Supplemental file for "Efficient learning of nonparametric directed acyclic graph with statistical guarantee"

Yibo Deng†, Xin He†, and Shaogao Lv‡*

† School of Statistics and Management
Shanghai University of Finance and Economics
† School of Statistics and Mathematics
Nanjing Audit University

Technical proofs

Proof of Theorem 1. For any $t = 0, ..., T-1$, given $\mathcal{S}_t$, it is always true that $E\,\mathrm{Var}(x_j|\mathbf{x}_{\mathcal{S}_t}) = E\,\mathrm{Var}(x_j|\mathbf{x}_{\mathrm{pa}_j}) = \sigma_j^2$ for any $j \in \mathcal{A}_t$, due to the fact that $\mathrm{pa}_j \subset \mathcal{S}_t$ if $j \in \mathcal{A}_t$. Moreover, for any $j \in \mathcal{V}\backslash\{\mathcal{S}_t \cup \mathcal{A}_t\}$, by total variance, we have

$$\begin{aligned} E\big[\,\mathrm{Var}(x_j|\mathbf{x}_{\mathcal{S}_t})\big] &= E\big[E[\mathrm{Var}(x_j|\mathbf{x}_{\mathrm{pa}_j})|\mathbf{x}_{\mathcal{S}_t}]\big] + E\big[\,\mathrm{Var}\,\big(E[x_j|\mathbf{x}_{\mathrm{pa}_j}]|\mathbf{x}_{\mathcal{S}_t}\big)\big] \\ &= \sigma_j^2 + E\big[\,\mathrm{Var}\,\big(E[x_j|\mathbf{x}_{\mathrm{pa}_j}]|\mathbf{x}_{\mathcal{S}_t}\big)\big]. \end{aligned}$$

This completes the first part of Theorem 1. Additionally, by Assumption 1 in the main text, for any $j, j' \in \mathcal{A}_t$, we have

$$E\big[\,\mathrm{Var}(x_j|\mathbf{x}_{\mathcal{S}_t})\big] = E\big[\,\mathrm{Var}(x_{j'}|\mathbf{x}_{\mathcal{S}_t})\big] := \sigma_{t,\min}^2,$$

and for any $k \in \mathcal{V}\backslash\{\mathcal{S}_t \cup \mathcal{A}_t\}$, we have

*Shaogao Lv is the corresponding author; the authors contributed equally to this paper and their names are listed in alphabetical ordering.

$$E\big[\,\mathrm{Var}(x_k|\mathbf{x}_{\mathcal{S}_t})\big] = \sigma_k^2 + E\big[\,\mathrm{Var}\,\big(E[x_k|\mathbf{x}_{\mathrm{pa}_k}]|\mathbf{x}_{\mathcal{S}_t}\big)\big] > \sigma_{t,\min}^2 + M_{\max}. \tag{1}$$

Clearly, all the nodes in $\mathcal{A}_t$ can be exactly identified by evaluating the expected conditional variance. This completes the proof. ■

Proof of Theorem 3. Note that the sample variance estimator

$$\widehat{\mathrm{Var}}(x_k) = \frac{1}{n-1}\sum_{i=1}^{n}\Big(x_{ik} - \frac{1}{n}\sum_{j=1}^{n} x_{jk}\Big)^2 = \frac{1}{\binom{n}{2}}\sum_{i<j}\frac{1}{2}(x_{ik}-x_{jk})^2$$

is a U-statistics with kernel $\frac{1}{2}(x_{ik}-x_{jk})^2$. By the definition of $C_{\mathcal{X}}$ that denotes the diameter of the support $\mathcal{X}$, then we have $\frac{1}{2}(x_{ik}-x_{jk})^2 \leq \frac{1}{2}C_{\mathcal{X}}^2$. Then, by McDiarmid's inequality, for any $\zeta > 0$ and $k \in \mathcal{V}$, there holds

$$P\Big(\big|\widehat{\mathrm{Var}}(x_k) - \mathrm{Var}(x_k)\big| > \zeta\Big) \leq 2\exp\Big(-\frac{n\zeta^2}{2C_{\mathcal{X}}^4}\Big). \tag{2}$$

Moreover, we define the following event

$$\mathcal{E}_0 = \Big\{\max_{k\in\mathcal{V}}\big|\widehat{\mathrm{Var}}(x_k) - \mathrm{Var}(x_k)\big| \leq \frac{M_{\max}}{4}\Big\},$$

and use the notation $\mathcal{E}_0^c$ to denote its complementary. By (2), we have

$$P(\mathcal{E}_0^c) \leq 2p\exp\Big(-\frac{nM_{\max}^2}{32C_{\mathcal{X}}^4}\Big). \tag{3}$$

Note that

$$
\begin{aligned}
P(\mathcal{A}_0 \neq \widehat{\mathcal{A}}_0) \leq & P(\mathcal{A}_0 \neq \widehat{\mathcal{A}}_0,\ \mathcal{E}_0) + P(\mathcal{E}_0^c) \\
\leq & P\Big(\exists\, k \in \mathcal{A}_0 \text{ such that} \big|\widehat{\mathrm{Var}}(x_k) - \widehat{\sigma}_{\min}^{(0)}\big| \geq \epsilon_0,\ \mathcal{E}_0\Big) \\
& + P\Big(\exists\, k \in \mathcal{V}\backslash\{\mathcal{A}_0\} \text{ such that} \big|\widehat{\mathrm{Var}}(x_k) - \widehat{\sigma}_{\min}^{(0)}\big| < \epsilon_0,\ \mathcal{E}_0\Big) + P(\mathcal{E}_0^c) \\
= & P_1 + P_2 + P(\mathcal{E}_0^c),
\end{aligned} \tag{4}
$$

where $\widehat{\sigma}_{\min}^{(0)} = \min_{j\in\mathcal{V}} \widehat{\mathrm{Var}}(x_j)$. For ease notation, we denote $k_0 = \mathrm{argmin}_{k\in\mathcal{V}} \widehat{\mathrm{Var}}(x_k)$, and it always holds true that $k_0 \in \mathcal{A}_0$. If not, suppose that $k_0 \in \mathcal{V}\backslash\{\mathcal{A}_0\}$ and for any $j \in \mathcal{A}_0$, under the event $\mathcal{E}_0$ and by Theorem 1 in the main text, we have

$$
\widehat{\mathrm{Var}}(x_{k_0}) > \mathrm{Var}(x_{k_0}) - \frac{M_{\max}}{2} > \mathrm{Var}(x_j) + \frac{M_{\max}}{2} > \widehat{\mathrm{Var}}(x_j),
$$

which contradicts the definition that $k_0 = \mathrm{argmin}_{k\in\mathcal{V}} \widehat{\mathrm{Var}}(x_k)$.

To bound P_1, we notice that under the event $\mathcal{E}_0$, for any $j \in \mathcal{A}_0$, there holds

$$
\begin{aligned}
\big|\widehat{\mathrm{Var}}(x_j) - \widehat{\mathrm{Var}}(x_{k_0})\big| &= \big|\widehat{\mathrm{Var}}(x_j) - \mathrm{Var}(x_j) + \mathrm{Var}(x_j) - \mathrm{Var}(x_{k_0}) + \mathrm{Var}(x_{k_0}) - \widehat{\mathrm{Var}}(x_{k_0})\big| \\
&\leq \big|\widehat{\mathrm{Var}}(x_j) - \mathrm{Var}(x_j)\big| + \big|\mathrm{Var}(x_j) - \mathrm{Var}(x_{k_0})\big| + \big|\mathrm{Var}(x_{k_0}) - \widehat{\mathrm{Var}}(x_{k_0})\big| \\
&\leq \frac{M_{\max}}{4} + 0 + \frac{M_{\max}}{4} = \frac{M_{\max}}{2},
\end{aligned}
$$

where the last inequity follows from Assumption 1 in the main text and the definition of $\mathcal{E}_0$. Thus, by taking $\epsilon_0 = \frac{M_{\max}}{2}$, we have $P_1 = 0$.

Next, we turn to bound P_2. Note that for any $k \in \mathcal{V}\backslash\{\mathcal{A}_0\}$, by Theorem 1 in the main text, there holds

$$
|\mathrm{Var}(x_k) - \mathrm{Var}(x_{k_0})| \geq M_{\max},
$$

and triangle inequality yields that

$$\big|\mathrm{Var}(x_k)-\mathrm{Var}(x_{k_0})\big| \le \big|\mathrm{Var}(x_k)-\widehat{\mathrm{Var}}(x_k)\big| + \big|\widehat{\mathrm{Var}}(x_k)-\widehat{\mathrm{Var}}(x_{k_0})\big| + \big|\widehat{\mathrm{Var}}(x_{k_0})-\mathrm{Var}(x_{k_0})\big|.$$

Then, under the event $\mathcal{E}_0$, we have

$$\begin{aligned}\big|\widehat{\mathrm{Var}}(x_k)-\widehat{\mathrm{Var}}(x_{k_0})\big| &\ge M_{\max} - \big|\mathrm{Var}(x_k)-\widehat{\mathrm{Var}}(x_k)\big| - \big|\widehat{\mathrm{Var}}(x_{k_0})-\mathrm{Var}(x_{k_0})\big| \\ &\ge M_{\max} - \frac{M_{\max}}{4} - \frac{M_{\max}}{4} = \frac{M_{\max}}{2}.\end{aligned}$$

Thus, by taking $\epsilon_0 = \frac{M_{\max}}{2}$, there holds $P_2 = 0$.

Clearly, we have

$$P(\mathcal{A}_0 \neq \widehat{\mathcal{A}}_0) \le P(\mathcal{E}_0^c) \le 2p\exp\Big(-\frac{nM_{\max}^2}{32C_{\mathcal{X}}^4}\Big), \tag{5}$$

by taking $\epsilon_0 = \frac{M_{\max}}{2}$. This completes the proof. ■

Lemma S1. *Suppose that Assumptions 1– 3 in the main text are satisfied. Given the events* $\{\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\}$ *and* $\mathcal{J}$ *and assume* $n\lambda \to \infty$*. Then, with probability at least* $1-\delta_n$*, for any* $j \in \mathcal{V}\backslash\{\mathcal{S}_t\}$*, there holds*

$$\|\widehat{f}_j - f^*_{j,\mathcal{S}_t}\|_\infty \le \frac{\kappa_1 C_{j0}}{\lambda\sqrt{n}}\log\frac{2}{\delta_n} + \kappa_1\lambda^{r-1/2}\|L_{K,t}^{-r} f^*_{j,\mathcal{S}_t}\|_2,$$

where $C_{j0} = 2\kappa_1 \max\Big\{C_{\mathcal{X}} + 2\kappa_1 R, \sqrt{2(2\kappa_1^2 R^2 + \sigma_j^2)}\Big\}$.

Proof of Lemma S1. To begin with, we define the sampling operator $S_{\mathbf{x}_{\mathcal{S}_t}} : \mathcal{H}_K \to \mathcal{R}^n$ associated with some copies of $\mathbf{x}_{\mathcal{S}_t} \in \mathcal{X}_t$ as

$$S_{\mathbf{x}_{\mathcal{S}_t}}(f) = \big(f(\mathbf{x}_{1\mathcal{S}_t}), ..., f(\mathbf{x}_{n\mathcal{S}_t})\big)^T,$$

and the adjoint of the sample operator as $S^T_{\mathbf{x}_{\mathcal{S}_t}} : \mathcal{R}^n \to \mathcal{H}_K$ as

$$S^T_{\mathbf{x}_{\mathcal{S}_t}} \mathbf{c} = \sum_{i=1}^n c_i K_{\mathbf{x}_{i\mathcal{S}_t}},$$

where $\mathbf{c} = (c_1, ..., c_n)^T \in \mathcal{R}^n$. Note that given the events $\{\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\}$ and $\mathcal{J}$, we have $\mathcal{S}_t = \widehat{\mathcal{S}}_t$, and $\|\widehat{f}_j\|_K \leq R$, for any $j \in \mathcal{V}\backslash\{\widehat{\mathcal{S}}_t\}$. Clearly, the solution of (3.2) in Section 3.1 of the main text can be written as

$$\widehat{f}_j = \Big(\frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} S_{\mathbf{x}_{\mathcal{S}_t}} + \lambda \mathbf{I}_n\Big)^{-1} \frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} \mathbf{x}_j,$$

where $\mathbf{x}_j = (x_{1j}, ..., x_{nj})^T$ and $\mathbf{I}_n \in \mathcal{R}^{n\times n}$ denotes the identity matrix.

Moreover, we define an immediate function $f_{\lambda,j}$ as

$$f_{\lambda,j} = \operatorname*{argmin}_{f_j \in \mathcal{H}_K} E\big[x_j - f_j(\mathbf{x}_{\mathcal{S}_t})\big]^2 + \lambda \|f_j\|_K^2. \tag{6}$$

Note that solving (6) equals solving the following problem that

$$f_{\lambda,j} = \operatorname*{argmin}_{f_j \in \mathcal{H}_K} \|f^*_{j,\mathcal{S}_t} - f_j\|^2_{\mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}})} + \lambda \|f_j\|_K^2, \tag{7}$$

by the fact that each node x_j is centered with mean zero and $Ef(\mathbf{x}) = 0$ for all $f \in \mathcal{H}_K$. Thus, the solution of (6) can be derived as

$$f_{\lambda,j} = \big(L_{K,t} + \lambda I\big)^{-1} L_{K,t} f^*_{j,\mathcal{S}_t},$$

where the integral operator $L_{K,t}$ is defined in Section 4 of the main text.

Simple algebra yields that

$$\begin{aligned}\widehat{f}_j - f_{\lambda,j} &= \Big(\frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} S_{\mathbf{x}_{\mathcal{S}_t}} + \lambda \mathbf{I}_n\Big)^{-1} \Big(\frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} \mathbf{x}_j - \frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} S_{\mathbf{x}_{\mathcal{S}_t}} f_{\lambda,j} - \lambda f_{\lambda,j}\Big) \\ &= \Big(\frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} S_{\mathbf{x}_{\mathcal{S}_t}} + \lambda \mathbf{I}_n\Big)^{-1} \Big(\frac{1}{n} \sum_{i=1}^n \big(x_{ij} - f_{\lambda,j}(\mathbf{x}_{i\mathcal{S}_t})\big) K_{\mathbf{x}_{i\mathcal{S}_t}} - L_{K,t}(f^*_{j,\mathcal{S}_t} - f_{\lambda,j})\Big).\end{aligned}$$

Thus, we have

$$\|\widehat{f}_j - f_{\lambda,j}\|_K \leq \frac{1}{\lambda} \Big\| \frac{1}{n} \sum_{i=1}^n \big(x_{ij} - f_{\lambda,j}(\mathbf{x}_{i\mathcal{S}_t})\big) K_{\mathbf{x}_{i\mathcal{S}_t}} - L_{K,t}(f^*_{j,\mathcal{S}_t} - f_{\lambda,j}) \Big\|_K.$$

For notation simplicity, we denote $\xi_i = \big(x_{ij} - f_{\lambda,j}(\mathbf{x}_{i\mathcal{S}_t})\big) K_{\mathbf{x}_{i\mathcal{S}_t}}$, which satisfies

$$\begin{aligned}&E[\xi_i] = L_{K,t}(f^*_{j,\mathcal{S}_t} - f_{\lambda,j}), \ \ \|\xi_i\|_K \leq \frac{\kappa_1}{2}\big(C_{\mathcal{X}} + 2\|f_{\lambda,j}\|_\infty\big) \\ &\text{and} \ \ E[\|\xi_i\|_K^2] \leq \kappa_1^2 \int \big(x_j - f_{\lambda,j}(\mathbf{x}_{\mathcal{S}_t})\big)^2 \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t \cup \{j\}}},\end{aligned}$$

and then by Lemma 2 of Smale and Zhou (2007), for any $\delta_n \in (0,1)$, with probability at least $1 - \delta_n$, we have

$$\|\widehat{f}_j - f_{\lambda,j}\|_K \leq \frac{\kappa_1\big(C_{\mathcal{X}} + 2\|f_{\lambda,j}\|_\infty\big)\log(2/\delta_n)}{\lambda n} + \frac{\kappa_1}{\lambda}\sqrt{\frac{2\int \big(x_j - f_{\lambda,j}(\mathbf{x}_{\mathcal{S}_t})\big)^2 \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t \cup \{j\}}} \log(2/\delta_n)}{n}}.$$

It is clear that by pluging $f_j = 0$ into (7) , we have $\|f_{\lambda,j}\|_K \leq \frac{\kappa_1 \|f^*_{j,\mathcal{S}_t}\|_K}{\lambda^{1/2}}$, and thus we have $\|f_{\lambda,j}\|_\infty \leq \kappa_1 \|f_{\lambda,j}\|_K < \frac{\kappa_1^2 \|f^*_{j,\mathcal{S}_t}\|_K}{\lambda^{1/2}}$. To bound $\int \big(x_j - f_{\lambda,j}(\mathbf{x}_{\mathcal{S}_t})\big)^2 \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t \cup \{j\}}}$, simple calculation yields that for any $f \in \mathcal{H}_K$,

$$\int \big(x_j - f(\mathbf{x}_{\mathcal{S}_t})\big)^2 \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t \cup \{j\}}} - \int \big(x_j - f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t})\big)^2 \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t \cup \{j\}}} = \|f - f^*_{j,\mathcal{S}_t}\|^2_{\mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}})}. \tag{8}$$

By taking $f = 0$, there holds

$$\int \left(x_j - f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t})\right)^2 \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t \cup \{j\}}} + \|0 - f^*_{j,\mathcal{S}_t}\|^2_{\mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}})} = E[f^*_j(\mathbf{x}_{\mathrm{pa}_j}) + n_j]^2 \leq \kappa_1^2 \|f^*_{j,\mathcal{S}_t}\|^2_K + \sigma_j^2,$$

where the last equality follows from the generating scheme of Model 1 in the main text and the last inequality follows from Assumption 3 in the main text. Moreover, we notice that from (7) and by the definition of $f_{\lambda,j}$, there holds

$$\|f^*_{j,\mathcal{S}_t} - f_{\lambda,j}\|^2_{\mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}})} + \lambda \|f_{\lambda,j}\|^2_K \leq \|f^*_{j,\mathcal{S}_t} - 0\|^2_{\mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t}, \rho_{\mathbf{x}_{\mathcal{S}_t}})} + \lambda \|0\|^2_K \leq \kappa_1^2 \|f^*_{j,\mathcal{S}_t}\|^2_K,$$

and then, by plugging $f = f_{\lambda,j}$ into (8), we have

$$\int \left(x_j - f_{\lambda,j}(\mathbf{x}_{\mathcal{S}_t})\right)^2 \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t \cup \{j\}}} = \int \left(x_j - f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t})\right)^2 \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t \cup \{j\}}} + \|f_{\lambda,j} - f^*_{j,\mathcal{S}_t}\|^2_2 \leq 2\kappa_1^2 \|f^*_{j,\mathcal{S}_t}\|^2_K + \sigma_j^2,$$

Therefore, with probability $1 - \delta_n$, we have

$$\begin{aligned} \|\widehat{f}_j - f_{\lambda,j}\|_K &\leq \frac{\kappa_1 (C_{\mathcal{X}} + 2\kappa_1^2 \|f^*_{j,\mathcal{S}_t}\|_K / \lambda^{1/2}) \log(2/\delta_n)}{\lambda n} + \frac{\kappa_1}{\lambda} \sqrt{\frac{2(2\kappa_1^2 \|f^*_{j,\mathcal{S}_t}\|^2_K + \sigma_j^2) \log(2/\delta_n)}{n}} \\ &\leq \frac{C_{j0} \log(2/\delta_n)}{\lambda \sqrt{n}}, \end{aligned} \tag{9}$$

where $C_{j0} = 2\kappa_1 \max\left\{C_{\mathcal{X}} + 2\kappa_1 \|f^*_{j,\mathcal{S}_t}\|_K, \sqrt{2(2\kappa_1^2 \|f^*_{j,\mathcal{S}_t}\|^2_K + \sigma_j^2)}\right\}$ and the last inequality follows from the fact that $n\lambda \to \infty$.

Then, we turn to bound $\|f_{\lambda,j} - f^*_{j,\mathcal{S}_t}\|_K$ following similar treatments as in Smale and Zhou (2005). Specifically, for the integral operator $L_{K,t}$ defined in Section 4 of the main text with normalized eigenpairs $\{(\mu_k, \phi_k)\}_{k=1}^\infty$, we have

$$L_{K,t}^{1/2} \phi_i = \sum_{j \geq 1} \mu_j^{1/2} \langle \phi_i, \phi_j \rangle_2 \phi_j = \mu_i^{1/2} \phi_i \in \mathcal{H}_K,$$

and

$$\|\mu_i^{1/2}\phi_i\|_K = \Big(\sum_{j\geq 1}\frac{\langle \mu_i^{1/2}\phi_i,\phi_j\rangle_2^2}{\mu_j}\Big)^{1/2} = \langle\phi_i,\phi_i\rangle_2 = 1.$$

Thus by Assumption 2 of the main text, there exists some function $h_{j,t} = \sum_{i\geq 1}\langle h_{j,t},\phi_i\rangle_2\phi_i \in \mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t},\rho_{\mathbf{x}_{\mathcal{S}_t}})$ such that $f^*_{j,\mathcal{S}_t} = L^r_{K,t}h_{j,t} = \sum_{i\geq 1}\mu_i^r\langle h_{j,t},\phi_i\rangle_2\phi_i \in \mathcal{H}_K$.

Therefore, we have

$$f_{\lambda,j} - f^*_{j,\mathcal{S}_t} = \big(L_{K,t}+\lambda I\big)^{-1}\big(-\lambda f^*_{j,\mathcal{S}_t}\big) = -\sum_{i\geq 1}\frac{\lambda}{\lambda+\mu_i}\mu_i^r\langle h_{j,t},\phi_i\rangle_2\phi_i,$$

and then, the K-norm of $f_{\lambda,j} - f^*_{j,\mathcal{S}_t}$ can be bounded as

$$\begin{aligned}\big\|f_{\lambda,j} - f^*_{j,\mathcal{S}_t}\big\|_K^2 &= \sum\nolimits_{i\geq 1}\Big(\frac{\lambda}{\lambda+\mu_i}\mu_i^{r-1/2}\langle h_{j,t},\phi_i\rangle_2\Big)^2\\ &= \lambda^{2r-1}\sum\nolimits_{i\geq 1}\Big(\frac{\lambda}{\lambda+\mu_i}\Big)^{3-2r}\Big(\frac{\mu_i}{\lambda+\mu_i}\Big)^{2r-1}\langle h_{j,t},\phi_i\rangle_2^2\\ &\leq \lambda^{2r-1}\sum\nolimits_{i\geq 1}\langle h_{j,t},\phi_i\rangle_2^2 = \lambda^{2r-1}\|h_{j,t}\|_2^2 = \lambda^{2r-1}\|L_{K,t}^{-r}f^*_{j,\mathcal{S}_t}\|_2^2. \end{aligned} \tag{10}$$

Combining (9) and (10), under the events $\{\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\}$ and $\mathcal{J}$, with probability at least $1-\delta_n$, we have

$$\begin{aligned}\|\widehat{f}_j - f^*_{j,\mathcal{S}_t}\|_K &\leq \|\widehat{f}_j - f_{\lambda,j}\|_K + \|f_{\lambda,j} - f^*_{j,\mathcal{S}_t}\|_K\\ &\leq \frac{C_{j0}}{\lambda\sqrt{n}}\log\frac{2}{\delta_n} + \lambda^{r-1/2}\|L_{K,t}^{-r}f^*_{j,\mathcal{S}_t}\|_2.\end{aligned}$$

Moreover, we notice that $\|\widehat{f}_j - f^*_{j,\mathcal{S}_t}\|_\infty \leq \kappa_1\|\widehat{f}_j - f^*_{j,\mathcal{S}_t}\|_K$ by the reproducing property and the requirement that $\|f^*_{j,\mathcal{S}_t}\|_K^2 \leq R/2$ in Section 4 of the main text. This completes the proof. ■

Proof of Theorem 4. Given the event $\{\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\}$, we have $\mathcal{S}_t = \widehat{\mathcal{S}}_t$. Then, for any $j \in \mathcal{V}\backslash\{\mathcal{S}_t\}$, there holds

$$
\begin{aligned}
&\left|\widehat{E\mathrm{Var}}(x_j|\mathbf{x}_{\mathcal{S}_t}) - E\,\mathrm{Var}(x_j|\mathbf{x}_{\mathcal{S}_t})\right| \\
&= \Big|\frac{1}{n}\sum_{i=1}^{n}(x_{ij})^2 - \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{f}_j(\mathbf{x}_{i\mathcal{S}_t})\big)^2 - E[x_j^2] + E\big[E[x_j|\mathbf{x}_{\mathcal{S}_t}]^2\big]\Big| \\
&\leq \Big|E[x_j^2] - \frac{1}{n}\sum_{i=1}^{n}(x_{ij})^2\Big| + \Big|E\big[E[x_j|\mathbf{x}_{\mathcal{S}_t}]^2\big] - \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{f}_j(\mathbf{x}_{i\mathcal{S}_t})\big)^2\Big|. \qquad (11)
\end{aligned}
$$

To bound the first term of (11), we notice that each x_j is required to be centered with mean zero in Section 2 of the main text, which implies that zero belong to the support $\mathcal{X}$, and then x_{ij}^2 are bounded by $\frac{C_{\mathcal{X}}^2}{4}$ from the definition of $C_{\mathcal{X}}$, which denotes the diameter of the support $\mathcal{X}$. Then by the Hoeffding's inequality, for any $\frac{\zeta}{2} > 0$, there holds

$$
P\Big(\Big|E[x_j^2] - \frac{1}{n}\sum_{i=1}^{n}(x_{ij})^2\Big| > \frac{\zeta}{2}\Big) \leq 2\exp\Big(-\frac{8n\zeta^2}{C_{\mathcal{X}}^4}\Big). \qquad (12)
$$

Next, the second term of (11) can be decomposed as

$$
\begin{aligned}
&\Big|E\big[E[x_j|\mathbf{x}_{\mathcal{S}_t}]^2\big] - \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{f}_j(\mathbf{x}_{i\mathcal{S}_t})\big)^2\Big| \\
&\leq \Big|E\big[E[x_j|\mathbf{x}_{\mathcal{S}_t}]^2 - \widehat{f}_j^2(\mathbf{x}_{\mathcal{S}_t})\big]\Big| + \Big|E\big[\widehat{f}_j^2(\mathbf{x}_{\mathcal{S}_t})\big] - \frac{1}{n}\sum_{i=1}^{n}\big(\widehat{f}_j(\mathbf{x}_{i\mathcal{S}_t})\big)^2\Big| \\
&= \Delta_1 + \Delta_2,
\end{aligned}
$$

and thus it suffices to bound Δ_1 and Δ_2 sequentially under the events $\{\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}\}$ and $\mathcal{J}$.

To bound Δ_1, we notice that

$$\begin{aligned}\Delta_1 &= \Big| E\big[f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t})\big(f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t}) - \widehat{f}_j(\mathbf{x}_{\mathcal{S}_t})\big)\big] + E\big[\widehat{f}_j(\mathbf{x}_{\mathcal{S}_t})\big(f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t}) - \widehat{f}_j(\mathbf{x}_{\mathcal{S}_t})\big)\big]\Big| \\
&\le \|f^*_{j,\mathcal{S}_t} - \widehat{f}_j\|_\infty \Big| \int |f^*_{j,\mathcal{S}_t}(\mathbf{x}_{\mathcal{S}_t})| d\rho_{\mathbf{x}_{\mathcal{S}_t}} + \int |\widehat{f}_j(\mathbf{x}_{\mathcal{S}_t})| d\rho_{\mathbf{x}_{\mathcal{S}_t}} \Big| \\
&\le 2\kappa_1 \max\{\|\widehat{f}_j\|_K, \|f^*_{j,\mathcal{S}_t}\|_K\} \|f^*_{j,\mathcal{S}_t} - \widehat{f}_j\|_\infty \le 2\kappa_1 R \|f^*_{j,\mathcal{S}_t} - \widehat{f}_j\|_\infty,\end{aligned}$$

where the last inequality follows from the reproducing property of RKHS, the requirement that $\|f^*_{j,\mathcal{S}_t}\|_K \le R/2$ in Section 4 of the main text and and under the event $\mathcal{J}$ that $\|\widehat{f}_j\|_K \le R$. Then, by Lemma S1, with probability at least $1 - \delta_n/2$, we have

$$\Delta_1 \le 2\kappa_1^2 R \Big(\frac{C_{j0}}{\lambda\sqrt{n}} \log \frac{4}{\delta_n} + \lambda^{r-1/2} \|L^{-r}_{K,t} f^*_{j,\mathcal{S}_t}\|_2 \Big). \tag{13}$$

To bound Δ_2, we notice that

$$\begin{aligned}\Delta_2 &= \Big| \int \widehat{f}_j(\mathbf{x}_{\mathcal{S}_t}) \langle \widehat{f}_j, K_{\mathbf{x}_{\mathcal{S}_t}} \rangle_K \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t}} - \frac{1}{n} \sum_{i=1}^n \widehat{f}_j(\mathbf{x}_{i\mathcal{S}_t}) \langle \widehat{f}_j, K_{\mathbf{x}_{i\mathcal{S}_t}} \rangle_K \Big| \\
&= \Big| \Big\langle \widehat{f}_j, \int \widehat{f}_j(\mathbf{x}_{\mathcal{S}_t}) K_{\mathbf{x}_{\mathcal{S}_t}} \mathrm{d}\rho_{\mathbf{x}_{\mathcal{S}_t}} \Big\rangle_K - \frac{1}{n} \Big\langle \widehat{f}_j, S^T_{\mathbf{x}_{\mathcal{S}_t}} S_{\mathbf{x}_{\mathcal{S}_t}} \widehat{f}_j \Big\rangle_K \Big| \\
&= \Big| \Big\langle \widehat{f}_j, \int \widehat{f}_j(\mathbf{x}_{\mathcal{S}_t}) K_{\mathbf{x}_{\mathcal{S}_t}} d\rho_{\mathbf{x}_{\mathcal{S}_t}} - \frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} S_{\mathbf{x}_{\mathcal{S}_t}} \widehat{f}_j \Big\rangle_K \Big| \\
&= \Big| \Big\langle \widehat{f}_j, \big(L_{K,t} - \frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} S_{\mathbf{x}_{\mathcal{S}_t}}\big) \widehat{f}_j \Big\rangle_K \Big| \le \|\widehat{f}_j\|_K \big\| L_{K,t} - \frac{1}{n} S^T_{\mathbf{x}_{\mathcal{S}_t}} S_{\mathbf{x}_{\mathcal{S}_t}} \big\|_{HS},\end{aligned}$$

where $S^T_{\mathbf{x}_{\mathcal{S}_t}}$ and $S_{\mathbf{x}_{\mathcal{S}_t}}$ denote the sampling operators defined in Lemma S1, and $\|\cdot\|_{HS}$ denotes the norm endowed with a Hilbert space $HS(K)$ containing all the Hilbert-Schmidt operators on $\mathcal{H}_K$ and satisfying $\|T\|_K \le \|T\|_{HS}$ for any $T \in HS(K)$. Note that under the event $\mathcal{J}$, we have $\|\widehat{f}_j\|_K \le R$. Moreover, by Lemma 18 of Rosasco et al. (2013), with probability at least $1 - \delta_n/2$, we have

$$\Big\|L_{K,t} - \frac{1}{n} S_{\mathbf{x}_{\mathcal{S}_t}}^T S_{\mathbf{x}_{\mathcal{S}_t}}\Big\|_{HS} \le \frac{2\sqrt{2}\kappa_1^2}{\sqrt{n}} \log\frac{4}{\delta_n}.$$

Clearly, with probability at least $1-\delta_n/2$, we have $\Delta_2 \le \frac{2R\sqrt{2}\kappa_1^2}{\sqrt{n}} \log\frac{4}{\delta_n}$.

Combining the upper bounds of Δ_1 and Δ_2, with probability at least $1-\delta_n$, there holds

$$\begin{aligned}
&\Big|E\big[E[x_j|\mathbf{x}_{\mathcal{S}_t}]^2\big] - \frac{1}{n}\sum_{i=1}^n \big(\widehat{f}_j(\mathbf{x}_{i\mathcal{S}_t})\big)^2\Big| \\
&\le 2\kappa_1^2 R\Big(\frac{C_{j0}}{\lambda\sqrt{n}}\log\frac{4}{\delta_n} + \lambda^{r-1/2}\|L_{K,t}^{-r} f_{j,\mathcal{S}_t}^*\|_2\Big) + \frac{2R\sqrt{2}\kappa_1^2}{\sqrt{n}}\log\frac{4}{\delta_n} \\
&\le 2\kappa_1^2 R\Big(\frac{C_{j0}+\sqrt{2}}{\lambda\sqrt{n}}\log\frac{4}{\delta_n} + \lambda^{r-1/2}\|L_{K,t}^{-r} f_{j,\mathcal{S}_t}^*\|_2\Big).
\end{aligned}$$

Then, by taking $\lambda = n^{-\frac{1}{2r+1}}$, for any $\delta_n \in (0,1)$, with probability at least $1-\delta_n$ there holds

$$\Big|E\big[E[x_j|\mathbf{x}_{\mathcal{S}_t}]^2\big] - \frac{1}{n}\sum_{i=1}^n \big(\widehat{f}_j(\mathbf{x}_{i\mathcal{S}_t})\big)^2\Big| \le C_{jt} n^{-\frac{2r-1}{2(2r+1)}} \log\frac{4}{\delta_n},$$

where $C_{jt} = 6\kappa_1^2 R \max\big\{C_{j0},\ \sqrt{2},\ \|L_{K,t}^{-r} f_{j,\mathcal{S}_t}^*\|_2\big\}$. Correspondingly, for any $\zeta > 0$, we have

$$\begin{aligned}
&P\Big(\Big|E\big[E[x_j|\mathbf{x}_{\mathcal{S}_t}]^2\big] - \frac{1}{n}\sum_{i=1}^n \big(\widehat{f}_j(\mathbf{x}_{i\mathcal{S}_t})\big)^2\Big| > \frac{\zeta}{2} \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}\Big) \\
&\le 4\exp\Big(-\frac{\zeta}{2C_{jt}} n^{\frac{2r-1}{2(2r+1)}}\Big). \qquad (14)
\end{aligned}$$

Combining (12) and (14), for any $\zeta > 0$, there holds

$$\begin{aligned}
&P\Big(\Big|E\mathrm{Var}(x_j|\mathbf{x}_{\mathcal{S}_t}) - \widehat{E}\widehat{\mathrm{Var}}(x_j|\mathbf{x}_{\mathcal{S}_t})\Big| > \zeta \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}\Big) \\
&\le 2\exp\Big(-\frac{8n\zeta^2}{C_{\mathcal{X}}^4}\Big) + 4\exp\Big(-\frac{\zeta}{2C_{jt}} n^{\frac{2r-1}{2(2r+1)}}\Big). \qquad (15)
\end{aligned}$$

This completes the proof of the first part of Theorem 4.

Next, we define the following event

$$\mathcal{E}_{1t} = \Big\{ \max_{j \in \mathcal{V} \backslash \{\mathcal{S}_t\}} \big| E\text{Var}(x_j|\mathbf{x}_{\mathcal{S}_t}) - \widehat{E\widehat{\text{Var}}}(x_j|\mathbf{x}_{\mathcal{S}_t}) \big| \leq \frac{M_{\max}}{4} \Big\},$$

and use the notation $\mathcal{E}_{1t}^c$ to denote its complementary. Then, by (15), we have

$$\begin{aligned} & P(\mathcal{E}_{1t}^c \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) \\ & \leq 2(p - |\mathcal{S}_t|) \exp\Big(- \frac{nM_{\max}^2}{2C_{\mathcal{X}}^4} \Big) + 4(p - |\mathcal{S}_t|) \exp\Big(- \frac{M_{\max} n^{\frac{2r-1}{2(2r+1)}}}{8C_{jt}} \Big). \end{aligned} \tag{16}$$

Note that

$$\begin{aligned} & P(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) \\ & \leq P(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t,\ \mathcal{E}_{1t} \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) + P(\mathcal{E}_{1t}^c \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) \\ & \leq P\big(\exists\, j \in \mathcal{A}_t \text{ such that} |\widehat{E\widehat{\text{Var}}}(x_j|\mathbf{x}_{\mathcal{S}_t}) - \widehat{\sigma}_{\min}^{(t)}| \geq \epsilon_t,\ \mathcal{E}_{1t} \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}\big) \\ & \quad + P\big(\exists\, j \in \mathcal{V}\backslash\{\mathcal{S}_t \cup \mathcal{A}_t\} \text{ such that } |\widehat{E\widehat{\text{Var}}}(x_j|\mathbf{x}_{\mathcal{S}_t}) - \widehat{\sigma}_{\min}^{(t)}| < \epsilon_t,\ \mathcal{E}_{1t} \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}\big) \\ & \quad + P(\mathcal{E}_{1t}^c \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) \\ & = P_3 + P_4 + P(\mathcal{E}_{1t}^c \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}). \end{aligned} \tag{17}$$

Note that following the similar treatments as that of P_1 and P_2 in the proof of Theorem 3 in the main text and by taking $\epsilon_t = \frac{M_{\max}}{2}$, we have $P_3 = 0$ and $P_4 = 0$. Finally, the bound (17) reduces to

$$\begin{aligned} & P(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) \\ & \leq P(\mathcal{E}_{1t}^c \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) \\ & \leq 2(p - |\mathcal{S}_t|) \exp\Big(- \frac{nM_{\max}^2}{2C_{\mathcal{X}}^4} \Big) + 4(p - |\mathcal{S}_t|) \exp\Big(- \frac{M_{\max} n^{\frac{2r-1}{2(2r+1)}}}{8C_{jt}} \Big). \end{aligned}$$

This completes the proof. ∎

Proof of Lemma 1. Given the event $\{\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_t = \mathcal{A}_t\}$, $t \geq 1$, we have $\mathcal{S}_t = \widehat{\mathcal{S}}_t$. At first, for some positive constant C_3, we define the following event

$$\mathcal{E}_{2t} = \Big\{ \max_{j\in\mathcal{A}_t, k\in\mathcal{S}_t} \big|\|\widehat{g}_{jk}\|_n^2 - \|g_{jk}^*\|_2^2\big| \leq C_3 n^{-\frac{2r-1}{2(2r+1)}} \log\big(4|\mathcal{S}_t|\max\{n, |\mathcal{S}_t|\}\big) \Big\},$$

and use the notation $\mathcal{E}_{2t}^c$ to denote its complementary.

We notice that

$$\begin{aligned} &P\Big(\big\{\mathcal{E}_j \neq \widehat{\mathcal{E}}_j : j \in \widehat{\mathcal{A}}_t\big\} \mid \mathcal{A}_0 = \widehat{\mathcal{A}}_0, ..., \mathcal{A}_t = \widehat{\mathcal{A}}_t, \mathcal{J}\Big) \\ &\leq P\Big(\big\{\mathcal{E}_j \neq \widehat{\mathcal{E}}_j : j \in \widehat{\mathcal{A}}_t\big\}, \mathcal{E}_{2t} \mid \mathcal{A}_0 = \widehat{\mathcal{A}}_0, ..., \mathcal{A}_t = \widehat{\mathcal{A}}_t, \mathcal{J}\Big) \\ &\quad + P\big(\mathcal{E}_{2t}^c \mid \mathcal{A}_0 = \widehat{\mathcal{A}}_0, ..., \mathcal{A}_t = \widehat{\mathcal{A}}_t, \mathcal{J}\big). \end{aligned} \tag{18}$$

Note that by the definition that $\widehat{\mathcal{E}}_j = \big\{k \to j, \|\widehat{g}_{jk}\|_n^2 > \upsilon_n^{(t)}, \text{for any } k \in \widehat{\mathcal{S}}_t\big\}$ and by Assumption 4 of the main text, for the first term of (18), there holds

$$\begin{aligned} &P\Big(\big\{\mathcal{E}_j \neq \widehat{\mathcal{E}}_j : j \in \widehat{\mathcal{A}}_t\big\}, \mathcal{E}_{2t} \mid \mathcal{A}_0 = \widehat{\mathcal{A}}_0, ..., \mathcal{A}_t = \widehat{\mathcal{A}}_t, \mathcal{J}\Big) \\ \leq &P\Big(\exists\, k \in \mathrm{pa}_j \text{ such that } \|\widehat{g}_{jk}\|_n^2 \leq \nu_n^{(t)}, \mathcal{E}_{2t} \mid \mathcal{A}_0 = \widehat{\mathcal{A}}_0, ..., \mathcal{A}_t = \widehat{\mathcal{A}}_t, \mathcal{J}\Big) \\ &\quad + P\Big(\exists\, k \in \widehat{\mathrm{pa}}_j \text{ such that } \|g_{jk}^*\|_2^2 = 0, \mathcal{E}_{2t} \mid \mathcal{A}_0 = \widehat{\mathcal{A}}_0, ..., \mathcal{A}_t = \widehat{\mathcal{A}}_t, \mathcal{J}\Big) \\ = &P_5 + P_6, \end{aligned} \tag{19}$$

where $\widehat{\mathrm{pa}}_j = \{k : k \to j \in \widehat{\mathcal{E}}_j\}$.

For the bound of P_5, by taking $\upsilon_n^{(t)} = \frac{C_2}{2} n^{-\frac{2r-1}{2(2r+1)}} \Big(\log\big(4|\mathcal{S}_t|\max\{n, |\mathcal{S}_t|\}\big)\Big)^{\beta}$ and Assumption 3 in the main text, we have

$$\big|\|\widehat{g}_{jk}\|_n^2 - \|g_{jk}^*\|_2^2\big| \geq \|g_{jk}^*\|_2^2 - \|\widehat{g}_{jk}\|_n^2 > 2\nu_n^{(t)} - \nu_n^{(t)} = \nu_n^{(t)},$$

which contradicts with $\mathcal{E}_{2t}$. Precisely, under $\mathcal{E}_{2t}$, we have

$$\max_{j\in\mathcal{A}_t,k\in\mathcal{S}_t}\left|\|\widehat{g}_{jk}\|_n^2-\|g^*_{jk}\|_2^2\right|\le C_3 n^{-\frac{2r-1}{2(2r+1)}}\log\big(4|\mathcal{S}_t|\max\{n,|\mathcal{S}_t|\}\big),$$

and then the different rates of convergence lead to the contradiction. Thus, when n is sufficiently large, $P_5=0$. To bound P_6, it is obvious that $\left|\|\widehat{g}_{jk}\|_n^2-\|g^*_{jk}\|_2^2\right|>\nu_n^{(t)}$, which contradicts with $\mathcal{E}_{2t}$ again, which yields that $P_6=0$.

Now, we turn to bound $P\big(\mathcal{E}^c_{2t}\mid\mathcal{A}_0=\widehat{\mathcal{A}}_0,...,\mathcal{A}_t=\widehat{\mathcal{A}}_t,\mathcal{J}\big)$. At first, we define the sample operators for gradients $\widehat{D}_{t,l}:\mathcal{H}_K\to\mathcal{R}^n$ as $(\widehat{D}_{t,l}f)_i=\big\langle f,\partial_l K_{\mathbf{x}_{i\mathcal{S}_t}}\big\rangle_K$, their adjoint operators $\widehat{D}^*_{t,l}:\mathcal{R}^n\to\mathcal{H}_K$ as $\widehat{D}^*_{t,l}\mathbf{c}=\frac{1}{n}\sum_{i=1}^n\partial_l K_{\mathbf{x}_{i\mathcal{S}_t}}c_i$, and define the integral operators for gradients $D_{t,l}:\mathcal{H}_K\to\mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t},\rho_{\mathbf{x}_{\mathcal{S}_t}})$ as $D_{t,l}f=\langle f,\partial_l K_{\mathbf{x}_{\mathcal{S}_t}}\rangle_K$, $D^*_{t,l}:\mathcal{L}^2(\mathcal{X}_{\mathcal{S}_t},\rho_{\mathbf{x}_{\mathcal{S}_t}})\to\mathcal{H}_K$ as $D^*_{t,l}f=\int\partial_l K_{\mathbf{x}_{\mathcal{S}_t}}f(\mathbf{x}_{\mathcal{S}_t})d\rho_{\mathbf{x}_{\mathcal{S}_t}}$. Then, we have

$$D^*_{t,l}D_{t,l}f^*_{j,\mathcal{S}_t}=\int\partial_l K_{\mathbf{x}_{\mathcal{S}_t}}g^*_{jl}(\mathbf{x}_{\mathcal{S}_t})d\rho_{\mathbf{x}_{\mathcal{S}_t}}\ \text{ and }\ \widehat{D}^*_{t,l}\widehat{D}_{t,l}f^*_{j,\mathcal{S}_t}=\frac{1}{n}\sum_{i=1}^n\partial_l K_{\mathbf{x}_{i\mathcal{S}_t}}g^*_{jl}(\mathbf{x}_{i\mathcal{S}_t}).$$

Note that $D^*_{t,l}D_{t,l}$ and $\widehat{D}^*_{t,l}\widehat{D}_{t,l}$ are the Hilbert-Schmidt operators belonging to a Hilbert space endowed with norm $\|\cdot\|_{HS}$.

Moreover, we notice that for any $j\in\mathcal{A}_t$ and $k\in\mathcal{S}_t$

$$\begin{aligned}\left|\|\widehat{g}_{jk}\|_n^2-\|g^*_{jk}\|_2^2\right|&=\Big|\frac{1}{n}\sum_{i=1}^n\big(\widehat{g}_{jk}(\mathbf{x}_{i\mathcal{S}_t})\big)^2-\int\big(g^*_{jk}(\mathbf{x}_{\mathcal{S}_t})\big)^2d\rho_{\mathbf{x}_{\mathcal{S}_t}}\Big|\\&=\Big|\big\langle\widehat{f}_j,\frac{1}{n}\sum_{i=1}^n\widehat{g}_{jk}(\mathbf{x}_{i\mathcal{S}_t})\partial_k K_{\mathbf{x}_{i\mathcal{S}_t}}\big\rangle_K-\big\langle f^*_{j,\mathcal{S}_t},\int g^*_{jk}(\mathbf{x}_{\mathcal{S}_t})\partial_k K_{\mathbf{x}_{\mathcal{S}_t}}d\rho_{\mathbf{x}_{\mathcal{S}_t}}\big\rangle_K\Big|\\&=\Big|\big\langle\widehat{f}_j-f^*_{j,\mathcal{S}_t},\widehat{D}^*_{t,k}\widehat{D}_{t,k}(\widehat{f}_j-f^*_{j,\mathcal{S}_t})\big\rangle_K+\langle\widehat{D}^*_{t,k}\widehat{D}_{t,k}f^*_{j,\mathcal{S}_t},\widehat{f}_j-f^*_{j,\mathcal{S}_t}\rangle_K\\&\quad+\langle f^*_{j,\mathcal{S}_t},\widehat{D}^*_{t,k}\widehat{D}_{t,k}(\widehat{f}_j-f^*_{j,\mathcal{S}_t})\rangle_K+\big\langle f^*_{j,\mathcal{S}_t},(\widehat{D}^*_{t,k}\widehat{D}_{t,k}-D^*_{t,k}D_{t,k})f^*_{j,\mathcal{S}_t}\big\rangle_K\Big|\\&\le\|\widehat{f}_j-f^*_{j,\mathcal{S}_t}\|_K^2\|\widehat{D}^*_{t,k}\widehat{D}_{t,k}\|_{HS}+2\|\widehat{f}_j-f^*_{j,\mathcal{S}_t}\|_K\|f^*_{j,\mathcal{S}_t}\|_K\|\widehat{D}^*_{t,k}\widehat{D}_{t,k}\|_{HS}\\&\quad+\|\widehat{D}^*_{t,k}\widehat{D}_{t,k}-D^*_{t,k}D_{t,k}\|_{HS}\|f^*_{j,\mathcal{S}_t}\|_K^2.\end{aligned}$$

By Assumption 3 in the main text, direct calculation yields that

$$\max_{k\in\mathcal{S}_t}\|\widehat{D}_{t,k}^*\widehat{D}_{t,k}\|_{HS}=\max_{k\in\mathcal{S}_t}\|\partial_k K_{\mathbf{x}_{\mathcal{S}_t}}\|_K^2\leq\kappa_2^2.$$

Moreover, by Lemma 18 of Rosasco et al. (2013), for any $\delta_n\in(0,1)$, with probability at least $1-\delta_n/2$, we have

$$\max_{k\in\mathcal{S}_t}\|\widehat{D}_{t,k}^*\widehat{D}_{t,k}-D_{t,k}^*D_{t,k}\|_{HS}\leq\frac{2\sqrt{2}\kappa_2^2}{\sqrt{n}}\log\frac{4|\mathcal{S}_t|}{\delta_n},$$

given $\{\widehat{\mathcal{A}}_0=\mathcal{A}_0,...,\widehat{\mathcal{A}}_t=\mathcal{A}_t\}$. Thus, by taking $\delta_n=(\max\{n,|\mathcal{S}_t|\})^{-1}$, with probability at least $1-\delta_n/2$, there holds

$$\max_{k\in\mathcal{S}_t}\|\widehat{D}_k^*\widehat{D}_k-D_k^*D_k\|_{HS}\leq\frac{2\sqrt{2}\kappa_2^2}{\sqrt{n}}\log\big(4|\mathcal{S}_t|\max\{n,|\mathcal{S}_t|\}\big).$$

When $\|\widehat{f}_j-f_{j,\mathcal{S}_t}^*\|_K$ is sufficiently small and by taking $\lambda=n^{-\frac{1}{2r+1}}$, with probability at least $1-\delta_n$, we have

$$\begin{aligned}
&\max_{j\in\mathcal{A}_t,k\in\mathcal{S}_t}\big|\|\widehat{g}_{jk}\|_n^2-\|g_{jk}\|_2^2\big|\\
&\leq\max\{\kappa_2^2,\kappa_2^2\|f_{j,\mathcal{S}_t}^*\|_K,\|f_{j,\mathcal{S}_t}^*\|_K^2\}\Big(3\max_{j\in\mathcal{A}_t,k\in\mathcal{S}_t}\|\widehat{f}_j-f_{j,\mathcal{S}_t}^*\|_K+\max_{k\in\mathcal{S}_t}\|\widehat{D}_k^*\widehat{D}_k-D_k^*D_k\|_{HS}\Big)\\
&\leq\max\{\kappa_2^2,\kappa_2^2\|f_{j,\mathcal{S}_t}^*\|_K,\|f_{j,\mathcal{S}_t}^*\|_K^2\}\Big(3\max_{j\in\mathcal{A}_t,k\in\mathcal{S}_t}\{C_{j0}+\|L_{K,t}^{-r}f_{j,\mathcal{S}_t}^*\|_2\}n^{-\frac{2r-1}{2(2r+1)}}\log\frac{4|\mathcal{S}_t|}{\delta_n}+\frac{2\sqrt{2}\kappa_2^2}{\sqrt{n}}\log\frac{4|\mathcal{S}_t|}{\delta_n}\Big)\\
&\leq C_3n^{-\frac{2r-1}{2(2r+1)}}\log\big(4|\mathcal{S}_t|\max\{n,|\mathcal{S}_t|\}\big),
\end{aligned}$$

where $C_3=3\max\{\kappa_2^2,\kappa_2^2\|f_{j,\mathcal{S}_t}^*\|_K,\|f_{j,\mathcal{S}_t}^*\|_K^2\}\max_{j\in\mathcal{A}_t,k\in\mathcal{S}_t}\big\{3C_{j0},3\|L_{K,t}^{-r}f_{j,\mathcal{S}_t}^*\|_2,2\sqrt{2}\kappa_2^2\big\}$. Thus, we have

$$P\big(\mathcal{E}_{2t}^c|\widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_t = \mathcal{A}_t, \mathcal{J}\big) \leq \frac{1}{\max\{n, |\mathcal{S}_t|\}},$$

Finally, by (18) and (19), we have

$$P\Big(\{\mathcal{E}_j = \widehat{\mathcal{E}}_j : j \in \widehat{\mathcal{A}}_t\} \mid \mathcal{A}_0 = \widehat{\mathcal{A}}_0, ..., \mathcal{A}_t = \widehat{\mathcal{A}}_t, \mathcal{J}\Big) \geq 1 - \frac{1}{\max\{n, |\mathcal{S}_t|\}}.$$

This completes the proof. ■

Proof for Theorem 5. Note that

$$P(\widehat{\mathcal{G}} \neq \mathcal{G}) = P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J}) + P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J}^c).$$

For $P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J}^c)$, we have

$$P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J}^c) \leq P(\mathcal{J}^c) \leq Tp \max_{1\leq t\leq T-1,\ j\in\mathcal{V}\backslash\{\mathcal{S}_t\}} P\big(\|\widehat{f}_j\|_K > R\big). \tag{20}$$

Note that by Theorem 1 and Lemma 3 of Smale and Zhou (2007), for any t and $j \in \mathcal{V}\backslash\{\mathcal{S}_t\}$, we have $P\big(\|\widehat{f}_j\|_K > R\big) \leq \frac{1}{n}$ if the sample size satisfies $n \geq \Big(\frac{C_4}{R}\log 2n\Big)^{\frac{2(2r+1)}{2r-1}}$ for some positive constant C_4 and the K-norm of the target function is upper bounded by $R/2$ as assumed in Section 4 of the main text, and thus $P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J}^c) \to 0$ as $n \to \infty$.

For $P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J})$, we have

$$
\begin{aligned}
P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J}) &\leq P\big(\cup_{t=0}^{T-1} \{\widehat{\mathcal{A}}_t \neq \mathcal{A}_t\} \cup \{\widehat{\mathcal{E}} \neq \mathcal{E}\}, \mathcal{J}\big) \\
&\leq P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0) + \sum_{t=1}^{T-1} P(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) + \\
&\qquad P(\widehat{\mathcal{E}} \neq \mathcal{E} \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{T-1} = \mathcal{A}_{T-1}, \mathcal{J}) \\
&\leq P(\widehat{\mathcal{A}}_0 \neq \mathcal{A}_0) + \sum_{t=1}^{T-1} P(\widehat{\mathcal{A}}_t \neq \mathcal{A}_t \mid \widehat{\mathcal{A}}_0 = \mathcal{A}_0, ..., \widehat{\mathcal{A}}_{t-1} = \mathcal{A}_{t-1}, \mathcal{J}) + \\
&\qquad \sum_{t=1}^{T-1} P\Big(\big\{\mathcal{E}_j = \widehat{\mathcal{E}}_j : j \in \widehat{\mathcal{A}}_t\big\} \mid \mathcal{A}_0 = \widehat{\mathcal{A}}_0, ..., \mathcal{A}_t = \widehat{\mathcal{A}}_t, \mathcal{J}\Big).
\end{aligned}
$$

Clearly, combining with Theorem 4 and Lemma 1 in the main text, we have $P(\widehat{\mathcal{G}} \neq \mathcal{G}, \mathcal{J}) \to 0$ as $n \to \infty$. This completes the proof. ■

References

Rosasco, L., S. Villa, S. Mosci, M. Santoro, and A. Verri (2013). Nonparametric sparsity and regularization. *Journal of Machine Learning Research* ***14***, 1665–1714.

Smale, S. and D. Zhou (2005). Shannon sampling ii: Connections to learning theory. *Applied and Computational Harmonic Analysis* ***19(3)***, 285–302.

Smale, S. and D. Zhou (2007). Learning theory estimates via integral operators and their approximations. *Constructive Approximation* ***26(2)***, 153–172.