Supplementary Appendix for Kernel Ordinary Differential Equations

XIAOWU DAI AND LEXIN LI

University of California at Berkeley

S1 Proofs

S1.1 Proof of Theorem 1

Denote the KODE objective in (10) by $\mathcal{A}(F_j)$:

$$\mathcal{A}(F_j) \equiv \frac{1}{n}\sum_{i=1}^{n}\left\{y_{ij}-\theta_{j0}-\int_0^{t_i}F_j(\widehat{x}(t))dt\right\}^2+\tau_{nj}\left(\sum_{k=1}^{p}\|\mathcal{P}^kF_j\|_{\mathcal{H}}+\sum_{k\neq l,k=1}^{p}\sum_{l=1}^{p}\|\mathcal{P}^{kl}F_j\|_{\mathcal{H}}\right).$$

Without loss of generality, let $\tau_{nj}=1$. Write $\mathcal{H}=\mathcal{H}^{(0)}\oplus\mathcal{H}^{(1)}$, where $\mathcal{H}^{(0)}\equiv\{1\}$ and $\mathcal{H}^{(1)}\equiv\sum_{k=1}^{p}\mathcal{H}_k\oplus\sum_{k\neq l,k=1}^{p}\sum_{l=1}^{p}[\mathcal{H}_k\otimes\mathcal{H}_l]$, where for any $F_j\in\mathcal{H}$ (Wahba et al., 1995),

$$\|F_j\|_{\mathcal{H}}^2=\|F_j\|_{\mathcal{H}^{(0)}}^2+\|F_j\|_{\mathcal{H}^{(1)}}^2,\quad\text{and}\quad\|F_j\|_{\mathcal{H}^{(1)}}^2=\sum_{k=1}^{p}\|\mathcal{P}^kF_j\|_{\mathcal{H}}^2+\sum_{k\neq l,k=1}^{p}\sum_{l=1}^{p}\|\mathcal{P}^{kl}F_j\|_{\mathcal{H}}^2.$$

Note that,

$$\begin{aligned}&\frac{p(p+1)}{2}\left(\sum_{k=1}^{p}\|\mathcal{P}^kF_j\|_{\mathcal{H}}^2+\sum_{k\neq l,k=1}^{p}\sum_{l=1}^{p}\|\mathcal{P}^{kl}F_j\|_{\mathcal{H}}^2\right)\\&\qquad\geq J_2^2(F_j)\geq\sum_{k=1}^{p}\|\mathcal{P}^kF_j\|_{\mathcal{H}}^2+\sum_{k\neq l,k=1}^{p}\sum_{l=1}^{p}\|\mathcal{P}^{kl}F_j\|_{\mathcal{H}}^2.\end{aligned}$$

Henceforth, for any $F_j\in\mathcal{H}^{(1)}$,

$$J_2(F_j)\geq\|F_j\|_{\mathcal{H}}. \tag{S1}$$

We next show the existence of the minimizer in three cases.

First, denote $\rho_j=\max_{i=1}^{n}(y_{ij}^2+|y_{ij}|+1)$. Let $K(\cdot,\cdot)$ be the reproducing kernel of $\mathcal{H}^{(1)}$, and let $\langle\cdot,\cdot\rangle_{\mathcal{H}^{(1)}}$ be the inner product in $\mathcal{H}^{(1)}$. Write $a=\sup_{t\in\mathcal{T}}K^{1/2}(\widehat{x}(t),\widehat{x}(t))$, where $\widehat{x}$ is obtained from (9). Consider the set

$$\mathcal{D}_j=\left\{F_j\in\mathcal{H}:F_j=b_j+F_j^{(1)},b_j\in\{1\},F_j^{(1)}\in\mathcal{H}^{(1)},J_2(F_j)\leq\rho_j,\ |b_j|\leq\rho^{1/2}+(a+1)\rho_j\right\}.$$

Then $\mathcal{D}_j$ is a closed and convex compact set. Note that both $J_2(F_j)$ and the functional $n^{-1}\sum_{i=1}^n \left\{ y_{ij} - \theta_{j0} - \int_0^{t_i} F_j(\widehat{x}(u))du \right\}^2$ are convex in F_j, and thus $\mathcal{A}(F_j)$ is convex. Therefore, there exists a minimizer of the convex optimization problem (10) in the convex set $\mathcal{D}_j$. Denote the minimizer by $\widehat{F}_j \in \mathcal{D}_j$. Then $\mathcal{A}(\widehat{F}_j) \le \mathcal{A}(0) \le n^{-1}\sum_{i=1}^n y_{ij}^2 < \rho_j$.

Second, for any $F_j \in \mathcal{H}$ with $J(F_j) > \rho_j$, then $F_j \notin \mathcal{D}_j$. However, $\mathcal{A}(F_j) \ge J(F_j) > \rho_j$, which implies that $\mathcal{A}(F_j) > \mathcal{A}(\widehat{F}_j)$.

Third, for any $F_j \in \mathcal{H}$ with $J_2(F_j) \le \rho_j$, $F_j = b_j + F_j^{(1)}$ with $b_j \in \{1\}$, $F_j^{(1)} \in \mathcal{H}^{(1)}$, and $|b_j| > \rho_j^{1/2} + (a+1)\rho_j$. By the reproducing property, for any $F_j^{(1)} \in \mathcal{H}^{(1)}$ and $t \in \mathcal{T}$,

$$
\begin{aligned}
\left| F_j^{(1)}(\widehat{x}(t)) \right| &= \left| \left\langle F_j^{(1)}(\cdot), K(\widehat{x}(t), \cdot) \right\rangle_{\mathcal{H}^{(1)}} \right| \le \left\| F_j^{(1)} \right\|_{\mathcal{H}^{(1)}} \langle K(\widehat{x}(t), \cdot), K(\widehat{x}(t), \cdot) \rangle_{\mathcal{H}^{(1)}}^{1/2} \\
&= \left\| F_j^{(1)} \right\|_{\mathcal{H}^{(1)}} K^{1/2}\left(\widehat{x}(t), \widehat{x}(t)\right) \le a J_2\left(F_j^{(1)}\right),
\end{aligned}
$$

where the last step is by (S1) and the definition of a. Hence, for any $i = 1, \ldots, n$, $t_i \in \mathcal{T}$,

$$
\begin{aligned}
\min_{C_{j0}} \left| C_{j0} + b_j t_i + \int_0^{t_i} F_j^{(1)}(\widehat{x}(u))du - y_{ij} \right| &\ge \left| b_j t_i + \int_0^{t_i} F_j^{(1)}(\widehat{x}(u))du - y_{ij} \right| \\
&> [\rho_j^{1/2} + (a+1)\rho_j] - a\rho_j - \rho_j = \rho_j^{1/2}.
\end{aligned}
$$

Therefore, $\mathcal{A}(F_j) > \rho$, and $\mathcal{A}(F_j) > \mathcal{A}(\widehat{F}_j)$. Consequently, for any $F_j \notin \mathcal{D}_j$, $\mathcal{A}(F_j) > \mathcal{A}(\widehat{F}_j)$, and $\widehat{F}_j$ is a minimizer of (10) in $\mathcal{H}$.

Next, we show that the minimizer $\widehat{F}_j$ is in a finite-dimensional space. Let $K_k(\cdot,\cdot)$ be the reproducing kernel of $\mathcal{H}_k$. Then $K_{kl} \equiv K_k K_l$ is the reproducing kernel of $\mathcal{H}_k \otimes \mathcal{H}_l$ (Aronszajn, 1950). Write $\widehat{F}_j = \widehat{b}_j + \sum_{k=1}^p \widehat{F}_{jk} + \sum_{k \ne l} \widehat{F}_{jkl}$, where $\widehat{F}_{jk} \in \mathcal{H}_k$, and $\widehat{F}_{jkl} \in \mathcal{H}_k \otimes \mathcal{H}_l$. Write $T_i(t) = \mathbf{1}\{0 \le t \le t_i\}$, and $\bar{T}(t) = n^{-1}\sum_{i=1}^n T_i(t)$. We have $\int_{\mathcal{T}} K(\widehat{x}(s), \widehat{x}(t)) T_i(t) dt \in \mathcal{H}$ (Cucker and Smale, 2002). Besides,

$$
\begin{aligned}
\left\langle \int_{\mathcal{T}} K(\widehat{x}(s), \widehat{x}(t)) T_i(t) dt,\ F_j(\widehat{x}(s)) \right\rangle_{\mathcal{H}} &= \int_{\mathcal{T}} \langle K(\widehat{x}(s), x(t)), F_j(\widehat{x}(s)) \rangle_{\mathcal{H}} T_i(t) dt \\
&= \int_{\mathcal{T}} F_j(\widehat{x}(t)) T_i(t) dt.
\end{aligned}
$$

Denote the projection of $\widehat{F}_{jk}$ onto the finitely spanned space

$$
\left\{ \int_{\mathcal{T}} K_k(\widehat{x}_k(t), \cdot) T_i(t) dt, i = 1, \ldots, n \right\} \subset \mathcal{H}_k
$$

as $\widehat{g}_{jk}$, and its orthogonal complement in $\mathcal{H}_k$ as $\widehat{h}_{jk}$. Similarly, denote the projection of $\widehat{F}_{jkl}$ onto the finitely spanned space

$$
\left\{ \int_{\mathcal{T}} K_k(\widehat{x}_k(t), \cdot) K_l(\widehat{x}_l(t), \cdot) T_i(t) dt, i = 1, \ldots, n \right\} \subset \mathcal{H}_k \otimes \mathcal{H}_l
$$

as $\widehat{g}_{kl}$, and its orthogonal complement in $\mathcal{H}_k \otimes \mathcal{H}_l$ as $\widehat{h}_{kl}$. Then $\widehat{F}_{jk} = \widehat{g}_{jk} + \widehat{h}_{jk}$, and $\widehat{F}_{jkl} = \widehat{g}_{jkl} + \widehat{h}_{jkl}$. Besides, $\|\widehat{F}_{jk}\|_{\mathcal{H}}^2 = \|\widehat{g}_{jk}\|_{\mathcal{H}}^2 + \|\widehat{h}_{jk}\|_{\mathcal{H}}^2$, and $\|\widehat{F}_{jkl}\|_{\mathcal{H}}^2 = \|\widehat{g}_{jkl}\|_{\mathcal{H}}^2 + \|\widehat{h}_{jkl}\|_{\mathcal{H}}^2$, for $k, l = 1, \dots, p, k \neq l$. Since $K = 1 + \sum_{k=1}^p K_k + \sum_{k \neq l} K_{kl}$ is the reproducing kernel of $\mathcal{H}$, by the orthogonal structure,

$$\begin{aligned}
\int_{\mathcal{T}} \widehat{F}_j(\widehat{x}(t)) T_i(t) dt = \Bigg\langle \int_{\mathcal{T}} \Bigg\{ 1 + \sum_{k=1}^p K_k(\widehat{x}_k(t), \cdot) + \sum_{k \neq l} K_k(\widehat{x}_k(t), \cdot) K_l(\widehat{x}_l(t), \cdot) \Bigg\} T_i(t) dt, \\
b_j + \sum_{k=1}^p \Big\{ \widehat{g}_{jk}(\widehat{x}_k(t)) + \widehat{h}_{jk}(\widehat{x}_k(t)) \Big\} + \sum_{k \neq l} \Big\{ \widehat{g}_{jkl}(\widehat{x}_k(t), \widehat{x}_l(t)) + \widehat{h}_{jkl}(\widehat{x}_k(t), \widehat{x}_l(t)) \Big\} \Bigg\rangle_{\mathcal{H}} \\
= b_j \int_{\mathcal{T}} T_i(t) dt + \sum_{k=1}^p \Big\langle \int_{\mathcal{T}} K_k(\widehat{x}_k(t), \cdot) T_i(t) dt,\ \widehat{g}_{jk}(\widehat{x}_k(t)) \Big\rangle_{\mathcal{H}} \\
+ \sum_{k \neq l} \Big\langle \int_{\mathcal{T}} K_k(\widehat{x}_k(t), \cdot) K_l(\widehat{x}_l(t), \cdot) T_i(t) dt,\ \widehat{g}_{jkl}(\widehat{x}_k(t), \widehat{x}_l(t)) \Big\rangle_{\mathcal{H}}.
\end{aligned}$$

Recall $\bar{y}_j = n^{-1} \sum_{i=1}^n y_{ij}$. Therefore, (10) can be written as

$$\begin{aligned}
\frac{1}{n} \sum_{i=1}^n \Bigg\{ (y_{ij} - \bar{y}_j) - b_j \int_{\mathcal{T}} [T_i(t) - \bar{T}(t)] dt \\
- \sum_{k=1}^p \Big\langle \int_{\mathcal{T}} K_k\left(\widehat{x}_k(s), \widehat{x}_k(t)\right) [T_i(t) - \bar{T}(t)] dt,\ \widehat{g}_{jk}(\widehat{x}_k(s)) \Big\rangle_{\mathcal{H}} \\
- \sum_{k \neq l} \Big\langle \int_{\mathcal{T}} K_k\left(\widehat{x}_k(s), \widehat{x}_k(t)\right) K_l\left(\widehat{x}_l(s), \widehat{x}_l(t)\right) [T_i(t) - \bar{T}(t)] dt,\ \widehat{g}_{jkl}\left(\widehat{x}_k(s), \widehat{x}_l(s)\right) \Big\rangle_{\mathcal{H}} \Bigg\}^2 \\
+ \tau_{nj} \Bigg\{ \sum_{k=1}^p (\|\widehat{g}_{jk}(\widehat{x}_k)\|_{\mathcal{H}}^2 + \|\widehat{h}_{jk}(\widehat{x}_k)\|_{\mathcal{H}}^2)^{1/2} + \sum_{k \neq l} (\|\widehat{g}_{jkl}(\widehat{x}_k, \widehat{x}_l)\|_{\mathcal{H}}^2 + \|\widehat{h}_{jkl}(\widehat{x}_k, \widehat{x}_l)\|_{\mathcal{H}}^2)^{1/2} \Bigg\}.
\end{aligned}$$

Therefore, the minimizer $\widehat{F}_j$ of (10) satisfies that $\widehat{h}_{jk} = \widehat{h}_{jkl} = 0$, for any $k, l = 1, \dots, p$ and $k \neq l$. This completes the proof of Theorem 1. □

S1.2 Proof of Theorem 2

We first prove that $c_0(x)$ does not depend on the true but unknown functional F_j. Consider

$$\begin{aligned}
\widetilde{F}_j(x) &= \theta_{j0} + \rho_j^{1/2} \mathcal{B}(x), \quad x = x(t), \\
Y_{ij} &= \mathcal{L}_i \widetilde{F}_j(x) + \epsilon_{ij}, \quad t \in \mathcal{T}.
\end{aligned}$$

where $\theta_{j0} \sim \mathcal{N}(0, aI)$ and $\epsilon_{ij} \sim \mathcal{N}(0, \sigma_j^2)$. The parameter $\rho_j = \sigma_j^2 / n\eta_{nj}$. The stochastic process $\mathcal{B}(\cdot)$ is a zero-mean Gaussian process with covariance $K_{\widehat{\theta}_{M_j}} = \sum_{k=1}^p \widehat{\theta}_{jk} K_k +$

$\sum_{k\neq l}\widehat{\theta}_{jkl}K_{kl}$. The bounded operator takes the form: $L_i\widetilde{F}_j(x) \equiv \int_{\mathcal{T}}\left\{T_i(t)-\bar{T}(t)\right\}\widetilde{F}_j(x(t))dt$, for any $\widetilde{F}_j \in \mathcal{H}$. It is shown that (Wahba, 1990),

$$\lim_{a\to\infty}\mathbb{E}\left\{\widetilde{F}_j(x)|Y_{ij}=y_{ij}-\bar{y}_j, i=1,\ldots,n\right\}=\widehat{F}_{j,\widehat{\theta}_{M_j}}(x),$$

and the covariance matrix of $\left(\mathcal{L}_1\widehat{F}_j,\ldots,\mathcal{L}_n\widehat{F}_j\right)$ is $\mathrm{Cov}(\mathcal{L}_1\widehat{F}_j,\ldots,\mathcal{L}_n\widehat{F}_j)=\sigma_j^2A_{M_j}$, where A_{M_j} is the smoothing matrix as defined in (14) with the kernel corresponding to $\widehat{\theta}_{M_j}$ (Wahba, 1983; Silverman, 1985). Consequently, the collection of all the quantities $A_{M_j}(y_j-\bar{y}_j)$ are jointly distributed as $N(0,\sigma_j^2A_{M_j})$, where A_{M_j} is independent of F_j. Henceforth, the joint distribution of the collection of ratios $|A_{M_j}(y_j-\bar{y}_j)|/\sigma_j$ is independent of F_j.

Next, we prove the coverage property. Observe that, for any $i=1,\ldots,n$,

$$\widehat{F}_{j,\widehat{\theta}_{M_j}}(\widehat{x}(t_i))-\mathbb{E}\left\{\widehat{F}_{j,\widehat{\theta}_{M_j}}(\widehat{x}(t_i))\right\}=\{A_{M_j}\}_{i\cdot}(y_j-\bar{y}_j).$$

We then have the following upper bound,

$$\left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j-\bar{y}_j)\right|/\sigma_j \le \max_{M_j\subseteq\mathcal{M}}\left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j-\bar{y}_j)\right|/\sigma_j.$$

By the choice of $c_0(\widehat{x})$ in (16), the coverage property holds.

Lastly, we show that there exists a unique $c_0(\widehat{x}_j(t_i))$ satisfying (16). Consider the maximum statistic, $\max_{M_j\subseteq\mathcal{M}}\left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j-\bar{y}_j)\right|/\sigma_j$, with the corresponding distribution $H(t)=\mathbb{P}\left[\max_{M_j\subseteq\mathcal{M}}\left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j-\bar{y}_j)\right|/\sigma_j\le t\right]$. We show that $H(t)=0$ for $t\le 0$, is continuous on $\mathbb{R}$, and is strictly increasing in $t\ge 0$.

Note that, for $t<0$, the event $\left\{\max_{M_j\subseteq\mathcal{M}}\left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j-\bar{y}_j)\right|/\sigma_j\le t\right\}$ is empty. For $t=0$, this event is an intersection of the sets $\left\{\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j-\bar{y}_j)=0\right\}$ for any $M_j\subseteq\mathcal{M}$, where at least one of these sets has a probability zero, given $y_j\neq\bar{y}_j$. Henceforth, $H(t)=0$ for $t\le 0$. To prove the continuity of H on $(0,\infty)$, we note that, for any $M_j\subseteq\mathcal{M}$ and $t\ge 0$, $\mathbb{P}\left[\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j-\bar{y}_j)/\sigma_j=t\right]=0$, since y_j is a continuous variable. Finally, we show the strict monotonicity that $H(t_1)<H(t_2)$ for any $0<t_1<t_2$. Toward that goal, suppose there exists $\widetilde{y}_j\in\mathbb{R}^n$ such that $\max_{M_j\subseteq\mathcal{M}}\left|\{\widetilde{A}_{M_j}\}_{i\cdot}\widetilde{y}_j\right|/\sigma_j=t_1$. There exists $M_j^\ddagger\subseteq\mathcal{M}$, such that $\{\widetilde{A}_{M_j^\ddagger}\}_{i\cdot}\widetilde{y}_j/\sigma_j=t_1$, which is obtained without loss of generality by changing the sign of t_1. Let $\mathcal{Y}$ be the set of all $y_j\in\mathbb{R}^n$ such that $\{\widetilde{A}_{M_j^\ddagger}\}_{i\cdot}\left[(y_j-\bar{y}_j)-\widetilde{y}_j\right]/\sigma_j>0$, and $\left|\{\widetilde{A}_{M_j}\}_{i\cdot}\left[(y_j-\bar{y}_j)-\widetilde{y}_j\right]\right|/\sigma_j<(t_2-t_1)/2$ for any $M_j\subseteq\mathcal{M}$. Then for any $y_j\in\mathcal{Y}$,

$$\begin{aligned}
&\max_{M_j\subseteq\mathcal{M}}\left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j-\bar{y}_j)\right|/\sigma_j\\
\le\ &\max_{M_j\subseteq\mathcal{M}}\left|\{\widetilde{A}_{M_j}\}_{i\cdot}\left[(y_j-\bar{y}_j)-\widetilde{y}_j\right]\right|/\sigma_j+\max_{M_j\subseteq\mathcal{M}}\left|\{\widetilde{A}_{M_j}\}_{i\cdot}\widetilde{y}_j\right|/\sigma_j\\
<\ &(t_2-t_1)/2+t_1<t_2.
\end{aligned}$$

Moreover, $\{\widetilde{A}_{M_j^\ddagger}\}_{i\cdot}(y_j - \bar{y}_j)/\sigma_j > \{\widetilde{A}_{M_j^\ddagger}\}_{i\cdot}\widetilde{y}_j/\sigma_j = t_1 > 0$. Therefore,

$$\max_{M_j \subseteq \mathcal{M}} \left| \{\widetilde{A}_{M_j}\}_{i\cdot}(y_j - \bar{y}_j) \right| /\sigma_j \geq \left| \{\widetilde{A}_{M_j^\ddagger}\}_{i\cdot}(y_j - \bar{y}_j) \right| /\sigma_j > t_1,$$

which implies that $H(t_2) - H(t_1) \geq \mathbb{P}(\mathcal{Y}) > 0$. Consequently, there exists a unique $c_0(\widehat{x}_j(t_i))$ satisfying (16), which is the $(1-\alpha)$th quantile of the distribution of $\max_{M_j \subseteq \mathcal{M}} \left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j - \bar{y}_j)\right|/\sigma_j$. This completes the proof of Theorem 2. □

S1.3 Proof of Proposition 1

Note that, for any $t \geq 0$,

$$\begin{aligned}
&\mathbb{P}_{n,F_j,\sigma_j}\left[\max_{M_j \subseteq \mathcal{M}} \left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j - \bar{y}_j)\right| /\sigma_j > t\right] \\
= \; &\mathbb{P}_{n,F_j,\sigma_j}\left[\max_{M_j \subseteq \mathcal{M}} \left|\{\widetilde{A}_{M_j}\}_{i\cdot}(y_j - \bar{y}_j)\right| /\|y_j - \bar{y}_j\|_{l_2} > (\sigma_j/\|y_j - \bar{y}_j\|_{l_2})\, t\right] \\
= \; &\mathbb{P}_{n,F_j,\sigma_j}\left(\max_{M_j \subseteq \mathcal{M}} |\{\widetilde{A}_{M_j}\}_{i\cdot}V| > t/U\right),
\end{aligned}$$

where V is uniformly distributed on the unit sphere in $\mathbb{R}^n$, and U is a nonnegative random variable such that U^2 follows an $\chi^2(n)$-distribution. Combining this result with the definition in (16) completes the proof. □

S1.4 Proof of Theorem 3

The upper bound of the convergence rate can be established following Cox (1983). The minimax lower bound of the convergence rate can be established following Tsybakov (2009). Moreover, the results hold for both fixed and random designs of $t \in \mathcal{T}$. □

S1.5 Proof of Theorem 4

We divide the proof of this theorem to three parts. To establish the minimax rate, we first prove the upper bound in Section S1.5.1, then prove the lower bound in Section S1.5.2. We give two auxiliary lemmas that are useful for the proof of this theorem in Section S1.5.3.

S1.5.1 Upper bound

For $j = 1, \ldots, p$, write $\widehat{F}_j(\widehat{x}) = \widehat{b}_j + \sum_{k=1}^p \widehat{F}_{jk}(\widehat{x}) + \sum_{k \neq l} \widehat{F}_{jkl}(\widehat{x})$, where $\sum_{i=1}^n \widehat{F}_{jk}(\widehat{x}_k(t_i)) = 0$, and $\sum_{i=1}^n \widehat{F}_{jkl}(\widehat{x}_k(t_i), \widehat{x}_l(t_i)) = 0$. Write $F_j(x) = b_j + \sum_{k=1}^p F_{jk}(x) + \sum_{k \neq l} F_{jkl}(x)$, where $\sum_{i=1}^n F_{jk}(x_k(t_i)) = 0$, and $\sum_{i=1}^n F_{jkl}(x_k(t_i), x_l(t_i)) = 0$. In light of the fact that $\widehat{\theta}_{j0}$ that minimizes (10) is given by $\widehat{\theta}_{j0} = \bar{y}_j - \int_{\mathcal{T}} \bar{T}(t)\widehat{F}_j(\widehat{x}(t))dt$, we focus our attention on $\widehat{F}_j(\widehat{x}(t))$ in the following proof, while the convergence rate of $\widehat{\theta}_{j0}$ is the same as that of $\widehat{F}_j(\widehat{x}(t))$.

Consider that $\widehat{F}_j$ is obtained from

$$\min_{F_j \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^{n} \left\{ y_{ij} - \int_0^{t_i} F_j(\widehat{x}(t))dt \right\}^2 + \tau_{nj} J_2(F_j) \right],$$

which implies that

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \left\{ \int_0^{t_i} F_j(x(t))dt + \epsilon_{ij} - \int_0^{t_i} \widehat{F}_j(\widehat{x}(t))dt \right\}^2 + \tau_{nj} J_2(\widehat{F}_j) \\ \leq \; & \frac{1}{n} \sum_{i=1}^{n} \left\{ \int_0^{t_i} F_j(x(t))dt + \epsilon_{ij} - \int_0^{t_i} F_j(\widehat{x}(t))dt \right\}^2 + \tau_{nj} J_2(F_j). \end{aligned}$$

With rearrangement of the terms, we have,

$$\begin{aligned} & \frac{1}{n} \sum_{i=1}^{n} \left[\int_0^{t_i} \left\{ F_j(x(t)) - \widehat{F}_j(\widehat{x}(t)) \right\} dt \right]^2 + \tau_{nj} J_2(\widehat{F}_j) \\ \leq \; & \frac{2}{n} \sum_{i=1}^{n} \epsilon_{ij} \left[\int_0^{t_i} \left\{ \widehat{F}_j(\widehat{x}(t)) - F_j(\widehat{x}(t)) \right\} dt \right] \\ & \qquad + \frac{1}{n} \sum_{i=1}^{n} \left[\int_0^{t_i} \left\{ F_j(x(t)) - F_j(\widehat{x}(t)) \right\} dt \right]^2 + \tau_{nj} J_2(F_j). \end{aligned} \tag{S2}$$

By Assumption 1 and the Taylor expansion,

$$(\widehat{F}_j - F_j)(\widehat{x}) = (\widehat{F}_j - F_j)(x) + \frac{\partial}{\partial t}(\widehat{F}_j - F_j)(x)(\widehat{x} - x) + o_p\left(\max_{k=1,\ldots,p} \|\widehat{x}_k - x_k\|_{L_2} \right),$$

where the Fréchet derivative of any $g(\cdot) \in \mathcal{H}$ is defined as,

$$\frac{\partial}{\partial t} g(x)(\widehat{x} - x) = \sum_{k=1}^{p} \frac{\partial g(x)}{\partial x_k}(\widehat{x}_k - x_k).$$

Then the first term on the right-hand-side of (S2) can be written as,

$$\begin{aligned} \frac{2}{n} \sum_{i=1}^{n} \epsilon_{ij} \left[\int_0^{t_i} \left\{ \widehat{F}_j(\widehat{x}(t)) - F_j(\widehat{x}(t)) \right\} dt \right] &= \frac{2}{n} \sum_{i=1}^{n} \epsilon_{ij} \left\{ \int_0^{t_i} (\widehat{F}_j - F_j)(x(t))dt \right\} \\ + \frac{2}{n} \sum_{i=1}^{n} \epsilon_{ij} & \left[\int_0^{t_i} \frac{\partial}{\partial t}(\widehat{F}_j - F_j)(x(t))\{\widehat{x}(t) - x(t)\}dt + o_p\left(\max_{k=1,\ldots,p} \|\widehat{x}_k - x_k\|_{L_2} \right) \right]. \end{aligned}$$

Meanwhile, by the Taylor expansion, the first term on the left-hand-side of (S2) can be written as,

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\left\{F_j(x(t))-\widehat{F}_j(\widehat{x}(t))\right\}dt\right]^2 \\
= & \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\left\{\widehat{F}_j(x(t))-F_j(x(t))\right\}dt+\int_0^{t_1}\frac{\partial}{\partial t}\widehat{F}_j(x(t))\{\widehat{x}(t)-x(t)\}dt\right. \\
& \qquad\qquad\qquad\qquad\qquad\qquad\qquad\left.-o_p\left(\max_{k=1,\ldots,p}\|\widehat{x}_k-x_k\|_{L_2}\right)\right]^2 \\
= & \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\left\{F_j(x(t))-\widehat{F}_j(x(t))\right\}dt\right]^2+\frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\frac{\partial}{\partial t}\widehat{F}_j(x(t))\{\widehat{x}(t)-x(t)\}dt\right]^2 \\
& \quad-\frac{2}{n}\sum_{i=1}^{n}\int_0^{t_i}\left\{F_j(x(t))-\widehat{F}_j(x(t))\right\}dt\int_0^{t_i}\frac{\partial}{\partial t}\widehat{F}_j(x(t))\{\widehat{x}+(t)-x(t)\}dt+R_1,
\end{aligned}
$$

where the remainder term R_1 is of the form,

$$
\begin{aligned}
R_1=\frac{1}{n}\sum_{i=1}^{n}&\left(o_p\left(\max_{k=1,\ldots,p}\|\widehat{x}_k-x_k\|_{L_2}^2\right)\right. \\
&\left.-o_p\left(\max_{k=1,\ldots,p}\|\widehat{x}_k-x_k\|_{L_2}\right)\int_0^{t_i}\left[F_j(x(t))-\widehat{F}_j(x(t))-\frac{\partial}{\partial t}\widehat{F}_j(x(t))\{\widehat{x}(t)-x(t)\}\right]dt\right).
\end{aligned}
$$

Therefore, the inequality (S2) is equivalent to

$$
\begin{aligned}
& \frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\left\{F_j(x(t))-\widehat{F}_j(x(t))\right\}dt\right]^2+\frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\frac{\partial}{\partial t}\widehat{F}_j(x(t))\{\widehat{x}(t)-x(t)\}dt\right]^2 \\
& +\frac{2}{n}\sum_{i=1}^{n}\int_0^{t_i}\left\{\widehat{F}_j(x(t))-F_j(x(t))\right\}dt\int_0^{t_i}\frac{\partial}{\partial t}\widehat{F}_j(x(t))\{\widehat{x}(t)-x(t)\}dt+R_1+\tau_{nj}J_2(\widehat{F}_j) \\
\leq\ & \frac{2}{n}\sum_{i=1}^{n}\epsilon_{ij}\left\{\int_0^{t_i}(\widehat{F}_j-F_j)(x(t))dt\right\}+\frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\{F_j(x(t))-F_j(\widehat{x}(t))\}dt\right]^2 \\
& +\frac{2}{n}\sum_{i=1}^{n}\epsilon_{ij}\left[\int_0^{t_i}\frac{\partial}{\partial t}(\widehat{F}_j-F_j)(x)\{\widehat{x}(t)-x(t)\}dt+o_p\left(\max_{k=1,\ldots,p}\|\widehat{x}_k-x_k\|_{L_2}\right)\right]+\tau_{nj}J_2(F_j)
\end{aligned}
\tag{S3}
$$

Write the left-hand side of (S3) as $\widetilde{\Delta}_1+\widetilde{\Delta}_2+\widetilde{\Delta}_3+R_1+\tau_{nj}J_2(\widehat{F}_j)$, and the right-hand side of (S3) as $\Delta_1+\Delta_2+\Delta_3+\tau_{nj}J_2(F_j)$. Our proof strategy is to derive the upper and lower bounds for the left and right-hand sides of (S3), respectively, then put them together.

Step 1: Bounding the right-hand-side of (S3). We first bound the three terms $\Delta_1,\Delta_2,\Delta_3$ on the right-hand-side of (S3).

For Δ_1, by Lemma 1 and the Minkowski inequality, we have,

$$\Delta_1 \le O_p\Big\{ \left\|\widehat{F}_j(x(t)) - F_j(x(t))\right\|_{L_2}^2 \log^{-2} \left\|\widehat{F}_j(x(t)) - F_j(x(t))\right\|_{L_2} \\ + \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{\log p}{n} + \sqrt{\frac{\log p}{n}} \left\|\widehat{F}_j(x(t)) - F_j(x(t))\right\|_{L_2} \Big\}.$$

For Δ_2, by the Taylor expansion and Assumption 1, we have,

$$\begin{aligned} \Delta_2 &\le \frac{c}{n}\sum_{i=1}^n \left[\int_0^{t_i} \frac{\partial}{\partial t} F_j(x(t))\{x(t) - \widehat{x}(t)\} + o_p\left(\max_{k=1,\ldots,p} \|x_k - \widehat{x}_k\|_{L_2}^2\right) dt\right]^2 \\ &\le \|F_j\|_{\mathcal{H}}^2 \max_{k=1,\ldots,p} \|x_k - \widehat{x}_k\|_{L_2}^2 + o_p\left(\max_{k=1,\ldots,p} \|x_k - \widehat{x}_k\|_{L_2}^2\right) = O_p\left(n^{\frac{-2\beta_1}{2\beta_1+1}}\right). \end{aligned} \tag{S4}$$

for some constant c, where the second step is by the Jensen's inequality, and the last step is due to Theorem 3.

For Δ_3, since $\beta_2 > 1$, $\partial K(x,\cdot)/\partial x_k \in \mathcal{H}$, and by the reproducing property, we have,

$$\frac{\partial(\widehat{F}_j - F_j)(x)}{\partial x_k} = \left\langle \widehat{F}_j - F_j, \frac{\partial K(x,\cdot)}{\partial x_k}\right\rangle_{\mathcal{H}} \le \|\widehat{F}_j - F_j\|_{\mathcal{H}}^{1/2} \left\|\frac{\partial K(x,\cdot)}{\partial x_k}\right\|_{\mathcal{H}}^{1/2} < \infty.$$

Hence, $\partial(\widehat{F}_j - F_j)(x)/\partial x_k \in \mathcal{H}$, and for any x, $|\partial(\widehat{F}_j - F_j)(x)/\partial x_k| \le \|\partial(\widehat{F}_j - F_j)(x)/\partial x_k\|_{\mathcal{H}} < \infty$, which together with Assumption 2, implies that $\max_k \Big\{|\partial(\widehat{F}_j - F_j)(x)/\partial x_k|\Big\} \le C\|\widehat{F}_j - F_j\|_{L_2}$ almost surely. By Assumption 1 and the Cauchy-Schwarz inequality, we have,

$$\begin{aligned} \Delta_3 \le\ & \frac{2c}{n}\sum_{i=1}^n |\epsilon_{ij}| \int_0^{t_i} C\|\widehat{F}_j(x(t)) - F_j(x(t))\|_{L_2} \max_{k=1,\ldots,p} |\widehat{x}_k(t) - x_k(t)| dt \\ & \qquad + o_p\left(n^{-1/2} \max_{k=1,\ldots,p} \|x_k - \widehat{x}_k\|_{L_2}^2\right) \\ \le\ & 2c \max_{k=1,\ldots,p} \|\widehat{x}_k - x_k\|_{L_2} \left\|\widehat{F}_j(x(t)) - F_j(x(t))\right\|_{L_2} \frac{1}{n}\sum_{i=1}^n |\epsilon_{ij} C| \\ & \qquad + o_p\left(n^{-1/2} \max_{k=1,\ldots,p} \|x_k - \widehat{x}_k\|_{L_2}^2\right) \\ =\ & O_p\left(n^{\frac{-\beta_1}{2\beta_1+1}} \|\widehat{F}_j(x(t)) - F_j(x(t))\|_{L_2}\right), \end{aligned}$$

for some constant c, where the last step is due to the strong law of large numbers.

Step 2: Bounding the left-hand-side of (S3). We next bound the terms $\widetilde{\Delta}_1, \widetilde{\Delta}_2, \widetilde{\Delta}_3$ and R_1 on the left-hand-side of (S3).

For $\widetilde{\Delta}_1$, by Lemma 2, with probability at least $1-2p^{-c_1}$, for some constant $C>0$,

$$\widetilde{\Delta}_1 \geq \left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2}^2 - C\Bigg\{ \left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2}^2 \log^{-2}\left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2} + \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + (c_1+1)\frac{\log p}{n} + \sqrt{(c_1+1)\frac{\log p}{n}}\left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2} + n^{-1/2}e^{-p}\Bigg\}. \tag{S5}$$

For $\widetilde{\Delta}_2$, we can drop this term, because $\widetilde{\Delta}_2 \geq 0$.

For $\widetilde{\Delta}_3$, by the Cauchy-Schwarz inequality,

$$\begin{aligned}
\widetilde{\Delta}_3 \geq & -2\left(\frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\left\{\widehat{F}_j(x(t))-F_j(x(t))\right\}dt\right]^2\right)^{1/2} \\
& \times \left(\frac{1}{n}\sum_{i=1}^{n}\left[\int_0^{t_i}\frac{\partial}{\partial t}\widehat{F}_j(x(t))\{\widehat{x}(t)-x(t)\}dt\right]^2\right)^{1/2} \\
\geq & -2\left\|\widehat{F}_j(x(t))-F_j(x(t))\right\|_{L_2}\|F_j\|_{\mathcal{H}}\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2} \\
= & \, O_p\left(n^{\frac{-\beta_1}{2\beta_1+1}}\left\|\widehat{F}_j(x(t))-F_j(x(t))\right\|_{L_2}\right),
\end{aligned}$$

where the second step is due to the Minkowski inequality.

For the remainder term R_1 on the left-hand-side of (S3), by Assumption 1 and the Cauchy-Schwarz inequality, we have,

$$\begin{aligned}
R_1 &= o_p\left(\max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}\left\|\widehat{F}_j(x(t))-F_j(x(t))\right\|_{L_2} + \max_{k=1,\ldots,p}\|x_k-\widehat{x}_k\|_{L_2}^2\|F_j\|_{\mathcal{H}}\right) \\
&= o_p\left(n^{\frac{-\beta_1}{2\beta_1+1}}\left\|\widehat{F}_j(x(t))-F_j(x(t))\right\|_{L_2}\right) + o_p\left(n^{\frac{-2\beta_1}{2\beta_1+1}}\right),
\end{aligned}$$

where the second step is again due to the Minkowski inequality.

Step 3: Putting the two bounds together. Combining the bounds for each term in (S3), we obtain that, for any $c_1>0$ and $c_2>1$, with probability at least $1-4p^{-c_1}$, there exists a constant $C>0$, such that

$$\begin{aligned}
&\|F_j(x(t))-\widehat{F}_j(x(t))\|_{L_2}^2 \\
\leq\ & C\Bigg[c_2^{-\frac{4\beta_2}{2\beta_2-1}}\left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2}^2\log^{-2}\left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2} + c_2^{\frac{4\beta_2}{4\beta_2+1}}\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} \\
& + (c_1+1)\frac{\log p}{n} + \sqrt{(c_1+1)\frac{\log p}{n}}\left\|F_j(x(t))-\widehat{F}_j(x(t))\right\|_{L_2} + n^{-1/2}e^{-p} \\
& + n^{\frac{-\beta_1}{2\beta_1+1}}\left\|\widehat{F}_j(x(t))-F_j(x(t))\right\|_{L_2} + n^{\frac{-2\beta_1}{2\beta_1+1}} + \tau_{nj}\left\{J_2(F_j)-J_2(\widehat{F}_j)\right\}\Bigg].
\end{aligned}$$

Taking c_2 large enough such that $Cc_2^{-4\beta_2/(2\beta_2-1)} \leq 1/2$, then

$$\left\|F_j(x(t)) - \widehat{F}_j(x(t))\right\|_{L_2}^2 \log^{-2}\left\|F_j(x(t)) - \widehat{F}_j(x(t))\right\|_{L_2} \leq 2C\left[c_2^{\frac{4\beta_2}{4\beta_2+1}}\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}}\right.$$
$$+ (c_1+1)\frac{\log p}{n} + \sqrt{(c_1+1)\frac{\log p}{n}}\left\|F_j(x(t)) - \widehat{F}_j(x(t))\right\|_{L_2}$$
$$\left. + n^{-1/2}e^{-p} + n^{\frac{-\beta_1}{2\beta_1+1}}\left\|\widehat{F}_j(x(t)) - F_j(x(t))\right\|_{L_2} + n^{\frac{-2\beta_1}{2\beta_1+1}} + \tau_{nj}\left\{J_2(F_j) - J_2(\widehat{F}_j)\right\}\right].$$

Therefore,

$$\left\|F_j(x(t)) - \widehat{F}_j(x(t))\right\|_{L_2}^2 = O_p\left\{\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} + \frac{\log p}{n} + n^{-\frac{2\beta_1}{2\beta_1+1}}\right\}.$$

This leads to the desired upper bound.

S1.5.2 Lower bound

We first construct a matrix $A_j \in \mathbb{R}^{p^2 \times \widetilde{N}}$ for each $j = 1, \ldots, p$, whose entry is chosen from $\{\pm 1, 0\}$, and is used to index a set of functions for establishing the lower bound. Here, the value of $\widetilde{N}$ is to be specified later. We choose $1 \leq s_j < \infty$ rows of A_j to be nonzero. By the Vershamov-Gilbert Lemma (Tsybakov, 2009), there exist a set $\{\zeta_1, \ldots, \zeta_{m_1}\} \subset \{0,1\}^{p^2}$ such that, (a) $\|\zeta_k\|_{l_1} = s_j$, for $k = 1, \ldots, m_1$; (b) $\|\zeta_{k_1} - \zeta_{k_2}\|_{l_1} \geq s_j/2$, for $k_1 \neq k_2$; and (c) $4\log m_1 \geq s_j \log(p^2/s_j)$. By the same lemma, there exist a set $\{\zeta_1^\dagger, \ldots, \zeta_{m_2}^\dagger\} \in \{-1,1\}^{s_j \times \widetilde{N}}$ such that, (a$'$) $\|\zeta_{k_1}^\dagger - \zeta_{k_2}^\dagger\|_F \geq \widetilde{N}s_j/2$, for $k_1 \neq k_2$; and (b$'$) $8\log m_2 \geq \widetilde{N}s_j$. We set the zero rows of A_j according to ζ_k, and set the nonzero rows of A_j according to $\zeta_k^\dagger$. As such, the matrix A_j is chosen from the set,

$$\mathbb{A} = \left\{A_j(\zeta_{k_1}, \zeta_{k_2}^\dagger) \in \mathbb{R}^{p^2 \times \widetilde{N}} : k_1 = 1, \ldots, m_1, k_2 = 1, \ldots, m_2\right\},$$

where $\mathrm{card}(\mathbb{A}) = m_1 m_2$. By the above constructions (c) and (b$'$), we have that,

$$\log \mathrm{card}(\mathbb{A}) \geq \frac{1}{4}s_j \log(p^2/s_j) + \frac{1}{8}\widetilde{N}s_j.$$

Next, we define functions of the form g_{A_j} with $A_j \in \mathbb{A}$. Note that, by the spectral theorem, the reproducing kernel K_j of the RKHS $\mathcal{H}_j$ admits the eigenvalue decomposition

$$K_j(\widehat{x}_j, \widehat{x}_j') = \sum_{\nu \geq 1} \gamma_{j\nu}\phi_{j\nu}(\widehat{x}_j)\phi_{j\nu}(\widehat{x}_j')$$

where $\gamma_1 \geq \gamma_2 \geq \cdots \geq 0$ are its eigenvalues, and $\{\phi_\nu : \nu \geq 1\}$ are the corresponding eigenfunctions that are orthonormal in L_2. Since $\mathcal{H}_j$ is embedded to a β_2th-order Sobolev space, the eigenvalues decays as $\gamma_{j\nu} \asymp \nu^{-2\beta_2}$ (Wahba, 1990). We define the function,

$$g_{A_j}(\widehat{x}_1, \ldots, \widehat{x}_p) = \widetilde{N}^{-1/2} \sum_{j=1}^{p} \sum_{\nu=1}^{\widetilde{N}} a_{j^2,\nu} \gamma_{j,\widetilde{N}+\nu}^{1/2} \phi_{j,\widetilde{N}+\nu}(\widehat{x}_j)$$

$$+ \widetilde{N}^{-1/2} \sum_{j,k=1,\ldots,p; j \neq k} \sum_{\nu=1}^{\widetilde{N}} a_{j \cdot k,\nu} \gamma_{j,\widetilde{N}+\nu}^{1/2} \gamma_{k,\widetilde{N}+\nu}^{1/2} \phi_{j,\widetilde{N}+\nu}(\widehat{x}_j) \phi_{k,\widetilde{N}+\nu}(\widehat{x}_k), \quad A_j \in \mathbb{A},$$

Let $\| \cdot \|_{\mathcal{H}}^0$ denote the ℓ_0-norm. Then, we have,

$$\|g_{A_j}\|_{\mathcal{H}}^0 \leq \sum_{j=1}^{p} \left\| \sum_{\nu=1}^{\tilde{N}} a_{j^2,\nu} \gamma_{j,\widetilde{N}+\nu}^{1/2} \phi_{j,\widetilde{N}+\nu}(\widehat{x}_j) \right\|_{\mathcal{H}}^0$$

$$+ \sum_{j,k=1,\ldots,p; j \neq k} \left\| \sum_{\nu=1}^{N} a_{j \cdot k,\nu} \gamma_{j,\widetilde{N}+\nu}^{1/2} \gamma_{k,\widetilde{N}+\nu}^{1/2} \phi_{j,\widetilde{N}+\nu}(\widehat{x}_j) \phi_{k,\widetilde{N}+\nu}(\widehat{x}_k) \right\|_{\mathcal{H}}^0 \leq s_j.$$

For any two matrices $A_j, B_j \in \mathbb{A}$, we have,

$$\|g_{A_j} - g_{B_j}\|_{L_2}^2 \geq C_1 \widetilde{N}^{-1} \sum_{j=1}^{p} \sum_{\nu=1}^{\widetilde{N}} \gamma_{j,\widetilde{N}+\nu} \left(a_{j^2,\nu} - b_{j^2,\nu}\right)^2$$

$$+ \ C_1 \widetilde{N}^{-1} \sum_{j,k=1,\ldots,p; j \neq k} \sum_{\nu=1}^{\widetilde{N}} \gamma_{j,\widetilde{N}+\nu} \gamma_{k,\widetilde{N}+\nu} \left(a_{j \cdot k,\nu} - b_{j \cdot k,\nu}\right)^2$$

$$\geq C_2 \widetilde{N}^{-1} (2\widetilde{N})^{-4\beta_2} \sum_{j,k=1,\ldots,p; j \neq k} \sum_{\nu=1}^{\widetilde{N}} \left(a_{j \cdot k,\nu} - b_{j \cdot k,\nu}\right)^2 \geq C_3 s \widetilde{N}^{-4\beta_2},$$

for some constants $C_1, C_2, C_3 > 0$, where the second and third steps are by the construction (a′). On the other hand, for any $A_j \in \mathbb{A}$, and by the Minkowski inequality,

$$\|g_{A_j}\|_{L_2}^2 \leq C_4 \widetilde{N}^{-1} \sum_{j=1}^{p} \left\| \sum_{\nu=1}^{N} a_{j^2,\nu} \gamma_{j,\widetilde{N}+\nu}^{1/2} \phi_{j,\widetilde{N}+\nu} \right\|_{L_2}^2$$

$$+ \ C_4 \widetilde{N}^{-1} \sum_{j,k=1,\ldots,p; j \neq k} \left\| \sum_{\nu=1}^{N} a_{j \cdot k,\nu} \gamma_{j,\widetilde{N}+\nu}^{1/2} \gamma_{k,\widetilde{N}+\nu}^{1/2} \phi_{j,\widetilde{N}+\nu} \phi_{k,\widetilde{N}+\nu} \right\|_{L_2}^2$$

$$\leq C_4 \widetilde{N}^{-1} \sum_{j=1}^{p} \sum_{\nu=1}^{\widetilde{N}} \gamma_{j,\widetilde{N}+\nu} a_{j^2,\nu}^2 \leq C_5 \widetilde{N}^{-1} \widetilde{N}^{-2\beta_2} \widetilde{N} s_j = C_5 s_j \widetilde{N}^{-2\beta_2},$$

for some constants $C_4, C_5 > 0$, where the second and third steps are by (a).

We are now ready to derive the lower bound. Let $\mathcal{Z}$ denote a random variable uniformly distributed on $\{1, 2, \ldots, \text{card}(\mathbb{A})\}$. Then for any $j = 1, \ldots, p$,

$$\inf_{\widetilde{F}_j} \sup_{F_j \in \mathcal{H}} \mathbb{P}\left\{\|\widetilde{F}_j(\widehat{x}) - F_j(\widehat{x})\|_{L_2}^2 \geq \frac{1}{4} \min_{A_j \neq B_j \in \mathbb{A}} \|g_{A_j} - g_{B_j}\|_{L_2}^2\right\} \geq \inf_{\widehat{\mathcal{Z}}} \mathbb{P}\left\{\widehat{\mathcal{Z}} \neq \mathcal{Z}\right\},$$

where the infimum on the right-hand-side is taken over all decision rules that are measurable functions of the data (Tsybakov, 2009). By the Fano's Lemma, we have,

$$\mathbb{P}\left\{\widehat{\mathcal{Z}} \neq \mathcal{Z} | t_1, \ldots, t_n\right\} \geq 1 - \frac{1}{\log\{\text{card}(\mathbb{A})\}}\left[\mathcal{I}_{t_1,\ldots,t_n}(y_{1j}, \ldots, y_{nj}; \mathcal{Z}) + \log 2\right],$$

where $\mathcal{I}_{t_1,\ldots,t_n}(y_{1j}, \ldots, y_{nj}; \mathcal{Z})$ is the mutual information between $\mathcal{Z}$ and $(y_{1j}, \ldots, y_{nj})$ conditioning on $(t_1, \ldots, t_n)$. Note that

$$\begin{aligned}\mathbb{E}_{t_1,\ldots,t_n}\left[\mathcal{I}_{t_1,\ldots,t_n}(y_{1j}, \ldots, y_{nj}; \mathcal{Z})\right] &\leq \frac{n}{\text{card}(\mathbb{A})\{\text{card}(\mathbb{A}) - 1\}} \sum_{A_j \neq B_j \in \mathbb{A}} \mathbb{E}_{t_1,\ldots,t_n} \|g_{A_j} - g_{B_j}\|_{L_2}^2 \\ &\leq \frac{n}{2} \max_{A_j \neq B_j \in \mathbb{A}} \|g_{A_j} - g_{B_j}\|_{L_2}^2 \leq 2n \max_{A_j \in \mathbb{A}} \|g_{A_j}\|_{L_2}^2 \leq 2C_5 n s_j \widetilde{N}^{-2\beta_2}.\end{aligned}$$

Henceforth,

$$\inf_{\widetilde{F}_j} \sup_{F_j \in \mathcal{H}} \mathbb{P}\left\{\|\widetilde{F}_j(\widehat{x}) - F_j(\widehat{x})\|_{L_2}^2 \geq C_3 s_j \widetilde{N}^{-4\beta_2}\right\} \geq \inf_{\widehat{\mathcal{Z}}} \mathbb{P}\{\widehat{\mathcal{Z}} \neq \mathcal{Z}\} \geq 1 - \frac{2C_5 n s_j \widetilde{N}^{-2\beta_2} + \log 2}{\frac{1}{4} s_j \log(p^2/s_j) + \frac{1}{8}\widetilde{N} s_j}.$$

Taking $\widetilde{N} = 1$ and $s_j = C_6 n^{-1} \log p$ for a sufficiently small constant C_6 yields that

$$\inf_{\widetilde{F}_j} \sup_{F_j \in \mathcal{H}} \mathbb{P}\left\{\|\widetilde{F}_j(\widehat{x}) - F_j(\widehat{x})|_{L_2}^2 \geq C_7 \frac{\log p}{n}\right\} \geq \frac{1}{2},$$

for some constant $C_7 > 0$. Meanwhile, taking $s_j = 1$ and $\widetilde{N} = C_8(n \log n^{-1})^{\frac{1}{4\beta_2+2}}$ for a sufficiently small $C_8 > 0$ yields that

$$\inf_{\widetilde{F}_j} \sup_{F_j \in \mathcal{H}} \mathbb{P}\left\{\|\widetilde{F}_j(\widehat{x}) - F_j(\widehat{x})\|_{L_2}^2 \geq C_9 (n \log^{-1} n)^{-\frac{2\beta_2}{2\beta_2+1}}\right\} \geq \frac{1}{2},$$

for some $C_9 > 0$. Therefore, we have

$$\inf_{\widetilde{F}_j} \sup_{F_j \in \mathcal{H}} \mathbb{P}\left[\|\widetilde{F}_j(\widehat{x}) - F_j(\widehat{x})\|_{L_2}^2 \geq C_{10}\left\{\frac{\log p}{n} + (n \log^{-1} n)^{-\frac{2\beta_2}{2\beta_2+1}}\right\}\right] \geq \frac{1}{2},$$

for some $C_{10} > 0$. Finally, note that, $\widehat{x}_j$ is an estimator of x_j satisfying that $\|\widehat{x}_j - x_j\|_{L_2}^2 = O_p\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\right)$. Then for any $F_j, \widetilde{F}_j \in \mathcal{H}$,

$$\mathbb{P}\left[\min\left\{\|F_j(x) - F_j(\widehat{x})\|_{L_2}^2, \|\widetilde{F}_j(x) - \widetilde{F}_j(\widehat{x})\|_{L_2}^2\right\} \geq C_{11} n^{-\frac{2\beta_1}{2\beta_1+1}}\right] \geq \frac{1}{2},$$

Therefore,

$$\inf_{\widetilde{F}_j}\sup_{F_j\in\mathcal{H}}\mathbb{P}\left[\|\widetilde{F}_j(x)-F_j(x)\|_{L_2}^2\geq C_{12}\left\{\frac{\log p}{n}+(n\log^{-1}n)^{-\frac{2\beta_2}{2\beta_2+1}}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right\}\right]\geq\frac{1}{2},$$

for some constant $C_{12}>0$, which completes the proof of Theorem 4. □

S1.5.3 Auxiliary lemmas for Theorem 4

For any $g\in\mathcal{H}$, define the norm, $\|g(x(t))\|_n=\sqrt{(1/n)\sum_{i=1}^n g^2(x(t_i))}$.

Lemma 1. *Suppose that $F_j\in\mathcal{H}$, and the errors $\{\epsilon_{ij}\}_{i=1}^n$ are i.i.d. Gaussian. Then there exists some constant $C>0$ such that, for any $c_1>0$ and $c_2>1$, with probability at least $1-2p^{-c_1}$,*

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\epsilon_{ij}F_j(x(t_i))\\
&\leq C\left\{c_2^{-\frac{4\beta_2}{2\beta_2-1}}\|F_j(x(t))\|_{L_2}^2\log^{-2}\|F_j(x(t))\|_{L_2}+\left(c_2^{-\frac{4\beta_2}{2\beta_2-1}}+c_2^{\frac{4\beta_2}{4\beta_2+1}}\right)\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}}\right.\\
&\qquad\left.+\left(c_2^{-\frac{4\beta_2}{2\beta_2-1}}+c_1+1\right)\frac{\log p}{n}+\sqrt{(c_1+1)\frac{\log p}{n}}\|F_j(x(t))\|_{L_2}+n^{-1/2}e^{-p}\right\}.
\end{aligned}$$

Proof of Lemma 1: Recall the RKHS $\mathcal{H}$ defined in (8). For notational simplicity, we denote $F_{jk}\equiv F_{jkk}$ for $k=1,\ldots,p$. It has been shown that the νth eigenvalue of the reproducing kernel of RKHS $\mathcal{H}$ is of order $(\nu\log^{-1}\nu)^{-2\beta_2}$, for $\nu\geq1$; see, e.g., Bach (2017). Since $\{\epsilon_{ij}\}_{i=1}^n$ are i.i.d. Gaussian, by Lemma 2.2 of Yuan and Zhou (2016) and Corollary 8.3 of van de Geer (2000), we have that, for any $c_1>0$, with probability at least $1-p^{-c_1}$,

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^n\epsilon_{ij}F_j(x(t_i))\\
&\leq 2C_1n^{-1/2}\sum_{k,l=1}^p\left(\|F_{jkl}(x(t))\|_n\log^{-1}\|F_{jkl}(x(t))\|_n\right)^{1-\frac{1}{2\beta_2}}\left(\|F_{jkl}\|_{\mathcal{H}}\log^{-1}\|F_{jkl}\|_{\mathcal{H}}\right)^{\frac{1}{2\beta_2}}\\
&\qquad+2C_1n^{-1/2}\sqrt{(c_1+1)\log p}\sum_{k,l=1}^p\|F_{jkl}(x(t))\|_n+2C_1n^{-1/2}e^{-p}\sum_{k,l=1}^p\|F_{jkl}\|_{\mathcal{H}}\\
&\equiv 2C_1(\Delta_4+\Delta_5+\Delta_6),
\end{aligned}\tag{S6}$$

for some constant C_1. Next, we bound the three terms $\Delta_4,\Delta_5,\Delta_6$ on the right-hand-side of (S6), respectively.

For Δ_4, by the Young's inequality, for any $c_2 > 1$, we have,

$$\begin{aligned}\Delta_4 \leq c_2^{-\frac{4\beta_2}{2\beta_2-1}} \sum_{k,l=1}^{p} \left(\|F_{jkl}(x(t))\|_n \log^{-1} \|F_{jkl}(x(t))\|_n\right)^2 \\ + c_2^{\frac{4\beta_2}{4\beta_2+1}} n^{-\frac{2\beta_2}{2\beta2+1}} \sum_{k,l=1}^{p} \left(\|F_{jkl}\|_{\mathcal{H}} \log^{-1} \|F_{jkl}\|_{\mathcal{H}}\right)^{\frac{2}{2\beta_2+1}} .\end{aligned}$$

Note that

$$\sum_{k,l=1}^{d} \left(\|F_{jk}\|_{\mathcal{H}} \log^{-1} \|F_{jk}\|_{\mathcal{H}}\right)^{\frac{2}{2\beta_2+1}} \leq C_2' \sum_{k,l=1}^{d} \left(\|F_{jk}\|_{\mathcal{H}} \log^{-1} \|F_{jk}\|_{\mathcal{H}}\right)^0 \leq C_2,$$

for some constants C_2', C_2, where the last step is due to Assumption 1 that the number of nonzero functional components of F_j is bounded. Henceforth,

$$\begin{aligned}n^{-1/2} \sum_{k,l=1}^{d} \left(\|F_{jkl}(x(t))\|_n \log^{-1} \|F_{jkl}(x(t))\|_n\right)^{1-\frac{1}{2\beta_2}} \left(\|F_{jkl}\|_{\mathcal{H}} \log^{-1} \|F_{jkl}\|_{\mathcal{H}}\right)^{\frac{1}{2\beta_2}} \\ \leq c_2^{-\frac{4\beta_2}{2\beta_2-1}} \sum_{k,l=1}^{p} \left(\|F_{jkl}(x(t))\|_n \log^{-1} \|F_{jkl}(x(t))\|_n\right)^2 + c_2^{\frac{4\beta_2}{4\beta_2+1}} n^{-\frac{2\beta_2}{2\beta2+1}} C_2.\end{aligned} \tag{S7}$$

By Theorem 4 of Koltchinskii and Yuan (2010), there exists some constant $C_3 > 0$ such that, with probability at least $1 - p^{-c_1}$,

$$\begin{aligned}\sum_{k,l=1}^{p} \left(\|F_{jkl}(x(t))\|_n \log^{-1} \|F_{jkl}(x(t))\|_n\right)^2 \leq 2C_3^2 \sum_{k,l=1}^{p} \left(\|F_{jkl}(x(t))\|_{L_2} \log^{-1} \|F_{jkl}(x(t))\|_{L_2}\right)^2 + \\ + 2C_3^2 \left\{ \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{(c_1+1)\log d}{n} \right\} \sum_{k,l=1}^{p} \left(\|F_{jkl}\|_{\mathcal{H}} \log^{-1} \|F_{jkl}\|_{\mathcal{H}}\right)^2 .\end{aligned}$$

Note that there exists some constant $c_3 > 1$, such that

$$\sum_{k,l=1}^{p} \left(\|F_{jkl}\|_{L_2} \log^{-1} \|F_{jkl}\|_{L_2}\right)^2 \leq c_3 \left(\|F_j\|_{L_2} \log^{-1} \|F_j\|_{L_2}\right)^2 ,$$

and

$$\sum_{k,l=1}^{p} \left(\|F_{jkl}\|_{\mathcal{H}} \log^{-1} \|F_{jkl}\|_{\mathcal{H}}\right)^2 \leq \sum_{k,l=1}^{p} (\|F_{jkl}\|_{\mathcal{H}} \log^{-1} \|F_{jkl}\|_{\mathcal{H}})^0 \leq C_2.$$

Then, we have

$$\begin{aligned}\sum_{k,l=1}^{p} \left(\|F_{jkl}(x(t))\|_n \log^{-1} \|F_{jkl}(x(t))\|_n\right)^2 \leq 2C_3^2 c_3 \left(\|F_j(x(t))\|_{L_2} \log^{-1} \|F_j(x(t))\|_{L_2}\right)^2 \\ + 2C_2C_3^2 \left\{ \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{(c_1+1)\log p}{n} \right\}.\end{aligned}$$

Inserting into (S7) yields that

$$\begin{aligned}\Delta_4 \leq\, & 2C_3^2 c_3 c_2^{-\frac{4\beta_2}{2\beta_2-1}} (\|F_j(x(t))\|_{L_2} \log^{-1} \|F_j(x(t))\|_{L_2})^2 \\ & + 2C_2 C_3^2 c_2^{-\frac{4\beta_2}{2\beta_2-1}} \left\{ \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{(c_1+1)\log p}{n} \right\} + C_2 c_2^{\frac{4\beta_2}{4\beta_2+1}} \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}}. \end{aligned} \tag{S8}$$

For Δ_5, by Theorem 4 of Koltchinskii and Yuan (2010) again, there exists a constant $C_4 > 0$, such that

$$\begin{aligned} & \sum_{k,l=1}^{p} \|F_{jkl}(x(t))\|_n \\ & \leq C_4 \sum_{k,l=1}^{p} \|F_{jkl}(x(t))\|_{L_2} + C_4 \left\{ \left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}} + \sqrt{\frac{(c_1+1)\log p}{n}} \right\} \sum_{k,l=1}^{p} \|F_{jkl}\|_{\mathcal{H}} \\ & \leq C_4 \sum_{k,l=1}^{p} \|F_{jkl}(x(t))\|_{L_2} + C_2 C_4 \left\{ \left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}} + \sqrt{\frac{(c_1+1)\log p}{n}} \right\}. \end{aligned}$$

Define the set $\mathcal{Q}_1 \equiv \left\{ k, l = 1, \dots, p : \|F_{jkl}(x(t))\|_{L_2} > \sqrt{n^{-1}\log p} \right\}$. By the Cauchy-Schwartz inequality, we have,

$$\begin{aligned} \sum_{k,l \in \mathcal{Q}_1} \|F_{jkl}(x(t))\|_{L_2} & \leq \operatorname{card}^{1/2}(\mathcal{Q}_1) \cdot \left(\sum_{k,l \in \mathcal{Q}_1} \|F_{jkl}(x(t))\|_{L_2}^2 \right)^{1/2} \\ \leq \sum_{k,l=1}^{p} \|F_{jkl}\|_{\mathcal{H}}^0 \cdot & \left(\sum_{k,l=1}^{p} \|F_{jkl}(x(t))\|_{L_2}^2 \right)^{1/2} \leq C_2 c_4 \|F_j(x(t))\|_{L_2}, \end{aligned}$$

where $c_4 > 1$ satisfies that $\sum_{k,l=1}^{p} \|F_{jkl}(x(t))\|_{L_2}^2 \leq c_4^2 \|F_j(x(t))\|_{L_2}^2$. Next, define the set $\mathcal{Q}_2 \equiv \{k, l = 1, \dots, p : \|F_{jkl}(x(t))\|_{L_2} \leq \sqrt{n^{-1}\log p}\}$. By definition,

$$\sum_{k,l \in \mathcal{Q}_2} \|F_{jkl}(x(t))\|_{L_2} \leq \sum_{k,l \in \mathcal{Q}_2} \|F_{jkl}(x(t))\|_{L_2}^0 \sqrt{\frac{\log p}{n}} \sqrt{\frac{\log p}{n}} \sum_{k,l=1}^{p} \|F_{jkl}(x(t))\|_{L_2}^0 \leq C_2 \sqrt{\frac{\log p}{n}}.$$

Combining $\mathcal{Q}_1$ and $\mathcal{Q}_2$ gives,

$$\begin{aligned} \sum_{k,l=1}^{p} \|F_{jkl}(x(t))\|_{L_2} & \leq \sum_{k,l \in \mathcal{Q}_1} \|F_{jkl}(x(t))\|_{L_2} + \sum_{k,l \in \mathcal{Q}_2} \|F_{jkl}(x(t))\|_{L_2} \\ & \leq C_2 c_4 \|F_j(x(t))\|_{L_2} + C_2 \sqrt{\frac{\log p}{n}}. \end{aligned}$$

Henceforth, we can bound Δ_5 as,

$$\Delta_5 \leq C_2 C_4 c_4 \sqrt{\frac{\log p}{n}} \|F_j(x(t))\|_{L_2} + C_2 C_4 \left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}} \sqrt{\frac{\log p}{n}} + 2C_2 C_4 \sqrt{(c_1+1)} \frac{\log p}{n}.$$

For Δ_6, it can be bounded as,

$$\Delta_6 \leq n^{-1/2} e^{-p} \sum_{k,l=1}^{p} \|F_{jk}\|_{\mathcal{H}}^0 \leq C_2 n^{-1/2} e^{-p}.$$

Combining the bounds for $\Delta_4, \Delta_5, \Delta_6$, and applying the Cauchy-Schwarz inequality completes the proof of Lemma 1. □

Lemma 2. *Suppose that $F_j \in \mathcal{H}$. Then there exists some constant $C > 0$ such that, for any $c_1 > 0$ and $c_2 > 1$, with probability at least $1 - 2p^{-c_1}$,*

$$\|F_j(x(t))\|_{L_2}^2 \leq \|F_j(x(t))\|_n^2 + C \left\{ c_2^{-\frac{4\beta_2}{2\beta_2 - 1}} \|F_j(x(t))\|_{L_2}^2 \log^{-2} \|F_j\|_{L_2} + c_2^{\frac{4\beta_2}{4\beta_2+1}} \left(\frac{n}{\log n} \right)^{-\frac{2\beta_2}{2\beta 2+1}} \right.$$
$$\left. + (c_1+1)\frac{\log p}{n} + \sqrt{(c_1+1)\frac{\log p}{n}} \|F_j(x(t))\|_{L_2} + n^{-1/2} e^{-p} \right\}.$$

Proof of Lemma 2: Note that

$$\|F_j(x(t))\|_{L_2}^2 - \|F_j(x(t))\|_n^2 \leq \sup_{\substack{g \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|g\|_n^2 \right).$$

By the Talagrand's concentration inequality (Talagrand, 1996), with probability at least $1 - e^{-c_1}$,

$$\sup_{\substack{g \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|g\|_n^2 \right) \leq 2 \left\{ \mathbb{E} \sup_{\substack{g \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|g\|_n^2 \right) + 4 \|F_j(x(t))\|_{L_2} \sqrt{\frac{c_1}{n}} + \frac{16 c_1}{n} \right\}. \tag{S9}$$

By the symmetrization inequality for the Rademacher process (van der Vaart and Wellner, 1996), there exists a constant $C_1 > 0$, such that

$$\begin{aligned} \mathbb{E} \sup_{\substack{g \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|g\|_n^2 \right) &\leq \mathbb{E} \sup_{\substack{g \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \omega_i g^2(x(t_i)) \right\} \\ &\leq C_1 \mathbb{E} \sup_{\substack{g \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \omega_i g(x(t_i)) \right\}, \end{aligned} \tag{S10}$$

where $\omega, \ldots, \omega_n$ are independent random variables drawn from the Rademacher distribution; i.e., $\mathbb{P}(\omega_i = 1) = \mathbb{P}(\omega_i = -1) = 1/2$, for $i = 1, \ldots, n$. The last inequality in (S10) is due to the contraction inequality, and the fact that g^2 is a Lipschitz function. Henceforth, with the Talagrand's concentration inequality, there exists a constant $C_2 > 0$, such that, with probability at least $1 - e^{-c_1}$,

$$
\begin{aligned}
&\mathbb{E} \sup_{\substack{g \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \left\{ \frac{1}{n} \sum_{i=1}^{n} \omega_i g(x(t_i)) \right\} \\
&\leq C_2 \left[\sup_{\substack{g = \sum_{k,l=1}^{p} g_{kl} \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \sum_{k,l=1}^{p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \omega_i g_{kl}(x(t_i)) \right\} + \|F_j\|_{L_2} \sqrt{\frac{c_1}{n}} + \frac{c_1}{n} \right].
\end{aligned}
\tag{S11}
$$

By Lemma 2.2 of Yuan and Zhou (2016), and the result that the νth eigenvalue of RKHS $\mathcal{H}$ is of order $(\nu \log^{-1} \nu)^{-2\beta_2}$, for $\nu \geq 1$ (Bach, 2017), there exists a constant $C_3 > 0$, such that, with probability at least $1 - d^{-c_1}$,

$$
\begin{aligned}
&\sum_{k,l=1}^{p} \left\{ \frac{1}{n} \sum_{i=1}^{n} \omega_i g_{kl}(x(t_i)) \right\} \\
&\leq C_3 n^{-1/2} \sum_{k,l=1}^{p} \Big\{ \left(\|g_{kl}\|_{\mathcal{H}} \log^{-1} \|g_{kl}\|_{\mathcal{H}} \right)^{\frac{1}{2\beta_2}} \left(\|g_{kl}\|_{L_2} \log^{-1} \|g_{kl}\|_{L_2} \right)^{1 - \frac{1}{2\beta_2}} \\
&\qquad + \|g_{kl}\|_{L_2} \sqrt{(c_1 + 1) \log p} + e^{-p} \|g_{kl}\|_{\mathcal{H}} \Big\},
\end{aligned}
$$

Following the arguments for bounding Δ_4 in (S6), there exists a constant $C_4 > 0$ and for any $c_2 > 1$, such that

$$
\begin{aligned}
&n^{-1/2} \sup_{\substack{g = \sum_{k,l=1}^{p} g_{kl} \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \sum_{k,l=1}^{p} \left(\|g_{jkl}\|_{\mathcal{H}} \log^{-1} \|g_{jkl}\|_{\mathcal{H}} \right)^{\frac{1}{2\beta_2}} \left(\|g_{kl}\|_{L_2} \log^{-1} \|g_{kl}\|_{L_2} \right)^{1 - \frac{1}{2\beta_2}} \\
&\leq C_4 c_2^{-\frac{4\beta_2}{2\beta_2 - 1}} \sup_{\substack{g = \sum_{k,l=1}^{p} g_{kl} \in \mathcal{H}, \|g\|_{\mathcal{H}}^0 \leq \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2} \leq \|F_j\|_{L_2}}} \sum_{k,l=1}^{p} \left(\|g_{kl}\|_{L_2} \log^{-1} \|g_{kl}\|_{L_2} \right)^2 + C_4 c_2^{\frac{4\beta_2}{4\beta_2 + 1}} \left(\frac{n}{\log n} \right)^{-\frac{2\beta_2}{2\beta2 + 1}} \\
&\leq C_4 c_2^{-\frac{4\beta_2}{2\beta_2 - 1}} C_5 \|F_j(x(t))\|_{L_2}^2 \log^{-2} \|F_j(x(t))\|_{L_2} + C_4 c_2^{\frac{4\beta_2}{4\beta_2 + 1}} \left(\frac{n}{\log n} \right)^{-\frac{2\beta_2}{2\beta2 + 1}},
\end{aligned}
$$

where the last step is due to $\sum_{k,l=1}^{p} \left(\|F_{jkl}\|_{L_2} \log^{-1} \|F_{jkl}\|_{L_2} \right)^2 \leq C_5 \left(\|F_j\|_{L_2} \log^{-1} \|F_j\|_{L_2} \right)^2$ for some constant $C_5 > 1$. Following the arguments for bounding Δ_5 in (S6), there exists a constant $C_6 > 0$, such that

$$
\sum_{k,l=1}^{p} \|g_{kl}\|_{L_2} \leq C_6 \left\{ \sqrt{\frac{\log p}{n}} + \|F_j(x(t))\|_{L_2} \right\}.
$$

Henceforth, for some constant $C_7 > 0$,

$$\begin{aligned}
&\sup_{\substack{g=\sum_{k,l=1}^{p} g_{kl}\in\mathcal{H}, \|g\|_{\mathcal{H}}^0 \le \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\le \|F_j\|_{L_2}}} \sum_{k,l=1}^{p} \left\{ \frac{1}{n}\sum_{i=1}^{n} \omega_i g_{kl}(x(t_i)) \right\} \\
&\le C_7 \left\{ c_2^{-\frac{4\beta_2}{2\beta_2-1}} \|F_j(x(t))\|_{L_2}^2 \log^{-2} \|F_j(x(t))\|_{L_2} + c_2^{\frac{4\beta_2}{4\beta_2+1}} \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} \right\} \\
&\quad + C_7 \sqrt{\frac{(c_1+1)\log p}{n}} \left\{ \sqrt{\frac{\log p}{n}} + \|F_j(x(t))\|_{L_2} \right\} + C_7 n^{-1/2} e^{-p}.
\end{aligned}$$

Together with (S9), (S10), and (S11), we have, with probability at least $1 - 2e^{-c_1}$,

$$\begin{aligned}
&\sup_{\substack{g\in\mathcal{H}, \|g\|_{\mathcal{H}}^0 \le \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\le \|F_j\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|g\|_n^2 \right) \\
&\le C_8 \left\{ c_2^{-\frac{4\beta_2}{2\beta_2-1}} \|F_j(x(t))\|_{L_2}^2 \log^{-2} \|F_j(x(t))\|_{L_2} + c_2^{\frac{4\beta_2}{4\beta_2+1}} \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} \right\} \\
&\quad + C_8 \sqrt{\frac{(c_1+1)\log p}{n}} \left\{ \sqrt{\frac{\log p}{n}} + \|F_j(x(t))\|_{L_2} \right\} + C_8 n^{-1/2} e^{-p} \\
&\quad + C_8 \left(\|F_j(x(t))\|_{L_2} \sqrt{\frac{c_1}{n}} + \frac{c_1}{n} \right),
\end{aligned}$$

for some constant $C_8 > 0$. Using the change of variable, the following result also holds. That is, with probability at least $1 - 2p^{-c_1}$, it holds that,

$$\begin{aligned}
&\|F_j(x(t))\|_{L_2}^2 - \|F_j(x(t))\|_n^2 \le \sup_{\substack{g\in\mathcal{H}, \|g\|_{\mathcal{H}}^0 \le \|F_j\|_{\mathcal{H}}^0 \\ \|g\|_{L_2}\le \|F_j\|_{L_2}}} \left(\|g\|_{L_2}^2 - \|g\|_n^2 \right) \\
&\le C_9 \left\{ c_2^{-\frac{4\beta_2}{2\beta_2-1}} \|F_j(x(t))\|_{L_2}^2 \log^{-2} \|F_j(x(t))\|_{L_2} + c_2^{\frac{4\beta_2}{4\beta_2+1}} \left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta 2+1}} \right\} \\
&\quad + C_9 \sqrt{\frac{(c_1+1)\log p}{n}} \left\{ \sqrt{\frac{\log p}{n}} + \|F_j(x(t))\|_{L_2} \right\} + C_8 n^{-1/2} e^{-p} \\
&\quad + C_9 \left\{ \|F_j(x(t))\|_{L_2} \sqrt{\frac{(c_1+1)\log p}{n}} + \frac{(c_1+1)\log p}{n} \right\},
\end{aligned}$$

for some constant $C_9 > 0$. This completes the proof of Lemma 2. □

S1.6 Proof of Theorem 5

We divide the proof of this theorem to three parts. We first present the main proof in Section S1.6.1. We then summarize some additional technical assumptions used during the proof in Section S1.6.2. We give an auxiliary lemma in Section S1.6.3.

S1.6.1 Main proof

We use the primal-dual witness method to prove that KODE selects all significant variables but includes no insignificant ones. The analysis here extends the techniques in Ravikumar et al. (2010) for the Ising model, where the pairwise interactions have a simple product form. Meanwhile, we also deal with measurement errors in variables.

Consider the optimization problem (11) that is equivalent to (10). Recall that, by the representer theorem (Wahba, 1990), the selection problem becomes (15); i.e.,

$$\min_{\theta_j} \left\{ (z_j - G\theta_j)^\top (z_j - G\theta_j) + n\kappa_{nj} \left(\sum_{k=1}^{p} \theta_{jk} + \sum_{k\neq l, k=1}^{p} \sum_{l=1}^{p} \theta_{jkl} \right) \right\}, \tag{S12}$$

subject to $\theta_k \geq 0, \theta_{kl} \geq 0, k, l = 1, \ldots, p, k \neq l$, where the "response" is $z_j = (y_j - \bar{y}_j) - (1/2)n\eta_{nj}c_j - Bb_j$, and the "predictor" is $G \in \mathbb{R}^{n\times p^2}$. The vector θ solves (S12) if it satisfies the Karush-Kuhn-Tucker (KKT) condition:

$$\frac{2}{n} G^\top (G\theta_j - z_j) + \kappa_{nj} g_j = 0, \quad j = 1, \ldots, p, \tag{S13}$$

where G contains errors in variable due to the estimated $\widehat{x}(t)$, and

$$g_j = \operatorname{sign}(\theta_j) \text{ if } \theta_j \neq 0, \quad \text{ and } \quad |g_j| \leq 1 \text{ otherwise.} \tag{S14}$$

To apply the primal-dual witness method, we next construct an oracle primal-dual pair $(\widehat{\theta}_j, \widehat{g}_j)$ satisfying the KKT conditions (S13) and (S14). Specifically,

(a) We set $\widehat{\theta}_{jkl} = 0$ for $(k, l) \notin S_j$, where S_j is defined as,

$$S_j \equiv \{1 \leq k \leq l \leq p: \text{ if } F_{jk} \neq 0, \text{ let } l = k; \text{ or if } F_{jkl} \neq 0 \text{ with } l \neq k \leq p\}.$$

The definition of S_j is similar to M_j defined in Section 3.2. However, S_j explicitly includes $\{(k, l) : k = l\}$. Let $s_j = \operatorname{card}(S_j)$.

(b) Let $\widehat{\theta}_{S_j}$ be the minimizer of the partial penalized likelihood,

$$(z_j - G_{S_j}\theta_{S_j})^\top (z_j - G_{S_j}\theta_{S_j}) + n\kappa_{nj} \left(\sum_{k=1}^{p} \theta_{jk} + \sum_{k\neq l, k=1}^{p} \sum_{l=1}^{p} \theta_{jkl} \right). \tag{S15}$$

(c) Let S_j^c be the complement of S_j in $\{(k, l) : 1 \leq k \leq l \leq p\}$. We obtain $\widehat{g}_{S_j^c}$ from (S13) by substituting in the values of $\widehat{\theta}_j$ and $\widehat{g}_{S_j}$.

Next, we verify the support recovery consistency; i.e.,

$$\max_{(k,l)\in S_j} \|\widehat{\theta}_{jkl} - \theta_{jkl}\|_{\ell_2} \leq \frac{2}{3}\theta_{\min},$$

which in turn implies that the oracle estimator $\widehat{\theta}_j$ recovers the support of θ_j exactly.

Note that the subgradient condition for the partial penalized likelihood (S15) is

$$2G_{S_j}^{\top}(G_{S_j}\widehat{\theta}_{S_j} - z_j) + n\kappa_{nj}\widehat{g}_{S_j} = 0,$$

which implies that

$$2G_{S_j}^{\top}(G_{S_j}\widehat{\theta}_{S_j} - G_{S_j}\theta_{S_j}) + 2G_{S_j}^{\top}(G_{S_j}\theta_{S_j} - z_j) + n\kappa_{nj}\widehat{g}_{S_j} = 0.$$

Define $\mathcal{R}_{S_j} \equiv 2G_{S_j}^{\top}G_{S_j}\theta_{S_j} - 2G_{S_j}^{\top}z_j$. Then,

$$\widehat{\theta}_{S_j} - \theta_{S_j} = -\left(2G_{S_j}^{\top}G_{S_j}\right)^{-1}(\mathcal{R}_{S_j} + n\kappa_{nj}\widehat{g}_{S_j}). \tag{S16}$$

For each (k,l), denote the corresponding column of G by G_{kl}. Then for $(k,l) \in S_j$,

$$\mathcal{R}_{kl} = 2G_{kl}^{\top}G_{S_j}\theta_{S_j} - 2G_{kl}^{\top}z_j. \tag{S17}$$

By Lemma 3, we have $\|\mathcal{R}_{kl}\|_{\ell_2} \leq \eta_{\mathcal{R}}$ for any $(k,l) \in S_j$. Then,

$$\|\mathcal{R}_{S_j}\|_{\ell_2} \leq \eta_{\mathcal{R}}\sqrt{s_j}. \tag{S18}$$

By Assumption 3 given in Section S1.6.2, we have $\Lambda_{\min}\left(G_{S_j}^{\top}G_{S_j}\right) \geq C_{\min}/2$, for some constant $C_{\min} > 0$. Henceforth,

$$\Lambda_{\max}\left\{\left(2G_{S_j}^{\top}G_{S_j}\right)^{-1}\right\} \leq \frac{1}{C_{\min}}.$$

Note that for any $(k,l) \in S_j$, $\|\widehat{g}_{jkl}\|_{\ell_2} \leq 1$, which implies that,

$$\|\widehat{g}_{S_j}\|_{\ell_2} \leq \sqrt{s_j}. \tag{S19}$$

Therefore,

$$\max_{(k,l)\in S_j} \|\widehat{\theta}_{jkl} - \theta_{jkl}\|_{\ell_2} \leq \|\widehat{\theta}_{S_j} - \theta_{S_j}\|_{\ell_2} \leq \frac{\eta_{\mathcal{R}}\sqrt{s_j}}{C_{\min}} + n\kappa_{nj}\frac{\sqrt{s_j}}{C_{\min}} \leq \frac{2}{3}\theta_{\min}.$$

where the last inequality is due to Assumption 5 in Section S1.6.2.

Next, we verify the strict dual feasibility; i.e.,

$$\max_{(k,l)\notin S_j} |\widehat{g}_{jkl}| < 1,$$

which in turn implies that the oracle estimator $\widehat{\theta}_j$ satisfies the KKT condition of the KODE optimization problem.

For any $(k,l) \notin S_j$, by (S13), we have,

$$2G_{kl}^{\top}(G_{S_j}\widehat{\theta}_{S_j} - z_j) + n\kappa_{nj}\widehat{g}_{jkl} = 0,$$

which implies that

$$2G_{kl}^{\top}(G_{S_j}\widehat{\theta}_{S_j} - G_{S_j}\theta_{S_j}) + 2G_{kl}^{\top}(G_{S_j}\theta_{S_j} - z_j) + n\kappa_{nj}\widehat{g}_{jkl} = 0.$$

By (S16) and (S17), we have,

$$n\kappa_{nj}\widehat{g}_{jkl} = G_{kl}^{\top}G_{S_j}(G_{S_j}^{\top}G_{S_j})^{-1}(\mathcal{R}_{S_j} + n\kappa_{nj}\widehat{g}_{S_j}) - \mathcal{R}_{kl}.$$

By Assumption 4 in Section S1.6.2, we have that,

$$\max_{(k,l)\notin S_j} \left\| G_{kl}^{\top}G_{S_j}(G_{S_j}^{\top}G_{S_j})^{-1} \right\|_{\ell_2} \leq \xi_G.$$

Then by (S18) and (S19), we have that

$$|\widehat{g}_{jkl}| \leq \frac{(\xi_G + 1)\sqrt{s_j}}{n\kappa_{nj}}\eta_{\mathcal{R}} + \xi_G\sqrt{s_j}, \quad (k,l) \notin S_j.$$

By Assumption 5 in Section S1.6.2 that

$$\frac{(\xi_G + 1)\sqrt{s_j}}{n\kappa_{nj}}\eta_{\mathcal{R}} + \xi_G\sqrt{s_j} < 1,$$

we obtain that,

$$|\widehat{g}_{jkl}| < 1, \quad \text{for any } (k,l) \notin S_j.$$

Finally, the selection consistency for S_j implies the selection consistency for S_j^0. This completes the proof of Theorem 5. □

S1.6.2 Additional technical assumptions

We summarize the additional assumptions used during the proof of Theorem 5.

Assumption 3. *Suppose there exists a constant $C_{\min} > 0$ such that the minimal eigenvalue of matrix $G_{S_j}^{\top}G_{S_j}$ satisfies,*

$$\Lambda_{\min}\left(G_{S_j}^{\top}G_{S_j}\right) \geq \frac{1}{2}C_{\min}.$$

Assumption 4. *Suppose there exists a constant* $0 \le \xi_G < 1$ *such that,*

$$\max_{(k,l)\notin S_j} \left\| G_{kl}^\top G_{S_j} (G_{S_j}^\top G_{S_j})^{-1} \right\|_{\ell_2} \le \xi_G.$$

Assumption 5. *Suppose the following inequalities hold:*

$$\frac{\eta_{\mathcal{R}}\sqrt{s_j}}{C_{\min}} + n\kappa_{nj}\frac{\sqrt{s_j}}{C_{\min}} \le \frac{2}{3}\theta_{\min}, \qquad \textit{and} \qquad \frac{(\xi_G+1)\sqrt{s_j}}{n\kappa_{nj}}\eta_{\mathcal{R}} + \xi_G\sqrt{s_j} < 1.$$

where $\theta_{\min} = \min_{(k,l)\in S_j} \|\theta_{jkl}\|_{\ell_2}$.

Assumption 3 ensures the identifiability among the s_j elements in the column set of G_{S_j}. The same condition has been used in Zhao and Yu (2006); Ravikumar et al. (2010); Chen et al. (2017). Assumption 4 reflects the intuition that the large number of irrelevant variables cannot exert an overly strong effect on the subset of relevant variables. This condition is standard in the literature of Lasso regressions (Meinshausen et al., 2006; Zhao and Yu, 2006; Ravikumar et al., 2010). Assumption 5 imposes some regularity on the minimum regulatory effect. The second inequality characterizes the relationship between the quantities ξ_G, the sparse tuning parameter κ_{nj}, and the sparsity level s_j. Similar assumptions have been used in Lasso regressions (Meinshausen et al., 2006; Zhao and Yu, 2006; Ravikumar et al., 2010).

Next, we detail Assumptions 4 and 5 in three specific examples, which help provide a better interpretation of these hypotheses. In particular, Meinshausen et al. (2006); Zhao and Yu (2006) provided some examples and results for the setting of classical regressions. We show that similar results hold for KODE for the dynamic system. Recall the definition of the "predictor" $G \in \mathbb{R}^{n\times p^2}$ in (15), where the first p columns of G are $\Sigma^k c_j$ with $k = 1, \dots, p$, and the last $p(p-1)$ columns of G are $\Sigma^{kl} c_j$ with $k, l = 1, \dots, p, k \ne l$. All diagonal elements of $G^\top G$ are assumed to be 1, which is equivalent to normalizing $\Sigma^{kl} c_j$ to the same scale for any $k, l = 1, \dots, p$, since Assumption 4 is invariant under a common scaling of $G^\top G$.

The first example considers bounded correlations of functional component estimates $\Sigma^{kl} c_j$ for all $k, l = 1, \dots, p$. It implies Assumption 4 holds even when p grows with n, as long as s_j remains fixed, which in turn ensures that KODE selects the true model asymptotically.

Example 1. *Suppose that the correlation of* $\Sigma^{kl} c_j$ *and* $\Sigma^{k'l'} c_j$ *is bounded by* $\xi_G/(2s_j - 1)$, $j, k, l, k', l' = 1, \dots, p$. *Then Assumption 4 holds.*

Proof: Recall that $C_{\min}/2$ is a lower bound of the minimum eigenvalue of $G_{S_j}^\top G_{S_j}$ defined in Assumption 3. We can bound $C_{\min}$ as follows. Let $u = (u_1, \dots, u_{s_j})^\top \in \mathbb{R}^{s_j}$. Since the correlation of $\Sigma^{kl} c_j$ and $\Sigma^{k'l'} c_j$ is bounded by $\frac{\xi_G}{2s_j - 1}$ for any $k, l, k', l' = 1, \dots, p$, any off-diagonal element of $G^\top G$ is bounded by $\xi_G/(2s_j - 1)$. Then

$$\begin{aligned}u^\top(G_{S_j}^\top G_{S_j})u &= 1+\sum_{1\le i_1\ne i_2\le s_j} u_{i_1}(G_{S_j}^\top G_{S_j})_{(i_1,i_2)}u_{i_2} \ge 1-\frac{\xi_G}{2s_j-1}\sum_{1\le i_1\ne i_2\le s_j}|u_{i_1}||u_{i_2}| \\ &\ge 1-\frac{1}{2s_j-1}\sum_{1\le i_1\ne i_2\le s_j}|u_{i_1}||u_{i_2}| \ge \frac{s_j}{2s_j-1}.\end{aligned}$$

Therefore, $C_{\min}\ge 2s_j/(2s_j-1)$, and

$$\left\|G_{kl}^\top G_{S_j}(G_{S_j}^\top G_{S_j})^{-1}\right\|_{\ell_2} \le \|G_{kl}^\top G_{S_j}\|_{\ell_2}\frac{2}{C_{\min}}\sqrt{s_j} \le \frac{\xi_G\sqrt{s_j}}{2s_j-1}\cdot\frac{2s_j-1}{s_j}\cdot\sqrt{s_j}=\xi_G.$$

This verifies Example 1. □

The second example gives two instances of Example 1, where Assumption 4 holds under some simplified structures.

Example 2. *Assumption 4 holds if (i) $s_j=1$, or (ii) $G^\top G$ is orthogonal.*

Proof: (i) If $s_j=1$, then $\xi_G=\max_{(k,l)\ne(k',l')}(\Sigma^{kl}c_j)^\top\Sigma^{k'l'}c_j<1$. Then the condition in Example 1 holds. (ii) If the matrix $G^\top G$ is orthogonal, the correlation of $\Sigma^{kl}c_j$ and $\Sigma^{k'l'}c_j$ is zero for any $(k,l)\ne(k',l')$, and thus the condition in Example 1 holds for any s_j and ξ_G.

Therefore, in both cases, Assumption 4 holds. This verifies Example 2. □

The third example illustrates Assumption 5 with a natural condition that the minimal signal term $\theta_{\min}$ does not decay too fast. In particular, it is necessary to have a gap between the decay rate of the minimal signal and $n^{-1/2}$. Since the noise aggregates at a rate of $n^{-1/2}$, this condition prevents the estimation and selection from being dominated by the noise.

Example 3. *Suppose that $\xi_G<s_j^{-1/2}$, and*

$$\theta_{\min}=O\left[(\log p)^{\epsilon_p}\left\{\left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}}+\left(\frac{\log p}{n}\right)^{\frac{1}{2}}+n^{-\frac{\beta_1}{2\beta_1+1}}\right\}\right],$$

with $\epsilon_p>0$. Here, $\theta_{\min}$ decays at the rate slower than $n^{-1/2}$ as n and p grow. Then there exists $\kappa_{nj}>0$, such that Assumption 5 holds.

Proof: Recall that

$$\eta_{\mathcal{R}}=O_p\left\{\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}}+\frac{\log p}{n}+n^{-\frac{2\beta_1}{2\beta_1+1}}\right\}.$$

Then $\mathbb{P}(\theta_{\min}\ge 3c_\theta\eta_{\mathcal{R}}/2)\to 1$ for any constant $c_\theta>0$ as n and p grow. Letting $\kappa_{nj}=c_\kappa n^{-1}\eta_{\mathcal{R}}$, with

$$0<c_\kappa\le\frac{c_\theta C_{\min}}{\sqrt{s_j}}-1, \tag{S20}$$

then the first inequality of Assumption 5 holds. Moreover, letting $\kappa_{nj} = c_\kappa n^{-1}\eta_{\mathcal{R}} > 0$, with

$$c_\kappa > \frac{(\xi_G + 1)\sqrt{s_j}}{1 - \xi_G\sqrt{s_j}}, \tag{S21}$$

then the second inequality of Assumption 5 also holds.

By setting

$$c_\theta > \frac{(1 + \sqrt{s_j})\sqrt{s_j}}{(1 - \xi_G\sqrt{s_j})C_{\min}},$$

there exists c_k satisfying both (S20) and (S21), as long as $\xi_G < 1/\sqrt{s_j}$. Therefore, there exists κ_{nj} given by $c_\kappa n^{-1}\eta_{\mathcal{R}}$, such that Assumption 5 holds. This verifies Example 3. □

S1.6.3 Auxiliary lemma for Theorem 5

We present a lemma that is useful for the proof of Theorem 5. It gives a bound similar to the deviation condition proposed by Loh and Wainwright (2012). The difference is that, the noise in variable $\widehat{x}(t)$ in our setting involves a nonlinear transformation through the kernel $K(\widehat{x}(t), \widehat{x}(s))$.

Lemma 3. *For $j = 1, \ldots, p$, we have,*

$$\|G_{kl}^\top G_{S_j}\theta_{S_j} - G_{kl}z_j\|_{\ell_2} \le \eta_{\mathcal{R}}, \quad \textit{where} \;\; \eta_{\mathcal{R}} = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}} + \left(\frac{\log p}{n}\right)^{1/2} + n^{-\frac{\beta_1}{2\beta_1+1}}\right).$$

Proof of Lemma 3: Similar to the "predictor" G defined in (15) in Section 3.1, we first construct a noiseless version of the predictor $\widetilde{G} \in \mathbb{R}^{n\times p^2}$, whose first p columns are $\widetilde{\Sigma}^k c_j$, and the last $p(p-1)$ columns are $\widetilde{\Sigma}^{kl}c_j$, and $\widetilde{\Sigma}^k = (\widetilde{\Sigma}^k_{ii'}), \widetilde{\Sigma}^{kl} = (\widetilde{\Sigma}^{kl}_{ii'})$ are both $n \times n$ matrices whose (i, i')th entries are,

$$\widetilde{\Sigma}^k_{ii'} = \int_{\mathcal{T}}\int_{\mathcal{T}} \{T_i(s) - \bar{T}(s)\}K_k(x(t), x(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt, \;\; 1 \le k \le p, 1 \le i, i' \le n,$$
$$\widetilde{\Sigma}^{kl}_{ii'} = \int_{\mathcal{T}}\int_{\mathcal{T}} \{T_i(s) - \bar{T}(s)\}K_{kl}(x(t), x(s))\{T_{i'}(t) - \bar{T}(t)\}dsdt, \;\; 1 \le k < l \le p, 1 \le i, i' \le n.$$

Next, we consider the term $\left\|G_{kl}^\top z_j - G_{kl}^\top G_{S_j}\theta_{S_j}\right\|_{\ell_2}$, which can be bounded as,

$$\begin{aligned}\left\|G_{kl}^\top z_j - G_{kl}^\top G_{S_j}\theta_{S_j}\right\|_{\ell_2} \le & \left\|G_{kl}^\top \mathbb{E}[z_j] - G_{kl}^\top \widetilde{G}_{S_j}\theta_{S_j}\right\|_{\ell_2} + \left\|G_{kl}^\top(\widetilde{G}_{S_j} - G_{S_j})\theta_{S_j}\right\|_{\ell_2} \\ & + \left\|G_{kl}^\top(z_j - \mathbb{E}[z_j])\right\|_{\ell_2} \equiv \Delta_7 + \Delta_8 + \Delta_9.\end{aligned} \tag{S22}$$

We next bound the three terms $\Delta_7, \Delta_8, \Delta_9$ on the right-hand-side of S22, respectively.

For Δ_7, by the Cauchy-Schwarz inequality, we have,

$$\Delta_7^2 \leq \left\|G_{kl}^\top\right\|_{\ell_2}^2 \left\|\mathbb{E}[z_j] - \widetilde{G}_{S_j}\theta_{S_j}\right\|_{\ell_2}^2 \leq C_1 \left\|\mathbb{E}[z_j] - \widetilde{G}_{S_j}\theta_{S_j}\right\|_{\ell_2}^2 = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{\log p}{n}\right),$$

for some constant $C_1 > 0$, where the last step is by (S5).

For Δ_8, again by the Cauchy-Schwarz inequality, we have,

$$\Delta_8^2 \leq \left\|G_{kl}^\top\right\|_{\ell_2}^2 \left\|\left(\widetilde{G}_{S_j} - G_{S_j}\right)\theta_{S_j}\right\|_{\ell_2}^2 \leq C_2 \left\|\left(\widetilde{G}_{S_j} - G_{S_j}\right)\right\|_{\infty}^2 \left\|\theta_{S_j}\right\|_{\ell_1}^2 = O_p\left(n^{-\frac{2\beta_1}{2\beta_1+1}}\right),$$

for some constants $C_2 > 0$, where the last step is by (S4) and the fact that $\|\theta_{S_j}\|_{\ell_1}$ is bounded.

For Δ_9, by Lemma 2, we have,

$$\Delta_9^2 = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{2\beta_2}{2\beta_2+1}} + \frac{\log p}{n}\right).$$

Combining the above three bounds, we obtain that,

$$\left\|G_{kl}^\top z_j - G_{kl}^\top G_{S_j}\theta_{S_j}\right\|_{\ell_2} = O_p\left(\left(\frac{n}{\log n}\right)^{-\frac{\beta_2}{2\beta_2+1}} + \left(\frac{\log p}{n}\right)^{1/2} + n^{-\frac{\beta_1}{2\beta_1+1}}\right),$$

which completes the proof of Lemma 3. □

S2 Additional numerical results

We report some additional numerical results. We begin with a comparison with a family of ODE solutions assuming a known functional F. We then carry out a sensitivity analysis to study the robustness of the choice of kernel function and initial parameters. Finally, we report the sparse recovery of the enzymatic regulatory network example studied in Section 5.2, and the gene regulatory network example studied in Section 6 of the paper.

S2.1 Comparison with alternative methods

We compare the proposed KODE method with a family of alternative ODE solutions, including González et al. (2014); Zhang et al. (2015b); Mikkelsen and Hansen (2017). Particularly, González et al. (2014) proposed a penalized log-likelihood approach in RKHS, where the ODE system is used as a penalty. Zhang et al. (2015b) studied a full predator-prey ODE model that takes a special form of two-dimensional rational ODE. Mikkelsen and Hansen

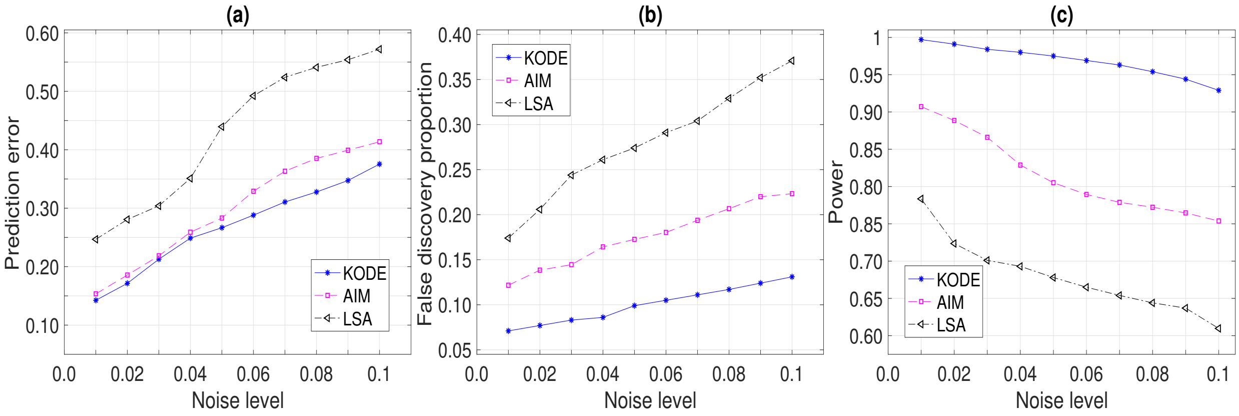

Figure S1: The prediction and selection performance of KODE, LSA and AIM with varying noise level. The results are averaged over 500 data replications. (a) Prediction error; (b) False discovery proportion; (c) Empirical power.

(2017) learned a class of polynomial or rational ODE systems. However, the main difference is that, all those solutions assumed the forms of the functional F are completely known, while KODE does not require so, but instead estimates the functional adaptively given the data. In addition, both González et al. (2014) and Zhang et al. (2015b) focused on the low-dimensional ODE, while our method works for both low-dimensional and high-dimensional ODE. Moreover, none of those solutions tackled post-selection inference, while we do. These differences clearly distinguish our proposal from those existing ones.

Next, we numerically compare KODE with the least squares approximation (LSA) method of Zhang et al. (2015b), and the adaptive integral matching (AIM) method of Mikkelsen and Hansen (2017). We did not include González et al. (2014) in our numerical comparison, since their code is not available. Moreover, they also assumed a known F similarly as the other two works. We implement LSA using the code provided by the Wiley Online Library, and AIM using the R package `episode` with the Lasso penalty and automatic adaptation of parameter scales.

Figure S1 reports the performance of KODE, LSA and AIM for the enzymatic regulatory network example in Section 5.2. The results are average over 500 data replications. It is seen that KODE clearly outperforms both LSA and AIM in terms of both prediction and selection accuracy. This suggests that the polynomial or rational forms of F imposed by LSA and AIM may not hold in this example. Table S1 reports the performance of the three methods for the gene regulatory network example in Section 6. The results are averaged over 100 data realizations for all ten combinations of network structures. Again, it is seen that KODE clearly outperforms LSA and AIM in all cases.

Table S1: The area under the ROC curve and the 95% confidence interval of KODE, LSA and AIM, for 10 combinations of network structures from GNW. The results are averaged over 100 data replications.

	$p = 10$			$p = 100$		
	KODE	AIM	LSA	KODE	AIM	LSA
E.coli1	**0.582**	0.558	0.437	**0.711**	0.698	0.614
	(0.577, 0.587)	(0.554, 0.562)	(0.428, 0.446)	(0.708, 0.714)	(0.694, 0.702)	(0.606, 0.622)
E.coli2	**0.662**	0.646	0.541	**0.685**	0.678	0.501
	(0.658, 0.666)	(0.641, 0.651)	(0.533, 0.549)	(0.681, 0.689)	(0.673, 0.683)	(0.488, 0.514)
Yeast1	**0.603**	0.539	0.413	**0.619**	0.599	0.527
	(0.599, 0.607)	(0.534, 0.544)	(0.401, 0.425)	(0.616, 0.622)	(0.594, 0.604)	(0.511, 0.543)
Yeast2	**0.599**	0.559	0.497	**0.606**	0.577	0.518
	(0.595, 0.603)	(0.554, 0.564)	(0.488, 0.506)	(0.603, 0.609)	(0.573, 0.581)	(0.505, 0.531)
Yeast3	**0.612**	0.563	0.451	**0.621**	0.609	0.577
	(0.608, 0.616)	(0.558, 0.567)	(0.440, 0.462)	(0.617, 0.625)	(0.604, 0.614)	(0.562, 0.592)

Together with the numerical results that compare KODE with linear ODE (Zhang et al., 2015a) and additive ODE (Chen et al., 2017) reported in the paper, it demonstrates that the proposed KODE is a competitive and useful tool for modeling complex dynamic systems.

S2.2 Sensitivity analysis

We carry out a sensitivity analysis to investigate the robustness of the choice of kernel function and initial parameters in KODE.

First, we consider three commonly used kernels and study their performances using the enzymatic regulatory network example in Section 5.2. Recall the first-order Matérn kernel used in our analysis in Section 5.2,

$$K_{\mathcal{F}}^{(1)}(x, x') = (1 + \sqrt{3}\|x - x'\|/\nu)\exp(-\sqrt{3}\|x - x'\|/\nu).$$

In addition, we also consider the second-order Matérn kernel,

$$K_{\mathcal{F}}^{(2)}(x, x') = (1 + \sqrt{5}\|x - x'\|/\nu + 5\|x - x'\|^2/3\nu^2)\exp(-\sqrt{5}\|x - x'\|/\nu),$$

and the Gaussian kernel,

$$K_{\mathcal{F}}^{(3)}(x, x') = \exp(-\|x - x'\|^2/2\nu^2).$$

It is known that the RKHS generated by $K_{\mathcal{F}}^{(1)}$ and $K_{\mathcal{F}}^{(2)}$ contains once differentiable and twice differentiable functions, respectively (Gneiting et al., 2010), while the RKHS generated by $K_{\mathcal{F}}^{(3)}$ contains infinitely differentiable functions (Lin and Brown, 2004). We couple the

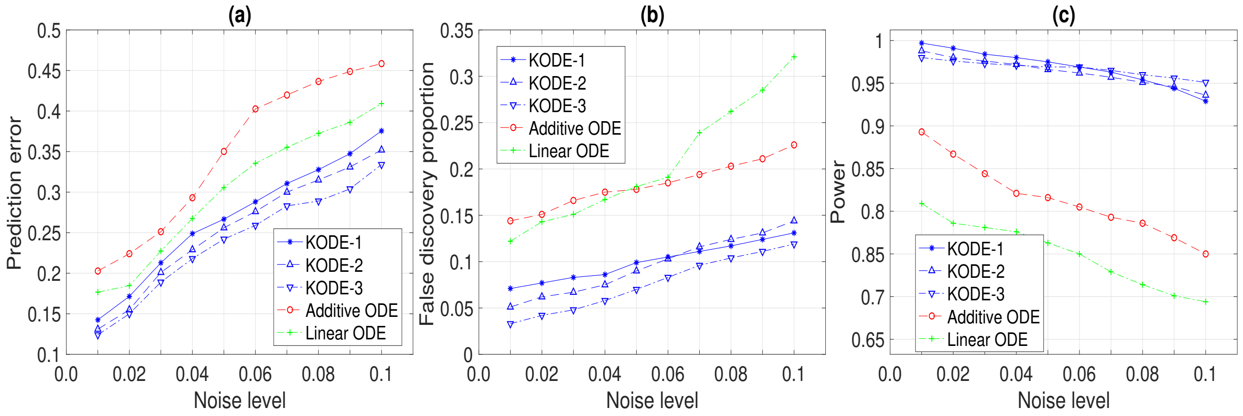

Figure S2: The prediction and selection performance with varying noise level for KODE with three different kernels, plus linear ODE and additive ODE. The results are averaged over 500 data replications. (a) Prediction error; (b) False discovery proportion; (c) Empirical power.

proposed KODE method with these three kernels: KODE-1 with the first-order Matérn kernel, KODE-2 with the second-order Matérn kernel, and KODE-3 with the Gaussian kernel. We continue to choose the bandwidth ν using tenfold cross-validation.

Figure S2 reports the prediction and selection performance of the KODE method with the three kernels, plus the linear and additive ODE methods. It is seen that the performances of the three KODE methods are fairly close. The relative prediction errors differ at most 14.9%, the false discovery proportions differ at most 4.0%, and the empirical powers differ at most 2.2%, across different noise levels. Besides, they all outperform the linear and additive ODE considerably. These results demonstrate that the proposed KODE is relatively robust to the choice of the kernel function.

We also make remarks about some general principles of choosing kernel functions. In practice, if there is prior knowledge about the smoothness of the trajectory x and the ODE system F, we may choose kernels such that the corresponding RKHS has the same order of smoothness as x and F, respectively. In this case, Theorems 3 and 4 ensure that KODE is minimax optimal for the estimation of x and F. On the other hand, if there is no such prior knowledge, we may then choose kernels with a higher-order smoothness. This recommendation is supported by the observation that the performance of KODE-3 based on the Gaussian kernel is slightly better than those of KODE-1 and KODE-2 based on the Matérn kernels when the noise level is large. This recommendation also agrees with the usual recommendation in classical kernel learning, e.g., density estimation (Hall and Marron, 1988), and nonparametric function estimation (Lin and Brown, 2004). Lastly, we comment that one can use cross-validation to choose the best performing kernel function as

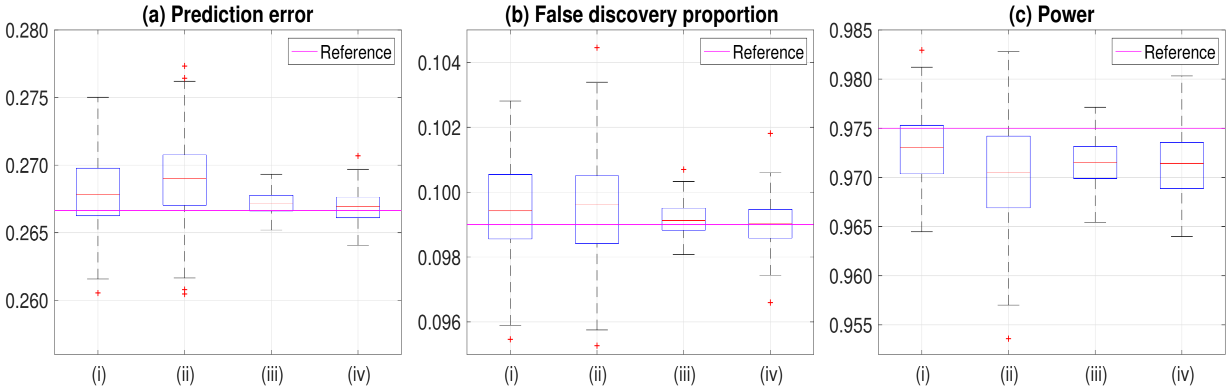

Figure S3: The prediction and selection performance of KODE with different initialization schemes. The boxes range from the lower to the upper quartile, and the whiskers extend to the most extreme data point that is no more than 1.5 times the interquartile range from the box. The solid horizontal lines denote the reference initialization scheme. The results are averaged over 500 data replications. (a) Prediction error; (b) False discovery proportion; (c) Empirical power.

well (Duan et al., 2003; Meyer et al., 2003).

Next, we study the sensitivity of KODE with respect to the choice of initial parameters. Toward that end, we divide the parameters in Algorithm 1 that require initialization into four subsets: (i) the parameters: $\{\theta_{jk}, j, k = 1, \ldots, p\}$; (ii) the parameters: $\{\theta_{jkl}, j, k, l = 1, \ldots, p, k \neq l\}$; (iii) the tuning parameters: $\{\eta_{nj}, j = 1, \ldots, p\}$; and (iv) the tuning parameters: $\{\kappa_{nj}, j = 1, \ldots, p\}$. We then consider the following initialization schemes: First, we uniformly draw 200 i.i.d. vectors from $[0.5, 1.5]^{p^2}$ as the initial values for $\{\theta_{jk}, j, k = 1, \ldots, p\}$ in (i), and fix the initial values of the parameters in (ii) to (iv) to 1. Second, we uniformly draw 200 i.i.d. vectors from $[0.5, 1.5]^{p^2(p-1)}$ as the initial values for $\{\theta_{jkl}, j, k, l = 1, \ldots, p, k \neq l\}$ in (ii), and fix the initial values of the rest of the parameters in (i), (iii) and (iv) to 1. Third, we uniformly draw 200 i.i.d. values from $[10^{-5}, 1]$ as the initial for $\{\eta_{nj}, j = 1, \ldots, p\}$ in (iii), and fix the initial values of the rest of parameters to 1. Fourth, we uniformly draw 200 i.i.d. values from $[10^{-5}, 1]$ as the initial for $\{\kappa_{nj}, j = 1, \ldots, p\}$ in (iv), and fix the initial values of the rest of parameters to 1. Finally, we initialize all the parameters in (i) to (iv) to 1, and take this as a reference. For each setting of parameter initialization, we apply Algorithm 1 to the enzymatic regulatory network example in Section 5.2 with the noise level $\sigma_j = 0.05, j = 1, 2, 3$.

Figure S3 reports the prediction and selection performance of KODE with different schemes of parameter initialization, and the results are averaged over 500 data replications. It is seen that the performances under different initialization schemes are close. For the

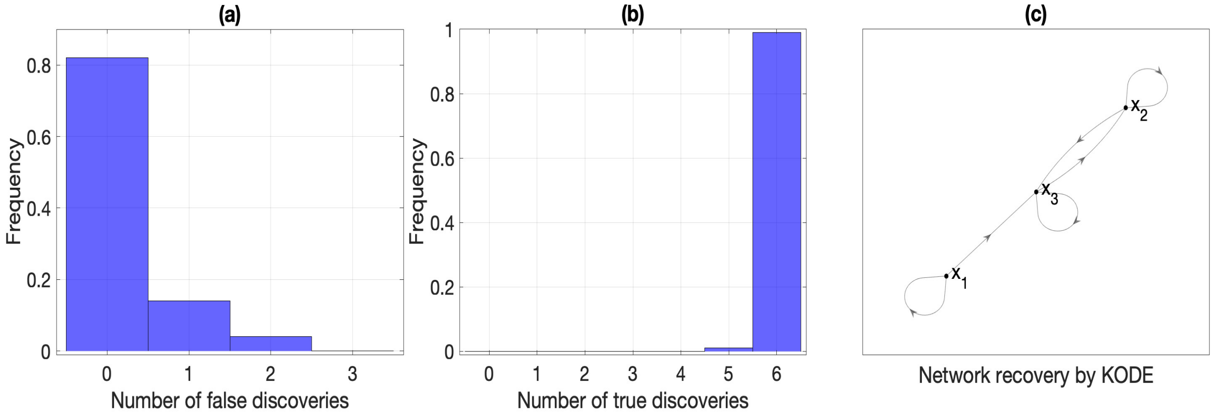

Figure S4: The selection performance of KODE for the enzymatic regulatory network, based on 500 data replications. (a) Number of false discoveries in the estimated model. (b) Number of true discoveries in the estimated model. (c) Network recovery thresholded at the 90% frequency.

majority of cases, the relative prediction errors to the reference differ at most 1.9%, the false discovery proportions differ at most 0.2% compared to the reference, and the empirical powers differ at most 0.8% compared to the reference. This example shows that the proposed KODE is fairly robust to the choice of the initial values. In the paper, we simply employ the reference initialization scheme, i.e., setting all the initial values to 1.

S2.3 Enzymatic regulatory network recovery by KODE

Figure 3(b)-(c) in Section 5.2 of the paper reported the selection performance of KODE in terms of false discovery proportion and power when recovering the enzymatic regulatory network under different noise levels. Here we present additional results about the recovery of this network at a given noise level. Figure S4 reports the results based on 500 data replications, where the noise level is set at $\sigma_j = 0.01, j = 1, 2, 3$. Figure S4(a) shows that, in more than 80% of the cases, KODE is able to recover the network without making any false discovery, whereas Figure S4(b) shows that, in more than 99% of the cases, KODE recovers all the true edges. Figure S4(c) reports a sparse recovery of the network, where an arrowed edge is drawn if it appears in more than 90% of the estimated networks out of 500 data replications. It is seen that KODE successfully recovers the true regulatory network.

S2.4 Gene regulatory network recovery by KODE

Table 1 in Section 6 of the paper reported the selection performance of KODE in terms of the area under the ROC curve for all 10 combinations of gene regulatory network structures

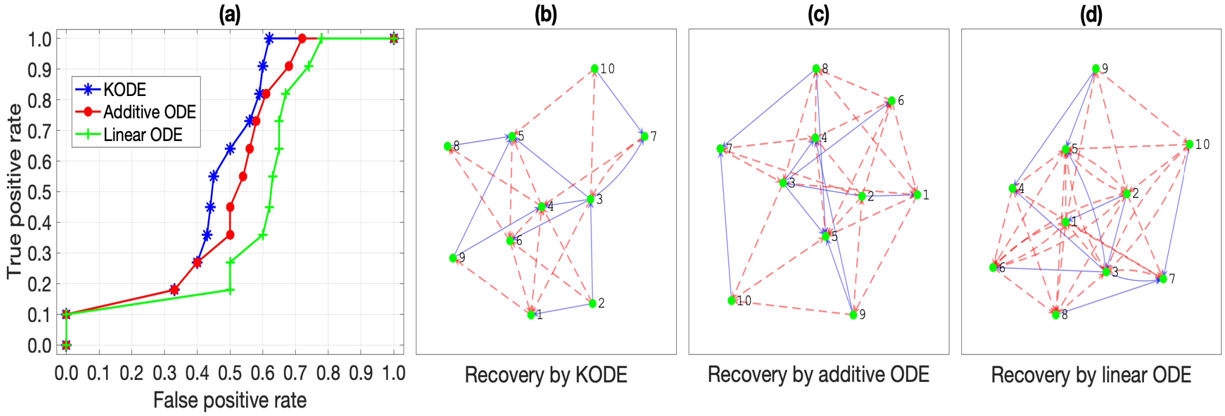

Figure S5: (a) Median ROC for recovering the 10-node gene regulatory network of *E.coli1*, based on 100 data replications. (b)-(d) The network recovery by KODE, additive ODE, and linear ODE. The solid and dashed arrowed lines denote the true and false discoveries, respectively.

from GeneNetWeaver. Here we present some additional results about the recovery of one such structure, the 10-node *E.coli1* network. Figure S5 reports the results based on 100 data replications. Figure S5(a) shows the median ROCs for KODE, additive ODE, and linear ODE. It is seen that KODE achieves the fastest recovery rate, as well as the largest AUC of 0.582, compared to 0.541 for additive ODE, and 0.460 for linear ODE. Figure S5(b)-(d) report the sparse recovery of the network, based on KODE, additive ODE, and linear ODE, respectively, under the 90% level of true positive rate. It is seen that KODE achieves a better selection accuracy compared to linear ODE and additive ODE.

References

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404.

Bach, F. (2017). On the equivalence between kernel quadrature rules and random feature expansions. *Journal of Machine Learning Research*, 18:714–751.

Chen, S., Shojaie, A., and Witten, D. M. (2017). Network reconstruction from high-dimensional ordinary differential equations. *Journal of the American Statistical Association*, 112:1697–1707.

Cox, D. D. (1983). Asymptotics for m-type smoothing splines. *Annals of Statistics*, 11:530–551.

Cucker, F. and Smale, S. (2002). On the mathematical foundations of learning. *American Mathematical Society Bulletin*, 39:1–49.

Duan, K., Keerthi, S. S., and Poo, A. N. (2003). Evaluation of simple performance measures for tuning svm hyperparameters. *Neurocomputing*, 51:41–59.

Gneiting, T., Kleiber, W., and Schlather, M. (2010). Matérn cross-covariance functions for multivariate random fields. *Journal of the American Statistical Association*, 105(491):1167–1177.

González, J., Vujačić, I., and Wit, E. (2014). Reproducing kernel hilbert space based estimation of systems of ordinary differential equations. *Pattern Recognition Letters*, 45:26–32.

Hall, P. and Marron, J. (1988). Choice of kernel order in density estimation. *Annals of Statistics*, pages 161–173.

Koltchinskii, V. and Yuan, M. (2010). Sparsity in multiple kernel learning. *Annals of Statistics*, 38:3660–3695.

Lin, Y. and Brown, L. D. (2004). Statistical properties of the method of regularization with periodic gaussian reproducing kernel. *Annals of Statistics*, 32(4):1723–1743.

Loh, P.-L. and Wainwright, M. J. (2012). High-dimensional regression with noisy and missing data: Provable guarantees with nonconvexity. *Annals of Statistics*, 40:1637–1664.

Meinshausen, N., Bühlmann, P., et al. (2006). High-dimensional graphs and variable selection with the lasso. *Annals of Statistics*, 34(3):1436–1462.

Meyer, D., Leisch, F., and Hornik, K. (2003). The support vector machine under test. *Neurocomputing*, 55(1-2):169–186.

Mikkelsen, F. V. and Hansen, N. R. (2017). Learning large scale ordinary differential equation systems. *arXiv preprint arXiv:1710.09308*.

Ravikumar, P., Wainwright, M. J., and Lafferty, J. (2010). High-dimensional ising model selection using l_1-regularized logistic regression. *Annals of Statistics*, 38:1287–1319.

Silverman, B. W. (1985). Some aspects of the spline smoothing approach to non-parametric regression curve fitting. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 47:1–21.

Talagrand, M. (1996). New concentration inequalities in product spaces. *Inventiones Mathematicae*, 126:505–563.

Tsybakov, A. B. (2009). *Introduction to Nonparametric Estimation.* Springer Science & Business Media.

van de Geer, S. (2000). *Empirical Processes in M-Estimation.* Cambridge University Press.

van der Vaart, A. W. and Wellner, J. A. (1996). *Weak Convergence and Empirical Processes.* Springer-Verlag, New York.

Wahba, G. (1983). Bayesian "confidence intervals" for the cross-validated smoothing spline. *Journal of the Royal Statistical Society. Series B (Statistical Methodology)*, 45:133–150.

Wahba, G. (1990). *Spline Models for Observational Data.* SIAM, Philadelphia.

Wahba, G., Wang, Y., Gu, C., Klein, R., and Klein, B. (1995). Smoothing spline ANOVA for exponential families, with application to the Wisconsin Epidemiological Study of Diabetic Retinopathy. *Annals of Statistics*, 23:1865–1895.

Yuan, M. and Zhou, D.-X. (2016). Minimax optimal rates of estimation in high dimensional additive models. *Annals of Statistics*, 44(6):2564–2593.

Zhang, T., Wu, J., Li, F., Caffo, B., and Boatman-Reich, D. (2015a). A dynamic directional model for effective brain connectivity using electrocorticographic (ECoG) time series. *Journal of the American Statistical Association*, 110:93–106.

Zhang, X., Cao, J., and Carroll, R. J. (2015b). On the selection of ordinary differential equation models with application to predator-prey dynamical models. *Biometrics*, 71(1):131–138.

Zhao, P. and Yu, B. (2006). On model selection consistency of lasso. *Journal of Machine Learning Research*, 7:2541–2563.