Lemma: parametric proof, smooth penalties

Suppose we observe training samples from the model

$$y = g(x|\boldsymbol{\theta}^*) + \epsilon$$

Suppose we are fitting parametric functions $g(\cdot|\boldsymbol{\theta})$ where $\boldsymbol{\theta} \in \mathbb{R}^p$.
The function class of interest are the minimizers of the penalized least squares criterion:

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

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

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

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

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

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

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

Moreover, 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

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

Proof

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

Define

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

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

By the KKT conditions, we have

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

Now we implicitly differentiate with respect to λ_ℓ for $\ell = 1, 2, ..., J$ (assuming everything is smooth)

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

By the product rule and chain rule, we have

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

Rearranging, we get

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

The first multiplicand is bounded by

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

since the squared loss is convex and the penalties are convex.
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

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

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

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

Let

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

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

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

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

Plugging in the inequality above, we get

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

After rearranging, we have

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

Therefore

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

That is,

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

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

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

Rearranging, we get

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

Lemma: Parametric Regression witih Nonsmooth Penalties

Suppose that training criterion satisfies Conditions 1, 2, and 3 from the Hillclimbing paper. Summarizing the conditions, we are supposing that for almost every λ,

Cond 1: The differentiable space of the training criterion at $\boldsymbol{\lambda}$, denoted

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

is a local optimality space.

Cond 2: The training criterion is twice-differentiable along directions spanned by the differentiable space.

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

As in the lemma above, we have the same assumptions regarding the convexity of the penalties and the function and the spectrum bound on the derivative of the penalties.

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

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

Moreover, 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

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

Proof

Let K be the set of knots

$$K = \left\{\boldsymbol{\lambda} : \nabla_\theta P_j(\theta)|_{\theta=\hat{\theta}(\boldsymbol{\lambda})} \text{ does not exist for some } j \in 1:J\right\}$$

By assumption, $\mu(K) = 0$. Let K^{ext} be the smallest set of $(J-1)$-dimensional planes that contain K. K^{ext} must satisfy $\mu(K^{ext}) = 0$. (Otherwise if $\mu(K^{ext}) > 0$, then K should also have $\mu(K) > 0$. A more rigorous argument might be needed here.)

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

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

Claim 1: The set

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

has measure $\mu(H) = 0$.

Proof of Claim 1: To see this, note that for every $(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \in H$, we must have that $\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}} \in K^{ext}$. Therefore $H \subseteq K^{ext} \times K^{ext}$. Since $\mu(K^{ext} \times K^{ext}) = 0$, then $\mu(H) = 0$.

Claim 2: For a given set of points $\{\ell^{(i)}\} \subset \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$, let $d(\{\ell^{(i)}\})$ denote the uncovered distance of $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ by the union of their differentiable spaces:

$$d(\{\ell^{(i)}\}) = \left\|\left[\cup_i \Omega^{L_T(\cdot, \ell^{(i)})}\left(\hat{\theta}(\ell^{(i)})\right)\right] \cap \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})\right\| - \|\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})\|$$

Define points $\left\{\ell^{(i)}_{min}\right\}$ as the minimizer of the uncovered distance:

$$\left\{\ell^{(i)}_{min}\right\} = \arg\min_{\{\ell^{(i)}\}} d(\{\ell^{(i)}\})$$

We claim that for all $(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \in H^C$, the $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ will be covered (almost everywhere) by the union of the differentiable spaces of $\left\{\ell^{(i)}_{min}\right\}$

$$d\left(\left\{\ell^{(i)}_{min}\right\}\right) = 0$$

Proof of Claim 2:
Suppose

$$d\left(\left\{\ell^{(i)}_{min}\right\}\right) > 0$$

Consider the points not covered by the differentiable spaces:

$$U = \left\{\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) - \left(\left[\cup_i \Omega^{L_T(\cdot, \ell^{(i)}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}))}\left(\hat{\theta}(\ell^{(i)}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}))\right)\right] \cap \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})\right)\right\}$$

If every $U \subseteq K$, then $\left\|\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \cap K\right\| \geq \left\|\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \cap U\right\| > 0$. This is clearly impossible since $(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}}) \in H^C$. Therefore there must be a point $\boldsymbol{p} \in U \backslash K$. Then we note that

$$\left\|\Omega^{L_T(\cdot, \boldsymbol{p})}\left(\hat{\boldsymbol{\theta}}(\boldsymbol{p})\right) \cap \mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})\right\| > 0$$

This implies that

$$d(\{\ell^{(i)}_{min}\}) > d(\{\ell^{(i)}_{min}\} \cup \{p\})$$

However contradicts the definition of $\left\{\ell_{min}^{(i)}\right\}$.

Combining Claim 1 and 2, we conclude that for almost every $(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$, there is a set of points $\left\{\ell_{min}^{(i)}\right\}$ such that their differentiable spaces cover $\mathcal{L}(\boldsymbol{\lambda^{(1)}}, \boldsymbol{\lambda^{(2)}})$ almost everywhere.

Hence we can choose a (potentially infinite) set of points $\left\{\boldsymbol{p^{(i)}}\right\}$ such that the differentiable space of the training criterion over the interval $[\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}]$ is constant (basically choose the points at the edge of the differentiable spaces). Within each interval $[\boldsymbol{p^{(i)}}, \boldsymbol{p^{(i+1)}}]$, we can apply the smooth parametric lemma above since the training criterion is smooth with respect to the differentiable space and the local optimality space is equal to the differentiable space. So for every i, we have

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

By the triangle inequality,

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

Example parametric penalties

Ridge, assuming $\sup_{\theta \in \mathcal{G}(T)} \|\theta\|_2 \leq G$:

$$\begin{aligned}
\frac{\partial}{\partial m}\|\theta + m\beta\|_2^2 &= \langle \theta + m\beta, \beta \rangle \\
&\leq G\|\beta\|_2
\end{aligned}$$

Lasso:

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

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

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

Group Lasso:

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