A REPRODUCING KERNEL HILBERT SPACE APPROACH TO FUNCTIONAL LINEAR REGRESSION

BY MING YUAN[1] AND T. TONY CAI[2]

Georgia Institute of Technology and University of Pennsylvania

We study in this paper a smoothness regularization method for functional linear regression and provide a unified treatment for both the prediction and estimation problems. By developing a tool on simultaneous diagonalization of two positive definite kernels, we obtain shaper results on the minimax rates of convergence and show that smoothness regularized estimators achieve the optimal rates of convergence for both prediction and estimation under conditions weaker than those for the functional principal components based methods developed in the literature. Despite the generality of the method of regularization, we show that the procedure is easily implementable. Numerical results are obtained to illustrate the merits of the method and to demonstrate the theoretical developments.

1. Introduction. Consider the following functional linear regression model where the response Y is related to a square integrable random function $X(\cdot)$ through

$$
(1) \qquad Y = \alpha_0 + \int_{\mathcal{T}} X(t)\beta_0(t)\,dt + \varepsilon.
$$

Here α_0 is the intercept, $\mathcal{T}$ is the domain of $X(\cdot)$, $\beta_0(\cdot)$ is an unknown slope function and ε is a centered noise random variable. The domain $\mathcal{T}$ is assumed to be a compact subset of an Euclidean space. Our goal is to estimate α_0 and $\beta_0(\cdot)$ as well as to retrieve

$$
(2) \qquad \eta_0(X) := \alpha_0 + \int_{\mathcal{T}} X(t)\beta_0(t)\,dt
$$

based on a set of training data $(x_1, y_1), \ldots, (x_n, y_n)$ consisting of n independent copies of (X, Y). We shall assume that the slope function β_0 resides in a reproducing kernel Hilbert space (RKHS) $\mathcal{H}$, a subspace of the collection of square integrable functions on $\mathcal{T}$.

In this paper, we investigate the method of regularization for estimating η_0, as well as α_0 and β_0. Let ℓ_n be a data fit functional that measures how well η fits the data and J be a penalty functional that assesses the "plausibility" of η. The method of regularization estimates η_0 by

$$(3) \qquad \hat{\eta}_{n\lambda} = \arg\min_{\eta}[\ell_n(\eta|\text{data}) + \lambda J(\eta)],$$

where the minimization is taken over

$$(4) \qquad \left\{\eta : \mathcal{L}_2(\mathcal{T}) \to \mathbb{R} | \eta(X) = \alpha + \int_{\mathcal{T}} X\beta : \alpha \in \mathbb{R}, \beta \in \mathcal{H}\right\},$$

and $\lambda \geq 0$ is a tuning parameter that balances the fidelity to the data and the plausibility. Equivalently, the minimization can be taken over (α, β) instead of η to obtain estimates for both the intercept and slope, denoted by $\hat{\alpha}_{n\lambda}$ and $\hat{\beta}_{n\lambda}$ hereafter. The most common choice of the data fit functional is the squared error

$$(5) \qquad \ell_n(\eta) = \frac{1}{n} \sum_{i=1}^{n} [y_i - \eta(x_i)]^2.$$

In general, ℓ_n is chosen such that it is convex in η and $E\ell_n(\eta)$ in uniquely minimized by η_0.

In the context of functional linear regression, the penalty functional can be conveniently defined through the slope function β as a squared norm or semi-norm associated with $\mathcal{H}$. The canonical example of $\mathcal{H}$ is the Sobolev spaces. Without loss of generality, assume that $\mathcal{T} = [0, 1]$, the Sobolev space of order m is then defined as

$$\mathcal{W}_2^m([0, 1]) = \{\beta : [0, 1] \to \mathbb{R} | \beta, \beta^{(1)}, \dots, \beta^{(m-1)} \text{ are absolutely}$$

$$\text{continuous and } \beta^{(m)} \in \mathcal{L}_2\}.$$

There are many possible norms that can be equipped with $\mathcal{W}_2^m$ to make it a reproducing kernel Hilbert space. For example, it can be endowed with the norm

$$(6) \qquad \|\beta\|_{\mathcal{W}_2^m}^2 = \sum_{q=0}^{m-1} \left(\int \beta^{(q)}\right)^2 + \int (\beta^{(m)})^2.$$

The readers are referred to Adams (1975) for a thorough treatment of this subject. In this case, a possible choice of the penalty functional is given by

$$(7) \qquad J(\beta) = \int_0^1 [\beta^{(m)}(t)]^2 \, dt.$$

Another setting of particular interest is $\mathcal{T} = [0, 1]^2$ which naturally occurs when X represents an image. A popular choice in this setting is the thin plate spline where J is given by

$$(8) \qquad J(\beta) = \int_0^1 \int_0^1 \left[\left(\frac{\partial^2 \beta}{\partial x_1^2}\right)^2 + 2\left(\frac{\partial^2 \beta}{\partial x_1 \partial x_2}\right)^2 + \left(\frac{\partial^2 \beta}{\partial x_2^2}\right)^2\right] dx_1 \, dx_2,$$

and (x_1, x_2) are the arguments of bivariate function β. Other examples of $\mathcal{T}$ include $\mathcal{T} = \{1, 2, \ldots, p\}$ for some positive integer p, and unit sphere in an Euclidean space among others. The readers are referred to Wahba (1990) for common choices of $\mathcal{H}$ and J in these as well as other contexts.

Other than the methods of regularization, a number of alternative estimators have been introduced in recent years for the functional linear regression [James (2002); Cardot, Ferraty and Sarda (2003); Ramsay and Silverman (2005); Yao, Müller and Wang (2005); Ferraty and Vieu (2006); Cai and Hall (2006); Li and Hsing (2007); Hall and Horowitz (2007); Crambes, Kneip and Sarda (2009); Johannes (2009)]. Most of the existing methods are based upon the functional principal component analysis (FPCA). The success of these approaches hinges on the availability of a good estimate of the functional principal components for $X(\cdot)$. In contrast, the aforementioned smoothness regularized estimator avoids this task and therefore circumvents assumptions on the spacing of the eigenvalues of the covariance operator for $X(\cdot)$ as well as Fourier coefficients of β_0 with respect to the eigenfunctions, which are required by the FPCA-based approaches. Furthermore, as we shall see in the subsequent theoretical analysis, because the regularized estimator does not rely on estimating the functional principle components, stronger results on the convergence rates can be obtained.

Despite the generality of the method of regularization, we show that the estimators can be computed rather efficiently. We first derive a representer theorem in Section 2 which demonstrates that although the minimization with respect to η in (3) is taken over an infinite-dimensional space, the solution can actually be found in a finite-dimensional subspace. This result makes our procedure easily implementable and enables us to take advantage of the existing techniques and algorithms for smoothing splines to compute $\hat{\eta}_{n\lambda}$, $\hat{\beta}_{n\lambda}$ and $\hat{\alpha}_{n\lambda}$.

We then consider in Section 3 the relationship between the eigen structures of the covariance operator for $X(\cdot)$ and the reproducing kernel of the RKHS $\mathcal{H}$. These eigen structures play prominent roles in determining the difficulty of the prediction and estimation problems in functional linear regression. We prove in Section 3 a result on simultaneous diagonalization of the reproducing kernel of the RKHS $\mathcal{H}$ and the covariance operator of $X(\cdot)$ which provides a powerful machinery for studying the minimax rates of convergence.

Section 4 investigates the rates of convergence of the smoothness regularized estimators. Both the minimax upper and lower bounds are established. The optimal convergence rates are derived in terms of a class of intermediate norms which provide a wide range of measures for the estimation accuracy. In particular, this approach gives a unified treatment for both the prediction of $\eta_0(X)$ and the estimation of β_0. The results show that the smoothness regularized estimators achieve the optimal rate of convergence for both prediction and estimation under conditions weaker than those for the functional principal components based methods developed in the literature.

The representer theorem makes the regularized estimators easy to implement. Several efficient algorithms are available in the literature that can be used for the numerical implementation of our procedure. Section 5 presents numerical studies to illustrate the merits of the method as well as demonstrate the theoretical developments. All proofs are relegated to Section 6.

2. Representer theorem. The smoothness regularized estimators $\hat{\eta}_{n\lambda}$ and $\hat{\beta}_{n\lambda}$ are defined as the solution to a minimization problem over an infinite-dimensional space. Before studying the properties of the estimators, we first show that the minimization is indeed well defined and easily computable thanks to a version of the so-called representer theorem.

Let the penalty functional J be a squared semi-norm on $\mathcal{H}$ such that the null space

$$(9) \qquad \mathcal{H}_0 := \{\beta \in \mathcal{H} : J(\beta) = 0\}$$

is a finite-dimensional linear subspace of $\mathcal{H}$ with orthonormal basis $\{\xi_1, \ldots, \xi_N\}$ where $N := \dim(\mathcal{H}_0)$. Denote by $\mathcal{H}_1$ its orthogonal complement in $\mathcal{H}$ such that $\mathcal{H} = \mathcal{H}_0 \oplus \mathcal{H}_1$. Similarly, for any function $f \in \mathcal{H}$, there exists a unique decomposition $f = f_0 + f_1$ such that $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$. Note $\mathcal{H}_1$ forms a reproducing kernel Hilbert space with the inner product of $\mathcal{H}$ restricted to $\mathcal{H}_1$. Let $K(\cdot, \cdot)$ be the corresponding reproducing kernel of $\mathcal{H}_1$ such that $J(f_1) = \|f_1\|_K^2 = \|f_1\|_{\mathcal{H}}^2$ for any $f_1 \in \mathcal{H}_1$. Hereafter we use the subscript K to emphasize the correspondence between the inner product and its reproducing kernel.

In what follows, we shall assume that K is continuous and square integrable. Note that K is also a nonnegative definite operator on $\mathcal{L}_2$. With slight abuse of notation, write

$$(10) \qquad (Kf)(\cdot) = \int_{\mathcal{T}} K(\cdot, s) f(s) \, ds.$$

It is known [see, e.g., Cucker and Smale (2001)] that $Kf \in \mathcal{H}_1$ for any $f \in \mathcal{L}_2$. Furthermore, for any $f \in \mathcal{H}_1$

$$(11) \qquad \int_{\mathcal{T}} f(t)\beta(t) \, dt = \langle Kf, \beta \rangle_{\mathcal{H}}.$$

This observation allows us to prove the following result which is important to both numerical implementation of the procedure and our theoretical analysis.

THEOREM 1. *Assume that ℓ_n depends on η only through $\eta(x_1), \eta(x_2), \ldots, \eta(x_n)$; then there exist $\mathbf{d} = (d_1, \ldots, d_N)' \in \mathbb{R}^N$ and $\mathbf{c} = (c_1, \ldots, c_n)' \in \mathbb{R}^n$ such that*

$$(12) \qquad \hat{\beta}_{n\lambda}(t) = \sum_{k=1}^{N} d_k \xi_k(t) + \sum_{i=1}^{n} c_i (Kx_i)(t).$$

Theorem 1 is a generalization of the well-known representer lemma for smoothing splines (Wahba, 1990). It demonstrates that although the minimization with respect to η is taken over an infinite-dimensional space, the solution can actually be found in a finite-dimensional subspace, and it suffices to evaluate the coefficients **c** and **d** in (12). Its proof follows a similar argument as that of Theorem 1.3.1 in Wahba (1990) where ℓ_n is assumed to be squared error, and is therefore omitted here for brevity.

Consider, for example, the squared error loss. The regularized estimator is given by

$$(13) \qquad (\hat{\alpha}_{n\lambda}, \hat{\beta}_{n\lambda}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathcal{H}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \left(\alpha + \int_{\mathcal{T}} x_i(t)\beta(t)\,dt \right) \right]^2 + \lambda J(\beta) \right\}.$$

It is not hard to see that

$$(14) \qquad \hat{\alpha}_{n\lambda} = \bar{y} - \int_{\mathcal{T}} \bar{x}(t)\hat{\beta}_{n\lambda}(t)\,dt,$$

where $\bar{x}(t) = \frac{1}{n}\sum_{i=1}^{n} x_i(t)$ and $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ are the sample average of x and y, respectively. Consequently, (13) yields

$$(15) \qquad \hat{\beta}_{n\lambda} = \underset{\beta \in \mathcal{H}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \int_{\mathcal{T}} (x_i(t) - \bar{x}(t))\beta(t)\,dt \right]^2 + \lambda J(\beta) \right\}.$$

For the purpose of illustration, assume that $\mathcal{H} = \mathcal{W}_2^2$ and $J(\beta) = \int(\beta'')^2$. Then $\mathcal{H}_0$ is the linear space spanned by $\xi_1(t) = 1$ and $\xi_2(t) = t$. A popular reproducing kernel associated with $\mathcal{H}_1$ is

$$(16) \qquad K(s, t) = \frac{1}{(2!)^2} B_2(s)B_2(t) - \frac{1}{4!} B_4(|s - t|),$$

where $B_m(\cdot)$ is the mth Bernoulli polynomial. The readers are referred to Wahba (1990) for further details. Following Theorem 1, it suffices to consider β of the following form:

$$(17) \qquad \beta(t) = d_1 + d_2 t + \sum_{i=1}^{n} c_i \int_{\mathcal{T}} [x_i(s) - \bar{x}(s)] K(t, s)\,ds$$

for some $\mathbf{d} \in \mathbb{R}^2$ and $\mathbf{c} \in \mathbb{R}^n$. Correspondingly,

$$\int_{\mathcal{T}} [X(t) - \bar{x}(t)]\beta(t)\,dt$$

$$= d_1 \int_{\mathcal{T}} [X(t) - \bar{x}(t)]\,dt + d_2 \int_{\mathcal{T}} [X(t) - \bar{x}(t)]t\,dt$$

$$+ \sum_{i=1}^{n} c_i \int_{\mathcal{T}} \int_{\mathcal{T}} [x_i(s) - \bar{x}(s)] K(t, s)[X(t) - \bar{x}(t)]\,ds\,dt.$$

Note also that for β given in (17)

$$(18) \qquad J(\beta) = \mathbf{c}' \Sigma \mathbf{c},$$

where $\Sigma = (\Sigma_{ij})$ is a $n \times n$ matrix with

$$(19) \qquad \Sigma_{ij} = \int_{\mathcal{T}} \int_{\mathcal{T}} [x_i(s) - \bar{x}(s)] K(t,s)[x_j(t) - \bar{x}(t)] \, ds \, dt.$$

Denote by $T = (T_{ij})$ an $n \times 2$ matrix whose (i, j) entry is

$$(20) \qquad T_{ij} = \int [x_i(t) - \bar{x}(t)] t^{j-1} \, dt$$

for $j = 1$, 2. Set $\mathbf{y} = (y_1, \ldots, y_n)'$. Then

$$(21) \qquad \ell_n(\eta) + \lambda J(\beta) = \frac{1}{n} \|\mathbf{y} - (T\mathbf{d} + \Sigma \mathbf{c})\|_{\ell_2}^2 + \lambda \mathbf{c}' \Sigma \mathbf{c},$$

which is quadratic in $\mathbf{c}$ and $\mathbf{d}$, and the explicit form of the solution can be easily obtained for such a problem. This computational problem is similar to that behind the smoothing splines. Write $W = \Sigma + n\lambda I$; then the minimizer of (21) is given by

$$\mathbf{d} = (T'W^{-1}T)^{-1} T'W^{-1}\mathbf{y},$$
$$\mathbf{c} = W^{-1}[I - T(T'W^{-1}T)^{-1} T'W^{-1}]\mathbf{y}.$$

3. Simultaneous diagonalization. Before studying the asymptotic properties of the regularized estimators $\hat{\eta}_{n\lambda}$ and $\hat{\beta}_{n\lambda}$, we first investigate the relationship between the eigen structures of the covariance operator for $X(\cdot)$ and the reproducing kernel of the functional space $\mathcal{H}$. As observed in earlier studies [e.g., Cai and Hall (2006); Hall and Horowitz (2007)], eigen structures play prominent roles in determining the nature of the estimation problem in functional linear regression.

Recall that K is the reproducing kernel of $\mathcal{H}_1$. Because K is continuous and square integrable, it follows from Mercer's theorem [Riesz and Sz-Nagy (1955)] that K admits the following spectral decomposition:

$$(22) \qquad K(s, t) = \sum_{k=1}^{\infty} \rho_k \psi_k(s) \psi_k(t).$$

Here $\rho_1 \geq \rho_2 \geq \cdots$ are the eigenvalues of K, and $\{\psi_1, \psi_2, \ldots\}$ are the corresponding eigenfunctions, that is,

$$(23) \qquad K\psi_k = \rho_k \psi_k, \qquad k = 1, 2, \ldots.$$

Moreover,

$$(24) \qquad \langle \psi_i, \psi_j \rangle_{\mathcal{L}_2} = \delta_{ij} \quad \text{and} \quad \langle \psi_i, \psi_j \rangle_K = \delta_{ij}/\rho_j,$$

where δ_{ij} is the Kronecker's delta.

Consider, for example, the univariate Sobolev space $\mathcal{W}_2^m([0,1])$ with norm (6) and penalty (7). Observe that

$$(25) \qquad \mathcal{H}_1 = \left\{ f \in \mathcal{H} : \int f^{(k)} = 0, k = 0, 1, \ldots, m-1 \right\}.$$

It is known that [see, e.g., Wahba (1990)]

$$(26) \qquad K(s,t) = \frac{1}{(m!)^2} B_m(s) B_m(t) + \frac{(-1)^{m-1}}{(2m)!} B_{2m}(|s-t|).$$

Recall that B_m is the mth Bernoulli polynomial. It is known [see, e.g., Micchelli and Wahba (1981)] that in this case, $\rho_k \asymp k^{-2m}$, where for two positive sequences a_k and b_k, $a_k \asymp b_k$ means that a_k/b_k is bounded away from 0 and ∞ as $k \to \infty$.

Denote by C the covariance operator for X, that is,

$$(27) \qquad C(s,t) = E\{[X(s) - E(X(s))][X(t) - E(X(t))]\}.$$

There is a duality between reproducing kernel Hilbert spaces and covariance operators [Stein (1999)]. Similarly to the reproducing kernel K, assuming that the covariance operator C is continuous and square integrable, we also have the following spectral decomposition

$$(28) \qquad C(s,t) = \sum_{k=1}^{\infty} \mu_k \phi_k(s) \phi_k(t),$$

where $\mu_1 \geq \mu_2 \geq \cdots$ are the eigenvalues and $\{\phi_1, \phi_2, \ldots\}$ are the eigenfunctions such that

$$(29) \qquad C\phi_k := \int_{\mathcal{T}} C(\cdot, t) \phi_k(t) \, dt = \mu_k \phi_k, \qquad k = 1, 2, \ldots.$$

The decay rate of the eigenvalues $\{\mu_k : k \geq 1\}$ can be determined by the smoothness of the covariance operator C. More specifically, when C satisfies the so-called Sacks–Ylvisaker conditions of order s where s is a nonnegative integer [Sacks and Ylvisaker (1966, 1968, 1970)], then $\mu_k \asymp k^{-2(s+1)}$. The readers are referred to the original papers by Sacks and Ylvisaker or a more recent paper by Ritter, Wasilkowski and Woźniakwski (1995) for detailed discussions of the Sacks–Ylvisaker conditions. The conditions are also stated in the Appendix for completeness. Roughly speaking, a covariance operator C is said to satisfy the Sacks–Ylvisaker conditions of order 0 if it is twice differentiable when $s \neq t$ but not differentiable when $s = t$. A covariance operator C satisfies the Sacks–Ylvisaker conditions of order r for an integer $r > 0$ if $\partial^{2r} C(s,t)/(\partial s^r \, \partial t^r)$ satisfies the Sacks–Ylvisaker conditions of order 0. In this paper, we say a covariance operator C satisfies the Sacks–Ylvisaker conditions if C satisfies the Sacks–Ylvisaker conditions of order r for some $r \geq 0$. Various examples of covariance functions are known to satisfy Sacks–Ylvisaker conditions. For example, the Ornstein–Uhlenbeck covariance function $C(s,t) = \exp(-|s-t|)$ satisfies

the Sacks–Ylvisaker conditions of order 0. Ritter, Wasilkowski and Woźniakowski (1995) recently showed that covariance functions satisfying the Sacks–Ylvisaker conditions are also intimately related to Sobolev spaces, a fact that is useful for the purpose of simultaneously diagonalizing K and C as we shall see later.

Note that the two sets of eigenfunctions $\{\psi_1, \psi_2, \ldots\}$ and $\{\phi_1, \phi_2, \ldots\}$ may differ from each other. The two kernels K and C can, however, be simultaneously diagonalized. To avoid ambiguity, we shall assume in what follows that $Cf \neq 0$ for any $f \in \mathcal{H}_0$ and $f \neq 0$. When using the squared error loss, this is also a necessary condition to ensure that $E\ell_n(\eta)$ is uniquely minimized even if β is known to come from the finite-dimensional space $\mathcal{H}_0$. Under this assumption, we can define a norm $\|\cdot\|_R$ in $\mathcal{H}$ by

$$(30) \quad \|f\|_R^2 = \langle Cf, f\rangle_{\mathcal{L}_2} + J(f) = \int_{\mathcal{T}\times\mathcal{T}} f(s)C(s,t)f(t)\,ds\,dt + J(f).$$

Note that $\|\cdot\|_R$ is a norm because $\|f\|_R^2$ defined above is a quadratic form and is zero if and only if $f = 0$.

The following proposition shows that when this condition holds, $\|\cdot\|_R$ is well defined on $\mathcal{H}$ and equivalent to its original norm, $\|\cdot\|_{\mathcal{H}}$, in that there exist constants $0 < c_1 < c_2 < \infty$ such that $c_1\|f\|_R \leq \|f\|_{\mathcal{H}} \leq c_2\|f\|_R$ for all $f \in \mathcal{H}$. In particular, $\|f\|_R < \infty$ if and only if $\|f\|_{\mathcal{H}} < \infty$.

PROPOSITION 2. *If $Cf \neq 0$ for any $f \in \mathcal{H}_0$ and $f \neq 0$, then $\|\cdot\|_R$ and $\|\cdot\|_{\mathcal{H}}$ are equivalent.*

Let R be the reproducing kernel associated with $\|\cdot\|_R$. Recall that R can also be viewed as a positive operator. Denote by $\{(\rho_k', \psi_k') : k \geq 1\}$ the eigenvalues and eigenfunctions of R. Then R is a linear map from $\mathcal{L}_2$ to $\mathcal{L}_2$ such that

$$(31) \quad R\psi_k' = \int_{\mathcal{T}} R(\cdot, t)\psi_k'(t)\,dt = \rho_k'\psi_k', \qquad k = 1, 2, \ldots.$$

The square root of the positive definite operator can therefore be given as the linear map from $\mathcal{L}_2$ to $\mathcal{L}_2$ such that

$$(32) \quad R^{1/2}\psi_k' = (\rho_k')^{1/2}\psi_k', \qquad k = 1, 2, \ldots.$$

Let $\nu_1 \geq \nu_2 \geq \cdots$ be the eigenvalues of the bounded linear operator $R^{1/2}CR^{1/2}$ and $\{\zeta_k : k = 1, 2, \ldots\}$ be the corresponding orthogonal eigenfunctions in $\mathcal{L}_2$. Write $\omega_k = \nu_k^{-1/2}R^{1/2}\zeta_k$, $k = 1, 2, \ldots$. Also let $\langle\cdot, \cdot\rangle_R$ be the inner product associated with $\|\cdot\|_R$, that is, for any $f, g \in \mathcal{H}$,

$$(33) \quad \langle f, g\rangle_R = \tfrac{1}{4}(\|f+g\|_R^2 - \|f-g\|_R^2).$$

It is not hard to see that

$$(34) \quad \langle\omega_j, \omega_k\rangle_R = \nu_j^{-1/2}\nu_k^{-1/2}\langle R^{1/2}\zeta_j, R^{1/2}\zeta_k\rangle_R = \nu_k^{-1}\langle\zeta_j, \zeta_k\rangle_{\mathcal{L}_2} = \nu_k^{-1}\delta_{jk},$$

and

$$
\begin{aligned}
\langle C^{1/2}\omega_j, C^{1/2}\omega_k\rangle_{\mathcal{L}_2} &= v_j^{-1/2} v_k^{-1/2} \langle C^{1/2} R^{1/2}\zeta_j, C^{1/2} R^{1/2}\zeta_k\rangle_{\mathcal{L}_2} \\
&= v_j^{-1/2} v_k^{-1/2} \langle R^{1/2} C R^{1/2}\zeta_j, \zeta_k\rangle_{\mathcal{L}_2} \\
&= \delta_{jk}.
\end{aligned}
$$

The following theorem shows that quadratic forms $\|f\|_R^2 = \langle f, f\rangle_R$ and $\langle Cf, f\rangle_{\mathcal{L}_2}$ can be simultaneously diagonalized on the basis of $\{\omega_k : k \geq 1\}$.

THEOREM 3.   *For any $f \in \mathcal{H}$,*

$$
f = \sum_{k=1}^{\infty} f_k \omega_k, \tag{35}
$$

in the absolute sense where $f_k = v_k\langle f, \omega_k\rangle_R$. Furthermore, if $\gamma_k = (v_k^{-1} - 1)^{-1}$, then

$$
\langle f, f\rangle_R = \sum_{k=1}^{\infty} (1 + \gamma_k^{-1}) f_k^2 \quad and \quad \langle Cf, f\rangle_{\mathcal{L}_2} = \sum_{k=1}^{\infty} f_k^2. \tag{36}
$$

Consequently,

$$
J(f) = \langle f, f\rangle_R - \langle Cf, f\rangle_{\mathcal{L}_2} = \sum_{k=1}^{\infty} \gamma_k^{-1} f_k^2. \tag{37}
$$

Note that $\{(\gamma_k, \omega_k) : k \geq 1\}$ can be determined jointly by $\{(\rho_k, \psi_k) : k \geq 1\}$ and $\{(\mu_k, \phi_k) : k \geq 1\}$. However, in general, neither γ_k nor ω_k can be given in explicit form of $\{(\rho_k, \psi_k) : k \geq 1\}$ and $\{(\mu_k, \phi_k) : k \geq 1\}$. One notable exception is the case when the operators C and K are commutable. In particular, the setting $\psi_k = \phi_k$, $k = 1, 2, \ldots$, is commonly adopted when studying FPCA-based approaches [see, e.g., Cai and Hall (2006); Hall and Horowitz (2007)].

PROPOSITION 4.   *Assume that $\psi_k = \phi_k$, $k = 1, 2, \ldots$, then $\gamma_k = \rho_k\mu_k$ and $\omega_k = \mu_k^{-1/2}\psi_k$.*

In general, when ψ_k and ϕ_k differ, such a relationship no longer holds. The following theorem reveals that similar asymptotic behavior of γ_k can still be expected in many practical settings.

THEOREM 5.   *Consider the one-dimensional case when $\mathcal{T} = [0, 1]$. If $\mathcal{H}$ is the Sobolev space $\mathcal{W}_2^m([0, 1])$ endowed with norm (6), and C satisfies the Sacks–Ylvisaker conditions, then $\gamma_k \asymp \mu_k\rho_k$.*

Theorem 5 shows that under fairly general conditions $\gamma_k \asymp \mu_k \rho_k$. In this case, there is little difference between the general situation and the special case when K and C share a common set of eigenfunctions when working with the system $\{(\gamma_k, \omega_k), k = 1, 2, \ldots\}$. This observation is crucial for our theoretical development in the next section.

4. Convergence rates. We now turn to the asymptotic properties of the smoothness regularized estimators. To fix ideas, in what follows, we shall focus on the squared error loss. Recall that in this case

$$(38) \qquad (\hat{\alpha}_{n\lambda}, \hat{\beta}_{n\lambda}) = \underset{\alpha \in \mathbb{R}, \beta \in \mathcal{H}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[y_i - \left(\alpha + \int_{\mathcal{T}} x_i(t)\beta(t)\, dt \right) \right]^2 + \lambda J(\beta) \right\}.$$

As shown before, the slope function can be equivalently defined as

$$(39) \qquad \hat{\beta}_{n\lambda} = \underset{\beta \in \mathcal{H}}{\arg\min} \left\{ \frac{1}{n} \sum_{i=1}^{n} \left[(y_i - \bar{y}) - \int_{\mathcal{T}} (x_i(t) - \bar{x}(t))\beta(t)\, dt \right]^2 + \lambda J(\beta) \right\},$$

and once $\hat{\beta}_{n\lambda}$ is computed, $\hat{\alpha}_{n\lambda}$ is given by

$$(40) \qquad \hat{\alpha}_{n\lambda} = \bar{y} - \int_{\mathcal{T}} \bar{x}(t)\hat{\beta}_{n\lambda}(t)\, dt.$$

In light of this fact, we shall focus our attention on $\hat{\beta}_{n\lambda}$ in the following discussion for brevity. We shall also assume that the eigenvalues of the reproducing kernel K satisfies $\rho_k \asymp k^{-2r}$ for some $r > 1/2$. Let $\mathcal{F}(s, M, K)$ be the collection of the distributions F of the process X that satisfy the following conditions:

(a) The eigenvalues μ_k of its covariance operator $C(\cdot, \cdot)$ satisfy $\mu_k \asymp k^{-2s}$ for some $s > 1/2$.

(b) For any function $f \in \mathcal{L}_2(\mathcal{T})$,

$$(41) \qquad \begin{aligned} & E\left(\int_{\mathcal{T}} f(t)[X(t) - E(X)(t)]\, dt \right)^4 \\ & \qquad \leq M \left[E\left(\int_{\mathcal{T}} f(t)[X(t) - E(X)(t)]\, dt \right)^2 \right]^2. \end{aligned}$$

(c) When simultaneously diagonalizing K and C, $\gamma_k \asymp \rho_k \mu_k$, where $\nu_k = (1 + \gamma_k^{-1})^{-1}$ is the kth largest eigenvalue of $R^{1/2}CR^{1/2}$ where R is the reproducing kernel associated with $\|\cdot\|_R$ defined by (30).

The first condition specifies the smoothness of the sample path of $X(\cdot)$. The second condition concerns the fourth moment of a linear functional of $X(\cdot)$. This condition is satisfied with $M = 3$ for a Gaussian process because $\int f(t)X(t)\, dt$ is normally distributed. In the light of Theorem 5, the last condition is satisfied by any covariance function that satisfies the Sacks–Ylvisaker conditions if $\mathcal{H}$ is taken to be $\mathcal{W}_2^m$ with norm (6). It is also trivially satisfied if the eigenfunctions of the covariance operator C coincide with those of K.

4.1. *Optimal rates of convergence.* We are now ready to state our main results on the optimal rates of convergence, which are given in terms of a class of intermediate norms between $\|f\|_K$ and

$$(42) \qquad \left(\int\int f(s)C(s,t)f(t)\,ds\,dt \right)^{1/2},$$

which enables a unified treatment of both the prediction and estimation problems. For $0 \le a \le 1$ define the norm $\|\cdot\|_a$ by

$$(43) \qquad \|f\|_a^2 = \sum_{k=1}^{\infty} (1 + \gamma_k^{-a}) f_k^2,$$

where $f_k = \nu_k \langle f, \omega_k \rangle_R$ as shown in Theorem 3. Clearly $\|f\|_0$ reduces to $\langle Cf, f \rangle_{\mathcal{L}_2}$ whereas $\|f\|_1 = \|f\|_R$. The convergence rate results given below are valid for all $0 \le a \le 1$. They cover a range of interesting cases including the prediction error and estimation error.

The following result gives the optimal rate of convergence for the regularized estimator $\hat{\beta}_{n\lambda}$ with an appropriately chosen tuning parameter λ under the loss $\|\cdot\|_a$.

THEOREM 6. *Assume that $E(\varepsilon_i) = 0$ and $\mathrm{Var}(\varepsilon_i) \le M_2$. Suppose the eigenvalues ρ_k of the reproducing kernel K of the RKHS $\mathcal{H}$ satisfy $\rho_k \asymp k^{-2r}$ for some $r > 1/2$. Then the regularized estimator $\hat{\beta}_{n\lambda}$ with*

$$(44) \qquad \lambda \asymp n^{-2(r+s)/(2(r+s)+1)}$$

satisfies

$$(45) \qquad \begin{aligned} \lim_{D\to\infty} \varlimsup_{n\to\infty} \sup_{F\in\mathcal{F}(s,M,K),\beta_0\in\mathcal{H}} & P\big(\|\hat{\beta}_{n\lambda} - \beta_0\|_a^2 > Dn^{-2(1-a)(r+s)/(2(r+s)+1)}\big) \\ & = 0. \end{aligned}$$

Note that the rate of the optimal choice of λ does not depend on a. Theorem 6 shows that the optimal rate of convergence for the regularized estimator $\hat{\beta}_{n\lambda}$ is $n^{-2(1-a)(r+s)/(2(r+s)+1)}$. The following lower bound result demonstrates that this rate of convergence is indeed optimal among all estimators, and consequently the upper bound in equation (45) cannot be improved. Denote by $\mathcal{B}$ the collection of all measurable functions of the observations $(X_1, Y_1), \ldots, (X_n, Y_n)$.

THEOREM 7. *Under the assumptions of Theorem 6, there exists a constant $d > 0$ such that*

$$(46) \qquad \lim_{n\to\infty} \inf_{\tilde{\beta}\in\mathcal{B}} \sup_{F\in\mathcal{F}(s,M,K),\beta_0\in\mathcal{H}} P\big(\|\tilde{\beta} - \beta_0\|_a^2 > dn^{-2(1-a)(r+s)/(2(r+s)+1)}\big) > 0.$$

Consequently, the regularized estimator $\hat{\beta}_{n\lambda}$ with $\lambda \asymp n^{-2(r+s)/(2(r+s)+1)}$ is rate optimal.

The results, given in terms of $\| \cdot \|_a$, provide a wide range of measures of the quality of an estimate for β_0. Observe that

$$(47) \qquad \|\tilde{\beta} - \beta_0\|_0^2 = E_{X^*}\left(\int \tilde{\beta}(t)X^*(t)\,dt - \int \beta_0(t)X^*(t)\,dt \right)^2,$$

where X^* is an independent copy of X, and the expectation on the right-hand side is taken over X^*. The right-hand side is often referred to as the prediction error in regression. It measures the mean squared prediction error for a random future observation on X. From Theorems 6 and 7, we have the following corollary.

COROLLARY 8. *Under the assumptions of Theorem 6, the mean squared optimal prediction error of a slope function estimator over $F \in \mathcal{F}(s, M, K)$ and $\beta_0 \in \mathcal{H}$ is of the order $n^{-\frac{2(r+s)}{2(r+s)+1}}$ and it can be achieved by the regularized estimator $\hat{\beta}_{n\lambda}$ with λ satisfying (44).*

The result shows that the faster the eigenvalues of the covariance operator C for $X(\cdot)$ decay, the smaller the prediction error.

When $\psi_k = \phi_k$, the prediction error of a slope function estimator $\tilde{\beta}$ can also be understood as the squared prediction error for a fixed predictor $x^*(\cdot)$ such that $|\langle x^*, \phi_k \rangle_{\mathcal{L}_2}| \asymp k^{-s}$ following the discussed from the last section. A similar prediction problem has also been considered by Cai and Hall (2006) for FPCA-based approaches. In particular, they established a similar minimax lower bound and showed that the lower bound can be achieved by the FPCA-based approach, but with additional assumptions that $\mu_k - \mu_{k+1} \geq C_0^{-1} k^{-2s-1}$, and $2r > 4s + 3$. Our results here indicate that both restrictions are unnecessary for establishing the minimax rate for the prediction error. Moreover, in contrast to the FPCA-based approach, the regularized estimator $\hat{\beta}_{n\lambda}$ can achieve the optimal rate without the extra requirements.

To illustrate the generality of our results, we consider an example where $\mathcal{T} = [0, 1]$, $\mathcal{H} = \mathcal{W}_2^m([0, 1])$ and the stochastic process $X(\cdot)$ is a Wiener process. It is not hard to see that the covariance operator of X, $C(s, t) = \min\{s, t\}$, satisfies the Sacks–Ylvisaker conditions of order 0 and therefore $\mu_k \asymp k^{-2}$. By Corollary 8, the minimax rate of the prediction error in estimating β_0 is $n^{-(2m+2)/(r2m+3)}$. Note that the condition $2r > 4s + 3$ required by Cai and Hall (2006) does not hold here for $m \leq 7/2$.

4.2. *The special case of $\phi_k = \psi_k$.* It is of interest to further look into the case when the operators C and K share a common set of eigenfunctions. As discussed in the last section, we have in this case $\phi_k = \psi_k$ and $\gamma_k \asymp k^{-2(r+s)}$ for all $k \geq 1$. In this context, Theorems 6 and 7 provide bounds for more general prediction problems. Consider estimating $\int x^*\beta_0$ where x^* satisfies $|\langle x^*, \phi_k \rangle_{\mathcal{L}_2}| \asymp k^{-s+q}$ for

some $0 < q < s - 1/2$. Note that $q < s - 1/2$ is needed to ensure that x^* is square integrable. The squared prediction error

$$(48) \qquad \left(\int \tilde{\beta}(t) x^*(t) \, dt - \int \beta_0(t) x^*(t) \, dt \right)^2$$

is therefore equivalent to $\|\tilde{\beta} - \beta_0\|_{(s-q)/(r+s)}$. The following result is a direct consequence of Theorems 6 and 7.

COROLLARY 9. *Suppose x^* is a function satisfying $|\langle x^*, \phi_k \rangle_{\mathcal{L}_2}| \asymp k^{-s+q}$ for some $0 < q < s - 1/2$. Then under the assumptions of Theorem 6,*

$$(49) \qquad \lim_{n \to \infty} \inf_{\tilde{\beta} \in \mathcal{B}} \sup_{F \in \mathcal{F}(s,M,K), \beta_0 \in \mathcal{H}} P \left\{ \left(\int \tilde{\beta}(t) x^*(t) \, dt - \int \beta_0(t) x^*(t) \, dt \right)^2 \right.$$
$$\left. > dn^{-2(r+q)/(2(r+s)+1)} \right\} > 0$$

for some constant $d > 0$, and the regularized estimator $\hat{\beta}_{n\lambda}$ with λ satisfying (44) achieves the optimal rate of convergence under the prediction error (48).

It is also evident that when $\psi_k = \phi_k$, $\| \cdot \|_{s/(r+s)}$ is equivalent to $\| \cdot \|_{\mathcal{L}_2}$. Therefore, Theorems 6 and 7 imply the following result.

COROLLARY 10. *If $\phi_k = \psi_k$ for all $k \geq 1$, then under the assumptions of Theorem 6*

$$(50) \qquad \lim_{n \to \infty} \inf_{\tilde{\beta} \in \mathcal{B}} \sup_{F \in \mathcal{F}(s,M,K), \beta_0 \in \mathcal{H}} P \left(\|\tilde{\beta} - \beta_0\|_{\mathcal{L}_2}^2 > dn^{-2r/(2(r+s)+1)} \right) > 0$$

for some constant $d > 0$, and the regularized estimate $\hat{\beta}_{n\lambda}$ with λ satisfying (44) achieves the optimal rate.

This result demonstrates that the faster the eigenvalues of the covariance operator for $X(\cdot)$ decay, the larger the estimation error. The behavior of the estimation error thus differs significantly from that of prediction error.

Similar results on the lower bound have recently been obtained by Hall and Horowitz (2007) who considered estimating β_0 under the assumption that $|\langle \beta_0, \phi_k \rangle_{\mathcal{L}_2}|$ decays in a polynomial order. Note that this slightly differs from our setting where $\beta_0 \in \mathcal{H}$ means that

$$(51) \qquad \sum_{k=1}^{\infty} \rho_k^{-1} \langle \beta_0, \psi_k \rangle_{\mathcal{L}_2}^2 = \sum_{k=1}^{\infty} \rho_k^{-1} \langle \beta_0, \phi_k \rangle_{\mathcal{L}_2}^2 < \infty.$$

Recall that $\rho_k \asymp k^{-2r}$. Condition (51) is comparable to, and slightly stronger than,

$$(52) \qquad |\langle \beta_0, \phi_k \rangle_{\mathcal{L}_2}| \leq M_0 k^{-r-1/2}$$

for some constant $M_0 > 0$. When further assuming that $2s + 1 < 2r$, and $\mu_k - \mu_{k+1} \geq M_0^{-1} k^{-2s-1}$ for all $k \geq 1$, Hall and Horowitz (2007) obtain the same lower bound as ours. However, we do not require that $2s + 1 < 2r$ which in essence states that β_0 is smoother than the sample path of X. Perhaps, more importantly, we do not require the spacing condition $\mu_k - \mu_{k+1} \geq M_0^{-1} k^{-2s-1}$ on the eigenvalues because we do not need to estimate the corresponding eigenfunctions. Such a condition is impossible to verify even for a standard RKHS.

4.3. *Estimating derivatives.* Theorems 6 and 7 can also be used for estimating the derivatives of β_0. A natural estimator of the qth derivative of β_0, $\beta_0^{(q)}$, is $\hat{\beta}_{n\lambda}^{(q)}$, the qth derivative of $\hat{\beta}_{n\lambda}$. In addition to $\phi_k = \psi_k$, assume that $\|\psi_k^{(q)}/\psi_k\|_\infty \asymp k^q$. This clearly holds when $\mathcal{H} = \mathcal{W}_2^m$. In this case

$$\tag{53} \|\tilde{\beta}^{(q)} - \beta_0^{(q)}\|_{\mathcal{L}_2} \leq C_0 \|\tilde{\beta} - \beta_0\|_{(s+q)/(r+s)}.$$

The following is then a direct consequence of Theorems 6 and 7.

COROLLARY 11. *Assume that $\phi_k = \psi_k$ and $\|\psi_k^{(q)}/\psi_k\|_\infty \asymp k^q$ for all $k \geq 1$. Then under the assumptions of Theorem 6, for some constant $d > 0$,*

$$\lim_{n \to \infty} \inf_{\tilde{\beta}^{(q)} \in \mathcal{B}} \sup_{F \in \mathcal{F}(s,M,K), \beta_0 \in \mathcal{H}} P\big(\|\tilde{\beta}^{(q)} - \beta_0^{(q)}\|_{\mathcal{L}_2}^2 > dn^{-2(r-q)/(2(r+s)+1)}\big)$$

$$\tag{54} > 0,$$

and the regularized estimate $\hat{\beta}_{n\lambda}$ with λ satisfying (44) achieves the optimal rate.

Finally, we note that although we have focused on the squared error loss here, the method of regularization can be easily extended to handle other goodness of fit measures as well as the generalized functional linear regression [Cardot and Sarda (2005) and Müller and Stadtmüller (2005)]. We shall leave these extensions for future studies.

5. Numerical results. The Representer Theorem given in Section 2 makes the regularized estimators easy to implement. Similarly to smoothness regularized estimators in other contexts [see, e.g., Wahba (1990)], $\hat{\eta}_{n\lambda}$ and $\hat{\beta}_{n\lambda}$ can be expressed as a linear combination of a finite number of known basis functions although the minimization in (3) is taken over an infinite-dimensional space. Existing algorithms for smoothing splines can thus be used to compute our regularized estimators $\hat{\eta}_{n\lambda}$, $\hat{\beta}_{n\lambda}$ and $\hat{\alpha}_{n\lambda}$.

To demonstrate the merits of the proposed estimators in finite sample settings, we carried out a set of simulation studies. We adopt the simulation setting of Hall and Horowitz (2007) where $\mathcal{T} = [0, 1]$. The true slope function β_0 is given by

$$\tag{55} \beta_0 = \sum_{k=1}^{50} 4(-1)^{k+1} k^{-2} \phi_k,$$

where $\phi_1(t) = 1$ and $\phi_{k+1}(t) = \sqrt{2}\cos(k\pi t)$ for $k \geq 1$. The random function X was generated as

$$(56) \qquad X = \sum_{k=1}^{50} \zeta_k Z_k \phi_k,$$

where Z_k are independently sampled from the uniform distribution on $[-\sqrt{3}, \sqrt{3}]$ and ζ_k are deterministic. It is not hard to see that ζ_k^2 are the eigenvalues of the covariance function of X. Following Hall and Horowitz (2007), two sets of ζ_k were used. In the first set, the eigenvalues are well spaced: $\zeta_k = (-1)^{k+1}k^{-\nu/2}$ with $\nu = 1.1, 1.5, 2$ or 4. In the second set,

$$(57) \qquad \zeta_k = \begin{cases} 1, & k = 1, \\ 0.2(-1)^{k+1}(1 - 0.0001k), & 2 \leq k \leq 4, \\ 0.2(-1)^{k+1}[(5\lfloor k/5 \rfloor)^{-\nu/2} - 0.0001(k \bmod 5)], & k \geq 5. \end{cases}$$

As in Hall and Horowitz (2007), regression models with $\varepsilon \sim N(0, \sigma^2)$ where $\sigma = 0.5$ and 1 were considered. To comprehend the effect of sample size, we consider $n = 50, 100, 200$ and 500. We apply the regularization method to each simulated dataset and examine its estimation accuracy as measured by integrated squared error $\|\hat{\beta}_{n\lambda} - \beta_0\|_{\mathcal{L}_2}^2$ and prediction error $\|\hat{\beta}_{n\lambda} - \beta_0\|_0^2$. For the purpose of illustration, we take $\mathcal{H} = \mathcal{W}_2^2$ and $J(\beta) = \int(\beta'')^2$, for which the detailed estimation procedure is given in Section 2. For each setting, the experiment was repeated 1000 times.

As is common in most smoothing methods, the choice of the tuning parameter plays an important role in the performance of the regularized estimators. Data-driven choice of the tuning parameter is a difficult problem. Here we apply the commonly used practical strategy of empirically choosing the value of λ through the generalized cross validation. Note that the regularized estimator is a linear estimator in that $\hat{\mathbf{y}} = H(\lambda)\mathbf{y}$ where $\hat{\mathbf{y}} = (\hat{\eta}_{n\lambda}(x_1), \ldots, \hat{\eta}_{n\lambda}(x_n))'$ and $H(\lambda)$ is the so-called hat matrix depending on λ. We then select the tuning parameter λ that minimizes

$$(58) \qquad \mathrm{GCV}(\lambda) = \frac{(1/n)\|\hat{\mathbf{y}} - \mathbf{y}\|_{\ell_2}^2}{(1 - \mathrm{tr}(H(\lambda))/n)^2}.$$

Denote by $\hat{\lambda}^{\mathrm{GCV}}$ the resulting choice of the tuning parameter.

We begin with the setting of well-spaced eigenvalues. The left panel of Figure 1 shows the prediction error, $\|\hat{\beta}_{n\lambda} - \beta_0\|_0^2$, for each combination of ν value and sample size when $\sigma = 0.5$. The results were averaged over 1000 simulation runs in each setting. Both axes are given in the log scale. The plot suggests that the estimation error converges at a polynomial rate as sample size n increases, which agrees with our theoretical results from the previous section. Furthermore, one can observe that with the same sample size, the prediction error tends to be smaller for larger ν. This also confirms our theoretical development which indicates that the
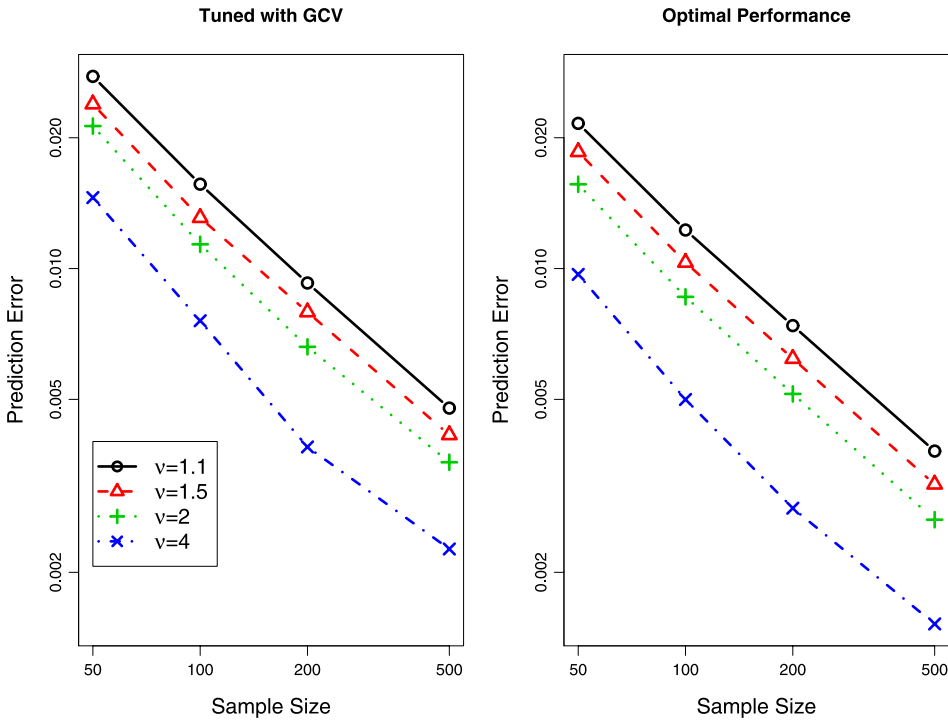
FIG. 1.   *Prediction errors of the regularized estimator* ($\sigma = 0.5$): *X was simulated with a covariance function with well-spaced eigenvalues. The results are averaged over* 1000 *runs. Black solid lines*, *red dashed lines*, *green dotted lines and blue dash-dotted lines correspond to* $\nu = 1.1$, 1.5, 2 *and* 4, *respectively. Both axes are in log scale.*

faster the eigenvalues of the covariance operator for $X(\cdot)$ decay, the smaller the prediction error.

To better understand the performance of the smoothness regularized estimator and the GCV choice of the tuning parameter, we also recorded the performance of an oracle estimator whose tuning parameter is chosen to minimize the prediction error. This choice of the tuning parameter ensures the optimal performance of the regularized estimator. It is, however, noteworthy that this is not a legitimate statistical estimator since it depends on the knowledge of unknown slope function β_0. The right panel of Figure 1 shows the prediction error associated with this choice of tuning parameter. It behaves similarly to the estimate with λ chosen by GCV. Note that the comparison between the two panels suggest that GCV generally leads to near optimal performance.

We now turn to the estimation error. Figure 2 shows the estimation errors, averaged over 1000 simulation runs, with λ chosen by GCV or minimizing the estimation error for each combination of sample size and ν value. Similarly to the prediction error, the plots suggest a polynomial rate of convergence of the estima-
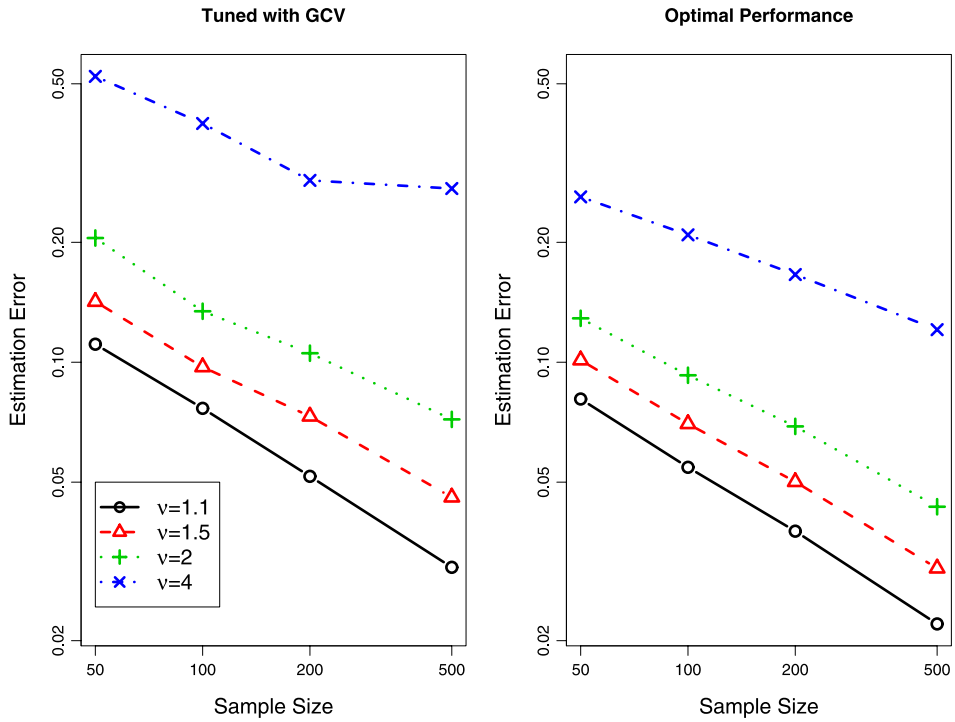
FIG. 2.   *Estimation errors of the regularized estimator ($\sigma = 0.5$): X was simulated with a covariance function with well-spaced eigenvalues. The results are averaged over* 1000 *runs. Black solid lines*, *red dashed lines*, *green dotted lines and blue dash-dotted lines correspond to* $v = 1.1, 1.5, 2$ *and* 4, *respectively. Both axes are in log scale.*

tion error when the sample size increases, and GCV again leads to near-optimal choice of the tuning parameter.

A comparison between Figures 1 and 2 suggests that when X is smoother (larger v), prediction (as measured by the prediction error) is easier, but estimation (as measured by the estimation error) tends to be harder, which highlights the difference between prediction and estimation in functional linear regression. We also note that this observation is in agreement with our theoretical results from the previous section where it is shown that the estimation error decreases at the rate of $n^{-2r/(2(r+s)+1)}$ which decelerates as s increases; whereas the prediction error decreases at the rate of $n^{-2(r+s)/(2(r+s)+1)}$ which accelerates as s increases.

Figure 3 reports the prediction and estimation error when tuned with GCV for the large noise ($\sigma = 1$) setting. Observations similar to those for the small noise setting can also be made. Furthermore, notice that the prediction errors are much smaller than the estimation error, which confirms our finding from the previous section that prediction is an easier problem in the context of functional linear regression.
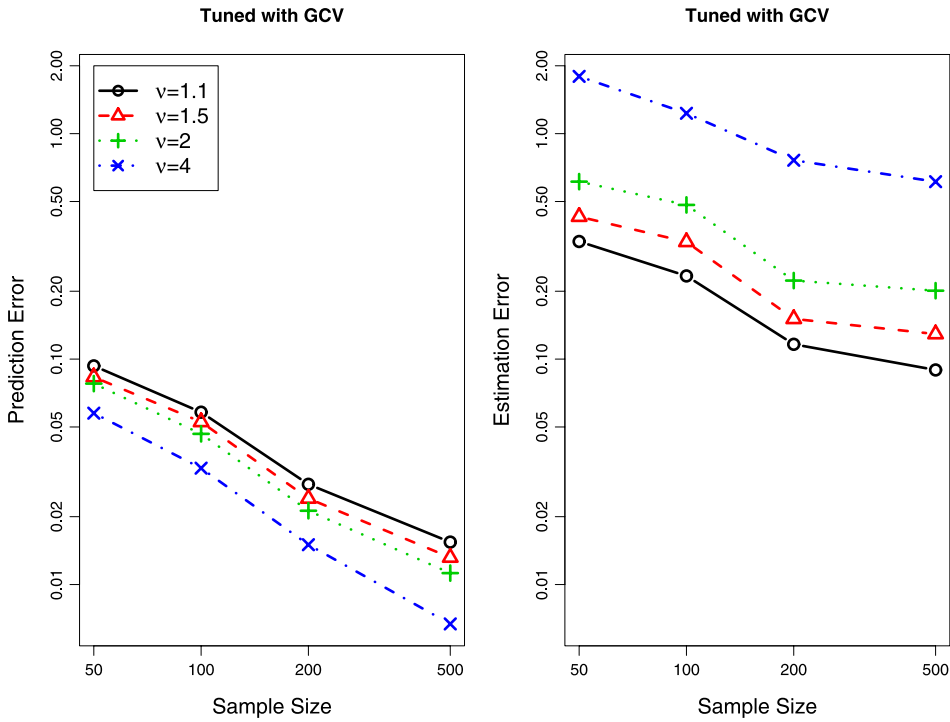
FIG. 3. *Estimation and prediction errors of the regularized estimator ($\sigma^2 = 1^2$): X was simulated with a covariance function with well-spaced eigenvalues. The results are averaged over 1000 runs. Black solid lines, red dashed lines, green dotted lines and blue dash-dotted lines correspond to $\nu = 1.1, 1.5, 2$ and 4, respectively. Both axes are in log scale.*

The numerical results in the setting with closely spaced eigenvalues are qualitatively similar to those in the setting with well-spaced eigenvalues. Figure 4 summarizes the results obtained for the setting with closely spaced eigenvalues.

We also note that the performance of the regularization estimate with λ tuned with GCV compares favorably with those from Hall and Horowitz (2007) using FPCA-based methods even though their results are obtained with optimal rather than data-driven choice of the tuning parameters.

6. Proofs.

6.1. *Proof of Proposition 2.* Observe that

$$\int_{\mathcal{T} \times \mathcal{T}} f(s) C(s, t) f(t) \, ds \, dt \leq \mu_1 \|f\|_{\mathcal{L}_2}^2 \leq c_1 \|f\|_{\mathcal{H}}^2 \tag{59}$$
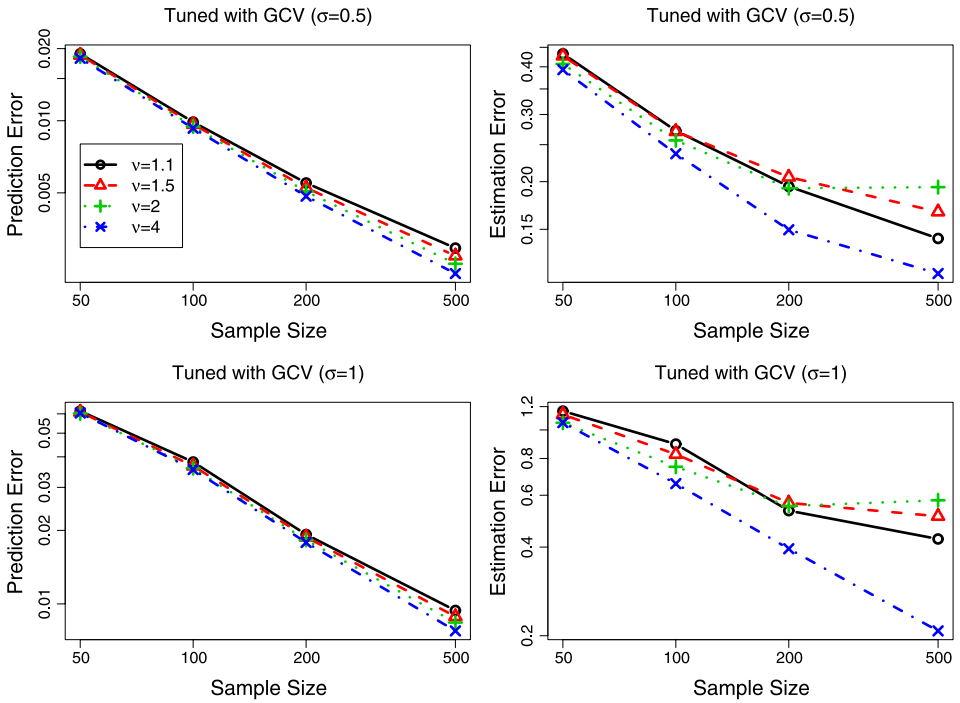
FIG. 4. *Estimation and prediction errors of the regularized estimator*: X *was simulated with a covariance function with closely-spaced eigenvalues. The results are averaged over* 1000 *runs. Both axes are in log scale. Note that y-axes are of different scales across panels.*

for some constant $c_1 > 0$. Together with the fact that $J(f) \leq \|f\|_{\mathcal{H}}^2$, we conclude that

$$(60) \qquad \|f\|_R^2 = \int_{\mathcal{T} \times \mathcal{T}} f(s)C(s,t)f(t)\,ds\,dt + J(f) \leq (c_1 + 1)\|f\|_{\mathcal{H}}^2.$$

Recall that ξ_k, $k = 1, \ldots, N$, are the orthonormal basis of $\mathcal{H}_0$. Under the assumption of the proposition, the matrix $\Sigma = (\langle C\xi_j, \xi_k \rangle_{\mathcal{H}})_{1 \leq j,k \leq N}$ is a positive definite matrix. Denote by $\mu_1' \geq \mu_2' \geq \cdots \geq \mu_N' > 0$ its eigenvalues. It is clear that for any $f_0 \in \mathcal{H}_0$

$$(61) \qquad \|f_0\|_R^2 \geq \mu_N' \|f_0\|_{\mathcal{H}}^2.$$

Note also that for any $f_1 \in \mathcal{H}_1$,

$$(62) \qquad \|f_1\|_{\mathcal{H}}^2 = J(f_1) \leq \|f_1\|_R^2.$$

For any $f \in \mathcal{H}$, we can write $f := f_0 + f_1$ where $f_0 \in \mathcal{H}_0$ and $f_1 \in \mathcal{H}_1$. Then

$$(63) \qquad \|f\|_R^2 = \int_{\mathcal{T} \times \mathcal{T}} f(s)C(s,t)f(t)\,ds\,dt + \|f_1\|_{\mathcal{H}}^2.$$

Recall that

(64) $$\|f_1\|_{\mathcal{H}}^2 \geq \rho_1^{-1}\|f_1\|_{\mathcal{L}_2}^2 \geq \rho_1^{-1}\mu_1^{-1}\int_{\mathcal{T}\times\mathcal{T}} f_0(s)C(s,t)f_0(t)\,ds\,dt.$$

For brevity, assume that $\rho_1 = \mu_1 = 1$ without loss of generality. By the Cauchy–Schwarz inequality,

$$
\begin{aligned}
\|f\|_R^2 &\geq \frac{1}{2}\int_{\mathcal{T}\times\mathcal{T}} f(s)C(s,t)f(t)\,ds\,dt + \|f_1\|_{\mathcal{H}}^2 \\
&\geq \frac{1}{2}\int_{\mathcal{T}\times\mathcal{T}} f_0(s)C(s,t)f_0(t)\,ds\,dt + \frac{3}{2}\int_{\mathcal{T}\times\mathcal{T}} f_1(s)C(s,t)f_1(t)\,ds\,dt \\
&\quad - \left(\int_{\mathcal{T}\times\mathcal{T}} f_0(s)C(s,t)f_0(t)\,ds\,dt\right)^{1/2} \\
&\quad\quad \times \left(\int_{\mathcal{T}\times\mathcal{T}} f_1(s)C(s,t)f_1(t)\,ds\,dt\right)^{1/2} \\
&\geq \frac{1}{3}\int_{\mathcal{T}\times\mathcal{T}} f_0(s)C(s,t)f_0(t)\,ds\,dt,
\end{aligned}
$$

where we used the fact that $3a^2/2 - ab \geq -b^2/6$ in deriving the last inequality. Therefore,

(65) $$\frac{3}{\mu_N'}\|f\|_R^2 \geq \|f_0\|_{\mathcal{H}}^2.$$

Together with the facts that $\|f\|_{\mathcal{H}}^2 = \|f_0\|_{\mathcal{H}}^2 + \|f_1\|_{\mathcal{H}}^2$ and

(66) $$\|f\|_R^2 \geq J(f_1) \geq \|f_1\|_{\mathcal{H}}^2,$$

we conclude that

(67) $$\|f\|_R^2 \geq (1 + 3/\mu_N')^{-1}\|f\|_{\mathcal{H}}^2.$$

The proof is now complete.

6.2. *Proof of Theorem* 3. First note that

$$
\begin{aligned}
R^{-1/2}f &= \sum_{k=1}^{\infty}\langle R^{-1/2}f, \zeta_k\rangle_{\mathcal{L}_2}\zeta_k = \sum_{k=1}^{\infty}\langle R^{-1/2}f, v_k^{1/2}R^{-1/2}\omega_k\rangle_{\mathcal{L}_2}v_k^{1/2}R^{-1/2}\omega_k \\
&= R^{-1/2}\left(\sum_{k=1}^{\infty}v_k\langle R^{-1/2}f, R^{-1/2}\omega_k\rangle_{\mathcal{L}_2}\omega_k\right) = R^{-1/2}\left(\sum_{k=1}^{\infty}v_k\langle f, \omega_k\rangle_R\omega_k\right).
\end{aligned}
$$

Applying bounded positive definite operator $R^{1/2}$ to both sides leads to

(68) $$f = \sum_{k=1}^{\infty}v_k\langle f, \omega_k\rangle_R\omega_k.$$

Recall that $\langle \omega_k, \omega_j \rangle_R = v_k^{-1} \delta_{kj}$. Therefore,

$$\|f\|_R^2 = \left\langle \sum_{k=1}^{\infty} v_k \langle f, \omega_k \rangle_R \omega_k, \sum_{j=1}^{\infty} v_j \langle f, \omega_j \rangle_R \omega_j \right\rangle_R$$

$$= \sum_{k,j=1} v_k v_j \langle f, \omega_k \rangle_R \langle f, \omega_j \rangle_R \langle \omega_k, \omega_j \rangle_R$$

$$= \sum_{k=1} v_k \langle f, \omega_k \rangle_R^2.$$

Similarly, because $\langle C\omega_k, \omega_j \rangle_{\mathcal{L}_2} = \delta_{kj}$,

$$\langle Cf, f \rangle_{\mathcal{L}_2} = \left\langle C\left(\sum_{k=1}^{\infty} v_k \langle f, \omega_k \rangle_R \omega_k \right), \sum_{j=1}^{\infty} v_j \langle f, \omega_j \rangle_R \omega_j \right\rangle_{\mathcal{L}_2}$$

$$= \left\langle \sum_{k=1}^{\infty} v_k \langle f, \omega_k \rangle_R C\omega_k, \sum_{j=1}^{\infty} v_j \langle f, \omega_j \rangle_R \omega_j \right\rangle_{\mathcal{L}_2}$$

$$= \sum_{k,j=1} v_k v_j \langle f, \omega_k \rangle_R \langle f, \omega_j \rangle_R \langle C\omega_k, \omega_j \rangle_{\mathcal{L}_2}$$

$$= \sum_{k=1} v_k^2 \langle f, \omega_k \rangle_R^2.$$

6.3. *Proof of Proposition* 4. Recall that for any $f \in \mathcal{H}_0$, $Cf \neq 0$ if and only if $f = 0$, which implies that $\mathcal{H}_0 \cap \text{l.s.}\{\phi_k : k \geq 1\}^\perp = \{0\}$. Together with the fact that $\mathcal{H}_0 \cap \mathcal{H}_1 = \{0\}$, we conclude that $\mathcal{H} = \mathcal{H}_1 = \text{l.s.}\{\phi_k : k \geq 1\}$. It is not hard to see that for any $f, g \in \mathcal{H}$,

$$(69) \qquad \langle f, g \rangle_R = \int_{\mathcal{T} \times \mathcal{T}} f(s) C(s, t) g(t) \, ds \, dt + \langle f, g \rangle_K.$$

In particular,

$$(70) \qquad \langle \psi_j, \psi_k \rangle_R = (\mu_k + \rho_k^{-1}) \delta_{jk},$$

which implies that $\{((\mu_k + \rho_k^{-1})^{-1}, \psi_k) : k \geq 1\}$ is also the eigen system of R, that is,

$$(71) \qquad R(s, t) = \sum_{k=1}^{\infty} (\mu_k + \rho_k^{-1})^{-1} \psi_k(s) \psi_k(t).$$

Then

$$(72) \qquad R\psi_k := \int_{\mathcal{T}} R(\cdot, t) \psi_k(t) \, dt = (\mu_k + \rho_k^{-1})^{-1} \psi_k, \qquad k = 1, 2, \ldots.$$

Therefore,

$$R^{1/2}CR^{1/2}\psi_k = R^{1/2}C\big((\mu_k + \rho_k^{-1})^{-1/2}\psi_k\big)$$
$$= R^{1/2}\big(\mu_k(\mu_k + \rho_k^{-1})^{-1/2}\psi_k\big) = (1 + \rho_k^{-1}\mu_k^{-1})^{-1}\psi_k,$$

which implies that $\zeta_k = \psi_k = \phi_k$, $\nu_k = (1 + \rho_k^{-1}\mu_k^{-1})^{-1}$ and $\gamma_k = \rho_k\mu_k$. Consequently,

$$(73) \qquad \omega_k = \nu_k^{-1/2}R^{1/2}\psi_k = \nu_k^{-1/2}(\mu_k + \rho_k^{-1})^{-1/2}\psi_k = \mu_k^{-1/2}\psi_k.$$

6.4. *Proof of Theorem* 5. Recall that $\mathcal{H} = \mathcal{W}_2^m$, which implies that $\rho_k \asymp k^{-2m}$. By Corollary 2 of Ritter, Wasilkowski and Woźniakowski (1995), $\mu_k \asymp k^{-2(s+1)}$. It therefore suffices to show $\gamma_k \asymp k^{-2(s+1+m)}$. The key idea of the proof is a result from Ritter, Wasilkowski and Woźniakowski (1995) indicating that the reproducing kernel Hilbert space associated with C differs from $\mathcal{W}_2^{s+1}([0,1])$ only by a finite-dimensional linear space of polynomials.

Denote by Q_r the reproducing kernel for $\mathcal{W}_2^r([0,1])$. Observe that $Q_r^{1/2}(\mathcal{L}_2) = \mathcal{W}_2^r$ [e.g., Cucker and Smale (2001)]. We begin by quantifying the decay rate of $\lambda_k(Q_m^{1/2}Q_{s+1}Q_m^{1/2})$. By Sobolev's embedding theorem, $(Q_{s+1}^{1/2}Q_m^{1/2})(\mathcal{L}_2) = Q_{s+1}^{1/2}(\mathcal{W}_2^m) = \mathcal{W}_2^{m+s+1}$. Therefore, $Q_m^{1/2}Q_{s+1}Q_m^{1/2}$ is equivalent to Q_{m+s+1}. Denote by $\lambda_k(Q)$ be the kth largest eigenvalue of a positive definite operator Q. Let $\{h_k : k \geq 1\}$ be the eigenfunctions of Q_{m+s+1}, that is, $Q_{m+s+1}h_k = \lambda_k(Q_{m+s+1})h_k$, $k = 1, 2, \ldots$. Denote by $\mathcal{F}_k$ and $\mathcal{F}_k^\perp$ the linear space spanned by $\{h_j : 1 \leq j \leq k\}$ and $\{h_j : j \geq k+1\}$, respectively. By the Courant–Fischer–Weyl min–max principle,

$$\lambda_k(Q_m^{1/2}Q_{s+1}Q_m^{1/2}) \geq \min_{f \in \mathcal{F}_k} \|Q_{s+1}^{1/2}Q_m^{1/2}f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2$$

$$\geq C_1 \min_{f \in \mathcal{F}_k} \|Q_{m+s+1}^{1/2}f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2$$

$$\geq C_1 \lambda_k(Q_{m+s+1})$$

for some constant $C_1 > 0$. On the other hand,

$$\lambda_k(Q_m^{1/2}Q_{s+1}Q_m^{1/2}) \leq \max_{f \in \mathcal{F}_{k-1}^\perp} \|Q_{s+1}^{1/2}Q_m^{1/2}f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2$$

$$\leq C_2 \min_{f \in \mathcal{F}_{k-1}^\perp} \|Q_{m+s+1}^{1/2}f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2$$

$$\leq C_2 \lambda_k(Q_{m+s+1})$$

for some constant $C_2 > 0$. In summary, we have $\lambda_k(Q_m^{1/2}Q_{s+1}Q_m^{1/2}) \asymp \times k^{-2(m+s+1)}$.

As shown by Ritter, Wasilkowski and Woźniakowski [(1995), Theorem 1, page 525], there exist D and U such that $Q_{s+1} = D + U$, D has at most $2(s+1)$

nonzero eigenvalues and $\|U^{1/2}f\|_{\mathcal{L}_2}$ is equivalent to $\|C^{1/2}f\|_{\mathcal{L}_2}$. Moreover, the eigenfunctions of D, denoted by $\{g_1, \ldots, g_d\}$ $(d \le 2(s+1))$ are polynomials of order no greater than $2s + 1$. Denote $\mathcal{G}$ the space spanned by $\{g_1, \ldots, g_d\}$. Clearly $\mathcal{G} \subset \mathcal{W}_2^m = Q_m^{1/2}(\mathcal{L}_2)$. Denote $\{\tilde{h}_j : j \ge 1\}$ the eigenfunctions of $Q_m^{1/2} Q_{s+1} Q_m^{1/2}$. Let $\tilde{\mathcal{F}}_k$ and $\tilde{\mathcal{F}}_k^\perp$ be defined similarly as $\mathcal{F}_k$ and $\mathcal{F}_k^\perp$. Then by the Courant–Fischer–Weyl min–max principle,

$$
\begin{aligned}
\lambda_{k-d}(Q_m^{1/2} U Q_m^{1/2}) &\ge \min_{f \in \tilde{\mathcal{F}}_k \cap Q_m^{-1/2}(\mathcal{G})^\perp} \|U^{1/2} Q_m^{1/2} f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2 \\
&= \min_{f \in \tilde{\mathcal{F}}_k \cap Q_m^{-1/2}(\mathcal{G})^\perp} \|Q_{s+1}^{1/2} Q_m^{1/2} f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2 \\
&= \min_{f \in \tilde{\mathcal{F}}_k} \|Q_{s+1}^{1/2} Q_m^{1/2} f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2 \\
&\ge C_1 \lambda_k(Q_{m+s+1})
\end{aligned}
$$

for some constant $C_1 > 0$. On the other hand,

$$
\begin{aligned}
\lambda_{k+d}(Q_m^{1/2} Q_{s+1} Q_m^{1/2}) &\le \max_{f \in \tilde{\mathcal{F}}_{k-1}^\perp \cap Q_m^{-1/2}(\mathcal{G})^\perp} \|U^{1/2} Q_m^{1/2} f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2 \\
&= \max_{f \in \tilde{\mathcal{F}}_{k-1}^\perp \cap Q_m^{-1/2}(\mathcal{G})^\perp} \|Q_{s+1}^{1/2} Q_m^{1/2} f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2 \\
&= \min_{f \in \tilde{\mathcal{F}}_{k-1}^\perp} \|Q_{s+1}^{1/2} Q_m^{1/2} f\|_{\mathcal{L}_2}^2 / \|f\|_{\mathcal{L}_2}^2 \\
&\le C_2 \lambda_k(Q_{m+s+1})
\end{aligned}
$$

for some constant $C_2 > 0$. Hence $\lambda_k(Q_m^{1/2} U Q_m^{1/2}) \asymp k^{-2(m+s+1)}$.

Because $Q_m^{1/2} U Q_m^{1/2}$ is equivalent to $R^{1/2} C R^{1/2}$, following a similar argument as before, by the Courant–Fischer–Weyl min–max principle, we complete the the proof.

6.5. *Proof of Theorem* 6. We now proceed to prove Theorem 6. The analysis follows a similar spirit as the technique commonly used in the study of the rate of convergence of smoothing splines [see, e.g., Silverman (1982); Cox and O'Sullivan (1990)]. For brevity, we shall assume that $EX(\cdot) = 0$ in the rest of the proof. In this case, α_0 can be estimated by $\bar{y}$ and β_0 by

$$
(74) \qquad \hat{\beta}_{n\lambda} = \arg\min_{\beta \in \mathcal{H}} \left[\frac{1}{n} \sum_{i=1}^n \left(y_i - \int_{\mathcal{T}} x_i(t) \beta(t) \, dt \right)^2 + \lambda J(\beta) \right].
$$

The proof below also applies to the more general setting when $EX(\cdot) \ne 0$ but with considerable technical obscurity.

Recall that

$$(75) \qquad \ell_n(\beta) = \frac{1}{n}\sum_{i=1}^{n}\left(y_i - \int_{\mathcal{T}} x_i(t)\beta(t)\,dt\right)^2.$$

Observe that

$$\ell_\infty(\beta) := E\ell_n(\beta) = E\left[Y - \int_{\mathcal{T}} X(t)\beta(t)\,dt\right]^2$$

$$= \sigma^2 + \int_{\mathcal{T}}\int_{\mathcal{T}}[\beta(s) - \beta_0(s)]C(s,t)[\beta(t) - \beta_0(t)]\,ds\,dt$$

$$= \sigma^2 + \|\beta - \beta_0\|_0^2.$$

Write

$$(76) \qquad \bar\beta_{\infty\lambda} = \arg\min_{\beta\in\mathcal{H}}\{\ell_\infty(\beta) + \lambda J(\beta)\}.$$

Clearly

$$(77) \qquad \hat\beta_{n\lambda} - \beta_0 = (\hat\beta_{n\lambda} - \bar\beta_{\infty\lambda}) + (\bar\beta_{\infty\lambda} - \beta_0).$$

We refer to the two terms on the right-hand side stochastic error and deterministic error, respectively.

6.5.1. *Deterministic error.* Write $\beta_0(\cdot) = \sum_{k=1}^{\infty} a_k\omega_k(\cdot)$ and $\beta(\cdot) = \sum_{k=1}^{\infty} b_k\omega_k(\cdot)$. Then Theorem 3 implies that

$$\ell_\infty(\beta) = \sigma^2 + \sum_{k=1}^{\infty}(b_k - a_k)^2, \qquad J(\beta) = \sum_{k=1}^{\infty}\gamma_k^{-1}b_k^2.$$

Therefore,

$$(78) \qquad \bar\beta_{\infty\lambda}(\cdot) = \sum_{k=1}^{\infty}\frac{a_k}{1 + \lambda\gamma_k^{-1}}\omega_k(\cdot) =: \sum_{k=1}^{\infty}\bar b_k\omega_k(\cdot).$$

It can then be computed that for any $a < 1$,

$$\|\bar\beta_{\infty\lambda} - \beta_0\|_a^2 = \sum_{k=1}^{\infty}(1 + \gamma_k^{-a})(\bar b_k - a_k)^2$$

$$= \sum_{k=1}^{\infty}(1 + \gamma_k^{-a})\left(\frac{\lambda\gamma_k^{-1}}{1 + \lambda\gamma_k^{-1}}\right)^2 a_k^2$$

$$\leq \lambda^2\sup_k\frac{(1 + \gamma^{-a})\gamma_k^{-1}}{(1 + \lambda\gamma_k^{-1})^2}\sum_{k=1}^{\infty}\gamma_k^{-1}a_k^2$$

$$= \lambda^2 J(\beta_0)\sup_k\frac{(1 + \gamma^{-a})\gamma_k^{-1}}{(1 + \lambda\gamma_k^{-1})^2}.$$

Now note that

$$\sup_{k} \frac{(1+\gamma^{-a})\gamma_k^{-1}}{(1+\lambda\gamma_k^{-1})^2} \le \sup_{x>0} \frac{(1+x^{-a})x^{-1}}{(1+\lambda x^{-1})^2}$$

$$\le \sup_{x>0} \frac{x^{-1}}{(1+\lambda x^{-1})^2} + \sup_{x>0} \frac{x^{-(a+1)}}{(1+\lambda x^{-1})^2}$$

$$= \frac{1}{\inf_{x>0}(x^{1/2}+\lambda x^{-1/2})^2} + \frac{1}{\inf_{x>0}(x^{(a+1)/2}+\lambda x^{-(1-a)/2})^2}$$

$$= \frac{1}{4\lambda} + C_0\lambda^{-(a+1)}.$$

Hereafter, we use C_0 to denote a generic positive constant. In summary, we have

LEMMA 12. *If λ is bounded from above, then*
$$\|\bar{\beta}_{\infty\lambda} - \beta_0\|_a^2 = O(\lambda^{1-a}J(\beta_0)).$$

6.5.2. *Stochastic error.* Next, we consider the stochastic error $\hat{\beta}_{n\lambda} - \bar{\beta}_{\infty\lambda}$. Denote

$$D\ell_n(\beta)f = -\frac{2}{n}\sum_{i=1}^{n}\left[\left(y_i - \int_{\mathcal{T}} x_i(t)\beta(t)\,dt\right)\int_{\mathcal{T}} x_i(t)f(t)\,dt\right],$$

$$D\ell_\infty(\beta)f = -2E_X\left(\int_{\mathcal{T}} X(t)[\beta_0(t)-\beta(t)]\,dt \int_{\mathcal{T}} X(t)f(t)\,dt\right)$$

$$= -2\int_{\mathcal{T}}\int_{\mathcal{T}}[\beta_0(s)-\beta(s)]C(s,t)f(t)\,ds\,dt,$$

$$D^2\ell_n(\beta)fg = \frac{2}{n}\sum_{i=1}^{n}\left[\int_{\mathcal{T}} x_i(t)f(t)\,dt \int_{\mathcal{T}} x_i(t)g(t)\,dt\right],$$

$$D^2\ell_\infty(\beta)fg = 2\int_{\mathcal{T}}\int_{\mathcal{T}} f(s)C(s,t)g(t)\,ds\,dt.$$

Also write $\ell_{n\lambda}(\beta) = \ell_n(\beta) + \lambda J(\beta)$ and $\ell_{\infty\lambda} = \ell_\infty(\beta) + \lambda J(\beta)$. Denote $G_\lambda = (1/2)D^2\ell_{\infty\lambda}(\bar{\beta}_{\infty\lambda})$ and

(79) $$\tilde{\beta} = \bar{\beta}_{\infty\lambda} - \tfrac{1}{2}G_\lambda^{-1}D\ell_{n\lambda}(\bar{\beta}_{\infty\lambda}).$$

It is clear that

(80) $$\hat{\beta}_{n\lambda} - \bar{\beta}_{\infty\lambda} = (\hat{\beta}_{n\lambda} - \tilde{\beta}) + (\tilde{\beta} - \bar{\beta}_{\infty\lambda}).$$

We now study the two terms on the right-hand side separately. For brevity, we shall abbreviate the subscripts of $\hat{\beta}$ and $\bar{\beta}$ in what follows. We begin with $\tilde{\beta} - \bar{\beta}$. Hereafter we shall omit the subscript for brevity if no confusion occurs.

LEMMA 13. *For any $0 \le a \le 1$,*

$$(81) \qquad E\|\tilde{\beta} - \bar{\beta}\|_a^2 \asymp n^{-1}\lambda^{-(a+1/(2(r+s)))}.$$

PROOF. Notice that $D\ell_{n\lambda}(\bar{\beta}) = D\ell_{n\lambda}(\bar{\beta}) - D\ell_{\infty\lambda}(\bar{\beta}) = D\ell_n(\bar{\beta}) - D\ell_\infty(\bar{\beta})$. Therefore

$$
\begin{aligned}
E[D\ell_{n\lambda}(\bar{\beta})f]^2 &= E[D\ell_n(\bar{\beta})f - D\ell_\infty(\bar{\beta})f]^2 \\
&= \frac{4}{n}\text{Var}\left[\left(Y - \int_\mathcal{T} X(t)\bar{\beta}(t)\,dt\right)\int_\mathcal{T} X(t)f(t)\,dt\right] \\
&\le \frac{4}{n}E\left[\left(Y - \int_\mathcal{T} X(t)\bar{\beta}(t)\,dt\right)\int_\mathcal{T} X(t)f(t)\,dt\right]^2 \\
&= \frac{4}{n}E\left(\int_\mathcal{T} X(t)[\beta_0(t) - \bar{\beta}(t)]\,dt\int_\mathcal{T} X(t)f(t)\,dt\right)^2 \\
&\quad + \frac{4\sigma^2}{n}E\left(\int_\mathcal{T} X(t)f(t)\,dt\right)^2,
\end{aligned}
$$

where we used the fact that $\varepsilon = Y - \int X\beta_0$ is uncorrelated with X. To bound the first term, an application of the Cauchy–Schwarz inequality yields

$$
\begin{aligned}
E&\left(\int_\mathcal{T} X(t)[\beta_0(t) - \bar{\beta}(t)]\,dt\int_\mathcal{T} X(t)f(t)\,dt\right)^2 \\
&\le \left\{E\left(\int_\mathcal{T} X(t)[\beta_0(t) - \bar{\beta}(t)]\,dt\right)^4 E\left(\int_\mathcal{T} X(t)f(t)\,dt\right)^4\right\}^{1/2} \\
&\le M\|\beta_0 - \bar{\beta}\|_0^2\|f\|_0^2,
\end{aligned}
$$

where the second inequality holds by the second condition of $\mathcal{F}(s, M, K)$. Therefore,

$$(82) \qquad E[D\ell_{n\lambda}(\bar{\beta})f]^2 \le \frac{4M}{n}\|\beta_0 - \bar{\beta}\|_0^2\|f\|_0^2 + \frac{4\sigma^2}{n}\|f\|_0^2,$$

which by Lemma 12 is further bounded by $(C_0\sigma^2/n)\|f\|_0^2$ for some positive constant C_0. Recall that $\|\omega_k\|_0 = 1$. We have

$$(83) \qquad E[D\ell_{n\lambda}(\bar{\beta})\omega_k]^2 \le C_0\sigma^2/n.$$

Thus, by the definition of $\tilde{\beta}$,

$$
\begin{aligned}
E\|\tilde{\beta} - \bar{\beta}\|_a^2 &= E\left\|\frac{1}{2}G_\lambda^{-1}D\ell_{n\lambda}(\bar{\beta})\right\|_a^2 \\
&= \frac{1}{4}E\left[\sum_{k=1}^\infty (1 + \gamma_k^{-a})(1 + \lambda\gamma_k^{-1})^{-2}(D\ell_{n\lambda}(\bar{\beta})\omega_k)^2\right]
\end{aligned}
$$

$$\leq \frac{C_0 \sigma^2}{4n} \sum_{k=1}^{\infty} (1 + \gamma_k^{-a})(1 + \lambda \gamma_k^{-1})^{-2}$$

$$\leq \frac{C_0 \sigma^2}{4n} \sum_{k=1}^{\infty} (1 + k^{2a(r+s)})(1 + \lambda k^{2(r+s)})^{-2}$$

$$\asymp \frac{C_0 \sigma^2}{4n} \int_1^{\infty} x^{2a(r+s)}(1 + \lambda x^{2(r+s)})^{-2} \, dx$$

$$\asymp \frac{C_0 \sigma^2}{4n} \int_1^{\infty} (1 + \lambda x^{2(r+s)/(2a(r+s)+1)})^{-2} \, dx$$

$$= \frac{C_0 \sigma^2}{4n} \lambda^{-(a+1/(2(r+s)))} \int_{\lambda^{a+1/(2(r+s))}}^{\infty} (1 + x^{2(r+s)/(2a(r+s)+1)})^{-2} \, dx$$

$$\asymp n^{-1} \lambda^{-(a+1/(2(r+s)))}.$$

The proof is now complete. $\square$

Now we are in position to bound $E \|\hat{\beta} - \tilde{\beta}\|_a^2$. By definition,

(84) $$G_\lambda(\hat{\beta} - \tilde{\beta}) = \tfrac{1}{2} D^2 \ell_{\infty \lambda}(\bar{\beta})(\hat{\beta} - \tilde{\beta}).$$

First-order condition implies that

(85) $$D\ell_{n\lambda}(\hat{\beta}) = D\ell_{n\lambda}(\bar{\beta}) + D^2 \ell_{n\lambda}(\bar{\beta})(\hat{\beta} - \bar{\beta}) = 0,$$

where we used the fact that $\ell_{n,\lambda}$ is quadratic. Together with the fact that

(86) $$D\ell_{n\lambda}(\bar{\beta}) + D^2 \ell_{\infty\lambda}(\bar{\beta})(\tilde{\beta} - \bar{\beta}) = 0,$$

we have

$$\begin{aligned} D^2 \ell_{\infty\lambda}(\bar{\beta})(\hat{\beta} - \tilde{\beta}) &= D^2 \ell_{\infty\lambda}(\bar{\beta})(\hat{\beta} - \bar{\beta}) + D^2 \ell_{\infty\lambda}(\bar{\beta})(\bar{\beta} - \tilde{\beta}) \\ &= D^2 \ell_{\infty\lambda}(\bar{\beta})(\hat{\beta} - \bar{\beta}) - D^2 \ell_{n\lambda}(\bar{\beta})(\hat{\beta} - \bar{\beta}) \\ &= D^2 \ell_{\infty}(\bar{\beta})(\hat{\beta} - \bar{\beta}) - D^2 \ell_n(\bar{\beta})(\hat{\beta} - \bar{\beta}). \end{aligned}$$

Therefore,

(87) $$(\hat{\beta} - \tilde{\beta}) = \tfrac{1}{2} G_\lambda^{-1} [D^2 \ell_{\infty}(\bar{\beta})(\hat{\beta} - \bar{\beta}) - D^2 \ell_n(\bar{\beta})(\hat{\beta} - \bar{\beta})].$$

Write

(88) $$\hat{\beta} = \sum_{k=1}^{\infty} \hat{b}_k \omega_k \quad \text{and} \quad \bar{\beta} = \sum_{k=0}^{\infty} \bar{b}_k \omega_k.$$

Then

$$\|\hat{\beta} - \tilde{\beta}\|_a^2$$

$$= \frac{1}{4} \sum_{k=1}^{\infty} (1 + \lambda \gamma_k^{-1})^{-2} (1 + \gamma_k^{-a})$$

$$\times \left[\sum_{j=1}^{\infty} (\hat{b}_j - \bar{b}_j) \int_{\mathcal{T}} \int_{\mathcal{T}} \omega_j(s) \left(\frac{1}{n} \sum_{i=1}^{n} x_i(t) x_i(s) - C(s,t) \right) \right.$$

$$\left. \times \omega_k(t) \, ds \, dt \right]^2$$

$$\leq \frac{1}{4} \sum_{k=1}^{\infty} (1 + \lambda \gamma_k^{-1})^{-2} (1 + \gamma_k^{-a}) \left[\sum_{j=1}^{\infty} (\hat{b}_j - \bar{b}_j)^2 (1 + \gamma_j^{-c}) \right]$$

$$\times \left(\sum_{j=1}^{\infty} (1 + \gamma_j^{-c})^{-1} \left[\int_{\mathcal{T}} \int_{\mathcal{T}} \omega_j(s) \left(\frac{1}{n} \sum_{i=1}^{n} x_i(t) x_i(s) - C(s,t) \right) \right. \right.$$

$$\left. \left. \times \omega_k(t) \, ds \, dt \right]^2 \right),$$

where the inequality is due to the Cauchy–Schwarz inequality.

Note that

$$E \left(\sum_{j=1}^{\infty} (1 + \gamma_j^{-c})^{-1} \left[\int_{\mathcal{T}} \omega_j(s) \left(\frac{1}{n} \sum_{i=1}^{n} x_i(t) x_i(s) - C(s,t) \right) \omega_k(t) \, ds \, dt \right]^2 \right)$$

$$= \frac{1}{n} \sum_{j=1}^{\infty} (1 + \gamma_j^{-c})^{-1} \operatorname{Var} \left(\int_{\mathcal{T}} \omega_j(t) X(t) \, dt \int_{\mathcal{T}} \omega_k(t) X(t) \, dt \right)$$

$$\leq \frac{1}{n} \sum_{j=1}^{\infty} (1 + \gamma_j^{-c})^{-1} E \left[\left(\int_{\mathcal{T}} \omega_j(t) X(t) \, dt \right)^2 \left(\int_{\mathcal{T}} \omega_k(t) X(t) \, dt \right)^2 \right]$$

$$\leq \frac{1}{n} \sum_{j=1}^{\infty} (1 + \gamma_j^{-c})^{-1} E \left[\left(\int_{\mathcal{T}} \omega_j(t) X(t) \, dt \right)^4 \right]^{1/2} E \left[\left(\int_{\mathcal{T}} \omega_k(t) X(t) \, dt \right)^4 \right]^{1/2}$$

$$\leq \frac{M}{n} \sum_{j=1}^{\infty} (1 + \gamma_j^{-c})^{-1} E \left[\left(\int_{\mathcal{T}} \omega_j(t) X(t) \, dt \right)^2 \right] E \left[\left(\int_{\mathcal{T}} \omega_k(t) X(t) \, dt \right)^2 \right]$$

$$= \frac{M}{n} \sum_{j=1}^{\infty} (1 + \gamma_j^{-c})^{-1} \asymp n^{-1},$$

provided that $c > 1/2(r + s)$. On the other hand,

$$\sum_{k=1}^{\infty} (1 + \lambda \gamma_k^{-1})^{-2} (1 + \gamma_k^{-a})$$

$$\leq C_0 \sum_{k=1}^{\infty} (1 + \lambda k^{2(r+s)})^{-2} (1 + k^{2a(r+s)})$$

$$\asymp \int_1^{\infty} (1 + \lambda x^{2(r+s)})^{-2} x^{2a(r+s)} \, dx$$

$$\asymp \int_1^{\infty} (1 + \lambda x^{2(r+s)/(2a(r+s)+1)})^{-2} \, dx$$

$$= \lambda^{-(a+1/(2(r+s)))} \int_{\lambda^{a+1/(2(r+s))}}^{\infty} (1 + x^{2(r+s)/(2a(r+s)+1)})^{-2} \, dx$$

$$\asymp \lambda^{-(a+1/(2(r+s)))}.$$

To sum up,

$$(89) \qquad \|\hat{\beta} - \tilde{\beta}\|_a^2 = O_p(n^{-1} \lambda^{-(a+1/(2(r+s)))} \|\hat{\beta} - \bar{\beta}\|_c^2).$$

In particular, taking $a = c$ yields

$$(90) \qquad \|\hat{\beta} - \tilde{\beta}\|_c^2 = O_p(n^{-1} \lambda^{-(c+1/(2(r+s)))} \|\hat{\beta} - \bar{\beta}\|_c^2).$$

If

$$(91) \qquad n^{-1} \lambda^{-(c+1/(2(r+s)))} \to 0,$$

then

$$(92) \qquad \|\hat{\beta} - \tilde{\beta}\|_c = o_p(\|\hat{\beta} - \bar{\beta}\|_c).$$

Together with the triangular inequality

$$(93) \qquad \|\tilde{\beta} - \bar{\beta}\|_c \geq \|\hat{\beta} - \bar{\beta}\|_c - \|\hat{\beta} - \tilde{\beta}\|_c = (1 - o_p(1)) \|\hat{\beta} - \bar{\beta}\|_c.$$

Therefore,

$$(94) \qquad \|\hat{\beta} - \bar{\beta}\|_c = O_p(\|\tilde{\beta} - \bar{\beta}\|_c)$$

Together with Lemma 13, we have

$$(95) \qquad \|\hat{\beta} - \bar{\beta}\|_c^2 = O_p(n^{-1} \lambda^{-(c+1/(2(r+s)))}) = o_p(1).$$

Putting it back to (89), we now have:

LEMMA 14. *If there also exists some $1/2(r + s) < c \leq 1$ such that $n^{-1} \times \lambda^{-(c+1/2(r+s))} \to 0$, then*

$$(96) \qquad \|\hat{\beta} - \tilde{\beta}\|_a^2 = o_p(n^{-1} \lambda^{-(a+1/2(r+s))}).$$

Combining Lemmas 12–14, we have

$$\lim_{D\to\infty} \overline{\lim_{n\to\infty}} \sup_{F\in\mathcal{F}(s,M,K),\beta_0\in\mathcal{H}} P\big(\|\hat{\beta}_{n\lambda} - \beta_0\|_a^2 > Dn^{-2(1-a)(r+s)/(2(r+s)+1)}\big)$$

(97)
$$= 0$$

by taking $\lambda \asymp n^{-2(r+s)/(2(r+s)+1)}$.

6.6. *Proof of Theorem* 7. We now set out to show that $n^{-2(1-a)(r+s)/(2(r+s)+1)}$ is the optimal rate. It follows from a similar argument as that of Hall and Horowitz (2007). Consider a setting where $\psi_k = \phi_k$, $k = 1, 2, \ldots$. Clearly in this case we also have $\omega_k = \mu_k^{-1/2}\phi_k$. It suffices to show that the rate is optimal in this special case. Recall that $\beta_0 = \sum a_k\phi_k$. Set

(98)
$$a_k = \begin{cases} L_n^{-1/2}k^{-r}\theta_k, & L_n + 1 \le k \le 2L_n, \\ 0, & \text{otherwise}, \end{cases}$$

where L_n is the integer part of $n^{1/(2(r+s)+1)}$, and θ_k is either 0 or 1. It is clear that

(99)
$$\|\beta_0\|_K^2 \le \sum_{k=L_n+1}^{2L_n} L_n^{-1} = 1.$$

Therefore $\beta_0 \in \mathcal{H}$. Now let X admit the following expansion: $X = \sum_k \xi_k k^{-s}\phi_k$ where ξ_ks are independent random variables drawn from a uniform distribution on $[-\sqrt{3}, \sqrt{3}]$. Simple algebraic manipulation shows that the distribution of X belongs to $\mathcal{F}(s, 3)$. The observed data are

(100)
$$y_i = \sum_{k=L_n+1}^{2L_n} L_n^{-1/2}k^{-(r+s)}\xi_{ik}\theta_k + \varepsilon_i, \qquad i = 1, \ldots, n,$$

where the noise ε_i is assumed to be independently sampled from $N(0, M_2)$. As shown in Hall and Horowitz (2007),

(101)
$$\lim_{n\to\infty} \inf_{L_n<j\le 2L_n} \inf_{\tilde{\theta}_j} \overset{*}{\sup} E(\tilde{\theta}_j - \theta_j)^2 > 0,$$

where $\overset{*}{\sup}$ denotes the supremum over all 2^{L_n} choices of $(\theta_{L_n+1}, \ldots, \theta_{2L_n})$, and $\inf_{\tilde{\theta}}$ is taken over all measurable functions $\tilde{\theta}_j$ of the data. Therefore, for any estimate $\tilde{\beta}$,

(102)
$$\overset{*}{\sup}\|\tilde{\beta} - \beta_0\|_a^2 = \overset{*}{\sup} \sum_{k=L_n+1}^{2L_n} L_n^{-1}k^{-2(1-a)(r+s)} E(\tilde{\theta}_j - \theta_j)^2$$

$$\ge Mn^{-2(1-a)(r+s)/(2(r+s)+1)}$$

for some constant $M > 0$.

Denote

$$
(103) \qquad \tilde{\tilde{\theta}}_k = \begin{cases} 1, & \tilde{\theta}_k > 1, \\ \tilde{\theta}_k, & 0 \le \tilde{\theta}_k \le 1, \\ 0, & \tilde{\theta}_k < 0. \end{cases}
$$

It is easy to see that

$$
\sum_{k=L_n+1}^{2L_n} L_n^{-1} k^{-2(1-a)(r+s)} (\tilde{\theta}_j - \theta_j)^2
$$

$$
(104)
$$

$$
\ge \sum_{k=L_n+1}^{2L_n} L_n^{-1} k^{-2(1-a)(r+s)} (\tilde{\tilde{\theta}}_j - \theta_j)^2.
$$

Hence, we can assume that $0 \le \tilde{\theta}_j \le 1$ without loss of generality in establishing the lower bound. Subsequently,

$$
\sum_{k=L_n+1}^{2L_n} L_n^{-1} k^{-2(1-a)(r+s)} (\tilde{\theta}_j - \theta_j)^2 \le \sum_{k=L_n+1}^{2L_n} L_n^{-1} k^{-2(1-a)(r+s)}
$$

$$
(105)
$$

$$
\le L_n^{-2(1-a)(r+s)}.
$$

Together with (102), this implies that

$$
(106) \qquad \lim_{n \to \infty} \inf_{\tilde{\beta}} \overset{*}{\sup} P\big(\|\tilde{\beta} - \beta\|_a^2 > d n^{-2(1-a)(r+s)/(2(r+s)+1)}\big) > 0
$$

for some constant $d > 0$.

APPENDIX: SACKS–YLVISAKER CONDITIONS

In Section 3, we discussed the relationship between the smoothness of C and the decay of its eigenvalues. More precisely, the smoothness can be quantified by the so-called Sacks–Ylvisaker conditions. Following Ritter, Wasilkowski and Woźniakowski (1995), denote

$$
(107) \qquad \begin{aligned} \Omega_+ &= \{(s, t) \in (0, 1)^2 : s > t\} \quad \text{and} \\ \Omega_- &= \{(s, t) \in (0, 1)^2 : s < t\}. \end{aligned}
$$

Let $\operatorname{cl}(A)$ be the closure of a set A. Suppose that L is a continuous function on $\Omega_+ \cup \Omega_-$ such that $L|_{\Omega_j}$ is continuously extendable to $\operatorname{cl}(\Omega_j)$ for $j \in \{+, -\}$. By L_j we denote the extension of L to $[0, 1]^2$, which is continuous on $\operatorname{cl}(\Omega_j)$, and on $[0, 1]^2 \setminus \operatorname{cl}(\Omega_j)$. Furthermore write $M^{(k,l)}(s, t) = (\partial^{k+l}/(\partial s^k \, \partial t^l)) M(s, t)$. We say that a covariance function M on $[0, 1]^2$ satisfies the Sacks–Ylvisaker conditions of order r if the following three conditions hold:

(A) $L = M^{(r,r)}$ is continuous on $[0, 1]^2$, and its partial derivatives up to order 2 are continuous on $\Omega_+ \cup \Omega_-$, and they are continuously extendable to $\text{cl}(\Omega_+)$ and $\text{cl}(\Omega_-)$.

(B)

$$\min_{0 \leq s \leq 1} \left(L_-^{(1,0)}(s, s) - L_+^{(1,0)}(s, s) \right) > 0. \tag{108}$$

(C) $L_+^{(2,0)}(s, \cdot)$ belongs to the reproducing kernel Hilbert space spanned by L and furthermore

$$\sup_{0 \leq s \leq 1} \left\| L_+^{(2,0)}(s, \cdot) \right\|_L < \infty. \tag{109}$$

REFERENCES

ADAMS, R. A. (1975). *Sobolev Spaces*. Academic Press, New York. MR0450957

CAI, T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. MR2291496

CARDOT, H., FERRATY, F. and SARDA, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. MR1997162

CARDOT, H. and SARDA, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.* **92** 24–41. MR2102242

COX, D. D. and O'SULLIVAN, F. (1990). Asymptotic analysis of penalized likelihood and related estimators. *Ann. Statist.* **18** 1676–1695. MR1074429

CRAMBES, C., KNEIP, A. and SARDA, P. (2009). Smoothing splines estimators for functional linear regression. *Ann. Statist.* **37** 35–72. MR2488344

CUCKER, F. and SMALE, S. (2001). On the mathematical foundations of learning. *Bull. Amer. Math. Soc.* **39** 1–49. MR1864085

FERRATY, F. and VIEU, P. (2006). *Nonparametric Functional Data Analysis: Methods, Theory, Applications and Implementations*. Springer, New York. MR2229687

HALL, P. and HOROWITZ, J. L. (2007). Methodology and convergence rates for functional linear regression. *Ann. Statist.* **35** 70–91. MR2332269

JAMES, G. (2002). Generalized linear models with functional predictors. *J. Roy. Statist. Soc. Ser. B* **64** 411–432. MR1924298

JOHANNES, J. (2009). Nonparametric estimation in functional linear models with second order stationary regressors. Unpublished manuscript. Available at http://arxiv.org/abs/0901.4266v1.

LI, Y. and HSING, T. (2007). On the rates of convergence in functional linear regression. *J. Multivariate Anal.* **98** 1782–1804. MR2392433

MICCHELLI, C. and WAHBA, G. (1981). Design problems for optimal surface interpolation. In *Approximation Theory and Applications* (Z. Ziegler, ed.) 329–347. Academic Press, New York. MR0615422

MÜLLER, H. G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. MR2163159

RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. MR2168993

RIESZ, F. and SZ-NAGY, B. (1955). *Functional Analysis*. Ungar, New York. MR0071727

RITTER, K., WASILKOWSKI, G. and WOŹNIAKOWSKI, H. (1995). Multivariate integeration and approximation for random fields satisfying Sacks–Ylvisaker conditions. *Ann. Appl. Probab.* **5** 518–540. MR1336881

SACKS, J. and YLVISAKER, D. (1966). Designs for regression problems with correlated errors. *Ann. Math. Statist.* **37** 66–89. MR0192601

SACKS, J. and YLVISAKER, D. (1968). Designs for regression problems with correlated errors; many parameters. *Ann. Math. Statist.* **39** 49–69. MR0220424

SACKS, J. and YLVISAKER, D. (1970). Designs for regression problems with correlated errors III. *Ann. Math. Statist.* **41** 2057–2074. MR0270530

SILVERMAN, B. W. (1982). On the estimation of a probability density function by the maximum penalized likelihood method. *Ann. Statist.* **10** 795–810. MR0663433

STEIN, M. (1999). *Statistical Interpolation of Spatial Data: Some Theory for Kriging*. Springer, New York. MR1697409

WAHBA, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia. MR1045442

YAO, F., MÜLLER, H. G. and WANG, J. L. (2005). Functional linear regression analysis for longitudinal data. *Ann. Statist.* **33** 2873–2903. MR2253106

MILTON STEWART SCHOOL OF INDUSTRIAL AND
    SYSTEMS ENGINEERING
GEORGIA INSTITUTE OF TECHNOLOGY
ATLANTA, GEORGIA 30332
USA
E-MAIL: myuan@isye.gatech.edu

DEPARTMENT OF STATISTICS
THE WHARTON SCHOOL
UNIVERSITY OF PENNSYLVANIA
PHILADELPHIA, PENNSYLVANIA 19104
USA
E-MAIL: tcai@wharton.upenn.edu