

Sobolev penalty:univariate

Given a function h , the Sobolev penalty for h is

$$P(h) = \int (h^{(r)}(x))^2 dx$$

Suppose $\sup_g \|g\|_\infty \leq G$.

We shall suppose for simplicity that the domain is $[0, 1]$.

Suppose we have the function class (so no additional ridge penalty)

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

Using the logic in Example 9.3.2 in Vandegeer, we can express any function in \mathcal{G} as

$$f + g$$

where

$$g = \sum_{k=1}^r \alpha_k \psi_k, f = \int_0^1 \beta_u \tilde{\phi}_u$$

where $\langle \psi_k, \tilde{\phi}_u \rangle_T = 0$ and $P(\psi_k) = 0$.

Suppose the observations were drawn from $y = f^*(x) + g^*(x) + \epsilon$ where ϵ are independent sub-gaussian random variables.

Now we have the function class

$$\hat{\mathcal{G}}(T) = \left\{ \hat{g}(\cdot|\lambda), \hat{f}(\cdot|\lambda) = \arg \min_{g \in \mathcal{G}} \frac{1}{2} \|y - (f + g)\|_T^2 + \lambda P(f) : \lambda \in \Lambda, g = \sum_{k=1}^r \alpha_k \psi_k, f = \int_0^1 \beta_u \tilde{\phi}_u \right\}$$

We will show that

$$\left\| \left(\hat{g}(\cdot|\lambda^{(1)}) + \hat{f}(\cdot|\lambda^{(1)}) \right) - \left(\hat{g}(\cdot|\lambda^{(2)}) + \hat{f}(\cdot|\lambda^{(2)}) \right) \right\|_\infty \leq |\lambda^{(1)} - \lambda^{(2)}| n^{\tau_{min}} \sqrt{\frac{n^{\tau_{min}}}{2} \|\epsilon\|_T^2 + P(f^*)} G$$

Proof

First by Vandegeer Example 9.3.2, we know that

$$\hat{g}(\cdot|\lambda) = \arg \min_{g = \sum \alpha_k \psi_k} -2\langle \epsilon, g - g^* \rangle_T + \|g - g^*\|_T^2$$

$$\hat{f}(\cdot|\lambda) = \arg \min_{f = \int_0^1 \beta_u \tilde{\phi}_u} -2\langle \epsilon, f - f^* \rangle_T + \|f - f^*\|_T^2 + \lambda P(f)$$

So $\hat{g}(\cdot|\lambda)$ is actually independent of λ and is therefore constant. We will just denote it \hat{g} from now on.

Now consider

$$h = c \left(\hat{f}(\cdot|\lambda^{(1)}) - \hat{f}(\cdot|\lambda^{(2)}) \right)$$

where c is some constant s.t. $P(h) = 1$.

We can assume that $P(h) \neq 0$. Otherwise, if

$$P \left(\hat{f}(\cdot|\lambda^{(1)}) - \hat{f}(\cdot|\lambda^{(2)}) \right) = 0$$

then we know that

$$\hat{f}(\cdot|\lambda^{(1)}) - \hat{f}(\cdot|\lambda^{(2)}) \in \text{span} \{\psi_k\}_{k=1}^r$$

This is true if and only if $\hat{f}(\cdot|\lambda^{(1)}) \equiv \hat{f}(\cdot|\lambda^{(2)})$ (by the fact that the function spaces are orthogonal). Consider the optimization problem

$$\hat{m}_h(\lambda) = \arg \min_m \frac{1}{2} \|y - (\hat{g} + \hat{f}(\cdot|\lambda^{(1)}) + mh)\|_T^2 + \lambda P(\hat{f}(\cdot|\lambda^{(1)}) + mh)$$

By implicit differentiation of the KKT conditions, we get

$$\left. \frac{\partial}{\partial \lambda} \hat{m}_h(\lambda) \right|_{\lambda=\lambda} = - \left(\|h\|_T^2 + \lambda \frac{\partial^2}{\partial m^2} P(\hat{f}(\cdot|\lambda^{(1)}) + mh) \right)^{-1} \left. \frac{\partial}{\partial m} P(\hat{f}(\cdot|\lambda^{(1)}) + mh) \right|_{m=\hat{m}_h(\lambda)}$$

Then the first multiplicand is bounded by

$$\begin{aligned} \left| \|h\|_T^2 + \lambda \frac{\partial^2}{\partial m^2} P(\hat{f}(\cdot|\lambda^{(1)}) + mh) \right|^{-1} &\leq n^{\tau_{min}} \frac{\partial^2}{\partial m^2} P(\hat{f}(\cdot|\lambda^{(1)}) + mh)^{-1} \\ &= \frac{n^{\tau_{min}}}{2P(h)} \end{aligned}$$

The equality follows from the Lemma Sobolev Facts (see below).

From the Lemma Sobolev Facts and by the fact that $P(h) = 1$, we have

$$\begin{aligned} \left| \frac{\partial}{\partial \lambda} \hat{m}_h(\lambda) \right|_{\lambda=\lambda} &\leq \frac{n^{\tau_{min}}}{P(h)} \sqrt{P(\hat{f}(\cdot|\lambda^{(1)}) + \hat{m}_h(\lambda)h) P(h)} \\ &= n^{\tau_{min}} \sqrt{P(\hat{f}(\cdot|\lambda^{(1)}) + \hat{m}_h(\lambda)h)} \end{aligned}$$

By the definition of $\hat{m}_h(\lambda)$ and $\hat{f}(\cdot|\lambda^{(1)})$, we have that

$$\begin{aligned} \lambda P(\hat{f}(\cdot|\lambda^{(1)}) + \hat{m}_h(\lambda)h) &\leq \frac{1}{2} \|y - (\hat{g} + \hat{f}(\cdot|\lambda^{(1)}))\|_T^2 + \lambda P(\hat{f}(\cdot|\lambda^{(1)})) \\ &= \frac{1}{2} \|y - (\hat{g} + \hat{f}(\cdot|\lambda^{(1)}))\|_T^2 + \lambda^{(1)} P(\hat{f}(\cdot|\lambda^{(1)})) + (\lambda - \lambda^{(1)}) P(\hat{f}(\cdot|\lambda^{(1)})) \\ &\leq \frac{1}{2} \|y - (g^* + f^*)\|_T^2 + \lambda^{(1)} P(f^*) + (\lambda - \lambda^{(1)}) P(\hat{f}(\cdot|\lambda^{(1)})) \end{aligned}$$

In addition, by definition of $\hat{f}(\cdot|\lambda^{(1)})$, we have

$$P(\hat{f}(\cdot|\lambda^{(1)})) \leq \frac{1}{2\lambda^{(1)}} \|y - (g^* + f^*)\|_T^2 + P(f^*)$$

Combining the two inequalities above, we have

$$\begin{aligned} \lambda P(\hat{f}(\cdot|\lambda^{(1)}) + \hat{m}_h(\lambda)h) &\leq \frac{1}{2} \|\epsilon\|_T^2 + \lambda^{(1)} P(f^*) + (\lambda - \lambda^{(1)}) \left(\frac{1}{2\lambda^{(1)}} \|\epsilon\|_T^2 + P(f^*) \right) \\ &\leq \frac{\lambda}{2\lambda^{(1)}} \|\epsilon\|_T^2 + \lambda P(f^*) \end{aligned}$$

Therefore

$$P(\hat{f}(\cdot|\lambda^{(1)}) + \hat{m}_h(\lambda)h) \leq \frac{n^{\tau_{min}}}{2} \|\epsilon\|_T^2 + P(f^*)$$

Then by the MVT, we have

$$\begin{aligned}
\|\hat{f}(\cdot|\lambda^{(1)}) - \hat{f}(\cdot|\lambda^{(2)})\|_\infty &= \| (m_h(\lambda^{(2)}) - m_h(\lambda^{(1)})) h \|_\infty \\
&= \|h\|_\infty \left| \lambda^{(2)} - \lambda^{(1)} \right| \left| \frac{\partial}{\partial \lambda} \hat{m}_h(\lambda) \right|_{\lambda \in [\lambda^{(1)}, \lambda^{(2)}]} \\
&\leq G \left| \lambda^{(1)} - \lambda^{(2)} \right| \left(\sup_{\lambda \in [\lambda^{(1)}, \lambda^{(2)}]} \left| \frac{\partial}{\partial \lambda} \hat{m}_h(\lambda) \right|_{\lambda=\lambda} \right) \\
&\leq \left| \lambda^{(1)} - \lambda^{(2)} \right| G n^{\tau_{min}} \sqrt{\frac{n^{\tau_{min}}}{2} \|\epsilon\|_T^2 + P(f^*)}
\end{aligned}$$

Sobolev penalty: multivariate

The function class of interest

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

We can show that

$$\|\hat{f}_\ell(\cdot|\lambda^{(1)}) - \hat{f}_\ell(\cdot|\lambda^{(2)})\|_\infty \leq G \left\| \lambda^{(1)} - \lambda^{(2)} \right\| \frac{n^{\tau_{min}}}{2} \sqrt{\frac{n^{\tau_{min}} + J n^{\tau_{max}} + 2\tau_{min}}{2} \|\epsilon\|_T^2 + n^{\tau_{max} + 2\tau_{min}} \sum_{j=1}^J P(f_j^*)}$$

A second approach gives a different bound:

$$\|\hat{f}_\ell(\cdot|\lambda^{(1)}) - \hat{f}_\ell(\cdot|\lambda^{(2)})\|_\infty \leq \left\| \lambda^{(1)} - \lambda^{(2)} \right\| \frac{G^2(2G + \|\epsilon\|_T) n^{3\tau_{min}}}{4}$$

Proof

First by Vandegeer Example 9.3.2, we know that

$$\begin{aligned}
\{\hat{g}_j(\cdot|\lambda)\}_{j=1}^J &= \arg \min_{g_j = \sum \alpha_k \psi_k} -2\langle \epsilon, \sum_{j=1}^J g_j - g_j^* \rangle_T + \left\| \sum_{j=1}^J g_j - g_j^* \right\|_T^2 \\
\{\hat{f}_j(\cdot|\lambda)\}_{j=1}^J &= \arg \min_{f_j = \int_0^1 \beta_u \tilde{\phi}_u} -2\langle \epsilon, \sum_{j=1}^J f_j - f_j^* \rangle_T + \left\| \sum_{j=1}^J f_j - f_j^* \right\|_T^2 + \sum_{j=1}^J \lambda_j P(f_j)
\end{aligned}$$

For every $j = 1 : J$, we again notice that $\hat{g}_j(\cdot|\lambda)$ is independent of λ . We will just denote it \hat{g}_j from now on.

For every $j = 1 : J$, define functions

$$h_j = c \left(\hat{f}_j(\cdot|\lambda^{(1)}) - \hat{f}_j(\cdot|\lambda^{(2)}) \right)$$

where c is some constant s.t. $P(h_j) = 1$.

We can assume that $P(h_j) \neq 0$. Otherwise, if

$$P \left(\hat{f}_j(\cdot|\lambda^{(1)}) - \hat{f}_j(\cdot|\lambda^{(2)}) \right) = 0$$

then we know that

$$\hat{f}_j(\cdot|\lambda^{(1)}) - \hat{f}_j(\cdot|\lambda^{(2)}) \in \text{span}\{\psi_k\}_{k=1}^r$$

This is true if and only if $\hat{f}_j(\cdot|\lambda^{(1)}) \equiv \hat{f}_j(\cdot|\lambda^{(2)})$ (by the fact that the function spaces are orthogonal). Now consider the optimization problem

$$\{\hat{m}_j(\lambda, h)\}_{j=1}^J = \arg \min_{m_j} \frac{1}{2} \|y - \sum_{j=1}^J (\hat{g}_j + \hat{f}_j(\cdot|\lambda^{(1)}) + m_j h_j)\|_T^2 + \sum_{j=1}^J \lambda_j P(\hat{f}_j(\cdot|\lambda^{(1)}) + m_j h_j)$$

(If $h_j \equiv 0$, then set $m_j = 0$ as a constant.) For simplicity, we will assume $h_j \neq 0$.

The KKT conditions give us for all $\ell = 1 : J$

$$\left\langle h_\ell, y - \left(\sum_{j=1}^J \hat{g}_j(\cdot|\lambda^{(1)}) + \hat{f}_j(\cdot|\lambda^{(1)}) + \hat{m}_j(\lambda, h) h_j \right) \right\rangle_T + \lambda_\ell \frac{\partial}{\partial m_\ell} P(\hat{f}_\ell(\cdot|\lambda^{(1)}) + m_\ell h) \Big|_{m_\ell = \hat{m}_\ell(\lambda, h)} = 0$$

For all $k = 1 : J$, by implicit differentiation of the KKT conditions with respect to λ_k , we get

$$\begin{aligned} \left\langle h_\ell, y - \sum_{j=1}^J h_j \frac{\partial}{\partial \lambda_k} \hat{m}_j(\lambda, h) \right\rangle_T + \lambda_\ell \frac{\partial^2}{\partial m_\ell^2} P(\hat{f}_\ell(\cdot|\lambda^{(1)}) + m_\ell h) \frac{\partial}{\partial \lambda_k} \hat{m}_\ell(\lambda, h) \\ + 1[\ell = k] \frac{\partial}{\partial m_\ell} P(\hat{f}_\ell(\cdot|\lambda^{(1)}) + m_\ell h) = 0 \end{aligned}$$

Define the following matrices

$$\begin{aligned} S : S_{ij} &= \langle h_j, h_i \rangle_T \\ D_1 &= \text{diag} \left(\lambda_\ell \frac{\partial^2}{\partial m_\ell^2} P(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell) \right) \\ D_3 &= \text{diag} \left(\frac{\partial}{\partial m_\ell} P(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell) \right) \\ M &= \begin{pmatrix} \frac{\partial \hat{m}_1(\lambda)}{\partial \lambda} & \frac{\partial \hat{m}_2(\lambda)}{\partial \lambda} & \dots & \frac{\partial \hat{m}_J(\lambda)}{\partial \lambda} \end{pmatrix} \end{aligned}$$

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

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

We know that S is a PSD matrix (since it can be written as $S = HH^T$ where $H_j = h_j$ evaluated at covariates T).

We are interested in bounding $\nabla_\lambda \hat{m}_\ell(\lambda, h)$, which is the ℓ -th column of M has norm. By Lemma PSD_Matrix_Inverse (see additive_models.pdf), we know that

$$\begin{aligned} \|\nabla_\lambda \hat{m}_\ell(\lambda, h)\| &= \|M e_\ell\| \\ &= \|D_3 (S + D_1)^{-1} e_\ell\| \\ &\leq \|D_3 D_1^{-1} e_\ell\| \\ &= \left\| \frac{\partial}{\partial m_\ell} P(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell) \right\| \left\| \lambda_\ell \frac{\partial^2}{\partial m_\ell^2} P(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell) \right\|^{-1} \end{aligned}$$

By Lemma Sobolev Facts (below), we have

$$\frac{\partial^2}{\partial m_\ell^2} P(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell) = 2P(h_\ell) = 2$$

Also by Lemma Sobolev Facts (below), we note that

$$\begin{aligned} \left| \frac{\partial}{\partial m_\ell} P \left(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell \right) \right| &\leq 2 \sqrt{P \left(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell \right) P(h_\ell)} \\ &= 2 \sqrt{P \left(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell \right)} \end{aligned}$$

By the definition of $\hat{m}_\ell(\lambda)$ and $\hat{f}(\cdot|\lambda^{(1)})$, we have

$$\begin{aligned} \lambda_\ell P \left(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell \right) &\leq \frac{1}{2} \|y - (\hat{g} + \hat{f}(\cdot|\lambda^{(1)}))\|_T^2 + \sum_{j=1}^J \lambda_j P \left(\hat{f}_j(\cdot|\lambda^{(1)}) \right) \\ &= \frac{1}{2} \|y - (g^* + f^*)\|_T^2 + \sum_{j=1}^J \lambda_j^{(1)} P(f_j^*) + \sum_{j=1}^J (\lambda_j - \lambda_j^{(1)}) P \left(\hat{f}_j(\cdot|\lambda^{(1)}) \right) \\ &\leq \frac{1}{2} \|\epsilon\|_T^2 + \lambda_{\max} \sum_{j=1}^J P(f_j^*) + J \lambda_{\max} \left[\max_{j=1:J} P \left(\hat{f}_j(\cdot|\lambda^{(1)}) \right) \right] \end{aligned}$$

By definition of $\hat{f}_j(\cdot|\lambda^{(1)})$, we know

$$\max_{j=1:J} P \left(\hat{f}_j(\cdot|\lambda^{(1)}) \right) \leq \frac{1}{\lambda_{\min}} \left(\frac{1}{2} \|\epsilon\|_T^2 + \sum_{j=1}^J \lambda_j^{(1)} P(f_j^*) \right)$$

So

$$P \left(\hat{f}_\ell(\cdot|\lambda^{(1)}) + m_\ell h_\ell \right) \leq \frac{n^{\tau_{\min}} + J n^{\tau_{\max} + 2\tau_{\min}}}{2} \|\epsilon\|_T^2 + n^{\tau_{\max} + 2\tau_{\min}} \sum_{j=1}^J P(f_j^*)$$

Then by the MVT, we have for some $\alpha \in (0, 1)$

$$\begin{aligned} \|\hat{f}_\ell(\cdot|\lambda^{(1)}) - \hat{f}_\ell(\cdot|\lambda^{(2)})\|_\infty &= \left\| \left(\hat{m}_\ell(\lambda^{(2)}, h) - \hat{m}_\ell(\lambda^{(1)}, h) \right) h_\ell \right\|_\infty \\ &= \|h_\ell\|_\infty \langle \lambda^{(1)} - \lambda^{(2)}, \nabla_\lambda \hat{m}_\ell(\lambda, h) \rangle_{\lambda=\alpha\lambda^{(1)}+(1-\alpha)\lambda^{(2)}} \\ &\leq G \left\| \lambda^{(1)} - \lambda^{(2)} \right\| \left\| \nabla_\lambda \hat{m}_\ell(\lambda, h) \right\| \\ &\leq G \left\| \lambda^{(1)} - \lambda^{(2)} \right\| \frac{n^{\tau_{\min}}}{2} \sqrt{\frac{n^{\tau_{\min}} + J n^{\tau_{\max} + 2\tau_{\min}}}{2} \|\epsilon\|_T^2 + n^{\tau_{\max} + 2\tau_{\min}} \sum_{j=1}^J P(f_j^*)} \end{aligned}$$

A second approach:

By the KKT conditions, we also know that

$$\begin{aligned} \left| \frac{\partial}{\partial m_\ell} P \left(\hat{f}_\ell(\cdot|\lambda^{(1)}) + \hat{m}_\ell(\lambda) h_\ell \right) \right| &= \frac{1}{\lambda_\ell} \left| \left\langle h_\ell, y - \left(\sum_{j=1}^J \hat{g}_j(\cdot|\lambda^{(1)}) + \hat{f}_j(\cdot|\lambda^{(1)}) + \hat{m}_j(\lambda, h) h_j \right) \right\rangle_T \right| \\ &\leq \frac{1}{\lambda_{\min}} \|h_\ell\|_T \left\| y - \left(\sum_{j=1}^J \hat{g}_j(\cdot|\lambda^{(1)}) + \hat{f}_j(\cdot|\lambda^{(1)}) + \hat{m}_j(\lambda, h) h_j \right) \right\|_T \\ &\leq G(2G + \|\epsilon\|_T) n^{\tau_{\min}} \end{aligned}$$

Hence

$$\|\nabla_\lambda \hat{m}_\ell(\lambda, h)\| \leq G(2G + \|\epsilon\|_T) n^{2\tau_{\min}} \frac{1}{2}$$

Then by the MVT, we have for some $\alpha \in (0, 1)$

$$\begin{aligned}
\|\hat{f}_\ell(\cdot|\lambda^{(1)}) - \hat{f}_\ell(\cdot|\lambda^{(2)})\|_\infty &= \|\left(\hat{m}_\ell(\lambda^{(2)}, h) - \hat{m}_\ell(\lambda^{(1)}, h)\right) h_\ell\|_\infty \\
&= \|h_\ell\|_\infty \langle \lambda^{(1)} - \lambda^{(2)}, \nabla_\lambda \hat{m}_\ell(\lambda, h) \rangle_{\lambda=\alpha\lambda^{(1)}+(1-\alpha)\lambda^{(2)}} \\
&\leq G \left\| \lambda^{(1)} - \lambda^{(2)} \right\| \left\| \nabla_\lambda \hat{m}_\ell(\lambda, h) \right\|_{\lambda=\alpha\lambda^{(1)}+(1-\alpha)\lambda^{(2)}} \\
&\leq \left\| \lambda^{(1)} - \lambda^{(2)} \right\| G^2 (2G + \|\epsilon\|_T) n^{3\tau_{min}} \frac{1}{4}
\end{aligned}$$

Lemma: Sobolev Facts

For any function h , we have

$$\begin{aligned}
\left| \frac{\partial}{\partial m} P(g + mh) \right| &= \left| 2 \int (g^{(r)}(x) + mh^{(r)}(x)) h^{(r)}(x) dx \right| \\
&\leq 2\sqrt{P(g + mh)P(h)}
\end{aligned}$$

and

$$\frac{\partial^2}{\partial m^2} P(g + mh) = 2 \int (h^{(r)}(x))^2 dx = 2P(h)$$