

# Proofs for Smoothness of Parametric Regression Models

November 7, 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

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

Technically, all the proofs require is that  $\boldsymbol{\theta}^* \in \Theta$  is fixed. In the convergence rate proofs, we will need  $\boldsymbol{\theta}^*$  to satisfy  $E[y|x] = g(x|\boldsymbol{\theta}^*)$ . We are interested in establishing inequalities of the form

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

If the functions are  $L$ -Lipschitz in their parameterization, we will also be able to bound the distance between the actual functions. That is, if there is a constant  $L > 0$  such that for all  $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$

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

Then

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

## Document Outline

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

1. Multiple penalties for a single model

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

## 2. Additive model

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

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:

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

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 twice-differentiable and convex wrt  $\boldsymbol{\theta}$ :

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

Suppose there is some constants  $K_1, K_0 > 0$  such that for all  $j = 1, \dots, J$  and any  $\boldsymbol{\theta}'$ , we have

$$|\nabla_{\boldsymbol{\theta}} P_j(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}| \leq K_1 \|\boldsymbol{\theta}'\|_2 + K_0$$

Let

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

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

$$\|\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(1)}} - \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}^{(2)}}\|_2 \leq \frac{1}{\lambda_{min} w J} \left( (K_1 + w) \sqrt{\frac{2}{\lambda_{min} w} C_{\boldsymbol{\theta}^*, \Lambda} + K_0} \right) \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2$$

Moreover, if  $g(\cdot | \boldsymbol{\theta})$  is  $L$ -Lipschitz wrt  $\|\cdot\|_{\infty}$ , then

$$\|g(\cdot | \boldsymbol{\theta}_1) - g(\cdot | \boldsymbol{\theta}_2)\|_{\infty} \leq L \frac{1}{\lambda_{min} w J} \left( (K_1 + w) \sqrt{\frac{2}{\lambda_{min} w} C_{\boldsymbol{\theta}^*, \Lambda} + K_0} \right) \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2$$

**Proof**

**1. We calculate  $\nabla_{\lambda} \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$  using the implicit differentiation trick.**

By the KKT conditions, we have

$$\nabla_{\boldsymbol{\theta}} \left( \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta}) \right) + \sum_{j=1}^J \lambda_j w \boldsymbol{\theta} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} = 0$$

Now we implicitly differentiate with respect to  $\boldsymbol{\lambda}$

$$\left[ \nabla_{\boldsymbol{\theta}}^2 \left( \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j(\boldsymbol{\theta}) + \frac{w}{2} \|\boldsymbol{\theta}\|^2 \right) \right) \nabla_{\lambda} \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) + \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) + w \boldsymbol{\theta} \vec{\mathbf{1}}_J^{\top} \right] \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} = 0$$

where

$$\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) = \{ \nabla_{\boldsymbol{\theta}} P_1(\boldsymbol{\theta}) \quad \dots \quad \nabla_{\boldsymbol{\theta}} P_J(\boldsymbol{\theta}) \}$$

Rearranging, we have for all  $\boldsymbol{\lambda} \in \Lambda$

$$\nabla_{\lambda} \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = - \left[ \nabla_{\boldsymbol{\theta}}^2 \left( \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j(\boldsymbol{\theta}) + \frac{w}{2} \|\boldsymbol{\theta}\|^2 \right) \right) \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}^{-1} \left( \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta}) + w \boldsymbol{\theta} \vec{\mathbf{1}}_J^{\top} \right) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}$$

**2. Bound  $\|\nabla_{\lambda} \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})\|$  for  $i = 1, \dots, p$**

We know that

$$\begin{aligned}
\|\nabla_{\lambda}\hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})\| &= \left\| e_i^{\top} \left[ \nabla_{\boldsymbol{\theta}}^2 \left( \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 \right) \right) \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}^{-1} \left( \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} + w\boldsymbol{\theta}\tilde{\mathbf{I}}_J^{\top} \right) \right\| \\
&= \left\| e_i^{\top} \left[ \nabla_{\boldsymbol{\theta}}^2 \left( \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 \right) \right) \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}^{-1} \left( \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} + w\boldsymbol{\theta}\tilde{\mathbf{I}}_J^{\top} \right) \right\| \\
&\leq \left\| \left[ \nabla_{\boldsymbol{\theta}}^2 \left( \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta}) \right) \right]_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} + \sum_{j=1}^J \lambda_j w I \right]^{-1} \right\| \left( \left\| \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\|_F + w \left\| \boldsymbol{\theta}\tilde{\mathbf{I}}_J^{\top} \right\| \right) \\
&\leq \left\| \left[ \sum_{j=1}^J \lambda_j w I \right]^{-1} \right\| \left( \left\| \nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\|_F + w\sqrt{J}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2 \right) \\
&\leq \frac{1}{J\lambda_{\min}w} \left( \sqrt{J} \left( K_1 \|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2 + K_0 \right) + w\sqrt{J}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2 \right) \\
&= \frac{(K_1 + w) \|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2 + K_0}{\lambda_{\min}w\sqrt{J}}
\end{aligned}$$

The second inequality follows from the assumption that  $\frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta})$  is convex in  $\boldsymbol{\theta}$ . The last inequality follows from the assumption  $\|\nabla_{\boldsymbol{\theta}} P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}\|_F \leq K\|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2$ .

We can use the definition of  $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$  to bound  $\|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2$ . By definition,

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

So

$$\|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2 \leq \sqrt{\frac{2}{J\lambda_{\min}w} C_{\boldsymbol{\theta}^*, \Lambda}}$$

Hence for all  $\lambda \in \Lambda$

$$\left\| \nabla_{\lambda} \hat{\theta}_i(\lambda) \right\| \leq \frac{1}{\lambda_{\min} w J} \left( (K_1 + w) \sqrt{\frac{2}{\lambda_{\min} w} C_{\theta^*, \Lambda} + K_0} \right)$$

#### 4. Put all the bounds together

By the mean value theorem, there is a  $\alpha \in (0, 1)$  such that

$$\begin{aligned} \left\| \hat{\theta}_i(\lambda^{(1)}) - \hat{\theta}_i(\lambda^{(2)}) \right\| &\leq \left\langle \nabla_{\lambda} \hat{\theta}_i(\lambda) \Big|_{\lambda = \alpha \lambda^{(1)} + (1-\alpha) \lambda^{(2)}}, \lambda^{(1)} - \lambda^{(2)} \right\rangle \\ &\leq \max_{\lambda \in \Lambda} \left\| \nabla_{\lambda} \hat{\theta}_i(\lambda) \right\| \left\| \lambda^{(1)} - \lambda^{(2)} \right\| \\ &\leq \frac{1}{\lambda_{\min} w J} \left( (K_1 + w) \sqrt{\frac{2}{\lambda_{\min} w} C_{\theta^*, \Lambda} + K_0} \right) \left\| \lambda^{(1)} - \lambda^{(2)} \right\| \end{aligned}$$

Moreover, if  $g(\cdot | \theta)$  is  $L$ -Lipschitz, then

$$\|g(\cdot | \theta_1) - g(\cdot | \theta_2)\|_{\infty} \leq L \|\theta_1 - \theta_2\|_2$$

So

$$\|g(\cdot | \theta_1) - g(\cdot | \theta_2)\|_{\infty} \leq L \frac{1}{\lambda_{\min} w J} \left( (K_1 + w) \sqrt{\frac{2}{\lambda_{\min} w} C_{\theta^*, \Lambda} + K_0} \right) \|\lambda^{(2)} - \lambda^{(1)}\|_2$$

## 2 Additive Model

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

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

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

Suppose that the penalties, functions  $g_j(x | \theta^{(j)})$  are twice-differentiable wrt  $\theta$  and for all  $j = 1, \dots, J$

- $\nabla_{\theta^{(j)}}^2 P_j(\theta^{(j)})$  are PSD matrices for all  $j = 1, \dots, J$  (so convex penalties)
- $g_j(x | \theta^{(j)})$  is convex in  $\theta^{(j)}$
- $\nabla_{\theta}^2 \|y - \sum_{j=1}^J g_j(x | \theta^{(j)})\|_T^2$  is a PSD matrix

Suppose there is a constant  $L > 0$  such that for all  $\boldsymbol{\theta}, \boldsymbol{\theta}'$  and all  $j = 1, \dots, J$ , we have

$$\|g_j(\cdot|\boldsymbol{\theta}) - g_j(\cdot|\boldsymbol{\theta}')\|_\infty \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2$$

Let

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

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

$$\|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)})\| \leq \frac{LJ^{3/2} \sqrt{2C_{\boldsymbol{\theta}^*, \Lambda}}}{w\lambda_{min}^2} \|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|$$

and

$$\left\| g(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)})) - g(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)})) \right\|_\infty \leq \frac{L^2 J^2 \sqrt{2C_{\boldsymbol{\theta}^*, \Lambda}}}{w\lambda_{min}^2} \|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|$$

### Proof

For simplicity, we write

$$g(\cdot|\boldsymbol{\theta}) = \sum_{i=1}^J g_i(\cdot|\boldsymbol{\theta}^{(i)})$$

and

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) = \left\{ \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}) \right\}_{j=1}^J$$

**1. Calculate  $\nabla_{\boldsymbol{\lambda}} \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda})$  using the implicit differentiation trick.**

By the KKT conditions, we have for all  $j = 1 : J$

$$\nabla_{\boldsymbol{\theta}^{(j)}} \left( \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right) + \lambda_j w \boldsymbol{\theta}^{(j)} \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} = 0$$

Now we implicitly differentiate with respect to  $\boldsymbol{\lambda}$

$$\nabla_{\boldsymbol{\lambda}} \left\{ \left[ \nabla_{\boldsymbol{\theta}^{(j)}} \left( \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right) + \lambda_j w \boldsymbol{\theta}^{(j)} \right] \Big|_{\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\} = 0$$

By the product rule and chain rule, we have

$$\left\{ \sum_{k=1}^J \left[ \nabla_{\boldsymbol{\theta}^{(k)}} \nabla_{\boldsymbol{\theta}^{(j)}} \left( \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 + 1[k=j]\lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right) + 1[k=j]\lambda_j w I \right] \nabla_{\lambda} \hat{\boldsymbol{\theta}}^{(k)}(\boldsymbol{\lambda}) \right\} + \left\{ \begin{matrix} \vec{0} & \dots & \vec{0} & \nabla_{\boldsymbol{\theta}^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) + w \hat{\boldsymbol{\theta}}^{(j)} & \vec{0} & \dots & \vec{0} \end{matrix} \right\} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} = 0$$

Define the following matrices

$$S : S_{jk} = \nabla_{\boldsymbol{\theta}}^2 \frac{1}{2} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}$$

$$D_1 = \text{diag} \left( \left\{ \nabla_{\boldsymbol{\theta}^{(j)}}^2 \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right\}_{j=1}^J \right) \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})}$$

$$D_2 = \text{diag} (w \lambda_j I^{p_j \times p_j}) \text{ where } \boldsymbol{\theta}^{(j)} \in \mathbb{R}^{p_j}$$

$$M = \left\{ \left[ \begin{matrix} \vec{0} \\ \nabla_{\boldsymbol{\theta}} P_j(\boldsymbol{\theta}^{(j)}) + w \boldsymbol{\theta}^{(j)} \\ \vec{0} \end{matrix} \right] \right\}_{j=1}^J \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \quad (\text{stack side by side})$$

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

$$\left( \nabla_{\lambda} \hat{\boldsymbol{\theta}}_1(\boldsymbol{\lambda}) \quad \nabla_{\lambda} \hat{\boldsymbol{\theta}}_2(\boldsymbol{\lambda}) \quad \dots \quad \nabla_{\lambda} \hat{\boldsymbol{\theta}}_p(\boldsymbol{\lambda}) \right) = -M^{\top} (S + D_1 + D_2)^{-1}$$

$S$  is a PSD matrix since 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 column in  $M$ :**

Rearranging the KKT conditions, we have

$$\begin{aligned} \nabla_{\boldsymbol{\theta}^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) + w \boldsymbol{\theta}^{(j)} \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} &= \frac{1}{2\lambda_j} \nabla_{\boldsymbol{\theta}^{(j)}} \|y - g(\cdot|\boldsymbol{\theta})\|_T^2 \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \\ &= \frac{1}{\lambda_j} \left\langle \nabla_{\boldsymbol{\theta}^{(j)}} g_j(\cdot|\boldsymbol{\theta}^{(j)}), y - g(\cdot|\boldsymbol{\theta}) \right\rangle_T \Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \end{aligned}$$

By the definition of  $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ , we have

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

Hence

$$\left\| y - g(\cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right\|_T \leq \sqrt{2C_{\boldsymbol{\theta}^*, \Lambda}}$$

Hence

$$\begin{aligned}
\left\| \nabla_{\boldsymbol{\theta}^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) + w \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}) \right\|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} &\leq \left\| \frac{1}{\lambda_j} \left\langle \nabla_{\boldsymbol{\theta}^{(j)}} g_j(\cdot | \boldsymbol{\theta}^{(j)}), y - g(\cdot | \boldsymbol{\theta}) \right\rangle \right\|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \\
&\leq \frac{1}{\lambda_{min} n_T} \sum_{i=1}^{n_T} \left\| \nabla_{\boldsymbol{\theta}^{(j)}} g_j(x_i | \boldsymbol{\theta}^{(j)}) \right\|_2 \left| y - g(x_i | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right| \\
&\leq \frac{1}{\lambda_{min} \sqrt{n_T}} \left\| y - g(\cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right\|_T \sqrt{\sum_{i=1}^{n_T} \left\| \nabla_{\boldsymbol{\theta}^{(j)}} g_j(x_i | \boldsymbol{\theta}^{(j)}) \right\|_2^2} \\
&\leq \frac{1}{\lambda_{min} \sqrt{n_T}} \sqrt{2C_{\boldsymbol{\theta}^*, \Lambda}} \sqrt{\sum_{i=1}^{n_T} \left\| \nabla_{\boldsymbol{\theta}^{(j)}} g_j(x_i | \boldsymbol{\theta}^{(j)}) \right\|_2^2}
\end{aligned}$$

In addition, since  $g_j(\cdot | \boldsymbol{\theta}^{(j)})$  is  $L$ -Lipschitz with respect to  $\|\cdot\|_\infty$ , we have that

$$\left\| \nabla_{\boldsymbol{\theta}^{(j)}} g_j(x | \boldsymbol{\theta}^{(j)}) \right\|_2 \leq L \quad \forall x$$

Putting all of this together, we get that for all  $j = 1, \dots, J$

$$\left\| \nabla_{\boldsymbol{\theta}^{(j)}} P_j(\boldsymbol{\theta}^{(j)}) \right\|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} + w \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda}) \leq \frac{L}{\lambda_{min}} \sqrt{2C_{\boldsymbol{\theta}^*, \Lambda}}$$

**3. We bound the norm of  $\nabla_{\lambda_k} \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$  for all  $k = 1, \dots, J$ .**



For every  $i = 1, \dots, p$ , we have

$$\begin{aligned}
\|\nabla_{\lambda} \hat{\theta}_i(\lambda)\| &= \|M^{\top} (S + D_1 + D_2)^{-1} e_k\| \\
&\leq \sum_{j=1}^J \|M_j\|_2 \left\| (S + D_1 + D_2)^{-1} \right\|_2 \\
&= \sum_{j=1}^J \left\| \nabla_{\theta^{(j)}} P_j(\theta^{(j)}) \Big|_{\theta=\hat{\theta}(\lambda)} + w \hat{\theta}^{(j)}(\lambda) \right\|_2 \left\| (S + D_1 + D_2)^{-1} \right\|_2 \\
&\leq J \left( \frac{L}{\lambda_{\min}} \sqrt{2C_{\theta^*, \Lambda}} \right) \left( \frac{1}{w \lambda_{\min}} \right)
\end{aligned}$$

Since the derivative of  $\hat{\theta}_i(\lambda)$  is bounded, then by Lemma 2 below,  $\hat{\theta}(\lambda)$  must be Lipschitz:

$$\left\| \hat{\theta}(\lambda) - \hat{\theta}(\lambda') \right\|_2 \leq \frac{LJ^{3/2} \sqrt{2C_{\theta^*, \Lambda}}}{w \lambda_{\min}^2} \|\lambda - \lambda'\|_2$$

#### 4. Put all the bounds together

Since each  $g_j(\cdot | \theta^{(j)})$  is Lipschitz in  $\theta^{(j)}$ , then

$$\begin{aligned}
\left\| g \left( \cdot | \hat{\theta}(\lambda^{(1)}) \right) - g \left( \cdot | \hat{\theta}(\lambda^{(2)}) \right) \right\|_{\infty} &\leq \sum_{j=1}^J \left\| g_j \left( \cdot | \hat{\theta}^{(j)}(\lambda^{(1)}) \right) - g_j \left( \cdot | \hat{\theta}^{(j)}(\lambda^{(2)}) \right) \right\|_{\infty} \\
&\leq \sum_{j=1}^J L \left\| \hat{\theta}^{(j)}(\lambda^{(1)}) - \hat{\theta}^{(j)}(\lambda^{(2)}) \right\|_2 \\
&\leq L \sqrt{J} \left\| \hat{\theta}(\lambda^{(1)}) - \hat{\theta}(\lambda^{(2)}) \right\|_2 \\
&\leq \frac{L^2 J^2 \sqrt{2C_{\theta^*, \Lambda}}}{w \lambda_{\min}^2} \|\lambda^{(1)} - \lambda^{(2)}\|
\end{aligned}$$

### 3 Nonsmooth Penalties

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

Recall that  $\Lambda \subseteq \mathbb{R}^J$ . Consider the measure space over  $\Lambda$  with respect to the Lebesgue measure  $\mu$ . We suppose that for a given dataset  $(X, y)$ , suppose the following three assumptions hold:

**Assumption (1):** Let the penalized training criterion be denoted  $L_T(\boldsymbol{\theta}, \boldsymbol{\lambda})$ . Denote the differentiable space of  $L_T(\cdot, \boldsymbol{\lambda})$  at any point  $\boldsymbol{\theta}$  as

$$\Omega^{L_T(\cdot, \boldsymbol{\lambda})}(\boldsymbol{\theta}) = \left\{ \boldsymbol{\eta} \mid \lim_{\epsilon \rightarrow 0} \frac{L_T(\boldsymbol{\theta} + \epsilon \boldsymbol{\eta}) - L_T(\boldsymbol{\theta})}{\epsilon} \text{ exists} \right\}$$

Suppose there is a set  $\Lambda_{smooth} \subseteq \Lambda$  such that  $\mu(\Lambda_{smooth}^C) = 0$  and for every  $\boldsymbol{\lambda} \in \Lambda_{smooth}$ , there exists a ball with nonzero radius centered at  $\boldsymbol{\lambda}$ , denoted  $B(\boldsymbol{\lambda})$ , such that the following conditions hold:

**Cond 1:** For all  $\boldsymbol{\lambda}' \in B(\boldsymbol{\lambda})$ , the training criterion  $L_T(\cdot, \cdot)$  is twice differentiable along directions in  $\Omega^{L_T(\cdot, \cdot)}(\hat{\boldsymbol{\theta}}_\lambda)$ . (So technically the twice-differentiable space is constant)

**Cond 2:**  $\Omega^{L_T(\cdot, \boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}_\lambda)$  is a local optimality space of  $B(\boldsymbol{\lambda})$ :

$$\arg \min_{\boldsymbol{\theta} \in \Theta} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}') = \arg \min_{\boldsymbol{\theta} \in \Omega^{L_T(\cdot, \boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}_\lambda)} L_T(\boldsymbol{\theta}, \boldsymbol{\lambda}') \quad \forall \boldsymbol{\lambda}' \in B(\boldsymbol{\lambda})$$

**Cond 3:** (Not necessary if we keep the ridge penalty) There is an orthonormal basis  $U_\lambda$  of  $\Omega^{L_T(\cdot, \boldsymbol{\lambda})}(\hat{\boldsymbol{\theta}}_\lambda)$  such that the Hessian of the training criterion taken along directions  $U_\lambda$  is invertible.

**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 specified in Section 1 and 2 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

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

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

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

Under these assumptions, the same Lipschitz conditions hold for dataset  $(X, y)$  and every  $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$ .

## Proof

Consider any  $\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)} \in \Lambda_{smooth}$ . The length of  $\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})$  covered by set  $A$  can be expressed as

$$\mu_1 \left( A \cap \mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \right)$$

where  $\mu_1$  is the Lebesgue measure over the line segment  $\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})$ . (So if  $A \cap \mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})$  is just a line segment, it is the length  $\|A \cap \mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})\|_2$ )

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

$$\max_{\{\ell^{(i)}\}_{i=1}^{\infty}} \mu_1 \left( \cup_{i=1}^{\infty} B(\ell^{(i)}) \cap \mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \right) = \left\| \mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \right\|_2$$

Let

$$\{\ell_{max}^{(i)}\}_{i=1}^{\infty} = \left\{ \arg \max_{\{\ell^{(i)}\}} \mu_1 \left( \cup_{i=1}^{\infty} B(\ell^{(i)}) \cap \mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \right) \right\} \cup \{\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}\}$$

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

$$P = \cup_{i=1}^{\infty} \text{Bd} B(\ell_{max}^{(i)}) \cap \mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)})$$

Every 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  $\alpha_p$ . By our assumptions, the differentiable space of the training criterion must be constant over the interior of line segment  $\mathcal{L}(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)})$  (so there might be bad behavior at the endpoints). Let the differentiable space over the interior of line segment  $\mathcal{L}(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)})$  be denoted  $\Omega_i$ .

By our assumptions, the differentiable space is also a local optimality space. Let  $U^{(i)}$  be an orthonormal basis of  $\Omega_i$ . For each  $i$ , we can express  $\hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}}$  for all  $\boldsymbol{\lambda} \in \text{Int} \{ \mathcal{L}(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)}) \}$  as

$$\begin{aligned} \hat{\boldsymbol{\theta}}_{\boldsymbol{\lambda}} &= U^{(i)} \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} \\ \hat{\boldsymbol{\beta}}_{\boldsymbol{\lambda}} &= \arg \min_{\boldsymbol{\beta}} L_T(U^{(i)} \boldsymbol{\beta}, \boldsymbol{\lambda}) \end{aligned}$$

Now apply the result in Section 1 or 2 over every line segment  $\mathcal{L}(\boldsymbol{p}^{(i)}, \boldsymbol{p}^{(i+1)})$ . To do this, we must modify the proofs to take directional derivatives along the columns of  $U^{(i)}$ . We can establish that there is a constant  $c > 0$  independent of  $i$  such that for all  $i = 1, 2, \dots$ , we have

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

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

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

### 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 differentiability” 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  $p \in \mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \setminus \left\{ \cup_{i=1}^{\infty} B(\boldsymbol{\ell}_{max}^{(i)}) \right\}$  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

$$\|\nabla_{\theta} P(\theta)\| \leq K_1 \|\theta\|_2 + K_0$$

for constants  $K_0, K_1 > 0$ .

**Ridge:**

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

**Lasso:**

$$\begin{aligned} \|\nabla_{\theta} \|\theta\|_1\| &= \|sgn(\theta)\| \\ &\leq p \end{aligned}$$

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

$$\begin{aligned} \|\nabla_{\theta} \|D\theta\|_1\| &= \|D^T sgn(D\theta)\| \\ &\leq G \|sgn(D\theta)\| \\ &\leq pG \end{aligned}$$

**Group Lasso:**

If we have un-pooled penalty parameters as follows

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

then we have the bound

$$\left\| \nabla_{\theta^{(j)}} \|\theta^{(j)}\|_2 \right\| = \frac{\|\theta^{(j)}\|_2}{\|\theta^{(j)}\|_2} = 1$$

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

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

then we have the bound

$$\begin{aligned}
\left\| \nabla_{\boldsymbol{\theta}} \sum_{j=1}^J \|\boldsymbol{\theta}^{(j)}\|_2 \right\| &= \sqrt{\sum_{j=1}^J \left\| \nabla_{\boldsymbol{\theta}^{(j)}} \|\boldsymbol{\theta}^{(j)}\|_2 \right\|^2} \\
&= \sqrt{\sum_{j=1}^J \left( \frac{\|\boldsymbol{\theta}^{(j)}\|_2}{\|\boldsymbol{\theta}^{(j)}\|_2} \right)^2} \\
&= J
\end{aligned}$$

## 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

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

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

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

where  $N_{T,j}$  is a matrix of the evaluations of the normalized B-spline basis at  $x_j$ .  $P_j(\boldsymbol{\theta}_j)$  is the Sobolev penalty and can be written as  $\boldsymbol{\theta}_j^T V_j \boldsymbol{\theta}_j$  for an appropriate penalty matrix  $V_j$ . We will not need to express anything in terms of  $V_j$  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

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

Let

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

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

$$\left\| \sum_{j=1}^J g_j \left( \cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}) \right) - g_j \left( \cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)}) \right) \right\|_\infty \leq \frac{BJ^3 \sqrt{2C_{\theta^*, \Lambda}}}{w\lambda_{min}^2} \|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|$$

### Proof

To apply the result from Section 2, we just need note that the  $\hat{g}_j(x|\boldsymbol{\theta}) = \sum_{i=1}^B \theta_i N_{j,i}(x)$  is  $\sqrt{B}$ -Lipschitz since  $N_{T,j}$  is a normalized B-spline and

$$\sup_x N_{j,i}(x) = 1$$

Hence for all  $j = 1, \dots, J$

$$\begin{aligned} \|\hat{g}_j(\cdot|\boldsymbol{\theta}) - \hat{g}_j(\cdot|\boldsymbol{\theta}')\|_\infty &= \sup_x \left| \sum_{i=1}^B (\theta_i - \theta'_i) N_{j,i}(x) \right| \\ &= \left| \sum_{i=1}^B |\theta_i - \theta'_i| \right| \\ &\leq \sqrt{B} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|_2 \end{aligned}$$

Apply the result from Section 2 to get the result for all  $j = 1, \dots, J$  that

$$\left\| g_j \left( \cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}) \right) - g_j \left( \cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)}) \right) \right\|_\infty \leq \frac{BJ^2 \sqrt{2C_{\theta^*, \Lambda}}}{w\lambda_{min}^2} \|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\|$$

The additive model then has the following Lipschitz bound

$$\begin{aligned} \left\| \sum_{j=1}^J g_j \left( \cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}) \right) - g_j \left( \cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)}) \right) \right\|_{\infty} &\leq \sum_{j=1}^J \left\| g_j \left( \cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(1)}) \right) - g_j \left( \cdot | \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}^{(2)}) \right) \right\|_{\infty} \\ &\leq \frac{BJ^3 \sqrt{2C_{\theta^*, \Lambda}}}{w\lambda_{min}^2} \|\boldsymbol{\lambda}^{(1)} - \boldsymbol{\lambda}^{(2)}\| \end{aligned}$$

## 5 Appendix

### Lemma lipschitz iff bounded gradient

Suppose  $g$  is convex in  $\boldsymbol{\theta}$ .

$$g(x|\boldsymbol{\theta}) \text{ is } L\text{-Lipschitz} \implies \|\nabla_{\boldsymbol{\theta}} g(x|\boldsymbol{\theta})\|_2 \leq \sqrt{p}L$$

(The other direction can also be proved. <https://homes.cs.washington.edu/~marcotcr/blog/lipschitz/>)

### Proof

Let  $\boldsymbol{\theta}' - \boldsymbol{\theta} = \arg \max_{\boldsymbol{\beta}} \langle \nabla_{\boldsymbol{\theta}} g(x|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}, \boldsymbol{\beta} \rangle = \|\nabla_{\boldsymbol{\theta}} g(x|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}\|_2$ .

Since  $g$  is convex in  $\boldsymbol{\theta}$ , then

$$\begin{aligned} g(x|\boldsymbol{\theta}) - g(x|\boldsymbol{\theta}') &\geq \left\langle \nabla_{\boldsymbol{\theta}} g(x|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}, \boldsymbol{\theta}' - \boldsymbol{\theta} \right\rangle \\ &= \|\nabla_{\boldsymbol{\theta}} g(x|\boldsymbol{\theta})|_{\boldsymbol{\theta}=\boldsymbol{\theta}'}\|_2 \end{aligned}$$

Also, by the Lipschitz assumption,

$$\left| g(x|\boldsymbol{\theta}) - g(x|\boldsymbol{\theta}') \right| \leq L\|\boldsymbol{\theta}' - \boldsymbol{\theta}\|$$

### Lemma 2: Bounded gradient implies lipschitz

Suppose  $\Lambda$  is a convex set. If  $\|\nabla_{\boldsymbol{\lambda}} \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\boldsymbol{\lambda}'}\| \leq B$  at all  $\boldsymbol{\lambda}'$  for all  $i = 1, \dots, J$

Then for all  $\boldsymbol{\lambda} \in \Lambda$ , we have

$$\left\| \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}') \right\| \leq \sqrt{J}B\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|$$



**Proof**

By the mean value theorem, there is some  $\alpha \in (0, 1)$  such that

$$\begin{aligned} \left| \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda}) - \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda}') \right| &= \left| \left\langle \nabla_{\boldsymbol{\lambda}} \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda}) \Big|_{\boldsymbol{\lambda}=\alpha\boldsymbol{\lambda}+(1-\alpha)\boldsymbol{\lambda}'}, \boldsymbol{\lambda} - \boldsymbol{\lambda}' \right\rangle \right| \\ &\leq \max_{\boldsymbol{\lambda} \in \Lambda} \|\nabla_{\boldsymbol{\lambda}} \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})\| \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\| \\ &\leq B \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\| \end{aligned}$$

Hence

$$\left\| \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}) - \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda}') \right\| \leq \sqrt{JB} \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|$$