Supplementary Materials for 'Discovering Model Structure for Partially Linear Model'

Xin He[†] and Junhui Wang[‡]

[†] School of Statistics and Management
Shanghai University of Finance and Economics

[‡] Department of Mathematics
City University of Hong Kong

1 Updating g and H

1.1 Update g

Denote $g_l(\mathbf{x}) = \sum_{i=1}^n \alpha_i^l K(\mathbf{x}_i, \mathbf{x}) = \boldsymbol{\alpha}_l^T \mathbf{K}_{\mathbf{x}}$, where $\boldsymbol{\alpha}_l = (\alpha_1^l, ..., \alpha_n^l)^T \in \mathcal{R}^n$. Let $\boldsymbol{\alpha} = (\boldsymbol{\alpha}_1^T, ..., \boldsymbol{\alpha}_p^T)^T \in \mathcal{R}^{np}$, we have $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), ..., g_p(\mathbf{x}))^T = (\boldsymbol{\alpha}_1^T \mathbf{K}_{\mathbf{x}}, ..., \boldsymbol{\alpha}_p^T \mathbf{K}_{\mathbf{x}})^T = \boldsymbol{\alpha}^T \mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}}$ and $\lambda_0 \sum_{l=1}^p \|g_l\|_{\mathcal{H}_K}^2 = \lambda_0 \boldsymbol{\alpha}^T \mathbf{I}_p \otimes \mathbf{K}\, \boldsymbol{\alpha}$, where $\mathbf{I}_p$ is the p-dimensional identity matrix. Assuming the current iteration time is t, then minimizing (4) in the main paper is equivalent to solve

$$
\begin{aligned}
0 &= \nabla_{\boldsymbol{\alpha}} \left(\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H}^t) + \lambda_0 \sum_{l=1}^p \|g_l\|_{\mathcal{H}_K}^2 \right) \\
&= \frac{2}{n(n-1)} \sum_{i,j=1}^n w_{ij} \Big(\big(\mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_i - \mathbf{x}_j)\big) \big(\mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_i - \mathbf{x}_j)\big)^T \Big) \boldsymbol{\alpha} + 2\lambda_0 \mathbf{I}_p \otimes \mathbf{K}\, \boldsymbol{\alpha} \\
&\qquad + \frac{2}{n(n-1)} \sum_{i,j=1}^n w_{ij} \Big(y_i - y_j + \frac{1}{2} (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^t(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j) \Big) \mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_j - \mathbf{x}_i).
\end{aligned}
$$

Then we get the explicit solution for $\boldsymbol{\alpha}$ as

$$
\begin{aligned}
\boldsymbol{\alpha}^{t+1} = &\left(\frac{1}{n(n-1)} \sum_{i,j=1}^n w_{ij} \left((\mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_i - \mathbf{x}_j))(\mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_i - \mathbf{x}_j))^T \right) + \lambda_0 \mathbf{I}_p \otimes \mathbf{K} \right)^{-1} \\
&\left(\frac{1}{n(n-1)} \sum_{i,j=1}^n w_{ij} \Big(y_i - y_j + \frac{1}{2} (\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}^t(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j) \Big) \mathbf{I}_p \otimes \mathbf{K}_{\mathbf{x}_i}(\mathbf{x}_i - \mathbf{x}_j) \right).
\end{aligned}
$$

1.2 Update $\mathbf{H}$

For the one-homogeneous functional $\Omega(\cdot)$, $\Omega(\vartheta\mathbf{H}) = \vartheta\Omega(\mathbf{H})$ for $\vartheta > 0$ and $\mathbf{H} \in \mathcal{H}_K^{p\times p}$, the equivalent relationship between the proximal operator and the projection operator is given by the Moreau identity (Combettes and Wajs, 2005),

$$\text{prox}_{\mu\Omega} = I - \pi_{\mu C_n}, \tag{S1}$$

where $C_n = (\nabla\Omega(\mathbf{0}))$ is the subdifferential of Ω at the origin, and $\pi_{\mu C_n} : \mathcal{H}_K^{p\times p} \to \mathcal{H}_K^{p\times p}$ is a projection on μC_n. Furthermore, applying the Proposition 2 of Rosasco et al. (2009) and the identity (S1), the proximal operator can be computed as

$$[I-\pi_{\mu C_n}(\mathbf{H})]_{ll'} = [\mathbf{H}]_{ll'} - \min\{\mu\lambda_{ll'}, \|[\mathbf{H}]_{ll'}\|_{\mathcal{H}_K}\}\frac{[\mathbf{H}]_{ll'}}{\|[\mathbf{H}]_{ll'}\|_{\mathcal{H}_K}} = \frac{[\mathbf{H}]_{ll'}}{\|[\mathbf{H}]_{ll'}\|_{\mathcal{H}_K}}\left(\|[\mathbf{H}]_{ll'}\|_{\mathcal{H}_K} - \mu\lambda_{ll'}\right)_+,$$

where $\lambda_{ll'} = \lambda_1\pi_{ll'}$. In our algorithm, we set $\mu = 1/D$, where D is the Lipschitz constant. Then following (S1), the proximal operator at the t-th iteration can be expressed explicitly as

$$[\mathbf{H}^{t+1}]_{ll'} = \frac{[\bar{\mathbf{H}}^t]_{ll'}}{\|[\bar{\mathbf{H}}]_{ll'}\|_{\mathcal{H}_K}}(\|[\bar{\mathbf{H}}]_{ll'}\|_{\mathcal{H}_K} - \frac{\lambda_{ll'}}{D})_+,$$

where $\bar{\mathbf{H}}^{t+1} = \widetilde{\mathbf{H}}^t - \frac{1}{D}\nabla_{\mathbf{H}}\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}^{t+1}, \widetilde{\mathbf{H}}^t)$, $\widetilde{\mathbf{H}}^t = \mathbf{H}^t + \frac{t-1}{t+2}(\mathbf{H}^t - \mathbf{H}^{t-1})$ and

$$\nabla_{\mathbf{H}}\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}^{t+1}, \mathbf{H}^t) = \frac{1}{n(n-1)}\sum_{i,j=1}^n w_{ij}\Big(y_i - y_j - \mathbf{g}^{t+1}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j) + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T\mathbf{H}^t(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j)\Big)$$
$$\mathbf{K}_{x_i}\Big(\big((\mathbf{x}_i - \mathbf{x}_j)\otimes\mathbf{1}_p\big)^T \odot \big(\mathbf{1}_p^T \otimes (\mathbf{x}_i - \mathbf{x}_j)^T\big)\Big).$$

Here $\odot$ denotes the componentwise product and $\mathbf{1}_p$ is a p-vector with all ones.

2 Technical proofs

Proposition 1. *Assume* $\mathbf{g}^* \in \mathcal{H}_K^p$ *and* $\mathbf{H}^* \in \mathcal{H}_K^{p\times p}$. *Let* $\varphi_1(\mathcal{Z}^n) = \mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - \mathcal{E}_{\mathcal{Z}^n}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}})$, $\varphi_2(\mathcal{Z}^n) = \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}^*, \mathbf{H}^*) - \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*)$ *and* $\Lambda(\lambda_0, \lambda_1, \mathbf{K}) = \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) + \lambda_0\sum_{l=1}^p \|g_l^*\|^2_{\mathcal{H}_K} + \lambda_1\sum_{l,l'=1}^p \pi_{ll'}\|H^*_{ll'}\|_{\mathcal{H}_K}$. *Then the following inequality holds*

$$\mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) + J(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) \le \varphi_1(\mathcal{Z}^n) + \varphi_2(\mathcal{Z}^n) + \Lambda(\lambda_0, \lambda_1, \mathbf{K}).$$

Proof of Proposition 1. Simple algebra yields that

$$\begin{aligned}
&\mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) + J(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) \\
&= \mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) + \mathcal{E}_{\mathcal{Z}^n}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - \mathcal{E}_{\mathcal{Z}^n}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) + J(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) \\
&\le \mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - \mathcal{E}_{\mathcal{Z}^n}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) + \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}^*, \mathbf{H}^*) + J(\mathbf{g}^*, \mathbf{H}^*) \\
&= \mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - \mathcal{E}_{\mathcal{Z}^n}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) + \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}^*, \mathbf{H}^*) + \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) - \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) + J(\mathbf{g}^*, \mathbf{H}^*) \\
&= \varphi_1(\mathcal{Z}^n) + \varphi_2(\mathcal{Z}^n) + \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) + \lambda_0\sum_{l=1}^p \|g_l^*\|^2_{\mathcal{H}_K} + \lambda_1\sum_{l,l'=1}^p \pi_{ll'}\|H^*_{ll'}\|_{\mathcal{H}_K},
\end{aligned}$$

where the inequality comes from the definition of $(\widehat{\mathbf{g}}, \widehat{\mathbf{H}})$. $\square$

Next, we consider the following function space

$$\mathcal{F}_{r_n} = \{\mathbf{g} \in \mathcal{H}_K^p, \mathbf{H} \in \mathcal{H}_K^{p\times p} : \lambda_0 \sum_{l=1}^p \|g_l\|_{\mathcal{H}_K} \le r_n, \lambda_1 \sum_{l,l'=1}^p \pi_{ll'} \|H_{ll'}\|_{\mathcal{H}_K} \le r_n\},$$

for some positive $r_n \ge \frac{1}{n(n-1)} \sum_{i,j=1}^n (y_i - y_j)^2$. By the definition of $(\widehat{\mathbf{g}}, \widehat{\mathbf{H}})$, $\mathcal{E}_{\mathcal{Z}^n}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) + J(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) \le \mathcal{E}_{\mathcal{Z}^n}(\mathbf{0}, \mathbf{0}) + J(\mathbf{0}, \mathbf{0}) \le \frac{1}{n(n-1)} \sum_{i,j=1}^n (y_i - y_j)^2$, implying that $(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) \in \mathcal{F}_{r_n}$. Denote

$$S(\mathcal{Z}^n, r_n) = \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} |\mathcal{E}(\mathbf{g}, \mathbf{H}) - \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H})|.$$

By McDiarmid's inequality (McDiarmid, 1989), we obtain the upper bound of $S(\mathcal{Z}^n, r_n)$.

Lemma 1. *Suppose that Assumption 3 is met. If $|y| \le M_n$ then for any $\epsilon > 0$,*

$$P(|S(\mathcal{Z}^n, r_n) - E(S(\mathcal{Z}^n, r_n))| \ge \epsilon) \le 2\exp\left(-\frac{n\epsilon^2}{32\Big(M_n + c_{\mathbf{x}}\kappa \left(p r_n \lambda_0^{-1}\right)^{1/2} + c_{\mathbf{x}}^2 \kappa r_n \lambda_1^{-1} c_2^{-1}\Big)^4}\right),$$

where $c_{\mathbf{x}} = \max_{\mathbf{x}\in\mathcal{X}} \| \mathbf{x} \|_\infty$ and $\kappa = \sup_{x\in\mathcal{X}} \left(K(\mathbf{x}, \mathbf{x})\right)^{1/2}$.

Proof of lemma 1. Denote $\mathcal{Z}^{n'}$ as a sample that is the same as $\mathcal{Z}^n$ except the i-th entry replaced by (x_i', y_i'), then we have

$$\begin{aligned} S(\mathcal{Z}^n, r_n) - S(\mathcal{Z}^{n'}, r_n) &= \sup_{(\mathbf{g},\mathbf{H})} |\mathcal{E}(\mathbf{g}, \mathbf{H}) - \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H})| - \sup_{(\mathbf{g},\mathbf{H})} |\mathcal{E}(\mathbf{g}, \mathbf{H}) - \mathcal{E}_{\mathcal{Z}^{n'}}(\mathbf{g}, \mathbf{H})| \\ &\le \sup_{(\mathbf{g},\mathbf{H})} \big(|\mathcal{E}(\mathbf{g}, \mathbf{H}) - \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H})| - |\mathcal{E}(\mathbf{g}, \mathbf{H}) - \mathcal{E}_{\mathcal{Z}^{n'}}(\mathbf{g}, \mathbf{H})|\big) \\ &\le \sup_{(\mathbf{g},\mathbf{H})} |\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H}) - \mathcal{E}_{\mathcal{Z}^{n'}}(\mathbf{g}, \mathbf{H})|, \end{aligned}$$

where the first inequality is trivial and the second inequality follows from the triangle inequality. Now we decompose $\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H})$ as

$$\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g}, \mathbf{H}) = \frac{1}{n(n-1)}\Big(\sum_{t\ne i, j\ne i}^n h(\mathbf{z}_t, \mathbf{z}_j) + \sum_{j=1, j\ne i}^n h(\mathbf{z}_i, \mathbf{z}_j) + \sum_{t=1, t\ne i}^n h(\mathbf{z}_t, \mathbf{z}_i)\Big),$$

where $h(\mathbf{z}_i, \mathbf{z}_j) = w_{ij}\big(y_i - y_j - \mathbf{g}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j) + \frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \mathbf{H}(\mathbf{x}_i)(\mathbf{x}_i - \mathbf{x}_j)\big)^2$ and $\mathbf{z}_i = (\mathbf{x}_i, y_i)$.

Then

$$
\begin{aligned}
&|\mathcal{E}_{\mathcal{Z}^n}(\mathbf{g},\mathbf{H}) - \mathcal{E}_{\mathcal{Z}^{n\prime}}(\mathbf{g},\mathbf{H})| \\
&= \frac{1}{n(n-1)} \left| \sum_{j=1,j\neq i}^{n} h(\mathbf{z}_i,\mathbf{z}_j) - \sum_{j=1,j\neq i'}^{n} h(\mathbf{z}_{i'},\mathbf{z}_j) + \sum_{t=1,t\neq i}^{n} h(\mathbf{z}_t,\mathbf{z}_i) - \sum_{t=1,t\neq i'}^{n} h(\mathbf{z}_t,\mathbf{z}_{i'}) \right| \\
&\leq \frac{4}{n} \left(M_n + c_{\mathbf{x}} \sum_{l=1}^{p} \|g_l(\mathbf{x})\|_\infty + c_{\mathbf{x}}^2 \sum_{l,l'} \|H_{ll'}(\mathbf{x})\|_\infty \right)^2 \\
&\leq \frac{8}{n} \left(M_n + c_{\mathbf{x}} \sum_{l=1}^{p} \langle g_l, \mathbf{K}_x \rangle_{\mathcal{H}_K} + c_{\mathbf{x}}^2 \sum_{l,l'=1}^{p} \langle H_{ll'}, \mathbf{K}_x \rangle_{\mathcal{H}_K} \right)^2 \\
&\leq \frac{8}{n} \left(M_n + c_{\mathbf{x}}\kappa \sum_{l=1}^{p} \|g_l\|_{\mathcal{H}_K} + c_{\mathbf{x}}^2 \kappa \sum_{l,l'=1}^{p} \|H_{ll'}\|_{\mathcal{H}_K} \right)^2 \\
&\leq \frac{8}{n} \left(M_n + c_{\mathbf{x}}\kappa \left(pr_n\lambda_0^{-1}\right)^{1/2} + c_{\mathbf{x}}^2 \kappa r_n \lambda_1^{-1} c_2^{-1} \right)^2,
\end{aligned}
$$

where the first inequality is trivial, the second inequality follows from the property of RKHS, and the last two inequalities follow from Cauchy-Schwartz inequality, the definition of the $\mathcal{F}_{r_n}$ and Assumption 3. Therefore, we have

$$
\left| \mathcal{E}_{\mathcal{Z}^n}(\mathbf{g},\mathbf{H}) - \mathcal{E}_{\mathcal{Z}^{n\prime}}(\mathbf{g},\mathbf{H}) \right| \leq \frac{8}{n} \left(M_n + c_{\mathbf{x}}\kappa \left(pr_n\lambda_0^{-1}\right)^{1/2} + c_{\mathbf{x}}^2 \kappa r_n \lambda_1^{-1} c_2^{-1} \right)^2.
$$

Finally, by McDiarmid's Inequality, we have

$$
P(|S(\mathcal{Z}^n, r_n) - E(S(\mathcal{Z}^n, r_n))| \geq \epsilon) \leq 2\exp\left(- \frac{n\epsilon^2}{32\left(M_n + c_{\mathbf{x}}\kappa \left(pr_n\lambda_0^{-1}\right)^{1/2} + c_{\mathbf{x}}^2 \kappa r_n \lambda_1^{-1} c_2^{-1} \right)^4} \right).
$$

This completes the proof of the desired lemma. □

Lemma 2. *If $|y| \leq M_n$, there exists a constant $b_{\kappa,\mathbf{x}}$ such that*

$$
\mathbb{E}\left(S(\mathcal{Z}^n, r_n)\right) \leq b_{\kappa,\mathbf{x}} n^{-1/2} \left(M_n + \left(pr_n\lambda_0^{-1}\right)^{1/2} + r_n\lambda_1^{-1} \right)^2.
$$

Proof of Lemma 2. For simplicity, denote

$$
\xi(\mathbf{x}, y, \mathbf{u}, v) = w(\mathbf{x}-\mathbf{u}) \left(y - v - \mathbf{g}(\mathbf{x})^T(\mathbf{x}-\mathbf{u}) + \frac{1}{2}(\mathbf{x}-\mathbf{u})^T \mathbf{H}(\mathbf{x})(\mathbf{x}-\mathbf{u}) \right)^2,
$$

where $(\mathbf{u}, v)$ is an independent copy of $(\mathbf{x}, y)$. Then

$$
\begin{aligned}
& S(\mathcal{Z}, r_n) = \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} |\mathcal{E}_{\mathcal{Z}}(\mathbf{g},\mathbf{H}) - \mathcal{E}(\mathbf{g},\mathbf{H})| \\
& \le \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \Big|\mathcal{E}(\mathbf{g},\mathbf{H}) - \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\xi(\mathbf{x}, y, \mathbf{x}_j, y_j)\Big| + \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \Big|\frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\xi(\mathbf{x}, y, \mathbf{x}_j, y_j) - \mathcal{E}_{\mathcal{Z}}(\mathbf{g},\mathbf{H})\Big| \\
& \le \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \mathbb{E}_{(\mathbf{x},y)} \Big|\mathbb{E}_{(\mathbf{u},v)}\xi(\mathbf{x}, y, \mathbf{u}, v) - \frac{1}{n}\sum_{j=1}^{n} \xi(\mathbf{x}, y, \mathbf{x}_j, y_j)\Big| \\
& + \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \frac{1}{n}\sum_{j=1}^{n} \Big|\mathbb{E}_{(\mathbf{x},y)}\xi(\mathbf{x}, y, \mathbf{x}_j, y_j) - \frac{1}{(n-1)}\sum_{i\neq j} \xi(\mathbf{x}_i, y_i, \mathbf{x}_j, y_j)\Big| \\
& \le \mathbb{E}_{(\mathbf{x},y)} \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \Big|\mathbb{E}_{(\mathbf{u},v)}\xi(\mathbf{x}, y, \mathbf{u}, v) - \frac{1}{n}\sum_{j=1}^{n} \xi(\mathbf{x}, y, \mathbf{x}_j, y_j)\Big| \\
& + \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \frac{1}{n}\sum_{j=1}^{n} \sup_{(\mathbf{u},v)\in\mathcal{X}} \Big|\mathbb{E}_{(\mathbf{x},y)}\xi(\mathbf{x}, y, \mathbf{u}, v) - \frac{1}{(n-1)}\sum_{i\neq j} \xi(\mathbf{x}_i, y_i, \mathbf{u}, v)\Big| \\
& \stackrel{\text{def}}{=} S_1(\mathcal{Z}) + S_2(\mathcal{Z}),
\end{aligned}
$$

where the first inequality follows from the triangle inequality, and the next two inequalities follow from the definition of $\mathcal{E}(\mathbf{g}, \mathbf{H})$ and Jensen's inequality, respectively.

Next, we use Rademacher complexity (Bartlett and Mendelson, 2002) to get the upper bounds of $E(S_1)$ and $E(S_2)$. In fact, there holds

$$
\begin{aligned}
& \mathbb{E}[S_1(\mathcal{Z})] = \mathbb{E}_{\mathcal{Z}}\mathbb{E}_{(\mathbf{x},\mathbf{y})} \sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \Big|\mathbb{E}_{(\mathbf{u},v)}\xi(\mathbf{x}, y, \mathbf{u}, v) - \frac{1}{n}\sum_{j=1}^{n} \xi(\mathbf{x}, y, \mathbf{x}_j, y_j)\Big| \\
& \le 2\mathbb{E}_{(\mathbf{x},\mathbf{y})}\mathbb{E}_{\mathcal{Z},\sigma}\Big(\sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \Big|\frac{1}{n}\sum_{j=1}^{n} \sigma_j \xi(\mathbf{x}, y, \mathbf{x}_j, y_j)\Big|\Big) \\
& \le 4\left(M_n + c_{\mathbf{x}}\kappa\left(p r_n \lambda_0^{-1} c_2^{-1}\right)^{1/2} + c_{\mathbf{x}}^2 \kappa r_n \lambda_1^{-1} c_2^{-1}\right) \mathbb{E}_{(\mathbf{x},\mathbf{y})}\mathbb{E}_{\mathcal{Z},\sigma}\left(\sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{j=1}^{n} \sigma_j \Big(y - y_j - \right.\right. \\
& \left.\left. \mathbf{g}(\mathbf{x})^T(\mathbf{x} - \mathbf{x}_j) + \frac{1}{2}(\mathbf{x} - \mathbf{x}_j)^T\mathbf{H}(\mathbf{x})(\mathbf{x} - \mathbf{x}_j)\Big)\right|\right) \\
& \le 4\left(M_n + c_{\mathbf{x}}\kappa\left(p r_n \lambda_0^{-1} c_2^{-1}\right)^{1/2} + c_{\mathbf{x}}^2 \kappa r_n \lambda_1^{-1} c_2^{-1}\right) \left(\mathbb{E}_{(\mathbf{x},\mathbf{y})}\mathbb{E}_{\mathcal{Z},\sigma}\left(\sup_{(\mathbf{g},\mathbf{H})\in\mathcal{F}_{r_n}} \left|\frac{1}{n}\sum_{j=1}^{n} \sigma_j \right.\right.\right. \\
& \left.\left.\left. \Big(\mathbf{g}(\mathbf{x})^T(\mathbf{x} - \mathbf{x}_j) - \frac{1}{2}(\mathbf{x} - \mathbf{x}_j)^T\mathbf{H}(\mathbf{x})(\mathbf{x} - \mathbf{x}_j)\Big)\right|\right) + 2n^{-1/2}M_n\right) \\
& \le \frac{b_{\kappa,\mathbf{x}}}{2} n^{-1/2}\left(M_n + \left(p r_n \lambda_0^{-1}\right)^{1/2} + r_n \lambda_1^{-1}\right)^2,
\end{aligned}
$$

where σ_j's are a sequence of Rademacher variables. Similarly, we have

$$\mathbb{E}S_2(\mathcal{Z}) \leq \frac{b_{\kappa,\mathbf{x}}}{2} n^{-1/2} \left(M_n + \left(pr_n\lambda_0^{-1}\right)^{1/2} + r_n\lambda_1^{-1} \right)^2,$$

which implies the desired inequality. □

Proposition 2. *If $|y| \leq M_n$ and $\frac{1}{n(n-1)}\sum_{i,j=1}^n (y_i - y_j)^2 \leq M_0$, there exists a constant b_1 such that with probability at least $1 - \frac{\delta_n}{2}$,*

$$\varphi_1(\mathcal{Z}^n) \leq b_1 \left(\frac{1}{n}\log\frac{4}{\delta_n}\right)^{1/2} \left(M_n + \left(pM_0\lambda_0^{-1}\right)^{1/2} + M_0\lambda_1^{-1} \right)^2.$$

Proof of Proposition 2. Since $\frac{1}{n(n-1)}\sum_{i,j=1}^n (y_i - y_j)^2 \leq M_0$, it implies that $(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) \in \mathcal{F}_{M_0}$. By Lemma 1, we have with probability at least $1 - \frac{\delta_n}{2}$,

$$\varphi_1(\mathcal{Z}^n) \leq E\big[S(\mathcal{Z}^n, r_n)\big] + \left(\frac{32}{n}\log\frac{4}{\delta_n}\right)^{1/2} \left(M_n + c_{\mathbf{x}}\kappa \left(pM_0\lambda_0^{-1}\right)^{1/2} + c_{\mathbf{x}}^2\kappa M_0\lambda_1^{-1}c_2^{-1} \right)^2.$$

The desired inequality follows immediately after Lemma 2. □

Now we derive the upper bound of $\mathcal{E}(\mathbf{g}^*, \mathbf{H}^*)$. By Assumption 1, for some positive constant b_2 we have

$$\mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) - 2\sigma_s^2 \leq \iint w(\mathbf{x},\mathbf{u}) c_0^2 \| \mathbf{x} - \mathbf{u} \|_2^6 d\rho_{\mathbf{x}} d\rho_{\mathbf{u}} \leq b_2 s^{p+6} \int e^{-\mathbf{t}^T\mathbf{t}} \mathbf{t}^T\mathbf{t} d\,\mathbf{t},$$

where $b_2 = c_0^2 c_4$, $\mathbf{t} = (\mathbf{u} - \mathbf{x})/s$, and the inequalities directly follow from Assumptions 1 and 2.

Lemma 3. *Suppose that the assumptions of Theorem 1 are met. If $|y_n| \leq M_n$ and $\frac{1}{n(n-1)}\sum_{i,j=1}^n (y_i - y_j)^2 \leq M_0$, there exists $b_3 > 0$ such that with probability at least $1 - \delta_n$,*

$$\mathcal{E}(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) + J(\widehat{\mathbf{g}}, \widehat{\mathbf{H}}) - 2\sigma_s^2 \leq b_3 \left(\log\frac{4}{\delta_n}\right)^{1/2} \left(n^{-1/2}M_n^2 + n^{-1/2}M_0\lambda_0^{-1} + n^{-1/2}M_0^2\lambda_1^{-2} + s^{p+6} + \lambda_0 + \lambda_1 \right).$$

Proof of Lemma 3. By Proposition 1, we have

$$\begin{aligned}
\Lambda(\lambda_0, \lambda_1, \mathbf{K}) - 2\sigma_s^2 &= \mathcal{E}(\mathbf{g}^*, \mathbf{H}^*) - 2\sigma_s^2 + \lambda_0 \sum_{l=1}^p \|g_l^*\|_{\mathcal{H}_K}^2 + \lambda_1 \sum_{l,l'}^p \pi_{ll'} \|H_{ll'}^*\|_{\mathcal{H}_K} \\
&\leq b_2 s^{p+6} \int e^{-\mathbf{t}^T\mathbf{t}} \mathbf{t}^T\mathbf{t} d\,\mathbf{t} + \lambda_0 \sum_{l=1}^p \|g_l^*\|_{\mathcal{H}_K}^2 + \lambda_1 \sum_{l,l'}^p \pi_{ll'} \|H_{ll'}^*\|_{\mathcal{H}_K} \\
&\leq b_4(s^{p+6} + \lambda_0 + \lambda_1),
\end{aligned}$$

where $b_4 = \max\{b_2 \int e^{-\mathbf{t}^T\mathbf{t}}\mathbf{t}^T\mathbf{t} d\,\mathbf{t}, \sum_{l=1}^p \|g_l^*\|_{\mathcal{H}_K}^2, \sum_{l,l'=1}^p \pi_{ll'}\|H_{ll'}^*\|_{\mathcal{H}_K}\}$. Following a similar proof of Lemma 1, we have with probability at least $1-\frac{\delta_n}{2}$,

$$\begin{aligned}\varphi_2(\mathcal{Z}^n) &\le \left(\frac{32}{n}\log\frac{4}{\delta_n}\right)^{1/2}\left(M_n + c_{\mathbf{x}}\kappa\sum_{l=1}^p\|g_l^*\|_{\mathcal{H}_K} + c_{\mathbf{x}}^2\kappa\sum_{l,l'=1}^p\|H_{ll'}^*\|_{\mathcal{H}_K}\right)^2\\ &\le 4\left(\frac{32}{n}\log\frac{4}{\delta_n}\right)^{1/2} M_n^2,\end{aligned}$$

where the second inequality follows from the fact that $\|g_l^*\|_{\mathcal{H}_K}$ and $\|H_{ll'}^*\|_{\mathcal{H}_K}$ are smaller than M_n when n is sufficient large. Together with Lemma 2 and Proposition 2, there exists a constant b_5 such that with probability at least $1-\delta_n$,

$$\begin{aligned}\mathcal{E}(\widehat{\mathbf{g}},\widehat{\mathbf{H}}) + J(\widehat{\mathbf{g}},\widehat{\mathbf{H}}) - 2\sigma_s^2 &\le \varphi_1(\mathcal{Z}^n)+\varphi_2(\mathcal{Z}^n)+\Lambda_n(\lambda_0,\lambda_1,\mathbf{K}) - 2\sigma_s^2\\ &\le b_5\left(\log\frac{4}{\delta_n}\right)^{1/2}\left(n^{-1/2}M_n^2 + n^{-1/2}pM_0\lambda_0^{-1} + n^{-1/2}M_0^2\lambda_1^{-2}c_2^{-2} + s^{p+6}+\lambda_0+\lambda_1\right),\end{aligned}$$

which immediately leads to the desired upper bound with some constant b_3. □

Lemma 4. *Suppose that Assumption 1 is met,* $\mathbf{g}^* \in \mathcal{H}_K^p$ *and* $\mathbf{H}^* \in \mathcal{H}_K^{p\times p}$*. There exists some constant* b_6 *such that for any* $\mathbf{g} \in \mathcal{H}_K^p$ *and* $\mathbf{H} \in \mathcal{H}_K^{p\times p}$*,*

$$\int_{\mathcal{X}_s}\|\mathbf{H}(\mathbf{x}) - \mathbf{H}^*(\mathbf{x})\|_F^2 d\rho_{\mathbf{x}} \le b_6\big(s + s^{-(p+5)}(\mathcal{E}(\mathbf{g},\mathbf{H}) - 2\sigma_s^2)\big),$$

where $\mathcal{X}_s = \{\mathbf{x}\in\mathcal{X} : d(\mathbf{x},\partial\mathcal{X}) > s, p(\mathbf{x}) > s + c_1 s^\theta\}$.

Proof of Lemma 4. Let $M_1(\mathbf{x},\mathbf{u}) = f^*(\mathbf{x}) - f^*(\mathbf{u}) - \mathbf{g}^*(\mathbf{x})^T(\mathbf{x}-\mathbf{u}) + \frac{1}{2}(\mathbf{x}-\mathbf{u})^T\mathbf{H}^*(\mathbf{x})(\mathbf{x}-\mathbf{u})$, and $M_2(\mathbf{x},\mathbf{u}) = (\mathbf{g}(\mathbf{x}) - \mathbf{g}^*(\mathbf{x}))^T(\mathbf{x}-\mathbf{u}) - \frac{1}{2}(\mathbf{x}-\mathbf{u})^T(\mathbf{H}(\mathbf{x}) - \mathbf{H}^*(\mathbf{x}))(\mathbf{x}-\mathbf{u})$. Then we have

$$\begin{aligned}\mathcal{E}(\mathbf{g},\mathbf{H}) - 2\sigma_s^2 &= \iint w(\mathbf{x},\mathbf{u})\left(M_1(\mathbf{x},\mathbf{u}) - M_2(\mathbf{x},\mathbf{u})\right)^2 d\rho_{\mathbf{u}} d\rho_{\mathbf{x}}\\ &\ge \iint w(\mathbf{x},\mathbf{u})(M_2(\mathbf{x},\mathbf{u}))^2\rho_{\mathbf{u}}\rho_{\mathbf{x}} - 2\iint w(\mathbf{x},\mathbf{u})M_1(\mathbf{x},\mathbf{u})M_2(\mathbf{x},\mathbf{u})d\rho_{\mathbf{u}}d\rho_{\mathbf{x}}.\end{aligned}$$

By Assumption 1, we have $|M_1(\mathbf{x},\mathbf{u})| \le c_0\|\,\mathbf{x}-\mathbf{u}\,\|^3$, and then

$$\iint w(\mathbf{x},\mathbf{u})(M_1(\mathbf{x},\mathbf{u}))^2 d\rho_{\mathbf{u}}d\rho_{\mathbf{x}} \le \iint w(\mathbf{x},\mathbf{u})c_0^2\|\,\mathbf{x}-\mathbf{u}\,\|^6 d\rho_{\mathbf{u}}d\rho_{\mathbf{x}} \le b_7^2 s^{p+6},$$

for some constant b_7. This inequality, together with Cauchy-Schwarz inequality, yields that

$$\begin{aligned}&\iint w(\mathbf{x},\mathbf{u})M_1(\mathbf{x},\mathbf{u})M_2(\mathbf{x},\mathbf{u})d\rho_{\mathbf{u}}d\rho_{\mathbf{x}}\\ &\le \left(\iint w(\mathbf{x},\mathbf{u})(M_1(\mathbf{x},\mathbf{u}))^2 d\rho_{\mathbf{u}}d\rho_{\mathbf{x}}\right)^{1/2}\left(\iint w(\mathbf{x},\mathbf{u})(M_2(\mathbf{x},\mathbf{u}))^2 d\rho_{\mathbf{u}}d\rho_{\mathbf{x}}\right)^{1/2}\\ &\le b_7 s^{p/2+3}\left(\iint w(\mathbf{x},\mathbf{u})(M_2(\mathbf{x},\mathbf{u}))^2 d\rho_{\mathbf{u}}d\rho_{\mathbf{x}}\right)^{1/2}.\end{aligned}$$

Next, we turn to bound $\iint w(\mathbf{x},\mathbf{u})(M_2(\mathbf{x},\mathbf{u}))^2 d\rho_{\mathbf{u}} d\rho_{\mathbf{x}}$. Specifically,

$$
\begin{aligned}
&Q(\mathbf{g},\mathbf{H}) = \iint w(\mathbf{x},\mathbf{u})(M_2(\mathbf{x},\mathbf{u}))^2 d\rho_{\mathbf{x}} d\rho_{\mathbf{u}} \\
&\geq \frac{1}{4}\int_{\mathcal{X}_s}\int_{\|\mathbf{u}-\mathbf{x}\|<s} w(\mathbf{x},\mathbf{u})\left((\mathbf{x}-\mathbf{u})^T(\mathbf{H}(\mathbf{x})-\mathbf{H}^*(\mathbf{x}))(\mathbf{x}-\mathbf{u})\right)^2 p(\mathbf{u})d\,\mathbf{u}\,d\rho_{\mathbf{x}}+ \\
&\int_{\mathcal{X}_s}\int_{\|\mathbf{u}-\mathbf{x}\|<s} w(\mathbf{x},\mathbf{u})\left((\mathbf{g}(\mathbf{x})-\mathbf{g}^*(\mathbf{x}))^T(\mathbf{x}-\mathbf{u})\right)^2 p(\mathbf{u})d\,\mathbf{u}\,d\rho_{\mathbf{x}}+ \\
&\int_{\mathcal{X}_s}\int_{\|\mathbf{u}-\mathbf{x}\|<s} w(\mathbf{x},\mathbf{u})\left((\mathbf{x}-\mathbf{u})^T(\mathbf{H}^*(\mathbf{x})-\mathbf{H}(\mathbf{x}))(\mathbf{x}-\mathbf{u})\right)\left((\mathbf{g}(\mathbf{x})-\mathbf{g}^*(\mathbf{x}))^T(\mathbf{x}-\mathbf{u})\right) p(\mathbf{u})d\,\mathbf{u}\,d\rho_{\mathbf{x}} \\
&= Q_1(\mathbf{g},\mathbf{H}) + Q_2(\mathbf{g},\mathbf{H}) + Q_3(\mathbf{g},\mathbf{H}).
\end{aligned}
$$

Then we bound $Q_1(\mathbf{g},\mathbf{H})$, $Q_2(\mathbf{g},\mathbf{H})$ and $Q_3(\mathbf{g},\mathbf{H})$ separately. Note that for any $\mathbf{x}\in\mathcal{X}_s$, it is clear that $\{\mathbf{u};\|\mathbf{u}-\mathbf{x}\|<s\}\subset\mathcal{X}$. Moreover, for any $\mathbf{u}\in\{\mathbf{u};\|\mathbf{u}-\mathbf{x}\|<s\}$, Assumption 2 implies that $p(\mathbf{u}) > p(\mathbf{x}) - c_1\|\mathbf{x}-\mathbf{u}\|_2^\theta > s + c_1 s^\theta - c_1 s^\theta = s$. For $Q_1(\mathbf{g},\mathbf{H})$, there exists some constant b_8 such that

$$
\begin{aligned}
Q_1(\mathbf{g},\mathbf{H}) &\geq s^{p+5}\sum_{l,l'=1}^{p}\int_{\mathcal{X}_s}(H_{ll'}(\mathbf{x})-H^*_{ll'}(\mathbf{x}))^2 d\rho_{\mathbf{x}}\int_{\|\mathbf{t}\|<1} e^{-\mathbf{t}^T\mathbf{t}}(t_l t_{l'})^2 d\,\mathbf{t} \\
&\geq b_8 s^{p+5}\sum_{l,l'=1}^{p}\int_{\mathcal{X}_s}(H_{ll'}(\mathbf{x})-H^*_{ll'}(\mathbf{x}))^2 d\rho_{\mathbf{x}} = b_8 s^{p+5}\int_{\mathcal{X}_s}\|\mathbf{H}(\mathbf{x})-\mathbf{H}^*(\mathbf{x})\|_F^2 d\rho_{\mathbf{x}},
\end{aligned}
$$

where the first inequality follows from the fact that $\sum_{l,l',s\neq k}^{p}\int_{\|\mathbf{t}\|<1} e^{-\mathbf{t}^T\mathbf{t}} t_l t_{l'} t_s t_k d\,\mathbf{t} = 0$, and $\|\cdot\|_F$ is the Frobenius norm. Similarly, there exists some constant b_9 such that $Q_2(\mathbf{g},\mathbf{H}) \geq b_9 s^{p+3}\int_{\mathcal{X}_s}\|\mathbf{g}(\mathbf{x})-\mathbf{g}^*(\mathbf{x})\|^2 d\rho_{\mathbf{x}}$ and $Q_3(\mathbf{g},\mathbf{H}) = 0$.

Since $\mathcal{E}(\mathbf{g},\mathbf{H}) - 2\sigma^2 \geq Q(\mathbf{g},\mathbf{H}) - 2b_7 s^{p/2+3}\left(Q(\mathbf{g},\mathbf{H})\right)^{1/2}$, solving the inequality equation yields that $Q(\mathbf{g},\mathbf{H}) \leq b_{10}(s^{6+p} + \mathcal{E}(\mathbf{g},\mathbf{H}) - 2\sigma_s^2)$ for some positive constant b_{10}. As $Q(\mathbf{g},\mathbf{H}) \geq Q_1(\mathbf{g},\mathbf{H}) \geq b_8 s^{p+5}\int_{\mathcal{X}_s}\|\mathbf{H}(\mathbf{x})-\mathbf{H}^*(\mathbf{x})\|_F^2 d\rho_{\mathbf{x}}$, combing these two inequalities yields that

$$
\int_{\mathcal{X}_s}\|\mathbf{H}(\mathbf{x})-\mathbf{H}^*(\mathbf{x})\|_F^2 d\rho_{\mathbf{x}} \leq b_6 s^{-(p+5)}(s^{6+p} + \mathcal{E}(\mathbf{g},\mathbf{H}) - 2\sigma_s^2).
$$

This completes the proof of the desired lemma. □

References

[1] Bartlett, P. and Mendelson, S. (2002). Rademacher and gaussian complexities: risk bounds and structural results. *Journal of Machine Learning Research*, **3**, 463–482.

[2] Combettes, P. and Wajs, V. (2005). Signal recovery by proximal forward-backward splitting. *Multiscale Modeling and Simulations*, **4**, 1168–1200.

[3] MCDIARMID, C. (1989). On the method of bounded differences. *In Surveys in Combinatorics*, 148–188. Cambridge University Press.

[4] ROSASCO, L., MOSCI, S., SANTORO, M., VERRI, A., AND VILLA, S. (2009). Iterative projection methods for structured sparsity regularization. *Computer Science and Artificial Intelligence Laboratory Technical Report*, MIT-CSAIL-TR-2009-050.