boldsymbol{\lambda} \in \Lambda \right\}$$

where $\Lambda = [\lambda_{min}, \lambda_{max}]^J$.
Suppose there is some constant $K > 0$ such that for all $j = 1, ..., J$ and all $\boldsymbol{\beta}, \boldsymbol{\theta}$,

$$\left| \frac{\partial}{\partial m} P_j(\boldsymbol{\theta} + m\boldsymbol{\beta}) \right| \le K\|\boldsymbol{\beta}\|_2$$

Let

$$C = \frac{1}{2}\|\boldsymbol{\epsilon}\|_T^2 + \lambda_{max} \sum_{j=1}^J \left(P_j(\boldsymbol{\theta}^*) + \frac{w}{2}\|\boldsymbol{\theta}^*\|_2^2 \right)$$

Then for any $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda$ we have

$$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(1)}}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(2)}}}\|_2 \le \|\boldsymbol{\lambda^{(2)}} - \boldsymbol{\lambda^{(1)}}\|_2 \left(wJ\lambda_{min} \right)^{-1} \left(K + w\sqrt{\frac{2}{J\lambda_{min} w} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}} \right) C} \right)$$

Moreover, if there are constants $L > 0$ and $r \in \mathbb{R}$, such that for all $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$

$$\|g(\cdot|\boldsymbol{\theta}_1) - g(\cdot|\boldsymbol{\theta}_2)\|_\infty \le Lp^r \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$$

Then

$$\|g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(1)}}}) - g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(2)}}})\|_\infty \le Lp^r \|\boldsymbol{\lambda^{(2)}} - \boldsymbol{\lambda^{(1)}}\|_2 \left(wJ\lambda_{min} \right)^{-1} \left(K + w\sqrt{\frac{2}{J\lambda_{min} w} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}} \right) C} \right)$$

Proof

Consider any $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda$. Let $\boldsymbol{\beta} = \boldsymbol{\theta}_{\boldsymbol{\lambda^{(1)}}} - \boldsymbol{\theta}_{\boldsymbol{\lambda^{(2)}}}$.
Define

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

By definition, we know that $\hat{m}_\beta(\boldsymbol{\lambda^{(2)}}) = 1$ and $\hat{m}_\beta(\boldsymbol{\lambda^{(1)}}) = 0$.
By the KKT conditions, we have

$$\frac{\partial}{\partial m} \left(\frac{1}{2}\|y - g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(1)}}} + m\boldsymbol{\beta})\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(1)}}} + m\boldsymbol{\beta}) \right) + \sum_{j=1}^J \lambda_j w \langle \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(1)}}} + m\boldsymbol{\beta} \rangle \Bigg|_{m = \hat{m}_\beta(\boldsymbol{\lambda})}$$

Now we implicitly differentiate with respect to λ_ℓ for $\ell = 1, 2, ..., J$ (assuming everything is smooth)

$$\frac{\partial}{\partial \lambda_\ell}\left\{\left[\frac{\partial}{\partial m}\left(\frac{1}{2}\|y-g(\cdot|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta)\|_T^2+\sum_{j=1}^J \lambda_j P_j(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta)\right)+\sum_{j=1}^J \lambda_j w\langle\boldsymbol\beta,\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta\rangle\right]\Bigg|_{m=\hat m_\beta(\boldsymbol\lambda)}\right\}=0$$

By the product rule and chain rule, we have

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

Rearranging, we get

$$\frac{\partial}{\partial\lambda_\ell}\hat m_{\boldsymbol\beta}(\boldsymbol\lambda)=-\left[\frac{\partial^2}{\partial m^2}\left(\frac{1}{2}\|y-g(\cdot|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta)\|_T^2+\sum_{j=1}^J \lambda_j P_j(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta)\right)+\sum_{j=1}^J \lambda_j w\|\boldsymbol\beta\|_2^2\right]^{-1}\left[\frac{\partial}{\partial m}P_\ell(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta)+w\right.$$

The first multiplicand is bounded by

$$\left|\frac{\partial^2}{\partial m^2}\left(\frac{1}{2}\|y-g(\cdot|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta)\|_T^2+\sum_{j=1}^J \lambda_j P_j(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta)\right)+\sum_{j=1}^J \lambda_j w\|\boldsymbol\beta\|_2^2\right|^{-1}\leq\left(wJ\lambda_{min}\|\boldsymbol\beta\|_2^2\right)^{-1}$$

since the squared loss is convex and the penalties are convex.
The first summand in the second multiplicand is bounded by assumption

$$\left|\frac{\partial}{\partial m}P_\ell(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+m\boldsymbol\beta)\right|\leq K\|\boldsymbol\beta\|_2$$

The second summand in the second multiplicand is bounded by

$$\left|w\langle\boldsymbol\beta,\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+\hat m_\beta(\boldsymbol\lambda)\boldsymbol\beta\rangle\right| \quad\leq\quad w\|\boldsymbol\beta\|_2\|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+\hat m_\beta(\boldsymbol\lambda)\boldsymbol\beta\|_2$$

We need to bound $\|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+\hat m_\beta(\boldsymbol\lambda)\boldsymbol\beta\|_2$. By definition of $\hat m_\beta(\boldsymbol\lambda)$ and $\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}$, we have

$$\begin{aligned}
\sum_{j=1}^J \lambda_j\frac{w}{2}\|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}+\hat m_\beta(\boldsymbol\lambda)\boldsymbol\beta\|_2^2 &\leq \frac{1}{2}\|y-g(\cdot|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}})\|_T^2+\sum_{j=1}^J\lambda_j\left(P_j(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}})+\frac{w}{2}\|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}\|_2^2\right)\\
&= \frac{1}{2}\|y-g(\cdot|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}})\|_T^2+\sum_{j=1}^J\lambda_j^{(1)}\left(P_j(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}})+\frac{w}{2}\|\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}}\|_2^2\right)+\sum_{j=1}^J\left(\lambda_j-\lambda_j^{(1)}\right)\left(P_j(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}})+\right.\\
&\leq \frac{1}{2}\|y-g(\cdot|\boldsymbol\theta^*)\|_T^2+\sum_{j=1}^J\lambda_j^{(1)}\left(P_j(\boldsymbol\theta^*)+\frac{w}{2}\|\boldsymbol\theta^*\|_2^2\right)+\sum_{j=1}^J\left(\lambda_j-\lambda_j^{(1)}\right)\left(P_j(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}})+\frac{w}{2}\|\hat{\boldsymbol\theta}_{\lambda^{(1)}}\right.\\
&\leq \frac{1}{2}\|y-g(\cdot|\boldsymbol\theta^*)\|_T^2+\sum_{j=1}^J\lambda_j^{(1)}\left(P_j(\boldsymbol\theta^*)+\frac{w}{2}\|\boldsymbol\theta^*\|_2^2\right)+\sum_{j=1}^J\left(\lambda_j-\lambda_j^{(1)}\right)\left[\max_{k=1:J}P_k(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}})+\frac{w}{2}\|\right.\\
&\leq \frac{1}{2}\|y-g(\cdot|\boldsymbol\theta^*)\|_T^2+\lambda_{max}\sum_{j=1}^J\left(P_j(\boldsymbol\theta^*)+\frac{w}{2}\|\boldsymbol\theta^*\|_2^2\right)+J\lambda_{max}\left[\max_{k=1:J}P_k(\hat{\boldsymbol\theta}_{\boldsymbol\lambda^{(1)}})+\frac{w}{2}\|\hat{\boldsymbol\theta}_{\lambda^{(1)}}\|_2^2\right.
\end{aligned}$$

Let

$$C=\frac{1}{2}\|\boldsymbol\epsilon\|_T^2+\lambda_{max}\sum_{j=1}^J\left(P_j(\boldsymbol\theta^*)+\frac{w}{2}\|\boldsymbol\theta^*\|_2^2\right)$$

To bound $\max_{k=1:J} P_k(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)}}\|_2^2$, we note that by the definition of $\hat{\boldsymbol{\theta}}_{\lambda^{(1)}}$, we have

$$\begin{aligned}
\lambda_{min}\left(\max_{k=1:J} P_k(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)}}\|_2^2\right) &\leq \sum_{j=1}^{J} \lambda_j^{(1)} \left(P_j(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)}}\|_2^2\right) \\
&\leq \frac{1}{2}\|y - g(\cdot|\boldsymbol{\theta}^*)\|_T^2 + \sum_{j=1}^{J} \lambda_j^{(1)} \left(P_j(\boldsymbol{\theta}^*) + \frac{w}{2}\|\boldsymbol{\theta}^*\|_2^2\right) \\
&\leq C
\end{aligned}$$

Plugging in the inequality above, we get

$$\sum_{j=1}^{J} \lambda_j \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}} + \hat{m}_\beta(\boldsymbol{\lambda})\boldsymbol{\beta}\|_2^2 \leq \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C$$

After rearranging, we have

$$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}} + \hat{m}_\beta(\boldsymbol{\lambda})\boldsymbol{\beta}\|_2 \leq \sqrt{\frac{2}{J\lambda_{min}w}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C}$$

Therefore

$$w\langle \boldsymbol{\beta}, \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}} + \hat{m}_\beta(\boldsymbol{\lambda})\boldsymbol{\beta}\rangle \leq w\|\boldsymbol{\beta}\|_2 \sqrt{\frac{2}{J\lambda_{min}w}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C}$$

That is,

$$\begin{aligned}
\left|\frac{\partial}{\partial \lambda_\ell}\hat{m}_\beta(\boldsymbol{\lambda})\right| &\leq \left(wJ\lambda_{min}\|\boldsymbol{\beta}\|_2^2\right)^{-1}\left(K\|\boldsymbol{\beta}\|_2 + w\|\boldsymbol{\beta}\|_2\sqrt{\frac{2}{J\lambda_{min}w}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C}\right) \\
&= \left(wJ\lambda_{min}\|\boldsymbol{\beta}\|_2\right)^{-1}\left(K + w\sqrt{\frac{2}{J\lambda_{min}w}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C}\right)
\end{aligned}$$

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

$$\begin{aligned}
\left|\hat{m}_\beta(\boldsymbol{\lambda}^{(2)}) - \hat{m}_\beta(\boldsymbol{\lambda}^{(1)})\right| &= \left|\left\langle \boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}, \nabla_\lambda \hat{m}_\beta(\boldsymbol{\lambda})\right\rangle\Big|_{\boldsymbol{\lambda}=\alpha\boldsymbol{\lambda}^{(1)}+(1-\alpha)\boldsymbol{\lambda}^{(2)}}\right| \\
&\leq \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2 \|\nabla_\lambda \hat{m}_\beta(\boldsymbol{\lambda})\|_{\boldsymbol{\lambda}=\alpha\boldsymbol{\lambda}^{(1)}+(1-\alpha)\boldsymbol{\lambda}^{(2)}} \\
&\leq \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2 \left(wJ\lambda_{min}\|\boldsymbol{\beta}\|_2\right)^{-1}\left(K + w\sqrt{\frac{2}{J\lambda_{min}w}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C}\right)
\end{aligned}$$

Rearranging, we get

$$\|\boldsymbol{\beta}\|_2 = \|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(2)}}\|_2 \leq \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2 \left(wJ\lambda_{min}\right)^{-1}\left(K + w\sqrt{\frac{2}{J\lambda_{min}w}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C}\right)$$

Lemma: Parametric Regression witih Nonsmooth Penalties

Suppose that training criterion satisfies Conditions 1, 2, and 3 from the Hillclimbing paper. Summarizing the conditions, we are supposing that for almost every λ,

Cond 1: The differentiable space at λ is a local optimality space.

Cond 2: The training criterion is twice-differentiable along directions spanned by the differentiable space.

Cond 3: There is an orthonormal basis of the differentiable space directions such that the Hessian of the training criterion is invertible.

Again suppose that there is some constant $K > 0$, such that for all $j = 1, ..., J$ and all $\boldsymbol{\beta}, \boldsymbol{\theta}$,

$$\left| \frac{\partial}{\partial m} P_j(\boldsymbol{\theta} + m\boldsymbol{\beta}) \right| \leq K \|\boldsymbol{\beta}\|_2$$

Then for any $\boldsymbol{\lambda^{(1)}, \lambda^{(2)}} \in \Lambda$ we have

$$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(1)}}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(2)}}}\|_2 \leq \|\boldsymbol{\lambda^{(2)}} - \boldsymbol{\lambda^{(1)}}\|_2 \left(wJ\lambda_{min}\right)^{-1} \left(K + w\sqrt{\frac{2}{J\lambda_{min} w} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C} \right)$$

Moreover, if there are constants $L > 0$ and $r \in \mathbb{R}$, such that for all $\boldsymbol{\theta_1, \theta_2}$

$$\|g(\cdot|\boldsymbol{\theta}_1) - g(\cdot|\boldsymbol{\theta}_2)\|_\infty \leq Lp^r \|\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2\|_2$$

Then

$$\|g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(1)}}}) - g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(2)}}})\|_\infty \leq Lp^r \|\boldsymbol{\lambda^{(2)}} - \boldsymbol{\lambda^{(1)}}\|_2 \left(wJ\lambda_{min}\right)^{-1} \left(K + w\sqrt{\frac{2}{J\lambda_{min} w} \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C} \right)$$

Proof

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

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

such that there is a finite set of points $\{\boldsymbol{\ell_i}\}_{i=1}^N \subset \mathcal{L}$ such that

(1) The union of their differentiable space $\Omega^{L_T(\cdot,\ell_i)}(g(\cdot|\hat{\theta}_{\ell_i}))$ satisfies

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

where $\boldsymbol{\ell_0} = \boldsymbol{\lambda^{(1)}}$ and $\boldsymbol{\ell_{N+1}} = \boldsymbol{\lambda^{(2)}}$.

(2) The differentiable space $\Omega^{L_T(\cdot,\ell_i)}(g(\cdot|\hat{\theta}_{\ell_i}))$ is also a local optimality space for the training criterion.

(3) The training criterion is twice-differentiable along the directions spanned by $\Omega^{L_T(\cdot,\ell_i)}(g(\cdot|\hat{\theta}_{\ell_i}))$.

(4) The Hessian of the training criterion along some orthogonal basis of $\Omega^{L_T(\cdot,\ell_i)}(g(\cdot|\hat{\theta}_{\ell_i}))$ is invertible.

To prove (1), consider any set of points along $\mathcal{L}$. Suppose that the union of their differentiable spaces do not cover $\mathcal{L}$. Then consider the shortest line segment $\mathcal{L}'$ that remains uncovered by the differentiable spaces. Consider the points at the two ends of $\mathcal{L}'$, which we denote as ℓ_s amd ℓ_e. The differentiable space of ℓ_s and ℓ_e must exist

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

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

and similarly

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

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 $\sup_{\theta\in\mathcal{G}(T)}\|\theta\|_2 \leq G$:

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

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:

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