

Proofs for Smoothness of Parametric Regression Models

October 9, 2016

Intro

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

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 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 $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2$

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

Then

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

Document Outline

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

1. Multiple penalties for a single model

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

2. Additive model

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

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

1 Multiple smooth penalties for a single model

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

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

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 smooth 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 $\nabla_{\boldsymbol{\theta}}^2 g(x|\boldsymbol{\theta})$ are PSD matrices for all x .

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

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

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

Result

Let

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

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

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

Proof

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

Define

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

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

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

By the KKT conditions, we have

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

Now we implicitly differentiate with respect to λ_{ℓ} for $\ell = 1, 2, \dots, J$

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

By the product rule and chain rule, we have

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

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

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

In vector notation, we have

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

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

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

2. Bound $\|\nabla_{\lambda} \hat{m}_{\beta}(\lambda)\|$

Bounding the first multiplicand:

The first multiplicand is bounded by

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

since the mean squared error and the penalty functions are convex.

Bounding the second multiplicand:

The first summand in the second multiplicand is bounded by assumption

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

The second summand in the second multiplicand is bounded by

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

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

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

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

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

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

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

Combining the above three inequalities, we get

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

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

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

Therefore

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

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

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

We can combine (4) with the fact that

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

to get

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

Plug the inequality above into (1) to get

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

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

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

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

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

3. Apply Mean Value Theorem

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

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

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

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

2 Additive Model

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

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

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

Suppose that the penalties and the function $g_j(x|\boldsymbol{\theta}_j)$ is convex wrt $\boldsymbol{\theta}_j$: $\nabla_{\boldsymbol{\theta}_j}^2 P_j(\boldsymbol{\theta}_j)$ for all $j = 1, \dots, J$ and $\nabla_{\boldsymbol{\theta}_j}^2 g_j(x|\boldsymbol{\theta}_j)$ are PSD matrices. Suppose there is some constant $K > 0$ such that for all $j = 1, \dots, J$ and all $\boldsymbol{\beta}, \boldsymbol{\theta}$,

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

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

Let

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

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

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

Proof

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

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 $\nabla_{\boldsymbol{\lambda}} \hat{\boldsymbol{m}}_k(\boldsymbol{\lambda})$ using the implicit differentiation trick.

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

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

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

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

By the product rule and chain rule, we have

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

Define the following matrices

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

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

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

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

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

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

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

S is a PSD matrix since the sum of convex functions is convex (so sum of g_j is convex) and the composition of convex functions is convex (so the composition of the mean squared error and the sum of g_j is convex).

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

2. We bound every diagonal element in D_3 :

By assumption, we know for every $k = 1, \dots, J$

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

Also,

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

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

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

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

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

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 for all $k = 1, \dots, J$

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

Rearranging, we get

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

Therefore

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

Let

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

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

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

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

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

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

Now let

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

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

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

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

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

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

$$\max_k \|\theta_{\lambda^{(1)},k} - \theta_{\lambda^{(2)},k}\| = \|\beta_{\ell_{max}}\| \leq \left\| \lambda^{(2)} - \lambda^{(1)} \right\| \left\| K + w \sqrt{\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right) \frac{2C}{\lambda_{min}w}} \right\| \lambda_{min}^{-1} w^{-1}$$

3 Nonsmooth Penalties

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

Assumption modification (1): Suppose there is some $K > 0$ such that for all $j = 1, \dots, J$ and any θ, β such that

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

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

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

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

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

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

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

Suppose that

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

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

Proof

Now define

$$L_{nonsmooth} = \{\text{line that passes through } \lambda_1, \lambda_2 : \lambda_1, \lambda_2 \in \Lambda_{smooth}^C\}$$

Unproven Claim: Since $\mu(\Lambda_{smooth}^C) = 0$, then $\mu(L_{nonsmooth}) = 0$. I don’t know how to prove this claim, but it seems true. Now denote the line segment between $\lambda^{(1)}, \lambda^{(2)}$ as

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

The set

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

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

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

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

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

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

Proof of Claim:

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

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

So

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

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

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

So

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

There are two cases:

- (1) $\mathcal{L}_{uncovered} \subseteq \Lambda_{smooth}^C$. Then $\|\mathcal{L}(\lambda^{(1)}, \lambda^{(2)}) \cap \Lambda_{smooth}^C\| \geq \|\mathcal{L}(\lambda^{(1)}, \lambda^{(2)}) \cap \mathcal{L}_{uncovered}\| > 0$. This is clearly impossible since $(\lambda^{(1)}, \lambda^{(2)}) \in H^C$.
- (2) There exists a point $\mathbf{p} \in \mathcal{L}_{uncovered} \setminus \Lambda_{smooth}^C$. Since $\mathbf{p} \in \Lambda_{smooth}$, then by Condition 2, then the neighborhood $B(\mathbf{p})$ is non-empty.

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

This implies that

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

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

End of Proof

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

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

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

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

Since every point can be expressed as $\alpha_{p^{(i)}} \lambda^{(1)} + (1 - \alpha_{p^{(i)}}) \lambda^{(2)}$ for some $\alpha_{p^{(i)}} \in [0, 1]$, we can order these points $\{\mathbf{p}^{(i)}\}$ by increasing $\alpha_{p^{(i)}}$. By definition of P and the Claim, the differentiable space of the training criterion over $(\mathbf{p}^{(i)}, \mathbf{p}^{(i+1)})$ must be constant.

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

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

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

where the training criterion is smooth over $(\mathbf{p}^{(i)}, \mathbf{p}^{(i+1)})$ 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

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

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

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

where c is some constant.

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

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

4 Example penalties that satisfy the conditions

Ridge:

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

Lasso:

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

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

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

Group Lasso:

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