

Lemma: Additive Models and Additive Penalties

Consider the problem

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

We suppose the penalty functions P_j are convex and twice-differentiable. (We do not need the semi-norm assumption.)

Suppose that $\sup_{g \in \mathcal{G}} \|g\|_D \leq G$.

For all $d > 0$, any $\lambda^{(1)}, \lambda^{(2)}$ that satisfy

$$\|\lambda^{(1)} - \lambda^{(2)}\| \leq \frac{dw}{2J} \left(\frac{n}{n_T} n^{\tau_{\min}} (2G + \|\epsilon\|_T) + wG + G \right)^{-1} n^{-\tau_{\min}}$$

we have

$$\|\hat{g}_j(\cdot|\lambda^{(1)}) - \hat{g}_j(\cdot|\lambda^{(2)})\|_D \leq d/J$$

Hence

$$\left\| \sum_{j=1}^J \hat{g}_j(\cdot|\lambda^{(1)}) - \hat{g}_j(\cdot|\lambda^{(2)}) \right\|_D \leq d$$

Proof

Let $h_j = \hat{g}_j(\cdot|\lambda^{(1)}) - \hat{g}_j(\cdot|\lambda^{(2)})$. Suppose for contradiction that for \tilde{k} , we have $\|h_{\tilde{k}}\|_D > d/J$.

Let

$$Z = \{j : \|h_j\| > 0\}$$

Consider the optimization problem

$$\{\hat{m}_j(\lambda)\}_{j \in Z} = \arg \min_m \frac{1}{2} \|y - \sum_{j=1}^J (g_j + m_j h_j)\|_T^2 + \sum_{j=1}^J \lambda_j P_j(g_j + m_j h_j) + \frac{w}{2} \|g_j + m_j h_j\|_D^2$$

Note that if $\|h_j\| = 0$, then we just set $m_j = 0$ as a constant.

Now by the KKT conditions, for all $\ell \in Z$, we have

$$\langle y - \sum_{j=1}^J (g_j + m_j h_j), h_\ell \rangle_T + \lambda_\ell \frac{\partial}{\partial m_\ell} P_\ell(g_\ell + m_\ell h_\ell) + \lambda_\ell w \langle h_\ell, g_\ell + m_\ell h_\ell \rangle_D = 0$$

It's implicit derivative with respect to λ_k is

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

Define the following matrices

$$S : S_{ij} = \langle h_j, h_\ell \rangle_T$$

$$D_1 = \text{diag} \left(\lambda_\ell \frac{\partial^2}{\partial m_\ell^2} P_\ell(g_\ell + m_\ell h_\ell) \right)$$

$$D_2 = \text{diag} (\lambda_\ell w \|h_\ell\|_D^2)$$

$$D_3 = \text{diag} \left(\frac{\partial}{\partial m_\ell} P_\ell(g_\ell + m_\ell h_\ell) + w \langle h_\ell, g_\ell + m_\ell h_\ell \rangle_D \right)$$

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

(You will have to omit certain columns/rows of the matrices if $m_j = 0$ is constant.)

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

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

We know that S is a PSD matrix (since it can be written as $S = HH^T$ where $H_j = h_j|_T$). Hence we can express $S = UD_4U^T$ (where U are orthonormal)

$$M = D_3 (UD_4U^T + D_1 + D_2)^{-1}$$

Let $D_{12} = D_1 + D_2$. By Woodbury's matrix identity,

$$M = D_3 \left(D_{12}^{-1} - D_{12}^{-1}U (\Sigma^{-1} + D_{12})^{-1} U^T D_{12}^{-1} \right)$$

Now consider the \tilde{k} th column of M :

$$\nabla_\lambda \hat{m}_{\tilde{k}}(\lambda) = D_3 \left(D_{12}^{-1} - D_{12}^{-1}U (\Sigma^{-1} + D_{12})^{-1} U^T D_{12}^{-1} \right) e_{\tilde{k}}$$

Note that the j th elements for $j \neq \tilde{k}$ must be zero since orthogonality is preserved by the orthonormal matrices U . Also note that $D_{12}^{-1}U (\Sigma^{-1} + D_{12})^{-1} U^T D_{12}^{-1}$ is a positive definite matrix. Hence we have

$$\left| \frac{\partial}{\partial \lambda_{\tilde{k}}} \hat{m}_{\tilde{k}}(\lambda) \right| \leq \left| \frac{\partial}{\partial m_{\tilde{k}}} P_{\tilde{k}}(g_{\tilde{k}} + m_{\tilde{k}} h_{\tilde{k}}) + w \langle h_{\tilde{k}}, g_{\tilde{k}} + m_{\tilde{k}} h_{\tilde{k}} \rangle_D \right| \lambda_{\tilde{k}}^{-1} w^{-1} \|h_{\tilde{k}}\|_D^{-2}$$

Note that from the KKT conditions, we have that

$$\begin{aligned} \left| \frac{\partial}{\partial m_{\tilde{k}}} P_{\tilde{k}}(g_{\tilde{k}} + m_{\tilde{k}} h_{\tilde{k}}) \right| &\leq \left| \frac{1}{\lambda_{\tilde{k}}} \langle y - \sum_{j=1}^J (g_j + m_j h_j), h_{\tilde{k}} \rangle_T + w \langle h_{\tilde{k}}, g_{\tilde{k}} + m_{\tilde{k}} h_{\tilde{k}} \rangle_D \right| \\ &\leq n^{\tau_{\min}} \|y - \sum_{j=1}^J (g_j + m_j h_j)\|_T \|h_{\tilde{k}}\|_T + w \|h_{\tilde{k}}\|_D \|g_{\tilde{k}} + m_{\tilde{k}} h_{\tilde{k}}\|_D \\ &\leq \left(\frac{n}{n_T} n^{\tau_{\min}} (2G + \|\epsilon\|_T) + wG \right) \|h_{\tilde{k}}\|_D \end{aligned}$$

Also

$$w \langle h_{\tilde{k}}, g_{\tilde{k}} + m_{\tilde{k}} h_{\tilde{k}} \rangle_D \leq w \|h_{\tilde{k}}\|_D G$$

Hence

$$\left| \frac{\partial}{\partial \lambda_{\tilde{k}}} \hat{m}_{\tilde{k}}(\lambda) \right| \leq \left(\frac{n}{n_T} n^{\tau_{\min}} (2G + \|\epsilon\|_T) + wG + G \right) n^{\tau_{\min}} w^{-1} \|h_{\tilde{k}}\|_D^{-1}$$

By the MVT, there is some $\alpha \in [0, 1]$ such that

$$\begin{aligned} \left| \hat{m}_{\tilde{k}}(\lambda^{(2)}) - \hat{m}_{\tilde{k}}(\lambda^{(1)}) \right| &= \left| \left\langle \lambda^{(2)} - \lambda^{(1)}, \frac{\partial}{\partial \lambda_{\tilde{k}}} \hat{m}_{\tilde{k}}(\lambda) \right\rangle_{\lambda = \alpha \lambda^{(1)} + (1-\alpha) \lambda^{(2)}} \right| \\ &\leq \|\lambda^{(2)} - \lambda^{(1)}\| \left(\frac{n}{n_T} n^{\tau_{\min}} (2G + \|\epsilon\|_T) + wG + G \right) n^{\tau_{\min}} \frac{J}{dw} \\ &= 1/2 \end{aligned}$$

But this is a contradiction since we know that $\hat{m}_{\tilde{k}}(\lambda^{(2)}) = 1$ and $\hat{m}_{\tilde{k}}(\lambda^{(1)}) = 0$.

Lemma: Additive Models and Additive Penalties, Nonsmooth

Same assumptions as above, but we allow the penalties to be nonsmooth.

Suppose for almost every λ , the differentiable space $\Omega^{L_T(\cdot, \lambda)}(\hat{g}(\cdot|\lambda))$ is a local optimality space.

Suppose for almost every λ , the penalty function is twice differentiable in the differentiable space.

The conclusions are the same as before.

For all $d > 0$, any $\lambda^{(1)}, \lambda^{(2)}$ that satisfy

$$\|\lambda^{(1)} - \lambda^{(2)}\| \leq \frac{dw}{2J} \left(\frac{n}{n_T} n^{\tau_{\min}} (2G + \|\epsilon\|_T) + wG + G \right)^{-1} n^{-\tau_{\min}}$$

we have

$$\|\hat{g}_j(\cdot|\lambda^{(1)}) - \hat{g}_j(\cdot|\lambda^{(2)})\|_D \leq d/J$$

Hence

$$\left\| \sum_{j=1}^J \hat{g}_j(\cdot|\lambda^{(1)}) - \hat{g}_j(\cdot|\lambda^{(2)}) \right\|_D \leq d$$

Proof

Let $\lambda^{(1)}, \lambda^{(2)}$ be the penalty parameters satisfying the distance constraint above. Let C be the constant defined in the assumption

$$\|\lambda^{(1)} - \lambda^{(2)}\| \leq dC$$

Under the assumptions about the differentiable space and the local optimality space, we know that for almost every pair $\lambda^{(1)}, \lambda^{(2)}$, there is a line

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

containing a finite set of points $\{\ell_i\}_{i=0}^{N+1} \subset \mathcal{L}$ where $\ell_0 = \lambda^{(1)}$ and $\ell_{N+1} = \lambda^{(2)}$ such that:

1. The differentiable spaces $\Omega^{L_T(\cdot, \ell_i)}(\hat{g}(\cdot|\ell_i))$ satisfy the condition that the differentiable space is a local optimality differentiable space conditions and

2. The union of the differentiable spaces contains the entire line \mathcal{L} :

$$\mathcal{L} \subset \bigcup_{i=0}^{N+1} \Omega^{L_T(\cdot, \ell_i)}(\hat{g}(\cdot|\ell_i))$$

Now we partition \mathcal{L} according to the differentiable spaces. We will partition with the centers of each differentiable space and points in the intersection of all the differentiable spaces. Let $\{\ell_{(i)}\}_{i=0}^N \subset \mathcal{L}$ be the points such that $\ell_{(i)}$ is in the differentiable space $\Omega^{L_T(\cdot, \ell_i)}(\hat{g}(\cdot|\ell_i))$ and $\Omega^{L_T(\cdot, \ell_{i+1})}(\hat{g}(\cdot|\ell_{i+1}))$. That is, we choose

$$\ell_{(i)} \in \Omega^{L_T(\cdot, \ell_i)}(\hat{g}(\cdot|\ell_i)) \cap \Omega^{L_T(\cdot, \ell_{i+1})}(\hat{g}(\cdot|\ell_{i+1}))$$

Hence the following points form a partition of \mathcal{L}

$$(\ell_0, \ell_{(0)}), (\ell_{(0)}, \ell_1), \dots, (\ell_N, \ell_{(N)}), (\ell_{(N)}, \ell_{N+1})$$

Note that

$$\|\ell_i - \ell_{(i)}\| \leq \frac{\|\ell_i - \ell_{(i)}\|}{\|\lambda^{(1)} - \lambda^{(2)}\|} dC$$

Applying the smooth lemma to the pairs of points above, we have that

$$\|g(\cdot|\ell_i) - g(\cdot|\ell_{(i)})\|_D \leq \frac{\|\ell_i - \ell_{(i)}\|}{\|\lambda^{(1)} - \lambda^{(2)}\|} d$$

Similarly,

$$\|g(\cdot|\ell_{i+1}) - g(\cdot|\ell_{(i)})\|_D \leq \frac{\|\ell_{i+1} - \ell_{(i)}\|}{\|\lambda^{(1)} - \lambda^{(2)}\|} d$$

Hence

$$\begin{aligned} \|g(\cdot|\lambda^{(1)}) - g(\cdot|\lambda^{(2)})\|_D &\leq \sum_{i=0}^N \|g(\cdot|\ell_i) - g(\cdot|\ell_{(i)})\|_D + \|g(\cdot|\ell_{i+1}) - g(\cdot|\ell_{(i)})\|_D \\ &\leq d \left(\sum_{i=0}^N \frac{\|\ell_{i+1} - \ell_{(i)}\|}{\|\lambda^{(1)} - \lambda^{(2)}\|} + \frac{\|\ell_i - \ell_{(i)}\|}{\|\lambda^{(1)} - \lambda^{(2)}\|} \right) \\ &= d \end{aligned}$$