Proofs for Smoothness of Parametric Regression Models

October 9, 2016

Intro

In this document, we consider parametric regression models $g(\cdot|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$. Throughout, we will suppose that the projection of the true model into the parametric model space is $g(x|\boldsymbol{\theta}^*)$.

We are interested in establishing inequalities of the form

$$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(2)}}\|_2 \le C\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2$$

If the functions are Lipschitz in their parameterization, we will also be able to bound the distance between the actual functions. That is, 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

$$\|g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}) - g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(2)}})\|_\infty \le Lp^rC\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2$$

Document Outline

First, we consider smooth training criterions and prove smoothness for two parametric regression examples:

1. Multiple penalties for a single model

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

2. Additive model

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

Then we extend these results to the situation where the penalty functions are non-smooth.

1 Multiple smooth penalties for a single model

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 $\Lambda = [\lambda_{min}, \lambda_{max}]^J$ and $w > 0$ is a fixed constant.
Suppose that the penalties and the function $g(x|\boldsymbol{\theta})$ are smooth and convex wrt $\boldsymbol{\theta}$:

- Suppose that $\nabla_\theta^2 P_j(\boldsymbol{\theta})$ are PSD matrices for all $j = 1, ..., J$.
- Suppose that $\nabla_\theta^2 g(x|\boldsymbol{\theta})$ are PSD matrices for all x.

Primary Assumption (rephrase?) : Suppose there is some $K > 0$ such that for all $j = 1, ..., J$ and any $\boldsymbol{\theta}, \boldsymbol{\beta}$, we have

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

(This is essentially bounding the spectrum of the penalty function)
Result
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

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

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

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

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

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

Now we implicitly differentiate with respect to λ_ℓ for $\ell=1,2,...,J$

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

By the product rule and chain rule, we have

$$\left\{\left[\frac{\partial^2}{\partial m^2}\left(\frac{1}{2}\|y-g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})\|_T^2+\sum_{j=1}^J \lambda_j P_j(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})\right)+\sum_{j=1}^J \lambda_j w\|\boldsymbol{\beta}\|_2^2\right]\frac{\partial}{\partial\lambda_\ell}\hat{m}_\beta(\boldsymbol{\lambda})+\frac{\partial}{\partial m}P_\ell(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})+w\langle\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta}\rangle\right\}\Bigg|_{m=\hat{m}_\beta(\boldsymbol{\lambda})}=0$$

Rearranging, for every $\ell=1,...,J$, we get

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

In vector notation, we have

$$\nabla_\lambda\hat{m}_{\boldsymbol{\beta}}(\boldsymbol{\lambda})=-\left[\frac{\partial^2}{\partial m^2}\left(\frac{1}{2}\|y-g(\cdot|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})\|_T^2+\sum_{j=1}^J \lambda_j P_j(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})\right)+\sum_{j=1}^J \lambda_j w\|\boldsymbol{\beta}\|_2^2\right]^{-1}\left[\nabla_m P(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})+w\langle\boldsymbol{\beta},\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta}\rangle\mathbf{1}\right]\Bigg|_{m=\hat{m}_\beta(\boldsymbol{\lambda})}$$

where $\nabla_m P(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})$ is the J-dimensional vector

$$\nabla_m P(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})=\begin{bmatrix}\frac{\partial}{\partial m}P_1(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})\\ \dots\\ \frac{\partial}{\partial m}P_J(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}+m\boldsymbol{\beta})\end{bmatrix}$$

Bounding the first multiplicand:

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 mean squared error and the penalty functions are convex.

Bounding the second multiplicand:

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

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

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

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

To bound the first part of the right hand side, use the definition of $\hat{\boldsymbol{\theta}}_{\lambda^{(1)}}$:

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

To bound the second part of the right hand side, note that

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

Combining the above three inequalities, we get

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

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

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

Therefore

$$\max_{k=1:J} P_k(\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}}) + \frac{w}{2} \|\hat{\boldsymbol{\theta}}_{\lambda^{(1)}}\|_2^2 \leq \frac{C}{\lambda_{min}} \tag{3}$$

Plugging (3) into (2) above, we get

$$\sum_{j=1}^{J} \lambda_j \frac{w}{2} \|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}} + \hat{m}_\beta(\boldsymbol{\lambda})\boldsymbol{\beta}\|_2^2 \quad \leq \quad \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C \tag{4}$$

We can combine (4) with the fact that

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

to get

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

Plug the inequality above into (1) to get

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

Finally we have bounded the derivative of $\frac{\partial}{\partial \lambda_\ell} \hat{m}_\beta(\boldsymbol{\lambda})$. For every $\ell = 1, ..., J$, we have

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

We can sum up these bounds to bound the norm of the gradient $\nabla_\lambda \hat{m}_\beta(\boldsymbol{\lambda})$:

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

Since the training criterion is smooth, then $\hat{m}_{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ is continuous and differentiable over the line segment $\left\{ \alpha \boldsymbol{\lambda}^{(1)} + (1-\alpha) \boldsymbol{\lambda}^{(2)} : \alpha \in [0,1] \right\}$. Therefore by MVT, there is some $\alpha \in (0,1)$ such that

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

Recall that $\hat{m}_\beta(\boldsymbol{\lambda}^{(2)}) - \hat{m}_\beta(\boldsymbol{\lambda}^{(1)}) = 1$. Rearranging, we get

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

2 Additive Model

The function class of interest are the minimizers of the penalized least squares criterion:

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

where $\Lambda = [\lambda_{min}, \lambda_{max}]^J$.
Suppose that the penalties and the function $g(x|\boldsymbol{\theta})$ is convex wrt $\boldsymbol{\theta}$: $\nabla_\theta P_j(\boldsymbol{\theta})$ for all $j = 1, ..., J$ and $\nabla_\theta g(x|\boldsymbol{\theta} + m\boldsymbol{\beta})$ are PSD matrices.
Suppose there is some constant $K > 0$ such that for all $j = 1, ..., J$ and all $\boldsymbol{\beta}, \boldsymbol{\theta}$,

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

(This is essentially bounding the spectrum of the penalty function)
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 for all $j = 1, ..., J$

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

Proof

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

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

By definition, we know that $\hat{\boldsymbol{m}}(\boldsymbol{\lambda}^{(2)}) = \mathbf{1}$ and $\hat{\boldsymbol{m}}(\boldsymbol{\lambda^{(1)}}) = \mathbf{0}$.
1. We calculate $\nabla_\lambda \hat{\boldsymbol{m}}_{\boldsymbol{k}}(\boldsymbol{\lambda})$ using the implicit differentiation trick.
By the KKT conditions, we have for all $j = 1 : J$

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

Now we implicitly differentiate with respect to λ_ℓ for $\ell = 1, 2, ..., J$

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

By the product rule and chain rule, we have

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

Define the following matrices

$$\begin{aligned} S : S_{jk} &= \frac{\partial^2}{\partial m_k \partial m_j} \frac{1}{2} \|y - \sum_{j=1}^{J} g(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + m_j \boldsymbol{\beta}_j)\|_T^2 \Big|_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})} \\ &= \sum_{i=1}^{T} \left[\frac{\partial}{\partial m_k} g(x|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_k \boldsymbol{\beta}_k) \right]_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})} \left[\frac{\partial}{\partial m_j} g(x|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + m_j \boldsymbol{\beta}_j) \right]_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})} \end{aligned}$$

$$D_1 = diag \left(\frac{\partial^2}{\partial m_j^2} \lambda_j P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + m_j \boldsymbol{\beta}_j) \right) \Big|_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})}$$

$$D_2 = diag \left(\lambda_j w \|\boldsymbol{\beta}_j\|_2^2 \right)$$

$$D_3 = diag \left(\frac{\partial}{\partial m_\ell} P_\ell(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},\ell} + m_\ell \boldsymbol{\beta}_\ell) + w\langle \boldsymbol{\beta}_\ell, \hat{\boldsymbol{\theta}}_{\lambda^{(1)},\ell} + m_\ell \boldsymbol{\beta}_\ell \rangle \right) \Big|_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})}$$

$$M = \begin{pmatrix} \nabla_\lambda \hat{m}_1(\lambda) & \nabla_\lambda \hat{m}_2(\lambda) & ... & \nabla_\lambda \hat{m}_J(\lambda) \end{pmatrix}$$

We can then combine all the equations into the following system of equations:

$$M = -D_3 \left(S + D_1 + D_2 \right)^{-1}$$

S and D_1 are PSD matrices since we've assumed that g and the penalty functions are convex.
2. We bound every diagonal element in D_3:
By assumption, we know for every $k = 1, ..., J$

$$\left| \frac{\partial}{\partial m_k} P_k(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_k \boldsymbol{\beta}_k) \right| \leq K \|\boldsymbol{\beta}_k\| \tag{5}$$

Also,

$$\left| w\langle \boldsymbol{\beta}_k, \hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + \hat{m}_k(\boldsymbol{\lambda})\boldsymbol{\beta}_k\rangle \right| \le w\|\boldsymbol{\beta}_k\|\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + \hat{m}_k(\boldsymbol{\lambda})\boldsymbol{\beta}_k\| \tag{6}$$

To bound $\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + \hat{m}_k(\boldsymbol{\lambda})\boldsymbol{\beta}_k\|$, we use the basic inequality for $\hat{m}_k(\boldsymbol{\lambda})$:

$$\begin{aligned}
\frac{\lambda_k w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + \hat{m}_k(\boldsymbol{\lambda})\boldsymbol{\beta}_k\|^2 &\le \frac{1}{2}\|y - \sum_{j=1}^J g(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j})\|_T^2 + \sum_{j=1}^J \lambda_j \left(P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2\right) \\
&= \frac{1}{2}\|y - \sum_{j=1}^J g(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j})\|_T^2 + \sum_{j=1}^J \lambda_j^{(1)} \left(P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2\right) + \sum_{j=1}^J (\lambda_j - \lambda_j^{(1)}) \left(P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2\right) \\
&\le C + J\lambda_{max} \max_{j=1:J} \left(P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2\right)
\end{aligned}$$

To bound the term $\max_{j=1:J} \left(P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2\right)$, we use the basic inequality for $\hat{\boldsymbol{\theta}}_{\lambda^{(1)}}$:

$$\begin{aligned}
\sum_{j=1}^J \lambda_j^{(1)} \left(P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2\right) &\le \frac{1}{2}\|y - \sum_{j=1}^J g(\cdot|\hat{\boldsymbol{\theta}}_j^*)\|_T^2 + \sum_{j=1}^J \lambda_j^{(1)} \left(P_j(\hat{\boldsymbol{\theta}}_j^*) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_j^*\|_2^2\right) \\
&\le C
\end{aligned}$$

Since

$$\lambda_{min}\left(\max_{j=1:J} P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2\right) \le \sum_{j=1}^J \lambda_j^{(1)} \left(P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2\right)$$

then we have that

$$\max_{j=1:J} P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}) + \frac{w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j}\|_2^2 \le \frac{C}{\lambda_{min}}$$

Therefore for all $k = 1, ..., J$

$$\frac{\lambda_k w}{2}\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_\ell\boldsymbol{\beta}_k\|^2 \le \left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) C$$

Rearranging, we get

$$\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_\ell\boldsymbol{\beta}_k\| \le \sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min} w}} \tag{7}$$

Therefore

$$\left| \frac{\partial}{\partial m_k} P_k(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_k\boldsymbol{\beta}_k) + w\langle \boldsymbol{\beta}_k, \hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_k\boldsymbol{\beta}_k\rangle \right|_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})} \le K\|\boldsymbol{\beta}_k\| + w\|\boldsymbol{\beta}_k\|\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min} w}}$$

Let

$$D_{3,upper} = \left(K + w\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}}\right) diag\left(\|\boldsymbol{\beta}_k\|\right)$$

We know that $D_{3,upper} \succeq D_3$.

3. We bound the norm of $\nabla_\lambda \hat{m}_k(\lambda)$ for all $k = 1, ..., J$.

$$\begin{aligned}
\|\nabla_\lambda \hat{m}_k(\lambda)\| &= \|Me_k\| \\
&= \|D_3 \left(S + D_1 + D_2\right)^{-1} e_k\| \\
&\leq \|D_{3,upper} \left(S + D_1 + D_2\right)^{-1} e_k\| \\
&\leq \left(K + w\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}}\right) \max_\ell \|\boldsymbol{\beta}_\ell\| \left\|\left(S + D_1 + D_2\right)^{-1} e_k\right\| \\
&\leq \left(K + w\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}}\right) \max_\ell \|\boldsymbol{\beta}_\ell\| \left\|D_2^{-1}e_k\right\| \qquad (8)
\end{aligned}$$

The last line follows from the matrix inverse lemma: Since $S + D_1$ is a PSD matrix, then

$$\left\|\left(S + D_1 + D_2\right)^{-1} e_k\right\| \leq \left\|D_2^{-1}e_k\right\|$$

Now let

$$\ell_{max} = \arg\max_\ell \|\boldsymbol{\beta}_\ell\|$$

If we consider (8) for $k = \ell_{max}$, then

$$\begin{aligned}
\|\nabla_\lambda \hat{m}_{\ell_{max}}(\lambda)\| &\leq \left(K + w\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}}\right) \|\boldsymbol{\beta}_{\ell_{max}}\| \left\|D_2^{-1}e_{\ell_{max}}\right\| \\
&= \left(K + w\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}}\right) \|\boldsymbol{\beta}_{\ell_{max}}\| \lambda_{\ell_{max}}^{-1} w^{-1} \|\boldsymbol{\beta}_{\ell_{max}}\|_2^{-2} \\
&= \left(K + w\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}}\right) \|\boldsymbol{\beta}_{\ell_{max}}\|^{-1} \lambda_{min}^{-1} w^{-1}
\end{aligned}$$

Since the training criterion is smooth, then $\hat{m}_{\ell_{max}}(\lambda)$ is a continuous, differentiable function.

By the MVT, we have that there exists an $\alpha \in (0,1)$ such that

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

We know that $\hat{m}_k(\boldsymbol{\lambda}^{(2)}) - \hat{m}_k(\boldsymbol{\lambda}^{(1)}) = \mathbf{1}$ for all $k = 1, .., J$. Rearranging the inequality above, we get

$$\max_k \|\boldsymbol{\theta}_{\lambda^{(1)},k} - \boldsymbol{\theta}_{\lambda^{(2)},k}\| = \|\boldsymbol{\beta}_{\ell_{max}}\| \le \left\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\right\| \left|K + w\sqrt{\left(1+\frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}}\right| \lambda_{min}^{-1} w^{-1}$$

3 Nonsmooth Penalties

Suppose we are dealing with parametric regression problems from Section 1 or 2. We will suppose all the same assumptions, except those that concern the smoothness of the penalties.

Assumption modification (1): Suppose there is some $K > 0$ such that for all $j = 1, ..., J$ and any $\boldsymbol{\theta}, \boldsymbol{\beta}$ such that

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

Assumption modification (2): The non-smooth training criterion satisfy the Conditions 1, 2, and 3 from the Hillclimbing paper. Denote the differentiable space of $L_T(\cdot, \boldsymbol{\lambda})$ at any point $\boldsymbol{\theta}$ as

$$\Omega^{L_T(\cdot,\boldsymbol{\lambda})}(\boldsymbol{\theta})$$

For every $\boldsymbol{\lambda} \in \Lambda_{smooth}$, we have

Cond 1: The differentiable space of the training criterion at $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$, denoted $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\right)$, is a local optimality space.

Cond 2: The training criterion $L_T(\cdot,\cdot)$ restricted to $\Omega^{L_T(\cdot,\cdot)}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}), \boldsymbol{\lambda}\right)$ is twice continuously differentiable within some ball centered $\boldsymbol{\lambda}$. Let "ball of differentiability" be denoted $B(\boldsymbol{\lambda})$.

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

Suppose that

$$\mu(\Lambda_{smooth}^C) = 0$$

Under these non-smooth conditions, the same Lipschitz condition will hold.

Proof

Now define

$$L_{nonsmooth} = \left\{\text{line that passes through } \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 : \boldsymbol{\lambda}_1, \boldsymbol{\lambda}_2 \in \Lambda_{smooth}^C\right\}$$

Unproven Claim: Since $\mu(\Lambda_{smooth}^C) = 0$, then $\mu(L_{nonsmooth}) = 0$. I don't know how to prove this clain, but it seems true.

Now denote the line segment between $\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}$ as

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

The set

$$H = \left\{(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}) : \left\|\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}) \cap \Lambda_{smooth}^C\right\| > 0\right\}$$

has measure $\mu(H) = 0$ since $H \subseteq L_{nonsmooth}$.

Now consider any line segment $\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})$ not in H^C. We want to show that there is a set of points $\{\boldsymbol{\ell}^{(i)}\}$ along $\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})$ such that the "balls of differentiability" $B(\boldsymbol{\ell}^{(i)})$ cover the entire line segment. We will define a function to measure this uncovered distance: For a given set of points $\{\boldsymbol{\ell}^{(i)}\} \subset \mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})$, let $d(\{\boldsymbol{\ell}^{(i)}\})$ denote the covered distance of $\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})$ by the union of their differentiable spaces:

$$d_{\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}}(\{\boldsymbol{\ell}^{(i)}\}) = \left\|\left[\cup_i B(\boldsymbol{\ell}^{(i)})\right] \cap \mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})\right\|$$

Claim: For all $(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}) \in H^C$, there is a set of points $\{\boldsymbol{\ell}^{(i)}\} \subseteq \mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})$ such that their "balls of differentiability" completely cover $\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})$:

$$\max d_{\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}}\left(\left\{\boldsymbol{\ell}^{(i)}\right\}\right) = \|\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})\|$$

Proof of Claim:

For contradiction, suppose that no set of points can cover the line segment. For notational convenience, let us write

$$\bar{\boldsymbol{\ell}}_{max} = \arg\max_{\{\ell^{(i)}\}} d_{\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}}(\{\boldsymbol{\ell}^{(i)}\})$$

So

$$d_{\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}}\left(\bar{\boldsymbol{\ell}}_{max}\right) < \|\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})\|$$

Let $\mathcal{L}_U$ be the set of points left uncovered:

$$\mathcal{L}_{uncovered} = \mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}) \backslash \left[\cup_{\ell \in \bar{\boldsymbol{\ell}}_{max}} B(\boldsymbol{\ell})\right]$$

So

$$\left\|\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})}) \cap U\right\| < \|\mathcal{L}(\boldsymbol{\lambda}^{(\mathbf{1})}, \boldsymbol{\lambda}^{(\mathbf{2})})\|$$

There are two cases:

(1) $\mathcal{L}_{uncovered} \subseteq \Lambda^C_{smooth}$. Then $\left\|\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \cap \Lambda^C_{smooth}\right\| \geq \left\|\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \cap \mathcal{L}_{uncovered}\right\| > 0$. This is clearly impossible since $(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \in H^C$.

(2) There exists a point $\boldsymbol{p} \in \mathcal{L}_{uncovered} \backslash \Lambda^C_{smooth}$. Since $\boldsymbol{p} \in \Lambda_{smooth}$, then by Condition 2, then the neighborhood $B(\boldsymbol{p})$ is non-empty.

$$\left\| B(\boldsymbol{p}) \cap \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \right\| > 0$$

This implies that

$$d_{\boldsymbol{\lambda^{(1)}},\boldsymbol{\lambda^{(2)}}}\left(\bar{\boldsymbol{\ell}}_{max}\right) < d_{\boldsymbol{\lambda^{(1)}},\boldsymbol{\lambda^{(2)}}}\left(\bar{\boldsymbol{\ell}}_{max} \cup \{\boldsymbol{p}\}\right)$$

However contradicts the definition of $\bar{\boldsymbol{\ell}}_{max}$ that it maximizes the covered distance.

End of Proof

From the claim above, let's consider any $(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \in H^C$. Let

$$\bar{\boldsymbol{\ell}}_{max} = \arg\max_{\left\{\boldsymbol{\ell}^{(i)}\right\}} d_{\boldsymbol{\lambda^{(1)}},\boldsymbol{\lambda^{(2)}}}\left(\left\{\boldsymbol{\ell}^{(i)}\right\}\right)$$

Then define the intersections of the edges of the "balls of differentiability" with the line segment $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$.

$$P = \left\{\text{The points at the edge of } B(\boldsymbol{\ell}) \text{ that intersect with } \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) : \boldsymbol{\ell} \in \bar{\boldsymbol{\ell}}_{max}\right\} \cup \{\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\}$$

Since every point can be expressed as $\alpha_{p^{(i)}}\boldsymbol{\lambda^{(1)}} + (1-\alpha_{p^{(i)}})\boldsymbol{\lambda^{(2)}}$ for some $\alpha_{p^{(i)}} \in [0,1]$, we can order these points $\{\boldsymbol{p}^{(i)}\}$ by increasing $\alpha_{p^{(i)}}$.

By definition of P and the Claim, the differentiable space of the training criterion over $\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$ must be constant.

We can apply the smoothness result in Section 1 or 2 over every interval $\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$ since we can come up with an equivalent definition for $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$: There is an orthonormal matrix $U^{(i)}$ such that for all $\boldsymbol{\lambda} \in \left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$

$$\hat{\boldsymbol{\theta}}_\lambda = U^{(i)}\hat{\boldsymbol{\beta}}_\lambda$$

$$\hat{\boldsymbol{\beta}}_\lambda = \arg\min_\beta L_T(U^{(i)}\boldsymbol{\beta}, \boldsymbol{\lambda})$$

where the training criterion is smooth over $\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$ wrt to the directional derivatives along the columns of $U^{(i)}$.

For example, in the case of Section 1, we would instead consider regression problems of the form

$$\begin{aligned}
\hat{\boldsymbol{\beta}}_\lambda &= \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|y - g(\cdot|U\boldsymbol{\beta})\|_T^2 + \sum_{j=1}^J \lambda_j \left(P_j(U\boldsymbol{\beta}) + \frac{w}{2}\|U\boldsymbol{\beta}\|_2^2\right) \\
&= \arg\min_{\boldsymbol{\beta}} \frac{1}{2}\|y - g(\cdot|U\boldsymbol{\beta})\|_T^2 + \sum_{j=1}^J \lambda_j \left(P_j(U\boldsymbol{\beta}) + \frac{w}{2}\|\boldsymbol{\beta}\|_2^2\right)
\end{aligned}$$

The proof from Sections 1 and 2 would need to be modified to take directional derivatives along the columns of U. Applying the Section 1 or 2 results to each interval $\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$, we would get Lipschitz conditions of the form

$$\|\hat{\boldsymbol{\beta}}_{p^{(i)}} - \hat{\boldsymbol{\beta}}_{p^{(i+1)}}\|_2 \le c\|\boldsymbol{p^{(i)}} - \boldsymbol{p^{(i+1)}}\|_2$$

where c is some constant.

Finally, we can sum up these inequalities to show smoothness of $\hat{\boldsymbol{\theta}}_\lambda$. By the triangle inequality,

$$\begin{aligned}
\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(1)}}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda^{(2)}}}\|_2 &\le \sum_{i=1} \|\hat{\boldsymbol{\theta}}_{p^{(i)}} - \hat{\boldsymbol{\theta}}_{p^{(i+1)}}\|_2 \\
&= \sum_{i=1} \|\hat{\boldsymbol{\beta}}_{p^{(i)}} - \hat{\boldsymbol{\beta}}_{p^{(i+1)}}\|_2 \\
&\le \sum_{i=1} c\|\boldsymbol{p^{(i)}} - \boldsymbol{p^{(i+1)}}\|_2 \\
&= c\|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|_2
\end{aligned}$$

4 Example penalties that satisfy the conditions

Ridge:

The perturbation isn't necessary if there is already a ridge penalty in the original penalized regression problem. Just set the penalties $P_j(\boldsymbol{\theta}) \equiv 0$ and fix $w = 2$.

Lasso:

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

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

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

Group Lasso:

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