Proofs for Smoothness of Parametric Regression Models

November 23, 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

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

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

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

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 $\boldsymbol{\theta_1}, \boldsymbol{\theta_2}$

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

Then

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

Document Outline

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

1. Multiple penalties for a single model

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

2. Additive model (no ridge!)

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

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:

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

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

Suppose that the penalties and the function $g(x|\boldsymbol{\theta})$ are twice-differentiable and convex wrt $\boldsymbol{\theta}$:

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

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

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

(If $\left| \nabla_\theta P_j\left(\boldsymbol{\theta}\right)|_{\boldsymbol{\theta}=\boldsymbol{\theta}'} \right| \le K_0$, then we can drop the additional ridge penalty!)

Let

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

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

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

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

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

Proof

1. **We calculate $\nabla_\lambda \hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ using the implicit differentiation trick.**
   By the KKT conditions, we have

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

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

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

   where

$$\nabla_\theta P(\boldsymbol{\theta}) = \left\{ \begin{array}{ccc} \nabla_\theta P_1(\boldsymbol{\theta}) & ... & \nabla_\theta P_J(\boldsymbol{\theta}) \end{array} \right\}$$

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

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

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

We know that

$$
\begin{aligned}
\left\|\nabla_\lambda \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})\right\| &= \left\| e_i^\top \left[\nabla^2_\theta \left(\frac{1}{2}\left\| y - g(\cdot|\boldsymbol{\theta})\right\|_T^2 + \sum_{j=1}^J \lambda_j \left(P_j(\boldsymbol{\theta}) + \frac{w}{2}\|\boldsymbol{\theta}\|^2 \right)\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right]^{-1} \left(\nabla_\theta P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} + w\boldsymbol{\theta}\vec{\mathbf{1}}_J^\top \right) \right\| \\
&= \left\| e_i^\top \left[\nabla^2_\theta \left(\frac{1}{2}\left\| y - g(\cdot|\boldsymbol{\theta})\right\|_T^2 + \sum_{j=1}^J \lambda_j \left(P_j(\boldsymbol{\theta}) + \frac{w}{2}\|\boldsymbol{\theta}\|^2 \right)\right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right]^{-1} \left(\nabla_\theta P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} + w\boldsymbol{\theta}\vec{\mathbf{1}}_J^\top \right) \right\| \\
&\le \left\| \left[\nabla^2_\theta \left(\frac{1}{2}\left\| y - g(\cdot|\boldsymbol{\theta})\right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta}) \right)_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} + \sum_{j=1}^J \lambda_j w I \right]^{-1} \right\| \left(\left\| \nabla_\theta P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\|_F + w \left\| \boldsymbol{\theta}\vec{\mathbf{1}}_J^\top \right\| \right) \\
&\le \left\| \left[\sum_{j=1}^J \lambda_j w I \right]^{-1} \right\| \left(\left\| \nabla_\theta P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \right\|_F + w\sqrt{J}\|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2 \right) \\
&\le \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 $\frac{1}{2}\left\| y - g(\cdot|\boldsymbol{\theta})\right\|_T^2 + \sum_{j=1}^J \lambda_j P_j(\boldsymbol{\theta})$ is convex in $\boldsymbol{\theta}$. The last inequality follows from the assumption $\nabla_\theta P(\boldsymbol{\theta})|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})} \le K_1 \|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})\|_2 + K_0$.

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

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

So

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

Hence for all $\boldsymbol{\lambda} \in \Lambda$

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

4. Put all the bounds together
By the mean value theorem, there is a $\alpha \in (0,1)$ such that

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

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

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

So

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

2 Additive Model

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

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

where $\Lambda = [\lambda_{min}, \lambda_{max}]^J$.
Suppose that the penalties, functions $g_j(x|\boldsymbol{\theta}^{(j)})$ are twice-differentiable wrt $\boldsymbol{\theta}$ and for all $j = 1, ..., J$

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

- There is a $m > 0$ such that the training criterion is m-strongly convex at the minimizer

$$\nabla^2_{\boldsymbol{\theta}} \left(\|y - \sum_{j=1}^{J} g_j(x|\boldsymbol{\theta}^{(j)})\|_T^2 + \sum_{j=1}^{J} \lambda_j P_j(\boldsymbol{\theta}^{(j)}) \right) \Bigg|_{\theta=\hat{\boldsymbol{\theta}}(\lambda)} \succeq mI$$

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

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

Let

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

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

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

and

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

Proof

For simplicity, we write

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

and

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

1. Calculate $\nabla_\lambda \hat{\boldsymbol{\theta}}^{(j)}(\boldsymbol{\lambda})$ using the implicit differentiation trick.
By the KKT conditions, we have for all $j = 1 : J$

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

Now we implicitly differentiate with respect to λ

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

By the product rule and chain rule, we have

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

Define the following matrices

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

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

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

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

$$\left(\begin{array}{cccc} \nabla_\lambda \hat{\boldsymbol{\theta}}_1(\boldsymbol{\lambda}) & \nabla_\lambda \hat{\boldsymbol{\theta}}_2(\boldsymbol{\lambda}) & ... & \nabla_\lambda \hat{\boldsymbol{\theta}}_p(\boldsymbol{\lambda}) \end{array} \right) = -M^\top (S+D)^{-1}$$

2. We bound every column in M:
Rearranging the KKT conditions, we have

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

Hence

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

We bound $\left\| y - g(\cdot|\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})) \right\|_T$. By the definition of $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$, we have

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

To bound $\left\| \nabla_{\theta^{(j)}} g_j(x_i|\boldsymbol{\theta}^{(j)}) \right\|_2^2$, note that since $g_j(\cdot|\boldsymbol{\theta}^{(j)})$ is L-Lipschitz with respect to $\|\cdot\|_\infty$, we have

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

Hence

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

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

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

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

For every $i = 1, ..., J$, we have

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

where we used the fact that $(S+D)^{-1} \preceq m^{-1} I$
Since the derivative of $\hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})$ is bounded, then by Lemma 2 below, $\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})$ must be Lipschitz:

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

4. Put all the bounds together
Since each $g_j(\cdot|\boldsymbol{\theta}^{(j)})$ is Lipschitz in $\boldsymbol{\theta}^{(j)}$, then

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

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 Λ with respect to the Lebesgue measure μ. 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

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

Suppose there is a set $\Lambda_{smooth} \subseteq \Lambda$ such that
Cond 1: For every $\boldsymbol{\lambda} \in \Lambda_{smooth}$, there exists a ball with nonzero radius centered at $\boldsymbol{\lambda}$, denoted $B(\boldsymbol{\lambda})$, such that

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

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

Cond 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 $\mathcal{L}(\boldsymbol{\lambda}^{(1)}, \boldsymbol{\lambda}^{(2)}) \cap \Lambda^C_{smooth}$ is countable.
Assumption modifications: Previously we bounded the derivative of P_j. Now we only need the bound to apply when the directional derivative exists. The condition on the derivative of the penalty is now

$$\|\nabla_{\boldsymbol{\theta}} P_j(\boldsymbol{\theta})\|_2 \leq K_1 \|\boldsymbol{\theta}\|_2 + K_0 \text{ if } \frac{\partial}{\partial m} P_j(\boldsymbol{\theta} + m\boldsymbol{\beta}) \text{ exists}$$

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 μ_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 $\cup_{i=1}^{\infty}\boldsymbol{\ell}^{(i)} \subset \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ such that the union of their "balls of differentiabilities" entirely cover $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$:

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

Let

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

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

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

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

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

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

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

Now apply the result in Section 1 or 2 over every line segment $\mathcal{L}\left(\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}\right)$. 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 $i = 1, 2...$, we have

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

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

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

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

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

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 $P_j(\boldsymbol{\theta}) \equiv 0$ and fix $w = 2$.

Lasso:

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

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

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

Group Lasso:

If we have un-pooled penalty parameters as follows

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

then we have the bound

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

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

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

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

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

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 $P_j(\boldsymbol{\theta}_j)$.

We will suppose that the training loss is $m-$strongly convex around its minimizer.

Let

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

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

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

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 $j = 1, .., J$

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

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

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

The additive model then has the following Lipschitz bound

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

5 Appendix

Lemma lipschitz iff bounded gradient

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

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

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

Proof

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

Since g is convex in θ, then

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

Also, by the Lipschitz assumption,

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

Lemma 2: Bounded gradient implies lipschitz

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

Let

$$\hat{\boldsymbol{\theta}}(\boldsymbol{\lambda})=\begin{pmatrix}\hat{\boldsymbol{\theta}}_1(\boldsymbol{\lambda}) & ... & \hat{\boldsymbol{\theta}}_J(\boldsymbol{\lambda})\end{pmatrix}$$

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

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

Proof

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

$$
\begin{aligned}
\left|\hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda}) - \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda}')\right| &\leq \max_{\lambda \in \Lambda} \|\nabla_{\boldsymbol{\lambda}} \hat{\boldsymbol{\theta}}_i(\boldsymbol{\lambda})\| \|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\| \\
&\leq B\|\boldsymbol{\lambda} - \boldsymbol{\lambda}'\|
\end{aligned}
$$

Hence

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