An Online Projection Estimator for Nonparametric Regression in Reproducing Kernel Hilbert Spaces

University of Washington

Supplementary Material
S1 Supplementary Discussion on RKHS

In the main text we gave two equivalent definitions of RKHS: one based on the reproducing property and another one based on the Mercer expansion of the kernel.

The proposed method directly works with the eigenfunctions $\psi_j$, and it does not directly approximate either the kernel function $K$ or the kernel matrix $\mathbb{K}$. Although in many cases we start with a Mercer kernel in hand and calculate its eigendecomposition afterwards, it is not uncommon to begin with features and then attempt to calculate a closed-form of an implied kernel. This situation suits perfectly with our method: for the well-known the smoothing spline method proposed in [Wahba (1990) Chapter 2], the author starts with $\psi_j(x) = \sin(2j\pi x), \cos(2j\pi x)$ and shows us how to get the closed-form of the reproducing kernel for periodic Sobolev space $W^0_m(\text{per})$. However, such a Bernoulli polynomial closed-form of the kernel is no longer available when $m$ is not an integer, which corresponds to a fractional Sobolev space case; when considering kernel space on sphere $S^2$, some effort is required to obtain the closed-form expression even for simple cases ([Kennedy et al. (2013), Michel (2012)]), but the features are just orthonormal spherical harmonics; for multiscale kernels defined by compactly-supported wavelet eigenfunctions ([Opfer 2006]) or Legendre polynomials
it is also simplest to work directly with features rather than attempting to identify a closed-form expression for the implied kernel.

In the main text we provide the Mercer expansion of a Sobolev space $W_1^0([0, 1])$. We also state the (correct) expansion for Gaussian kernel (there are several versions in the literature that are not correctly normalized):

When $\bar{\rho}_X$ has density (w.r.t Lebesgue measure on $\mathbb{R}$) $\bar{\rho}_X = \frac{\alpha}{\sqrt{\pi}} \exp(-\alpha^2 x^2)$, we have the expansion of Gaussian kernel $K(x, z) = \exp(-\epsilon^2|x - z|^2)$ with

$$
\lambda_j = \sqrt{\frac{\alpha^2}{\alpha^2 + \delta^2 + \epsilon^2}} \left(\frac{\epsilon^2}{\alpha^2 + \delta^2 + \epsilon^2}\right)^{j-1} (S1.1)
$$

$$
\psi_j(x) = \gamma_j \exp(-\delta^2 x^2) H_{j-1}(\alpha \beta x) (S1.2)
$$

where the $H_j$ are Hermite polynomials of degree $j$, and

$$
\beta = \left(1 + \frac{(2\epsilon)^2}{\alpha}\right)^{1/4}, \quad \gamma_j = \sqrt{\frac{\beta}{2^{j-1} \Gamma(j)}} (S1.3)
$$

The multivariate Gaussian kernel’s eigenfunctions and eigenvalues are just the tensor product of the 1-dimension Gaussian kernel. Formally, the multivariate Gaussian kernel $K(x, z) = \exp(-\epsilon^2||x - z||^2)$ has the following expansion:

$$
K(x, z) = \sum_{j \in \mathbb{N}^d} \lambda_j^* \psi_j^*(x) \psi_j^*(z) (S1.4)
$$

where the eigenvalues and eigenfunctions are related to (S1.1) as

$$
\lambda_j^* = \prod_{l=1}^d \lambda_{j_l}, \quad \psi_j^*(x) = \prod_{l=1}^d \psi_{j_l}(x^{(l)}),
$$
where \(x^{(l)}\) is the \(l\)-th component of \(x \in \mathbb{R}^d\). There are also available numerical methods (independent of \((X_i, Y_i)'\)'s) for approximating kernel eigenfunctions in cases where analytical forms are not available, see Rakotch et al. (1975); Santin and Schaback (2016), (Rasmussen 2003, Section 4.3), Cai and Vassilevski (2020) and (Fasshauer and McCourt 2015, Chapter 12).

There is also an interesting formal similarity between Mercer expansions and Bonchner’s theorem (see, e.g. Rahimi and Recht (2007)) which gives rise to random Fourier feature-based methods. On one hand, we have the Mercer expansion:

\[
K(x, z) = \sum_{j=1}^{\infty} \lambda(j) \psi(x, j) \psi(z, j)
\]

On the other hand, the positive-definite (real-valued) kernel has a convolutional representation by Bonchner’s theorem (Rahimi and Recht, 2007):

\[
K(x, z) = \int_{X \times [0, 2\pi]} p(\omega, b) \cos(\omega^T x + b) \cos(\omega^T z + b) \, d\omega db
\]

The random Fourier feature expansion (S1.6) uses a set of basis functions (cosines) that is not sensitive to the expanded kernel. Only the probability distribution we sample \(\omega\) from depends on the kernel. Such a choice may bring some convenience in application, but at the price of using an approximation that converges to the kernel much slower. Another difference is in the basis selection strategy: For the Mercer expansion it is very straight-
forward – we choose the eigenfunctions corresponding to larger eigenvalues. By this strategy, we can ensure the features we choose are more important and orthogonal to each other w.r.t. RKHS inner product. For random feature-based methodologies, one has to sample from a probability distribution because there are uncountably infinitely many \( \omega \) (versus countably infinite \( j \)) and there is less we can say about the geometric properties of random features (Yu et al., 2016).

Our readers can also find expansions of various kernels in Wainwright (2019); Wahba (1990); Fasshauer (2012); Williams and Seeger (2000); Shi et al. (2009); Liang (2014); Fornberg and Piret (2008). There are also several existing online nonparametric learning methods not mentioned in the main text, e.g. Kivinen et al. (2001); Ying and Zhou (2006); Rudi and Rosasco (2017); Alaouie and Mahoney (2015); Xiong and Wang (2019).

**S2 Proof of Theorem 3**

We can decompose the \( L^2_{\rho_X} \)-distance (i.e. \( \| \cdot \|_2 \)-distance) between \( \hat{f}_{n,N} \) and \( f_{\rho} \) into two parts by inserting a \( f_N \) function in between. Recall the definition
of the previous two are:

\[
\hat{f}_n := \arg\min_{f \in \mathcal{F}_N} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2
\]

\[f_\rho := \arg\min_{f \in L^2_{\rho_X}} \int_{X \times \mathbb{R}} (Y - f(X))^2 \, d\rho(X,Y)\]  

(S2.1)

where \(\mathcal{F}_N\) is a subset of the \(N\)-dimension vector space spanned by \(\psi_1, ..., \psi_N\):

\[
\mathcal{F}_N = \mathcal{F}_N(M) := \{ f \in L^2_{\rho_X} \mid f \in \text{span}(\psi_1, ..., \psi_N), \|f\|_\infty < M \} \tag{S2.2}
\]

We insert a deterministic function \(f_N\) in-between to facilitate the use of the triangle inequality.

\[
f_N := \arg\min_{f \in \mathcal{F}_N} \int_{X \times \mathbb{R}} (Y - f(X))^2 \, d\rho(X,Y)\]  

(S2.3)

So we have the following decomposition of \(L^2_{\rho_X}\) distance:

\[
E \|\hat{f}_{n,N} - f_\rho\|_2 \leq E \|\hat{f}_{n,N} - f_N\|_2 + \|f_N - f_\rho\|_2 \]  

(S2.4)

If we can bound the two terms at the correct rates separately at the desired order, combining them together would give the result in Theorem 3.

\textbf{S2.1 Bound } \|f_N - f_\rho\|_2

We first handle the second term in (S2.4). It is a deterministic quantity which represents the approximation error of our estimator. In the main text, we given two equivalent definitions of RKHS, respectively based on
the reproducing property and the Mercer expansion. We will use the second one to explicitly calculate the approximation error. Let $H$ denote the native space of $K$ (the RKHS of interest).

**Lemma S2.1.** Assume (A1),(A2),(A4), we have

$$\|f_N - f_\rho\|_2 \leq (D\|f_\rho\|_H \lambda_N)^{1/2} \tag{S2.5}$$

where $\| \cdot \|_H$ is the RKHS-norm. If we further assume (A3) and choose $N = \Theta(n^{d^2/2\alpha})$, then

$$\|f_N - f_\rho\|_2 = O(n^{-\frac{\alpha}{2\alpha+1}}) \tag{S2.6}$$

**Proof.** Since $f \in H$ by assumption, we know $f_\rho$ has the following expansion w.r.t $\psi_j$: $f_\rho = \sum_{j=1}^{\infty} \theta_j \psi_j$. Recall that we defined $(\lambda_j, \psi_j)$ as the eigen-system of operator $T_{k,\bar{\rho}_X}$ in Section 2. By the definition of RKHS in Proposition 2, the condition $\|f_\rho\|_H < \infty$ in (A2) can be rewritten as:

$$\|f_\rho\|_H^2 = \sum_{j=1}^{\infty} \left( \frac{\theta_j}{\sqrt{\lambda_j}} \right)^2 < \infty \tag{S2.7}$$

Define $f_{\rho,N} = \sum_{j=1}^{N} \theta_j \psi_j \in \mathcal{F}_N$ to be a truncated approximation of $f_\rho$ (which does not depend on data). We know that $\|f_N - f_\rho\|_2$ is smaller than $\|f_{\rho,N} - f_\rho\|_2$ because $f_N$ is the minimizer of $\|f - f_\rho\|_2$ over $f \in \mathcal{F}_N$. 

So we have:

\[ \| f_N - f_\rho \|_2 \leq \| f_{\rho,N} - f_\rho \|_2 \]

\[ = \left( \int_{\mathcal{X}} (f_{\rho,N}(x) - f_\rho(x))^2 \, d\rho_X(x) \right)^{1/2} \]

\[ \overset{(1)}{\leq} D^{1/2} \left( \int_{\mathcal{X}} (f_{\rho,N}(x) - f_\rho(x))^2 \, d\bar{\rho}_X(x) \right)^{1/2} \]

\[ \overset{(2)}{=} \left( D \sum_{j=N+1}^{\infty} \theta_j^2 \right)^{1/2} \]

\[ \leq \left( D \lambda_N \sum_{j=N+1}^{\infty} \theta_j^2 \lambda_j^{-1} \right)^{1/2} \]

\[ \leq (D\| f_\rho \|_H \lambda_N)^{1/2} \]

In (1) we use assumption (A4) about the relationship between \( \rho_X \) and \( \bar{\rho}_X \).

In (2) we use Parseval’s identity noting that \( \psi_j \)'s are orthonormal w.r.t. \( \bar{\rho}_X \).

If we take \( N = \Theta(n^{\frac{1}{2\alpha+d}}) \) and assume \( \lambda_j = \Theta(j^{-2\alpha/d}) \), we have \( \lambda_N = \Theta(n^{-\frac{2\alpha}{2\alpha+d}}) \), therefore \( \| f_N - f_\rho \|_2 = O(n^{-\frac{2\alpha}{2\alpha+d}}) \). Thus we have proven the first part of the Lemma.

S2.2 Bound \( \mathbb{E}\| \hat{f}_{n,N} - f_N \|_2 \)

In this section we bound the term associated with the stochastic error. Our proof engages the following steps: We first show the hypothesis space is a VC-class, then use this property to bound its localized Rademacher complexity. This will further lead us to the final convergence rate because \( \hat{f}_{n,N} \) is an M-estimator (ERM of the negative loss) over this hypothesis.
space. We use the novel result presented in Han et al. (2019) to bound the multiplier process with a Rademacher process, which allows us to quantify the interplay between hypothesis space size and the level of noise.

**Proposition S2.2.** Let $\mathcal{F}_N$ be the $N$-dimension linear space defined in (S2.2), then we know $\mathcal{F}_N$ is VC-subgraph class with index less than or equal to $N + 2$.

**Proof.** The definition of VC-subgraph class, together with the fact that a $N$-dimension vector space $\mathcal{F}_N$ of measurable functions is a VC-class of index no more than $N + 2$, can be found in (Van Der Vaart and Wellner, 1996, Lemma 2.6.15) or (Wainwright, 2019, Proposition 4.20).

Now we use the fact that $\mathcal{F}_N$ is a VC-class to get an upper bound on its covering number. For this, we need the following result.

**Proposition S2.3.** For a VC-subgraph class of functions $\mathcal{F}$. One has for any probability measure $Q$:

$$\mathcal{N}(\epsilon \|F\|_{Q,2}, \mathcal{F}, L_Q^2) \leq CN(16e)^N \left(\frac{1}{\epsilon}\right)^{2(N-1)}$$  \tag{S2.9}$$

where $N$ is the VC-dimension of $\mathcal{F}$ and $0 < \epsilon < 1$. And $F$ is the envelope function of $\mathcal{F}$, i.e. $|f(x)| \leq F(x)$ for any $x \in \mathcal{X}, f \in \mathcal{F}$.

**Proof.** One can find the proof of a slightly more general version in (Van Der Vaart and Wellner, 1996, Theorem 2.6.7).
For a function space $\mathcal{F}$, define the localized uniform entropy integral as:

$$J(\delta, \mathcal{F}, L_2) := \int_0^\delta \sup_Q \sqrt{1 + \log \mathcal{N}(\epsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\epsilon$$  \hspace{1cm} (S2.10)

Applying this to the space $\mathcal{F}_N$, we have the following result:

**Lemma S2.4.** Let $\mathcal{F}_N$ be the function space defined in (S2.2), we have

$$J(\delta, \mathcal{F}_N, L_2) \leq C_M \sqrt{N\delta^2 \log \left(\frac{1}{\delta}\right)}$$  \hspace{1cm} (S2.11)

for sufficiently small $\delta$. The constant $C_M$ only depends on $M$.

**Proof.** We first note $\mathcal{F}_N$ is a subset of an $N$-dimension vector space with envelope $F(x) = M$. By Proposition S2.2 and Proposition S2.3, we have

$$\mathcal{N}(\epsilon M, \mathcal{F}_N, L^2(Q)) \leq C N (16\epsilon)^N \left(\frac{1}{M\epsilon}\right)^{2N-2} \text{ for any measure } Q$$

$$\Rightarrow J(\delta, \mathcal{F}, L_2) \leq C \int_0^\delta \sqrt{N \log \left(\frac{1}{M\epsilon}\right)} d\epsilon \text{ for sufficiently small } \delta$$

$$\leq C \sqrt{N} \int_0^{\frac{1}{3M}} \sqrt{\log \frac{u}{M^2u^2}} du$$

$$\leq C M \delta \sqrt{N \log \left(\frac{1}{M\delta}\right)}$$  \hspace{1cm} (S2.12)

We can see for the linear space $\mathcal{F}_N$, the localized uniform entropy is basically $O(\sqrt{N}\delta)$ (if we omit the $\sqrt{\log(1/\delta)}$ term). When we construct the
online projection estimator, the dimension of hypothesis space $N$ increases with sample size (we can also call $\mathcal{F}_N$ a sieve). As we will see later, the local diameter $\delta = \delta_n$ we consider decreases to zero at rate $\Theta(n^{-\frac{\alpha}{2\alpha+d}})$.

We use $\epsilon_i = Y_i - g_\rho(X_i), i = 1, 2, .., n$ to denote the i.i.d zero-mean noise variables and use $e_i, i = 1, 2, .., n$ to denote $n$ i.i.d. Rademacher variable, that is $\mathbb{P}(e_1 = 1) = \mathbb{P}(e_1 = -1) = \frac{1}{2}$.

In the following Proposition we require the noise to have a finite $\|\epsilon_i\|_{m,1}$-moment, which is defined as

$$\|\epsilon\|_{m,1} := \int_0^\infty \mathbb{P}(|\epsilon| > t)^{1/m} dt \quad (S2.13)$$

Let $\Delta > 0$, it is known that if $\epsilon_1$ has a finite $m + \Delta$-th moment, then it has a finite $\|\cdot\|_{m,1}$-moment ([Ledoux and Talagrand, 2013, Chapter 10]). So requiring having a finite $\|\cdot\|_{m,1}$, as assumed in (A1), is only slightly stronger than requiring a finite $m$-th moment.

Now we state and prove a proposition that connects the bounds on the multiplier/Rademacher process to the convergence rate of our M-estimator. This proposition is essentially the same as Theorem 3.4.1 in [Van Der Vaart and Wellner, 1996] and is a slight generalization of Proposition 2 in [Han et al., 2019]. In Proposition S2.5 for better presentation we drop the subscript of $\mathcal{F}_N$ and simply denote it as $\mathcal{F}$. But we should keep in mind that $\mathcal{F}$ is a function space that depends on $n$. 
Proposition S2.5. Denote $\mathcal{F} - f_\rho := \{ f - f_\rho \mid f \in \mathcal{F} \}$ and $\mathcal{F} - f_N := \{ f - f_N \mid f \in \mathcal{F} \}$. Assume $(\mathcal{F} - f_\rho) \cup (\mathcal{F} - f_N)$ has an envelope function $F(x) \leq 1$. Let $X_i \sim \rho_X$ and assume $\epsilon_i$ are i.i.d. with finite $\|\epsilon_1\|_{m,1}$-norm for some $m > 1$. Assume that for any $\delta \geq 0$, for each $f^* \in \{ f_\rho, f_N \}$,

$$\mathbb{E} \sup_{f \in \mathcal{F} : \|f - f^*\|_2 \leq \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i (f - f^*) (X_i) \right| = O \left( \phi_n(\delta) \right)$$

(S2.14)

and

$$\mathbb{E} \sup_{f \in \mathcal{F} : \|f - f^*\|_2 \leq \delta} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i (f - f^*) (X_i) \right| = O \left( \phi_n(\delta) \right)$$

(S2.15)

for some $\phi_n$ such that $\delta \mapsto \phi_n(\delta)/\delta$ is nonincreasing. Further assume that $\|f_N - f_\rho\|_2 \leq C\delta_n$.

Then

$$\left\| \hat{f}_{n,N} - f_N \right\|_2 = O_P (\delta_n)$$

(S2.16)

for any $\delta_n \geq n^{-\frac{1}{2} + \frac{1}{m}}$ such that $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$. If $\epsilon_1$ has a finite $m$-th moment for some $m \geq 2$, then:

$$\mathbb{E} \left[ \left\| \hat{f}_{n,N} - f_N \right\|_2 \right] = O (\delta_n)$$

(S2.17)

Proof. The proof is a slight generalization of Proposition 2 in Han et al. (2019). The distance we are going to bound is not between $\hat{f}_{n,N}$ and $f_\rho$ but between $\hat{f}_{n,N}$ and $f_N$ (the population risk minimizer over $\mathcal{F}$). We first
define a random process and its mean functional:

\[ M_n f := \frac{2}{n} \sum_{i=1}^{n} (f - f_\rho)(X_i) \epsilon_i - \frac{1}{n} \sum_{i=1}^{n} (f - f_\rho)^2 (X_i) \]  
(S2.18)

\[ M f := \mathbb{E}[M_n(f)] = -P (f - f_\rho)^2 \]

We have the following property of \( M(\cdot) \). For any \( f \in \{ f \in \mathcal{F} | \| f - f_N \|_2 \geq 4\| f_N - f_\rho \|_2 \} \), \( M f - M f_N \leq -\frac{1}{4} \| f - f_N \|_2^2 \). For the proof of this elementary inequality, see p.337 Exercise 5 in \textit{Van Der Vaart and Wellner (1996)}, taking their \( x = f, y = f_N, z = f_\rho \).

Our proof is a standard peeling argument. Let

\[ \mathcal{F}_j := \{ f \in \mathcal{F} : 2^{j-1}t \delta_n \leq \| f - f_N \|_2 < 2^j t \delta_n \} \]  
(S2.19)

We choose a fixed \( t \) large enough such that \( t \delta_n \geq 4\| f_N - f_\rho \|_2 \), we use the ERM property of \( \hat{f}_{n,N} \):

\[
\mathbb{P} \left( \| \hat{f}_{n,N} - f_N \|_2 \geq t \delta_n \right) \leq \sum_{j \geq 1} \mathbb{P} \left( \sup_{f \in \mathcal{F}_j} (M_n(f) - M_n(f_N)) \geq 0 \right) \\
\leq \sum_{j \geq 1} \mathbb{P} \left( \sup_{f \in \mathcal{F}_j} (M_n(f) - M_n(f_N) - M(f) + M(f_N)) \geq 2^{2j-2}t^2 \delta_n^2 \right) 
\]  
(S2.20)

We write \( (M_n(f) - M_n(f_N) - M(f) + M(f_N)) \) explicitly:

\[
M_n(f) - M_n(f_N) - M(f) + M(f_N) \\
= \frac{2}{n} \sum_{i=1}^{n} (f - f_N)(X_i) \epsilon_i + (P - \mathbb{P}_n)(f - f_\rho)^2 + (\mathbb{P}_n - P)(f_N - f_\rho)^2 
\]  
(S2.21)
Then we can continue the peeling argument:

\[
\mathbb{P}\left( \left\| \hat{f}_{n,N} - f_N \right\|_2 \geq t\delta_n \right) \\
\leq \sum_{j \geq 1} \mathbb{P}\left( \sup_{f \in \mathcal{F} : \| f - f_N \|_2 \leq 2^j t\delta_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f - f_N)(X_i) \epsilon_i \right\| \geq 2^{2j-5} t^2 \sqrt{n\delta_n^2} \right) + \\
\mathbb{P}\left( \sup_{f \in \mathcal{F} : \| f - f_N \|_2 \leq 2^j t\delta_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f - f_\rho)^2(X_i) - \mathbb{E}(f - f_\rho)^2 \right\| \geq 2^{2j-4} t^2 \sqrt{n\delta_n^2} \right) + \\
\mathbb{P}\left( \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f_N - f_\rho)^2(X_i) - \mathbb{E}(f_N - f_\rho)^2 \right\| \geq 2^{2j-4} t^2 \sqrt{n\delta_n^2} \right) \\
\leq \sum_{j \geq 1} \mathbb{P}\left( \sup_{f \in \mathcal{F} : \| f - f_N \|_2 \leq 2^j t\delta_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f - f_N)(X_i) \epsilon_i \right\| \geq 2^{2j-5} t^2 \sqrt{n\delta_n^2} \right) + \\
2\mathbb{P}\left( \sup_{f \in \mathcal{F} : \| f - f_N \|_2 \leq 2^j t\delta_n} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f - f_\rho)^2(X_i) - \mathbb{E}(f - f_\rho)^2 \right\| \geq 2^{2j-4} t^2 \sqrt{n\delta_n^2} \right)
\]

(S2.22)

The first term is the multiplier process that contains the noise variable \( \epsilon_i \)'s, for which we have bound (given by our assumptions). The second term can be related to the Rademacher process by standard symmetrization and contraction principles (Van Der Vaart and Wellner 1996). There is still a miss-match between the supremum and the random variable to be bounded, to fix this we need to use the condition \( \| f_N - f_\rho \|_2 \leq C\delta_n \):

\[
\| f - f_\rho \|_2 \leq \| f - f_N \| + \| f_N - f_\rho \|_2 \\
\leq \| f - f_N \| + C\delta_n
\]

\[
\Rightarrow \{ f \in \mathcal{F} : \| f - f_N \| \leq 2^j t\delta_n \} \subset \{ f \in \mathcal{F} : \| f - f_\rho \|_2 \leq (2^j t + C)\delta_n \}
\]

(S2.23)
Therefore the second term is bounded by

\[ 2 \mathbb{P} \left( \sup_{f \in \mathcal{F}} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (f - f_\rho)^2(X_i) - \mathbb{E}(f - f_\rho)^2 \right| \geq 2^{2j} t^2 \sqrt{n} \delta_n \right) \]

(S2.24)

And the rest of the proof is the same as Proposition 2 in \cite{Han2019}.

\[ \square \]

When \( \epsilon_i \) is sub-Gaussian noise (note that sub-Gaussian/sub-exponential random variables have finite moments of all orders), the bound on the empirical process terms (S2.14) and (S2.15) usually only depend on the entropy of \( \mathcal{F}_N \): Thus the convergence rate will only depend on the entropy as well. However if we only assume moment conditions, then \( \phi_n(\delta) \) will depend on both the entropy and the moment order \cite[Lemma 9]{Han2019}:

Thus the convergence rate would depend on both as well when \( m \) is not large enough.

Now we state the following Lemma to complete our bound of \( \mathbb{E}\|\hat{f}_{n,N} - f_N\|_2 \). Its proof is postponed to after we conclude the main result.

**Lemma S2.6.** Assume (A1) and \( \hat{f}_{n,N} \in \mathcal{F}_N \) defined in (S2.2). We select \( N = \Theta \left( \frac{n^{\frac{d}{2m+\alpha}}} {\log n} \right) \). (Recall that \( \alpha \) is the smoothness parameter, \( d \) is the dimension of \( X_i \) and \( m \) is the moment index of \( \epsilon_i \)
Then with $\delta_n = \Theta \left( n^{-\frac{\alpha}{2d+\alpha}} \lor n^{-\frac{1}{2d+1}} \right)$, for each $f^* \in \{ f_N, f_\rho \}$ we have

$$
\mathbb{E} \sup_{f \in \mathcal{F}_N: \|f-f^*\|_2 \leq \delta_n} \left| \sum_{i=1}^{n} \epsilon_i (f-f^*) (X_i) \right| \lor \mathbb{E} \sup_{f \in \mathcal{F}_N: \|f-f^*\|_2 \leq \delta_n} \left| \sum_{i=1}^{n} \epsilon_i (f-f^*) (X_i) \right| \\
\leq C_{\alpha} \begin{cases} 
\frac{n}{\alpha + 1} \sqrt{\log n} \left( 1 \lor \|\epsilon_1\|_{2\alpha+1,1} \right), & m \geq 2\alpha/d + 1 \\
\frac{1}{m} \sqrt{\log n} \left( 1 \lor \|\epsilon_1\|_{m,1} \right), & 1 \leq m < 2\alpha/d + 1
\end{cases}
$$

(S2.25)

where $\|\epsilon_1\|_{2\alpha+1}$ is the $2\alpha + 1$-th moment of $\epsilon_1$.

In light of Proposition S2.5, (S2.25) can be written as

$$
\mathbb{E} \sup_{f \in \mathcal{F}_N: \|f-f^*\|_2 \leq \delta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i (f-f^*) (X_i) \right| \lor \mathbb{E} \sup_{f \in \mathcal{F}_N: \|f-f^*\|_2 \leq \delta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i (f-f^*) (X_i) \right| \leq \phi_n(\delta_n)
$$

(S2.26)

where

$$
\phi_n(\delta) = \begin{cases} 
C_{\alpha} \sqrt{\log n / n\delta^{-1/\alpha}} \left( 1 \lor \|\epsilon_1\|_{1+2\alpha,1} \right), & m \geq 1 + 2\alpha \\
C_{\alpha} \sqrt{\log n / n\delta^{-2/(m-1)}} \left( 1 \lor \|\epsilon_1\|_{m,1} \right), & 1 \leq m < 1 + 2\alpha
\end{cases}
$$

(S2.27)

Lemma S2.7. Assume (A1) and $\hat{f}_{n,N} \in \mathcal{F}_N$. Choosing $N = \Theta(n^{\frac{d}{2d+\alpha}})$,

$$
E[\|\hat{f}_{n,N} - f_N\|_2] = O \left( n^{-\frac{\alpha}{2d+\alpha}} \sqrt{\log n} \lor n^{-\frac{1}{2d+1}} \sqrt{\log n} \right)
$$

(S2.28)

Proof. We use the result of Lemma S2.6 as conditions of Proposition S2.5, and then identify the smallest $\delta_n$ satisfying $\phi_n(\delta_n) \leq \sqrt{n}\delta_n^2$, which will give the stated convergence rate.
Proof of Theorem 3. We need only combine the bounds in Lemma S2.1 and Lemma S2.7 using the triangle inequality.

We now return to proving Lemma S2.6. We first state two results, Propositions S2.8 and S2.9, from the literature which we will use to prove our Lemma. We begin with a standard result connecting Rademacher complexity and the entropy integral.

**Proposition S2.8** (Theorem 2.1, Van Der Vaart and Wellner (2011)).

Suppose that $G$ has a finite envelope $G(x) \leq 1$ and $X_1, \ldots, X_n$'s are i.i.d. random variables with law $P$.

Then with $G(\delta) := \{g \in G : Pg^2 < \delta^2\}$,

$$
\mathbb{E} \sup_{g \in G(\delta)} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \epsilon_i g(X_i) \right| = O \left( J(\delta, G, L_2) \left( 1 + \frac{J(\delta, G, L_2)}{\sqrt{n} \delta^2 \|G\|_{P,2}} \|G\|_{P,2} \right) \right)
$$

(S2.29)

We next give a recent inequality established in Han et al. (2019). This allows us to relax common subgaussian assumptions to only moment conditions on the $\epsilon_i$'s.

**Proposition S2.9** (Theorem 1, Han et al. (2019)). Suppose $X_i$'s, $\epsilon_i$'s are all i.i.d. random variables and $X_i$'s are independent of $\epsilon_i$'s. Let $\{G_k\}_{k=1}^{n}$ be a sequence of function classes such that $G_k \supseteq G_n$ for any $1 \leq k \leq n$. Assume further that there exists a nondecreasing concave function $\psi_n : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$.
with $\psi_n(0) = 0$ such that

\[ \mathbb{E} \sup_{f \in G_k} \left| \sum_{i=1}^{k} e_i f(X_i) \right| \leq \psi_n(k) \] (S2.30)

holds for all $1 \leq k \leq n$. Then

\[ \mathbb{E} \sup_{f \in G_n} \left| \sum_{i=1}^{n} e_i f(X_i) \right| \leq 4 \int_0^\infty \psi_n \left( \sum_{i=1}^{n} \mathbb{P}(|\epsilon_i| > t) \right) \, dt \] (S2.31)

With these two results in hand, we are now ready to prove Lemma S2.6.

**Proof of Lemma S2.6.** We need to show the result for both $f^* = f_N$ and $f^* = f_\rho$. We will explicitly show the result for $f^* = f_N$: The proof in the case $f^* = f_\rho$ is exactly the same.

Denote

\[ \mathcal{F}_N(\delta_k) := \{ f \in \mathcal{F}_N \mid \|f - f_N\|_2^2 \leq \delta_k^2 \} \] (S2.32)

We first combine Proposition S2.8 with the entropy bound we established in Lemma S2.4 to derive

\[ \mathbb{E} \sup_{f \in \mathcal{F}_N(\delta_k)} \left| \sum_{i=1}^{k} e_i f(X_i) \right| \leq C\delta_k k^{\frac{d}{2 \alpha + d}} + \frac{1}{2} \sqrt{\log k} \] (S2.33)

where $\delta_k = k^{-\frac{\alpha}{2 \alpha + d}} \lor k^{-\frac{1}{2} + \frac{1}{2m}}$.

When $m \geq 2\alpha/d + 1$ (recall $m$ is the moment index for $\epsilon_i$’s), $k^{-\frac{\alpha}{2 \alpha + d}} > k^{-\frac{1}{2} + \frac{1}{2m}}$, so the above bound becomes

\[ \mathbb{E} \sup_{f \in \mathcal{F}_N(\delta_k)} \left| \sum_{i=1}^{k} e_i f(X_i) \right| \leq Ck^{\frac{d}{2 \alpha + d}} \sqrt{\log k} \] (S2.34)
Using (S2.34) we see that the conditions of Proposition S2.9 are satisfied, thus giving us

\[
E \sup_{f \in \mathcal{F}_N(\delta_k)} \left| \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \leq C \int_{0}^{\infty} \left( \sum_{i=1}^{n} P(|\epsilon_i| > t) \right)^{\frac{d}{2\alpha+3}} \sqrt{ \log \left( \sum_{i=1}^{n} P(|\epsilon_i| > t) \right) } \, dt
\]

\[
= Cn^{\frac{d}{2\alpha+3}} \sqrt{ \log n (1 \vee \|\epsilon_1\|_{2\alpha+1,1}) }
\]

(S2.35)

Note that we used \( \epsilon_i \)'s are i.i.d. random variables.

When \( 1 < m < 2\alpha/d + 1 \), (S2.33) becomes

\[
E \sup_{f \in \mathcal{F}_N(\delta_k)} \left| \sum_{i=1}^{k} \epsilon_i f(X_i) \right| \leq Ck^{\frac{1}{m}} \sqrt{\log k}.
\]

(S2.36)

Plugging this in to Proposition S2.9 we get

\[
E \sup_{f \in \mathcal{F}_N(\delta_k)} \left| \sum_{i=1}^{n} \epsilon_i f(X_i) \right| \leq Cn^{\frac{1}{m}} \sqrt{\log n (1 \vee \|\epsilon_1\|_{m,1})}
\]

(S2.37)

This completes the proof.

S3 Online Projection Estimator and Functional Stochastic Gradient Descent

The computational expense of Algorithm 2 is a dramatic improvement compared with SGD based algorithms, whose expense is \( O(n) \) per updating. We also note that the computational expense of Algorithm 2 depends on our assumption of the spectrum of operator \( T_K \). The larger \( \alpha \) is, the
stronger our statistical assumption is, the faster our algorithm is. However, the expense of SGD-based algorithm is not sensitive to the statistical assumptions.

In this section we use the same notation as in Section 3 in the main text. We define $\hat{\theta}_{N,n}$ as the minimizer of the empirical loss

$$\min_{\theta \in \mathbb{R}^N} \sum_{i=1}^{n} (Y_i - \theta^\top \psi_i^N(X_i))^2$$  (S3.1)

Here we use double subscript to emphasize that $\hat{\theta}_{N,n}$ is calculated with $N$ basis function and $n$ data. Similarly, we can define $\hat{\theta}_{N,n-1}$ as the minimizer when there is one less sample $(X_n, Y_n)$ (but keep the other samples the same). There is actually a recursive relationship between $\hat{\theta}_{N,n}$ and $\hat{\theta}_{N,n-1}$:

$$\hat{\theta}_{N,n} = \hat{\theta}_{N,n-1} + \Phi_n \psi_n \left[ Y_n - \hat{f}_{n-1,N}(X_n) \right]$$  (S3.2)

See Ljung and Söderström (1983) p.18-20 for the derivation. This formula tells us how $\hat{\theta}_{N,n}$ changes when one additional data-pair is observed. If we see $\hat{\theta}_{N,n}$ as an update of $\hat{\theta}_{N,n-1}$ with $(X_n, Y_n)$, the step size will scale in proportion to the prediction error $|Y_n - \hat{f}_{n-1,N}(X_n)|$, and the direction is $\Phi_n \psi_n$ (which, in general, is not equal to $\psi_n$)

Similarly, we can derive a recursive relationship for how $\hat{\theta}_{N,n}$ changes when
one more basis function $\psi_{N+1}$ is added in. Specifically,

$$
\hat{\theta}_{N+1,n} = \begin{bmatrix}
\hat{\theta}_{N,n} \\
0
\end{bmatrix}
+ \frac{\left(\psi_{N+1}^\top\Delta_n\right)}{\| (I - P_n) \psi_{N+1} \|^2} \begin{bmatrix}
-P_n \psi_{N+1} \\
1
\end{bmatrix}
$$

(S3.3)

Where $\Delta_n$ is the residual vector, whose i-th component is defined by:

$$
\Delta_n^{(i)} = Y_i - \hat{f}_{n,N}(X_i)
$$

(S3.4)

and $P_n = (\Psi_n^\top \Psi_n)^{-1} \Psi_n^\top$ is the projection matrix of the column space of design matrix $\Psi_n$ with $N$ features. We give the derivation in the later part of this section.

The influence of a new feature on the regression coefficients is quantitatively associated with how much the residual can be explained by the new feature (represented by the term $\left(\psi_{N+1}^\top\Delta_n\right)$) and how orthogonal the new feature is to the old features (represented by $P_n \psi_{N+1}$).

However, if we use parametric stochastic gradient descent to solve the problem (S3.1), then the updating rule should be:

$$
\hat{\theta}_{N,n} = \theta_{N,n-1} + \epsilon_n \psi_n \left[ Y_n - \hat{f}_{n-1,N}(X_n) \right]
$$

(S3.5)

where we usually choose $\epsilon_n \approx \frac{1}{n}$.

Comparing (S3.5) with (S3.2), we see that it replaces the structured matrix $\Phi_n$ with a diagonal matrix $\epsilon_n I$. By doing so it omits the information
of the correlation between features, this can help to illustrate why the SGD-based estimator (S3.5) usually has a larger generalization error than the empirical risk minimizer (S3.2).

**S3.1 Proof of recursive formula (S3.3)**

*Proof. In this proof, we use a double subscript to indicate the dimension of the matrices. By definition of OLS estimator:*

\[
\hat{\theta}_{N+1,n} = \Phi_{(N+1)\times(N+1)} \cdot \Psi_{n \times (N+1)}^T \cdot Y_n \\
= \Phi_{(N+1)\times(N+1)} \cdot \left( \sum_{i=1}^{n} Y_i \left[ \psi_1(X_i), \ldots, \psi_{N+1}(X_i) \right]^T \right) \\
= \Phi_{(N+1)\times(N+1)} \cdot \left[ \begin{array}{c} \sum_{i=1}^{n} \psi_N(X_i) Y_i \\ \sum_{i=1}^{n} \psi_{N+1}(X_i) Y_i \end{array} \right] \\
\overset{(1)}{=} \Phi_{(N+1)\times(N+1)} \cdot \left[ \begin{array}{c} \Phi_{N \times N}^{-1} \cdot \hat{\theta}_{N,n} \\ \sum_{i=1}^{n} \psi_{N+1}(X_i) Y_i \end{array} \right] \\
\overset{(2)}{=} \left( \begin{array}{cc} \Phi_{N \times N} & 0 \\ 0 & 0 \end{array} \right) + A \cdot \left( \begin{array}{c} \Phi_{N \times N}^{-1} \cdot \hat{\theta}_{N,n} \\ \sum_{i=1}^{n} \psi_{N+1}(X_i) Y_i \end{array} \right)
\]

where

\[
A = \left[ \begin{array}{cc} \frac{1}{k} \Phi_{n-1} b b^T & \Phi_{n-1} - \frac{1}{k} \Phi_{n-1} b \\ -\frac{1}{k} b^T \Phi_{n-1} & \frac{1}{k} \end{array} \right]
\]
\[ b = \Psi_{n-1}^T \psi_{N+1} \]

\[ k = \psi_{N+1}^T \psi_{N+1} - b^T \Phi_{n-1} b \]

In (1) we use the definition of \( \hat{\theta}_{N,n} \) and in (2) use the block matrix inversion formula.

\[
\hat{\theta}_{N+1,n} = \begin{bmatrix} \hat{\theta}_{N,n} \\ 0 \end{bmatrix} + \frac{1}{k} \begin{bmatrix} \Phi_{N \times N} b \left( b^T \hat{\theta}_{N,n} - \sum_{i=1}^{n} \psi_{N+1} (X_i) Y_i \right) \\ \left( \sum_{i=1}^{n} \psi_{N+1} (X_i) Y_i - b^T \hat{\theta}_{N,n} \right) \end{bmatrix}
\] (S3.6)

Note that

\[
b^T \hat{\theta}_{N,n} = \sum_{i=1}^{n} \psi_{N+1} (X_i) \sum_{j=1}^{N} \psi_j (X_i) \hat{\theta}^{(j)}_{N,n} = \sum_{i=1}^{n} \psi_{N+1} (X_i) \hat{f}_{n,N} (X_i)
\] (S3.7)

So

\[
\sum_{i=1}^{n} \psi_{N+1} (X_i) Y_i - b^T \hat{\theta}_{N,n} = \sum_{i=1}^{n} \psi_{N+1} (X_i) (Y_i - f_{n,N} (X_i))
\] (S3.8)

Continuing, we see that

\[
\hat{\theta}_{N+1,n} = \begin{bmatrix} \hat{\theta}_{N,n} \\ 0 \end{bmatrix} + \frac{\psi_{N \times N}^T \Delta_n}{k} \begin{bmatrix} -\Phi_{N \times N} b \\ 1 \end{bmatrix}
\]

Now we expand \( k \):

\[
k = \psi_{N+1}^T \psi_{N+1} - \psi_{N+1}^T \Psi_{n \times N} \Phi_{N \times N} \psi_{n \times N} \psi_{N+1} \]

\[
= \psi_{N+1}^T \left( I - \Psi_{n \times N} \left( \Psi_{n \times N}^T \Psi_{n \times N} \right)^{-1} \Psi_{n \times N} \right) \psi_{N+1}
\]

\[
= \| (I - P_n) \psi_{N+1} \|^2
\]
And use the definition of $b$:

$$
\hat{\theta}_{N+1,n} = \begin{bmatrix}
\hat{\theta}_{N,n} \\
0
\end{bmatrix} + \frac{\psi_{N+1}^\top \Delta_n}{\|(I - P_n) \psi_{N+1}\|^2} \begin{bmatrix}
\psi_{N+1}^\top \\
-\Phi_{N \times N} \\
\vdots \\
\psi_{N}^\top \\
1
\end{bmatrix}
$$

(S3.9)

S4 Regression in Additive Models

In the main text we discussed estimation in multivariate RKHS and how it suffers from the curse of dimensionality. For $X_i \in \mathbb{R}^d$, it is also quite common to impose an extra additive structure on the model, in other words, we assume

$$
f_\rho(x_i) = \sum_{k=1}^d f_{\rho,k}(x_{i}^{(k)}) \tag{S4.1}
$$

where the component functions $f_{\rho,i}$ belong to a RKHS $\mathcal{H}$ (in general they can belong to different spaces), and $x_{i}^{(k)}$ is the $k$-th entry of $x_i$. Such a model
S4. REGRESSION IN ADDITIVE MODELS

is a generalization of the multivariate linear model. It balances modeling flexibility with tractability of estimation. See e.g. Hastie et al. (2009) and Yuan and Zhou (2016) for further discussion.

The projection estimator for an additive model is obtained by solving the following least-squares problem in Euclidean space (which is essentially the same as solving the problem (S3.1)).

\[
\min_{\theta \in \mathbb{R}^{N \times d}} \sum_{i=1}^{n} (Y_i - \sum_{k=1}^{d} \sum_{j=1}^{N} \theta_{jk} \psi_j(x_i^{(k)}))^2
\]  

(S4.2)

here \( N \) still needs to be chosen of order \( n^{\frac{1}{2\alpha+1}} \), when \( \lambda_j = \Theta(j^{-2\alpha}) \). The online projection estimator in an additive model is

\[
\hat{f}_{n,N} = \sum_{k=1}^{d} \sum_{j=1}^{N} \hat{\theta}_{jk} \psi_j
\]  

(S4.3)

For a fixed \( d \), the minimax rate for estimating an additive model is identical (losing a constant \( d \)) to the minimax rate in the analogous one-dimension nonparametric regression problem working with the same hypothesis space \( \mathcal{H} \) (Raskutti et al., 2009).

The design matrix of (S4.2) now is of dimension \( n \times (Nd) \). When a new data point is collected, our design matrix grows by one row. When we need to increase the model capacity however, we need to add one feature for each dimension (in total \( d \) columns). Updating such estimators when \( X_i \in \mathbb{R}^d \) has a computational expense of order \( O(d^2 n^{\frac{2\alpha}{2\alpha+1}}) \), by a argument similar
to that presented in Section 3.4. To clarify, in Section 3.4 we are assuming
the eigenvalue $\lambda_j = \Theta(j^{-2\alpha/d})$ (for example, the RKHS is $d$-dimension,
$\alpha$-th order Sobolev space); however in this section we are discussing $d$-
dimension additive model, each component lies in a 1-dimension RKHS
whose $\lambda_j = \Theta(j^{-2\alpha})$. The additive model is more restrictive, therefore
we have better statistical and computational guarantee when the model is
well-specified.

\textbf{S4.1 Additive Model Application}

We chose a 10-dimension additive function to illustrate the efficacy of our
method for fitting additive models. In this example, the components of the
$f_\rho$ in each dimension are Doppler-like functions. For $x \in \mathbb{R}^{10}$,

$$f_\rho(x) = \sum_{k=1}^{10} f_{\rho,k}(x^{(k)}) = \sum_{k=1}^{10} \left\{ \sin \left( \frac{2\pi}{(x^{(k)} + 0.1)^{1/20}} \right) - \sin \left( \frac{2\pi}{0.1^{1/20}} \right) \right\} \quad (S4.4)$$

Similar functions are used in \cite{Sadhanala2019}. The
kernel (for each dimension) we consider is

$$K(s, t) = \sum_{m=1}^{2} s^{m} t^{m} + B_4(\{s - t\}) \quad (S4.5)$$

In Figure 1, we compare the method in this paper with the additive smoothing
spline estimator calculated with back fitting using R package 'gam'
Both of the methods achieve rate-optimal convergence, but we note the smoothing spline method takes dramatically more time as an offline estimator.

Figure 1: Additive model: generalization error and CPU time. (A) Both smoothing spline and online projection estimator achieve the optimal rate $O(n^{-4/5})$. The black line has slope $-4/5$. Each curve is based on 15 independent runs. (B) The CPU time decreases as $\alpha$ becomes larger (repetitions=10).

**S5 Details of simulation studies**

In the main text we gave important details on of the settings of our simulation studies. To help our readers replicate our result, we now list all details
S5.1 Notation and general setting

The $\|\hat{f}_{n,N} - f_\rho\|_2^2$ on the y-axis of Figure 2 is estimated with 1,000 $X$ generated from $\rho_X$. The estimator based on kernel ridge regression (KRR) is defined as the minimizer of penalized mean-square error

$$\min_{f \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} (Y_i - f(X_i))^2 + \lambda_{n,KRR} \|f\|_H^2$$

for a closed form solution and theoretical optimal selection of $\lambda_{n,KRR}$, see 12.5.2 and Theorem 13.7 of Wainwright (2019).

In the main text, we slightly simplify the update rule for nonparametric SGD estimator without losing the essential principles. In all the simulation study of this paper, the SGD estimator we use is the version with Polyak averaging (p.1375-1376 of Dieuleveut and Bach (2016))

$$\tilde{f}_n = \tilde{f}_{n-1} + \gamma_{n,SGD} \left[ Y_n - \tilde{f}_{n-1}(X_n) \right] K_{X_n}$$

$$\hat{f}_n = \frac{1}{n+1} \sum_{k=0}^{n} \tilde{f}_k$$

The nonparametric SGD estimator we use is $\hat{f}_n$. To update such an estimator, the computational cost is also $O(n)$. 
All the simulation study examples are coded in R version 3.5.1.

S5.2 One Dimension Example Settings

We give the details of example 1 (resp. example 2) in Table 1 (resp. Table 2).

Table 1: Settings of example 1. See Wahba (1990) and Dieuleveut and Bach (2016).

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>( f_\rho )</td>
<td>( B_4(x) = x^4 - 2x^3 + x^2 - \frac{1}{30} )</td>
</tr>
<tr>
<td>( \epsilon )</td>
<td>( \text{Unif}([-0.02, 0.02]) )</td>
</tr>
<tr>
<td>( p_X(x) )</td>
<td>( 1_{[0,1]}(x) )</td>
</tr>
<tr>
<td>( K(s,t) )</td>
<td>( \frac{1}{2\pi}B_4{s-t} = \sum_{j=1}^{\infty} \frac{2}{(2\pi j)^4} \left[ \cos(2\pi js) \cos(2\pi jt) + \sin(2\pi js) \sin(2\pi jt) \right] )</td>
</tr>
<tr>
<td>RKHS ( \mathcal{H} )</td>
<td>( W_2^{\text{per}} = \left{ f \in L^2([0,1])</td>
</tr>
<tr>
<td>( \lambda_j )</td>
<td>( \frac{2}{(2\pi j)^4} = O(j^{-4}) )</td>
</tr>
<tr>
<td>( \psi_j(x) )</td>
<td>( \sin(2\pi jx) ) and ( \cos(2\pi jx) )</td>
</tr>
<tr>
<td>basis adding step</td>
<td>( n = \lfloor 0.2N^5 \rfloor )</td>
</tr>
<tr>
<td>Hyperparameter KRR ( \lambda_{n,KRR} )</td>
<td>( \lambda_{n,KRR} = 10^{-3}n^{-4/5} )</td>
</tr>
<tr>
<td>Learning rate ( \gamma_{n,SGD} )</td>
<td>( \gamma_{n,SGD} = 128n^{-0.5} )</td>
</tr>
</tbody>
</table>

S5.3 Additive Model Example

We use the function `gam()` in R package `gam` (Hastie, 2019) to fit the additive model with smoothing spline. The degrees of freedom parameter
Table 2: Settings of example 2. See Wainwright (2019, Chap. 12) for more discussion on the kernel space $W^0_1$.

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Setting</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_\rho$</td>
<td>$(6x - 3) \sin(12x - 6) + \cos^2(12x - 6)$</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Normal$(0, 5)$</td>
</tr>
<tr>
<td>$p_\rho(x)$</td>
<td>$(x + 0.5)1_{[0, 1]}(x)$</td>
</tr>
<tr>
<td>$K(s, t)$</td>
<td>$\min{s, t} = \sum_{j=1}^{\infty} \frac{8}{(2j-1)^2\pi^2} \sin\left(\frac{(2j-1)s\pi}{2}\right) \sin\left(\frac{(2j-1)t\pi}{2}\right)$</td>
</tr>
<tr>
<td>RKHS $\mathcal{H}$</td>
<td>$W^0_1 = {f \in L^2([0, 1])</td>
</tr>
<tr>
<td>$\lambda_j$</td>
<td>$\frac{2}{(2j-1)^2\pi^2} = O(j^{-2})$</td>
</tr>
<tr>
<td>$\psi_j(x)$</td>
<td>$2 \sin\left(\frac{(2j-1)x\pi}{2}\right)$</td>
</tr>
<tr>
<td>Basis adding step</td>
<td>$n = \lfloor 0.5N^3 \rfloor$</td>
</tr>
<tr>
<td>Hyperparameter KRR $\lambda_{n,KRR}$</td>
<td>$\lambda_{n,KRR} = 0.1n^{-2/3}$</td>
</tr>
<tr>
<td>Learning rate $\gamma_{n,SGD}$</td>
<td>$\gamma_{n,SGD} = 5n^{-0.5}$</td>
</tr>
</tbody>
</table>
S6. A NOTE FOR APPLICATION AND ADDITIONAL EXAMPLES

used in the \( s() \) function were selected to increase with \( n \). The details for the additive model example (including parameter selection) are given in Table 3.

<table>
<thead>
<tr>
<th>( f_{\rho} )</th>
<th>( \sum_{k=1}^{10} \left{ \sin \left( \frac{2\pi}{(X^k + 0.1)^{0.1} \pi} \right) - \sin \left( \frac{2\pi}{0.1^{0.1} \pi} \right) } )</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \epsilon )</td>
<td>Normal(0,5)</td>
</tr>
<tr>
<td>( p_{\rho}(X_1, ..., X_{10}) )</td>
<td>( \prod_{k=1}^{10} 1_{[0,1]}(X^{(k)}) )</td>
</tr>
<tr>
<td>( K(s, t) ) (for each dimension)</td>
<td>( \sum_{m=1}^{2} s^m t^m + B_4({s - t}) )</td>
</tr>
<tr>
<td>RKHS ( \mathcal{H} )</td>
<td>( W_2 = \left{ f \in L^2([0, 1]) \mid \int_0^1 (f''(u))^2 du &lt; \infty \right} )</td>
</tr>
<tr>
<td>( \lambda_j )</td>
<td>( \frac{2}{(2\pi)^{j^4}} = O \left(j^{-4}\right) )</td>
</tr>
<tr>
<td>( \psi_j(x) )</td>
<td>( x, x^2, \sin(2\pi j x), \cos(2\pi j x) )</td>
</tr>
<tr>
<td>basis adding step</td>
<td>( n = \lfloor 0.2N^{5} \rfloor )</td>
</tr>
<tr>
<td>df for smoothing spline</td>
<td>( 2</td>
</tr>
</tbody>
</table>

**Table 3: Settings of Additive model example.**

S6   A Note for Application and Additional Examples

The hypothesis spaces used so far in this paper have been well-studied in previous work, and are relatively easy to engage with: Their kernel functions have a closed form, and their eigenfunctions can also be explicitly written out with respect to some special measures \( \bar{\rho} \).
However, they are usually equipped with some undesirable boundary conditions. For example, in example 2, it is more interesting to consider the space

\[ W_1 = \left\{ f \in L^2([0,1]) \mid \int_0^1 (f'(u))^2 \, du < \infty \right\} \quad (S6.1) \]

rather than the one we use in our simulation study

\[ W_0^1 = \left\{ f \in L^2([0,1]) | f(0) = 0, \int_0^1 (f'(u))^2 \, du < \infty \right\} \quad (S6.2) \]

Although it is known that \( W_1 \) is also an RKHS \cite{Wainwright2019} with kernel \( \tilde{K}(s,t) = 1 + \min\{s,t\} \), it takes extra analytical work to get the form of eigenfunctions for \( \tilde{K} \).

For practical purposes, it is enough to consider functions of the following form as estimator:

\[ \hat{f}_{n,N}(x) = \theta_0 \cdot 1 + \sum_{j=1}^{N} \theta_j \psi_j(x) \quad (S6.3) \]

where \( \psi_j = \sqrt{2} \sin\left(\frac{(2j-1)\pi x}{2}\right) \) as stated in Table 1. Because the difference between \( W_0^1 \) and \( W_1 \) is merely a constant function in the sense that

\[ W_1 = \{1\} \oplus W_0^1 \quad (S6.4) \]

When a new sample comes in, we update \( \hat{f}_{n,N} \) (and potentially add a new basis function) in an online manner as in Algorithm 2. Similarly, in example
1, the more interesting space is

\[ W_2 = \left\{ f \in L^2([0, 1]) \mid \int_0^1 (f^{(2)}(u))^2 \, du < \infty \right\} \]  

(S6.5)

Note that

\[ W_2 = \{1\} \oplus \{x\} \oplus \{x^2\} \oplus W_{2}^{\text{per}} \]  

(S6.6)

So the projection estimator can be of the form

\[ \hat{f}_{n,N}(x) = 2\sum_{k=0}^{2} \tilde{\theta}_k x^k + \sum_{j=1}^{N} \theta_j \psi_j(x) \]  

(S6.7)

where \( \psi_j \)'s are the trigonometric functions listed in Table 1.

Figure 2: Generalization error for additional examples. (A) Example A.1, black line has slope \(-2/3\) (B) Example A.2, the black line has slope \(-4/5\). Both estimators achieve the minimax rates in \( W_1 \) and \( W_2 \). Each curve is based on 15 independent repetitions.

The settings for our two additional examples are given in Table 4.
Table 4: Settings of additional examples.

<table>
<thead>
<tr>
<th>Example</th>
<th>Example A.1</th>
<th>Example A.2</th>
</tr>
</thead>
<tbody>
<tr>
<td>$f_\rho$</td>
<td>$1 + (x - 0.5)^1_{[0.5,1]}(x)$</td>
<td>$1 + (6x - 3)\sin(12x - 6) + \cos^2(12x - 6)$</td>
</tr>
<tr>
<td></td>
<td>$+2(x - 0.2)^1_{[0.2,1]}(x)$</td>
<td>$+10(x - 0.5)^2^1_{[0.5,1]}(x)$</td>
</tr>
<tr>
<td>$\epsilon$</td>
<td>Normal(0,1)</td>
<td>Unif(-5,5)</td>
</tr>
<tr>
<td>$p_\rho(x)$</td>
<td>$(x + 0.5)^1_{[0.1]}(x)$</td>
<td>$1_{[0,1]}(x)$</td>
</tr>
<tr>
<td>RKHS</td>
<td>$W_1$</td>
<td>$W_2$</td>
</tr>
<tr>
<td>basis function</td>
<td>$1, \sin\left(\frac{(2j-1)\pi x}{2}\right), j = 1, 2, \ldots$</td>
<td>$1, x, x^2, \sin(2\pi j x), \cos(2\pi j x), j = 1, 2, \ldots$</td>
</tr>
</tbody>
</table>
| basis adding step | $n = \lfloor 0.5N^3 \rfloor$                               | $n = \lfloor \frac{1}{30}N^5 \rfloor$ }
Bibliography


Rudi, A. and Rosasco, L. (2017). Generalization properties of learning with


Williams, C. and Seeger, M. (2000). The effect of the input density distri-


