Gradient-induced Model-free Variable Selection with Composite Quantile Regression

Xin He[†], Junhui Wang[†] and Shaogao Lv[‡]

[†] *City University of Hong Kong and* [‡] *Zhejiang Gongshang University*

Supplementary Material

This note contains technical proofs.

S1 Technical proofs

We start with proving some preparatory propositions and lemmas.

Proposition 1. Assume $\mathbf{Q}^* \in \mathcal{H}_K^m$. Let $\varphi_1(\mathcal{Z}) = \mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - \mathcal{E}_{\mathcal{Z}}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}),\ \ \varphi_2(\mathcal{Z}) = \mathcal{E}_{\mathcal{Z}}(\mathbf{Q}^*, \mathbf{g}^*) - \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*)$ and $\Lambda_n(\lambda_0, \lambda_1, \mathbf{K}) = \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*) + \frac{\lambda_0}{m}\sum_{k=1}^m \|Q^*_{\tau_k}\|^2_{\mathcal{H}_K} + \lambda_1 \sum_{l=1}^p \pi_l \|\mathbf{g}^{*l}\|_{\mathcal{H}_K^m}$. Then the following inequality holds

$$\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) + \frac{\lambda_0}{m}\sum_{k=1}^m \|\widehat{Q}_{\tau_k}\|^2_{\mathcal{H}_K} + \lambda_1 \sum_{l=1}^p \pi_l \|\widehat{\mathbf{g}}^l\|_{\mathcal{H}_K} \le \varphi_1(\mathcal{Z}) + \varphi_2(\mathcal{Z}) + \Lambda_n(\lambda_0, \lambda_1, \mathbf{K}).$$

Proof of Proposition 1: Since $\mathbf{Q}^* \in \mathcal{H}_k^m,\ \mathbf{g}^{*l} \in \mathcal{H}_K^m,$ for any l (Zhou, 2007). Direct calculation yields that

$$\begin{aligned}
& \mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) + J(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) \\
= \ & \mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - \mathcal{E}_{\mathcal{Z}}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) + \mathcal{E}_{\mathcal{Z}}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) + \frac{\lambda_0}{m}\sum_{k=1}^m \|\widehat{Q}_{\tau_k}\|^2_{\mathcal{H}_K} + \lambda_1 \sum_{l=1}^p \pi_l \|\widehat{\mathbf{g}}^l\|_{\mathcal{H}_K^m} \\
\le \ & \mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - \mathcal{E}_{\mathcal{Z}}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) + \mathcal{E}_{\mathcal{Z}}(\mathbf{Q}^*, \mathbf{g}^*) - \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*) + \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*) \\
& \qquad\qquad\qquad\qquad + \frac{\lambda_0}{m}\sum_{k=1}^m \|Q^*_{\tau_k}\|^2_{\mathcal{H}_K} + \lambda_1 \sum_{l=1}^p \pi_l \|\mathbf{g}^{*l}\|_{\mathcal{H}_K^m} \\
= \ & \varphi_1(\mathcal{Z}) + \varphi_2(\mathcal{Z}) + \Lambda_n(\lambda_0, \lambda_1, \mathbf{K}),
\end{aligned}$$

where the first inequality follows from the definition of $(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}})$. ∎

Correspondence to: Shaogao Lv (lvsg716@swufe.edu.cn)

Recall that

$$\mathcal{F}_{r_n} = \{(\mathbf{Q}, \mathbf{g}) \in \mathcal{H}_K^{m(p+1)} : \ \ \frac{\lambda_0}{m} \sum_{k=1}^{m} \|Q_{\tau_k}\|_{\mathcal{H}_K}^2 \leq r_n, \lambda_1 \sum_{l=1}^{p} \pi_l \|\mathbf{g}_\tau^l\|_{\mathcal{H}_K} \leq r_n\},$$

with r_n defined as in Assumption 5 in the main text. Then denote

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

Now we bound $\mathcal{S}(\mathcal{Z}, r_n)$ using the McDiarmid's inequality.

Lemma 1. *(McDiarmid's Inequality) Let $Z_1, ..., Z_n$ be independent random variables taking values in a set $\mathcal{Z}$, and assume that $\mathbf{f} : \mathcal{Z}^n \to \mathbb{R}$ satisfies*

$$\sup_{z_1,...,z_n,z_i' \in \mathcal{Z}} |\mathbf{f}(z_1, ..., z_n) - \mathbf{f}(z_1, ..., z_i', ..., z_n)| \leq C_i,$$

for every $i \in \{1, 2, ..., n\}$. Then, for every $t > 0$,

$$P\big(|\mathbf{f}(z_1, ..., z_n) - E\left(\mathbf{f}(z_1, ..., z_n)\right)| \geq t\big) \leq 2\exp\left(-\frac{2t^2}{\sum_{i=1}^n C_i^2}\right).$$

The following Lemma 2 can be easily proved using the McDiarmid's Inequality.

Lemma 2. *Supposed Assumptions 1-3 in the main text are met. If $|y| \leq M_n$, then for any r_n and $\varepsilon > 0$, there holds*

$$P(|\mathcal{S}(\mathcal{Z}, r_n) - \mathbb{E}(\mathcal{S}(\mathcal{Z}, r_n))| \geq \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{8\left(M_n + \kappa\sqrt{\frac{r_n}{\lambda_0}} + \frac{c_{\mathbf{x}}\kappa r_n}{c_3\lambda_1}\right)^2}\right). \tag{S1.1}$$

In addition,

$$P(|\mathcal{E}_{\mathcal{Z}}(\mathbf{Q}^*, \mathbf{g}^*) - \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*)| \geq \varepsilon) \leq 2\exp\left(-\frac{n\varepsilon^2}{8\left(M_n + \frac{1}{m}\sum\limits_{k=1}^{m}\|Q^*\|_\infty + \frac{c_{\mathbf{x}}}{m}\sum\limits_{k=1}^{m}\sum\limits_{l=1}^{p}\|g_{\tau_k}^{*l}\|_\infty\right)^2}\right), \tag{S1.2}$$

where $c_{\mathbf{x}} = \max_{\mathbf{x}\in\mathcal{Z}} \| \mathbf{x} \|_\infty$ and $\kappa = \sup_{\mathbf{x}\in\mathcal{Z}} \sqrt{K(\mathbf{x}, \mathbf{x})}$.

Proof of Lemma 2: Let $(\mathbf{x}_i', y_i')$ be a sample point drawn from the distribution $\rho(\mathbf{x}, y)$ and independent of $(\mathbf{x}_i, y_i)$. Denote by $\mathcal{Z}'$ the modified training sample which is the same as $\mathcal{Z}$ except that the i-th entry $(\mathbf{x}_i, y_i)$ is replaced by $(\mathbf{x}_i', y_i')$. By the triangle inequality

$$\begin{aligned}
\mathcal{S}(\mathcal{Z}, r_n) - \mathcal{S}(\mathcal{Z}', r_n) &= \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} |\mathcal{E}_{\mathcal{Z}}(\mathbf{Q},\mathbf{g}) - \mathcal{E}(\mathbf{Q},\mathbf{g})| - \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} |\mathcal{E}_{\mathcal{Z}'}(\mathbf{Q},\mathbf{g}) - \mathcal{E}(\mathbf{Q},\mathbf{g})| \\
&\leq \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} |\mathcal{E}_{\mathcal{Z}}(\mathbf{Q},\mathbf{g}) - \mathcal{E}_{\mathcal{Z}'}(\mathbf{Q},\mathbf{g})|.
\end{aligned}$$

Note that $\mathcal{E}_{\mathcal{Z}}(\mathbf{Q},\mathbf{g})$ can be decomposed as

$$\mathcal{E}_{\mathcal{Z}}(\mathbf{Q},\mathbf{g}) = \frac{1}{mn(n-1)} \sum_{k=1}^{m} \left(\sum_{t\neq i, j\neq i}^{n} h_k(\mathbf{z}_t, \mathbf{z}_j) + \sum_{j=1}^{n} h_k(\mathbf{z}_i, \mathbf{z}_j) + \sum_{t=1}^{n} h_k(\mathbf{z}_t, \mathbf{z}_i) \right),$$

where $h_k(\mathbf{z}_i, \mathbf{z}_j) = w_{ij} L_{\tau_k}(y_i - Q_{\tau_k}(\mathbf{x}_j) - \mathbf{g}_{\tau_k}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j))$ with any fixed $(\mathbf{Q},\mathbf{g}) \in \mathcal{H}_K^{m(p+1)}$. Therefore, for any $(\mathbf{Q},\mathbf{g}) \in \mathcal{F}_{r_n}$, $\mathcal{E}_{\mathcal{Z}}(\mathbf{Q},\mathbf{g}) - \mathcal{E}_{\mathcal{Z}'}(\mathbf{Q},\mathbf{g})$ can be simplified as

$$\begin{aligned}
\mathcal{E}_{\mathcal{Z}}(\mathbf{Q},\mathbf{g}) - \mathcal{E}_{\mathcal{Z}'}(\mathbf{Q},\mathbf{g}) &= \frac{1}{mn(n-1)} \sum_{k=1}^{m} \Big(\sum_{j=1, j\neq i}^{n} h_k(\mathbf{z}_i, \mathbf{z}_j) - \\
&\quad \sum_{j=1, j\neq i'}^{n} h_k(\mathbf{z}_i', \mathbf{z}_j) + \sum_{t=1, t\neq i}^{n} h_k(\mathbf{z}_t, \mathbf{z}_i) - \sum_{t=1, t\neq i'}^{n} h_k(\mathbf{z}_t, \mathbf{z}_i') \Big) \\
&\leq \frac{4}{n} \left(M_n + \kappa \sqrt{\frac{r_n}{\lambda_0}} + \frac{c_{\mathbf{x}} \kappa r_n}{c_3 \lambda_1} \right).
\end{aligned}$$

The last inequality follows from the following derivation,

$$\begin{aligned}
& \frac{1}{mn(n-1)} \sum_{k=1}^{m} \sum_{j=1, j\neq i}^{n} h_k(\mathbf{z}_i, \mathbf{z}_j) \\
\leq\ & \frac{1}{mn(n-1)} \sum_{k=1}^{m} \sum_{j=1, j\neq i}^{n} |y_i - Q_{\tau_k}(\mathbf{x}_j) - \mathbf{g}_{\tau}(\mathbf{x}_i)^T(\mathbf{x}_i - \mathbf{x}_j)| \\
\leq\ & \frac{M_n}{n} + \frac{1}{mn} \sum_{k=1}^{m} \|Q_{\tau_k}\|_\infty + \frac{2c_{\mathbf{x}}}{mn} \sum_{k=1}^{m} \sum_{l=1}^{p} \|g_{\tau_k}^l\|_\infty \\
=\ & \frac{M_n}{n} + \frac{1}{mn} \sum_{k=1}^{m} \sup_{\mathbf{x}\in\mathcal{X}} |\langle Q_{\tau_k}, \mathbf{K}_x \rangle_{\mathcal{H}_K}| + \frac{2c_{\mathbf{x}}}{mn} \sum_{k=1}^{m} \sum_{l=1}^{p} \sup_{\mathbf{x}\in\mathcal{X}} |\langle g_{\tau_k}^l, \mathbf{K}_x \rangle_{\mathcal{H}_K}| \\
\leq\ & \frac{1}{n} \left(M_n + \kappa \sqrt{\frac{r_n}{\lambda_0}} + \frac{2c_{\mathbf{x}} \kappa r_n}{c_3 \lambda_1} \right),
\end{aligned}$$

where the first inequality follows from $L_\tau(u) \le |u|$ and $|w| \le 1$, the second one follows from the triangle inequality, the equality follows from the reproducing property, and the last one is based on Assumptions 1 and 3 in the main text, the Cauchy-Schwartz inequality and the fact that $(\mathbf{Q}, \mathbf{g}) \in \mathcal{F}_{r_n}$.

Interchanging the roles of $\mathcal{Z}$ and $\mathcal{Z}'$ yields that

$$|\mathcal{E}_{\mathcal{Z}}(\mathbf{Q}, \mathbf{g}) - \mathcal{E}_{\mathcal{Z}'}(\mathbf{Q}, \mathbf{g})| \le \frac{4}{n}\left(M_n + \kappa\sqrt{\frac{r_n}{\lambda_0}} + \frac{c_{\mathbf{x}}\kappa r_n}{c_3\lambda_1}\right).$$

Then applying the McDiarmid's inequality, we have the desired probability upper bound in (S1.1). Similarily, the proof of (S1.2) can be obtained by directly applying the McDiarmid's inequality. ■

Lemma 3. *If $|y| \le M_n$, there exists a constant $a_{\kappa,\mathbf{x}}$ such that*

$$\mathbb{E}[S(\mathcal{Z}, r_n)] \le a_{\kappa,\mathbf{x}} \frac{\left(M_n + \sqrt{\frac{r_n}{\lambda_0}} + \frac{r_n}{c_3\lambda_1}\right)}{\sqrt{n}}.$$

Proof of Lemma 3: Denote $\xi_k(\mathbf{x}, y, \mathbf{u}) = w(\mathbf{x}-\mathbf{u})L_{\tau_k}\left(y - Q_{\tau_k}(\mathbf{u}) - \mathbf{g}_{\tau_k}^T(\mathbf{x})(\mathbf{x}-\mathbf{u})\right)$, then

$$\begin{aligned}
& S(\mathcal{Z}, r_n) = \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} |\mathcal{E}_{\mathcal{Z}}(\mathbf{Q}, \mathbf{g}) - \mathcal{E}(\mathbf{Q}, \mathbf{g})| \\
\le\ & \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} \left|\mathcal{E}(\mathbf{Q}, \mathbf{g}) - \frac{1}{mn}\sum_{k=1}^m\sum_{j=1}^n \mathbb{E}\big(\xi_k(\mathbf{x}, y, \mathbf{x}_j)\big)\right| + \left|\frac{1}{mn}\sum_{k=1}^m\sum_{j=1}^n \mathbb{E}\big(\xi_k(\mathbf{x}, y, \mathbf{x}_j)\big) - \mathcal{E}_{\mathcal{Z}}(\mathbf{Q}, \mathbf{g})\right| \\
\le\ & \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} \mathbb{E}_{(\mathbf{x},y)}\left|\frac{1}{m}\mathbb{E}_{\mathbf{u}}\sum_{k=1}^m \xi_k(\mathbf{x}, y, \mathbf{u}) - \frac{1}{mn}\sum_{j=1}^n\sum_{k=1}^m \xi_k(\mathbf{x}, y, \mathbf{x}_j)\right| \\
& + \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} \frac{1}{mn}\sum_{j=1}^n \left|\mathbb{E}_{(\mathbf{x},y)}\big(\sum_{k=1}^m \xi_k(\mathbf{x}, y, \mathbf{x}_j)\big) - \frac{1}{(n-1)}\sum_{i\neq j}\sum_{k=1}^m \xi_k(\mathbf{x}_i, y_i, \mathbf{x}_j)\right| \\
\le\ & \mathbb{E}_{(\mathbf{x},y)} \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} \frac{1}{m}\left|\mathbb{E}_{\mathbf{u}}\big(\sum_{k=1}^m \xi_k(\mathbf{x}, y, \mathbf{u})\big) - \frac{1}{n}\sum_{j=1}^n\sum_{k=1}^m \xi_k(\mathbf{x}, y, \mathbf{x}_j)\right| \\
& + \sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} \frac{1}{mn}\sum_{j=1}^n \sup_{\mathbf{u}\in\mathcal{X}} \left|\mathbb{E}_{(\mathbf{x},y)}\big(\sum_{k=1}^m \xi_k(\mathbf{x}, y, \mathbf{u})\big) - \frac{1}{(n-1)}\sum_{i\neq j}\sum_{k=1}^m \xi_k(\mathbf{x}_i, y_i, \mathbf{u})\right| \\
\stackrel{\text{def}}{=}\ & S_1(\mathcal{Z}) + S_2(\mathcal{Z}),
\end{aligned}$$

where the first inequality follows from the triangle inequality, and the second and third inequalities obtain from the definition of expected error and Jensen's inequality, respectively.

Then, we apply the Rademacher complexities to obtain the upper bounds of $E(S_1)$ and $E(S_2)$ separately. In fact, there holds

$$\begin{aligned}
\mathbb{E}[S_1(\mathcal{Z})] &= \mathbb{E}_{\mathcal{Z}}\mathbb{E}_{(\mathbf{x},\mathbf{y})}\Big(\sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}}\frac{1}{m}\big|\mathbb{E}_{\mathbf{u}}\sum_{k=1}^{m}\xi_k(\mathbf{x},y,\mathbf{u})-\frac{1}{n}\sum_{j=1}^{n}\sum_{k=1}^{m}\xi_k(\mathbf{x},y,\mathbf{x}_j)\big|\Big)\\
&\le \frac{2}{m}\mathbb{E}_{(\mathbf{x},\mathbf{y})}\mathbb{E}_{\mathcal{Z},\sigma}\Big(\sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}}\big|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\sum_{k=1}^{m}\xi_k(\mathbf{x},y,\mathbf{x}_j)\big|\Big)\\
&\le \frac{4}{m}\mathbb{E}_{(\mathbf{x},\mathbf{y})}\mathbb{E}_{\mathcal{Z},\sigma}\Big(\sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}}\big|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\sum_{k=1}^{m}w(\mathbf{x}-\mathbf{x}_j)\left(y-Q_{\tau_k}(\mathbf{x}_j)-\mathbf{g}_{\tau_k}^T(\mathbf{x})(\mathbf{x}-\mathbf{x}_j)\right)\big|\Big)\\
&\le \frac{4}{m}\mathbb{E}_{(\mathbf{x},\mathbf{y})}\mathbb{E}_{\mathcal{Z},\sigma}\Big(\sup_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}}\big|\frac{1}{n}\sum_{j=1}^{n}\sigma_j\sum_{k=1}^{m}w(\mathbf{x}-\mathbf{x}_j)\left(Q_{\tau_k}(\mathbf{x}_j)+\mathbf{g}_{\tau_k}^T(\mathbf{x})(\mathbf{x}-\mathbf{x}_j)\right)\big|\Big)+\frac{4M_n}{\sqrt{n}}\\
&\le a_{\kappa,\mathbf{x}}\frac{\left(M_n+\sqrt{\frac{r_n}{\lambda_0}}+\frac{r_n}{c_3\lambda_1}\right)}{\sqrt{n}},
\end{aligned}$$

where σ_j's are a sequence of Rademacher variables. Here, the first inequality follows from the Rademacher averages, the second, the third and the last inequalities are based on the fact that the absolute function $|\cdot|:\mathcal{R}\to\mathcal{R}$ is Lipschitz and the basic properties of Rademacher complexity (Bartlett, 2002). Similarly, we have

$$\mathbb{E}[S_2(\mathcal{Z})]\le a_{\kappa,x}\frac{\left(M_n+\sqrt{\frac{r_n}{\lambda_0}}+\frac{r_n}{c_3\lambda_1}\right)}{\sqrt{n}}.$$

Then the desired result follows immediately. ■

Proposition 2. If $|y|\le M_n$, there exists a constant a_1, such that with probability at least $1-\frac{\delta}{2}$,

$$\varphi_t(\mathcal{Z})\le a_1\sqrt{\frac{1}{n}\log\frac{4}{\delta}}\Big(M_n+\sqrt{\frac{r_n}{\lambda_0}}+\frac{r_n}{c_3\lambda_1}\Big),\quad \text{for any } \ t=1,2.$$

Proof of Proposition 2: The above proposition can be obtained by using Lemma 2, Lemma 3, and the fact that $\varphi_1(\mathcal{Z})\le S(\mathcal{Z},r_n)$ for any r_n defined above. ■

Now we derive the upper bound of $\mathcal{E}(\mathbf{Q}^*,\mathbf{g}^*)$. Based on the Assumption 1 in the main text, we have

$$\begin{aligned}
\mathcal{E}(\mathbf{Q}^*,\mathbf{g}^*) &\le s_n+\frac{1}{m}\sum_{k=1}^{m}\int\int w(\mathbf{x}-\mathbf{u})|Q^*_{\tau_k}(\mathbf{x})-Q^*_{\tau_k}(\mathbf{u})-\mathbf{g}^{*T}_{\tau_k}(\mathbf{x})(\mathbf{x}-\mathbf{u})|p_{\mathbf{X}}(\mathbf{u})d\,\mathbf{u}\,d\rho_{\mathbf{X}}\\
&\le s_n+\frac{1}{m}\sum_{k=1}^{m}\int\int w(\mathbf{x}-\mathbf{u})c_1\|\,\mathbf{x}-\mathbf{u}\,\|^2p_{\mathbf{X}}(\mathbf{u})p_{\mathbf{X}}(\mathbf{x})d\,\mathbf{u}\,d\,\mathbf{x}\\
&\le s_n+\sigma_n^{p+2}c_1c_5\int e^{-\mathbf{t}^T\mathbf{t}}\mathbf{t}^T\mathbf{t}d\mathbf{t},
\end{aligned}$$

where $\mathbf{t} = \frac{\mathbf{x}-\mathbf{u}}{\sigma_n}$ and $s_n = \frac{1}{m}\sum_{k=1}^{m}\int\int w(\mathbf{x}-\mathbf{u})|y - Q^*_{\tau_k}(\mathbf{x})|d\rho_{\mathbf{X}}d\rho_{(\mathbf{X},Y)}$. Here the first inequality follows from the triangle inequality, and the last two follow from Assumption 1 in the main text.

Proposition 3. Suppose that the assumptions of Theorem 1 are met. If $|y| \leq M_n$, there exists $a_2 > 0$ such that for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) + J(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - s_n \leq a_2\sqrt{\log\frac{4}{\delta}}\Big(n^{-\frac{1}{2}}M_n + \sqrt{\frac{M_n}{\lambda_0 n}} + \frac{M_n}{\sqrt{n}\lambda_1} + \sigma_n^{p+2} + \lambda_0 + \lambda_1\Big).$$

Proof of Proposition 3: Since

$$\mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*) - s_n \leq c_1 c_5 \sigma_n^{p+2}\int e^{-\mathbf{t}^T\mathbf{t}}\mathbf{t}^T\,\mathbf{t}\,d\,\mathbf{t}, \tag{S1.3}$$

we have

$$\begin{aligned}
\Lambda_n(\lambda_0,\lambda_1,\mathbf{K}) - s_n &= \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*) - s_n + \frac{\lambda_0}{m}\sum_{k=1}^{m}\|Q^*_{\tau_k}\|^2_{\mathcal{H}_K} + \lambda_1\sum_{l=1}^{p}\pi_l\|\mathbf{g}^{*l}\|_{\mathcal{H}_K^m} \\
&\leq c_1 c_5 \sigma_n^{p+2}\int e^{-\mathbf{t}^T\mathbf{t}}\mathbf{t}^T\,\mathbf{t}\,d\,\mathbf{t} + \frac{\lambda_0}{m}\sum_{k=1}^{m}\|Q^*_{\tau_k}\|^2_{\mathcal{H}_K} + \lambda_1\sum_{l=1}^{p_0}\pi_l\|\mathbf{g}^{*l}\|_{\mathcal{H}_K^m} \\
&\leq a_3\left(\sigma_n^{p+2} + \lambda_0 + \lambda_1\right),
\end{aligned}$$

where a_3 is a constant large than $\max\{c_1c_5\int e^{-\mathbf{t}^T\mathbf{t}}\mathbf{t}^T\mathbf{t}d\mathbf{t},\ \max_{1\leq k\leq m}\|Q^*_{\tau_k}\|^2_{\mathcal{H}_K},\ \max_{l\leq p_0} c_4\|\mathbf{g}^{*l}\|_{\mathcal{H}_K^m}\}$. Together with Proposition 2 and Lemma 3, there exists a constant a_2 such that with probability at least $1-\delta$

$$\begin{aligned}
\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) + J(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - s_n &\leq \varphi_1(\mathcal{Z}) + \varphi_2(\mathcal{Z}) + \Lambda_n(\lambda_0,\lambda_1,\mathbf{K}) \\
&\leq a_2\sqrt{\log\frac{4}{\delta}}(n^{-\frac{1}{2}}M_n + \sqrt{\frac{r_n}{\lambda_0 n}} + \frac{r_n}{\sqrt{n}\lambda_1} + \sigma_n^{p+2} + \lambda_0 + \lambda_1).
\end{aligned}$$

Furthermore, based on the definition of $(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}})$, we have that

$$\mathcal{E}_{\mathcal{Z}}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) + J(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) \leq \mathcal{E}_{\mathcal{Z}}(\mathbf{0},\mathbf{0}) + J(\mathbf{0},\mathbf{0}) \leq \frac{1}{mn(n-1)}\sum_{k=1}^{m}\sum_{i,j=1}^{n} w_{ij}|y_i| \leq M_n.$$

The desired upper bound can be obtained by setting $r_n = M_n$ in the above inequality. ■

Proof of Theorem 1: For given constant $a_4 > 0$, denote

$$\mathcal{C} = \{\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - s_n \geq a_4\sqrt{\log\frac{4}{\delta}}(n^{-\frac{1}{4}} + n^{-\frac{3}{8}}\lambda_0^{-\frac{1}{2}} + n^{-\frac{1}{4}}\lambda_1^{-1} + \sigma_n^{p+2} + \lambda_0 + \lambda_1)\}. \tag{S1.4}$$

Then we split $\mathcal{C}$ into two different events

$$\begin{aligned} P(\mathcal{C}) &= P(\mathcal{C} \cap \{|y| \geq n^{\frac{1}{4}}\}) + P(\mathcal{C} \cap \{|y| \leq n^{\frac{1}{4}}\}) \\ &\leq P(|y| \geq n^{\frac{1}{4}}) + P(\mathcal{C} \cap \{|y| \leq n^{\frac{1}{4}}\}). \end{aligned}$$

For the first probability, $P(|y| \geq n^{\frac{1}{4}}) = P(|y|^2 \geq n^{\frac{1}{2}}) \leq \frac{E(|y|^2)}{n^{\frac{1}{2}}} = O(n^{-\frac{1}{2}})$. Then we turn to bound the second probability. Within the set $\{|y| \leq n^{\frac{1}{4}}\}$, by Proposition 3, we have with probability at least $1 - \delta$

$$\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - s_n \leq a_3\sqrt{\log\frac{4}{\delta}}(n^{-\frac{1}{4}} + \sqrt{\frac{1}{\lambda_0 n^{\frac{3}{4}}}} + \frac{n^{-\frac{1}{4}}}{\lambda_1} + \sigma_n^{p+2} + \lambda_0 + \lambda_1),$$

where M_n is set as $n^{\frac{1}{4}}$. This implies that $P(\mathcal{C}) \leq \delta$.

Then with the choice of $\lambda_0 = n^{-\frac{1}{4}}$, $\lambda_1 = n^{-\frac{\theta}{2(p+2+2\theta)}}$ and $\sigma_n = n^{-\frac{\theta}{2(p+2+2\theta)}}$, there exists a constant a_4, such that with probability at least $1 - \delta$,

$$\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - s_n \leq a_4\sqrt{\log\frac{4}{\delta}}n^{-\Theta},$$

with $\Theta = \min\{\frac{p+2}{4(p+2+2\theta)}, \frac{\theta}{2(p+2+2\theta)}\}$.Together with (S1.3), there exists a constant c_6 such that with probability at least $1 - \delta$, $|\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*)| \leq c_6\sqrt{\log\frac{4}{\delta}}n^{-\Theta}$. ■

Proof of Theorem 2: First we show that $\sum_{k=1}^m \|\widehat{\mathbf{g}}_{\tau_k}^l\|_1 = 0$ for any $l > p_0$. Note that $\sum_{k=1}^m \|\widehat{\boldsymbol{\alpha}}_k^l\|_1 = 0$ implies that all $\widehat{\boldsymbol{\alpha}}_k^l$ are exactly zero and thus $\sum_{k=1}^m \|\widehat{\mathbf{g}}_{\tau_k}^l\|_1 = 0$ based on the representer theorem in the RKHS. Therefore, it suffices to show that $\sum_{k=1}^m \|\widehat{\boldsymbol{\alpha}}_k^l\|_1 = 0$ for any $l > p_0$.

Suppose $\sum_{k=1}^m \|\widehat{\boldsymbol{\alpha}}_k^l\|_1 > 0$ for some $l > p_0$. As the check loss function is not differentiable at 0, the sub-differential of (1) with respect to $\widehat{\boldsymbol{\alpha}}_k^l$ is

$$\widetilde{R}(\widehat{\boldsymbol{\alpha}}_k^l) = \Big[B_1(\widehat{\boldsymbol{\alpha}}_k^l) + A(\widehat{\boldsymbol{\alpha}}_k^l),\ B_2(\widehat{\boldsymbol{\alpha}}_k^l) + A(\widehat{\boldsymbol{\alpha}}_k^l)\Big],$$

where $A(\widehat{\boldsymbol{\alpha}}_k^l) = \lambda_1 \frac{\pi_l \mathbf{K}\widehat{\boldsymbol{\alpha}}_k^l}{\sqrt{m\sum_{k=1}^m (\widehat{\boldsymbol{\alpha}}_k^l)^T \mathbf{K}\widehat{\boldsymbol{\alpha}}_k^l}}$, $B_1(\widehat{\boldsymbol{\alpha}}_k^l) = \frac{1}{mn(n-1)}\sum_{i,j=1}^n w_{ij}(x_{jl} - x_{il})\mathbf{K}_{x_i}(\tau_k - 1)$ and $B_2(\widehat{\boldsymbol{\alpha}}_k^l) = \frac{1}{mn(n-1)}\sum_{i,j=1}^n w_{ij}(x_{jl} - x_{il})\mathbf{K}_{x_i}\tau_k$.

On one hand, there exists a constant a_5 such that

$$\sum_{k=1}^{m} \|n^{-\frac{1}{2}} A(\widehat{\boldsymbol{\alpha}}_k^l)\|_2 = n^{-\frac{1}{2}}\lambda_1 \frac{\pi_l \sum_{k=1}^m \sqrt{(\widehat{\boldsymbol{\alpha}}_k^l)^T \mathbf{K}^2 \widehat{\boldsymbol{\alpha}}_k^l}}{\sqrt{m \sum_{k=1}^m (\widehat{\boldsymbol{\alpha}}_k^l)^T \mathbf{K} \widehat{\boldsymbol{\alpha}}_k^l}}$$
$$\geq m^{-\frac{1}{2}} n^{-\frac{1}{2}} \lambda_1 a_5 \pi_l \psi_{min} \psi_{max}^{-\frac{1}{2}} \frac{\sum_{k=1}^m \|\widehat{\boldsymbol{\alpha}}_k^l\|_2}{\sqrt{\sum_{k=1}^m \|\widehat{\boldsymbol{\alpha}}_k^l\|_2^2}} \geq a_5 m^{-\frac{1}{2}} n^{-\frac{1}{2}} \lambda_1 \pi_l \psi_{min} \psi_{max}^{-\frac{1}{2}}.$$

By Assumption 4 in the main text, $n^{-\frac{1}{2}}\lambda_1 \pi_l \psi_{min}\psi_{max}^{-\frac{1}{2}} \to \infty$ as n diverges. Therefore, with an appropriately selected m, we can assure that $\|n^{-\frac{1}{2}} A(\widehat{\boldsymbol{\alpha}}_k^l)\|_2 \to \infty$ for some k, and thus at least one component of $A(\widehat{\boldsymbol{\alpha}}_k^l)$ will diverge to infinity.

On the other hand,

$$\sum_{k=1}^{m} |B_1(\widehat{\boldsymbol{\alpha}}_k^l)| = \sum_{k=1}^{m} \Big| \frac{1}{mn(n-1)} \sum_{i,j=1}^{n} w_{ij}(x_{jl} - x_{il}) \mathbf{K}_{\mathbf{x}_i} (\tau_k - 1) \Big| \leq 2 c_{\mathbf{x}} \mathbf{K}_{\mathbf{x}},$$
$$\sum_{k=1}^{m} |B_2(\widehat{\boldsymbol{\alpha}}_k^l)| = \sum_{k=1}^{m} \Big| \frac{1}{mn(n-1)} \sum_{i,j=1}^{n} w_{ij}(x_{jl} - x_{il}) \mathbf{K}_{\mathbf{x}_i} \tau_k \Big| \leq 2 c_{\mathbf{x}} \mathbf{K}_{\mathbf{x}},$$

where the above inequalities between vectors are component-wise. By Assumption 1 in the main text, all elements of $\mathbf{K}_{\mathbf{x}}$ are bounded and thus all components of $B_1(\widehat{\boldsymbol{\alpha}}_k^l)$ and $B_2(\widehat{\boldsymbol{\alpha}}_k^l)$ are also bounded. Combining the above results, it is then obvious that $\mathbf{0} \notin \widetilde{R}(\widehat{\boldsymbol{\alpha}}_k^l)$ for some k, which contradicts with the fact that $\widehat{\boldsymbol{\alpha}}$ is a minimizer of (3) in the main text. Therefore, $\sum_{k=1}^m \|\widehat{\boldsymbol{\alpha}}_l^k\|_1 = 0$ for all $l > p_0$, implying $\sum_{k=1}^m \|\widehat{\mathbf{g}}_{\tau_k}^l\|_1 = 0$ for all $l > p_0$.

Next, we show that $\sum_{k=1}^m \|\widehat{\mathbf{g}}_{\tau_k}^l\|_1 \neq 0$ for every $l \leq p_0$. Suppose $\sum_{k=1}^m \|\widehat{\mathbf{g}}_{\tau_k}^l\|_1 = 0$ for some $l \leq p_0$, then

$$\sum_{k=1}^{m} \int_{\overline{\mathcal{X}}_{\sigma_n}} \left(g_{\tau_k}^{*l}(\mathbf{x})\right)^2 d\rho_X(\mathbf{x}) \leq \sum_{k=1}^{m} \int_{\overline{\mathcal{X}}_{\sigma_n}} \|\widehat{\mathbf{g}}_{\tau_k}(\mathbf{x}) - \mathbf{g}_{\tau_k}^*(\mathbf{x})\|_1^2 d\rho_X(\mathbf{x}) \leq \sum_{k=1}^{m} \|\widehat{\mathbf{g}}_{\tau_k} - \mathbf{g}_{\tau_k}^*\|_1^2, \tag{S1.5}$$

where $\overline{\mathcal{X}}_{\sigma_n} = \{\mathbf{x} \in \mathcal{X} : d(\mathbf{x}, \partial\mathcal{X}) > \sigma_n, p(\mathbf{x}) > \sigma_n + c_5\sigma_n^\theta\}$.

On one hand, by Assumption 5 in the main text and Theorem 1, we have as n diverges,

$$\sum_{k=1}^{m} \|\widehat{\mathbf{g}}_{\tau_k} - \mathbf{g}_{\tau_k}^*\|_1^2 \leq \frac{m}{c_8} \inf_{(\mathbf{Q},\mathbf{g})\in\mathcal{F}_{r_n}} |\mathcal{E}(\widehat{\mathbf{Q}}, \widehat{\mathbf{g}}) - \mathcal{E}(\mathbf{Q}^*, \mathbf{g}^*)| \to 0,$$

with an appropriately selected m. On the other hand, by Assumption 6 in the main text, there exist t and $\tau_0 \in (0,1)$

such that $\int_{\mathcal{X}\setminus\mathcal{X}_t}\left(g^{*l}_{\tau_0}(\mathbf{x})\right)^2 d\rho_X(\mathbf{x}) > 0$, and there exists τ_{k_0} such that

$$\sup_{\mathbf{x},l}|\mathbf{g}^{*l}_{\tau_0}(\mathbf{x}) - \mathbf{g}^{*l}_{\tau_{k_0}}(\mathbf{x})| \le c_9|\tau_0 - \tau_{k_0}|^{\zeta} \to 0,$$

as $m \to \infty$. Therefore, as $\sigma_n \to 0$

$$\sum_{k=1}^{m}\int_{\overline{\mathcal{X}}_{\sigma_n}}\left(g^{*l}_{\tau_k}(\mathbf{x})\right)^2 d\rho_X(\mathbf{x}) \ge \int_{\mathcal{X}\setminus\mathcal{X}_t}\left(g^{*l}_{\tau_{k_0}}(\mathbf{x}) - g^{*l}_{\tau_0}(\mathbf{x})\right)^2 d\rho_X(\mathbf{x}) + \int_{\mathcal{X}\setminus\mathcal{X}_t}\left(g^{*l}_{\tau_0}(\mathbf{x})\right)^2 d\rho_X(\mathbf{x}) + 2\int_{\mathcal{X}\setminus\mathcal{X}_t}\left(g^{*l}_{\tau_{k_0}}(\mathbf{x}) - g^{*l}_{\tau_0}(\mathbf{x})\right) g^{*l}_{\tau_0}(\mathbf{x}) d\rho_X(\mathbf{x}) > 0.$$

Clearly, it contradicts to the inequality in (S1.5), and thus $\sum_{k=1}^{m}\|\widehat{\mathbf{g}}^{l}_{\tau_k}\|_1 \ne 0$ for every $l \le p_0$. Finally, combining the above two results, we show that the proposed method can exactly recover the true active set with probability tending to 1. ■

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] Zhou, D. (2007). Derivative reproducing properties for kernel methods in learning theory. *Journal of Computational and Applied Mathematics*, **220**, 456–463.