

# Proofs for Smoothness of Non-Parametric Regression Models

October 9, 2016

## Intro

In this document, we consider nonparametric regression models  $g$  from function class  $\mathcal{G}$ . Throughout, we will suppose that the projection of the true model into the model space  $\mathcal{G}$  is  $g^*$ .

We are interested in establishing inequalities of the form

$$\|\hat{g}(\cdot|\boldsymbol{\lambda}^{(2)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})\|_D \leq C\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2$$

## Document Outline

Let  $D$  be some set of observed covariates (it could be the training and validation sets combined or just the validation set).

We prove smoothness for two nonparametric regression examples:

1. Additive model

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

2. Multiple penalties for a single model

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

- (a) This regression problem is complicated and we may want to just leave it out. Depending on the situation, we get smoothness of the form

$$\|\hat{g}(\cdot|\boldsymbol{\lambda}^{(2)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})\|_D^2 \leq C\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2$$

or

$$\|\hat{g}(\cdot|\boldsymbol{\lambda}^{(2)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})\|_D \leq C\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|_2$$

## 1 Additive Model

Consider the problem

$$\mathcal{G}(T) = \left\{ \hat{g}(\cdot|\boldsymbol{\lambda}) = \arg \min_{g \in \mathcal{G}} \frac{1}{2} \left\| y - \sum_{j=1}^J g_j \right\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j(g_j) + \frac{w}{2} \|g_j\|_D^2 \right) \right\}$$

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

For all  $j = 1, \dots, J$ , suppose the penalty functions  $P_j$  are convex and twice-differentiable: For any functions  $g, h$ , the following second-derivative exists and the inequality holds:

$$\frac{\partial^2}{\partial m^2} P_j(g + mh) \geq 0 \forall j = 1, \dots, J$$

Let

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

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

$$\|\hat{g}_j(\cdot|\boldsymbol{\lambda}^{(2)}) - \hat{g}_j(\cdot|\boldsymbol{\lambda}^{(1)})\|_D \leq \left\| \boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)} \right\| \left( \frac{1}{\lambda_{min}} \sqrt{\frac{n_D}{n_T}} + 2\sqrt{\frac{w}{\lambda_{min}}} \right) \sqrt{2C \left( 1 + \frac{J\lambda_{max}}{\lambda_{min}} \right)} \lambda_{min}^{-1} w^{-1}$$

### Proof

For every  $j = 1, \dots, J$ , let  $h_j = \hat{g}_j(\cdot|\boldsymbol{\lambda}^{(2)}) - \hat{g}_j(\cdot|\boldsymbol{\lambda}^{(1)})$ . For notational convenient, let  $\hat{g}_{1,j}(\cdot) = \hat{g}_j(\cdot|\boldsymbol{\lambda}^{(1)})$ .

Let the set of additive components with nonzero differences be denoted

$$H_{nonzero} = \{j : \|h_j\|_D > 0\}$$

We consider the optimization problem restricted to the set of non-zero differences

$$\hat{\mathbf{m}}(\boldsymbol{\lambda}) = \{\hat{m}_j(\boldsymbol{\lambda})\}_{j \in H_{\text{nonzero}}} = \arg \min_{m_j: j \in H_{\text{nonzero}}} \frac{1}{2} \|y - \sum_{j=1}^J (\hat{g}_{1,j} + m_j h_j)\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j(\hat{g}_{1,j} + m_j h_j) + \frac{w}{2} \|\hat{g}_{1,j} + m_j h_j\|_D^2 \right)$$

### 1. Calculate $\nabla_{\lambda} \hat{m}_j(\lambda)$

By the gradient optimality conditions, the gradient of the objective with respect to  $m_\ell$  for all  $\ell \in H_{\text{nonzero}}$

$$\begin{aligned} & \frac{\partial}{\partial m_\ell} \left[ \frac{1}{2} \|y - \sum_{j=1}^J (\hat{g}_{1,j} + m_j h_j)\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j(\hat{g}_{1,j} + m_j h_j) + \frac{w}{2} \|\hat{g}_{1,j} + m_j h_j\|_D^2 \right) \right]_{m=\hat{\mathbf{m}}(\lambda)} \\ &= \left\langle y - \sum_{j=1}^J (\hat{g}_{1,j} + m_j h_j), h_\ell \right\rangle_T + \lambda_\ell \frac{\partial}{\partial m_\ell} P_\ell(\hat{g}_{1,\ell} + m_\ell h_\ell) + \lambda_\ell w \langle h_\ell, \hat{g}_{1,\ell} + m_\ell h_\ell \rangle_D \Big|_{m=\hat{\mathbf{m}}(\lambda)} \\ &= 0 \end{aligned}$$

Now we implicitly differentiate with respect to  $\lambda_k$  for all  $k \in H_{\text{nonzero}}$  to get

$$\begin{aligned} & \frac{\partial}{\partial \lambda_k} \left[ \left\langle y - \sum_{j=1}^J (\hat{g}_{1,j} + m_j h_j), h_\ell \right\rangle_T + \lambda_\ell \frac{\partial}{\partial m_\ell} P_\ell(\hat{g}_{1,\ell} + m_\ell h_\ell) + \lambda_\ell w \langle h_\ell, \hat{g}_{1,\ell} + m_\ell h_\ell \rangle_D \right]_{m=\hat{\mathbf{m}}(\lambda)} \\ &= \sum_{j=1}^J \left[ \langle h_j, h_\ell \rangle_T + 1[\ell = j] \left( \lambda_\ell \frac{\partial^2}{\partial m_\ell^2} P_\ell(\hat{g}_{1,\ell} + m_\ell h_\ell) + \lambda_\ell w \|h_\ell\|_D^2 \right) \right] \frac{\partial \hat{m}_j(\lambda)}{\partial \lambda_k} + 1[\ell = k] \left( \frac{\partial}{\partial m_\ell} P_\ell(\hat{g}_{1,\ell} + m_\ell h_\ell) + w \langle h_\ell, \hat{g}_{1,\ell} + m_\ell h_\ell \rangle_D \right) \Big|_{m=\hat{\mathbf{m}}(\lambda)} \\ &= 0 \end{aligned}$$

Define the following square matrices

$$\begin{aligned} S &: S_{ij} = \langle h_j, h_\ell \rangle_T \forall \ell, j \in H_{\text{nonzero}} \\ D_1 &= \text{diag} \left( \lambda_\ell \frac{\partial^2}{\partial m_\ell^2} P_\ell(\hat{g}_{1,\ell} + m_\ell h_\ell) \Big|_{m=\hat{\mathbf{m}}(\lambda)} \quad \forall \ell \in H_{\text{nonzero}} \right) \\ D_2 &= \text{diag} (\lambda_\ell w \|h_\ell\|_D^2 \forall \ell \in H_{\text{nonzero}}) \\ D_3 &= \text{diag} \left( \frac{\partial}{\partial m_\ell} P_\ell(\hat{g}_{1,\ell} + m_\ell h_\ell) + w \langle h_\ell, \hat{g}_{1,\ell} + m_\ell h_\ell \rangle_D \Big|_{m=\hat{\mathbf{m}}(\lambda)} \quad \forall \ell \in H_{\text{nonzero}} \right) \end{aligned}$$

$M$  : column  $M_j = \nabla_{\lambda} \hat{m}_j(\lambda) \forall j \in H_{nonzero}$

From the implicit differentiation equations, we have the following system of equations:

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

## 2. We bound every diagonal element in $D_3$ :

We first bound  $\left| \frac{\partial}{\partial m_k} P_k(\hat{g}_{1,k} + m_k h_k) \right|$  for all  $k \in H_{nonzero}$ .

Note that from the gradient optimality conditions, we have that

$$\begin{aligned} \left| \frac{\partial}{\partial m_k} P_k(\hat{g}_{1,k} + m_k h_k) \right|_{m=\hat{m}(\lambda)} &= \left| \frac{1}{\lambda_k} \left\langle y - \sum_{j=1}^J (\hat{g}_{1,j} + \hat{m}_j(\lambda) h_j), h_k \right\rangle_T + w \langle h_k, \hat{g}_{1,k} + \hat{m}_k(\lambda) h_k \rangle_D \right| \\ &\leq \frac{1}{\lambda_{min}} \left\| y - \sum_{j=1}^J (\hat{g}_{1,j} + \hat{m}_j(\lambda) h_j) \right\|_T \|h_k\|_T + w \|h_k\|_D \|\hat{g}_{1,k} + \hat{m}_k(\lambda) h_k\|_D \\ &\leq \left( \frac{1}{\lambda_{min}} \sqrt{\frac{n_D}{n_T}} \left\| y - \sum_{j=1}^J (\hat{g}_{1,j} + \hat{m}_j(\lambda) h_j) \right\|_T + w \|\hat{g}_{1,k} + \hat{m}_k(\lambda) h_k\|_D \right) \|h_k\|_D \end{aligned}$$

where the last line uses the fact that

$$n_T \|h_k\|_T^2 \leq n_D \|h_k\|_D^2 \implies \|h_k\|_T \leq \sqrt{\frac{n_D}{n_T}} \|h_k\|_D$$

We can bound  $\left\| y - \sum_{j=1}^J (\hat{g}_{1,j} + \hat{m}_j(\boldsymbol{\lambda})h_j) \right\|_T$  using the basic inequality

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

To bound  $\max_{j=1:J} \left( P_j(\hat{g}_{1,j}) + \frac{w}{2} \|\hat{g}_{1,j}\|_D^2 \right)$ , we also use the basic inequality

$$\begin{aligned}
\lambda_{min} \max_{j=1:J} \left( P_j(\hat{g}_{1,j}) + \frac{w}{2} \|\hat{g}_{1,j}\|_D^2 \right) &\leq \frac{1}{2} \left\| y - \sum_{j=1}^J \hat{g}_{1,j} \right\|_T^2 + \sum_{j=1}^J \lambda_j^{(1)} \left( P_j(\hat{g}_{1,j}) + \frac{w}{2} \|\hat{g}_{1,j}\|_D^2 \right) \\
&\leq C
\end{aligned}$$

Putting the two above inequalities together, we get

$$\frac{1}{2} \left\| y - \sum_{j=1}^J (\hat{g}_{1,j} + \hat{m}_j(\boldsymbol{\lambda})h_j) \right\|_T^2 \leq C \left( 1 + \frac{J\lambda_{max}}{\lambda_{min}} \right) \implies \left\| y - \sum_{j=1}^J (\hat{g}_{1,j} + \hat{m}_j(\boldsymbol{\lambda})h_j) \right\|_T \leq \sqrt{2C \left( 1 + \frac{J\lambda_{max}}{\lambda_{min}} \right)}$$

and

$$\lambda_{min} \frac{w}{2} \|\hat{g}_{1,k} + \hat{m}_k(\boldsymbol{\lambda})h_k\|_D^2 \leq C \left( 1 + \frac{J\lambda_{max}}{\lambda_{min}} \right) \implies \|\hat{g}_{1,k} + \hat{m}_k(\boldsymbol{\lambda})h_k\|_D \leq \sqrt{\frac{2C}{\lambda_{min}w} \left( 1 + \frac{J\lambda_{max}}{\lambda_{min}} \right)}$$

So

$$\left| \frac{\partial}{\partial m_k} P_k(\hat{g}_{1,k} + m_k h_k) \right|_{m=\hat{m}(\boldsymbol{\lambda})} \leq \left( \frac{1}{\lambda_{min}} \sqrt{\frac{n_D}{n_T}} + \sqrt{\frac{w}{\lambda_{min}}} \right) \sqrt{2C \left( 1 + \frac{J\lambda_{max}}{\lambda_{min}} \right)} \|h_k\|_D$$

Next we bound  $|w\langle h_k, g_k + \hat{m}_k(\boldsymbol{\lambda})h_k \rangle_D|$  for all  $k \in H_{nonzero}$ . By Cauchy Schwarz

$$\begin{aligned} |w\langle h_k, g_k + \hat{m}_k(\boldsymbol{\lambda})h_k \rangle_D| &\leq w\|h_k\|_D\|g_k + \hat{m}_k(\boldsymbol{\lambda})h_k\|_D \\ &\leq w\|h_k\|_D\sqrt{\frac{2C}{\lambda_{min}w}\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)} \end{aligned}$$

Define the matrix  $D_{3,upper}$  which bounds the diagonal elements of  $D_3$

$$D_{3,upper} = \left(\frac{1}{\lambda_{min}}\sqrt{\frac{n_D}{n_T}} + 2\sqrt{\frac{w}{\lambda_{min}}}\right)\sqrt{2C\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)}diag(\|h_k\|_D)$$

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

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

Hence

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

Now let

$$\ell_{max} = \arg \max_{\ell} \|h_{\ell}\|_D$$

Then for  $k = \ell_{max}$  in the inequality above, we get

$$\begin{aligned} \|\nabla_{\lambda}\hat{m}_{\ell_{max}}(\lambda)\| &\leq \left(\frac{1}{\lambda_{min}}\sqrt{\frac{n_D}{n_T}} + 2\sqrt{\frac{w}{\lambda_{min}}}\right)\sqrt{2C\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)}\|h_{\ell_{max}}\|_D\|D_2^{-1}e_k\| \\ &= \left(\frac{1}{\lambda_{min}}\sqrt{\frac{n_D}{n_T}} + 2\sqrt{\frac{w}{\lambda_{min}}}\right)\sqrt{2C\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)}\|h_{\ell_{max}}\|_D\lambda_{\ell_{max}}^{-1}w^{-1}\|h_{\ell_{max}}\|_D^{-2} \\ &\leq \left(\frac{1}{\lambda_{min}}\sqrt{\frac{n_D}{n_T}} + 2\sqrt{\frac{w}{\lambda_{min}}}\right)\sqrt{2C\left(1 + \frac{J\lambda_{max}}{\lambda_{min}}\right)}\lambda_{min}^{-1}w^{-1}\|h_{\ell_{max}}\|_D^{-1} \end{aligned}$$

#### 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

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

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

$$\max_j \|\hat{g}_j(\cdot|\lambda^{(2)}) - \hat{g}_j(\cdot|\lambda^{(1)})\|_D = \|h_{\ell_{max}}\|_D \leq \left\| \lambda^{(2)} - \lambda^{(1)} \right\| \left( \frac{1}{\lambda_{min}} \sqrt{\frac{n_D}{n_T}} + 2\sqrt{\frac{w}{\lambda_{min}}} \right) \sqrt{2C \left( 1 + \frac{J\lambda_{max}}{\lambda_{min}} \right) \lambda_{min}^{-1} w^{-1}}$$

## 2 Multiple smooth penalties for a single model

Consider the problem

$$\mathcal{G}(T) = \left\{ \hat{g}(\cdot|\lambda) = \arg \min_{g \in \mathcal{G}} \frac{1}{2} \|y - g\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j^{v_j}(g) + \frac{w}{2} \|g\|_D^2 \right) \right\}$$

where  $\Lambda = [\lambda_{min}, \lambda_{max}]^J$  and  $v_j > 1$  for all  $j = 1, \dots, J$ .

For all  $j = 1, \dots, J$ , suppose the penalty functions  $P_j$  are convex and twice-differentiable: For any functions  $g, h$ , the following second-derivative exists and the inequality holds:

$$\frac{\partial^2}{\partial m^2} P_j(g + mh) \geq 0 \forall j = 1, \dots, J$$

Also, suppose that the penalty functions  $P_j$  are semi-norms: for all functions  $a, b$ , the triangle inequality is satisfied

$$P_j(a) + P_j(b) \geq P_j(a + b)$$

For  $\lambda^{(1)}, \lambda^{(2)} \in \Lambda$  where  $\|\lambda^{(1)} - \lambda^{(2)}\|$  is sufficiently small, we have

$$\|\hat{g}(\cdot|\lambda^{(2)}) - \hat{g}(\cdot|\lambda^{(1)})\|_D^2 \leq \|\lambda^{(2)} - \lambda^{(1)}\| \left( w\sqrt{J}\lambda_{min} \right)^{-1} C_0$$

where  $C_0$  is a constant.

### Proof

Let  $h = \hat{g}(\cdot|\boldsymbol{\lambda}^{(2)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})$ . For notational convenient, let  $\hat{g}_1(\cdot) = \hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})$ . Suppose  $\|h\|_D > 0$ .

We consider the optimization problem restricted to the set of non-zero differences

$$\hat{m}(\boldsymbol{\lambda}) = \arg \min_m \frac{1}{2} \|y - (\hat{g}_1 + mh)\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j^{v_j}(\hat{g}_1 + mh) + \frac{w}{2} \|\hat{g}_1 + mh\|_D^2 \right)$$

#### 1. Calculate $\frac{\partial \hat{m}(\lambda)}{\partial \lambda}$

By the gradient optimality conditions, we have that

$$\left. \langle y - (\hat{g}_1 + mh), h \rangle_T + \sum_{j=1}^J \lambda_j \left( \frac{\partial}{\partial m} P_j^{v_j}(\hat{g}_1 + mh) + w \langle h, \hat{g}_1 + mh \rangle_D \right) \right|_{m=\hat{m}(\lambda)} = 0$$

Now we implicitly differentiate with respect to  $\lambda_k$  to get

$$\begin{aligned} & \frac{\partial}{\partial \lambda_k} \left[ \langle y - (\hat{g}_1 + mh), h \rangle_T + \sum_{j=1}^J \lambda_j \left( \frac{\partial}{\partial m} P_j^{v_j}(\hat{g}_1 + mh) + w \langle h, \hat{g}_1 + mh \rangle_D \right) \right] \Big|_{m=\hat{m}(\lambda)} \\ &= \left[ \|h\|_T^2 + \sum_{j=1}^J \lambda_j \left( \frac{\partial^2}{\partial m^2} P_j^{v_j}(\hat{g}_1 + mh) + w \|h\|_D^2 \right) \right] \Big|_{m=\hat{m}(\lambda)} \frac{\partial \hat{m}(\lambda)}{\partial \lambda_k} + \left( \frac{\partial}{\partial m} P_k^{v_k}(\hat{g}_1 + mh) + w \langle h, \hat{g}_1 + mh \rangle_D \right) \Big|_{m=\hat{m}(\lambda)} \\ &= 0 \end{aligned}$$

So

$$\frac{\partial \hat{m}(\lambda)}{\partial \lambda_k} = - \left[ \|h\|_T^2 + \sum_{j=1}^J \lambda_j \left( \frac{\partial^2}{\partial m^2} P_j^{v_j}(\hat{g}_1 + mh) + w \|h\|_D^2 \right) \right]^{-1} \left( \frac{\partial}{\partial m} P_k^{v_k}(\hat{g}_1 + mh) + w \langle h, \hat{g}_1 + mh \rangle_D \right) \Big|_{m=\hat{m}(\lambda)}$$

#### 2. Bound $\frac{\partial \hat{m}(\lambda)}{\partial \lambda_k}$

The first multiplicand is bounded by

$$\left| \|h\|_T^2 + \sum_{j=1}^J \lambda_j \left( \frac{\partial^2}{\partial m^2} P_j^{v_j}(\hat{g}_1 + mh) + w \|h\|_D^2 \right) \right|^{-1} \leq (wJ\lambda_{\min} \|h\|_D^2)^{-1}$$



since the penalty functions are convex.

By Lemma Semi-norm derivatives (Appendix), we have that since  $P_k$  is a semi-norm, then

$$\begin{aligned} \left| \frac{\partial}{\partial m} P_j(\hat{g}_1 + mh) \right| &\leq P_j(h) \\ &= P_j\left(\hat{g}(\cdot|\boldsymbol{\lambda}^{(2)}) - \hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})\right) \\ &\leq P_j\left(\hat{g}(\cdot|\boldsymbol{\lambda}^{(2)})\right) + P_j\left(\hat{g}(\cdot|\boldsymbol{\lambda}^{(1)})\right) \end{aligned}$$

By the basic inequality, we know that

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

where

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

Therefore

$$\left| \frac{\partial}{\partial m} P_j(\hat{g}_1 + mh) \right| \leq 2C/\lambda_{\min}$$

Also by the definition of  $\hat{m}(\boldsymbol{\lambda})$ ,

$$\begin{aligned} \lambda_{\min} \left( P_k^{v_k}(\hat{g}_1 + \hat{m}(\boldsymbol{\lambda})h) + \frac{w}{2} \|\hat{g} + \hat{m}(\boldsymbol{\lambda})h\|_D^2 \right) &\leq \frac{1}{2} \|y - \hat{g}_1\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j^{v_j}(\hat{g}_1) + \frac{w}{2} \|\hat{g}_1\|_D^2 \right) \\ &= \frac{1}{2} \|y - \hat{g}_1\|_T^2 + \sum_{j=1}^J \lambda_j^{(1)} \left( P_j^{v_j}(\hat{g}_1) + \frac{w}{2} \|\hat{g}_1\|_D^2 \right) + \sum_{j=1}^J \left( \lambda_j - \lambda_j^{(1)} \right) \left( P_j^{v_j}(\hat{g}_1) + \frac{w}{2} \|\hat{g}_1\|_D^2 \right) \\ &\leq \frac{1}{2} \|y - g^*\|_T^2 + \sum_{j=1}^J \lambda_j^{(1)} \left( P_j^{v_j}(g^*) + \frac{w}{2} \|g^*\|_D^2 \right) + J\lambda_{\max} \max_j \left( P_j^{v_j}(\hat{g}_1) + \frac{w}{2} \|\hat{g}_1\|_D^2 \right) \\ &\leq C + J\lambda_{\max} \max_j \left( P_j^{v_j}(\hat{g}_1) + \frac{w}{2} \|\hat{g}_1\|_D^2 \right) \end{aligned}$$

And by the definition of  $\hat{g}_1$ ,

$$\lambda_{\min} \max_j \left( P_j^{v_j}(\hat{g}_1) + \frac{w}{2} \|\hat{g}_1\|_D^2 \right) \leq \frac{1}{2} \|y - g^*\|_T^2 + \sum_{j=1}^J \lambda_j^{(1)} \left( P_j^{v_j}(g^*) + \frac{w}{2} \|g^*\|_D^2 \right) \leq C$$

Therefore

$$\lambda_{\min} P_k^{v_k}(\hat{g}_1 + \hat{m}(\boldsymbol{\lambda})h) \leq C \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \implies P_k^{v_k-1}(\hat{g}_1 + \hat{m}(\boldsymbol{\lambda})h) \leq \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k}$$

and

$$\lambda_{\min} \frac{w}{2} \|\hat{g} + \hat{m}(\boldsymbol{\lambda})h\|_D^2 \leq C \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \implies \|\hat{g} + \hat{m}(\boldsymbol{\lambda})h\|_D \leq \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)}$$

Hence

$$\begin{aligned} \left| \frac{\partial}{\partial m} P_k^{v_k}(\hat{g}_1 + mh) + w \langle h, \hat{g} + mh \rangle_D \right| &\leq \left| \frac{\partial}{\partial m} P_k^{v_k}(\hat{g}_1 + mh) \right| + w \|h\|_D \|\hat{g} + mh\|_D \\ &\leq \frac{2Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k} + w \|h\|_D \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)} \end{aligned}$$

Therefore

$$\left| \frac{\partial \hat{m}(\boldsymbol{\lambda})}{\partial \lambda_k} \right| \leq (wJ\lambda_{\min} \|h\|_D^2)^{-1} \left[ \frac{2Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k} + w \|h\|_D \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)} \right]$$

Therefore

$$\begin{aligned} \|\nabla_{\boldsymbol{\lambda}} \hat{m}(\boldsymbol{\lambda})\| &\leq \sqrt{J} \left[ (wJ\lambda_{\min} \|h\|_D^2)^{-1} \left[ \frac{2Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k} + w \|h\|_D \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)} \right] \right] \\ &= (w\sqrt{J}\lambda_{\min} \|h\|_D^2)^{-1} \left[ \frac{2Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k} + w \|h\|_D \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)} \right] \end{aligned}$$

### 3. Apply the Mean Value Theorem

Assuming that the penalty functions are smooth, then  $\hat{m}(\boldsymbol{\lambda})$  is continuous and differentiable. Then by the MVT, there is an  $\alpha \in (0, 1)$  such that

$$\begin{aligned}
\left| \hat{m}(\boldsymbol{\lambda}^{(2)}) - \hat{m}(\boldsymbol{\lambda}^{(1)}) \right| &= \left\langle \boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}, \nabla_{\boldsymbol{\lambda}} \hat{m}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\alpha\boldsymbol{\lambda}^{(1)}+(1-\alpha)\boldsymbol{\lambda}^{(2)}} \right\rangle \\
&\leq \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\| \left\| \nabla_{\boldsymbol{\lambda}} \hat{m}(\boldsymbol{\lambda})|_{\boldsymbol{\lambda}=\alpha\boldsymbol{\lambda}^{(1)}+(1-\alpha)\boldsymbol{\lambda}^{(2)}} \right\| \\
&\leq \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\| \left( w\sqrt{J}\lambda_{\min} \|h\|_D^2 \right)^{-1} \left[ \frac{2Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k} + w\|h\|_D \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)} \right]
\end{aligned}$$

Since  $\hat{m}(\boldsymbol{\lambda}^{(2)}) - \hat{m}(\boldsymbol{\lambda}^{(1)}) = 1$ , then we have

$$\|h\|_D^2 \leq \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\| \left( w\sqrt{J}\lambda_{\min} \right)^{-1} \left[ \frac{2Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k} + w\|h\|_D \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)} \right]$$

**Case 1:**

Suppose  $\frac{2Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k} \geq w\|h\|_D \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)}$ .

Then

$$\|h\|_D^2 \leq \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\| \left( w\sqrt{J}\lambda_{\min} \right)^{-1} \frac{4Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k}$$

**Case 2:**

Suppose  $\frac{2Cv_k}{\lambda_{\min}} \left[ \frac{C}{\lambda_{\min}} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right) \right]^{(v_k-1)/v_k} \leq w\|h\|_D \sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)}$ .

$$\|h\|_D \leq \|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\| \left( \sqrt{J}\lambda_{\min} \right)^{-1} 2\sqrt{\frac{2C}{\lambda_{\min}w} \left( 1 + \frac{J\lambda_{\max}}{\lambda_{\min}} \right)}$$

Unfortunately, for  $\|\boldsymbol{\lambda}^{(2)} - \boldsymbol{\lambda}^{(1)}\|$  sufficiently small, then  $\|h\|_D$  will be sufficiently small such that we will always be in Case 1.

### 3 Examples

Sobolev satisfies the conditions in Section 2

$$\arg \min_{g \in \mathcal{G}} \frac{1}{2} \left\| y - \sum_{j=1}^J g_j \right\|_T^2 + \sum_{j=1}^J \lambda_j \left( P_j(g_j) + \frac{w}{2} \|g_j\|_D^2 \right)$$

where

$$P_j(g_j) = \int \left( g_j^{(r_j)}(x) \right)^2 dx$$

Note that the Sobolev penalty is convex since

$$\frac{\partial^2}{\partial m^2} P_j(g + mh) = P_j(h) \geq 0$$

## 4 Appendix

**Lemma: Bounding the derivative of a semi-norm**

Let  $P$  be a semi-norm. Then

$$\left| \frac{\partial}{\partial m} P(a + mb) \right| \leq P(b)$$

**Proof**

By triangle inequality, we know

$$|P(a + mb) - P(a)| \leq |m|P(b)$$

Therefore as we take  $m \rightarrow 0$ , we have

$$\left| \frac{\partial}{\partial m} P(a + mb) \right| \leq P(b)$$