Noname manuscript No.
(will be inserted by the editor)

Optimal Learning Rates for Kernel Partial Least Squares

Shao-Bo Lin · **Ding-Xuan Zhou**

Received: date / Accepted: date

Abstract We study two learning algorithms generated by kernel partial least squares (KPLS) and kernel minimal residual (KMR) methods. In these algorithms, regularization against overfitting is obtained by early stopping, which makes stopping rules crucial to their learning capabilities. We propose a stopping rule for determining the number of iterations based on cross-validation, without assuming a priori knowledge of the underlying probability measure, and show that optimal learning rates can be achieved. Our novel analysis consists of a nice bound for the number of iterations in a priori knowledge-based stopping rule for KMR and a stepping stone from KMR to KPLS. Technical tools include a recently developed integral operator approach based on a second order decomposition of inverse operators and an orthogonal polynomial argument.

Keywords Learning theory · Kernel partial least squares · Kernel minimal residual · Cross validation

Mathematics Subject Classification (2000) 68T05 · 94A20 · 41A35

1 Introduction

The *partial least squares* (PLS) method [19] is a popular and effective tool for solving many statistical and learning problems (e.g. [6]). Its idea is to make use of correlations between input and output vectors for creating orthogonal components while

The work described in this paper is partially supported by the NSFC/RGC Joint Research Scheme [RGC Project No. N_CityU120/14 and NSFC Project No. 11461161006] and by the National Natural Science Foundation of China [Grant No. 61502342].

S. B. Lin
College of Mathematics and Information Science, Wenzhou University, Wenzhou, China
E-mail: sblin1983@gmail.com

D. X. Zhou
Department of Mathematics, City University of Hong Kong, Kowloon, Hong Kong, China
E-mail: mazhou@cityu.edu.hk

keeping variances of input vectors. This is essentially different from the classical principle component regression which finds orthogonal components solely by input vectors. Kernelizing PLS [17] benefits in allowing nonlinear features of data, while avoiding to solve nonlinear optimization problems. The method of kernel partial least squares (KPLS) has been widely used in gene sequence analysis, image processing, face recognition, and many other applications [13]. We refer the readers to [11,17,2] for the existing literature on its theory, implementations, and applications.

Some error analysis of KPLS for regression has been carried out in the literature [2,3]. This paper aims at optimal learning rates of KPLS and answering an open question raised in [3]. As a kernel method [8] stated in terms of a *reproduced kernel Hilbert space* (RKHS) $(\mathscr{H}_K, \|\cdot\|_K)$ induced by a Mercer kernel $K : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ on a compact metric space $\mathscr{X}$ (input space), KPLS can be defined [2] for a sample $D = \{(x_i, y_i)\}_{i=1}^N \subset \mathscr{X} \times \mathscr{Y}$ with the output space $\mathscr{Y} \subset \mathbb{R}$ and a nonnegative integer parameter m by $f_{D,0}^{[0]} = 0$ and

$$f_{D,m}^{[0]} = \arg\min_{f \in \mathscr{K}_m(K,D)} \|\mathbf{y} - f(\mathbf{x})\|_{\ell^2}, \qquad m \geq 1, \tag{1}$$

where the ℓ^2-norm is taken in $\mathbb{R}^N$ for the difference of the vectors $\mathbf{y} := (y_1, \ldots, y_N)^T$, and $f(\mathbf{x}) := (f(x_1), \ldots, f(x_N))^T$, and the minimization is taken over the Krylov subspace

$$\mathscr{K}_m(K,D) := \operatorname{span}\left\{ f_{K,D}, L_{K,D} f_{K,D}, \ldots, (L_{K,D})^{m-1} f_{K,D} \right\} \tag{2}$$

of $\mathscr{H}_K$ generated by the data dependent initial function

$$f_{K,D} := \frac{1}{N} \sum_{i=1}^N y_i K(\cdot, x_i) \tag{3}$$

and an *empirical integral operator* $L_{K,D}$ defined by

$$L_{K,D} f = \frac{1}{N} \sum_{i=1}^N f(x_i) K(\cdot, x_i), \qquad f \in \mathscr{H}_K. \tag{4}$$

KPLS consists of a sequence $\{f_{D,m}^{[0]}\}_m$ of output functions defined by (1). A major advantage of this method is its iterative nature [2] and easy implementation (compared with kernel regularized least squares [4] and kernel principle components analysis [15]): the output functions as linear combinations of $\{K_{x_i}\}_{i=1}^N$ can be computed in terms of their coefficient vectors by only using forward multiplication of vectors by the Gramian matrix $\mathbb{K} := [K(x_i, x_j)]_{i,j=1}^N$.

A crucial component of KPLS is to determine m, the number of iterations, by some stopping rules which is equivalent to a regularization, as for some other iterative methods such as kernel gradient descent algorithms [21]. Two nice stopping rules were proposed in [2] to ensure universal consistency of the algorithm. However, there lacks concrete learning rates, which was raised as an open question in [3]. Different from previous work on early stopping of iterative algorithms [21,3,16,10] requiring priori knowledge of the regression problem, we devote in this paper our main analysis to learning rates of KPLS equipped with a posteriori selection of m based on cross-validation, and answering the open question raised in [3].

Our error analysis is based on an equivalence between KPLS and a kernel conjugate gradient (KCG) algorithm [7,2], an orthogonal polynomial approach of KCG [11], and a recently developed integral operator approach [14,10]. Our key novelty is a stepping stone to error bounds for KPLS from those for the *kernel minimal residual* (KMR) method [11, Section 2.2], a special kernel conjugate gradient method defined with the ℓ^2 norm in (1) replaced by the K-norm as $f_{D,0}^{[1]} = 0$ and

$$f_{D,m}^{[1]} = \arg\min_{f \in \mathscr{K}_m(K,D)} \|L_{K,D} f - f_{K,D}\|_K, \qquad m \geq 1, \tag{5}$$

and a nice bound for the number of iterations in a priori knowledge-based stopping rule for KMR. As a byproduct, we derive the optimal learning rate of KMR equipped with cross-validation, which improves the learning rate in [3] from an almost optimal one (with a logarithmic factor) to an optimal one.

2 Main Results

Our main results are stated in a standard learning theory framework for regression [8]. Let $D = \{z_i\}_{i=1}^N = \{(x_i, y_i)\}_{i=1}^N$ be drawn independently according to a Borel probability measure ρ on $\mathscr{Z} := \mathscr{X} \times \mathscr{Y}$. Let ρ_X be the marginal distribution of ρ on $\mathscr{X}$ and $(L^2_{\rho_X}, \|\cdot\|_\rho)$ be the Hilbert space of ρ_X square integrable functions on $\mathscr{X}$. The primary objective of our study on KPLS for regression is to investigate the convergence of the KPLS estimator, measured in the $L^2_{\rho_X}$-distance, to the regression function f_ρ defined by $f_\rho(x) = \int_{\mathscr{Y}} y d\rho(y|x)$, where $\rho(y|x)$ denotes the conditional distribution at x induced by ρ. Throughout the paper we assume for some constant $M > 0$ that $|y| \leq M$ almost surely. This implies that f_ρ is supported on $[-M, M]$. It is then natural for us to project an output function $f : \mathscr{X} \to \mathbb{R}$ onto the interval $[-M, M]$ by the projection operator π_M defined [20] by

$$\pi_M f(x) = \begin{cases} f(x), \text{ if } -M \leq f(x) \leq M, \\ M, \quad \text{if } f(x) > M, \\ -M, \text{ if } f(x) < -M. \end{cases}$$

We now give our *stopping rule* for KPLS and KMR by using a *cross-validation* method [5] to determine the number of iterations in (1) or (5). Throughout the paper we assume that the data size N is even and the data set D is the disjoint union of two data subsets, D_1 (the training set) and D_2 (the validation set), of equal cardinality $|D_1| = |D_2| = N/2$.

Definition 1 Let $v \in \{0,1\}$ and $\{f_{D_1,m}^{[v]}\}_{m=0}^{2N-1}$ be given with the training data subset D_1 by (1) for KPLS or (5) for KMR. We define the stopping rule as the stopping iteration m^* by means of the validation set D_2 by

$$m^* = \arg \min_{0 \leq m \leq 2N-1} \frac{1}{|D_2|} \sum_{z_i \in D_2} \left(\pi_M f_{D_1,m}^{[v]}(x_i) - y_i \right)^2. \tag{6}$$

The final estimator for regression is given by $\pi_M f^{[v]}_{D_1,m^*}$. Our error analysis estimates the convergence of this estimator to f_ρ under assumptions on the regularity of the target function f_ρ and complexity of the hypothesis space $\mathscr{H}_K$.

Our *regularization condition* for f_ρ is defined in terms of the integral operator L_K on $L^2_{\rho_X}$ associated with the Mercer kernel K given by

$$L_K(f) = \int_X f(x)K(\cdot,x)d\rho_X.$$

It takes the following form as in [1,4,18,12]

$$f_\rho = L_K^r(h_\rho) \ \text{ for some } r>0 \text{ and } h_\rho \in L^2_{\rho_X}, \tag{7}$$

where L_K^r denotes the r-th power of the compact positive operator L_K on $L^2_{\rho_X}$.

The *complexity* of $\mathscr{H}_K$ with respect to ρ_X is measured by the *effective dimension* $\mathscr{N}(\lambda)$ defined to be the trace of the operator $(\lambda I + L_K)^{-1}L_K$, i.e.,

$$\mathscr{N}(\lambda) = \mathrm{Tr}((\lambda I + L_K)^{-1}L_K), \qquad \lambda > 0.$$

Our complexity condition is given by quantitative increment of the effective dimension $\mathscr{N}(\lambda)$ with a parameter $0 < s \le 1$ and a constant $C_0 > 0$ as

$$\mathscr{N}(\lambda) \le C_0\lambda^{-s}, \qquad \forall \lambda > 0. \tag{8}$$

Now we can state our main results which will be proved in Section 5. Here the error estimates with the index $v=0$ are for the KPLS estimator $\pi_M f^{[0]}_{D_1,m^*}$ while those with $v=1$ are for the KMR estimator $\pi_M f^{[1]}_{D_1,m^*}$.

Theorem 1 *Assume (7) with $r \ge 1/2$ and (8) with $0 < s \le 1$. Let $v \in \{0,1\}$ and choose m^* by (6). Then for any $0<\delta<1$, with confidence at least $1-\delta$, there holds*

$$\|\pi_M f^{[v]}_{D_1,m^*} - f_\rho\|^2_\rho \le C\left(N^{-\frac{2r}{2r+s}}\log^6\frac{12}{\delta} + \frac{\log N}{N}\right), \tag{9}$$

where C is a constant depending only on $\|h_\rho\|_\rho, r, s, C_0, M$, and $\kappa := \sqrt{\sup_{x\in\mathscr{X}} K(x,x)}$. Moreover, we have

$$E\left[\|\pi_M f^{[v]}_{D_1,m^*} - f_\rho\|^2_\rho\right] = \mathscr{O}\left(N^{-\frac{2r}{2r+s}}\right). \tag{10}$$

By [4, Theorem 3], the learning rate in (10) with $v=0$ for KPLS is optimal in the minimax sense.

Consider the case $v=1$ for KMR. A learning rate which is almost optimal was presented in [3]. To be more specific, under conditions (7) and (8), it was proved in [3] that if $\tilde{m}$ is the smallest positive integer satisfying

$$\|L_{K,D}f^{[1]}_{D,m} - f_{K,D}\|_K \le c_0 M\kappa\left(\frac{4\sqrt{c_0}\kappa^{-s}}{\sqrt{|D|}}\log\frac{6}{\delta}\right)^{\frac{2r+1}{2r+s}} \tag{11}$$

for some constant $c_0 > 3/2$, then with confidence $1-\delta$, there holds

$$\| f_{D,\tilde{m}}^{[1]} - f_\rho \|_\rho^2 \le C' |D|^{-2r/(2r+s)} \log^{\frac{4r}{2r+s}}(6/\delta), \tag{12}$$

where C' is a constant independent of δ or $|D|$. Since the stopping rule (11) depends on the confidence level δ, the error bound (12) yields only an almost optimal learning rate with an additional logarithmic factor rather than the optimal learning rate for KMR. This was pointed out in [3, Section 5]. Furthermore, two open questions were raised in [3] about how to remove δ in the stopping rule (11) while achieving the optimal learning rate and how to get optimal learning rates for KPLS. In Theorem 1, we show that if a cross-validation based parameter selection strategy is employed, then both questions are answered successfully and optimal learning rates are achieved.

3 Stepping Stone from KMR to KPLS

In this section, we present a stepping stone from KMR to KPLS (Theorem 2 below) which will help us to achieve optimal learning rates for KPLS from those for KMR. It aims at bounding the norm $\| f_{D,\hat{m}}^{[0]} - f_\rho \|_\rho$ for the error of KPLS by $\| f_{D,\hat{m}}^{[1]} - f_\rho \|_\rho$ for the error of KMR. Here $\hat{m} = \hat{m}_{\rho,K,D,\lambda}$ is an early stopping rule depending on priori knowledge of the probability measure ρ and an auxiliary positive number λ which plays the role of a regularization parameter to be determined for getting explicit learning rates later.

To introduce the priori knowledge-based early stopping rule $\hat{m}$, we need three quantities defined in terms of a parameter $\lambda > 0$ as

$$\mathscr{P}_{D,\lambda} = \| (L_K + \lambda I)^{-1/2} (f_{K,D} - L_{K,D} f_\rho)) \|_K, \tag{13}$$

$$\mathscr{Q}_{D,\lambda} = \| (L_K + \lambda I)(L_{K,D} + \lambda I)^{-1} \|^{1/2}, \tag{14}$$

$$\mathscr{R}_D = \| L_{K,D} - L_K \|. \tag{15}$$

Here $L_{K,D}$ and L_K are regarded as positive compact operators on $\mathscr{H}_K$. Then under the regularity condition (7), $\hat{m}$ is selected to be the smallest nonnegative integer m satisfying

$$\| L_{K,D} f_{D,m}^{[1]} - f_{K,D} \|_K \le \lambda^{\frac{1}{2}} \Lambda_{\rho,\lambda,r}, \tag{16}$$

where $\Lambda_{\rho,\lambda,r}$ is a quantity depending on ρ, K, D, λ, r given by

$$\Lambda_{\rho,\lambda,r} = \begin{cases} \max\left\{ 3\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}, 3(4r+2)^{r+1/2} \mathscr{Q}_{D,\lambda}^{2r-1} \| h_\rho \|_\rho \lambda^r \right\}, & \text{if } 1/2 \le r \le 3/2, \\ \max\left\{ 3\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}, 3(4r+2)^{r+1/2} \| h_\rho \|_\rho \lambda^r, \right. & \\ \qquad \left. 6(r-1/2)\kappa^{2r-3} \| h_\rho \|_\rho \lambda^{1/2} \mathscr{R}_D \right\}, & \text{if } r > 3/2. \end{cases} \tag{17}$$

Now our stepping stone can be stated as follows.

Theorem 2 *Assume the regularity condition (7). If $\hat{m}$ is the smallest nonnegative integer satisfying (16), then we have*

$$\|f_{D,\hat{m}}^{[0]} - f_\rho\|_\rho \le \|f_{D,\hat{m}}^{[1]} - f_\rho\|_\rho + 20\mathscr{D}_{D,\lambda}\Lambda_{\rho,\lambda,r}. \tag{18}$$

Theorem 2 will be proved at the end of this section by an orthogonal polynomial argument described for conjugate gradient type methods in [11] and for KPLS and KMR in [3]. The existence of m satisfying (16) follows from $\lim_{m\to\infty}\|L_{K,D}f_{D,m}^{[1]} - f_{K,D}\|_K = 0$, a limit proved in [11, Chapter 3] under the condition that $f_{K,D}$ lies in the range of the operator $L_{K,D}$. This condition is verified in the following lemma which is well understood in the literature [1,15,9]. We give the proof for completeness.

Lemma 1 *Denote $\mathscr{H}_{K,\mathbf{x}} = span\{K(\cdot,x_i)\}_{i=1}^N$ for $D \in \mathscr{Z}^N$. Then $f_{K,D} \in \mathscr{H}_{K,\mathbf{x}}$. The space $\mathscr{H}_{K,\mathbf{x}}$ equals the range of $L_{K,D}$ and is spanned by all eigenfunctions of $L_{K,D}$ with positive eigenvalues. Its dimension equals the rank $d_{\mathbf{x}}$ of the Gramian matrix $\mathbb{K}$.*

Proof Consider the linear map $\mathscr{F}$ from $\mathbb{R}^N$ to $(\mathscr{H}_{K,\mathbf{x}}, \|\cdot\|_K)$ defined by $\mathscr{F}(\mathbf{c}) = \sum_{i=1}^N c_i K(\cdot,x_i)$. It is onto, so its range is $\mathscr{H}_{K,\mathbf{x}}$. A vector $\mathbf{c}$ lies in the kernel of this map if and only if $\|\sum_{i=1}^N c_i K(\cdot,x_i)\|_K^2 = \mathbf{c}^T\mathbb{K}\mathbf{c} = 0$, i.e., $\mathbf{c}$ is an eigenvector of $\mathbb{K}$ with eigenvalue 0. So the kernel of this map is exactly the same as the eigenspace of the matrix $\mathbb{K}$ with eigenvalue 0 which has dimension $N - d_{\mathbf{x}}$. It follows that the range of $\mathscr{F}$, $\mathscr{H}_{K,\mathbf{x}}$, has dimension $d_{\mathbf{x}}$.

Let $\{(\sigma_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_i$ be a set of normalized eigenpairs of $L_{K,D}$ with the eigenfunctions $\{\phi_i^{\mathbf{x}}\}_i$ forming an orthonormal basis of $\mathscr{H}_K$ and $\sigma_1^{\mathbf{x}} \ge \sigma_2^{\mathbf{x}} \ge \ldots \ge 0$. It is well known (see e.g. [9]) that the rank $d_{\mathbf{x}}$ of $\mathbb{K}$ is the same as the number of positive eigenvalues of $L_{K,D}$. That is, $\sigma_1^{\mathbf{x}} \ge \sigma_2^{\mathbf{x}} \ge \sigma_{d_{\mathbf{x}}}^{\mathbf{x}} > 0$ and $\sigma_i^{\mathbf{x}} = 0$ for $i \ge d_{\mathbf{x}} + 1$. Hence for each $i \in \{1,\ldots,d_{\mathbf{x}}\}$, $\phi_i^{\mathbf{x}} = \frac{1}{\sigma_i^{\mathbf{x}}} L_{K,D}(\phi_i^{\mathbf{x}})$ lies in the range of $L_{K,D}$ which is contained in $\mathscr{H}_{K,\mathbf{x}}$ by the definition of $L_{K,D}$. Therefore, the range of $L_{K,D}$ equals $\mathscr{H}_{K,\mathbf{x}}$ and $\{\phi_i^{\mathbf{x}}\}_{i=1}^{d_{\mathbf{x}}}$ forms an orthonormal basis.

To prove Theorem 2, we introduce another auxiliary function defined by regularizing (1) and (5) as $f_{D,0}^{[2]} = 0$ and

$$f_{D,m}^{[2]} = \arg\min_{f\in\mathscr{H}_m(K,D)} \|L_{K,D}^{1/2}(L_{K,D}f - f_{K,D})\|_K, \qquad m \ge 1. \tag{19}$$

As an element in $\mathscr{H}_m(K,D)$, each of $f_{D,m}^{[0]}, f_{D,m}^{[1]}, f_{D,m}^{[2]}$ can be expressed as

$$f_{D,m}^{[u]} = q_{m-1}^{[u]}(L_{K,D})f_{K,D}, \qquad u \in \{0,1,2\},\ m \in \mathbb{N}_0, \tag{20}$$

where $q_{-1}^{[u]} = 0$, $q_{m-1}^{[u]} = q_{m-1,D}^{[u]}$ is a polynomial of degree at most $m-1$ and $\mathbb{N}_0 = \mathbb{N}\cup\{0\}$ is the set of nonnegative integer. The polynomial $q_{m-1}^{[u]}$ depends on the sample value vector $\mathbf{y}$, which makes KPLS or KMR essentially different from kernel spectral algorithms [15].

For each $u \in \{0,1,2\}$, let

$$p_m^{[u]}(t) = 1 - tq_{m-1}^{[u]}(t), \qquad m \in \mathbb{N}_0 \tag{21}$$

be the companion polynomial of $q_{m-1}^{[u]}$. Note that the constant term of $p_m^{[u]}$ is 1. We have

$$\|L_{K,D} f_{D,m}^{[u]} - f_{K,D}\|_K^2 = \left\| \left(I - L_{K,D} q_{m-1}^{[u]}(L_{K,D}) \right) f_{K,D} \right\|_K^2 = [p_m^{[u]}, p_m^{[u]}]_{[0]}, \tag{22}$$

where the inner product $[\cdot,\cdot]_{[0]}$ is defined (with $v=0$) for polynomials ϕ and ψ by

$$[\phi,\psi]_{[v]} = \langle \phi(L_{K,D}) f_{K,D}, L_{K,D}^v \psi(L_{K,D}) f_{K,D} \rangle_K, \qquad v \in \{0,1,2\}. \tag{23}$$

The following two lemmas found in [11] describe some properties of $p_m^{[u]}$ and $q_{m-1}^{[u]}$.

Lemma 2 *Let $m \in \mathbb{N}_0$. The following identities hold*

$$(p_m^{[1]})'(0) - (p_{m+1}^{[1]})'(0) = \frac{[p_m^{[1]}, p_m^{[1]}]_{[0]} - [p_{m+1}^{[1]}, p_{m+1}^{[1]}]_{[0]}}{[p_m^{[2]}, p_m^{[2]}]_{[1]}}, \tag{24}$$

$$(p_{m+1}^{[1]})'(0) - (p_{m+1}^{[0]})'(0) = \frac{[p_{m+1}^{[1]}, p_{m+1}^{[1]}]_{[0]}}{[p_m^{[2]}, p_m^{[2]}]_{[1]}}, \tag{25}$$

$$p_m^{[2]}(t) = \frac{p_{m+1}^{[1]}(t) - p_{m+1}^{[0]}(t)}{t[(p_{m+1}^{[1]})'(0) - (p_{m+1}^{[0]})'(0)]}, \qquad \forall\, 0 < t \le \kappa^2, \tag{26}$$

where $(p_{m+1}^{[1]})'(0) \neq (p_{m+1}^{[0]})'(0)$.

The above three identities are stated in Corollary 2.6, Corollary 2.9, and Proposition 2.8 of [11].

Lemma 3 *Let $u \in \{0,1,2\}, m \in \mathbb{N}$, and $\{t_{k,m}^{[u]}\}_{k=1}^m$ be the simple zeros of $p_m^{[u]}$ in the increasing order. Then the following statements hold*

$$0 < t_{k,m}^{[u]} < t_{k,m-1}^{[u]} < t_{k+1,m}^{[u]}, \qquad \text{for } m \ge 2, \tag{27}$$

$$t_{k,m}^{[0]} < t_{k,m}^{[1]} < t_{k,m}^{[2]}, \tag{28}$$

$$q_{m-1}^{[u]}(0) = -(p_m^{[u]})'(0) = \sum_{k=1}^m (t_{k,m}^{[u]})^{-1} = \max_{0 \le t \le t_{1,m}^{[u]}} q_{m-1}^{[u]}(t), \tag{29}$$

$$q_{m-1}^{[u]}(0) \le q_m^{[u]}(0) \le (t_{1,m+1}^{[u]})^{-1} + q_{m-1}^{[u]}(0). \tag{30}$$

The first two statements above are stated in Corollary 2.7 of [11], while the last two follow from the first statement and the representation of $p_m^{[u]}$ in terms of its constant term 1 and zeros as

$$p_m^{[u]}(t) = \prod_{k=1}^m \left(1 - t/t_{k,m}^{[u]}\right), \qquad m \in \mathbb{N}. \tag{31}$$

To prove Theorem 2, we need the following proposition which will be proved in the appendix.

Proposition 1 *Assume (7). Let $\lambda > 0$ and $\hat{m}$ be the smallest nonnegative integer satisfying (16). If $\hat{m} \geq 1$, then*

$$|(p_{\hat{m}-1}^{[1]})'(0)| \leq \frac{3}{\lambda}, \tag{32}$$

and for $v \in \{1,2\}$, with $\varepsilon = \lambda/3$, we have

$$\left\| F_\varepsilon \left[p_{\hat{m}-1}^{[v]}(L_{K,D}) f_{K,D} \right] \right\|_K \leq \frac{1}{2} [p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]}^{1/2},$$

where for $\mu > 0$, F_μ denotes the orthogonal projection onto the subspace of $\mathscr{H}_K$ spanned by the eigenvectors of $L_{K,D}$ associated with eigenvalues strictly less than μ.

We are now in a position to prove the main result of this section.

Proof of Theorem 2. It is obvious that (18) holds when $\hat{m} = 0$, since $f_{D,0}^{[0]} = f_{D,0}^{[1]} = 0$. We then prove (18) when $\hat{m} \geq 1$. Observe from the identity (26) that

$$q_{\hat{m}-1}^{[0]}(t) - q_{\hat{m}-1}^{[1]}(t) = \frac{p_{\hat{m}}^{[1]}(t) - p_{\hat{m}}^{[0]}(t)}{t} = \left[(p_{\hat{m}}^{[1]})'(0) - (p_{\hat{m}}^{[0]})'(0) \right] p_{\hat{m}-1}^{[2]}(t).$$

It follows that

$$\begin{aligned} f_{D,\hat{m}}^{[0]} - f_{D,\hat{m}}^{[1]} &= \left(q_{\hat{m}-1}^{[0]}(L_{K,D}) - q_{\hat{m}-1}^{[1]}(L_{K,D}) \right) f_{K,D} \\ &= \left[(p_{\hat{m}}^{[1]})'(0) - (p_{\hat{m}}^{[0]})'(0) \right] p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D}. \end{aligned}$$

Combining this with (25) yields

$$f_{D,\hat{m}}^{[0]} - f_{D,\hat{m}}^{[1]} = \frac{[p_{\hat{m}}^{[1]}, p_{\hat{m}}^{[1]}]_{[0]}}{[p_{\hat{m}-1}^{[2]}, p_{\hat{m}-1}^{[2]}]_{[1]}} p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D}. \tag{33}$$

By the identity $\|f\|_\rho = \|L_K^{1/2} f\|_K$ for $f \in L^2_{\rho_X}$ and the relation $\|L_K^{1/2}(L_K + \lambda I)^{-1/2}\| \leq 1$, we have

$$\begin{aligned} \left\| p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D} \right\|_\rho &= \left\| L_K^{1/2} p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D} \right\|_K \\ &\leq \left\| (L_K + \lambda I)^{1/2} p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D} \right\|_K. \end{aligned}$$

Recall (see e.g. [3, Lemma A.7]) that for any positive operators U and V on a Hilbert space, there holds

$$\|U^\alpha V^\alpha\| \leq \|UV\|^\alpha = \|VU\|^\alpha, \qquad \forall \alpha \in (0,1]. \tag{34}$$

Applying this with $\alpha = 1/2$ yields

$$\begin{aligned} \left\| p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D} \right\|_\rho &\leq \mathscr{Q}_{D,\lambda} \left\| (L_{K,D}^{1/2} + \lambda^{1/2} I) p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D} \right\|_K \\ &\leq \mathscr{Q}_{D,\lambda} \left\{ [p_{\hat{m}-1}^{[2]}, p_{\hat{m}-1}^{[2]}]_{[1]}^{1/2} + \lambda^{1/2} [p_{\hat{m}-1}^{[2]}, p_{\hat{m}-1}^{[2]}]_{[0]}^{1/2} \right\}, \end{aligned} \tag{35}$$

where we have expressed $\|p^{[2]}_{\hat{m}-1}(L_{K,D})f_{K,D}\|_K$ and $\|L^{1/2}_{K,D}p^{[2]}_{\hat{m}-1}(L_{K,D})f_{K,D}\|_K$ by the definition of $[\cdot,\cdot]_{[u]}$. To continue, we bound $[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]_{[0]}$ by $[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]_{[1]}$ by applying Proposition 1 with $\nu=2$ and obtain with $\varepsilon=\lambda/3$,

$$\begin{aligned}
[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{1/2}_{[0]} &= \|p^{[2]}_{\hat{m}-1}(L_{K,D})f_{K,D}\|_K \\
&\le \|F_\varepsilon p^{[2]}_{\hat{m}-1}(L_{K,D})f_{K,D}\|_K + \|F^\perp_\varepsilon p^{[2]}_{\hat{m}-1}(L_{K,D})f_{K,D}\|_K \\
&\le \frac{1}{2}[p^{[1]}_{\hat{m}-1},p^{[1]}_{\hat{m}-1}]^{1/2}_{[0]} + \varepsilon^{-1/2}\|F^\perp_\varepsilon p^{[2]}_{\hat{m}-1}(L_{K,D})L^{1/2}_{K,D}f_{K,D}\|_K \\
&\le \frac{1}{2}[p^{[1]}_{\hat{m}-1},p^{[1]}_{\hat{m}-1}]^{1/2}_{[0]} + \varepsilon^{-1/2}[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{1/2}_{[1]},
\end{aligned}$$

where $F^\perp_\mu = I - F_\mu$. Since $f^{[1]}_{D,\hat{m}-1}$ minimizes $\|L_{K,D}f - f_{K,D}\|_K$ over $\mathscr{K}_{\hat{m}-1}(K,D)$ and $f^{[2]}_{D,\hat{m}-1}\in\mathscr{K}_{\hat{m}-1}(K,D)$, we know that

$$[p^{[1]}_{\hat{m}-1},p^{[1]}_{\hat{m}-1}]_{[0]} = \|L_{K,D}f^{[1]}_{D,\hat{m}-1} - f_{K,D}\|^2_K \le \|L_{K,D}f^{[2]}_{D,\hat{m}-1} - f_{K,D}\|^2_K = [p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]_{[0]}.$$

It follows that

$$[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{1/2}_{[0]} \le \frac{1}{2}[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]_{[0]} + \varepsilon^{-1/2}[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{1/2}_{[1]}$$

which implies

$$[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{1/2}_{[0]} \le 2\varepsilon^{-1/2}[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{1/2}_{[1]}. \tag{36}$$

This together with (35) yields

$$\left\|p^{[2]}_{\hat{m}-1}(L_{K,D})f_{K,D}\right\|_\rho \le 5\mathscr{Q}_{D,\lambda}[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{1/2}_{[1]}.$$

Putting this estimate into (33), we get

$$\left\|f^{[0]}_{D,\hat{m}} - f^{[1]}_{D,\hat{m}}\right\|_\rho \le \frac{[p^{[1]}_{\hat{m}},p^{[1]}_{\hat{m}}]_{[0]}}{[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]_{[1]}}5\mathscr{Q}_{D,\lambda}[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{1/2}_{[1]}.$$

Recall that $f^{[1]}_{D,\hat{m}}$ minimizes $\|L_{K,D}f - f_{K,D}\|_K$ over $\mathscr{K}_{\hat{m}}(K,D)$ and $f^{[2]}_{D,\hat{m}-1}\in\mathscr{K}_{\hat{m}}(K,D)$, we see again that

$$[p^{[1]}_{\hat{m}},p^{[1]}_{\hat{m}}]^{\frac{1}{2}}_{[0]} = \|L_{K,D}f^{[1]}_{D,\hat{m}} - f_{K,D}\|_K \le \|L_{K,D}f^{[2]}_{D,\hat{m}-1} - f_{K,D}\|_K = [p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{\frac{1}{2}}_{[0]}.$$

It can be further bounded by $4\lambda^{-\frac{1}{2}}[p^{[2]}_{\hat{m}-1},p^{[2]}_{\hat{m}-1}]^{\frac{1}{2}}_{[1]}$ according to (36). Hence,

$$\|f^{[0]}_{D,\hat{m}} - f^{[1]}_{D,\hat{m}}\|_\rho \le [p^{[1]}_{\hat{m}},p^{[1]}_{\hat{m}}]^{\frac{1}{2}}_{[0]}20\mathscr{Q}_{D,\lambda}\lambda^{-\frac{1}{2}}.$$

Finally, we use the choice of $\hat{m}$ by (16) and bound the norm $[p^{[1]}_{\hat{m}},p^{[1]}_{\hat{m}}]^{\frac{1}{2}}_{[0]}$ as

$$[p^{[1]}_{\hat{m}},p^{[1]}_{\hat{m}}]^{\frac{1}{2}}_{[0]} = \|L_{K,D}f^{[1]}_{D,\hat{m}} - f_{K,D}\|_K \le \lambda^{1/2}\Lambda_{\rho,\lambda,r}.$$

Then the desired estimate follows from the triangle inequality

$$\|f^{[0]}_{D,\hat{m}} - f_\rho\|_\rho \le \|f^{[1]}_{D,\hat{m}} - f_\rho\|_\rho + \|f^{[0]}_{D,\hat{m}} - f^{[1]}_{D,\hat{m}}\|_\rho.$$

The proof of Theorem 2 is complete. □

4 Learning Rates of Priori Knowledge Based Algorithms

In this section, we use our recently developed integral operator approach [14,10] to derive the following learning rates for the KPLS (with $v=0$) and KMR (with $v=1$) algorithms with the early stopping rule (16) based on priori knowledge.

Theorem 3 *Assume (7) with $r \geq 1/2$. Let $\hat{m}$ be the smallest nonnegative integer satisfying (16) with $\lambda = \kappa^2 |D|^{-1/(2r+s)}$. If (8) is satisfied for some $0 < s \leq 1$, then for $v \in \{0,1\}$ and any $0 < \delta < 1$, with confidence at least $1-\delta$, there holds*

$$\| f_{D,\hat{m}}^{[v]} - f_\rho \|_\rho \leq \hat{C} \log^3(6/\delta) |D|^{-r/(2r+s)}. \tag{37}$$

Here $\hat{C}$ is a constant depending only on κ, C_0, s, $\|h_\rho\|_\rho$, M, and r.

Theorem 3 will be proved at the end of this section: we shall first provide the learning rates for KMR, and then apply the stepping stone from KMR to KPLS established in the previous section to get the learning rates for KPLS.

The main tools of our analysis include the following upper bound for $|(p_{\hat{m}}^{[1]})'(0)|$, to be proved in the appendix, and an error decomposition technique developed in [3].

Proposition 2 *Assume the regularization condition (7) with $r \geq 1/2$. Let $\lambda > 0$ and $\hat{m}$ be the smallest nonnegative integer satisfying (16). Then we have*

$$|(p_{\hat{m}}^{[1]})'(0)| \leq 15 \lambda^{-1} \tag{38}$$

and

$$\| f_{D,\hat{m}}^{[1]} - f_\rho \|_\rho \leq 32 \mathscr{Q}_{D,\lambda} \Lambda_{\rho,\lambda,r}.$$

To prove Theorem 3, we need the following bounds for $\mathscr{Q}_{D,\lambda}$, $\mathscr{P}_{D,\lambda}$ and $\mathscr{R}_D$.

Lemma 4 *Let D be a sample drawn independently according to ρ and $0 < \delta < 1$. Then each of the following estimates holds with confidence at least $1-\delta$,*

$$\mathscr{Q}_{D,\lambda} \leq \frac{2\sqrt{2}(\kappa^2+\kappa)\mathscr{A}_{D,\lambda} \log \frac{2}{\delta}}{\sqrt{\lambda}} + \sqrt{2}, \tag{39}$$

$$\mathscr{P}_{D,\lambda} \leq 2(\kappa^2+\kappa)\mathscr{A}_{D,\lambda} \log\big(2/\delta\big), \tag{40}$$

$$\mathscr{R}_D \leq \frac{2\kappa^2}{\sqrt{|D|}} \log \frac{2}{\delta}, \tag{41}$$

where

$$\mathscr{A}_{|D|,\lambda} = \frac{1}{\sqrt{|D|}} \left\{ \frac{1}{\sqrt{|D|\lambda}} + \sqrt{\mathscr{N}(\lambda)} \right\}. \tag{42}$$

The proofs of (40), (41) and (39) can be found in [4], [21] and [10], respectively. Compared with the suboptimal learning rates in [3], the main reason why we can derive optimal learning rates for KMR is the tight norm estimate (39) for $\mathscr{Q}_{D,\lambda}$ based on a second order decomposition of operator differences presented in [14]. In fact, an upper bound for $\mathscr{Q}_{D,\lambda}$ was presented in [3, Lemma A.5] asserting that if

$$\|(L_K+\lambda I)^{-1/2}(L_{K,D}-L_K)(L_K+\lambda I)^{-1/2}\|<1-\eta \tag{43}$$

for some $0<\eta<1$, then

$$\|(\lambda I+L_K)^{1/2}(\lambda I+L_{K,D})^{-1/2}\|\le\frac{1}{\sqrt{\eta}}.$$

Since $(L_K+\lambda I)^{-1/2}(L_{K,D}-L_K)(L_K+\lambda I)^{-1/2}$ is a random variable, (43) holds only with confidence. So the restriction $0<\eta<1$ requires the sample size $|D|$ to be large enough, which imposes a logarithmic factor in the error estimates for KMR given in [3]. Our estimate (39) removes this restriction and provides a powerful tool to derive optimal learning rates for KMR.

Proof of Theorem 3. Since $r+s\ge 1/2$, we have from $\lambda=\kappa^2|D|^{-1/(2r+s)}$ and (8) that

$$\mathscr{A}_{D,\lambda}\le\frac{1}{\sqrt{|D|}}\left\{|D|^{-\frac{1}{2}}|D|^{\frac{1}{4r+2s}}\kappa^{-1}+\sqrt{C_0}|D|^{\frac{s}{4r+2s}}\kappa^{-s}\right\}\le(\kappa^{-1}+\kappa^{-s}\sqrt{C_0})|D|^{-\frac{r}{2r+s}}.$$

It follows from $r\ge 1/2$ and (39) that there exists a subset $\mathscr{Z}_\delta^N$ of $\mathscr{Z}^N$ with measure at least $1-\delta$ such that for $D\in\mathscr{Z}_\delta^N$,

$$\mathscr{Q}_{D,\lambda}\le 2\sqrt{2}(\kappa+1)(\kappa^{-1}+\kappa^{-s})|D|^{-\frac{r+1/2}{2r+s}}\log\frac{2}{\delta}+\sqrt{2}\le C_1\log\frac{2}{\delta},$$

where

$$C_1:=2\sqrt{2}(\sqrt{C_0}+1)(\kappa+1)(\kappa^{-1}+\kappa^{-s})+\sqrt{2}.$$

Furthermore, (40) and (41) tell us that there exist two subsets $(\mathscr{Z}_\delta^N)'$ and $(\mathscr{Z}_\delta^N)''$ of $\mathscr{Z}^N$ with measures at least $1-\delta$ such that

$$\begin{aligned}\mathscr{P}_{D,\lambda}&\le C_2|D|^{-\frac{r}{2r+s}}\log\frac{2}{\delta},\qquad \forall D\in(\mathscr{Z}_\delta^N)',\\ \mathscr{R}_D&\le 2\kappa^2|D|^{-1/2}\log\frac{2}{\delta},\qquad \forall D\in(\mathscr{Z}_\delta^N)'',\end{aligned}$$

where

$$C_2:=2\kappa(\kappa+1)(\kappa^{-1}+\kappa^{-s})(\sqrt{C_0}+1).$$

Putting all the above bounds for $\mathscr{Q}_{D,\lambda},\mathscr{P}_{D,\lambda},\mathscr{R}_D$ into the expression (17) for $\Lambda_{\rho,\lambda,r}$, we know that for $\frac{1}{2}\le r\le\frac{3}{2}$ and $D\in\mathscr{Z}_\delta^N\cap(\mathscr{Z}_\delta^N)'$,

$$\mathscr{Q}_{D,\lambda}\Lambda_{\rho,\lambda,r}\le 3(\mathscr{Q}_{D,\lambda}^2\mathscr{P}_{D,\lambda}+(4r+2)^{r+1/2}\mathscr{Q}_{D,\lambda}^{2r}\lambda^r\|h_\rho\|_\rho)\le C_3|D|^{-\frac{r}{2r+s}}\log^3\frac{2}{\delta},$$

while for $r > 3/2$ and $D \in \mathscr{Z}_\delta^N \cap (\mathscr{Z}_\delta^N)' \cap (\mathscr{Z}_\delta^N)''$,

$$\mathscr{Q}_{D,\lambda}\Lambda_{\rho,\lambda,r} \le 3\mathscr{Q}_{D,\lambda}^2\mathscr{P}_{D,\lambda} + 3(4r+2)^{r+1/2}\|h_\rho\|_\rho\lambda^r\mathscr{Q}_{D,\lambda} + 6(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho\lambda^{1/2}\mathscr{Q}_{D,\lambda}\mathscr{R}_D \le C_4|D|^{-\frac{r}{2r+s}}\log^3\frac{2}{\delta},$$

where

$$C_3 := 3(C_1^2C_2 + (4r+2)^{r+1/2}\kappa^{2r}\|h_\rho\|_\rho C_1^{2r}),$$
$$C_4 := 3C_1^2C_2 + 3(4r+2)^{r+1/2}C_1\kappa^{2r}\|h_\rho\|_\rho + 12(r-\frac{1}{2})\kappa^{3r-1}\|h_\rho\|_\rho.$$

For $v = 1$ corresponding to KMR, we apply Proposition 2 with the stopping rule (16) and find

$$\|f_{D,\hat{m}}^{[1]} - f_\rho\|_\rho \le 32\mathscr{Q}_{D,\lambda}\Lambda_{\rho,\lambda,r}.$$

For $v = 0$ corresponding to KPLS, we apply Theorem 2 and Proposition 2 and find

$$\|f_{D,\hat{m}}^{[0]} - f_\rho\|_\rho \le \|f_{D,\hat{m}}^{[1]} - f_\rho\|_\rho + 20\mathscr{Q}_{D,\lambda}\Lambda_{\rho,\lambda,r} \le 52\mathscr{Q}_{D,\lambda}\Lambda_{\rho,\lambda,r}.$$

But our derived bounds for $\mathscr{Q}_{D,\lambda}\Lambda_{\rho,\lambda,r}$ can be stated as

$$\mathscr{Q}_{D,\lambda}\Lambda_{\rho,\lambda,r} \le \begin{cases} C_3|D|^{-\frac{r}{2r+s}}\log^3\frac{2}{\delta}, & \forall D \in \mathscr{Z}_\delta^N \cap (\mathscr{Z}_\delta^N)', \quad \text{if } \frac{1}{2} \le r \le \frac{3}{2}, \\ C_4|D|^{-\frac{r}{2r+s}}\log^3\frac{2}{\delta}, & \forall D \in \mathscr{Z}_\delta^N \cap (\mathscr{Z}_\delta^N)' \cap (\mathscr{Z}_\delta^N)'', \text{ if } r > \frac{3}{2}. \end{cases}$$

Observe that the subset $\mathscr{Z}_\delta^N \cap (\mathscr{Z}_\delta^N)' \cap (\mathscr{Z}_\delta^N)''$ of $\mathscr{Z}^N$ has measure at least $1-3\delta$. Scaling δ to $\delta/3$ and setting $\hat{C} = 52\max\{C_3, C_4\}$, we know that for $v \in \{0,1\}$, with confidence at least $1-\delta$, there holds for $r \ge 1/2$,

$$\|f_{D,\hat{m}}^{[v]} - f_\rho\|_\rho \le \hat{C}|D|^{-\frac{r}{2r+s}}\log^3\frac{6}{\delta}.$$

This verifies the learning rates for KMR ($v = 1$) and KPLS ($v = 0$). The proof of Theorem 3 is complete. □

5 Proving Main Results

Theorem 3 gave learning rates of the KMR and KPLS algorithms with the stopping rule (16) based on priori knowledge. In this section we use these learning rates and the following lemma stated in [5, Proposition 11] to prove the optimal learning rates for KPLS and KMR equipped with cross-validation.

Lemma 5 *Let $\{\xi_i\}_{i=1}^n$ be a sequence of real valued independent random variables with mean μ, satisfying $|\xi_i| \le B$ and $E[(\xi_i-\mu)^2] \le \tau^2$ for $i \in \{1,2,\ldots,n\}$. Then for any $a > 0$ and $\varepsilon > 0$, there hold*

$$\mathbf{P}\left[\frac{1}{n}\sum_{i=1}^n \xi_i - \mu \ge a\tau^2 + \varepsilon\right] \le e^{-\frac{6na\varepsilon}{3+4aB}},$$

and

$$\mathbf{P}\left[\mu - \frac{1}{n}\sum_{i=1}^{n}\xi_i \geq a\tau^2 + \varepsilon\right] \leq e^{-\frac{6na\varepsilon}{3+4aB}}.$$

We also need the following inequalities for the zeros $\{t_{k,m}^{[1]}\}_{k=1}^{m}$ of the polynomial $p_m^{[1]}$ of degree $m \in \mathbb{N}$ and the norm κ^2 of the operator $L_{K,D}$ which can be found in [11, page 18]:

$$0 < t_{1,m}^{[1]} < \ldots < t_{m,m}^{[1]} \leq \kappa^2. \tag{44}$$

We are in a position to prove our main results.

Proof of Theorem 1. Recall that the data set D of even size N is the disjoint union of two data subsets, D_1 and D_2, of equal cardinality $|D_1| = |D_2| = N/2$. We divide our proof into three steps. The training set D_1 is used for defining the sequences $\{f_{D_1,m}^{[v]}\}_{m\in\mathbb{N}}$ by (1) for KPLS with $v = 0$ and by (5) for KMR with $v = 1$. The validation set D_2 will be used for estimating the sample error in the second step. In the first step, we set $\lambda = \kappa^2|D_1|^{-\frac{1}{2r+s}}$ in the priori knowledge-based stopping rule (16). Let $v \in \{0,1\}$.

Step 1. Bounding $\hat{m}$ by $2N-1$. It is sufficient for us to bound $\hat{m}$ by $2N-1$ when $\hat{m} \geq 1$. In this case, we apply Proposition 1 and find

$$|(p_{\hat{m}-1}^{[1]})'(0)| \leq \frac{3}{\lambda} = 3\kappa^{-2}|D_1|^{\frac{1}{2r+s}}.$$

On the other hand, $|(p_{\hat{m}-1}^{[1]})'(0)|$ can be expressed in terms of the zeros $\{t_{k,\hat{m}-1}^{[1]}\}_{k=1}^{\hat{m}-1}$ of the polynomial $p_{\hat{m}-1}^{[1]}$ as (29). It then follows from (44) that

$$|(p_{\hat{m}-1}^{[1]})'(0)| = \sum_{k=1}^{\hat{m}-1}(t_{k,\hat{m}-1}^{[1]})^{-1} \geq (\hat{m}-1)\kappa^{-2}.$$

Combining the above upper and lower bounds for $|(p_{\hat{m}-1}^{[1]})'(0)|$ yields

$$\hat{m} \leq \kappa^2|(p_{\hat{m}-1}^{[1]})'(0)| + 1 \leq 3|D_1|^{\frac{1}{2r+s}} + 1 < 3\frac{|D|}{2} + 1 \leq 2|D| = 2N.$$

So $\hat{m} \leq 2N-1$. This proves that $\hat{m}$ is bounded by $2N$.

Step 2. Bounding the sample error for deriving learning rates. Fix D_1 and $m \in \{0,1,\ldots,2N-1\}$. Write the validation set as $D_2 = \{(x_{i+N/2}, y_{i+N/2})\}_{i=1}^{N/2}$. Define a sequence $\{\xi_i\}_{i=1}^{N/2}$ of real valued independent random variables by

$$\xi_i = (\pi_M f_{D_1,m}^{[v]}(x_{i+N/2}) - y_{i+N/2})^2 - (f_\rho(x_{i+N/2}) - y_{i+N/2})^2.$$

They have the same mean

$$E[\xi_i] = \mu_m := \int_{\mathscr{Z}}(\pi_M f_{D_1,m}^{[v]}(x) - y)^2 d\rho - \int_{\mathscr{Z}}(f_\rho(x) - y)^2 d\rho = \left\|\pi_M f_{D_1,m}^{[v]} - f_\rho\right\|_\rho^2.$$

Clearly, $|\xi_i| \le 4M^2$ almost surely and

$$E[(\xi_i-\mu_m)^2] \le E[\xi_i^2] = \int_{\mathscr{Z}} (\pi_M f_{D_1,m}^{[v]}(x)-y+f_\rho(x)-y)^2 (\pi_M f_{D_1,m}^{[v]}(x)-f_\rho(x))^2 d\rho \le 16M^2\mu_m.$$

Applying Lemma 5 with $a=1/(32M^2)$, $B=4M^2$ and $\tau^2=16M^2\mu$, we know that for any $\varepsilon>0$, there exists a subset $\mathscr{Z}_{\varepsilon,m}^{N/2}$ of $\mathscr{Z}^{N/2}$ with measure at least $1-2\exp\left\{-\frac{3N\varepsilon}{128M^2}\right\}$ such that for every $D_2 \in \mathscr{Z}_{\varepsilon,m}^{N/2}$, there holds

$$\left| \frac{2}{N} \sum_{z_i \in D_2} \left\{ \left(\pi_M f_{D_1,m}^{[v]}(x_i)-y_i \right)^2 - (f_\rho(x_i)-y_i)^2 \right\} - \mu_m \right| \le \frac{\mu_m}{2} + \varepsilon. \qquad (45)$$

Now we let m run over $\{0,1,\ldots,2N-1\}$, and know that for every $D_2 \in \cap_{m=0}^{2N-1} \mathscr{Z}_{\varepsilon,m}^{N/2}$, the bound (45) holds true for every $m \in \{0,1,\ldots,2N-1\}$.

We first choose $m=m^* \in \{0,\ldots,2N-1\}$, and see from (45) that for every $D_2 \in \cap_{m=0}^{2N-1} \mathscr{Z}_{\varepsilon,m}^{N/2}$,

$$\mu_{m^*} = \|\pi_M f_{D_1,m^*}^{[v]} - f_\rho\|_\rho^2 \le \frac{4}{N} \sum_{z_i \in D_2} \left\{ \left(\pi_M f_{D_1,m^*}^{[v]}(x_i)-y_i \right)^2 - (f_\rho(x_i)-y_i)^2 \right\} + 2\varepsilon.$$

According to the stopping rule (6), m^* minimizes $\frac{4}{N}\sum_{z_i\in D_2} \left(\pi_M f_{D_1,m}^{[v]}(x_i)-y_i \right)^2$ over $m \in \{0,\ldots,2N-1\}$. In particular, since $\hat{m} \in \{0,\ldots,2N-1\}$ as proved in the first step, we have

$$\|\pi_M f_{D_1,m^*}^{[v]} - f_\rho\|_\rho^2 \le \frac{4}{N} \sum_{z_i \in D_2} \left\{ \left(\pi_M f_{D_1,\hat{m}}^{[v]}(x_i)-y_i \right)^2 - (f_\rho(x_i)-y_i)^2 \right\} + 2\varepsilon.$$

We then choose $m=\hat{m}$ in (45) and see that for every $D_2 \in \cap_{m=0}^{2N-1} \mathscr{Z}_{\varepsilon,m}^{N/2}$,

$$\frac{2}{N} \sum_{z_i \in D_2} \left\{ \left(\pi_M f_{D_1,\hat{m}}^{[v]}(x_i)-y_i \right)^2 - (f_\rho(x_i)-y_i)^2 \right\} \le \frac{3\mu_{\hat{m}}}{2} + \varepsilon.$$

Combining the above two estimates and noting $\mu_{\hat{m}} = \left\| \pi_M f_{D_1,\hat{m}}^{[v]} - f_\rho \right\|_\rho^2$, we know that for every $D_2 \in \cap_{m=0}^{2N-1} \mathscr{Z}_{\varepsilon,m}^{N/2}$,

$$\|\pi_M f_{D_1,m^*}^{[v]} - f_\rho\|_\rho^2 \le 3 \left\| \pi_M f_{D_1,\hat{m}}^{[v]} - f_\rho \right\|_\rho^2 + 4\varepsilon.$$

Since the subset $\cap_{m=0}^{2N-1} \mathscr{Z}_{\varepsilon,m}^{N/2}$ of $\mathscr{Z}^{N/2}$ has measure at least $1-4N\exp\left\{-\frac{3N\varepsilon}{128M^2}\right\}$, we take

$$\varepsilon = \frac{128M^2}{3N} \log \frac{4N}{\delta},$$

and see that with confidence at least $1-\delta$, there holds

$$\|\pi_M f_{D_1,m^*}^{[v]} - f_\rho\|_\rho^2 \le 3\|\pi_M f_{D_1,\hat{m}}^{[v]} - f_\rho\|_\rho^2 + \frac{171M^2}{N} \log \frac{4N}{\delta}.$$

This together with Theorem 3 applied to the data subset D_1 implies that with confidence at least $1-2\delta$,

$$\|\pi_M f^{[v]}_{D_1,m^*} - f_\rho\|^2_\rho \le 6\hat{C}^2 N^{-2r/(2r+s)} \log^6(6/\delta) + \frac{171M^2}{N}\log\frac{4N}{\delta}.$$

Since $\log\frac{4N}{\delta} \le \log N + \log\frac{4}{\delta}$ and $N^{-1} \le N^{-2r/(2r+s)}$, after scaling 2δ to δ, we know that with confidence at least $1-\delta$, the bound (9) holds true where the constant C is given by

$$C = 6\hat{C}^2 + 171M^2.$$

Step 3. Proving the learning rate (10) in expectation. From the confidence-based error bound (9), we know that the nonnegative random variable $\xi = \|\pi_M f^{[v]}_{D_1,m^*} - f_\rho\|^2_\rho$ satisfies

$$\text{Prob}\,[\xi > t] \le 12\exp\left\{-\left[C(N^{-\frac{2r}{2r+s}} + \frac{\log N}{N})\right]^{-1/6} t^{\frac{1}{6}}\right\}$$

for any $t > C(\log 12)^6\left(N^{-2r/(2r+s)} + \frac{\log N}{N}\right)$. Applying this bound to the formula

$$E[\xi] = \int_0^\infty \text{Prob}[\xi > t]dt$$

for nonnegative random variables, we obtain

$$\begin{aligned} E\left[\|\pi_M f^{[v]}_{D_1,m^*} - f_\rho\|^2_\rho\right] &\le C(\log 12)^6\left(N^{-\frac{2r}{2r+s}} + \frac{\log N}{N}\right) \\ &\quad + 12\int_0^\infty \exp\left\{-\left[C(N^{-\frac{2r}{2r+s}} + \frac{\log N}{N})\right]^{-1/6} t^{\frac{1}{6}}\right\} dt. \end{aligned}$$

By a change of variable, we see that the above integration equals

$$6\left[C(N^{-\frac{2r}{2r+s}} + N^{-1}\log N\right]\int_0^\infty u^5\exp\{-u\}\,du = 6!\left[C(N^{-\frac{2r}{2r+s}} + \frac{\log N}{N}\right].$$

Hence

$$E\left[\|\pi_M f^{[v]}_{D_1,m^*} - f_\rho\|^2_\rho\right] \le \left((\log 12)^6 + 12\cdot 6!\right)C\left(N^{-\frac{2r}{2r+s}} + \frac{\log N}{N}\right).$$

Since $s > 0$, for some constant $C_{r,s}$ depending only on r and s we have $N^{\frac{2r}{2r+s}-1}\log N \le C_{r,s}$ for any $N \ge 2$. It follows that

$$E\left[\|\pi_M f^{[v]}_{D_1,m^*} - f_\rho\|^2_\rho\right] \le \left((\log 12)^6 + 12\cdot 6!\right)C(1+C_{r,s})N^{-\frac{2r}{2r+s}}.$$

This proves (10). The proof of Theorem 1 is complete. □

6 Appendix

This appendix provides technical proofs of two propositions concerning the priori knowledge based learning algorithms.

Proof of Proposition 1. We start with proving the first statement (32). Since $p_0^{[1]}(t)=1$ and $(p_0^{[1]})'(t)=0$ for all $t\in[0,\kappa^2]$, (32) holds obviously for $\hat{m}=1$. It then suffices to prove (32) for $\hat{m}\geq 2$. It was presented in [11, p. 41] (see also [3, p. 16]) that

$$\|L_{K,D}f_{D,\hat{m}-1}^{[1]}-f_{K,D}\|_K\leq\|F_{t_{1,\hat{m}-1}^{[1]}}\phi_{\hat{m}-1}^{[1]}(L_{K,D})f_{K,D}\|.$$

Here $F_{t_{1,\hat{m}-1}^{[1]}}\phi_{\hat{m}-1}^{[1]}(L_{K,D})$ is the linear operator on $\mathscr{H}_K$ defined in terms of the orthonormal basis $\{\phi_j^{\mathbf{x}}\}_j$ and the orthogonal projection $F_{t_{1,\hat{m}-1}^{[1]}}$ by spectral calculus as

$$F_{t_{1,\hat{m}-1}^{[1]}}\phi_{\hat{m}-1}^{[1]}(L_{K,D})\left(\sum_j b_j\phi_j^{\mathbf{x}}\right)=\sum_{\sigma_j^{\mathbf{x}}<t_{1,\hat{m}-1}^{[1]}}\phi_{\hat{m}-1}^{[1]}(\sigma_j^{\mathbf{x}})b_j\phi_j^{\mathbf{x}},$$

where $\phi_{\hat{m}-1}^{[1]}(t)$ is the function defined on $[0,t_{1,\hat{m}-1}^{[1]})$ by

$$\phi_{\hat{m}-1}^{[1]}(t)=p_{\hat{m}-1}^{[1]}(t)\left(\frac{t_{1,\hat{m}-1}^{[1]}}{t_{1,\hat{m}-1}^{[1]}-t}\right)^{1/2},\qquad 0\leq t<t_{1,\hat{m}-1}^{[1]}.$$

Then we decompose $f_{K,D}$ as $f_{K,D}-L_{K,D}f_\rho+L_{K,D}f_\rho$ and bound the norm as

$$\begin{aligned}\|L_{K,D}f_{D,\hat{m}-1}^{[1]}-f_{K,D}\|_K&\leq\|F_{t_{1,\hat{m}-1}^{[1]}}\phi_{\hat{m}-1}^{[1]}(L_{K,D})(f_{K,D}-L_{K,D}f_\rho)\|_K\\&\quad+\|F_{t_{1,\hat{m}-1}^{[1]}}\phi_{\hat{m}-1}^{[1]}(L_{K,D})L_{K,D}f_\rho\|_K=:I+II.\end{aligned}\tag{46}$$

We continue our estimates by bounding the first term I. Applying (34) with $\alpha=1/2$ gives

$$\begin{aligned}I&=\|F_{t_{1,\hat{m}-1}^{[1]}}\phi_{\hat{m}-1}^{[1]}(L_{K,D})(L_{K,D}+\lambda I)^{1/2}(L_{K,D}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}\\&\quad(L_K+\lambda I)^{-1/2}(f_{K,D}-L_{K,D}f_\rho)\|_K\\&\leq\|F_{t_{1,\hat{m}-1}^{[1]}}\phi_{\hat{m}-1}^{[1]}(L_{K,D})(L_{K,D}+\lambda I)^{1/2}\|\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}\\&\leq\left(\sup_{t\in[0,t_{1,\hat{m}-1}^{[1]})}t^{1/2}|\phi_{\hat{m}-1}^{[1]}(t)|+\lambda^{1/2}\sup_{t\in[0,t_{1,\hat{m}-1}^{[1]})}|\phi_{\hat{m}-1}^{[1]}(t)|\right)\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}.\end{aligned}$$

Furthermore, the representation (31) for $p_{\hat{m}-1}^{[1]}$ and the definition of $\phi_{\hat{m}-1}^{[1]}$ yield

$$|\phi_{\hat{m}-1}^{[1]}(t)|=\left|(1-t/t_{1,\hat{m}-1}^{[1]})^{1/2}\prod_{k=2}^{\hat{m}-1}(1-t/t_{k,\hat{m}-1}^{[1]})\right|\leq 1,\qquad\forall\,0\leq t<t_{1,\hat{m}-1}^{[1]}.$$

It was shown in [11, Equation (3.10)] that for an arbitrary $v > 0$,

$$\sup_{t\in[0,t^{[1]}_{1,\hat{m}-1})} t^v(\phi^{[1]}_{\hat{m}-1}(t))^2 \le v^v|(p^{[1]}_{\hat{m}-1})'(0)|^{-v}. \tag{47}$$

Combining the above three bounds yields an estimate for the first term of (46) as

$$I \le (|(p^{[1]}_{\hat{m}-1})'(0)|^{-1/2} + \lambda^{1/2})\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}. \tag{48}$$

We now turn to the second term II of (46). By the regularity condition (7) for $f_\rho = L_K^r h_\rho$ and the identity $\|L_K^{1/2}h_\rho\|_K = \|h_\rho\|_\rho$, we find

$$II \le \|F_{t^{[1]}_{1,\hat{m}-1}}\phi^{[1]}_{\hat{m}-1}(L_{K,D})L_{K,D}L_K^{r-1/2}\|\|h_\rho\|_\rho \le \widetilde{II}\|h_\rho\|_\rho, \tag{49}$$

where for simplicity we denote the norm as

$$\widetilde{II} := \|F_{t^{[1]}_{1,\hat{m}-1}}\phi^{[1]}_{\hat{m}-1}(L_{K,D})L_{K,D}(L_K+\lambda I)^{r-1/2}\|.$$

When $1/2 \le r \le 3/2$, we express $(L_K+\lambda I)^{r-1/2}$ as $(L_{K,D}+\lambda I)^{r-1/2}(L_{K,D}+\lambda I)^{1/2-r}(L_K+\lambda I)^{r-1/2}$ and apply (34) with $\alpha = r-1/2$ to get

$$\widetilde{II} \le \|F_{t^{[1]}_{1,\hat{m}-1}}\phi^{[1]}_{\hat{m}-1}(L_{K,D})L_{K,D}(L_{K,D}+\lambda I)^{r-1/2}\|\mathscr{Q}_{D,\lambda}^{2r-1}. \tag{50}$$

When $r > 3/2$, we decompose the operator $(L_K+\lambda I)^{r-1/2}$ in $\widetilde{II}$ as

$$(L_{K,D}+\lambda I)^{r-1/2} + \left\{(L_K+\lambda I)^{r-1/2} - (L_{K,D}+\lambda I)^{r-1/2}\right\}.$$

The bounds $\|L_{K,D}\| \le \kappa^2$, $\|L_K\| \le \kappa^2$, and the Lipschitz property of the function $x \mapsto x^{r-1/2}$ imply

$$\|L_{K,D}^{r-1/2} - L_K^{r-1/2}\| \le (r-1/2)\kappa^{2r-3}\|L_{K,D}-L_K\|. \tag{51}$$

Hence

$$\begin{aligned}\widetilde{II} \le\ & \|F_{t^{[1]}_{1,\hat{m}-1}}\phi^{[1]}_{\hat{m}-1}(L_{K,D})L_{K,D}(L_{K,D}+\lambda I)^{r-1/2}\| \\ & + \|F_{t^{[1]}_{1,\hat{m}-1}}\phi^{[1]}_{\hat{m}-1}(L_{K,D})L_{K,D}\|(r-1/2)\kappa^{2r-3}\mathscr{R}_D.\end{aligned}$$

Combining this with (50) and the following norm estimate with $\gamma,\beta \ge 0$,

$$\begin{aligned}\|F_{t^{[1]}_{1,\hat{m}-1}}\phi^{[1]}_{\hat{m}-1}(L_{K,D})L_{K,D}^{\gamma}(L_{K,D}+\lambda I)^{\beta}\| &= \sup_{t\in[0,t^{[1]}_{1,\hat{m}-1})}\left\{t^\gamma(t+\lambda)^\beta\left|\phi^{[1]}_{\hat{m}-1}(t)\right|\right\} \\ &\le 2^\beta\max\left\{(2\gamma+2\beta)^{\gamma+\beta}|(p^{[1]}_{\hat{m}-1})'(0)|^{-(\gamma+\beta)}, \lambda^\beta(2\gamma)^\gamma|(p^{[1]}_{\hat{m}-1})'(0)|^{-\gamma}\right\}\end{aligned}$$

derived from spectral calculus and the inequality (47), we have with the notation $\mathscr{I} = \lambda |(p_{\hat{m}-1}^{[1]})'(0)|$,

$$\widetilde{II} \le \begin{cases} \left(2^{r-\frac{1}{2}}(2r+1)^{r+\frac{1}{2}}\mathscr{I}^{-(r+\frac{1}{2})} + 2^{r+\frac{1}{2}}\mathscr{I}^{-1}\right)\lambda^{r+\frac{1}{2}}\mathscr{Q}_{D,\lambda}^{2r-1}, \text{ when } \frac{1}{2} \le r \le \frac{3}{2}, \\ 2^{r-\frac{1}{2}}(2r+1)^{r+\frac{1}{2}}\mathscr{I}^{-(r+\frac{1}{2})}\lambda^{r+\frac{1}{2}} \\ +\mathscr{I}^{-1}\left(2^{r+\frac{1}{2}}\lambda^{r+\frac{1}{2}} + 2(r-1/2)\kappa^{2r-3}\lambda\mathscr{R}_D\right), \qquad \text{when } r > \frac{3}{2}. \end{cases}$$

This together with (49), the bound (48) for I, (46), and the definition (16) of the quantity $\Lambda_{\rho,\lambda,r}$ tells us that

$$\|L_{K,D} f_{D,\hat{m}-1}^{[1]} - f_{K,D}\|_K \le \left(\frac{1}{3}\mathscr{I}^{-\frac{1}{2}} + \frac{1}{3} + \frac{1}{6}\mathscr{I}^{-(r+\frac{1}{2})} + \frac{1}{3}\mathscr{I}^{-1}\right)\lambda^{\frac{1}{2}}\Lambda_{\rho,\lambda,r}. \quad (52)$$

On the other hand, $\hat{m} \ge 2$ is the smallest nonnegative integer satisfying (16), so for the smaller integer $\hat{m}-1$, we must have

$$\|L_{K,D} f_{D,\hat{m}-1}^{[1]} - f_{K,D}\|_K > \lambda^{\frac{1}{2}}\Lambda_{\rho,\lambda,r}.$$

This together with (52) implies

$$\lambda^{\frac{1}{2}}\Lambda_{\rho,\lambda,r} \le \left(\frac{1}{3}\mathscr{I}^{-\frac{1}{2}} + \frac{1}{3} + \frac{1}{6}\mathscr{I}^{-(r+\frac{1}{2})} + \frac{1}{3}\mathscr{I}^{-1}\right)\lambda^{\frac{1}{2}}\Lambda_{\rho,\lambda,r}$$

and thereby

$$\mathscr{I}^{-\frac{1}{2}} + \frac{1}{2}\mathscr{I}^{-(r+\frac{1}{2})} + \mathscr{I}^{-1} \ge 2.$$

One of the above terms in the summation is at least $\frac{2}{3}$. It follows that $\mathscr{I} \le \frac{9}{4} < 3$. It follows that $|(p_{\hat{m}-1}^{[1]})'(0)| = \frac{\mathscr{I}}{\lambda} < \frac{3}{\lambda}$. This proves the first statement (32).

To prove the second statement, we first claim that for $v \in \{1,2\}$,

$$\|F_\varepsilon[p_{\hat{m}-1}^{[v]}(L_{K,D})f_{K,D}]\|_K \le \|F_\varepsilon[f_{K,D}]\|_K. \quad (53)$$

This claim is obviously true for $\hat{m} = 1$ with equality valid since in this case $p_{\hat{m}-1}^{[v]} \equiv 1$ and $p_{\hat{m}-1}^{[v]}(L_{K,D})$ is the identity operator.

Consider the case $\hat{m} \ge 2$. Since $\varepsilon = \lambda/3$, we have from (32), (29) and (28) that

$$\varepsilon = \frac{\lambda}{3} \le |(p_{\hat{m}-1}^{[1]})'(0)|^{-1} = \left[\sum_{k=1}^{\hat{m}-1}(t_{k,\hat{m}-1}^{[1]})^{-1}\right]^{-1} \le t_{1,\hat{m}-1}^{[1]} < t_{1,\hat{m}-1}^{[2]}. \quad (54)$$

It follows from (31) that for $v \in \{1,2\}$, there holds

$$\max_{0\le t\le\varepsilon} p_{\hat{m}-1}^{[v]}(t) \le \max_{0\le t\le t_{1,\hat{m}-1}^{[v]}} p_{\hat{m}-1}^{[v]}(t) = \max_{0\le t\le t_{1,\hat{m}-1}^{[v]}} \Pi_{k=1}^{\hat{m}-1}\left(1 - t/t_{k,\hat{m}-1}^{[v]}\right) \le 1.$$

Recall the eigenpairs $\{(\sigma_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_i$ of $L_{K,D}$. Expressing $f_{K,D} = \sum_j c_j \phi_j^{\mathbf{x}}$ implies

$$\begin{aligned}\|F_\varepsilon[p_{\hat{m}-1}^{[v]}(L_{K,D})f_{K,D}]\|_K &= \left\| F_\varepsilon \left[\sum_j p_{\hat{m}-1}^{[v]}(\sigma_j^{\mathbf{x}}) c_j \phi_j^{\mathbf{x}} \right] \right\|_K \\ &= \left\| \sum_{j:\sigma_j^{\mathbf{x}}<\varepsilon} p_{\hat{m}-1}^{[v]}(\sigma_j^{\mathbf{x}}) c_j \phi_j^{\mathbf{x}} \right\|_K = \sqrt{\sum_{j:\sigma_j^{\mathbf{x}}<\varepsilon} [p_{\hat{m}-1}^{[v]}(\sigma_j^{\mathbf{x}}) c_j]^2} \\ &\le \sqrt{\sum_{j:\sigma_j^{\mathbf{x}}<\varepsilon} c_j^2} = \|F_\varepsilon[f_{K,D}]\|_K.\end{aligned}$$

So the claim (53) is also true in the case $\hat{m} \ge 2$. This proves the claim.

To prove the second statement of the proposition, we estimate the norm $\|F_\varepsilon[f_{K,D}]\|_K$. Under the condition (7),

$$\begin{aligned}&\|F_\varepsilon[f_{K,D}]\|_K \le \|F_\varepsilon[f_{K,D} - L_{K,D} f_\rho]\|_K + \|F_\varepsilon[L_{K,D} f_\rho]\|_K \\ &\le \|F_\varepsilon[(L_{K,D}+\lambda I)^{1/2}]\| \|(L_{K,D}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}\| \\ &\times \|(L_K+\lambda I)^{-1/2}(f_{K,D} - L_{K,D} f_\rho)\|_K + \|F_\varepsilon L_{K,D} L_K^{r-1/2}\| \|L_K^{1/2} h_\rho\|_K \\ &\le (\varepsilon+\lambda)^{1/2} \mathscr{Q}_{D,\lambda} \mathscr{P}_{D,\lambda} + \|F_\varepsilon L_{K,D} L_K^{r-1/2}\| \|h_\rho\|_\rho, \end{aligned} \tag{55}$$

where the operators $F_\varepsilon(L_{K,D}+\lambda I)^{1/2}$ and $F_\varepsilon L_{K,D} L_K^{r-1/2}$ are defined by spectral calculus.

When $1/2 \le r \le 3/2$, we have

$$\|F_\varepsilon L_{K,D} L_K^{r-1/2}\| \le \|F_\varepsilon L_{K,D}(L_{K,D}+\lambda I)^{r-1/2}\| \mathscr{Q}_{D,\lambda}^{2r-1} \le \varepsilon(\lambda+\varepsilon)^{r-1/2} \mathscr{Q}_{D,\lambda}^{2r-1}.$$

When $r > 3/2$, it follows from (51) that

$$\begin{aligned}\|F_\varepsilon L_{K,D} L_K^{r-1/2}\| &\le \|F_\varepsilon L_{K,D} L_{K,D}^{r-1/2}\| + \|F_\varepsilon L_{K,D}(L_K^{r-1/2} - L_{K,D}^{r-1/2})\| \\ &\le \varepsilon^{r+1/2} + (r-1/2)\kappa^{2r-3}\varepsilon \mathscr{R}_D.\end{aligned}$$

Combining the above bounds for $\|F_\varepsilon L_{K,D} L_K^{r-1/2}\|$ with (55) and noticing the choice $\varepsilon = \lambda/3$ and the definition (17) of the quantity $\Lambda_{\rho,\lambda,r}$, we find

$$\|F_\varepsilon[p_{\hat{m}-1}^{[v]}(L_{K,D})f_{K,D}]\|_K \le \frac{1}{2}\lambda^{1/2}\Lambda_{\rho,\lambda,r}.$$

But $\hat{m}$ is the smallest nonnegative integer satisfying (16), the integer $\hat{m}-1$ does not satisfy (16). Hence (22) implies

$$\lambda^{1/2}\Lambda_{\rho,\lambda,r} \le \|L_{K,D} f_{D,\hat{m}-1}^{[1]} - f_{K,D}\|_K = [p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]}^{1/2}.$$

Then the desired statement of the proposition is verified. The proof of Proposition 1 is completed. □

Proof of Proposition 2. We first prove (38). Since $|(p_0^{[1]})'(0)|=0$, (38) obviously holds for $\hat{m}=0$. We then consider the case $\hat{m}\geq 1$. By (29),

$$|(p_{\hat{m}}^{[1]})'(0)| = -(p_{\hat{m}}^{[1]})'(0) = (p_{\hat{m}-1}^{[1]})'(0) - (p_{\hat{m}}^{[1]})'(0) + |(p_{\hat{m}-1}^{[1]})'(0)|. \tag{56}$$

From (24), we have

$$(p_{\hat{m}-1}^{[1]})'(0) - (p_{\hat{m}}^{[1]})'(0) = \frac{[p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]} - [p_{\hat{m}}^{[1]}, p_{\hat{m}}^{[1]}]_{[0]}}{[p_{\hat{m}-1}^{[2]}, p_{\hat{m}-1}^{[2]}]_{[1]}}.$$

Therefore,

$$(p_{\hat{m}-1}^{[1]})'(0) - (p_{\hat{m}}^{[1]})'(0) \leq \frac{[p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]}}{[p_{\hat{m}-1}^{[2]}, p_{\hat{m}-1}^{[2]}]_{[1]}}. \tag{57}$$

Then, it follows from (54), (22) and (5) that

$$\begin{aligned}
[p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_0^{1/2} &= \|p_{\hat{m}-1}^{[1]}(L_{K,D}) f_{K,D}\|_K = \|L_{K,D} f_{D,\hat{m}-1}^{[1]} - f_{K,D}\|_K \\
&\leq \|L_{K,D} f_{D,\hat{m}-1}^{[2]} - f_{K,D}\|_K = \|p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D}\|_K \\
&\leq \|F_\varepsilon[p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D}]\|_K + \|F_\varepsilon^\perp[p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D}]\|_K.
\end{aligned}$$

But Proposition 1 with $\nu=2$ gives

$$\|F_\varepsilon[p_{\hat{m}-1}^{[2]}(L_{K,D}) f_{K,D}]\|_K \leq \frac{1}{2}[p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]}^{1/2}.$$

Hence

$$\begin{aligned}
[p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]}^{1/2} &\leq \frac{1}{2}[p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]}^{1/2} + \varepsilon^{-1/2}\|p_{\hat{m}-1}^{[2]}(L_{K,D}) L_{K,D}^{1/2} f_{K,D}\|_K \\
&= \frac{1}{2}[p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]}^{1/2} + \varepsilon^{-1/2}[p_{\hat{m}-1}^{[2]}, p_{\hat{m}-1}^{[2]}]_{[1]}^{1/2}.
\end{aligned}$$

Therefore,

$$[p_{\hat{m}-1}^{[1]}, p_{\hat{m}-1}^{[1]}]_{[0]}^{1/2} \leq 2\varepsilon^{-1/2}[p_{\hat{m}-1}^{[2]}, p_{\hat{m}-1}^{[2]}]_{[1]}^{1/2},$$

which together with (56), (57) and Proposition 1 yields

$$|(p_{\hat{m}}^{[1]})'(0)| \leq 3\lambda^{-1} + 12\lambda^{-1} = 15\lambda^{-1}.$$

This proves the first statement (38) of Proposition 2.

To prove the second statement, we denote $\varepsilon_0 = \lambda/15$ and

$$f_{D,\hat{m}}^{[1]*} = \begin{cases} q_{\hat{m}-1}^{[1]}(L_{K,D}) L_{K,D} f_\rho, & \text{if } \hat{m}\geq 1, \\ 0, & \text{if } \hat{m}=0. \end{cases} \tag{58}$$

We can decompose $\|f_{D,\hat{m}}^{[1]} - f_\rho\|_\rho$ as

$$\begin{aligned}
&\|f_{D,\hat{m}}^{[1]} - f_\rho\|_\rho = \|L_K^{1/2}(f_{D,\hat{m}}^{[1]} - f_\rho)\|_K \le \|(L_K+\lambda I)^{1/2}(f_{D,\hat{m}}^{[1]} - f_\rho)\|_K \\
&\le \mathscr{Q}_{D,\lambda}\|F_{\varepsilon_0}[(L_{K,D}+\lambda I)^{1/2}(f_{D,\hat{m}}^{[1]} - f_{D,\hat{m}}^{[1]*})]\|_K + \mathscr{Q}_{D,\lambda}\|F_{\varepsilon_0}[(L_{K,D}+\lambda I)^{1/2}(f_{D,\hat{m}}^{[1]*} - f_\rho)]\|_K \\
&+ \mathscr{Q}_{D,\lambda}\|F_{\varepsilon_0}^{\perp}[(L_{K,D}+\lambda I)^{1/2}(f_{D,\hat{m}}^{[1]} - f_\rho)]\|_K \\
&=: \mathscr{Q}_{D,\lambda}(A_1+A_2+A_3). \qquad (59)
\end{aligned}$$

Due to (29) and (38), we have

$$\varepsilon_0 = \lambda/15 \le |(p_{\hat{m}}^{[1]})'(0)|^{-1} \le \left[\sum_{k=1}^{\hat{m}}(t_{k,\hat{m}})^{-1}\right]^{-1} \le t_{1,\hat{m}}^{[1]}. \qquad (60)$$

Note that $A_1 = 0$ and $f_{D,\hat{m}}^{[1]*} - f_\rho = -f_\rho = -p_{\hat{m}}^{[1]}(L_{K,D})f_\rho$ when $\hat{m} = 0$. If $\hat{m} \ge 1$, we use (20), (58), (34), the definitions of $\mathscr{P}_{D,\lambda}$ and $\mathscr{Q}_{D,\lambda}$ to bound A_1 as

$$\begin{aligned}
A_1 &= \|F_{\varepsilon_0}[(L_{K,D}+\lambda I)^{1/2}q_{\hat{m}-1}^{[1]}(L_{K,D})(f_{K,D} - L_{K,D}f_\rho)]\|_K \\
&\le \|F_{\varepsilon_0}[(L_{K,D}+\lambda I)^{1/2}q_{\hat{m}-1}^{[1]}(L_{K,D})(L_K+\lambda I)^{1/2}]\|\|(L_K+\lambda I)^{-1/2}(f_{K,D} - L_{K,D}f_\rho)\|_K \\
&\le \mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}\max_{0\le t<\varepsilon_0}|(t+\lambda)q_{\hat{m}-1}^{[1]}(t)|.
\end{aligned}$$

By (31), for $0 \le t < \varepsilon_0 \le t_{1,\hat{m}}^{[1]}$, we have

$$|tq_{\hat{m}-1}^{[1]}(t)| = |1 - p_{\hat{m}}^{[1]}(t)| \le 1.$$

Furthermore, (31), (29), (60) and (38) imply

$$\max_{0\le t<\varepsilon_0}|q_{\hat{m}-1}^{[1]}(t)| \le q_{\hat{m}-1}^{[1]}(0) = |(p_{\hat{m}}^{[1]})'(0)| \le 15\lambda^{-1}.$$

Therefore, the first term in (59) can be bounded as

$$A_1 \le 16\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}. \qquad (61)$$

We then bound the second term A_2 in two cases involving r.

When $1/2 \le r \le 3/2$, we have $r - 1/2 \le 1$, and the bound $\sup_{0\le t<\varepsilon_0\le t_{1,\hat{m}}^{[1]}}|p_{\hat{m}}^{[1]}(t)| \le 1$ together with (34) and the regularization condition (7) yields

$$\begin{aligned}
A_2 &\le \|F_{\varepsilon_0}(L_{K,D}+\lambda I)^{1/2}p_{\hat{m}}^{[1]}(L_{K,D})L_K^{r-1/2}\|\|h_\rho\|_\rho \\
&\le \mathscr{Q}_{D,\lambda}^{2r-1}\|F_{\varepsilon_0}(L_{K,D}+\lambda I)^{1/2}(L_{K,D}+\lambda I)^{r-1/2}\|\|h_\rho\|_\rho \\
&\le \mathscr{Q}_{D,\lambda}^{2r-1}\|h_\rho\|_\rho(\varepsilon_0+\lambda)^r. \qquad (62)
\end{aligned}$$

When $r > 3/2$, we use (21), (58) and the regularization condition (7) to get

$$\begin{aligned}
A_2 &\le \|F_{\varepsilon_0}(L_{K,D}+\lambda I)^{1/2}p_{\hat{m}}^{[1]}(L_{K,D})(L_{K,D}^{r-1/2} - L_K^{r-1/2})\|\|h_\rho\|_\rho \\
&+ \|F_{\varepsilon_0}(L_{K,D}+\lambda I)^{1/2}p_{\hat{m}}^{[1]}(L_{K,D})L_{K,D}^{r-1/2}\|\|h_\rho\|_\rho.
\end{aligned}$$

Since $|p^{[1]}_{\hat{m}}(t)| \le 1$ for all $0 \le t < \varepsilon_0 \le t^{[1]}_{1,\hat{m}}$, we get

$$\|F_{\varepsilon_0}(L_{K,D}+\lambda I)^{1/2} p^{[1]}_{\hat{m}}(L_{K,D}) L_{K,D}^{r-1/2}\| \le \varepsilon_0^{r-1/2}(\varepsilon_0+\lambda)^{1/2}.$$

Combining these with (51) and the definition of $\mathscr{R}_D$ yields

$$A_2 \le (\varepsilon_0^{r-1/2} + (r-1/2)\kappa^{2r-3}\mathscr{R}_D)(\varepsilon_0+\lambda)^{1/2}\|h_\rho\|_\rho. \tag{63}$$

Finally, we turn to bound A_3. From Lemma 1 and $f_\rho \in \mathscr{H}_K$, we obtain that f_ρ is in the range of $L_{K,D}$. Since $(a+b)^{1/2} \le a^{1/2}+b^{1/2}$ for $a,b>0$, we then have

$$\begin{aligned}
A_3 &\le \|F^{\perp}_{\varepsilon_0}[L^{1/2}_{K,D}(f^{[1]}_{D,\hat{m}} - f_\rho)]\|_K + \lambda^{1/2}\|F^{\perp}_{\varepsilon_0}(f^{[1]}_{D,\hat{m}} - f_\rho)\|_K \\
&\le \left(\frac{(\varepsilon_0+\lambda)^{1/2}}{\varepsilon_0{}^{1/2}} + \lambda^{1/2}\frac{(\varepsilon_0+\lambda)^{1/2}}{\varepsilon_0}\right)\|F^{\perp}_{\varepsilon_0}(L_{K,D}+\lambda I)^{-1/2}L_{K,D}(f^{[1]}_{D,\hat{m}} - f_\rho)\|_K \\
&\le \left(1+\frac{\lambda^{1/2}}{\varepsilon_0{}^{1/2}}\right)\left(1+\frac{\lambda}{\varepsilon_0}\right)^{1/2}\|F^{\perp}_{\varepsilon_0}(L_{K,D}+\lambda I)^{-1/2}(L_{K,D}f^{[1]}_{D,\hat{m}} - f_{K,D})\|_K \\
&+ \left(1+\frac{\lambda^{1/2}}{\varepsilon_0{}^{1/2}}\right)\left(1+\frac{\lambda}{\varepsilon_0}\right)^{1/2}\|F^{\perp}_{\varepsilon_0}(L_{K,D}+\lambda I)^{-1/2}(f_{K,D} - L_{K,D(x)}f_\rho)\|_K \\
&\le \left(1+\frac{\lambda^{1/2}}{\varepsilon_0{}^{1/2}}\right)\varepsilon_0{}^{-1/2}\|L_{K,D}f^{[1]}_{D,\hat{m}} - f_{K,D}\|_K + \sqrt{2}\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}\left(1+\frac{\lambda}{\varepsilon_0}\right).
\end{aligned}$$

But $\hat{m}$ satisfies (16). It follows that

$$A_3 \le \left(1+\frac{\lambda^{1/2}}{\varepsilon_0{}^{1/2}}\right)\varepsilon_0{}^{-1/2}\lambda^{1/2}\Lambda_{\rho,\lambda,r} + \sqrt{2}\mathscr{Q}_{D,\lambda}\mathscr{P}_{D,\lambda}\left(1+\frac{\lambda}{\varepsilon_0}\right). \tag{64}$$

Inserting (61), (62), (63) and (64) into (59) and noticing $\varepsilon_0 = \lambda/15$, we obtain

$$\|f^{[1]}_{D,\hat{m}} - f_\rho\|_\rho \le 32\mathscr{Q}_{D,\lambda}\Lambda_{\rho,\lambda,r}.$$

This verifies the second statement of the proposition. The proof of Proposition 2 is complete. □

References

1. F. Bauer, S. Pereverzev, and L. Rosasco, On regularization algorithms in learning theory, J. Complex. 23, 52-72 (2007).
2. G. Blanchard, N. Krämer, Kernel partial least square is universally consistent. Proceeding of the 13th International Conference on Artificial Intelligence and Statistics, JMLR Workshop & Conference Proceedings, 9, 57-64 (2010).
3. G. Blanchard, N. Krämer, Optimal learning rates for kernel conjugate gradient regression, NIPs, 226-234 (2010).
4. A. Caponnetto, E. DeVito, Optimal rates for the regularized least squares algorithm, Found. Comput. Math., 7, 331-368 (2007).
5. A. Caponnetto, Y. Yao, Cross-validation based adaptation for regularization operators in learning theory, Anal. Appl., 8, 161-183 (2010).

6. H. Chun, S. Keles, Sparse partial least squares for simultaneous dimension reduction and variable selection, J. Royal Statis. Society: Series B, 72, 3-25 (2010).
7. H. Engle, M. Hanke, A. Neubauer, Regularization of inverse problems, Amsterdam: Kluwer Academic, (2000).
8. T. Evgeniou, M. Pontil, T. Poggio, Regularization networks and support vector machines, Adv. Comput. Math., 13, 1-50 (2000).
9. X. Guo, D. X. Zhou, An empirical feature-based learning algorithm producing sparse approximations, Appl. Comput. Harmonic Anal. 32, 389–400 (2012).
10. Z. C. Guo, S. B. Lin, D. X. Zhou, Optimal learning rates for spectral algorithm, manuscript, (2016).
11. M. Hanke, Conjugate Gradient Type Methods for Ill-posed Problems, Pitman Research Notes in Mathematics Series, 327 (1995).
12. T. Hu, J. Fan, Q. Wu, and D. X. Zhou, Regularization schemes for minimum error entropy principle, Anal. Appl., 13, 437-455 (2015).
13. S. Li, C. Liao, J. Kwok, Gene feature extraction using T-test statistics and kernel partial least squares, Neural Information Processing, Springer Berlin Heidelberg, 11-20 (2006)
14. S. B. Lin, X. Guo, D. X. Zhou, Distributed learning with regularization schemes, manuscript, (2015).
15. L. Lo Gerfo, L. Rosasco, F. Odone, E. De Vito, A. Verri, Spectral algorithms for supervised learning, Neural Comput., 20, 1873-1897 (2008).
16. G. Raskutti, M. Wainwright, B. Yu, Early stopping and non-parametric regression: an optimal data-dependent stopping rule, J. Mach. Learn. Res., 15, 335-366 (2014).
17. R. Rosipal, L. Trejo, Kernel partial least squares regression in reproducing kernel Hilbert spaces, J. Mach. Learn. Res., 2, 97-123 (2001).
18. S. Smale and D.X. Zhou, Learning theory estimates via integral operators and their approximations, Constructive Approx., 26, 153-172 (2007).
19. H. Wold, Path models with latent variables: the NIPALS approach. In et al., Editor, Quantitative Sociology: International Perspectives on Mathematical and Statistical Model Building, pages 307-357. Academic Press, (1975).
20. Q. Wu, Y. M. Ying, D. X. Zhou, Learning rates of least square regularized regression. Found. Comput. Math., 6, 171-192 (2006).
21. Y. Yao, L. Rosasco, A. Caponnetto, On early stopping in gradient descent learning, Constr. Approx., 26, 289-315 (2007).