AN RKHS APPROACH TO ESTIMATE INDIVIDUALIZED TREATMENT RULES BASED ON FUNCTIONAL PREDICTORS

JUN FAN AND FUSHENG LV

Department of Mathematics
Hong Kong Baptist University, Kowloon, Hong Kong, China

LEI SHI*

School of Mathematical Sciences
Shanghai Key Laboratory for Contemporary Applied Mathematics
Fudan University, Shanghai, 200433, China

(Communicated by Yunlong Feng)

Abstract. In recent years there has been massive interest in precision medicine, which aims to tailor treatment plans to the individual characteristics of each patient. This paper studies the estimation of individualized treatment rules (ITR) based on functional predictors such as images or spectra. We consider a reproducing kernel Hilbert space (RKHS) approach to learn the optimal ITR which maximizes the expected clinical outcome. The algorithm can be conveniently implemented although it involves infinite-dimensional functional data. We provide convergence rate for prediction under mild conditions, which is jointly determined by both the covariance kernel and the reproducing kernel.

1. Introduction. In recent years precision medicine has attracted growing attention in both statistics and machine learning communities. One goal of precision medicine is to identify the optimal individualized treatment rules (ITR) by incorporating patients’ individual-level clinical information. Different approaches on how to estimate the optimal ITR via maximizing the expected clinical outcome have been considered, mainly including Q-learning, A-learning, D-learning and outcome weighted learning, see details in [9, 10, 11, 14, 17] and references therein.

Denote the treatment assignment $A \in \mathcal{A} := \{-1, 1\}$ i.e., we only consider two treatment options labeled $-1$ and $1$, and denote the clinical outcome by $Y \in \mathcal{Y} \subset \mathbb{R}_+ := [0, +\infty)$. Let $L^2(\mathcal{T})$ be the Hilbert space of square-integrable functions from $\mathcal{T}$ to $\mathbb{R}$, where $\mathcal{T}$ is a compact subset of a Euclidean space. We are interested in maximizing the expected clinical outcome given a set of functional predictors $X \in \mathcal{X} \subset L^2(\mathcal{T})$ such as brain images [12], which has been introduced as the observation of one patient’s symptom and discussed in [2, 8] together with its importance for applications. In clinical trials, patients are randomly assigned to one treatment of the two options. Therefore, we assume that the observation $X$ is independent of $A$ and $\text{Prob}(A = 1) = \text{Prob}(A = -1) = \frac{1}{2}$. The goal is to learn an ITR $f : \mathcal{X} \rightarrow \mathcal{A}$ based on a sample set $\{(x_i, a_i, y_i)\}_{i=1}^n$ such that the expected clinical

2010 Mathematics Subject Classification. Primary: 68T05; Secondary: 62J02.

Key words and phrases. Learning theory, individualized treatment rules, functional data, reproducing kernel Hilbert space.

* Corresponding author: Lei Shi.
outcome $\mathbb{E}[Y|A = f(X)]$ is maximized, where the sample data $\{(x_i, a_i, y_i)\}_{i=1}^n$ are observations from independent and identically distributed (i.i.d.) random copies of $(X,A,Y)$, which all have the same distribution $\rho$ on $\mathcal{X}\times\mathcal{A}\times\mathcal{Y}$. Note that maximizing the expected clinical outcome $\mathbb{E}[Y|A = f(X)]$ is equivalent to minimizing

$$\mathcal{R}(f) = \mathbb{E}[Y|A \neq f(X)],$$

where $I(\cdot)$ is the indicator function. Then the optimal ITR can be characterized by

$$f^*(X) = \arg\min_{f: \mathcal{X} \to \mathcal{A}} \mathcal{R}(f) = \text{sign}(\eta_0(X)) := \begin{cases} 1, & \text{if } \eta_0(X) \geq 0, \\ -1, & \text{otherwise}, \end{cases}$$

where $\eta_0(X) = \frac{1}{2}[\mathbb{E}(Y|X, A = 1) - \mathbb{E}(Y|X, A = -1)],$ which is the relative difference in outcome between two treatment assignments.

In this paper, we propose to adopt the D-learning framework [11] to directly learn the optimal ITR $f^*$ based on functional predictors. Denote by $\|\cdot\|_{L^2(\mathcal{T})}$ the norm of $L^2(\mathcal{T})$ induced by the inner product $\langle f, g \rangle_{L^2(\mathcal{T})} := \int_\mathcal{T} f(t)g(t)dt$. For brevity we simply denote the norm and the inner product by $\|\cdot\|$ and $\langle \cdot, \cdot \rangle$ respectively. For $\beta \in L^2(\mathcal{T})$, the induced linear functional is given by

$$L_{\beta}(x) = \langle \beta, x \rangle = \int_\mathcal{T} \beta(t)x(t)dt, \quad \forall x \in L^2(\mathcal{T}).$$

As in [11], we assume that the clinical outcome $Y$ can be characterized as

$$Y = \mu_0(X) + L_{\beta_0}(X)A + \epsilon,$$  \hspace{1cm} (1)

where $\mu_0, \beta_0 \in L^2(\mathcal{T})$ are unknown slope functions and $\epsilon$ is the random noise independent of $X$ and $A$. Without loss of generality, we will assume $\mathbb{E}(X) = 0$ throughout the paper. By simple calculation, we have $L_{\mu_0}(X) = \frac{1}{2}[\mathbb{E}(Y|X, A = 1) + \mathbb{E}(Y|X, A = -1)]$ and $L_{\beta_0}(X) = \frac{1}{2}[\mathbb{E}(Y|X, A = 1) - \mathbb{E}(Y|X, A = -1)],$ which follows

$$\eta_0(X) = L_{\beta_0}(X) = \mathbb{E}[AY|X].$$

For $\beta \in L^2(\mathcal{T})$, define the expected risk functional

$$\mathcal{E}(\beta) = \mathbb{E}[\|L_{\beta}(X) - AY\|^2],$$

and it is easy to check that $\beta_0 = \arg\min_{\beta: \mathcal{T} \to \mathbb{R}} \mathcal{E}(\beta)$. The goal of prediction is to recover the linear functional $L_{\beta_0}$ based on a training sample $\mathbf{z} = \{(x_i, a_i, y_i)\}_{i=1}^n$ consisting of $n$ independent copies of $(X, A, Y)$.

Let $K: \mathcal{T} \times \mathcal{T} \to \mathbb{R}$ be a continuous, symmetric and positive semi-definite kernel function. We say $K$ is positive semi-definite, meaning that for all $s \in \mathbb{N}$, the $s \times s$ kernel matrix $[K(t_i, t_j)]_{i,j=1}^s$ evaluated on any finite set of distinct points $\{t_1, \cdots, t_s\} \subset \mathcal{T}$ is positive semi-definite. The kernel function $K$ uniquely induces a reproducing kernel Hilbert space (RKHS) $H_K$, which is defined to be the closure of the linear span of the set of functions $\{K_t := K(t, \cdot) : t \in \mathcal{T}\}$ with the inner product $\langle \cdot, \cdot \rangle_K$ satisfying $\langle K_s, K_t \rangle_K = K(s,t), \forall s, t \in \mathcal{T}$ and the induced norm $\|\cdot\|_K$. In this paper we consider the regularized estimator of $\eta_0 = L_{\beta_0}$ given by $\hat{\eta} = L_{\hat{\beta}}$ with

$$\hat{\beta} = \arg\min_{\beta \in H_K} \left\{ \frac{1}{n} \sum_{i=1}^n (L_{\beta}(x_i) - a_iy_i)^2 + \lambda \|\beta\|_K^2 \right\},$$  \hspace{1cm} (2)

where $\lambda > 0$ is a regularization parameter controlling overfitting and can be selected via generalized cross validation. By virtue of the Representer Theorem, the above optimization problem is conveniently implementable [1, 15]. More specifically, we
have \( \hat{\beta}(t) = \sum_{i=1}^{n} \hat{c}_i \int_T x_i(s)K(s,t)ds \), where \( \hat{c} = (\hat{c}_1, ..., \hat{c}_n)^T \in \mathbb{R}^n \) is the solution of
\[
\min_{c \in \mathbb{R}^n} \left\{ \frac{1}{n} \|y - Wc\|_2^2 + \lambda c^TWc \right\},
\]
with \( y = (a_1y_1, ..., a_ny_n) \) and \( W = (W_{ij}) \) being a \( n \times n \) matrix defined by
\[
W_{ij} = \int_T \int_T x_i(s)K(s,t)x_j(t)dsdt.
\]
Note that the prediction accuracy can be measured by the excess risk
\[
\mathcal{E}(\beta) - \mathcal{E}(\beta_0) = \mathbb{E}[L_\beta(X) - AY]^2 - \mathbb{E}[L_{\beta_0}(X) - AY]^2
\]
\[
= \mathbb{E}[L_\beta(X) - L_{\beta_0}(X)]^2
\]
\[
= \int_T \int_T [\beta(s) - \beta_0(s)]C(s,t)[\beta(t) - \beta_0(t)]dsdt,
\]
where \( C : \mathcal{T} \times \mathcal{T} \rightarrow \mathbb{R} \) is the covariance kernel defined as
\[
C(s,t) = \mathbb{E}[X(s)X(t)], \quad \forall s,t \in \mathcal{T}.
\]
Define an integral operator
\[
L_C : L^2(\mathcal{T}) \rightarrow L^2(\mathcal{T})
\]
as
\[
L_C(f)(s) = \int_T C(s,t)f(t)dt, \quad \forall s \in \mathcal{T}, f \in L^2(\mathcal{T}).
\]
Throughout the paper we assume \( \mathbb{E}[\|X\|^2] < \infty \), which ensures that \( L_C \) is a compact operator on \( L^2(\mathcal{T}) \). We can similarly define an integral operator associated with the kernel function \( K \), which is given by
\[
L_K(f)(s) = \int_T K(s,t)f(t)dt, \quad \forall s \in \mathcal{T}, f \in L^2(\mathcal{T}).
\]
Note that both \( L_C \) and \( L_K \) are self-adjoint, positive and compact operators on \( L^2(\mathcal{T}) \). Thus the \( r \)-th power of \( L_C \) and \( L_K \) can be defined by spectral theorem for \( r > 0 \). For notational simplicity, we denote
\[
T_0 = L_C^{1/2}L_K^{1/2} \quad \text{and} \quad T_1 = L_C^{1/2}L_K^{1/2}.
\]
From (3) and the definition (4) of \( L_C \), the excess risk can be rewritten as
\[
\mathcal{E}(\hat{\beta}) - \mathcal{E}(\beta_0) = \langle L_C(\hat{\beta} - \beta_0), \hat{\beta} - \beta_0 \rangle = \|L_C^{1/2}(\hat{\beta} - \beta_0)\|^2.
\]
Thus we can evaluate the performance of algorithm (2) via estimating the norm \( \|L_C^{1/2}(\hat{\beta} - \beta_0)\| \), which will be discussed in next section. We first state the following generalized margin condition for our theoretical analysis. Recall that in our setting, the distribution \( \rho \) on \((\mathcal{X}, \mathcal{A}, \mathcal{Y})\) can be factorized as \( \rho(x,a,y) = \rho(y|x,a)\rho_A(a)\rho_X(x) \), where \( \rho_A \) and \( \rho_X \) are the marginal distributions of \( \rho \) on \( \mathcal{A} \) and \( \mathcal{X} \) respectively.

**Assumption 1.** There exist constants \( C_0 > 0 \) and \( \gamma \geq 0 \) such that
\[
\rho_X \{ 0 < |\rho_0(x)| \leq C_0t \} \leq t^\gamma,
\]
where \( \rho_X \) is the marginal distribution of \( X \).

**Theorem 1.1.** For any distribution \( \rho \) of \((X, A, Y)\) satisfying Assumption 1 and \( \eta = L_\beta \) for \( \beta \in L^2(\mathcal{T}) \), we have
\[
\mathcal{R}(\text{sign}(\eta)) - \mathcal{R}(f^*) \leq 4C_0^{-\frac{\gamma}{1+\gamma}}(\mathcal{E}(\beta) - \mathcal{E}(\beta_0))^{(1+\gamma)/(2+\gamma)}.
\]
Proof. Denote $X_\eta = \{x \in \mathcal{X} : \text{sign}(\eta)(x) \neq f^*(x)\}$. Note that $\rho_A$ is a binary distribution satisfying $\rho_A(+1) = \rho_A(-1) = 1/2$. It is easy to check that
\[
\mathcal{R}(\text{sign}(\eta)) - \mathcal{R}(f^*) = \int_{X_\eta} |2\eta_0(x)|d\rho_X(x).
\]

We separate the set $X_\eta$ into two parts one with $|\eta_0(x)| \leq C_0t$ and the other with $|\eta_0(x)| > C_0t$ where $|\eta_0(x)| \leq |\eta_0(x)|^2/(C_0t) \leq |x - \eta_0(x)|^2/(C_0t)$. We find that
\[
\int_{X_\eta} |2\eta_0(x)|d\rho_X(x) \\
\leq \int_{\{x \in X_\eta : |\eta_0(x)| \leq C_0t\}} 2C_0td\rho_X(x) \\
+ \int_{\{x \in X_\eta : |\eta_0(x)| > C_0t\}} 2|x - \eta_0(x)|^2/(C_0t)d\rho_X(x) \\
\leq 2C_0t\rho\{x \in \mathcal{X} : |\eta_0(x)| \leq C_0t\} + 2(\mathcal{E}(\beta) - \mathcal{E}(\beta_0))/(C_0t) \\
\leq 2C_0t + 2(\mathcal{E}(\beta) - \mathcal{E}(\beta_0))/(C_0t) = 4C_0(\mathcal{E}(\beta) - \mathcal{E}(\beta_0)/C_0^2)^{\frac{1}{2}},
\]
where the last equality follows by taking $t = (\mathcal{E}(\beta) - \mathcal{E}(\beta_0))/C_0^2$.

\[
\square
\]

Remark 1. When $\gamma = \infty$, we have $|\eta_0(x)| \geq C_0$ and hence $|\eta_0(x)| \leq |\eta_0(x)|^2/C_0$ for almost every $x \in \mathcal{X}$. So $\mathcal{R}(\text{sign}(\eta)) - \mathcal{R}(f^*) = \int_{X_\eta} |2\eta_0(x)|d\rho_X(x) \leq \int_{X_\eta} 2|x - \eta_0(x)|^2/(C_0t)d\rho_X(x) = 2(\mathcal{E}(\beta) - \mathcal{E}(\beta_0))/C_0$.

Theorem 1.1 indicates that we can learn the optimal ITR via minimizing the excess risk. Before establishing error bounds for the excess risk, we state some assumptions as follows, which are common conditions in statistical learning [3, 6] and functional data analysis [1, 15].

Assumption 2. Assume that $\|L^2_{\mathcal{X}}\mu_0\| \leq c_0$, $E\epsilon = 0$, $E\epsilon^4 \leq c_1$ and there exists a constant $c_2 > 0$ such that $E(X, f)^4 \leq c_2 \left(E(X, f)^2\right)^2$ for any $f \in L^2(T)$.

Assumption 3. Assume $L^2_{\mathcal{X}}\beta_0 = T_0^\theta(\gamma_0)$ for some $0 < \theta \leq \frac{1}{2}$ and $\gamma_0 \in L^2(T)$.

Assumption 4. Assume $\mathcal{N}(\lambda) = O(\lambda^{-\frac{\alpha}{2}})$ for some $\alpha > 0$, where
\[
\mathcal{N}(\lambda) := \text{Tr}((\lambda I + T_0)^{-1}T_0), \quad \lambda > 0,
\]
which is known as the effective number in learning theory literature [5, 7, 16].

Theorem 1.2. Suppose Assumptions 2, 3 and 4 hold, then by taking $\lambda = n^{-\frac{\alpha (1 + \gamma + 1)}{4\alpha (1 + 1)}}$, we have
\[
\mathcal{E}(\hat{\beta}) - \mathcal{E}(\beta_0) = O\left(n^{-\frac{12\beta_0}{4\alpha (1 + 1)}}\right). \tag{9}
\]

If Assumption 1 also holds, then we have
\[
\mathcal{R}(\text{sign}(\hat{\eta})) - \mathcal{R}(f^*) = O\left(n^{-\frac{12\beta_0 (1 + 1)}{4\alpha (1 + 1) (\gamma + 1)}}\right). \tag{10}
\]

Remark 2. Particularly, if $\beta_0 \in H_K$, then by taking $\lambda = n^{-\frac{2\beta_0}{2\alpha + 1}}$, we have $\mathcal{E}(\hat{\beta}) - \mathcal{E}(\beta_0) = O\left(n^{-\frac{2\beta_0}{2\alpha + 1}}\right)$, which achieves the minimax optimal learning rate [1, 4, 15] and follows $\mathcal{R}(\text{sign}(\hat{\eta})) - \mathcal{R}(f^*) = O\left(n^{-\frac{2\beta_0 (1 + 1)}{4\alpha (1 + 1) (\gamma + 1)}}\right)$. 

2. Preliminaries. Recall that $L^{1/2}_K$ is an isomorphism from $H_K$, the closure of $H_K$ in $L^2(T)$, to $H_K$, i.e., for each $\beta \in H_K$, there exists some $f \in H_K \subseteq L^2(T)$ such that $\beta = L^{1/2}_K f$ and $\|\beta\|_K = \|f\|$. Therefore, we have $\hat{\beta} = L^{1/2}_K \hat{f}$, where

$$
\hat{f} = \arg\min_{f \in L^2(T)} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left( \langle L^{1/2}_K f, x_i \rangle - a_i y_i \right)^2 + \lambda \|f\|^2 \right\}.
$$

We first give an explicit expression for $\hat{f}$. Given an input data set $x = \{x_i\}_{i=1}^{n}$, define two operators $S_x : \mathbb{R}^n \to L^2(T)$ and $T_x : L^2(T) \to L^2(T)$ by

$$
S_x c = \sum_{i=1}^{n} c_i L^{1/2}_K x_i, \quad c \in \mathbb{R}^n,
$$

and

$$
T_x = \frac{1}{n} \sum_{i=1}^{n} L^{1/2}_K x_i \otimes L^{1/2}_K x_i.
$$

Here, for any $f, g \in L^2(T)$, $f \otimes g : L^2(T) \to L^2(T)$ denotes the rank-one operator defined by $f \otimes g(h) = \langle f, h \rangle g$ for any $h \in L^2(T)$. One can check that $T_x = \frac{1}{n} S_x^* S_x$ where $S_x^* : L^2(T) \to \mathbb{R}^n$ is the adjoint operator of $S_x$. Then the estimator $\hat{f}$ can be expressed as

$$
\hat{f} = (\lambda I + T_x)^{-1} \frac{1}{n} S_x y
$$

(11)

where $y = (a_1 y_1, \cdots, a_n y_n)$ is the vector consisting of the output data. To provide some intuition behind $\hat{f}$, for any $f \in L^2(T)$, consider the random variable $\xi(x) := \langle L^{1/2}_K x, f \rangle L^{1/2}_K x$ with values in $L^2(T)$. We see that

$$
\frac{1}{n} \sum_{i=1}^{n} \xi(x_i) = \frac{1}{n} \sum_{i=1}^{n} \langle L^{1/2}_K x_i, f \rangle L^{1/2}_K x_i = T_x(f),
$$

(12)

$$
\mathbb{E}(\xi) = \mathbb{E}(\langle x, L^{1/2}_K f \rangle L^{1/2}_K x) = L^{1/2}_K L^{1/2}_C L^{1/2}_K (f) = T_0(f),
$$

which shows that $T_x$ is a finite sample approximation of $L^{1/2}_K L^{1/2}_C L^{1/2}_K$. Analogously, if we consider the random variable $\xi(z) := ay L^{1/2}_K x$ with value in $L^2(T)$, where $z = (x, a, y)$ is generated from the model (1). Then we have $\mathbb{E}(ay|x) = L_{\beta_0}(x) = \langle \beta_0, x \rangle$ and

$$
\frac{1}{n} \sum_{i=1}^{n} \xi(z_i) = \frac{1}{n} S_x y,
$$

(13)

$$
\mathbb{E}(\xi) = \mathbb{E}(\langle \beta_0, x \rangle L^{1/2}_K x) = L^{1/2}_K L^{1/2}_C \beta_0,
$$

meaning that $\frac{1}{n} S_x y$ is a finite sample approximation of $L^{1/2}_K L^{1/2}_C \beta_0$. Based on these observations, define an intermediate function

$$
f_\lambda = \left( \lambda I + L^{1/2}_K L^{1/2}_C L^{1/2}_K \right)^{-1} L^{1/2}_K L^{1/2}_C \beta_0,
$$

(14)

which is actually a population version of $\hat{f}$. By introducing $f_\lambda$, we can decompose $\|L^{1/2}_C (\hat{\beta} - \beta_0)\|$ into two terms, namely, sample error and approximation error. In the followings, the adjoint of an operator $T$ is denoted by $T^\ast$.  

Proposition 1. Let $\lambda > 0$ and $f_\lambda$ be defined by (14). Then a solution of (2) is given by $\hat{\beta} = L_K^{1/2} \hat{f}$ with $\hat{f}$ expressed as (11) and the following error decomposition holds
\[
\|L_C^{1/2}(\hat{\beta} - \beta_0)\| \leq S(z, \lambda) + A(\lambda),
\] (15)
where
\[
S(z, \lambda) := \|T_0^{1/2}(\hat{f} - f_\lambda)\|,
\]
\[
A(\lambda) := \|T_1(\lambda I + T_1)^{-1} L_C^{1/2} \beta_0 - L_C^{1/2} \beta_0\|.
\]

Proof. We can bound $\|L_C^{1/2}(\hat{\beta} - \beta_0)\|$ as
\[
\|L_C^{1/2}(\hat{\beta} - \beta_0)\| \leq \|L_C^{1/2}L_K^{1/2}(\hat{f} - f_\lambda)\| + \|L_C^{1/2}(L_K^{1/2} f_\lambda - \beta_0)\|. \tag{16}
\]
Observe that $T_0 = (L_C^{1/2}L_K^{1/2})^* (L_C^{1/2}L_K^{1/2})$ and $T_1 = (L_C^{1/2}L_K^{1/2})(L_C^{1/2}L_K^{1/2})^*$. Then based on the polar decomposition of compact operators, there holds
\[
L_C^{1/2}L_K^{1/2} = UT_0^{1/2} = T_1^{1/2}U^*
\]
where $U$ is an isometry from the range of $T_0^{1/2}$ to the range of $T_1^{1/2}$. It follows that
\[
L_K^{1/2}L_C^{1/2} = (L_C^{1/2}L_K^{1/2})^* = (T_1^{1/2}U^*)^* = UT_1^{1/2},
\]
\[
T_0 = (UT_1^{1/2})(T_1^{1/2}U^*) = UT_1U^*,
\]
and
\[
\|L_C^{1/2}L_K^{1/2}(\hat{f} - f_\lambda)\| = \|UT_0^{1/2}(\hat{f} - f_\lambda)\| \leq \|T_0^{1/2}(\hat{f} - f_\lambda)\|. \tag{17}
\]
Recall that
\[
f_\lambda = (\lambda I + L_K^{1/2}L_CL_K^{1/2})^{-1} L_K^{1/2}L_C\beta_0 = (\lambda I + T_0)^{-1} L_K^{1/2}L_C\beta_0.
\]
Then we obtain
\[
L_C^{1/2}L_K^{1/2} f_\lambda = T_1^{1/2}U^*(\lambda I + UT_1U^*)^{-1}UT_1^{1/2}L_C^{1/2} \beta
\]
\[
= T_1^{1/2}(\lambda I + T_1)^{-1} T_1^{1/2}L_C^{1/2} \beta = T_1(\lambda I + T_1)^{-1} L_C^{1/2} \beta_0. \tag{18}
\]
Substituting (17) and (18) into (16), we get the desired bounds. This completes the proof of Proposition 1.

The first term $S(z, \lambda)$ in (15) is called the sample error, which is caused by randomly sampling. As we discussed, the intermediate function $f_\lambda$ is a population version of the estimator $\hat{f}$, one would expect $S(z, \lambda)$ tends to zero as the sample size $n$ becomes large. An elegant error analysis of $S(z, \lambda)$ will be carried out in detail in next section by characterizing the approximations of $(\lambda I + T_k)^{-1}$ to $(\lambda I + T_0)^{-1}$ and $\frac{1}{n} S_n \mathbf{y}$ to $L_C^{1/2}L_C\beta_0$ in some suitable sense. The second term $A(\lambda)$ in (15) is called the approximation error, which depends on the regularities of slope function $\beta_0$, the kernel $K$ and the covariance kernel $C$. Though $A(\lambda)$ is independent of randomized sampling, the decay of $A(\lambda)$ is crucial to bound the first term $S(z, \lambda)$. According to the error decomposition in Proposition 1, to bound the prediction error of algorithm (2), it suffices to bound $S(z, \lambda)$ and $A(\lambda)$ respectively.
3. Error analysis. In this section, we will estimate $S(z, \lambda)$ and $A(\lambda)$ respectively, and then derive an upper bound for $\|L_{C}^{1/2}(\hat{\beta} - \beta_{0})\|$. Recall the definition (6) of $T_{0}$ and $T_{1}$. Both $T_{0}$ and $T_{1}$ are positive, self-adjoint and compact. It should be pointed out that these two operators shares the same eigenvalues, which we denote by $(\lambda_{j})_{j \geq 1}$ and is sorted in descending order. Then the spectral theorem of compact operators claims that

$$
T_{0} = \sum_{j \geq 1} \lambda_{j} \phi_{j} \otimes \phi_{j},
$$

$$
T_{1} = \sum_{j \geq 1} \lambda_{j} \psi_{j} \otimes \psi_{j},
$$

(19)

where $\{\phi\}_{j \geq 1}$ and $\{\psi\}_{j \geq 1}$ are two sets of orthonormal basis of $L^{2}(T)$. The spectral decomposition of $T_{0}$ and $T_{1}$ are very important in our error analysis. We first estimate the approximation error $A(\lambda)$ based on the spectral decomposition of $T_{1}$.

**Proposition 2.** If $L_{C}^{1/2} \beta_{0} = T_{0}^{\theta}(\gamma_{0})$ for some $0 < \theta \leq \frac{1}{2}$ and $\gamma_{0} \in L^{2}(T)$, then

$$
A(\lambda) \leq \|\gamma_{0}\| |\lambda|^{\theta}, \quad \lambda > 0.
$$

(20)

**Proof.** According to the spectral decomposition of $T_{1}$ in (19), there exists in $L^{2}(T)$ an orthonormal basis $\{\psi\}_{j \geq 1}$ consisting the eigenfunctions of $T_{1}$ associated with eigenvalues $(\lambda_{j})_{j \geq 1}$. Since $L_{C}^{1/2} \beta_{0} = T_{0}^{\theta}(\gamma_{0})$, we know by writing $\gamma_{0} = \sum_{j \geq 1} d_{j} \psi_{j}$ with $\|\gamma_{0}\|^{2} = \sum_{j \geq 1} d_{j}^{2}$ that

$$
L_{C}^{1/2} \beta_{0} = \sum_{j \geq 1} \lambda_{j}^{\theta} d_{j} \psi_{j}
$$

and

$$
T_{1}(\lambda I + T_{1})^{-1} L_{C}^{1/2} \beta_{0} - L_{C}^{1/2} \beta_{0} = \sum_{j \geq 1} \frac{\lambda_{j}^{1+\theta}}{\lambda + \lambda_{j}} d_{j} \psi_{j} - \sum_{j \geq 1} \lambda_{j}^{\theta} d_{j} \psi_{j} = \sum_{j \geq 1} \frac{\lambda_{j}^{\theta}}{\lambda + \lambda_{j}} d_{j} \psi_{j}.
$$

Thus we have

$$
A^{2}(\lambda) = \left\| T_{1}(\lambda I + T_{1})^{-1} L_{C}^{1/2} \beta_{0} - L_{C}^{1/2} \beta_{0} \right\|^{2}
$$

$$
= \sum_{j \geq 1} \frac{\lambda^{2} \lambda_{j}^{2\theta}}{(\lambda + \lambda_{j})^{2}} d_{j}^{2} \leq \lambda^{2} \left( \sup_{t \geq 0} \frac{t^{\theta}}{\lambda + t} \right)^{2} \sum_{j \geq 1} d_{j}^{2}.
$$

Simple calculations show that for $0 < \theta \leq \frac{1}{2}$,

$$
\sup_{t \geq 0} \frac{t^{\theta}}{\lambda + t} = \theta^{\theta}(1 - \theta)^{1-\theta} \lambda^{\theta-1} \leq \lambda^{\theta-1}.
$$

This inequality together with $\|\gamma_{0}\|^{2} = \sum_{j \geq 1} d_{j}^{2}$ leads to the bound (20) for $A(\lambda)$. Thus we have completed the proof of Proposition 2. 

Next we focus on estimating the sample error $S(z, \lambda)$. Our mathematical analysis involves approximation of operators. So we introduce some notations in operator theory. Let $A : L^{2}(T) \to L^{2}(T)$ be a (linear) bounded operator and $(e_{j})_{j \geq 1}$ an orthonormal basis of $L^{2}(T)$. The operator norm is defined as $\|A\| = \sup_{\|f\| = 1} \|Af\|$. A bounded operator $A$ is called Hilbert-Schmidt if $\sum_{j \geq 1} \|Ae_{j}\|^{2} < \infty$, where $(e_{j})_{j \geq 1}$ is an orthonormal basis of $L^{2}(T)$. The space of Hilbert-Schmidt operators is a
Hilbert space with respect to the inner product $\langle A, B \rangle_{HS} = \sum_{j \geq 1} \langle A e_j, B e_j \rangle$ and we denote the corresponding norm by $\| \cdot \|_{HS}$. If $A$ is Hilbert-Schmidt, then

$$\| A \| \leq \| A \|_{HS}. \quad (21)$$

And for any bounded operator $B$, there holds

$$\| AB \|_{HS} \leq \| A \|_{HS} \| B \|, \quad \| BA \|_{HS} \leq \| B \| \| A \|_{HS}.$$

We denote by $\text{Tr}(A)$ the trace of $A$, which is defined as $\text{Tr}(A) = \sum_{j \geq 1} \langle (A^* A)^{1/2} e_j, e_j \rangle$. In particular, if $A$ is positive, self-adjoint and compact, we can choose the orthonormal basis $(e_j)_{j \geq 1}$ to be the eigenfunctions of $A$ with corresponding eigenvalues $\{\gamma_j\}_{j \geq 1}$. Then $\text{Tr}(A) = \sum_{j \geq 1} (\langle A e_j, e_j \rangle = \sum_{j \geq 1} \gamma_j$. Now we are in the position to bound the sample error $S(z, \lambda)$.

**Proposition 3.** Assume that $\| L_{E_0}^2 \| \leq c_0, \, E \epsilon = 0$ and $E \epsilon^4 \leq c_1$. Suppose that there exists a constant $c_2 > 0$ such that for any $f \in L^2(T)$,

$$\mathbb{E}(X, f)^4 \leq c_2 (\mathbb{E}(X, f)^2)^2. \quad (22)$$

Then for $0 < \delta < 1$, with confidence $1 - \delta$,

$$S(z, \lambda) \leq c_3 (1 + A(\lambda))^{1/2} \delta^{-3/2} \sqrt{N(\lambda)} \left( n^{-1/2} + \sqrt{\frac{N(\lambda)}{\lambda}} n^{-1} + \frac{N(\lambda)}{\lambda} n^{-3/2} \right), \quad (23)$$

where $c_3$ is a constant depending on $c_0, c_1, c_2$ and $\text{Tr}(T_0)$.

**Proof.** We split the proof in several steps.

**Step 1.** Error decomposition for $S(z, \lambda)$. The definition (11) of $\hat{f}$ gives that

$$\hat{f} - f_\lambda = (\lambda I + T_x)^{-1} \Delta(z, \lambda),$$

where

$$\Delta(z, \lambda) := \frac{1}{n} S_x y - T_x f_\lambda - \lambda f_\lambda.$$

Then we can decompose $\hat{f} - f_\lambda$ as

$$\hat{f} - f_\lambda = \left[ (\lambda I + T_x)^{-1} - (\lambda I + T_0)^{-1} \right] \Delta(z, \lambda) + (\lambda I + T_0)^{-1} \Delta(z, \lambda).$$

We apply the second order decomposition to $(\lambda I + T_x)^{-1} - (\lambda I + T_0)^{-1}$, which gives that

$$(\lambda I + T_x)^{-1} - (\lambda I + T_0)^{-1} = (\lambda I + T_0)^{-1} (T_0 - T_x) (\lambda I + T_0)^{-1} + (\lambda I + T_0)^{-1} (T_0 - T_x) (\lambda I + T_x)^{-1} (T_0 - T_x) (\lambda I + T_0)^{-1}.$$

From the two equations above and the bound $\| T_0^{1/2} (\lambda I + T_0)^{-1/2} \| \leq 1$, we obtain

$$S(z, \lambda) = \| T_0^{1/2} (\hat{f} - f_\lambda) \| \leq S_1(z, \lambda) + S_2(z, \lambda) + S_3(z, \lambda),$$

where

$$S_1(z, \lambda) := \| (\lambda I + T_0)^{-1/2} (T_0 - T_x) (\lambda I + T_0)^{-1} \Delta(z, \lambda) \|,$$

$$S_2(z, \lambda) := \| (\lambda I + T_0)^{-1/2} (T_0 - T_x) (\lambda I + T_x)^{-1} (T_0 - T_x) (\lambda I + T_0)^{-1} \Delta(z, \lambda) \|,$$

$$S_3(z, \lambda) := \| (\lambda I + T_0)^{-1/2} \Delta(z, \lambda) \|.$$
Step 2. Bound on $S_3(z, \lambda)$. Note that
\[
\frac{1}{n} S_x y - T_x f_\lambda = \frac{1}{n} \sum_{i=1}^{n} (a_i y_i - \langle L_K^{1/2} x_i, f_\lambda \rangle) L_K^{1/2} x_i
\]
and by the definition (14) of $f_\lambda$,
\[
\lambda f_\lambda = L_K^{1/2} L_C \beta_0 - T_0 f_\lambda.
\]
Define a random variable
\[
\xi_\lambda(z) := (\lambda y - \langle L_K^{1/2} x, f_\lambda \rangle) L_K^{1/2} x
\]
with values in $L^2(\mathcal{T})$, where $z = (x, a, y)$ is generated from model (1). Then from (12) and (13),
\[
E \xi_\lambda = L_K^{1/2} L_C \beta_0 - T_0 f_\lambda.
\]
It then follows that
\[
\Delta(z, \lambda) := \frac{1}{n} S_x y - T_x f_\lambda - \lambda f_\lambda = \frac{1}{n} \sum_{i=1}^{n} \xi_\lambda(z_i) - E(\xi_\lambda).
\]
We further define
\[
\tilde{\xi}_\lambda(z) = (\lambda I + T_0)^{-1/2} \xi_\lambda(z).
\]
Then one can see that
\[
(\lambda I + T_0)^{-1/2} \Delta(z, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \tilde{\xi}_\lambda(z_i) - E(\tilde{\xi}_\lambda).
\]
The Cheby-
shew's inequality gives that, for all $\epsilon > 0$,
\[
\text{Prob}_x \left\{ \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\xi}_\lambda(z_i) - E(\tilde{\xi}_\lambda) \right\| \geq \epsilon \right\} \leq \frac{E(\|\tilde{\xi}_\lambda\|^2) \epsilon^2}{n^2}.
\]
Note that $S_3(z, \lambda) = \left\| \frac{1}{n} \sum_{i=1}^{n} \tilde{\xi}_\lambda(z_i) - E(\tilde{\xi}_\lambda) \right\|$. For any $0 < \delta < 1$ and by taking $\epsilon = \sqrt{\frac{E(\|\tilde{\xi}_\lambda\|^2)}{n^2}}$ in the inequality above, it follows that with confidence $1 - \delta$,
\[
S_3(z, \lambda) \leq \sqrt{\frac{E(\|\tilde{\xi}_\lambda\|^2)}{n^2}}.
\]
So in order to bound $S_3(z, \lambda)$, we need to estimate $E(\|\tilde{\xi}_\lambda\|^2)$. Recall that the sample $z = (x, y)$ is generated from model (1). Then
\[
ay = L_{\beta_0}(x) + a(L_{\mu_0}(x) + \epsilon) = \langle \beta_0, x \rangle + a(\mu_0, x) + \epsilon.
\]
This equation together with the definition (24) of $\tilde{\xi}_\lambda(z)$ yields that
\[
\tilde{\xi}_\lambda(z) = \langle \beta_0 - L_K^{1/2} f_\lambda, x \rangle (\lambda I + T_0)^{-1/2} L_K^{1/2} x + a(\mu_0, x) + \epsilon (\lambda I + T_0)^{-1/2} L_K^{1/2} x.
\]
Thus we have
\[
\|\tilde{\xi}_\lambda\|^2 \leq 2(\langle \mu_0, x \rangle + \epsilon)^2 (\lambda I + T_0)^{-1/2} L_K^{1/2} x \parallel^2 + 2(\langle \beta_0 - L_K^{1/2} f_\lambda, x \rangle) (\lambda I + T_0)^{-1/2} L_K^{1/2} x \parallel^2.
\]
The Cauchy-Schwartz inequality gives that
\[
E(\parallel\tilde{\xi}_\lambda\|^2) \leq 8 \left\{ E(\langle \mu_0, x \rangle^4) + E(\epsilon^4) \right\}^{1/2} \left\{ E(\|\lambda I + T_0\|^{-1/2} L_K^{1/2} x \parallel^4) \right\}^{1/2} + 2 \left\{ E(\langle \beta_0 - L_K^{1/2} f_\lambda, x \rangle^4) \right\}^{1/2} \left\{ E(\|\lambda I + T_0\|^{-1/2} L_K^{1/2} x \parallel^4) \right\}^{1/2}.
\]
Now we use condition (22) to obtain $E(\mu_0, x)^4 \leq c_2\|L^{1/2}_C \mu_0\|^4$ and

$$E(\beta_0 - L^{1/2}_K f_\lambda, x)^4 \leq c_2 \left[ E(\beta_0 - L^{1/2}_K f_\lambda, x)^2 \right]^2$$

$$= c_2 \left[ \langle L_C (\beta_0 - L^{1/2}_K f_\lambda), \beta_0 - L^{1/2}_K f_\lambda \rangle \right]^2$$

$$= c_2 \| L^{1/2}_C \beta_0 - L^{1/2}_C L^{1/2}_K f_\lambda \|^4 = c_2 A^2(\lambda),$$

where the last equality follows the same argument in the proof of Proposition 1. Next we shall show that

$$E \| (\lambda + T_0)^{-1/2} L^{1/2}_K x \|^4 \leq 2c_2\mathcal{N}(\lambda),$$

where $\mathcal{N}(\lambda)$ are defined by (8). According to the spectral decomposition of $T_0$ in (19), there exists in $L^2(T)$ an orthonormal basis $\{\phi_j\}_{j \geq 1}$ consisting the eigenfunctions of $T_0$ associated with eigenvalues $\{\lambda_j\}_{j \geq 1}$. Hence

$$E \| (\lambda + T_0)^{-1/2} L^{1/2}_K x \|^4 = \sum_{j \geq 1} \langle (\lambda + T_0)^{-1/2} L^{1/2}_K x, \phi_j \rangle^2$$

$$\leq \sum_{j \geq 1} \frac{1}{(\lambda + \lambda_j)^2} E \left[ \sum_{j \neq k} \frac{c_2}{(\lambda + \lambda_j)(\lambda + \lambda_k)} \right]$$

$$\leq \sum_{j \geq 1} \frac{c_2}{(\lambda + \lambda_j)^2} \sum_{j \neq k} \frac{c_2}{(\lambda + \lambda_j)(\lambda + \lambda_k)}$$

$$\leq 2c_2\mathcal{N}(\lambda),$$

where we use condition (22) to obtain the first inequality. Substituting the estimates (27) and (28) to (26), we obtain

$$E \| \tilde{c}_{\lambda} \|^2 \leq (8\sqrt{2c_1c_2} + 8\sqrt{2c_2c_0} + 2\sqrt{2c_2})(1 + A(\lambda))\mathcal{N}(\lambda).$$

Hence inequality (25) gives that, with confidence $1 - \delta$,

$$S_3(\mathbf{z}, \lambda) \leq c_4 \delta^{-1/2}(1 + A(\lambda))^{1/2} \sqrt{\mathcal{N}(\lambda)n}^{-1/2}$$

where $c_4 := (8\sqrt{2c_1c_2} + 8\sqrt{2c_2c_0} + 2\sqrt{2c_2})^{1/2}$.

**Step 3.** Bound on $S_1(\mathbf{z}, \lambda)$. For convenience, we introduce the notation

$$D(\mathbf{z}, \lambda) := \| (\lambda + T_0)^{-1/2} (T_0 - T_\mathbf{z}) \| = \| (T_0 - T_\mathbf{z}) (\lambda + T_0)^{-1/2} \|.$$  

We use the bounds $\| (\lambda + T_0)^{-1} \| \leq \lambda^{-1}$ and $\| (\lambda + T_0)^{-1/2} \| \leq \lambda^{-1/2}$ to obtain that

$$S_1(\mathbf{z}, \lambda) \leq \lambda^{-1/2} D(\mathbf{z}, \lambda) S_3(\mathbf{z}, \lambda).$$

We next show that with confidence $1 - \delta/2$,

$$D(\mathbf{z}, \lambda) \leq [2c_2 \text{Tr}(T_0)]^{1/2} \delta^{-1/2} \sqrt{\mathcal{N}(\lambda)n}^{-1/2}.$$
To derive the above bound, we define a random variable  
\[ \eta_\lambda(x) := (\lambda I + T_0)^{-1/2} L_{\mathcal{K}}^{1/2} x \].  
It takes values in the space of Hilbert-Schmidt operators on \( L^2(\mathcal{T}) \). From (12), one can check that  
\[ (\lambda I + T_0)^{-1/2} T_x = \frac{1}{n} \sum_{i=1}^{n} \eta_\lambda(x_i) \]  
and  
\[ E \eta_\lambda = (\lambda I + T_0)^{-1/2} T_0. \]  
Then  
\[ D(z, \lambda) = \frac{1}{n} \sum_{i=1}^{n} \eta_\lambda(x_i) - E \eta_\lambda \].  
Following the same argument in step 2, we see that with confidence \( 1 - \delta \),  
\[ D(z, \lambda) \leq \sqrt{\frac{E \eta_\lambda^2}{n \delta}}. \]  
Next we use the spectral decomposition of \( T_0 \) in (19) to derive a bound on \( E \eta_\lambda^2 \). By (21), there holds  
\[ E \eta_\lambda^2 \leq E \eta_\lambda^2_{HS} = E \left[ \sum_{j \geq 1} \langle \eta_\lambda \phi_j, \eta_\lambda \phi_j \rangle \right] = \sum_{j \geq 1} E \langle L_{\mathcal{K}}^{1/2} x, \phi_j \rangle^2 \langle (\lambda I + T_0)^{-1/2} L_{\mathcal{K}}^{1/2} x \rangle^2 \leq c_2 E \| (\lambda I + T_0)^{-1/2} L_{\mathcal{K}}^{1/2} x \|^2 \sum_{j \geq 1} E \langle L_{\mathcal{K}}^{1/2} x, \phi_j \rangle^2 \]  
where the last inequality is from Cauchy-Schwartz inequality and condition (22). Note that  
\[ E \| (\lambda I + T_0)^{-1/2} L_{\mathcal{K}}^{1/2} x \|^2 = \sum_{j \geq 1} E \langle (\lambda I + T_0)^{-1/2} L_{\mathcal{K}}^{1/2} x, \phi_j \rangle^2 = \sum_{j \geq 1} \lambda_j \frac{1}{\lambda + \lambda_j} \langle L_{\mathcal{K}} L_{\mathcal{K}}^{1/2} \phi_j, L_{\mathcal{K}}^{1/2} \phi_j \rangle = \sum_{j \geq 1} \frac{\lambda_j}{\lambda + \lambda_j} = \mathcal{N}(\lambda) \]  
and  
\[ \sum_{j \geq 1} E \langle L_{\mathcal{K}}^{1/2} x, \phi_j \rangle^2 = \sum_{j \geq 1} \langle L_{\mathcal{K}} L_{\mathcal{K}}^{1/2} \phi_j, L_{\mathcal{K}}^{1/2} \phi_j \rangle = \text{Tr}(T_0). \]  
Combining the estimates above, we derive bound (32). The estimation in step 2 gives that, with confidence \( 1 - \delta/2 \),  
\[ S_3(z, \lambda) \leq \sqrt{2c_4 \delta^{-1/2}(1 + \mathcal{A}(\lambda))^{1/2}} \sqrt{\mathcal{N}(\lambda)n^{-1/2}}. \]  
Using this inequality and (32), we obtain with confidence \( 1 - \delta \),  
\[ S_1(z, \lambda) \leq c_5 \delta^{-1}(1 + \mathcal{A}(\lambda))^{1/2}\mathcal{N}(\lambda)\lambda^{-1/2}n^{-1} \]  
(33)  
where \( c_5 := (4c_2 \text{Tr}(T_0))^{1/2}c_4 \).  
**Step 4.** Derive total bound on \( S(z, \lambda) \). We first bound \( S_2(z, \lambda) \). Recall the definition (30) of \( D(z, \lambda) \). Then the bound  
\[ \| (\lambda I + T_x)^{-1} \| \leq \lambda^{-1} \]  
gives that  
\[ S_2(z, \lambda) \leq \lambda^{-1} [D(z, \lambda)]^2 S_3(z, \lambda). \]
From the estimates in step 3, we know that with confidence $1 - \delta/2$, there holds
\[
D(z, \lambda)S_3(z, \lambda) \leq \sqrt{2c_3}\delta^{-1}(1 + A(\lambda))^{1/2}N(\lambda)n^{-1}.
\]
The two inequalities above together with estimates (32) gives that with confidence $1 - \delta$,
\[
S_2(z, \lambda) \leq c_6\delta^{-3/2}(1 + A(\lambda))^{1/2}[N(\lambda)]^{3/2}\lambda^{-1}n^{-3/2}
\]
where $c_6 := c_5(4c_2\text{Tr}(T_0))^{1/2}$. Finally, combining bounds (33), (34) and (29), we obtain with the desired bound (23) with $c_3 = \sqrt{3}c_4 + 3c_5 + 3\sqrt{3}c_6$.

**Proof of Theorems 1.2**: Combining Propositions 1, 2 and 3, we have
\[
\|L_C^\frac{1}{2}(\hat{\beta} - \beta_0)\| \\
\leq \|\gamma_0\|\lambda^\theta + c_3(1 + A(\lambda))^{1/2}\delta^{-3/2}\sqrt{N(\lambda)}\left(n^{-1/2} + \sqrt{\frac{N(\lambda)}{\lambda}}n^{-1} + \frac{N(\lambda)}{\lambda}n^{-3/2}\right) \\
\leq \left(\|\gamma_0\| + \sqrt{2c_3}\delta^{-\frac{3}{2}}\right)\left[\lambda^\theta + \lambda^{-1/4}\alpha\left(n^{-1/2} + \lambda^{-\frac{2\alpha + 1}{4\alpha}}n^{-1} + \lambda^{-\frac{2\alpha + 1}{2\alpha}}n^{-3/2}\right)\right] \\
\leq \left(\|\gamma_0\| + \sqrt{2c_3}\delta^{-\frac{3}{2}}\right)n^{-\frac{4\alpha \theta}{(4\alpha + 4\alpha + 3)}},
\]
where the second inequality follows from $N(\lambda) = O(\lambda^{-\frac{1}{4}})$ and the third inequality follows by taking $\lambda = n^{-\frac{4\alpha}{(4\alpha + 4\alpha + 3)}}$. This together with Theorem 1.1 leads to the desired conclusion.

4. **Discussions.** In this paper, we consider the estimation of individualized treatment rules (ITR) based on functional predictors within the framework of D-learning. We consider an RKHS approach to learn the optimal ITR which maximizes the expected clinical outcome and derive fast convergence rates for prediction. There are several possible extensions to this work. First, it would be interesting to extend our approach to the cases involving multiple treatments instead of just two options. Second, to take into account patients’ various information, it would be worthwhile to consider scalar and functional predictors in the model simultaneously. Another interesting question is how to employ distributed learning scheme [5, 7, 13] to make it possible to deal with large-scale dataset efficiently.

**Acknowledgments.** The work of J. Fan described in this paper is partially supported by the startup and FRG grants from Hong Kong Baptist University, Hong Kong RGC Early Career Schemes 22303518, and the NSF grant of China (No. 11801478). The work of L. Shi described in this paper is supported by the NSF grants of China (No.11571078, 11631015) and the Program of Shanghai Subject Chief Scientist (No.18XD1400700).

**REFERENCES**

[1] T. T. Cai and M. Yuan, Minimax and adaptive prediction for functional linear regression, *Journal of the American Statistical Association*, 107 (2012), 1201–1216.
[2] A. Ciarleglio, E. Petkova, T. Tarpey and R. T. Ogden, Flexible functional regression methods for estimating individualized treatment regimes, *Stat*, 5 (2016), 185–199.
[3] J. Fan, T. Hu, Q. Wu and D. X. Zhou, Consistency analysis of an empirical minimum error entropy algorithm, *Applied and Computational Harmonic Analysis*, 41 (2016), 164–189.
[4] X. Guo, J. Fan and D. X. Zhou, Sparsity and error analysis of empirical feature-based regularization schemes, *Journal of Machine Learning Research*, 17 (2016), 3058–3091.
[5] Z. C. Guo, S. B. Lin and D. X. Zhou, Learning theory of distributed spectral algorithms, *Inverse Problems*, 33 (2017), 0740009, 29pp.
[6] T. Hu, J. Fan, Q. Wu and D. X. Zhou, Regularization schemes for minimum error entropy principle, Analysis and Applications, 13 (2015), 437–455.
[7] S. B. Lin, X. Guo and D. X. Zhou, Distributed learning with regularized least squares, Journal of Machine Learning Research, 18 (2017), Paper No. 92, 31 pp.
[8] I. McKeague and M. Qian, Estimation of treatment policies based on functional predictors, Statistica Sinica, 24 (2014), 1461–1485.
[9] S. A. Murphy, Optimal dynamic treatment regimes, Journal of the Royal Statistical Society: Series B (Statistical Methodology), 65 (2003), 331–366.
[10] S. A. Murphy, An experimental design for the development of adaptive treatment strategies, Statistics in Medicine, 24 (2005), 1455–1481.
[11] Z. L. Qi and Y. F. Liu, D-learning to estimate optimal individual treatment rules, Electronic Journal of Statistics, 12 (2018), 3601–3638.
[12] J. O. Ramsay and B. W. Silverman, Applied Functional Data Analysis, Springer, New York, 2002.
[13] L. Shi, Distributed Learning with Indefinite Kernels, Analysis and Applications, 2019.
[14] R. Song, W. Wang, D. Zeng and M. R. Kosorok, Penalized q-learning for dynamic treatment regimes, Statistica Sinica, 25 (2015), 901–920.
[15] M. Yuan and T. T. Cai, A reproducing kernel Hilbert space approach to functional linear regression, The Annals of Statistics, 38 (2010), 3412–3444.
[16] T. Zhang, Learning bounds for kernel regression using effective data dimensionality, Neural Computation, 17 (2005), 2077–2098.
[17] Y. Zhao, D. Zeng, A. J. Rush and M. R. Kosorok, Estimating individualized treatment rules using outcome weighted learning, Journal of the American Statistical Association, 107 (2012), 1106–1118.

E-mail address: junfan@hkbu.edu.hk
E-mail address: lyufusheng@hkbu.edu.hk
E-mail address: leishi@fudan.edu.cn