Proofs for Smoothness of Parametric Regression Models

November 3, 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 $\boldsymbol{\theta}^*$ is the model such that

$$\boldsymbol{\theta}^* = \arg\min_{\theta} E_{x,y}\left[(y - g(x|\boldsymbol{\theta}))^2\right]$$

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

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

Document Outline

First, we consider smooth training criteria 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

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

Then we will extend these results to non-smooth penalty functions.

Finally we will consider examples of parametric penalty functions. This includes a deep dive into the Sobolev penalty.

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 $\boldsymbol{\theta}, \boldsymbol{\beta}, m$, 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

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

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

where

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

Proof

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

Define

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

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

1. We calculate $\nabla_\lambda \hat{m}_{\boldsymbol{\beta}}(\boldsymbol{\lambda})$ using the implicit differentiation trick.

By the KKT conditions, we have

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

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

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

By the product rule and chain rule, we have

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

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

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

In vector notation, we have

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

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

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

2. Bound $\|\nabla_\lambda \hat{m}_{\boldsymbol{\beta}}(\boldsymbol{\lambda})\|$
Bounding the first multiplicand:
The first multiplicand is bounded by

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

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

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

The second summand in the second multiplicand is bounded by

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

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

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

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

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

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

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

We can combine (4) with the fact that

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

to get

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

Plug the inequality above into (1) to get

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

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

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

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

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

3. Apply Mean Value Theorem

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

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

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:

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

where $\Lambda = [\lambda_{min}, \lambda_{max}]^J$.

Suppose that the penalties and the mean squared error $\|y - \sum_{j=1}^J g_j(x|\boldsymbol{\theta}_j)\|_T^2$ are twice-differentiable and convex wrt $\boldsymbol{\theta}$: $\nabla^2_{\boldsymbol{\theta}_j} P_j(\boldsymbol{\theta}_j)$ for all $j = 1, ..., J$ and $\nabla^2_{\boldsymbol{\theta}}\|y - \sum_{j=1}^J g_j(x|\boldsymbol{\theta}_j)\|_T^2$ are PSD matrices.

Suppose for each $j = 1, ..., J$, there is a constant $K_j \ge 0$ such that for all $\boldsymbol{\beta}, \boldsymbol{\theta}$, we either have

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

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

$$\|\nabla_\theta g_j(X_{T,j}|\boldsymbol{\theta})\| \le K_j$$

(We are bounding the spectral norm of $\nabla_\theta g_j(X_{T,j}|\boldsymbol{\theta})$. This case is most relevant to linear models I believe.)

Let

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

and

$$d_{max} = \max_{k=1,\dots,J} d_k$$

where

$$d_k = \begin{cases} \left(K + w\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min} w}}\right) & \text{if } \left|\frac{\partial}{\partial m} P_k(\boldsymbol{\theta} + m\boldsymbol{\beta})\right| \le K\|\boldsymbol{\beta}\| \\ \frac{1}{\lambda_{min}} K\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min} w}} & \text{if } \|\nabla_{\boldsymbol{\theta}_k} g_k\left(X_{T,k}|\boldsymbol{\theta}_k\right)\| \le K \end{cases}$$

Then for any $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda$ we have for all $j = 1, ..., J$

$$\|\boldsymbol{\theta}_{\lambda^{(1)},j} - \boldsymbol{\theta}_{\lambda^{(2)},j}\| \le \left\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\right\| \lambda_{min}^{-1} w^{-1} d_{max}$$

Proof

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

Define

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

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 $j = 1 : J$

$$\frac{\partial}{\partial m_j}\left(\frac{1}{2}\left\| y - \sum_{j=1}^{J} g_j(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + m_j\boldsymbol{\beta}_j)\right\|_T^2 + \lambda_j P_j(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + m_j\boldsymbol{\beta}_j)\right) + \lambda_j w\langle\boldsymbol{\beta}_j, \hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + m_j\boldsymbol{\beta}_j\rangle\Bigg|_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})} = 0 \quad (5)$$

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

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

By the product rule and chain rule, we have

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

$$+1[j=\ell] \left\{ \frac{\partial}{\partial m_\ell} P_\ell(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},\ell} + m_\ell \boldsymbol{\beta}_\ell) + w\langle \boldsymbol{\beta}_\ell, \hat{\boldsymbol{\theta}}_{\lambda^{(1)},\ell} + m_\ell \boldsymbol{\beta}_\ell \rangle \right\} \Bigg|_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})} = 0$$

Define the following matrices

$$S : S_{jk} = \frac{\partial^2}{\partial m_k \partial m_j} \frac{1}{2} \|y - \sum_{j=1}^{J} g_j(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + m_j \boldsymbol{\beta}_j)\|_T^2 \Bigg|_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})}$$

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

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

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

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

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 is a PSD matrix since the sum of convex functions is convex (so sum of g_j is convex) and the composition of a convex function with an affine function is convex.

D_1 is a PSD matrix since the penalty functions are convex.

2. We bound every diagonal element in D_3:

By Cauchy-Schwarz,

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

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

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

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 $\hat{\boldsymbol{\theta}}_{\lambda^{(1)}}$:

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

Since

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

then we have that

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

Therefore

$$\left\| y - \sum_{j=1}^{J} g_j(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + \hat{m}_j(\boldsymbol{\lambda})\boldsymbol{\beta}_j) \right\|_T^2 + \sum_{j=1}^{J} \lambda_j \left(P_j \left(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + \hat{m}_k(\boldsymbol{\lambda})\boldsymbol{\beta}_k \right) + \frac{w}{2} \|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + \hat{m}_k(\boldsymbol{\lambda})\boldsymbol{\beta}_k\|^2 \right) \leq \left(1 + \frac{J\lambda_{max}}{\lambda_{min}} \right) C$$

This implies that

$$\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_\ell \boldsymbol{\beta}_k\| \leq \sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}} \right) \frac{2C}{\lambda_{min} w}} \tag{6}$$

and

$$\left\| y - \sum_{j=1}^{J} g_j(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + \hat{m}_j(\boldsymbol{\lambda})\boldsymbol{\beta}_j) \right\|_T \leq \sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min} w}} \tag{7}$$

If combine the assumption

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

with (6), we get

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

On the other hand, suppose the other assumption is satisfied:

$$\|\nabla_{\boldsymbol{\theta}_k} g_k\left(X_{T,k}|\boldsymbol{\theta}_k\right)\| \leq K$$

Then we will need to use the implicit differentiation equation (5). Rearranging, we get

$$\begin{aligned}
\frac{\partial}{\partial m_k}\left(P_k(\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_k\boldsymbol{\beta}_k)\right)\Big|_{\boldsymbol{m}=\hat{\boldsymbol{m}}(\boldsymbol{\lambda})} + w\langle \boldsymbol{\beta}_k, \hat{\boldsymbol{\theta}}_{\lambda^{(1)},k} + m_k\boldsymbol{\beta}_k \rangle &= \frac{1}{\lambda_k}\left\langle \nabla_{\boldsymbol{\theta}_k} g_k\left(\cdot|\boldsymbol{\theta}_k\right)|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k}+m_k\boldsymbol{\beta}_k} \boldsymbol{\beta}_k, y - \sum_{j=1}^{J} g_j(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + m_j\boldsymbol{\beta}_j) \right\rangle_T \\
&\leq \frac{1}{\lambda_{min}}\|\boldsymbol{\beta}_k\| \left\| \nabla_{\boldsymbol{\theta}_k} g_k\left(X_{T,k}|\boldsymbol{\theta}_k\right)|_{\boldsymbol{\theta}_k=\hat{\boldsymbol{\theta}}_{\lambda^{(1)},k}+m_k\boldsymbol{\beta}_k} \right\| \left\| y - \sum_{j=1}^{J} g_j(\cdot|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} + \hat{m}_j(\boldsymbol{\lambda})\boldsymbol{\beta}_j) \right\|_T
\end{aligned}$$

Plugging in (7), we get

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

Using these upper bounds, we can bound D_3 by the diagonal matrix

$$d_{max} diag\left(\{\|\beta\|_k\}_{k=1}^J\right) \succeq D_3$$

where

$$d_{max} = \max_{k=1,\dots,J} d_k$$

and

$$d_k = \begin{cases} \left(K + w\sqrt{\left(1+\frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}}\right) & \text{if } \left|\frac{\partial}{\partial m}P_k(\boldsymbol{\theta}+m\boldsymbol{\beta})\right| \leq K\|\boldsymbol{\beta}\| \\ \frac{1}{\lambda_{min}}K\sqrt{\left(1+\frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min}w}} & \text{if } \|\nabla_{\boldsymbol{\theta}_k} g_k\left(X_{T,k}|\boldsymbol{\theta}_k\right)\| \leq K \end{cases}$$

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

$$\begin{aligned} \|\nabla_\lambda \hat{m}_k(\lambda)\| &= \|Me_k\| \\ &= \|D_3\left(S+D_1+D_2\right)^{-1}e_k\| \\ &\leq \left\|d_{max}diag\left(\{\|\beta\|_k\}_{k=1}^J\right)\left(S+D_1+D_2\right)^{-1}e_k\right\| \\ &\leq d_{max}\max_\ell \|\boldsymbol{\beta}_\ell\| \left\|\left(S+D_1+D_2\right)^{-1}e_k\right\| \\ &\leq d_{max}\max_\ell \|\boldsymbol{\beta}_\ell\| \left\|D_2^{-1}e_k\right\| \end{aligned} \tag{8}$$

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 consider (8) for

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

We have

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

4. Apply the Mean Value Theorem
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

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

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

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

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.

Suppose $\Lambda \subseteq \mathbb{R}^p$. We suppose that for almost every dataset (X, y), the following hold:

Assumption (1): Let the penalized training criterion be denoted $L_T(\cdot, \boldsymbol{\lambda})$. Denote the differentiable space of $L_T(\cdot, \boldsymbol{\lambda})$ at any point $\boldsymbol{\theta}$ as $\Omega^{L_T(\cdot,\boldsymbol{\lambda})}(\boldsymbol{\theta})$. Suppose there is a set $\Lambda_{smooth} \subseteq \Lambda$ such that for every $\boldsymbol{\lambda} \in \Lambda_{smooth}$, the following conditions hold

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

Cond 2: The training criterion $L_T(\cdot, \cdot)$ restricted to $\Omega^{L_T(\cdot,\cdot)}\left(\hat{\boldsymbol{\theta}}_\lambda, \boldsymbol{\lambda}\right)$ 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.

Furthermore, suppose that

$$\mu\left(\Lambda \backslash \Lambda_{smooth}\right) = 0$$

where μ is the Lebesgue measure in p-dimensions.

Assumption (2): For every $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$, let the line segment between the two points be denoted

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

Suppose the intersection $\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \cap \Lambda_{smooth}^C$ is countable.

Assumption (3): All the conditions that bound the spectrum of P_j or g_j only need to apply when the directional derivatives exist. That is, the condition on the spectrum of the penalty derivative is now

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

Similarly, we would change the condition on the spectrum of the function derivative to

$$\|\nabla_{\boldsymbol{\theta}} g_j\left(\boldsymbol{\theta}\right)\| \le K \text{ if } \nabla_{\boldsymbol{\theta}} g_j\left(\boldsymbol{\theta}\right) \text{exists}$$

Under these assumptions, the same Lipschitz conditions will hold.

Proof

Consider any $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda_{smooth}$. We define the length of $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ covered by set A as

$$d_{\boldsymbol{\lambda^{(1)}},\boldsymbol{\lambda^{(2)}}}(A) = \mu\left(A \cap \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})\right)$$

where μ is the Lebesgue measure over the line segment $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$.

By the Differentiability Cover Lemma below, there exists a countable set of points $\cup_{i=1}^{\infty} \boldsymbol{\ell}^{(i)} \subset \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ such that the union of their "balls of differentiabilities" entirely cover $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$:

$$\max_{\{\boldsymbol{\ell}^{(i)}\}_{i=1}^{\infty}} d_{\boldsymbol{\lambda^{(1)}},\boldsymbol{\lambda^{(2)}}}\left(\cup_{i=1}^{\infty} B(\boldsymbol{\ell}^{(i)})\right) = \mu\left(\mathcal{L}\left(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}\right)\right)$$

Let

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

Let P be the intersections of the boundary of $B(\boldsymbol{\ell}^{(i)})$ with the line segment $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$:

$$P = \left\{\cup_{i=1}^{\infty} \mathrm{bd} B(\boldsymbol{\ell}^{(i)}) \cap \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})\right\} \cup \{\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\}$$

The point $p \in P$ can be expressed as $\alpha_p \boldsymbol{\lambda^{(1)}} + (1-\alpha_p)\boldsymbol{\lambda^{(2)}}$ for some $\alpha_p \in [0,1]$. This means we can order these points $\{\boldsymbol{p}^{(i)}\}_{i=1}^{\infty}$ by increasing α_p. By our assumptions, the differentiable space of the training criterion over line segment $\mathcal{L}\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$ must be constant.

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

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

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

where the training criterion is smooth over $\mathcal{L}\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$ 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. We can establish that there is a constant $c > 0$ such that for every tuple of points $\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$ from $i = 1, 2...$, we have

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

Finally, we can sum up these inequalities. By the triangle inequality,

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

Lemma - Differentiability Cover

For any $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda_{smooth}$, there exists a countable set of points $\cup_{i=1}^{\infty} \boldsymbol{\ell}^{(i)} \subset \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ such that the union of their "balls of differentiabilities" entirely cover $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$

$$\max_{\{\boldsymbol{\ell}^{(i)}\}_{i=1}^{\infty}} d_{\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}} \left(\cup_{i=1}^{\infty} B(\boldsymbol{\ell}^{(i)})\right) = \left\|\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})\right\|$$

Proof

We prove this by contradiction. Let

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

and for contradiction, suppose that the covered length is less than the length of the line segment:

$$d_{\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}} \left(\cup_{i=1}^{\infty} B(\boldsymbol{\ell}_{max}^{(i)})\right) < \left\|\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})\right\|$$

By assumption (2), since $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \cap \Lambda_{smooth}^C$ is countable, there must exist a point $p \in \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \backslash \left\{\cup_{i=1}^{\infty} B(\boldsymbol{\ell}_{max}^{(i)})\right\}$ such that $p \notin \Lambda_{smooth}^C$. However if we consider the set of points $\left\{\boldsymbol{\ell}_{max}^{(i)}\right\}_{i=1}^{\infty} \cup \{p\}$, then

$$d_{\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}} \left(\cup_{i=1}^{\infty} B(\boldsymbol{\ell}_{max}^{(i)})\right) < d_{\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}} \left(\cup_{i=1}^{\infty} B(\boldsymbol{\ell}_{max}^{(i)}) \cup B(p)\right)$$

This is a contradiction of the definition of $\{\boldsymbol{\ell}_{max}^{(i)}\}$. Therefore we should always be able to cover $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ with "balls of differentiability."

4 Example

4.1 Penalties that satisfy the conditions

We will show penalties that satisfy the condition

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

for some constant $K > 0$.

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:

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

so $K = p$ in this case.

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

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

so $K = pG$ in this case.

Group Lasso:

If we have un-pooled penalty parameters as follows

$$\sum_{j=1}^{J} \lambda_j \|\boldsymbol{\theta}^{(j)} + m^{(j)}\boldsymbol{\beta}^{(j)}\|_2$$

then we need the following bound for every $j = 1, ..., J$

$$\begin{aligned} \frac{\partial}{\partial m^{(j)}}\|\boldsymbol{\theta}^{(j)} + m^{(j)}\boldsymbol{\beta}^{(j)}\|_2 &= \left\langle \frac{\boldsymbol{\theta}^{(j)} + m^{(j)}\boldsymbol{\beta}^{(j)}}{\|\boldsymbol{\theta}^{(j)} + m^{(j)}\boldsymbol{\beta}^{(j)}\|_2}, \boldsymbol{\beta}^{(j)} \right\rangle \\ &\leq \|\boldsymbol{\beta}^{(j)}\|_2 \end{aligned}$$

So $K = 1$ in this case.

If there is a single penalty parameter for the entire group laso penalty as follows

$$\lambda \sum_{j=1}^{J} \|\boldsymbol{\theta}^{(j)} + m\boldsymbol{\beta}^{(j)}\|_2$$

then

$$\begin{aligned} \frac{\partial}{\partial m} \sum_{j=1}^{J} \|\boldsymbol{\theta}^{(j)} + m\boldsymbol{\beta}^{(j)}\|_2 &= \sum_{j=1}^{J} \left\langle \frac{\boldsymbol{\theta}^{(j)} + m\boldsymbol{\beta}^{(j)}}{\|\boldsymbol{\theta}^{(j)} + m\boldsymbol{\beta}^{(j)}\|_2}, \boldsymbol{\beta}^{(j)} \right\rangle \\ &\le \sum_{j=1}^{J} \|\boldsymbol{\beta}^{(j)}\|_2 \\ &\le \sqrt{J}\|\boldsymbol{\beta}\|_2 \end{aligned}$$

and $K = \sqrt{J}$.

4.2 Sobolev

Given a function h, the Sobolev penalty for h is

$$P(h) = \int (h^{(r)}(x))^2 dx$$

The Sobolev penalty is used in nonparametric regression models, but such nonparametric regression models can be re-expressed in parametric form. We will use this to understand the smoothness of models fitted in this manner.

Consider the class of smoothing splines

$$\left\{ \hat{g}(\cdot|\lambda) = \arg\min_{g \in \mathcal{G}} \frac{1}{2} \|y - \sum_{j=1}^{J} g_j(x_j)\|_T^2 + \sum_{j=1}^{J} \lambda_j P(g_j) : \lambda \in \Lambda \right\}$$

Each function $\hat{g}_j(\cdot|\lambda)$ is a spline that can be expressed as the weighted sum of B normalized B-splines of degree $r+1$ for a given set of knots:

$$\hat{g}_j(x|\lambda) = \sum_{i=1}^{B} \theta_i N_{j,i}(x)$$

Note that the normalized B-splines have the property that they sum up to one at all points within the boundary of the knots. Also recall that B-splines are non-negative.

Therefore we can re-express the class of smoothing splines as a set of function parameters

$$\left\{\hat{\boldsymbol{\theta}}_\lambda = \arg\min_\theta \frac{1}{2}\|y - \sum_{j=1}^J N_{T,j}\boldsymbol{\theta}_j\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta}_j) : \lambda \in \Lambda\right\}$$

where $N_{T,j}$ is the normalized B-spline basis for the given set of knots evaluated at the observed x_j in the training set. $P_j(\boldsymbol{\theta_j})$ is the Sobolev penalty and can be written as $\boldsymbol{\theta}_j^T \Omega_j \boldsymbol{\theta}_j$ for an appropriate penalty matrix Ω_j. We will not need to express anything in terms of Ω_j so the penalty will be just written as $P_j(\boldsymbol{\theta}_j)$.

Instead of considering the original smoothing spline problem with the roughness penalty, we will add a ridge penalty on the function parameters

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

Let

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

Then for any $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in \Lambda$ we have for all $j = 1, ..., J$

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

Moreover,

$$\left\|\sum_{j=1}^J \hat{g}_j(x_j|\lambda^{(1)}) - \hat{g}_j(x_j|\lambda^{(2)})\right\|_\infty \le \left\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\right\|_2 J\sqrt{B}\lambda_{min}^{-1} w^{-1}\left(\frac{1}{\lambda_{min}} B\sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)\frac{2C}{\lambda_{min} w}}\right)$$

Proof

To apply the result from Section 2, we just need to bound the spectral norm

$$\|\nabla_\theta g_j(X_{T,j}|\boldsymbol{\theta})\| = \|N_{T,j}\|$$

Note that the eigenvalue of $N_{T,j}$ is bounded by B since the maximum eigenvalue of a non-negative matrix is bounded by its maximum row sum. In the case of $N_{T,j}$, since it is the values of normalized B-splines, each row is at most the number of B-spline basis functions. That is, we have for all $j = 1, ..., J$

$$\|\nabla_\theta g_j(X_{T,j}|\boldsymbol{\theta})\| = \|N_{T,j}\| \le B$$

Apply the result from Section 2 to get the result

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

The "moreover" statement follows from the fact that for any point $\boldsymbol{x}$, we have

$$\begin{aligned}
\left|\sum_{j=1}^{J} \hat{g}_j(x_j|\boldsymbol{\lambda}^{(1)}) - \hat{g}_j(x_j|\boldsymbol{\lambda}^{(2)})\right| &= \left|\sum_{j=1}^{J}\sum_{i=1}^{B} \left(\hat{\theta}_{\lambda^{(1)},j,i} - \hat{\theta}_{\lambda^{(2)},j,i}\right) N_{j,i}(x_j)\right| \\
&\leq \sum_{j=1}^{J}\sum_{i=1}^{B} \left|\left(\hat{\theta}_{\lambda^{(1)},j,i} - \hat{\theta}_{\lambda^{(2)},j,i}\right) N_{j,i}(x_j)\right| \\
&\leq \sum_{j=1}^{J}\sum_{i=1}^{B} \left|\hat{\theta}_{\lambda^{(1)},j,i} - \hat{\theta}_{\lambda^{(2)},j,i}\right| \\
&\leq \sum_{j=1}^{J} \|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} - \hat{\boldsymbol{\theta}}_{\lambda^{(2)},j}\|_1 \\
&\leq \sqrt{B}\sum_{j=1}^{J} \|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} - \hat{\boldsymbol{\theta}}_{\lambda^{(2)},j}\|_2
\end{aligned}$$

where the second inequality uses the fact that normalized B-splines have value at most 1. Therefore

$$\left\|\sum_{j=1}^{J} \hat{g}_j(x_j|\lambda^{(1)}) - \hat{g}_j(x_j|\lambda^{(2)})\right\|_\infty \leq \sqrt{B}\sum_{j=1}^{J} \left\|\hat{\boldsymbol{\theta}}_{\lambda^{(1)},j} - \hat{\boldsymbol{\theta}}_{\lambda^{(2)},j}\right\|$$