Journal of Machine Learning Research 20 (2019) 1-36 Submitted 2/18; Revised 12/18; Published 2/19

Boosted Kernel Ridge Regression: Optimal Learning Rates and Early Stopping

Shao-Bo Lin SBLIN1983@GMAIL.COM
Department of Mathematics
Wenzhou University
Wenzhou, China

Yunwen Lei LEIYW@SUSTC.EDU.CN
Department of Computer Science and Engineering
Southern University of Science and Technology
Shenzhen, China

Ding-Xuan Zhou MAZHOU@CITYU.EDU.HK
School of Data Science and Department of Mathematics
City University of Hong Kong
Kowloon, Hong Kong, China

Editor: Arthur Gretton

Abstract

In this paper, we introduce a learning algorithm, boosted kernel ridge regression (BKRR), that combines L_2-Boosting with the kernel ridge regression (KRR). We analyze the learning performance of this algorithm in the framework of learning theory. We show that BKRR provides a new bias-variance trade-off via tuning the number of boosting iterations, which is different from KRR via adjusting the regularization parameter. A (semi-)exponential bias-variance trade-off is derived for BKRR, exhibiting a stable relationship between the generalization error and the number of iterations. Furthermore, an adaptive stopping rule is proposed, with which BKRR achieves the optimal learning rate without saturation.

Keywords: learning theory, kernel ridge regression, boosting, integral operator

1. Introduction

Supervised learning aims at learning function relationships between input and output variables, based on input-output pair samples. Kernel ridge regression (KRR) is a classical and standard approach for supervised learning due to its easy implementation and theoretical optimality (Evgeniou et al., 2000; Caponnetto and De Vito, 2007; Steinwart et al., 2009; Lin et al., 2017) and thus has triggered enormous research activities in the statistical and machine learning communities (Bauer et al., 2007; Caponnetto and De Vito, 2007; Cucker and Zhou, 2007; Smale and Zhou, 2007; Steinwart et al., 2009; Lin et al., 2017). However, KRR suffers from a so-called saturation phenomenon (Gerfo et al., 2008) meaning that its learning rate cannot be improved once the target (regression) function goes beyond a certain level of regularity. Furthermore, there lacks efficient parameter-selection strategies for KRR to realize its theoretically optimal learning performance. Many users' spirit is dampened by these drawbacks and then turns to other learning algorithms such as the kernel-based

gradient descent (Yao et al., 2007), kernel-based conjugate gradient descent (Blanchard and Krämer, 2016) and kernel-based partial least squares (Lin and Zhou, 2018b).

Boosting, originally proposed by Schapire (1990); Freund (1995), is devoted to producing a strong composite learner from a given class of weak learners. Some boosting algorithms can be interpreted from a viewpoint of statistical gradient descent to solve optimization problems with different loss functions (Friedman et al., 2000; Friedman, 2001). In this way, a special boosting algorithm, L_2-Boosting, was interpreted as an iterative least squares fitting of residuals (Friedman, 2001; Bühlmann and Yu, 2003). L_2-Boosting was utilized by Park et al. (2009) to improve the learning performance of Nadaya-Watson kernel estimators by overcoming saturation and was also proved in Bühlmann and Yu (2003) to be almost over-fitting resistant by exhibiting its exponential bias-variance trade-off for linear regression, reducing the difficulty of model selection.

The aim of this paper is to combine L_2-Boosting with KRR to overcome the saturation and reduce the difficulty of model selection of KRR. Let $(\mathcal{H}_K, \|\cdot\|_K)$ be the reproducing kernel Hilbert space (RKHS) induced by a Mercer kernel K on a metric (input) space $\mathcal{X}$ and $D = \{(x_i, y_i)\}_{i=1}^N \subset \mathcal{X} \times \mathcal{Y}$ be a sample with $\mathcal{Y} \subseteq \mathbb{R}$ the output space. KRR is defined by

$$f_{D,\lambda}^{(1)} := \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{(x,y)\in D} (f(x) - y)^2 + \lambda \|f\|_K^2 \right\}, \tag{1}$$

where $\lambda > 0$ is a regularization parameter and $|D| = N$ is the cardinality of D. Implementing KRR needs the inverse of the matrix $\mathbb{K} + \lambda |D| \mathbb{I}$ and thus requires $\mathcal{O}(|D|^3)$ computational complexity in time for a fixed λ, where $\mathbb{K}$ is the kernel matrix $(K(x_i, x_j))_{i,j=1}^{|D|}$ and $\mathbb{I}$ is the $|D| \times |D|$ identity matrix. The performance of KRR is sensitive to regularization parameters, which need be carefully tuned to achieve satisfactory learning rates close to the optimal one.

Boosted KRR (BKRR) studied in this paper iteratively defines an estimator $f_{D,\lambda}^{(k)}$ by running KRR on the data set $\{(x_i, y_i - f_{D,\lambda}^{(k-1)}(x_i))\}_{(x_i,y_i)\in D}$, whose outputs are residuals of the previous iteration $f_{D,\lambda}^{(k-1)}$. To be detailed, the estimator at the k-th boosting iteration is given by

$$f_{D,\lambda}^{(k)} := f_{D,\lambda}^{(k-1)} + f_{D,\lambda}^{(k)\diamond}, \qquad k > 1, \tag{2}$$

where

$$f_{D,\lambda}^{(k)\diamond} := \arg\min_{f \in \mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{(x,y)\in D} \left\{ f(x) - [y - f_{D,\lambda}^{(k-1)}(x)] \right\}^2 + \lambda \|f\|_K^2 \right\}, \qquad k > 1. \tag{3}$$

An advantage of BKRR over KRR is the flexibility of choosing some relatively large regularization parameter λ. Satisfactory learning rates are achieved by the boosting iterations. As the inputs $\{x_i\}_{i=1}^N$ of KRR and $f_{D,\lambda}^{(k)\diamond}$ are the same and the inverse of $\mathbb{K} + \lambda |D| \mathbb{I}$ has already been derived in the first step, it only requires $\mathcal{O}(|D|^2)$ computational complexity to solve (3). Then, BKRR with k iterations only needs $\mathcal{O}(|D|^3 + k|D|^2)$ computational complexity, which does not bring additional computational burden over KRR.

As pointed out by Friedman (2001), boosting leads to over-fitting if the weak learners are already over-fitting. So the regularization parameter of (1) must be relatively large,

implying under-fitting of the original KRR. BKRR then tunes k to reduce the bias, which enlarges the variance, reflecting the bias-variance trade-off. Our first main result is to deduce a (semi-)exponential bias-variance trade-off of BKRR: the bias of BKRR decreases exponentially with respect to k, while the variance increases by an exponentially diminishing amount as k gets large for the in-sample estimate and by an algebraic diminishing amount with respect to k for the out-sample estimate. The (semi-)exponential bias-variance trade-off shows that BKRR can reach its optimal learning performance with a relatively small number of iterations. It also exhibits that moderately large k does not degrade the learning performance of BKRR very much, making the model selection much easier than that of other iteration-based learning algorithms such as kernel-based gradient descent, kernel-based conjugate gradient descent and kernel-based partial least squares for which only polynomial bias-variance trade-off is obtained (Yao et al., 2007; Blanchard and Krämer, 2016; Lin and Zhou, 2018b).

The exponential bias-variance trade-off does not mean that over-fitting never happens for BKRR, and it requires a stopping rule of high quality. Our second main result is to propose an adaptive stopping rule based on an empirical effective dimension (Lu et al., 2018; Mücke, 2018), with which we prove that BKRR achieves the optimal learning rate without saturation. In a nutshell, our analysis shows that BKRR reduces the difficulty of model selection of KRR in terms of providing a stable relationship between the learning performance and model selection. Furthermore, BKRR with an adaptive stopping rule can improve the learning performance of KRR via overcoming the saturation. The main tools of our analysis are detailed spectral analysis of BKRR, the recently developed integral operator approach (Lin et al., 2017; Guo et al., 2017) and a tight bound for the number of iterations of the stopping rule.

2. Main Results

Our analysis is conducted in a standard learning theory framework for regression (Cucker and Zhou, 2007), in which the samples in D are independently drawn according to ρ, a Borel probability measure on $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$. The purpose of regression is to derive an estimator based on D to approximate the regression function $f_\rho(x) := \int_{\mathcal{Y}} y d\rho(y|x)$ with $\rho(\cdot|x)$ being the conditional distribution of ρ induced at $x \in \mathcal{X}$. Let ρ_X be the marginal distribution of ρ on $\mathcal{X}$ and $L^2_{\rho_X}$ be the space of ρ_X square integrable functions endowed with norm $\|\cdot\|_\rho$. Throughout this paper, we assume $\mathcal{X}$ is compact, which implies $\kappa := \sqrt{\sup_{x\in\mathcal{X}} K(x,x)} < \infty$.

2.1 Optimal Learning Rates

Before presenting the exponential bias-variance trade-off and stopping rule, we derive optimal learning rates for BKRR with a priori knowledge involving λ and k to show the necessity of our studies. For this purpose, some assumptions on the decay of the outputs, regularity of the regression function and capacity of $\mathcal{H}_K$ are needed. To be detailed, we assume $\int_{\mathcal{Y}} y^2 d\rho < \infty$ and the following output decay condition

$$\int_{\mathcal{Y}} \left(e^{\frac{|y-f_\rho(x)|}{M}} - \frac{|y-f_\rho(x)|}{M} - 1 \right) d\rho(y|x) \leq \frac{\gamma^2}{2M^2}, \qquad \forall x \in \mathcal{X}, \tag{4}$$

where M and γ are positive constants. Condition (4) is satisfied if the noise is uniformly bounded, Gaussian or sub-Gaussian (Caponnetto and De Vito, 2007). In particular, if $|y| \leq \mathcal{B}$ almost surely for some $\mathcal{B} > 0$, then (4) holds with $\gamma/2 = M = \mathcal{B}$.

Let $L_K : \mathcal{H}_K \to \mathcal{H}_K$ (or $L^2_{\rho_X} \to L^2_{\rho_X}$) be the integral operator defined by

$$L_K f = \int_{\mathcal{X}} K_x f(x) d\rho_X(x)$$

with $K_x := K(\cdot, x)$. We assume that f_ρ satisfies the standard regularity condition

$$f_\rho = L_K^r h_\rho, \quad \text{for some } r > 0 \text{ and } h_\rho \in L^2_{\rho_X}, \tag{5}$$

where L_K^r is the r-th power of $L_K : L^2_{\rho_X} \to L^2_{\rho_X}$. The regularity condition (5) describes the regularity of f_ρ and has been adopted in a large literature to quantify learning rates for some algorithms (Smale and Zhou, 2007; Bauer et al., 2007; Caponnetto and De Vito, 2007; Caponnetto and Yao, 2010; Shi et al., 2011; Blanchard and Krämer, 2016; Guo et al., 2017; Lin et al., 2017; Lin and Zhou, 2018b; Ying and Zhou, 2017) .

We also introduce the effective dimension $\mathcal{N}(\lambda) := \mathrm{Tr}[(L_K + \lambda I)^{-1} L_K]$ to measure the capacity of $\mathcal{H}_K$. Here $\mathrm{Tr}(A)$ denotes the trace of an operator A with a detailed definition to be given in Appendix B. In particular, we assume with a parameter $0 < s \leq 1$ and a constant $C_0 > 0$ that

$$\mathcal{N}(\lambda) \leq C_0 \lambda^{-s}, \qquad \forall \lambda > 0. \tag{6}$$

Condition (6) with $s = 1$ is always satisfied by taking $C_0 = \mathrm{Tr}(L_K) \leq \kappa^2$. For $0 < s < 1$, it was shown in Guo et al. (2017, Page 7) that (6) is slightly more general than the eigenvalue decaying assumption in the literature (Caponnetto and De Vito, 2007) and has been extensively employed to derive fast learning rates for some algorithms (Caponnetto and De Vito, 2007; Blanchard and Krämer, 2016; Guo et al., 2017; Lin et al., 2017; Lin and Zhou, 2018a,b). With these assumptions, we present in the following theorem the optimal learning rates for BKRR.

Theorem 1 *Let $0 < \delta < 1$, $k \leq \sqrt{|D|}$ and $\lambda = \left(k^2/|D|\right)^{1/(2\min\{k,r\}+s)}$ with $r > 3/2$ and $0 < s \leq 1$. Under assumptions (4), (5) and (6), with confidence $1 - \delta$ there holds*

$$\left\| f_{D,\lambda}^{(k)} - f_\rho \right\|_\rho \leq \tilde{C} (k^2/|D|)^{\frac{\min\{k,r\}}{2\min\{k,r\}+s}} \log^3 \frac{8}{\delta}, \tag{7}$$

where $\tilde{C}$ is a constant independent of $|D|$, k or δ.

For an iteration number $k \geq r$ independent of $|D|$, Theorem 1 presents an optimal learning rate for BKRR since (7) achieves the minimax lower bound established by Caponnetto and De Vito (2007). However, as shown in (7), too large k may worsen the learning rate. An extreme case is $k = \sqrt{|D|}$ which leads to a constant upper bound. In a nutshell, small k (smaller than r) suffers from the saturation but too large k leads to slow learning rates. Thus, it is important to derive an adaptive stopping rule on selecting k. Furthermore, Theorem 1 implies that as long as $k \geq r$, the boosting iteration with large k does not have any benefits in the learning process, which contradicts the boosting theory developed in Friedman (2001); Bühlmann and Yu (2003) at first glance. Thus, it requires a more

delicate analysis to explore the power of iterations, especially when $k \geq r$. Based on this observation, we conduct a detailed bias-variance analysis in the following subsection and find a so-called exponential bias-variance trade-off of BKRR.

Remark 2 *In Theorem 1 as well as Theorem 4 below, we require $r > \frac{3}{2}$. We believe that similar optimal learning rates can be derived for $r \geq 1/2$ by using the technique in Caponnetto and Yao (2010); Guo et al. (2017). Since one of the main advantages of BKRR is to conquer the saturation, we focus on relatively large r in this paper. Throughout this paper, we assume $r \geq 1/2$, implying $f_\rho \in \mathcal{H}_K$. For $0 < r < \frac{1}{2}$, i.e. $f_\rho \notin \mathcal{H}_K$, like KRR (Caponnetto and Yao, 2010; Chang et al., 2017), BKRR usually requires additional unlabeled data to achieve the optimal learning rates.*

2.2 Exponential Bias-variance Trade-off

One of the most important advantages of L_2 boosting in linear regression is its almost over-fitting resistance (Bühlmann and Yu, 2003) meaning that for the in-sample error estimate, the bias decreases exponentially fast and the variance increases with exponentially diminishing terms as k increases. In this subsection, we will show that BKRR also possesses this property and show the power of boosting iteration for $k \geq r$.

Let $S_D : \mathcal{H}_K \to \mathbb{R}^{|D|}$ be the sampling operator (Smale and Zhou, 2004) defined by

$$S_D f := (f(x))_{(x,y)\in D}.$$

Its scaled adjoint $S_D^T : \mathbb{R}^{|D|} \to \mathcal{H}_K$ (or $\mathbb{R}^{|D|} \to L^2_{\rho_X}$) is given by

$$S_D^T \mathbf{c} := \frac{1}{|D|} \sum_{i=1}^{|D|} c_i K_{x_i}, \qquad \mathbf{c} := (c_1, c_2, \ldots, c_{|D|})^T \in \mathbb{R}^{|D|}.$$

Define a discretization of the integral operator L_K by

$$L_{K,D} f := S_D^T S_D f = \frac{1}{|D|} \sum_{(x,y)\in D} f(x) K_x.$$

Our bias-variance trade-off will be stated in terms of some quantities involving the difference between the compact and positive operators L_K and $L_{K,D}$ given by

$$\mathcal{Q}_{D,\lambda} := \|(L_{K,D} + \lambda I)^{-1/2}(L_K + \lambda I)^{1/2}\|, \qquad \mathcal{R}_D := \|L_{K,D} - L_K\|_{HS} \tag{8}$$

and the difference between $S_D^T y_D$ and $L_{K,D} f_\rho$ given by

$$\mathcal{P}_{D,\lambda} := \left\|(L_K + \lambda I)^{-1/2}(L_{K,D} f_\rho - S_D^T y_D)\right\|_K, \tag{9}$$

where $\|A\|_{HS}$ denotes the Hilbert-Schmidt norm of a Hilbert-Schmidt operator A, $\|\cdot\|$ denotes the operator norm, $y_D := (y_1, \ldots, y_{|D|})^T$ and I is the identity operator. We refer the readers to Appendix B for some basic definitions of linear operators. Let $\{(\sigma_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}$ be a set of normalized eigenpairs of $L_{K,D}$ with the eigenfunctions $\{\phi_i^{\mathbf{x}}\}_i$ forming an orthonormal basis of $\mathcal{H}_K$ and the eigenvalue sequence $\{\sigma_i^{\mathbf{x}}\}$ non-increasing. Since $L_{K,D}$ is a positive

operator of rank at most $|D|$, we have $\sigma_k^{\mathbf{x}} = 0$ for $k \geq |D|+1$. Denote by $\sigma_{min}^{\mathbf{x}}$ the minimum positive eigenvalue of $L_{K,D}$. Denote by $\|f\|_D^2 := \frac{1}{|D|}\sum_{i=1}^{|D|} f^2(x_i)$. The following theorem shows a trade-off between the bias and variance of BKRR under the $\|\cdot\|_D$ semi-norm.

Theorem 3 *Let $k \geq 1$. Then, under condition (5) with $r \geq 1/2$, there holds*

$$\left\|f_{D,\lambda}^{(k)} - f_\rho\right\|_D \leq \frac{(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho}{\sqrt{k}}\lambda^{1/2}\mathcal{R}_D + \mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda} \tag{10}$$
$$+ \quad \mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}\sum_{j=1}^{k-1}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{j-1/2} + 4(\mathcal{Q}_{D,\lambda}^2+1)\lambda^k\|h_\rho\|_\rho \begin{cases} (\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k}, & \text{if } k > r, \\ \kappa^{2r-2k}+\lambda^{r-k}, & \text{if } k \leq r. \end{cases}$$

The dominant terms on the right-hand side of (10) are the third and fourth terms and we call them the variance and bias for BKRR, respectively. For $0 < \lambda \leq 1$, it follows from Theorem 3 that the bias of BKRR decreases exponentially fast while the variance increases with exponentially diminishing terms as k increases, showing the exponential bias-variance trade-off. Bound (10) also exhibits a sudden change of the rate of bias decay when k is around the regularity level r of f_ρ. Its rate drops from λ^k to $\lambda^k(\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k} = \lambda^r\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r}$. Since Proposition 17 below shows that the variance of KRR can be bounded by $\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}$, the additional term in the variance of BKRR, $\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}\sum_{j=1}^{k-1}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{j-1/2}$, implies that BKRR degrades the learning performance of KRR if their regularization parameters are identical. This coincides with the consensus that boosting is not worthwhile if the learner is already complex. Hence, in BKRR, a large λ should be chosen to guarantee under-fitting of the original KRR, i.e., large bias and small variance. The trade-off can be achieved via an appropriately tuned k such that the bias and variance are close.

Theorem 3 presents an error estimate for BKRR in terms of the empirical semi-norm $\|\cdot\|_D$. In the following theorem, we present error analysis for BKRR in terms of the $\|\cdot\|_\rho$ norm.

Theorem 4 *Let $k \geq 1$. Assume condition (5) with $r > 3/2$. Then there holds*

$$\|f_{D,\lambda}^{(k)} - f_\rho\|_\rho \leq 2\mathcal{Q}_{D,\lambda}(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho\lambda^{1/2}\mathcal{R}_D + 2k\mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda}$$
$$+ \quad \mathcal{Q}_{D,\lambda}\lambda^k\|h_\rho\|_\rho \begin{cases} \lambda^{r-k} + \kappa^{2r-2k-1}(\kappa+\lambda^{\frac{1}{2}}), & \text{if } k \leq r, \\ 2(\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k}, & \text{if } k > r. \end{cases} \tag{11}$$

Different from the exponential bias-variance trade-off of the error estimate in terms of the $\|\cdot\|_D$ semi-norm shown in Theorem 3, there exhibits a semi-exponential bias-variance trade-off for the error estimate in terms of the $\|\cdot\|_\rho$ norm. To be detailed, the bias decreases exponentially, while the variance increases polynomially as k increases. Based on Theorem 3 and Theorem 4, we can derive the following corollary.

Corollary 5 *Under condition (5) with $r \geq 1/2$, $\|L_{K,D}f_{D,\lambda}^{(k)} - S_D^T y_D\|_K$ decreases with respect to k. Moreover,*

$$\lim_{k\to\infty}\|L_{K,D}f_{D,\lambda}^{(k)} - S_D^T y_D\|_K \leq \lambda^{\frac{1}{2}}\mathcal{P}_{D,\lambda}\mathcal{Q}_{D,\lambda} \tag{12}$$

and

$$\lim_{k\to\infty}\left\|f_{D,\lambda}^{(k)}-f_\rho\right\|_D\leq\left(1+\frac{(\sigma_{min}^{\mathbf{x}}+\lambda)^{1/2}\lambda^{1/2}}{\sigma_{min}^{\mathbf{x}}}\right)\mathcal{P}_{D,\lambda}\mathcal{Q}_{D,\lambda}. \tag{13}$$

Corollary 5 exhibits an almost over-fitting resistance phenomenon of BKRR (neglecting the constant) for some kernels, since the sample error of KRR is bounded by $\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}$ (see Proposition 17 below). The behavior of the boosting iteration in (13) is different from that of the kernel-based (conjugate) gradient descent (Blanchard and Krämer, 2016; Lin and Zhou, 2018a), where the generalization error becomes ∞ for an arbitrary kernel, as the iteration number tends to infinity.

2.3 Adaptive Stopping Rule

We present in this subsection an adaptive stopping rule for BKRR to guarantee its optimal learning rates. To introduce the stopping rule, a user-friendly measurement of the capacity, empirical effective dimension (Lu et al., 2018; Mücke, 2018), defined by

$$\mathcal{N}_D(\lambda)=\mathrm{Tr}[(L_{K,D}+\lambda I)^{-1}L_{K,D}]=\mathrm{Tr}[(\lambda|D|I+\mathbb{K})^{-1}\mathbb{K}] \tag{14}$$

is needed. Denote

$$\mathcal{W}_{D,\lambda}\ =\frac{16\sqrt{2}(\kappa^2+\kappa+1)(\kappa M+\gamma)}{\sqrt{|D|}}\left(\frac{(\sqrt{|D|\lambda}+9)\sqrt{\max\{\mathcal{N}_D(\lambda),1\}}}{|D|\lambda}+1\right)\frac{(\sqrt{|D|\lambda}+9)\sqrt{\max\{\mathcal{N}_D(\lambda),1\}}}{\sqrt{|D|\lambda}}. \tag{15}$$

If $\delta\in(0,1)$ is the parameter corresponding to the confidence level, the boosting iteration will stop at the first positive integer $\hat{k}:=\hat{k}_{D,\lambda,\delta,K}$ satisfying

$$\|L_{K,D}f_{D,\lambda}^{(\hat{k})}-S_D^Ty_D\|_K\leq\lambda^{\frac{1}{2}}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}. \tag{16}$$

Since

$$L_{K,D}f_{D,\lambda}^{(\hat{k})}-S_D^Ty_D=\frac{1}{|D|}\sum_{i=1}^{|D|}(f_{D,\lambda}^{(\hat{k})}(x_i)-y_i)K_{x_i},$$

we have

$$\|L_{K,D}f_{D,\lambda}^{(\hat{k})}-S_D^Ty_D\|_K^2=\frac{1}{|D|^2}(f_{D,\lambda}^{(\hat{k})}(\mathbf{x})-\mathbf{y})^T\mathbb{K}(f_{D,\lambda}^{(\hat{k})}(\mathbf{x})-\mathbf{y}),$$

where $f_{D,\lambda}^{(\hat{k})}(\mathbf{x})-\mathbf{y}$ is the vector $\left(f_{D,\lambda}^{(\hat{k})}(x_i)-y_i\right)_{i=1}^{|D|}$. This together with (14) shows that the stopping rule in (16) is implementable. Moreover, Lemma 23 in Appendix A shows that

$$\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}\leq\frac{1}{2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta} \tag{17}$$

holds with confidence $1-\delta$. Then Corollary 5 verifies the existence of $\hat{k}$ with high probability since (16) is satisfied for sufficiently large $\hat{k}$ with high probability.

We are now in a position to present our second main result in the following theorem.

Theorem 6 *Let $\delta \in (0,1)$. Under conditions (4), (6) with $0 < s \le 1$ and condition (5) with $r \ge 1/2$, if $\lambda = (c/|D|)^{1/(2r+s)}$ for some $c \ge 1$, and $\hat{k}$ is the smallest positive integer satisfying (16), then with confidence at least $1-\delta$, there holds*

$$\|f_{D,\lambda}^{(\hat{k})} - f_\rho\|_\rho \le C|D|^{-\frac{r}{2r+s}} \log^{10} \frac{16}{\delta}, \tag{18}$$

where C is a constant independent of δ or $|D|$.

Theorem 6 shows that BKRR equipped with the stopping rule (16) achieves the same optimal learning rate without saturation, that is, the optimal learning rate holds for an arbitrary $r \ge 1/2$ rather than $\frac{1}{2} \le r \le 1$ shown by Caponnetto and De Vito (2007) and Lin et al. (2017) for KRR. Theorem 4 and Theorem 6 state that BKRR provides a novel semi-exponential bias-variance trade-off achieved by the boosting iteration, and the stopping rule (16) can realize its good performance. It follows from Corollary 5 and Theorem 6 that combining L_2 boosting with KRR reduces the difficulty of model selection (almost over-fitting resistance) and overcomes the saturation of KRR.

Remark 7 *Theoretically, a more delicate stopping rule for BKRR should be the first positive integer satisfying*

$$\|L_{K,D} f_{D,\lambda}^{(\hat{k})} - S_D^T y_D\|_K \le 2\lambda^{\frac{1}{2}} \mathcal{Q}_{D,\lambda} \mathcal{P}_{D,\lambda}.$$

Since the quantities $\mathcal{Q}_{D,\lambda}$ and $\mathcal{P}_{D,\lambda}$ cannot be implemented, we have to present a bound for them and thus get a confidence-dependent stopping rule (16). It should be noted that the constant in the definition of $\mathcal{W}_{D,\lambda}$ is not tight, which makes the algorithm stop much earlier than the optimal one. Due to the (semi-)exponential bias-variance trade-off presented in the previous subsection, a relatively large number of iterations does not degrade the generalization ability of BKRR very much. We thus multiply by a small factor to make the algorithm stop later. In a word, we implement the stopping rule (16) as

$$\begin{aligned} & \|L_{K,D} f_{D,\lambda}^{(\hat{k})} - S_D^T y_D\|_K \\ \le\ & \frac{\theta\sqrt{\lambda}}{\sqrt{|D|}} \left(\frac{(\sqrt{|D|\lambda}+1)\sqrt{\max\{\mathcal{N}_D(\lambda),1\}}}{|D|\lambda} + 1 \right) \frac{(\sqrt{|D|\lambda}+1)\sqrt{\max\{\mathcal{N}_D(\lambda),1\}}}{\sqrt{|D|\lambda}} \end{aligned} \tag{19}$$

for some small θ such as $\theta = 0.05$ (or other values).

Remark 8 *In Theorem 6, although k can be adaptively determined by (16), λ depends on r and s. It should be noted in Theorem 1 that $\lambda \sim |D|^{-1/(2r+s)}$ is the optimal regularization parameter for BKRR to achieve the optimal learning rate. The reason for this phenomenon is that we do not impose additional restrictions to the kernel K and the marginal distribution ρ_X other than (6). In particular, we use $\frac{\lambda}{\lambda+\sigma_{min}^{\mathbf{x}}} \le 1$ directly in the proof. For some particular kernel and ρ_X, the minimum positive eigenvalue $\sigma_{min}^{\mathbf{x}}$ for the matrix $\mathbb{K}/|D|$ can be computed. Then, we can derive similar error estimates by Theorem 4. In this way, optimal learning rates of BKRR hold for large values of λ. It would be interesting to determine the kernel and marginal distribution ρ_X, with which BKRR achieves the similar optimal*

learning rates as Theorem 6 for large λ. The other reason that we do not focus on the selection of λ is that boosting theory usually requires large λ to keep the algorithm underfitting and using iteration to reduce the bias. Thus, we can select a relatively large λ in advance numerically. Our experimental results in Section 6 show that the generalization ability of BKRR is not very sensitive to λ provided it is larger than some value.

Remark 9 *The constant exhibited in (18) is a bit pessimistic, compared with the classical results in the literature (Caponnetto and De Vito, 2007; Caponnetto and Yao, 2010; Blanchard and Krämer, 2016; Lin et al., 2017; Guo et al., 2017). One of the reasons for this pessimistic estimate is that we do not impose any restriction on the relation between $|D|$ and δ. In particular, as shown in our proof, if we assume $2\log(16/\delta) \leq \sqrt{|D|\lambda}$, i.e. $\delta \geq 16\exp\left\{-\frac{1}{2}c^{-1/(4r+2s)}|D|^{\frac{2r+s-1}{4r+2s}}\right\}$, then the exponent should be reduced from 10 to 6. Since the optimal constant is difficult to obtain, we only pursue the optimal learning rate in Theorem 6.*

3. Related Work

In this section, we discuss some related work in the literature and show the novelty of our results.

3.1 Boosting

A functional gradient descent viewpoint in statistics (Friedman et al., 2000; Friedman, 2001) reformulates boosting as a family of stage-wise optimization problems with different loss functions. Gradient boosting requires computing the negative gradient vector and line search in each boosting iteration. For L_2-Boosting, the gradient computation and line search can be unified in solving least squares fitting of residuals (Bühlmann and Yu, 2003). Thus, L_2-Boosting is essentially iterative least squares of residuals. An important advantage of boosting is its almost resistance to over-fitting (e.g. Friedman (2001) and its discussion papers), showing an easy way for model selection.

In Bühlmann and Yu (2003), an exponential bias-variance trade-off for linear regression was derived to illustrate the almost resistance to over-fitting for L_2-Boosting in a fixed design setting. In particular, Theorem 1 in Bühlmann and Yu (2003) shows that as the boosting iteration goes on, the bias decreases exponentially with a quantity depending on the minimum eigenvalue of the data matrix, while the variance increases with exponentially diminishing terms. Our Theorem 3 presents a similar result as Theorem 1 of Bühlmann and Yu (2003) for taking KRR as weak learners in L_2-Boosting, but highlights the importance of the regularity of the regression function by showing a sudden change of bias decaying. In Theorem 4, we also analyze the changes of bias and variance in a random design setting and show a semi-exponential bias-variance trade-off.

In Park et al. (2009), the learning performance of L_2-Boosting whose weak learners are Nadaraya-Watson kernel estimates was analyzed in the same framework as ours. It was shown in Theorem 2 and Theorem 3 of Park et al. (2009) that L_2-Boosting overcomes the saturation of Nadaraya-Watson kernel estimates (Györfi et al., 2002, Chapter 5). Differently, we utilize KRR as the weak learners instead of the kernel estimates, requiring totally different analysis. Furthermore, we present an adaptive stopping rule to select the number

of boosting iterations, while Park et al. (2009) requires an a priori knowledge-dependent number of boosting iterations. In this paper, we are concerned with combining L_2-Boosting with KRR. It would be interesting to consider boosted versions of other algorithms such as the kernel-based gradient descent (Yao et al., 2007) and more generally the kernel-based spectral algorithms (Gerfo et al., 2008).

3.2 Iterated Tikhonov Regularization

From Lemma 12 below, we find that for a fixed k, BKRR can be regarded as a special spectral algorithm, the iterated Tikhonov regularization (Gerfo et al., 2008). In this framework, the learning rate of BKRR with a fixed k may be derived directly from general results for spectral algorithms (Bauer et al., 2007; Caponnetto and Yao, 2010; Guo et al., 2017,b).

Different from the iterated Tikhonov regularization, BKRR focuses on fixed but relatively large λ and parameterizes the number of iterations, though they possess the same spectral representation (see (27) below). It follows from Theorem 3 that BKRR has an eventually stable relationship between the generalization error and boosting iteration in the sense that the generalization error does not increase much with the boosting iteration after some k. Theorem 6 shows that BKRR with adaptive stopping rule (16) can overcome the saturation of KRR, just as iterated Tikhonov regularization does but with an a priori knowledge-dependent selected and fixed k.

In a recent paper (Wu, 2017), a bias correction algorithm was proposed for ridge regression and detailed analysis was provided for the changes of bias and variance. It was found that one-step iteration can reduce the bias without increasing the variance much. The analysis in Wu (2017) is carried out in a more general framework than that in this paper. It should be pointed out that with the same setting in this paper, the algorithm in (Wu, 2017) possesses the spectral representation (27) below with $k = 1$.

Iterated Tikhonov regularization is closely related to BKRR and widely used in the community of inverse problems. Analysis of iterated Tikhonov regularization in solving ill-posed inverse problems can be dated back to the 1970's (e.g. King and Chillingworth (1979)). The optimal convergence rates and parameter selection of iterated Tikhonov regularization are important topics in inverse problems (Engl, 1987; Jin and Hou, 1997; Hanke and Groetsch, 1998; Jin and Stals, 2012). In particular, our stopping rule (16) is motivated by the discrepancy principle provided in Hanke and Groetsch (1998).

3.3 Iteration-based Learning Schemes and Stopping Rules

Saturation is a well known design-flaw of KRR (Gerfo et al., 2008) and limits its usage. Due to this phenomenon, researchers turn to other iteration-based learning schemes such as the kernel-based gradient descent (Yao et al., 2007), kernel-based conjugate gradient descent (Blanchard and Krämer, 2016) and kernel-based partial least squares (Lin and Zhou, 2018b). The theoretical results in Blanchard and Krämer (2016); Lin and Zhou (2018a,b) showed that these strategies can reach the optimal learning rates without saturation.

As an iteration-based algorithm, the bias and variance of the kernel-based gradient descent algorithm were analyzed in Lin and Zhou (2018a) and a polynomial bias-variance trade-off was exhibited. In particular, as the iteration goes on, its bias decreases as $\mathcal{O}(k^{-r})$ and its variance increases as a polynomial of k. Similar results on polynomial

bias-variance trade-off of the kernel-based conjugate gradient descent and kernel-based partial least squares were derived in Blanchard and Krämer (2016) and Lin and Zhou (2018b), respectively. Different from these iteration-based algorithms, BKRR shows an exponential bias-variance trade-off, reducing the difficulty of model selection.

Stopping rules play an important role in iteration-based learning schemes. Learning rates of iteration-based algorithms were built in Bauer et al. (2007); Yao et al. (2007); Guo et al. (2017); Blanchard and Krämer (2016) upon prior knowledge-based stopping rules. In Caponnetto and Yao (2010); Lin and Zhou (2018b), cross-validation based stopping rules were presented for general spectral algorithms and kernel-based partial least squares. In Raskutti et al. (2014), an adaptive stopping rule was deduced for the kernel-based gradient descent algorithm under the regularity condition (5) with $r = 1/2$. More recently, another adaptive stopping rule based on a balancing principle for general spectral algorithms was presented in Lu et al. (2018). Different from these results, our stopping rule presented in (16) requires neither dividing the sample set (compared with the cross-validation), nor computing estimators with various λ (compared with the balancing principle). Compared with Raskutti et al. (2014), our results hold under condition (5) with all $r \geq 1/2$, i.e., we adaptively select r rather than fixing it to be $1/2$. At first glance, the dependence of the confidence level in (16) may make the stopping rule not so stable. However, the (semi-) exponential bias-variance trade-off of BKRR compensates this instability by showing that the learning performance remains stable for a large range of k. It would be interesting to derive a confidence-independent stopping rule for BKRR.

4. Operator Representations and Error Estimates

We analyze the learning performance of BKRR by using the integral operator approach (Smale and Zhou, 2007; Lin et al., 2017; Guo et al., 2017). The novelties of our proof are special operator representations of BKRR, special spectral properties of BKRR and a tight bound for $\hat{k}$ defined by (16). In Subsections 4.1 and 4.2, we provide detailed spectral analysis for BKRR, which is crucial for deriving the bias and variance estimates in Subsections 4.3 and 4.4. In Subsection 4.5, we provide a tight bound on the number of boosting iteration defined by (16) by utilizing the special spectral properties of BKRR.

4.1 Special Operator Representations of BKRR

Define the noise-free version of $f_{D,\lambda}^{(k)}$ by

$$f_{D,\lambda}^{(1,*)} := \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{|D|} \sum_{(x,y)\in D} (f(x) - f_\rho(x))^2 + \lambda \|f\|_K^2 \right\} \tag{20}$$

and

$$f_{D,\lambda}^{(k,*)} := f_{D,\lambda}^{(k-1,*)} + f_{D,\lambda}^{(k,*)\diamond}, \qquad k > 1, \tag{21}$$

where

$$f_{D,\lambda}^{(k,*)\diamond} := \arg\min_{f\in\mathcal{H}_K}\left\{\frac{1}{|D|}\sum_{(x,y)\in D}[f(x)-(f_\rho(x)-f_{D,\lambda}^{(k-1,*)}(x))]^2+\lambda\|f\|_K^2\right\}, \qquad k>1. \tag{22}$$

For KRR, the classical result in Smale and Zhou (2007) shows

$$f_{D,\lambda}^{(1)} = (L_{K,D}+\lambda I)^{-1} S_D^T y_D, \quad \text{and} \quad f_{D,\lambda}^{(1,*)} = (L_{K,D}+\lambda I)^{-1} L_{K,D} f_\rho. \tag{23}$$

Similar to (23), the following Lemma 10 presents operator representations for $f_{D,\lambda}^{(k)}$ and $f_{D,\lambda}^{(k,*)}$.

Lemma 10 *Let $k \geq 2$. We have*

$$f_{D,\lambda}^{(k)} = [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]f_{D,\lambda}^{(k-1)} + f_{D,\lambda}^{(1)}, \tag{24}$$

$$L_{K,D}f_{D,\lambda}^{(k)} = S_D^T y_D - [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k S_D^T y_D \tag{25}$$

and

$$f_{D,\lambda}^{(k,*)} = [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]f_{D,\lambda}^{(k-1,*)} + f_{D,\lambda}^{(1,*)} = f_\rho - [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k f_\rho. \tag{26}$$

Proof. Since $f_{D,\lambda}^{(k)\diamond}$ is the solution to KRR (1) with data $\{x_i, y_i - f_{D,\lambda}^{(k-1)}(x_i)\}_{(x_i,y_i)\in D}$, it follows from (2), (23) and the definition $L_{K,D} = S_D^T S_D$ that

$$\begin{aligned} f_{D,\lambda}^{(k)} &= f_{D,\lambda}^{(k-1)} + (L_{K,D}+\lambda I)^{-1} S_D^T (y_D - S_D f_{D,\lambda}^{(k-1)}) \\ &= [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]f_{D,\lambda}^{(k-1)} + f_{D,\lambda}^{(1)}. \end{aligned}$$

This verifies (24). Combining this with (23) yields

$$\begin{aligned} & L_{K,D}f_{D,\lambda}^{(k)} - S_D^T y_D = L_{K,D}[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]f_{D,\lambda}^{(k-1)} + L_{K,D}f_{D,\lambda}^{(1)} - S_D^T y_D \\ = \;& [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]L_{K,D}f_{D,\lambda}^{(k-1)} + [(L_{K,D}+\lambda I)^{-1}L_{K,D} - I]S_D^T y_D \\ = \;& [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}][L_{K,D}f_{D,\lambda}^{(k-1)} - S_D^T y_D]. \end{aligned}$$

Applying this relation iteratively and using (23) give

$$\begin{aligned} L_{K,D}f_{D,\lambda}^{(k)} - S_D^T y_D &= [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^{k-1}[L_{K,D}f_{D,\lambda}^{(1)} - S_D^T y_D] \\ &= -[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k S_D^T y_D. \end{aligned}$$

This proves (25). As for deriving (26), we have

$$f_{D,\lambda}^{(k,*)} = [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]f_{D,\lambda}^{(k-1,*)} + (L_{K,D}+\lambda I)^{-1}L_{K,D}f_\rho.$$

It follows that

$$f_{D,\lambda}^{(k,*)} - f_\rho = [I-(L_{K,D}+\lambda I)^{-1}L_{K,D}][f_{D,\lambda}^{(k-1,*)} - f_\rho]$$

and by iterations,

$$\begin{aligned} f_{D,\lambda}^{(k,*)} - f_\rho &= [I - (L_{K,D} + \lambda I)^{-1} L_{K,D}]^{k-1} [f_{D,\lambda}^{(1,*)} - f_\rho] \\ &= -[I - (L_{K,D} + \lambda I)^{-1} L_{K,D}]^k f_\rho. \end{aligned}$$

This completes the proof of Lemma 10. ■

From Lemma 10, we have

$$f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)} = [I - (L_{K,D} + \lambda I)^{-1} L_{K,D}](f_{D,\lambda}^{(k-1)} - f_{D,\lambda}^{(k-1,*)}) + f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)},$$

from which the following expression is obtained by iterations.

Lemma 11 *For $k \in \mathbb{N}$, we have*

$$f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)} = \sum_{j=0}^{k-1} [I - (L_{K,D} + \lambda I)^{-1} L_{K,D}]^j [f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)}].$$

4.2 Special Spectral Properties of BKRR

Our analysis depends on some spectral analysis of BKRR, viewed as a special class of spectral algorithms. It follows iteratively from the identity

$$f_{D,\lambda}^{(k)} = \lambda (L_{K,D} + \lambda I)^{-1} f_{D,\lambda}^{(k-1)} + (L_{K,D} + \lambda I)^{-1} S_D^T y_D$$

obtained from (24) by writing $L_{K,D} = L_{K,D} + \lambda I - \lambda I$.

Lemma 12 *For $k \in \mathbb{N}$, we have*

$$f_{D,\lambda}^{(k)} = g_\lambda^{(k)}(L_{K,D}) S_D^T y_D, \tag{27}$$

where $g_\lambda^{(k)}(L_{K,D})$ is an operator on $\mathcal{H}_K$ defined by spectral calculus and

$$g_\lambda^{(k)}(\sigma) = \sum_{j=0}^{k-1} (\lambda(\sigma + \lambda)^{-1})^{k-j-1} (\sigma + \lambda)^{-1} = \sum_{j=0}^{k-1} \frac{\lambda^{k-1-j}}{(\sigma + \lambda)^{k-j}}. \tag{28}$$

Based on Lemma 12, we derive the following two lemmas, showing some special spectral properties of BKRR.

Lemma 13 *Let $g_\lambda^{(k)}$ be defined by (28), then we have*

$$I - L_{K,D} g_\lambda^{(k)}(L_{K,D}) = \lambda^k (L_{K,D} + \lambda I)^{-k}, \tag{29}$$

$$\|L_{K,D} g_\lambda^{(k)}(L_{K,D})]\| \le 1, \qquad \lambda \|g_\lambda^{(k)}(L_{K,D})\| \le k, \tag{30}$$

and for all $u > v > 0$, there holds

$$\|L_{K,D}^v \lambda^u (L_{K,D} + \lambda I)^{-u}\| \le v^v \left(\frac{\lambda}{u}\right)^v. \tag{31}$$

Proof. Observe from (28) that

$$\begin{aligned}\sigma g_\lambda^{(k)}(\sigma) &= (\sigma+\lambda-\lambda)\sum_{j=0}^{k-1}((\sigma+\lambda)^{-1})^{-j}\lambda^{k-j-1}(\sigma+\lambda)^{-k} \\ &= \left\{\sum_{j=0}^{k-1}(\sigma+\lambda)^{j+1}\lambda^{k-j-1}-\sum_{j=0}^{k-1}(\sigma+\lambda)^j\lambda^{k-j}\right\}(\sigma+\lambda)^{-k} \\ &= \{(\sigma+\lambda)^k-\lambda^k\}(\sigma+\lambda)^{-k}.\end{aligned}$$

Hence

$$L_{K,D}g_\lambda^{(k)}(L_{K,D}) = \left[(L_{K,D}+\lambda I)^k-\lambda^k I\right](L_{K,D}+\lambda I)^{-k} \tag{32}$$

and (29) follows. Then spectral analysis with the eigenpairs $\{(\sigma_i^{\mathbf{x}},\phi_i^{\mathbf{x}})\}$ of $L_{K,D}$ verifies the first inequality of (30). The second inequality of (30) follows directly from (28). Set a function $h_{v,u}$ on $[0,\infty)$ by

$$h_{v,u}(\sigma) = \frac{\sigma^v\lambda^u}{(\sigma+\lambda)^u}.$$

Since $u>v$, we have $h_{v,u}(0)=h_{v,u}(\infty)=0$. It is easy to check that $\sigma=\frac{v\lambda}{u-v}$ is the unique maximum point of $h_{v,u}$ on $(0,\infty)$. Thus, $\|L_{K,D}\|\le\kappa^2$ yields

$$\begin{aligned}&\|L_{K,D}^v\lambda^u(L_{K,D}+\lambda I)^{-u}\| \le \max_{0\le\sigma\le\kappa^2} h_{v,u}(\sigma) \le \max_{0\le\sigma<\infty} h_{v,u}(\sigma) \le h_{v,u}\left(\frac{v\lambda}{u-v}\right) \\ = \;& \lambda^v v^v\frac{(u-v)^u}{(u-v)^v u^u} = v^v\left(\frac{\lambda}{u}\right)^v\left(\frac{u-v}{u}\right)^{u-v} \le v^v\left(\frac{\lambda}{u}\right)^v.\end{aligned}$$

This completes the proof of Lemma 13. ∎

Lemma 14 *Let $u>0$ and $\ell\in\mathbb{N}_0:=\{0\}\cup\mathbb{N}$. Then for $f\in\mathcal{H}_K$, we have*

$$\|L_{K,D}^u[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^\ell f\|_K \le \begin{cases}\kappa^{2(u-\ell)}\lambda^\ell\|f\|_K, & \text{if } \ell\le u,\\ \lambda^u\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{\ell-u}\|f\|_K, & \text{if } \ell>u.\end{cases} \tag{33}$$

Proof. Due to spectral calculus with $f=\sum_i\langle f,\phi_i^{\mathbf{x}}\rangle_K\phi_i^{\mathbf{x}}$, we have

$$\begin{aligned}&\|L_{K,D}^u[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^\ell f\|_K^2 = \sum_i\frac{(\sigma_i^{\mathbf{x}})^{2u}\lambda^{2\ell}}{(\sigma_i^{\mathbf{x}}+\lambda)^{2\ell}}|\langle f,\phi_i^{\mathbf{x}}\rangle_K|^2 \\ =\;& \sum_{\sigma_i^{\mathbf{x}}>0}\frac{(\sigma_i^{\mathbf{x}})^{2u}\lambda^{2\ell}}{(\sigma_i^{\mathbf{x}}+\lambda)^{2\ell}}|\langle f,\phi_i^{\mathbf{x}}\rangle_K|^2.\end{aligned} \tag{34}$$

If $\ell\le u$, we have from $\max_i\sigma_i^{\mathbf{x}}\le\kappa^2$ and (34) that

$$\|L_{K,D}^u[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^\ell f\|_K^2 \le \lambda^{2\ell}\kappa^{4(u-\ell)}\sum_{\sigma_i^{\mathbf{x}}>0}|\langle f,\phi_i^{\mathbf{x}}\rangle_K|^2.$$

Then

$$\|L_{K,D}^u[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^\ell f\|_K \le \kappa^{2(u-\ell)}\lambda^\ell\|f\|_K.$$

If $\ell > u$, we get from (34) that

$$\begin{aligned}\|L_{K,D}^u[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^\ell f\|_K^2 &\le \lambda^{2u}\sum_{\sigma_i^{\mathbf{x}}>0}\frac{\lambda^{2\ell-2u}}{(\sigma_i^{\mathbf{x}}+\lambda)^{2\ell-2u}}|\langle f,\phi_i^{\mathbf{x}}\rangle_K|^2\\ \le\ \lambda^{2u}\frac{\lambda^{2\ell-2u}}{(\sigma_{min}^{\mathbf{x}}+\lambda)^{2\ell-2u}}\sum_{\sigma_i^{\mathbf{x}}>0}|\langle f,\phi_i^{\mathbf{x}}\rangle_K|^2 &\le \lambda^{2u}\frac{\lambda^{2\ell-2u}}{(\sigma_{min}^{\mathbf{x}}+\lambda)^{2\ell-2u}}\sum_i|\langle f,\phi_i^{\mathbf{x}}\rangle_K|^2.\end{aligned}$$

Hence,

$$\|L_{K,D}^u[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^\ell f\|_K \le \lambda^u\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{\ell-u}\|f\|_K.$$

This completes the proof of Lemma 14. ■

4.3 Bounding the Bias

Our error decomposition will be carried out by bounding the two terms, bias and variance, as follows

$$\|f_{D,\lambda}^{(k)}-f_\rho\| \le \|f_{D,\lambda}^{(k,*)}-f_\rho\| + \|f_{D,\lambda}^{(k)}-f_{D,\lambda}^{(k,*)}\|,$$

where $\|\cdot\|$ denotes either the $\|\cdot\|_D$ semi-norm or $\|\cdot\|_\rho$ norm. For $f\in\mathcal{H}_K$, it is easy to check that

$$\|f-f_\rho\|_D = \|L_{K,D}^{1/2}(f-f_\rho)\|_K. \tag{35}$$

In this subsection, we present two bounds for the bias term $f_{D,\lambda}^{(k,*)}-f_\rho$ in terms of the $\|\cdot\|_D$ semi-norm and $\|\cdot\|_\rho$ norm.

Proposition 15 *Let $0\le\nu\le 1/2$. Under condition (5) with $r>3/2$, we have*

$$\begin{aligned}\|L_{K,D}^\nu(f_{D,\lambda}^{(k,*)}-f_\rho)\|_K &\le \frac{(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho}{k^\nu}\lambda^\nu\mathcal{R}_D\\ +\ &\begin{cases}\kappa^{2(r+\nu-1/2-k)}\lambda^k\|h_\rho\|_\rho, & \text{if } k\le r+\nu-1/2,\\ \lambda^{r+\nu-1/2}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r-\nu+1/2}\|h_\rho\|_\rho, & \text{if } k>r+\nu-1/2.\end{cases}\end{aligned} \tag{36}$$

Proof. Since $r>3/2$, from (5) and (26) we find

$$\begin{aligned}&\|L_{K,D}^\nu(f_{D,\lambda}^{(k,*)}-f_\rho)\|_K = \|L_{K,D}^\nu[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k L_K^{r-1/2}L_K^{1/2}h_\rho\|_K\\ =\ &\left\|L_{K,D}^\nu[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k\left(L_K^{r-1/2}-L_{K,D}^{r-1/2}+L_{K,D}^{r-1/2}\right)L_K^{1/2}h_\rho\right\|_K\\ \le\ &\left\|L_{K,D}^{r+\nu-1/2}[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k L_K^{1/2}h_\rho\right\|_K\\ +\ &\left\|L_{K,D}^\nu[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k\left(L_K^{r-1/2}-L_{K,D}^{r-1/2}\right)L_K^{1/2}h_\rho\right\|_K\\ =:\ &A_1+A_2.\end{aligned} \tag{37}$$

We first estimate A_2. Since $r > 3/2$, the bounds $\|L_{K,D}\| \le \kappa^2$, $\|L_K\| \le \kappa^2$, we get from (68) in Appendix A that

$$\|L_{K,D}^{r-1/2} - L_K^{r-1/2}\|_{HS} \le (r-1/2)\kappa^{2r-3}\|L_{K,D} - L_K\|_{HS}. \tag{38}$$

When $\nu > 0$, we apply (31) with $v = \nu$ to obtain

$$\begin{aligned} A_2 &\le \|L_{K,D}^{\nu}[I - (L_{K,D} + \lambda I)^{-1}L_{K,D}]^k\| \left\|L_K^{r-1/2} - L_{K,D}^{r-1/2}\right\| \|L_K^{1/2}h_\rho\|_K \\ &\le \frac{(r-1/2)\nu^\nu \kappa^{2r-3}\|h_\rho\|_\rho}{k^\nu}\lambda^\nu \mathcal{R}_D. \end{aligned}$$

If $\nu = 0$, we can also obtain from $\|[I - (L_{K,D} + \lambda I)^{-1}L_{K,D}]^k\| \le 1$ that

$$A_2 \le (r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho \mathcal{R}_D.$$

Then we estimate A_1 by applying (33) with $u = r + \nu - 1/2$ and $f = L_K^{1/2}h_\rho$ to get

$$A_1 \le \begin{cases} \kappa^{2(r+\nu-1/2-k)}\lambda^k\|h_\rho\|_\rho, & \text{if } k \le r+\nu-1/2, \\ \lambda^{r+\nu-1/2}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r-\nu+1/2}\|h_\rho\|_\rho, & \text{if } k > r+\nu-1/2. \end{cases}$$

Plugging the estimates of A_1 and A_2 into (37), we obtain (36), which completes the proof of Proposition 15. ■

Proposition 16 *Under condition (5) with $r > 3/2$, we have*

$$\begin{aligned} \|f_{D,\lambda}^{(k,*)} - f_\rho\|_\rho &\le 2\mathcal{Q}_{D,\lambda}(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho \lambda^{\frac{1}{2}}\mathcal{R}_D \\ + \quad \mathcal{Q}_{D,\lambda}\|h_\rho\|_\rho &\begin{cases} \kappa^{2r-2k-1}(\lambda^{\frac{1}{2}}+\kappa)\lambda^k, & \text{if } k \le r-1/2, \\ \lambda^r\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r+1/2} + \kappa^{2r-2k}\lambda^k, & \text{if } r-1/2 < k \le r, \\ 2\lambda^r\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r}, & \text{if } k > r. \end{cases} \end{aligned}$$

Proof. Since $f_{D,\lambda}^{(k,*)} - f_\rho \in \mathcal{H}_K$, we have from Lemma 24 in Appendix A that

$$\|f_{D,\lambda}^{(k,*)} - f_\rho\|_\rho \le \mathcal{Q}_{D,\lambda}\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k,*)} - f_\rho)\|_K + \mathcal{Q}_{D,\lambda}\lambda^{1/2}\|f_{D,\lambda}^{(k,*)} - f_\rho\|_K.$$

For $k \le r - 1/2$, it follows from (36) with $\nu = 1/2$ and $\nu = 0$ that

$$\begin{aligned} \|f_{D,\lambda}^{(k,*)} - f_\rho\|_\rho &\le \mathcal{Q}_{D,\lambda}\left(\frac{(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho}{\sqrt{k}}\lambda^{1/2}\mathcal{R}_D + \kappa^{2(r-k)}\lambda^k\|h_\rho\|_\rho\right) \\ + \quad \lambda^{1/2}\mathcal{Q}_{D,\lambda}&\left((r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho\mathcal{R}_D + \kappa^{2(r-1/2-k)}\lambda^k\|h_\rho\|_\rho\right). \end{aligned}$$

For $r - 1/2 < k \le r$,

$$\begin{aligned} \|f_{D,\lambda}^{(k,*)} - f_\rho\|_\rho &\le \mathcal{Q}_{D,\lambda}\left(\frac{(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho}{\sqrt{k}}\lambda^{1/2}\mathcal{R}_D + \kappa^{2(r-k)}\lambda^k\|h_\rho\|_\rho\right) \\ + \quad \lambda^{1/2}\mathcal{Q}_{D,\lambda}&\left((r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho\mathcal{R}_D + \lambda^{r-1/2}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r+1/2}\|h_\rho\|_\rho\right). \end{aligned}$$

For $k > r$,

$$\|f_{D,\lambda}^{(k,*)} - f_\rho\|_\rho \le \mathcal{Q}_{D,\lambda}\left(\frac{(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho\lambda^{1/2}}{\sqrt{k}}\mathcal{R}_D + \lambda^r\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r}\|h_\rho\|_\rho\right)$$
$$+ \quad \lambda^{1/2}\mathcal{Q}_{D,\lambda}\left((r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho\mathcal{R}_D + \lambda^{r-1/2}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r+1/2}\|h_\rho\|_\rho\right).$$

This completes the proof of Proposition 16. ■

4.4 Bounding the Variance

In this subsection, we present the bounds for the variance term $f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)}$.

Proposition 17 *Let $0 < \nu \le 1/2$. We have*

$$\left\|L_{K,D}^\nu(f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)})\right\|_K \le \lambda^{\nu-1/2}\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}\left(1 + \sum_{j=1}^{k-1}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{j-\nu}\right). \tag{39}$$

Proof. Due to Lemma 11, we get

$$L_{K,D}^\nu(f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)}) = \sum_{j=0}^{k-1} L_{K,D}^\nu[I - (L_{K,D}+\lambda I)^{-1}L_{K,D}]^j[f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)}].$$

It then follows from (33) with $u = \nu$, $\ell = 1, 2, \ldots, k-1$ and $f = f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)}$ that

$$\left\|L_{K,D}^\nu(f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)})\right\|_K \le \sum_{j=0}^{k-1}\left\|L_{K,D}^\nu[I - (L_{K,D}+\lambda I)^{-1}L_{K,D}]^j[f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)}]\right\|_K$$
$$\le \quad \|L_{K,D}^\nu(f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)})\|_K + \sum_{j=1}^{k-1}\lambda^\nu\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{j-\nu}\|f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)}\|_K.$$

But (23) implies that for $0 \le u \le 1/2$, there holds

$$\|L_{K,D}^u(f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)})\|_K = \|L_{K,D}^u(L_{K,D}+\lambda I)^{-1}(S_D^T y_D - L_{K,D}f_\rho)\|_K$$
$$\le \quad \lambda^{-1/2+u}\|(L_{K,D}+\lambda I)^{-1/2}(S_D^T y_D - L_{K,D}f_\rho)\|_K \le \lambda^{-1/2+u}\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}. \tag{40}$$

Applying this inequality with $u = \nu$ and $u = 0$ yields

$$\left\|L_{K,D}^\nu(f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)})\right\|_K \quad \le \quad \lambda^{\nu-1/2}\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda} + \sum_{j=1}^{k-1}\lambda^{\nu-1/2}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{j-\nu}\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}.$$

This completes the proof of Proposition 17. ■

Different from Proposition 15, Proposition 17 does not hold for $\nu = 0$, which makes the bound of variance in the out-sample case totally different from that in the in-sample case.

Proposition 18 *For $k \in \mathbb{N}$, we have*

$$\|f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)}\|_\rho \leq 2k\mathcal{Q}_{D,\lambda}^2 \mathcal{P}_{D,\lambda}.$$

Proof. We obtain from Lemma 11 and (40) with $u = 0$ that

$$\|f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)}\|_K \leq \sum_{j=0}^{k-1} \|f_{D,\lambda}^{(1)} - f_{D,\lambda}^{(1,*)}\|_K \leq k\lambda^{-\frac{1}{2}} \mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}.$$

Then it follows from Lemma 24 in Appendix A and (39) with $\nu = 1/2$ that

$$\begin{aligned} \|f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)}\|_\rho &\leq \mathcal{Q}_{D,\lambda}\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)})\|_K + \mathcal{Q}_{D,\lambda}\lambda^{1/2}\|f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)}\|_K \\ \leq \quad & \mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda}\left(1 + \sum_{j=1}^{k-1}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}} + \lambda}\right)^{j-1/2}\right) + k\mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda} \leq 2k\mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda}. \end{aligned}$$

This completes the proof of Proposition 18. ■

4.5 Bounding the Number of Boosting Iterations

We first show the important role of the stopping rule (16) in controlling the bias.

Lemma 19 *Let $\delta \in (0,1)$ and $\lambda > 0$. If $\hat{k}$ is the smallest positive integer satisfying (16), then with confidence $1 - \delta$, there holds*

$$\|L_{K,D}f_\rho - L_{K,D}f_{D,\lambda}^{(\hat{k},*)}\|_K \leq \frac{3}{2}\lambda^{\frac{1}{2}}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}, \tag{41}$$

and

$$\|L_{K,D}f_\rho - L_{K,D}f_{D,\lambda}^{(\hat{k}-1,*)}\|_K \geq \frac{1}{2}\lambda^{\frac{1}{2}}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}, \qquad \text{if } \hat{k} \geq 2. \tag{42}$$

Proof. For $k \in \mathbb{N}$, we have

$$L_{K,D}f_\rho - L_{K,D}f_{D,\lambda}^{(k,*)} = S_D^T y_D - L_{K,D}f_{D,\lambda}^{(k)} + L_{K,D}f_{D,\lambda}^{(k)} - L_{K,D}f_{D,\lambda}^{(k,*)} + L_{K,D}f_\rho - S_D^T y_D.$$

But (27) and

$$f_{D,\lambda}^{(k,*)} = g_\lambda^{(k)}(L_{K,D})L_{K,D}f_\rho \tag{43}$$

yield

$$f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)} = g_\lambda^{(k)}(L_{K,D})(S_D^T y_D - L_{K,D}f_\rho). \tag{44}$$

It then follows from (32) that

$$\begin{aligned} L_{K,D}f_\rho - L_{K,D}f_{D,\lambda}^{(k,*)} &= S_D^T y_D - L_{K,D}f_{D,\lambda}^{(k)} + [L_{K,D}g_\lambda^{(k)}(L_{K,D}) - I](S_D^T y_D - L_{K,D}f_\rho) \\ &= S_D^T y_D - L_{K,D}f_{D,\lambda}^{(k)} - \lambda^k(L_{K,D} + \lambda I)^{-k}(S_D^T y_D - L_{K,D}f_\rho). \end{aligned} \tag{45}$$

Moreover, (17) implies

$$\begin{aligned} & \|\lambda^k(L_{K,D}+\lambda I)^{-k}(S_D^T y_D - L_{K,D}f_\rho)\|_K \le \|\lambda^k(L_{K,D}+\lambda I)^{-k}(L_{K,D}+\lambda I)^{1/2}\| \\ \times \quad & \|(L_{K,D}+\lambda I)^{-1/2}(L_K+\lambda I)^{1/2}\|\|(L_K+\lambda I)^{-1/2}(S_D^T y_D - L_{K,D}f_\rho)\|_K \\ \le \quad & \lambda^{1/2}\mathcal{P}_{D,\lambda}\mathcal{Q}_{D,\lambda} \le \frac{1}{2}\lambda^{1/2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta} \end{aligned} \tag{46}$$

with confidence $1-\delta$. Combining this with (16) and (45), we have that with confidence $1-\delta$, there holds

$$\|L_{K,D}f_\rho - L_{K,D}f_{D,\lambda}^{(\hat{k},*)}\|_K \le \frac{3}{2}\lambda^{1/2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta},$$

which proves (41). To prove (42), the definition of $\hat{k}$ implies

$$\|S_D^T y_D - L_{K,D}f_{D,\lambda}^{(\hat{k}-1)}\|_K > \lambda^{1/2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}, \qquad \text{if } \hat{k}\ge 2.$$

It follows from (45) with $k=\hat{k}-1$ when $\hat{k}\ge 2$ that

$$\|L_{K,D}f_\rho - L_{K,D}f_{D,\lambda}^{(\hat{k}-1,*)}\|_K \ge \lambda^{1/2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta} - \frac{1}{2}\lambda^{1/2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta} = \frac{1}{2}\lambda^{1/2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}$$

holds with confidence at least $1-\delta$. This completes the proof of Lemma 19. ■

Based on the above important lemma, we derive the following bound for $\hat{k}$.

Proposition 20 *Let $\delta\in(0,1)$ and $\hat{k}$ be the smallest positive integer satisfying (16). We have with confidence $1-\delta$ that*

$$\begin{aligned} \hat{k} \quad \le \quad & (4r+2) + 4\mathcal{W}_{D,\lambda}^{-1}\left(\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2+1\right)\lambda^r\|h_\rho\|_\rho\log^{-2}\frac{16}{\delta} \\ + \quad & (4r-2)\mathcal{W}_{D,\lambda}^{-1}\frac{4\kappa^{2r-1}}{\sqrt{|D|}}\lambda^{1/2}\|h_\rho\|_\rho\log^{-3}\frac{16}{\delta}, \end{aligned} \tag{47}$$

where

$$\mathcal{A}_{D,\lambda} := \frac{1}{\sqrt{|D|}}\left(\frac{1}{\sqrt{|D|\lambda}}+\sqrt{\mathcal{N}(\lambda)}\right). \tag{48}$$

Proof. If $\hat{k}\le r+3/2$, (47) obviously holds. Now we prove (47) for $\hat{k}>r+3/2$. Due to Lemma 12, Lemma 19 and (43), we have with confidence at least $1-\delta$ that

$$\begin{aligned} & \lambda^{\frac{1}{2}}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta} \le 2\|L_{K,D}(f_\rho - f_{D,\lambda}^{(\hat{k}-1,*)})\|_K \\ = \quad & 2\left\|L_{K,D}\left(g_\lambda^{(\hat{k}-1)}(L_{K,D})L_{K,D}-I\right)f_\rho\right\|_K \le 2\|L_{K,D}\lambda^{\hat{k}-1}(L_{K,D}+\lambda I)^{-\hat{k}+1}L_K^{r-1/2}\|\|h_\rho\|_\rho, \end{aligned}$$

where the last inequality is due to (32).

If $\frac{1}{2} \le r \le \frac{3}{2}$, we have from Lemma 13 and $\hat{k} - r - \frac{1}{2} > 1$ that

$$\begin{aligned} & \lambda^{\frac{1}{2}} \mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} \le 2\|L_{K,D}\lambda^{\hat{k}-1}(L_{K,D}+\lambda I)^{-\hat{k}+1}(L_{K,D}+\lambda I)^{r-1/2}\| \mathcal{Q}_{D,\lambda}^{2r-1}\|h_\rho\|_\rho \\ = \;& 2\mathcal{Q}_{D,\lambda}^{2r-1}\|h_\rho\|_\rho \lambda^{r-1/2}\|L_{K,D}\lambda^{\hat{k}-1-r+1/2}(L_{K,D}+\lambda I)^{-\hat{k}+1+r-1/2}\| \\ \le \;& 2\max(\mathcal{Q}_{D,\lambda}^2, 1)\|h_\rho\|_\rho \lambda^{r+1/2}(\hat{k}-r-1/2)^{-1}. \end{aligned}$$

Thus, it follows from Lemma 22 in Appendix A that with confidence at least $1-\delta$

$$\hat{k}\mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} \le (r+1/2)\mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} + 4\log^2 \frac{16}{\delta}\left(\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2 + 1\right)\lambda^r \|h_\rho\|_\rho,$$

which implies (47).

If $r > 3/2$, it follows from (38), Lemma 13 and $\hat{k} > r + 3/2$ that

$$\begin{aligned} & \lambda^{1/2} \mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} \le 2\|L_{K,D}\lambda^{\hat{k}-1}(L_{K,D}+\lambda I)^{-\hat{k}+1} L_{K,D}^{r-1/2}\| \|h_\rho\|_\rho \\ + \;& 2\|L_{K,D}\lambda^{\hat{k}-1}(L_{K,D}+\lambda I)^{-\hat{k}+1}(L_K^{r-1/2} - L_{K,D}^{r-1/2})\| \|h_\rho\|_\rho \\ \le \;& 2\|h_\rho\|_\rho \lambda^{r-1/2}\|L_{K,D}\lambda^{\hat{k}-1-r+1/2}(L_{K,D}+\lambda I)^{-\hat{k}+1+r-1/2}\| \\ + \;& (2r-1)\kappa^{2r-3}\|h_\rho\|_\rho \|L_{K,D}\lambda^{\hat{k}-1}(L_{K,D}+\lambda I)^{-\hat{k}+1}\| \|L_K - L_{K,D}\| \\ \le \;& 2\|h_\rho\|_\rho \lambda^{r+1/2}(\hat{k}-r-1/2)^{-1} + (2r-1)\kappa^{2r-3}\|h_\rho\|_\rho \frac{\lambda}{\hat{k}-1}\mathcal{R}_D. \end{aligned}$$

If

$$\lambda^{r+1/2}(\hat{k}-r-1/2)^{-1} \le (r-1/2)\kappa^{2r-3}\frac{\lambda}{\hat{k}-1}\mathcal{R}_D,$$

we have

$$\lambda^{\frac{1}{2}} \mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} \le (4r-2)\kappa^{2r-3}\frac{\lambda}{\hat{k}-1}\mathcal{R}_D\|h_\rho\|_\rho,$$

which together with Lemma 22 in Appendix A yields with confidence $1-\delta$

$$\hat{k}\mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} \le \mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} + (4r-2)\frac{4\kappa^{2r-1}}{\sqrt{|D|}}\lambda^{1/2}\|h_\rho\|_\rho \log \frac{16}{\delta}.$$

Thus, (47) holds. If

$$\lambda^{r+1/2}(\hat{k}-r-1/2)^{-1} > (r-1/2)\kappa^{2r-3}\frac{\lambda}{\hat{k}-1}\mathcal{R}_D,$$

we get

$$\lambda^{\frac{1}{2}} \mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} \le 4\lambda^{r+1/2}(\hat{k}-r-1/2)^{-1}\|h_\rho\|_\rho.$$

Hence,

$$\hat{k}\mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} \le (r+\frac{1}{2})\mathcal{W}_{D,\lambda} \log^4 \frac{16}{\delta} + 4\lambda^r\|h_\rho\|_\rho,$$

which together with $\log \frac{16}{\delta} > 1$ yields (47). The proof of Proposition 20 is completed. ■

5. Proofs of Main Results

Based on the previous bounds we can now prove our main results.

Proof of Theorem 3. If $1/2 \le r \le 3/2$, we get from (5) and (26) that

$$\begin{aligned}
&\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k,*)} - f_\rho)\|_K = \|L_{K,D}^{1/2}[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k L_K^r h_\rho\|_K \\
= \;& \|L_{K,D}^{1/2}[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k (L_{K,D}+\lambda I)^{r-1/2}(L_{K,D}+\lambda I)^{1/2-r} L_K^r h_\rho\|_K \\
\le \;& 2^{r-1/2}\|L_{K,D}^{r}[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k (L_{K,D}+\lambda I)^{1/2-r} L_K^r h_\rho\|_K \\
+ \;& 2^{r-1/2}\lambda^{r-1/2}\|L_{K,D}^{1/2}[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k (L_{K,D}+\lambda I)^{1/2-r} L_K^r h_\rho\|_K.
\end{aligned}$$

Here we have used the inequality

$$\|(L_{K,D}+\lambda I)^r f\|_K \le 2^{r-1/2}\big[\|L_{K,D}^r f\|_K + \lambda^r \|f\|_K\big], \quad \forall f \in \mathcal{H}_K$$

which follows by means of the normalized eigenpairs $\{(\sigma_i^{\mathbf{x}}, \phi_i^{\mathbf{x}})\}_i$ of $L_{K,D}: \mathcal{H}_K \to \mathcal{H}_K$ as

$$\begin{aligned}
\|(L_{K,D}+\lambda I)^r f\|_K^2 &= \sum_i \big(\sigma_i^{\mathbf{x}}+\lambda\big)^{2r} |\langle \phi_i^{\mathbf{x}}, f\rangle_K|^2 \\
&\le 2^{2r-1}\sum_i \big[(\sigma_i^{\mathbf{x}})^{2r}|\langle \phi_i^{\mathbf{x}}, f\rangle_K|^2 + \lambda^{2r}|\langle \phi_i^{\mathbf{x}}, f\rangle_K|^2\big].
\end{aligned}$$

Since $0 \le r - \frac{1}{2} \le 1$, (67) in Appendix B shows

$$\|(L_{K,D}+\lambda I)^{1/2-r} L_K^r h_\rho\|_K \le \mathcal{Q}_{D,\lambda}^{2r-1}\|h_\rho\|_\rho.$$

It then follows from (33) with $u = r$, $f = (L_{K,D}+\lambda I)^{1/2-r} L_K^r h_\rho$ and $u = 1/2$, $f = (L_{K,D}+\lambda I)^{1/2-r} L_K^r h_\rho$ that

$$\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k,*)} - f_\rho)\|_K \le 2\sqrt{2}\mathcal{Q}_{D,\lambda}^{2r-1}\|h_\rho\|_\rho \lambda^r \left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r}, \qquad \text{if } 1/2 \le r < 1,$$

$$\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k,*)} - f_\rho)\|_K \le 2\mathcal{Q}_{D,\lambda}^{2r-1}\lambda\|h_\rho\|_\rho(\kappa^{2r-2}+\lambda^{r-1}), \qquad \text{if } 1 \le r \le 3/2, k = 1,$$

and

$$\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k,*)} - f_\rho)\|_K \le 4\mathcal{Q}_{D,\lambda}^{2r-1}\lambda^r\|h_\rho\|_\rho \left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-r}, \qquad \text{if } 1 \le r \le 3/2, k \ge 2.$$

The above estimates together with (36) with $\nu = 1/2$ yield for $r \ge 1/2$,

$$\begin{aligned}
\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k,*)} - f_\rho)\|_K &\le \frac{(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho}{\sqrt{k}}\lambda^{1/2}\mathcal{R}_D \\
&+ 4(\mathcal{Q}_{D,\lambda}^2+1)\lambda^k\|h\|_\rho \begin{cases} (\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k}, & \text{if } k > r, \\ \kappa^{2r-2k}+\lambda^{r-k}, & \text{if } k \le r. \end{cases}
\end{aligned}$$

Furthermore, (39) with $\nu = 1/2$ implies

$$\Big\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k)} - f_{D,\lambda}^{(k,*)})\Big\|_K \le \mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}\left(1 + \sum_{j=1}^{k-1}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{j-1/2}\right).$$

Hence

$$\begin{aligned}
&\left\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k)}-f_\rho)\right\|_K \le \|L_{K,D}^{1/2}(f_{D,\lambda}^{(k,*)}-f_\rho)\|_K + \left\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k)}-f_{D,\lambda}^{(k,*)})\right\|_K \\
\le\ & \mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}\sum_{j=1}^{k-1}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{j-1/2} + \frac{(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho}{\sqrt{k}}\lambda^{1/2}\mathcal{R}_D + \mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda} \\
+\ & 4(\mathcal{Q}_{D,\lambda}^2+1)\lambda^k\|h\|_\rho\begin{cases}(\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k}, & \text{if } k>r,\\ \kappa^{2r-2k}+\lambda^{r-k}, & \text{if } k\le r.\end{cases}
\end{aligned}$$

This completes the proof of Theorem 3. ■

Proof of Theorem 4. Since $r > 3/2$, we get from Proposition 16 that

$$\begin{aligned}
&\|f_{D,\lambda}^{(k,*)}-f_\rho\|_\rho \le 2\mathcal{Q}_{D,\lambda}(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho\lambda^{1/2}\mathcal{R}_D \\
+\ & \mathcal{Q}_{D,\lambda}\lambda^k\|h_\rho\|_\rho\begin{cases}\lambda^{r-k}+\kappa^{2r-2k-1}(\kappa+\lambda^{\frac{1}{2}}), & \text{if } k\le r,\\ 2(\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k}, & \text{if } k>r.\end{cases}
\end{aligned}$$

Furthermore, it follows from Proposition 18 that

$$\|f_{D,\lambda}^{(k)}-f_{D,\lambda}^{(k,*)}\|_\rho \le 2k\mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda}.$$

Then

$$\begin{aligned}
&\|f_{D,\lambda}^{(k)}-f_\rho\|_\rho \le \|f_{D,\lambda}^{(k,*)}-f_\rho\|_\rho + \|f_{D,\lambda}^{(k)}-f_{D,\lambda}^{(k,*)}\|_\rho \\
\le\ & 2\mathcal{Q}_{D,\lambda}(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho\lambda^{1/2}\mathcal{R}_D \\
+\ & 2k\mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda} + \mathcal{Q}_{D,\lambda}\lambda^k\|h_\rho\|_\rho\begin{cases}\lambda^{r-k}+\kappa^{2r-2k-1}(\kappa+\lambda^{\frac{1}{2}}), & \text{if } k\le r,\\ 2(\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k}, & \text{if } k>r.\end{cases}
\end{aligned}$$

This completes the proof of Theorem 4. ■

Proof of Theorem 1. We get from Lemma 22 in Appendix A, (48), $r > 3/2$, (6) and $\lambda = \left(\frac{k^2}{|D|}\right)^{\frac{1}{2\min\{k,r\}+s}}$ with $k \le \sqrt{|D|}$ that with confidence $1-\delta$,

$$\mathcal{R}_D \le \frac{4\kappa^2}{\sqrt{|D|}}\log\frac{8}{\delta} \le \frac{4\kappa^2}{\sqrt{|D|\lambda^s}}\log\frac{8}{\delta},$$

$$\mathcal{Q}_{D,\lambda}^2 \le 2\left(2(\kappa^2+\kappa)\left(\frac{1}{|D|\lambda}+\sqrt{\frac{\mathcal{N}(\lambda)}{|D|\lambda}}\right)\log\frac{8}{\delta}\right)^2 + 2 \le 8(\kappa^2+\kappa+1)^2(1+\sqrt{C_0})^2\log^2\frac{8}{\delta},$$

and

$$\begin{aligned}
&\mathcal{P}_{D,\lambda}\mathcal{Q}_{D\lambda} \le \frac{4\sqrt{2}(\kappa^2+\kappa+1)(\kappa M+\gamma)}{\sqrt{|D|}}\left(\frac{1}{\lambda|D|}+\frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{|D|\lambda}}+1\right)\left(\frac{1}{\sqrt{\lambda|D|}}+\sqrt{\mathcal{N}(\lambda)}\right)\log^2\frac{8}{\delta} \\
\le\ & \frac{4\sqrt{2}(\kappa^2+\kappa+1)(\kappa M+\gamma)(2+\sqrt{C_0})(1+\sqrt{C_0})}{\sqrt{|D|\lambda^s}}\log^2\frac{8}{\delta}.
\end{aligned}$$

Here we have used $\frac{1}{\sqrt{\lambda|D|}} + \sqrt{\mathcal{N}(\lambda)} \leq \frac{\sqrt{C_0}+1}{\sqrt{\lambda^s}}$. Applying these three estimates to Theorem 4, we get

$$\begin{aligned}
&\|f_{D,\lambda}^{(k)} - f_\rho\|_\rho \leq 4\sqrt{2}(\kappa^2+\kappa+1)(1+\sqrt{C_0})(r-1/2)\kappa^{2r-3}\|h_\rho\|_\rho \lambda^{1/2}\frac{4\kappa^2}{\sqrt{|D|\lambda^s}}\log^2\frac{8}{\delta} \\
+ \quad & 32(2+\sqrt{C_0})^3(\kappa^2+\kappa+1)^2(\kappa M+\gamma)\frac{k}{\sqrt{|D|\lambda^s}}\log^3\frac{8}{\delta} \\
+ \quad & 2\sqrt{2}(1+\sqrt{C_0})(\kappa^2+\kappa+1)\log\frac{8}{\delta}\lambda^k\|h_\rho\|_\rho \begin{cases} \lambda^{r-k}+\kappa^{2r-2k-1}(\kappa+\lambda^{\frac{1}{2}}), & \text{if } k\leq r, \\ 2(\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k}, & \text{if } k>r. \end{cases} \\
\leq \quad & \frac{\tilde{C}k}{2\sqrt{|D|\lambda^s}}\log^3\frac{8}{\delta} + \frac{\tilde{C}\log\frac{8}{\delta}}{2}\lambda^k \begin{cases} 1, & \text{if } k\leq r, \\ (\sigma_{min}^{\mathbf{x}}+\lambda)^{r-k}, & \text{if } k>r, \end{cases}
\end{aligned}$$

where $\tilde{C}$ is a constant independent of $|D|, k$ or δ given by

$$\begin{aligned}
\tilde{C} = 2(\kappa^2+\kappa+1)\max\{&16\sqrt{2}(1+\sqrt{C_0})(r-1/2)\kappa^{2r-1}\|h_\rho\|_\rho + 32(2+\sqrt{C_0})^3(\kappa^2+\kappa+1)(\kappa M+\gamma), \\
&2\sqrt{2}(1+\sqrt{C_0})\|h_\rho\|_\rho(2+\kappa^{2r-2k}+\kappa^{2r-2k-1})\}.
\end{aligned}$$

Plugging $\lambda = (k^2/|D|)^{\frac{1}{2\min\{k,r\}+s}}$ into the above estimate and noting $\sigma_{min}^{\mathbf{x}} > 0$, we get that

$$\left\|f_{D,\lambda}^{(k)} - f_\rho\right\|_\rho \leq \tilde{C}(k^2/|D|)^{\frac{\min\{k,r\}}{2\min\{k,r\}+s}}\log^3\frac{8}{\delta},$$

holds with confidence $1-\delta$. This completes the proof of Theorem 1. ■

Proof of Corollary 5. It follows from (25) that

$$\begin{aligned}
&\|L_{K,D}f_{D,\lambda}^{(k+1)} - S_D^T y_D\|_K = \left\|[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^{k+1}S_D^T y_D\right\|_K \\
\leq \quad & \|I-(L_{K,D}+\lambda I)^{-1}L_{K,D}\|\|[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k S_D^T y_D\|_K \\
\leq \quad & \|L_{K,D}f_{D,\lambda}^{(k)} - S_D^T y_D\|_K.
\end{aligned}$$

Thus $\|L_{K,D}f_{D,\lambda}^{(k)} - S_D^T y_D\|_K$ decreases with respect to k. Furthermore, we have from (25) again and Lemma 14 that for $k>1$

$$\begin{aligned}
&\|L_{K,D}f_{D,\lambda}^{(k)} - S_D^T y_D\|_K \leq \|[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k(S_D^T y_D - L_{K,D}f_\rho)\|_K \\
+ \quad & \|[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k L_{K,D}f_\rho\|_K \\
\leq \quad & \|[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k(L_{K,D}+\lambda I)^{1/2}\|\|(L_{K,D}+\lambda I)^{-1/2}(S_D^T y_D - L_{K,D}f_\rho)\|_K \\
+ \quad & \|[I-(L_{K,D}+\lambda I)^{-1}L_{K,D}]^k L_{K,D}f_\rho\|_K \\
\leq \quad & \lambda^{1/2}\|(L_{K,D}+\lambda I)^{-1/2}(L_{K,D}f_\rho - S_D^T y_D)\|_K + \lambda\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{k-1}\|f_\rho\|_K.
\end{aligned}$$

Thus,

$$\begin{aligned}
\lim_{k\to\infty}\|L_{K,D}f_{D,\lambda}^{(k)} - S_D^T y_D\|_K &\leq \lambda^{1/2}\|(L_{K,D}+\lambda I)^{-1/2}(L_{K,D}f_\rho - S_D^T y_D)\|_K \\
&\leq \lambda^{1/2}\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}.
\end{aligned}$$

This verifies (12). To prove (13), it follows from Theorem 3 that

$$\lim_{k\to\infty}\left\|L_{K,D}^{1/2}(f_{D,\lambda}^{(k)}-f_\rho)\right\|_K \le \mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}+\mathcal{Q}_{D,\lambda}\mathcal{P}_{D,\lambda}\lim_{k\to\infty}\sum_{j=1}^{k-1}\left(\frac{\lambda}{\sigma_{min}^{\mathbf{x}}+\lambda}\right)^{j-1/2}$$

$$= \left(1+\frac{(\sigma_{min}^{\mathbf{x}}+\lambda)^{1/2}\lambda^{1/2}}{\sigma_{min}^{\mathbf{x}}}\right)\mathcal{P}_{D,\lambda}\mathcal{Q}_{D,\lambda}.$$

This completes the proof of Corollary 5. ■

Proof of Theorem 6. Since

$$\|f_{D,\lambda}^{(\hat{k})}-f_\rho\|_\rho \le \|f_{D,\lambda}^{(\hat{k},*)}-f_\rho\|_\rho+\|f_{D,\lambda}^{(\hat{k})}-f_{D,\lambda}^{(\hat{k},*)}\|_\rho =: A(D,\lambda,\hat{k})+S(D,\lambda,\hat{k}). \tag{49}$$

We divide the proof into four steps.

Step 1: Bounding $A(D,\lambda,\hat{k})$. Define

$$\widetilde{\mathcal{Q}}_{D,\lambda}=\|(L_{K,D}+\lambda I)^{-1}(L_K+\lambda I)\|. \tag{50}$$

We obtain from $\|L_K^{1/2}(L_K+\lambda I)^{-1/2}\|\le 1$ and $\|(L_K+\lambda I)^{-\frac{1}{2}}\|\le\lambda^{-\frac{1}{2}}$ that

$$\begin{aligned} A(D,\lambda,\hat{k}) &= \|L_K^{1/2}(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)\|_K \le \lambda^{-1/2}\|(L_K+\lambda I)(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)\|_K \\ &\le \lambda^{-1/2}\widetilde{\mathcal{Q}}_{D,\lambda}\|(L_{K,D}+\lambda I)(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)\|_K \\ &\le \lambda^{-1/2}\widetilde{\mathcal{Q}}_{D,\lambda}\|L_{K,D}(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)\|_K+\lambda^{1/2}\widetilde{\mathcal{Q}}_{D,\lambda}\|f_{D,\lambda}^{(\hat{k},*)}-f_\rho\|_K. \end{aligned} \tag{51}$$

By Lemma 19, with confidence at least $1-\delta$, there holds

$$\|L_{K,D}(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)\|_K \le \frac{3}{2}\lambda^{1/2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}. \tag{52}$$

Let F_λ be the orthogonal projection onto the subspace of $\mathcal{H}_K$ spanned by the eigenvectors of $L_{K,D}$ associated with eigenvalues less than λ and $F_\lambda^\perp=I-F_\lambda$. We have

$$\|f_{D,\lambda}^{(\hat{k},*)}-f_\rho\|_K \le \lambda^{-1}\|F_\lambda[\lambda(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)]\|_K+\lambda^{-1}\|F_\lambda^\perp[\lambda(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)]\|_K =: A_1+A_2. \tag{53}$$

By the definition of $F_\lambda^\perp$ and Lemma 19, it follows with confidence $1-\delta$,

$$A_2 \le \lambda^{-1}\|F_\lambda^\perp[L_{K,D}(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)]\|_K \le \lambda^{-1}\|L_{K,D}(f_{D,\lambda}^{(\hat{k},*)}-f_\rho)\|_K \le \frac{3}{2}\lambda^{-1/2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}. \tag{54}$$

Due to (43) and (5), we have

$$A_1 \le \|F_\lambda[g_\lambda^{(\hat{k})}(L_{K,D})L_{K,D}f_\rho-f_\rho)]\|_K \le \|F_\lambda[\lambda^{\hat{k}}(L_{K,D}+\lambda I)^{-\hat{k}}L_K^{r-1/2}]\|\|h_\rho\|_\rho.$$

If $\frac{1}{2}\le r\le\frac{3}{2}$, we have

$$A_1 \le \|F_\lambda[\lambda^{\hat{k}}(L_{K,D}+\lambda I)^{-\hat{k}}(L_{K,D}+\lambda I)^{r-1/2}]\|(\widetilde{\mathcal{Q}}_{D,\lambda})^{r-1/2}\|h_\rho\|_\rho \le (\widetilde{\mathcal{Q}}_{D,\lambda})^{r-1/2}\|h_\rho\|_\rho\lambda^{r-1/2}. \tag{55}$$

If $r > 3/2$, (38) implies

$$\begin{aligned} A_1 &\leq \|F_\lambda[\lambda^{\hat{k}}(L_{K,D}+\lambda I)^{-\hat{k}}(L_K^{r-1/2}-L_{K,D}^{r-1/2})]\|\|h_\rho\|_\rho + \|F_\lambda[\lambda^{\hat{k}}(L_{K,D}+\lambda I)^{-\hat{k}}L_{K,D}^{r-1/2}]\|\|h_\rho\|_\rho \\ &\leq \|h_\rho\|_\rho\left((r-1/2)\kappa^{2r-3}\mathcal{R}_D+\lambda^{r-1/2}\right). \end{aligned} \tag{56}$$

Inserting (56), (55) and (54) into (53) and then plugging (53) and (52) into (51), we have from Lemma 22 in Appendix A that with confidence $1-\delta$,

$$\begin{aligned} A(D,\lambda,\hat{k}) &\leq 3\log^6\frac{16}{\delta}\left[\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]\mathcal{W}_{D,\lambda} \\ &+ 4\lambda^{\frac{1}{2}}\|h_\rho\|_\rho\log^4\frac{16}{\delta}\left[\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]^2\left(\lambda^{r-\frac{1}{2}}+\frac{(4r-2)\kappa^{2r-1}}{\sqrt{|D|}}\right). \end{aligned} \tag{57}$$

Step 2: Bounding $S(D,\lambda,\hat{k})$. It follows from (30), (44) and the definitions of $\mathcal{Q}_{D,\lambda}$ and $\mathcal{P}_{D,\lambda}$ that

$$\begin{aligned} S(D,\lambda,\hat{k}) &= \|L_K^{1/2}[f_{D,\lambda}^{(\hat{k})}-f_{D,\lambda}^{(\hat{k},*)}]\|_K \\ &\leq \|(L_K+\lambda I)^{1/2}g_\lambda^{(\hat{k})}(L_{K,D})(L_{K,D}+\lambda I)^{1/2}(L_{K,D}+\lambda I)^{-1/2}(L_{K,D}f_\rho-S_D^Ty_D)\|_K \\ &\leq \mathcal{Q}_{D,\lambda}^2\|g_\lambda^{(\hat{k})}(L_{K,D})(L_{K,D}+\lambda I)\|\mathcal{P}_{D,\lambda} \\ &\leq \mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda}\left[\|g_\lambda^{(\hat{k})}(L_{K,D})L_{K,D}\|+\lambda\|g_\lambda^{(\hat{k})}(L_{K,D})\|\right] \\ &\leq (\hat{k}+1)\mathcal{Q}_{D,\lambda}^2\mathcal{P}_{D,\lambda}. \end{aligned}$$

This together with (17) and Lemma 22 in Appendix A implies with confidence $1-\delta$

$$S(D,\lambda,\hat{k}) \leq \frac{\hat{k}+1}{2}\mathcal{W}_{D,\lambda}\log^5\frac{16}{\delta}\left[\sqrt{2}\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)+\sqrt{2}\right].$$

Combining the above inequality with (47) yields

$$\begin{aligned} S(D,\lambda,\hat{k}) &\leq \left[\sqrt{2}\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)+\sqrt{2}\right]\log\frac{16}{\delta}\left\{(2r+1)\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}\right. \\ &+ 2\left[\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}}\right)^2+1\right]\lambda^r\|h_\rho\|_\rho\log^2\frac{16}{\delta} \\ &+ \left.(2r-1)\frac{4\kappa^{2r-1}}{\sqrt{|D|}}\lambda^{1/2}\|h_\rho\|_\rho\log\frac{16}{\delta}\right\}. \end{aligned} \tag{58}$$

Step 3: Bounding $\mathcal{A}_{D,\lambda}$ and $\mathcal{W}_{D,\lambda}$. Since $r \geq 1/2$, and $\lambda = (c/|D|)^{1/(2r+s)}$ with $c \geq 1$, it follows from (6), (48) and $r \geq \frac{1}{2}$ that

$$\mathcal{A}_{D,\lambda} = \frac{1}{\sqrt{|D|}}\left\{\frac{1}{\sqrt{|D|\lambda}}+\sqrt{\mathcal{N}(\lambda)}\right\} \leq (1+\sqrt{C_0})c^{-s/(4r+2s)}|D|^{-r/(2r+s)}. \tag{59}$$

This implies

$$\mathcal{A}_{D,\lambda}^2/\lambda \leq (1+\sqrt{C_0})^2 c^{(-s-1)/(2r+s)}|D|^{\frac{1-2r}{2r+s}}$$

and

$$\frac{\mathcal{A}_{D,\lambda}}{\sqrt{\lambda}} + 1 \leq \tilde{C}_1 \tag{60}$$

with $\tilde{C}_1 := (1+\sqrt{C_0})c^{\frac{-s-1}{4r+2s}}+1$. Now we turn to bound $\mathcal{W}_{D,\lambda}$. If $\eta_{\delta/4} := 2\log(16/\delta)/\sqrt{|D|\lambda} \leq 1$, we have from Lemma 22 in Appendix A that

$$\sqrt{\max\{\mathcal{N}_D(\lambda),1\}} \leq 5\sqrt{\max\{\mathcal{N}(\lambda),1\}}.$$

Then, it follows from (6) and $\lambda = (c/|D|)^{1/(2r+s)}$ that

$$(\sqrt{|D|\lambda}+9)\sqrt{\max\{\mathcal{N}_D(\lambda),1\}} \leq 5\max\{c^{1/(4r+2s)},9\}\max\{C_0^{1/2}c^{-s/(4r+2s)},1\}|D|^{\frac{2r+2s-1}{4r+2s}}.$$

Thus, it follows from (15) that

$$\mathcal{W}_{D,\lambda} \leq \tilde{C}_2'|D|^{-\frac{r}{2r+s}}, \tag{61}$$

where

$$\tilde{C}_2' := 16\sqrt{2}(\kappa^2+\kappa+1)(\kappa M+\gamma)\left(5(c^{-1/(2r+s)}+1)(c^{1/(4r+2s)}+9)(C_0^{1/2}c^{-s/(4r+2s)}+1)+1\right)^2.$$

If $\eta_{\delta/4} > 1$, we get from Lemma 22 in Appendix A again and $\lambda = (c/|D|)^{1/(2r+s)}$ that with confidence $1-\delta$

$$\sqrt{\max\{\mathcal{N}_D(\lambda),1\}} \leq (1+16c^{-1/(2r+s)})\sqrt{\max\{\mathcal{N}(\lambda),1\}}\log^2\frac{16}{\delta}.$$

The same argument as above shows that with confidence $1-\delta$,

$$\mathcal{W}_{D,\lambda} \leq \tilde{C}_2''|D|^{-\frac{r}{2r+s}}\log^4\frac{16}{\delta}, \tag{62}$$

where

$$\begin{aligned}\tilde{C}_2'' &:= 16\sqrt{2}(\kappa^2+\kappa+1)(\kappa M+\gamma)\\ &\times \left((1+16c^{-1/(2r+s)})(c^{-1/(2r+s)}+1)(c^{1/(4r+2s)}+9)(C_0^{1/2}c^{-s/(4r+2s)}+1)+1\right)^2.\end{aligned}$$

Combining (61) with (62), we obtain with confidence $1-\delta$

$$\mathcal{W}_{D,\lambda} \leq \tilde{C}_2|D|^{-\frac{r}{2r+s}}\log^4\frac{16}{\delta} \tag{63}$$

with $\tilde{C}_2 := \max\{\tilde{C}_2', \tilde{C}_2''\}$.

Step 4: Deriving the learning rate. Plugging (63) and (60) into (57) and (58), we have with confidence $1-\delta$ that

$$A(D,\lambda,\hat{k}) \leq \tilde{C}_3|D|^{-\frac{r}{2r+s}}\log^{10}\frac{16}{\delta} \tag{64}$$

with $\tilde{C}_3 = 3(4(\kappa^2+\kappa)^2\tilde{C}_1^2+1)\tilde{C}_2 + 4(4(\kappa^2+\kappa)^2\tilde{C}_1^2+1)^2\|h_\rho\|\big(c^{\frac{r}{2r+s}} + (4r-2)\kappa^{2r-1}c^{\frac{1}{4r+2s}}\big)$ and

$$S(D,\lambda,\hat{k}) \leq \tilde{C}_4|D|^{-\frac{r}{2r+s}}\log^9\frac{16}{\delta} \tag{65}$$

with $\tilde{C}_4 = \sqrt{2}(2(\kappa^2+\kappa)\tilde{C}_1+1)\big[(2r+1)\tilde{C}_2 + 2(2(\kappa^2+\kappa)^2\tilde{C}_1^2+1)c^{\frac{r}{2r+s}}\|h_\rho\|_\rho + (8r-4)\kappa^{2r-1}\|h_\rho\|_\rho c^{\frac{1}{4r+2s}}\big]$. Putting (64) and (65) into (49), we have

$$\|f_{D,\lambda}^{(\hat{k})} - f_\rho\|_\rho \leq (\tilde{C}_3+\tilde{C}_4)|D|^{-\frac{r}{2r+s}}\log^{10}\frac{16}{\delta}.$$

This completes the proof of Theorem 6 with the constant $C := \tilde{C}_3 + \tilde{C}_4$. ■

6. Empirical Studies

In this section, we report experimental results to study the behavior of BKRR and the adaptive stopping rule (16) in practice. We consider two regression problems. For the j-th regression problem ($j = 1, 2$), we assume that training examples are independently drawn from the regression model $y_i = g_j(x_i) + \xi_i, i = 1, \ldots, |D|$, where $\{x_i\}_{i=1}^{|D|}$ are drawn from the uniform distribution on the (hyper)-cube $[0,1]^{d_j}$ (d_j is the input dimension) and $\{\xi_i\}_{i=1}^{|D|}$ are noise components independently drawn from the Gaussian distribution $\mathcal{N}(0, 1/5)$. For the j-th problem, we build the estimator by applying BKRR in the RKHS induced by a Mercer kernel K_j. We consider the following two regression functions

$$g_1(x) = \min(x, 1-x), \quad x \in [0,1],$$

$$g_2(x) = (1-\|x\|_2)_+^6(35\|x\|_2^2 + 18\|x\|_2 + 3), \quad x \in [0,1]^3.$$

The two Mercer kernels are $K_1 : \mathbb{R}\times\mathbb{R} \mapsto \mathbb{R}, K_2 : \mathbb{R}^3\times\mathbb{R}^3 \mapsto \mathbb{R}$ defined by

$$K_1(x,\tilde{x}) = 1 + \min(x,\tilde{x}) \quad \text{and} \quad K_2(x,\tilde{x}) = g_3(x-\tilde{x}),$$

where $g_3(x) = (1-\|x\|_2)_+^4(4\|x\|_2+1), x \in [0,1]^3$. It can be found in Chang et al. (2017) that $g_1 \in \mathcal{H}_{K_1}$ with exponent $r = 1/2$ in (5) and $g_2 \in \mathcal{H}_{K_2}$ with exponent $r > 1/2$. We repeat each experiment 40 times and report the average of these experimental results.

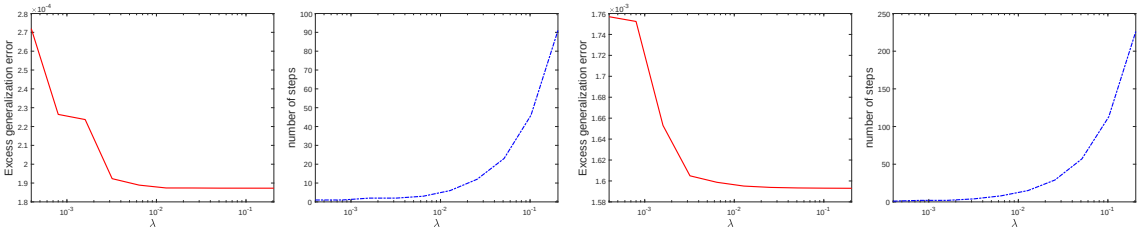
Our numerical results are divided into three parts. In the first part, we study the relation between the generalization ability of BKRR and the regularization parameter to verify our motivation to combine KRR with boosting. In the second part, we validate the empirical behavior of BKRR and its comparison with iterated Tikhonov regularization (ITR). In the last part we show the effectiveness of adaptive stopping rule (16) in practical regression problems.

6.1 Regularization Parameters in BKRR

A common consensus in boosting theory (Friedman, 2001) is that the weak learners should be under-fitting and the boosting iteration will reduce the bias and increase the variance accordingly. For estimators with high-level under-fitting, more boosting iterations are imposed and lead to a similar learning performance as other efficient algorithms. If KRR is

used to build up a weak learner, this argument then shows that we should select a relatively large λ, larger than some value, so that BKRR with a suitable number of iterations can reach a similar learning performance as KRR. Our first simulation is to verify this argument and show a relation between the generalization ability and regularization parameters in BKRR.

In this simulation, We traverse the regularization parameter λ over the set $0.0002 \times \{1, 2, 2^2, \ldots, 2^{10}\}$. For each regularization parameter, we run BKRR until k reaches 150 for $f_\rho = g_1$ and 300 for $f_\rho = g_2$, respectively. For each considered k and λ, we estimate the excess generalization error (EGE) $\mathcal{E}(f_{D,\lambda}^{(k)}) - \mathcal{E}(f_\rho)$ by $\frac{1}{2000}\sum_{i=1}^{2000}[f_{D,\lambda}^{(k)}(x_i') - f_\rho(x_i')]^2$, where $\{x_i'\}_{i=1}^{2000}$ are independently drawn from the uniform distribution on the corresponding input space. For each λ, we record the optimal k (selected to be optimal to the test data directly) and the corresponding EGE. Figure 1 reports EGEs and iteration numbers versus regularization parameters. It is shown in Figure 1 that if λ is larger than some value (near 10^{-2} in this simulation), then BKRR with different λ possesses similar learning performances provided the number of iterations is appropriately selected. Figure 1 also shows that the more high-level of under-fitting, the more boosting iterations required, which verifies the previous common consensus. All these results show that using KRR to build up a weak learner for boosting is reasonable and the selection of λ does not affect the generalization ability very much.



(a) $f_\rho = g_1$. (b) $f_\rho = g_2$.

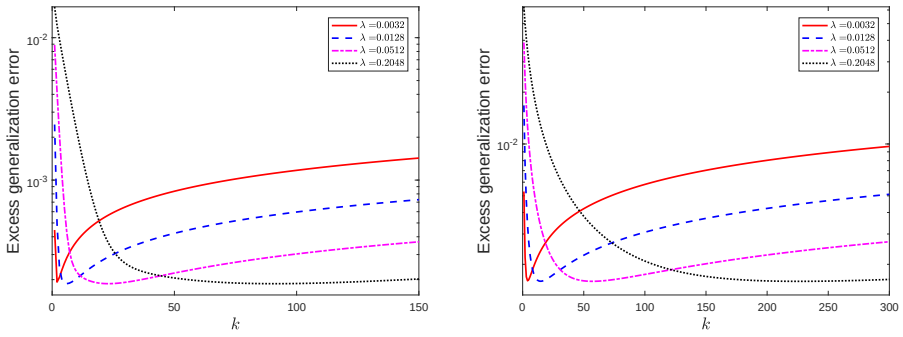
Figure 1: Excess generalization errors of BKRR versus regularization parameters. We also plot the iteration number at which the optimal EGEs are achieved.

6.2 Behavior of BKRR

In this subsection, we want to validate the empirical behavior of BKRR and its comparison with ITR. It should be mentioned that BKRR focuses on fixed regularization parameters and varying iteration numbers, while ITR focuses on fixed iteration numbers and varying regularization parameters. The aim of this simulation is to verify an advantage of BKRR over ITR in the parameter selection, showing that selecting an appropriate number of iterations in BKRR is easier than selecting an appropriate regularization parameter in ITR.

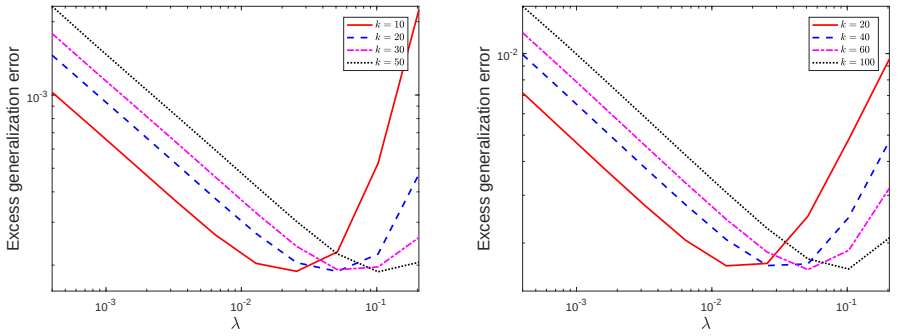
We first study how BKRR would behave along the iterations. In the first experiment, we aim to study how EGEs would change as a function of iteration numbers. We fix regularization parameters $\lambda \in \{0.0032, 0.0128, 0.0512, 0.2048\}$, and show in Figure 2 EGEs versus the iteration number for two regression problems. From Figure 2 we see that EGEs

would typically decrease first as k increases from 1 to some number, after which it increases slowly as a function of k. To be detailed, the increasing curve behaves as a concave function with respect to k, which validates our arguments in Theorem 3 and Theorem 4 that the variance increases with exponentially diminishing terms as k increases. Furthermore, for some large λ ($\lambda = 0.2048$ for example), BKRR shows a rather stable relationship between the generalization performance and iteration numbers. This is consistent with our theoretical findings in Theorem 3 and Theorem 4. We can also see clearly that the iteration number k at which EGEs achieve the minimal value would increase as λ increases.



(a) $f_\rho = g_1$. (b) $f_\rho = g_2$.

Figure 2: Excess generalization errors versus the number of iterations for different regularization parameters. We consider four λ and two regression problems with the regression function being g_1 and g_2 in panel (a) and panel (b), respectively.



(a) $f_\rho = g_1$. (b) $f_\rho = g_2$.

Figure 3: Excess generalization errors versus regularization parameters for different number of iterations. We consider four k and two regression problems with the regression function being g_1 and g_2 in panel (a) and panel (b), respectively.

We then illustrate the behavior of ITR with fixed k versus regularization parameters. We fix the iteration numbers to some predefined values, and report EGEs of $f_{D,\lambda}^{(k)}$ as a function of regularization parameters in Figure 3. According to Figure 3, we see that EGEs first decrease and then increase as a function of λ. As compared to the stable relationship between EGEs and the number of iterations for BKRR, EGEs change more rapidly with the regularization parameters for ITR. This gives empirical evidence that the selection of iteration number in BKRR is more tractable than the selection of regularization parameters in ITR. It is also clear that the optimal regularization parameter becomes larger as k increases, which is consistent with Theorem 1.

In Table 1, we record the optimal EGEs achieved by BKRR over all iterates for different regularization parameters. We list in the first row the considered regularization parameters. In the second and third rows, we report the optimal EGEs achieved by BKRR with the fixed regularization parameters on two regression problems. In Table 2, we record the optimal EGEs achieved by ITR over all regularization parameters for different iteration numbers. We list in the first and third rows the considered number of iterations. In the second and fourth rows, we report the optimal EGEs achieved by ITR with fixed iteration number on two regression problems. It can be found that BKRR can achieve similar accuracies to ITR with a much easier parameter-selection strategy.

λ	0.0004	0.0008	0.0016	0.0032	0.0064	0.0128	0.0256	0.0512	0.1024	0.2048
g_1	2.72e–4	2.26e–4	2.24e–4	1.92e–4	1.89e–4	1.87e–4	1.87e–4	1.87e–4	1.87e–4	1.87e–4
g_2	1.76e–3	1.75e–3	1.65e–3	1.60e–3	1.60e–3	1.60e–3	1.59e–3	1.59e–3	1.59e–3	1.59e–3

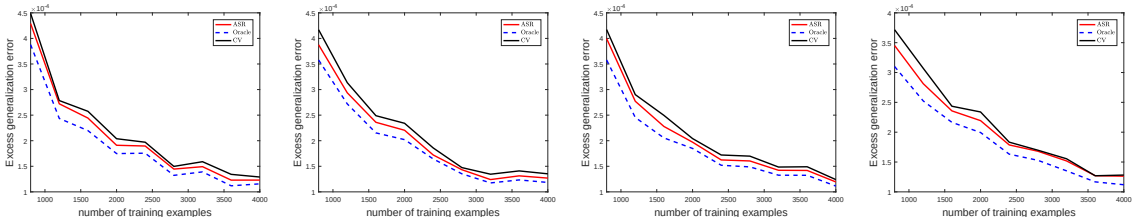
Table 1: Excess generalization errors of BKRR with different regularization parameters.

k	15	30	45	60	75	90	105	120	135	150
g_1	1.91e–4	1.92e–4	1.87e–4	1.92e–4	1.89e–4	1.87e–4	1.88e–4	1.92e–4	1.97e–4	2.02e–4
k	30	60	90	120	150	180	210	240	270	300
g_2	1.59e–3	1.59e–3	1.63e–3	1.59e–3	1.63e–3	1.62e–3	1.60e–3	1.59e–3	1.61e–3	1.63e–3

Table 2: Excess generalization errors of ITR with different iteration numbers.

6.3 BKRR with Early Stopping

In this subsection, we aim to validate the effectiveness of adaptive stopping rule (19) in practical regression problems for some $\theta \in \mathbb{R}_+$. It should be noted that (19) is independent of the confidence level and also different from (16) in the constant term. As discussed in Remark 7, such a modification is reasonable due to the (semi) exponential bias-variance trade-off of BKRR. We apply BKRR to regression problems with different sample sizes ($|D| \in \{800, 1200, 1600, 2000, 2400, 2800, 3200, 3600, 4000\}$) and different regularization parameters ($\lambda \in \{0.016, 0.032, 0.064, 0.128\}$). For each sample size and regularization parameter, we run BKRR with several iterations to get a sequence of candidate models. We record the iteration number $\hat{k}_{\mathrm{ASR}}$ selected by the adaptive stopping rule (ASR) (19) with $\theta = 0.05$, the iteration number $\hat{k}_{\mathrm{CV}}$ selected by the five-fold cross validation (CV) and the iteration number $\hat{k}_{\mathrm{Oracle}}$ with the minimal generalization error over all candidate models.



(a) $\lambda = 0.016$ (b) $\lambda = 0.032$ (c) $\lambda = 0.064$ (d) $\lambda = 0.128$

Figure 4: Excess generalization errors of the model selected by ASR (red color), the model selected by CV (black color) and the best candidate model (blue color) versus the number of training examples. We consider the regression problem $y = g_1(x) + \epsilon$ and four regularization parameters: $\lambda = 0.016$, $\lambda = 0.032$, $\lambda = 0.064$ and $\lambda = 0.128$.

In Figure 4, we fix different regularization parameters and plot the EGEs of $f_{D,\lambda}^{\hat{k}_{\rm ASR}}$, $f_{D,\lambda}^{\hat{k}_{\rm CV}}$ and $f_{D,\lambda}^{\hat{k}_{\rm Oracle}}$ versus the number of training examples for the regression problem with the regression function g_1. According to Figure 4, it is clear that ASR (19) works well in selecting a good model with EGEs comparable to the best candidate model. Furthermore, ASR also behaves slightly better than the CV widely used in practical learning problems. It should be mentioned that the five-fold CV requires the training of an additional five models based on different assignments of validation sets, which can be time-consuming. This repeated training is not required in ASR and therefore ASR requires significantly less computational costs than CV.

Acknowledgments

The authors would like to thank two anonymous referees for their constructive suggestions. The work of Shao-Bo Lin is supported partially by the National Natural Science Foundation of China (Grant Numbers 61876133, 11771021). The work of Yunwen Lei is supported partially by the National Natural Science Foundation of China (Grant No. 61806091) and the Shenzhen Peacock Plan (Grant No. KQTD2016112514355531). The work of Ding-Xuan Zhou is supported partially by the Research Grants Council of Hong Kong [Project No. CityU 11303915] and by National Natural Science Foundation of China under Grant 11461161006. Part of the work was done when the last author visited Shanghai Jiaotong University (SJTU), for which the support from SJTU and the Ministry of Education is greatly appreciated. The corresponding author is Yunwen Lei.

Appendix A. Auxiliary Lemmas

In this appendix, we present some useful lemmas. The first one (Mücke, 2018, Corollary 2.2) (see also Lu et al. (2018)) describes the difference between the effective dimension and its empirical counterpart.

Lemma 21 *For any $0 < \delta < 1$, with confidence $1-\delta$, there holds*

$$(1+4\eta_\delta)^{-1}\sqrt{\max\{\mathcal{N}(\lambda),1\}} \le \sqrt{\max\{\mathcal{N}_D(\lambda),1\}} \le (1+4\max\{\sqrt{\eta_\delta},\eta_\delta^2\})\sqrt{\max\{\mathcal{N}(\lambda),1\}},$$

where $\eta_\delta := 2\log(4/\delta)/\sqrt{|D|\lambda}$.

Based on Lemma 21, we can get the following lemma.

Lemma 22 *Let D be a sample drawn independently according to ρ and $0 < \delta < 1$. With confidence at least $1-\delta$,*

$$\begin{aligned}
\mathcal{Q}_{D,\lambda}^2 &\le \widetilde{\mathcal{Q}}_{D,\lambda} \le 2\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}\log\frac{8}{\delta}}{\sqrt{\lambda}}\right)^2 + 2,\\
\mathcal{R}_D &\le \frac{4\kappa^2}{\sqrt{|D|}}\log\frac{8}{\delta},\\
\mathcal{P}_{D,\lambda} &\le 2(\kappa M+\gamma)\mathcal{A}_{D,\lambda}\log\big(8/\delta\big),\\
(1+4\eta_{\delta/4})^{-1}\sqrt{\max\{\mathcal{N}(\lambda),1\}} &\le \sqrt{\max\{\mathcal{N}_D(\lambda),1\}} \le (1+4\sqrt{\eta_{\delta/4}}\vee\eta_{\delta/4}^2)\sqrt{\max\{\mathcal{N}(\lambda),1\}}
\end{aligned}$$

hold simultaneously.

Proof. From Guo et al. (2017, Proposition 1) and (67) in Appendix B below, there exists a subset $\mathcal{Z}_{\delta,1}^{|D|}$ of $\mathcal{Z}^{|D|}$ with measure at least $1-\delta$ such that for all $D \in \mathcal{Z}_{\delta,1}^{|D|}$

$$\mathcal{Q}_{D,\lambda}^2 \le \widetilde{\mathcal{Q}}_{D,\lambda} \le 2\left(\frac{2(\kappa^2+\kappa)\mathcal{A}_{D,\lambda}\log\frac{2}{\delta}}{\sqrt{\lambda}}\right)^2 + 2.$$

From Yao et al. (2007, Proposition 5.3), there exists a subset $\mathcal{Z}_{\delta,2}^{|D|}$ of $\mathcal{Z}^{|D|}$ with measure at least $1-\delta$ such that for all $D \in \mathcal{Z}_{\delta,2}^{|D|}$

$$\mathcal{R}_D \le \frac{4\kappa^2}{\sqrt{|D|}}\log\frac{2}{\delta}.$$

It also follows from Blanchard and Krämer (2016, Lemma 5.1) that there exists a subset $\mathcal{Z}_{\delta,3}^{|D|}$ of $\mathcal{Z}^{|D|}$ with measure at least $1-\delta$ such that for all $D \in \mathcal{Z}_{\delta,3}^{|D|}$

$$\mathcal{P}_{D,\lambda} \le 2(\kappa M+\gamma)\mathcal{A}_{D,\lambda}\log\big(2/\delta\big).$$

According to Lemma 21, there exists a subset $\mathcal{Z}_{\delta,4}^{|D|}$ of $\mathcal{Z}^{|D|}$ with measure at least $1-\delta$ such that for all $D \in \mathcal{Z}_{\delta,4}^{|D|}$

$$(1+4\eta_\delta)^{-1}\sqrt{\max\{\mathcal{N}(\lambda),1\}} \le \sqrt{\max\{\mathcal{N}_D(\lambda),1\}} \le (1+4\sqrt{\eta_\delta}\vee\eta_\delta^2)\sqrt{\max\{\mathcal{N}(\lambda),1\}}.$$

Thus, for $D \in \mathcal{Z}_{\delta,1}^{|D|} \cap \mathcal{Z}_{\delta,2}^{|D|} \cap \mathcal{Z}_{\delta,3}^{|D|} \cap \mathcal{Z}_{\delta,4}^{|D|}$, the above four inequalities hold simultaneously. Then Lemma 22 follows by scaling δ to $\delta/4$. ■

From Lemma 22, we derive the following estimate.

Lemma 23 *For any $0 < \delta < 1$, with confidence $1 - \delta$, there holds*

$$\mathcal{P}_{D,\lambda}\mathcal{Q}_{D\lambda} \leq \frac{1}{2}\mathcal{W}_{D,\lambda}\log^4\frac{16}{\delta}.$$

Proof. It follows from Lemma 22 and (48) that with confidence $1-\delta$, there holds

$$\begin{aligned}
\mathcal{P}_{D,\lambda}\mathcal{Q}_{D\lambda} &\leq \frac{4\sqrt{2}(\kappa^2+\kappa+1)(\kappa M+\gamma)}{\sqrt{|D|}}\left(\frac{1}{\lambda|D|}+\frac{\sqrt{\mathcal{N}(\lambda)}}{\sqrt{|D|\lambda}}+1\right)\left(\frac{1}{\sqrt{\lambda|D|}}+\sqrt{\mathcal{N}(\lambda)}\right)\log^2\frac{16}{\delta}\\
&\leq \frac{4\sqrt{2}(\kappa^2+\kappa+1)(\kappa M+\gamma)}{\sqrt{|D|}}\left(\frac{(\sqrt{|D|\lambda}+9)\sqrt{\max\{\mathcal{N}_D(\lambda),1\}}}{|D|\lambda}+1\right)\\
&\times\frac{(\sqrt{|D|\lambda}+9)\sqrt{\max\{\mathcal{N}_D(\lambda),1\}}}{\sqrt{|D|\lambda}}\log^4\frac{16}{\delta}.
\end{aligned}$$

This together with (15) completes the proof of Lemma 23. ∎

Our final lemma establishes a relation between the in-sample norm and out-sample norm of functions in $\mathcal{H}_K$.

Lemma 24 *Let $f \in \mathcal{H}_K$. Then*

$$\|f\|_\rho \leq \mathcal{Q}_{D,\lambda}\|L_{K,D}^{1/2}f\|_K + \mathcal{Q}_{D,\lambda}\lambda^{1/2}\|f\|_K. \tag{66}$$

Proof. Since $f \in \mathcal{H}_K$, it follows from the definition of $\mathcal{Q}_{D,\lambda}$ that

$$\begin{aligned}
\|f\|_\rho &= \|L_K^{1/2}f\|_K \leq \|(L_K+\lambda I)^{1/2}f\|_K \leq \mathcal{Q}_{D,\lambda}\|(L_{K,D}+\lambda I)^{1/2}f\|_K\\
&\leq \mathcal{Q}_{D,\lambda}\|L_{K,D}^{1/2}f\|_K + \mathcal{Q}_{D,\lambda}\lambda^{1/2}\|f\|_K.
\end{aligned}$$

This finishes the proof of Lemma 24. ∎

Appendix B. Some Inequalities for Positive Linear Operators

In this part, we recall some basic definitions and properties for linear operators which can be found in Bhatia (2013). Let $\mathcal{H}_1$ and $\mathcal{H}_2$ be two Hilbert spaces. Let $\mathcal{L}(\mathcal{H}_1,\mathcal{H}_2)$ be the space of all bounded linear operators from $\mathcal{H}_1$ to $\mathcal{H}_2$. The adjoint of an operator A is the unique operator A^* in $\mathcal{L}(\mathcal{H}_2,\mathcal{H}_1)$ that satisfies the relation

$$\langle g, Af\rangle_{\mathcal{H}_2} = \langle A^*g, f\rangle_{\mathcal{H}_1}$$

for all $f \in \mathcal{H}_1$ and $g \in \mathcal{H}_2$. For the space $\mathcal{L}(\mathcal{H},\mathcal{H})$, we use the more compact notation $\mathcal{L}(\mathcal{H})$. For $A \in \mathcal{L}(\mathcal{H})$, if $A = A^*$, we then call A a self-adjoint operator. A self-adjoint operator is said to be positive, if $\langle f, Af\rangle_{\mathcal{H}} \geq 0$ for all $f \in \mathcal{H}$. If $\langle f, Af\rangle_{\mathcal{H}} > 0$ for all nonzero f, we say A is strictly positive.

For $A \in \mathcal{L}(\mathcal{H})$, the operator norm is defined by

$$\|A\| = \sup_{\|f\|_{\mathcal{H}}=1}\|Af\|_{\mathcal{H}}.$$

If A is compact and positive, there exists a normalized eigenpairs of A, denoted by $\{(\lambda_i, \varphi_i)\}_{i=1}^{\infty}$, with eigenvectors $\{\varphi_i\}$ forming an orthonormal basis for $\mathcal{H}$ and eigenvalues satisfying $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$. There also holds $\|A\| = \lambda_1$. Write the trace of the positive operator A by

$$\mathrm{Tr}(A) = \sum_{i=1}^{\infty} \lambda_i.$$

The Hilbert-Schmidt norm of A is then defined by

$$\|A\|_{HS} = (\mathrm{Tr}(A^2))^{1/2} = \left(\sum_{j=1}^{\infty} \lambda_j^2 \right)^{1/2}.$$

If $\|A\|_{HS} < \infty$, we then call A a Hilbert-Schmidt operator. If A is Hilbert-Schmidt, it follows from the definition that $\|A\| \leq \|A\|_{HS}$. For $F : \mathbf{R}_+ \cap \{0\} \to \mathbf{R}$, we define the operator

$$F(A) = \sum_{i=1}^{\infty} F(\lambda_i) \varphi_i \otimes \varphi_i = \sum_{i=1}^{\infty} F(\lambda_i) \langle \cdot, \varphi_i \rangle_{\mathcal{H}} \varphi_i$$

by spectral calculus. For positive operators A and B, there holds $\|AB\| = \|BA\|$. We also need the following two important inequalities, which can be found in Lemma 1 and Lemma 4 in Guo et al. (2017) (see also Bhatia (2013, Lemma VII.5.5) for the second one).

Lemma 25 *Let A and B be positive operators. Then for any $0 \leq \tau \leq 1$, there holds*

$$\|A^\tau B^\tau\| \leq \|AB\|^\tau. \tag{67}$$

If in addition A and B are Hilbert-Schmidt and $\max\{\|A\|, \|B\|\} \leq \kappa$, then for arbitrary $\mu \geq 1$, there holds

$$\|A^\mu - B^\mu\|_{HS} \leq \mu \kappa^{\mu-1} \|A - B\|_{HS}. \tag{68}$$

References

F. Bauer, S. Pereverzev, and L. Rosasco. On regularization algorithms in learning theory. *J. Complex.*, 23(1):52–72, 2007.

R. Bhatia. *Matrix Analysis*. Vol. 169, Springer Science & Business Media, 2013.

G. Blanchard and N. Krämer. Convergence rates of kernel conjugate gradient for random design regression. *Anal. Appl.*, 14(06):763–794, 2016.

P. Bühlmann and B. Yu. Boosting with the l_2 loss: regression and classification. *J. Amer. Satis. Assoc.*, 98(462):324–339, 2003.

A. Caponnetto and E. De Vito. Optimal rates for the regularized least-squares algorithm. *Found. Comp. Math.*, 7(3):331–368, 2007.

A. Caponnetto and Y. Yao. Cross-validation based adaptation for regularization operators in learning theory. *Anal. Appl.*, 8(02):161–183, 2010.

X. Chang, S. B. Lin, and D. X. Zhou. Distributed semi-supervised learning with kernel ridge regression. *J. Mach. Learn. Res.*, 18(46): 1-22, 2017.

F. Cucker and D. X. Zhou. *Learning Theory: An Approximation Theory Viewpoint*. Cambridge University Press, 2007.

E. De Vito, V. Umanità, and S. Villa. An extension of Mercer theorem to vector-valued measurable kernels. *Appl. Comput. Harmonic Anal.* 34:339-351, 2013.

H. W. Engl. On the choice of the regularization parameter for iterated tikhonov regularization of ill-posed problems. *J. Approx. Theory*, 49(1):55–63, 1987.

T. Evgeniou, M. Pontil, and T. Poggio. Regularization networks and support vector machines. *Adv. Comput. Math.*, 13(1):1–50, 2000.

Y. Freund. Boosting a weak learning algorithm by majority. *Inform. & Comput.*, 121(2): 256–285, 1995.

J. H. Friedman. Greedy function approximation: a gradient boosting machine. *Ann. Stat.*, 29(5):1189–1232, 2001.

J. H. Friedman, T. Hastie, and R. Tibshirani. Additive logistic regression: a statistical view of boosting. *Ann. Stat.*, 28(2):337–407, 2000.

L. L. Gerfo, L. Rosasco, F. Odone, E. De Vito, and A. Verri. Spectral algorithms for supervised learning. *Neural Comput.*, 20(7):1873–1897, 2008.

Z. C. Guo, S. B. Lin, and D. X. Zhou. Learning theory of distributed spectral algorithms. *Inverse Probl.*, 33(7):074009, 2017.

Z.-C. Guo, D. H. Xiang, X. Guo, and D. X. Zhou. Thresholded spectral algorithms for sparse approximations. *Anal. Appl.*, 15(03):433–455, 2017b.

L. Györfi, M. Kohler, A. Krzyzak, and H. Walk. *A Distribution-Free Theory of Nonparametric Regression*. Springer, Berlin, 2002.

M. Hanke and C. W. Groetsch. Nonstationary iterated tikhonov regularization. *J. Optimiz. Theory Appl.*, 98(1):37–53, 1998.

Q. Jin and L. Stals. Nonstationary iterated tikhonov regularization for ill-posed problems in banach spaces. *Inverse Probl.*, 28(10):104011, 2012.

Q. N. Jin and Z. Y. Hou. On the choice of the regularization parameter for ordinary and iterated tikhonov regularization of nonlinear ill-posed problems. *Inverse Probl.*, 13(3): 815, 1997.

J. T. King and D. Chillingworth. Approximation of generalized inverses by iterated regularization. *Numer. Funct. Anal. Optim.*, 1(5):499–513, 1979.

S. B. Lin and D. X. Zhou. Distributed kernel-based gradient descent algorithms. *Constr. Approx.*, 47(2):249–276, 2018.

S. B. Lin and D. X. Zhou. Optimal learning rates for kernel partial least squares. *J. Fourier Anal. Appl.*, 24(3):908–933, 2018.

S. B. Lin, X. Guo, and D. X. Zhou. Distributed learning with regularized least squares. *J. Mach. Learn. Res.*, 18(92):1–31, 2017.

S. Lu, P. Mathé, and S. Pereverzyev. Balancing principle in supervised learning for a general regularization scheme. *Appl. Comput. Harmon. Anal.*, In Press. 2018

N. Mücke. Adaptivity for Regularized Kernel Methods by Lepskii's Principle. arXiv preprint arXiv:1804.05433, 2018.

B. U. Park, Y. K. Lee, and S. Ha. L_2 boosting in kernel regression. *Bernoulli*, 15(3): 599–613, 2009.

G. Raskutti, M. Wainwright, and B. Yu. Early stopping and non-parametric regression: an optimal data-dependent stopping rule. *J. Mach. Learn. Res.*, 15(1):335–366, 2014.

R. E. Schapire. The strength of weak learnability. *Mach. Learn.*, 5(2):197–227, 1990.

L. Shi, Y. L. Feng, and D. X. Zhou. Concentration estimates for learning with l_1-regularizer and data dependent hypothesis spaces. *Appl. Comput. Harmonic Anal.*, 31(2):286–302, 2011.

S. Smale and D. X. Zhou. Shannon sampling and function reconstruction from point values. *Bull. Amer. Math. Soc.*, 41(3):279–305, 2004.

S. Smale and D. X. Zhou. Learning theory estimates via integral operators and their approximations. *Constr. Approx.*, 26(2):153–172, 2007.

I. Steinwart, D. Hush, and C. Scovel. Optimal rates for regularized least squares regression. In S. Dasgupta and A. Klivan, editors, *Annual Conference on Learning Theory*, pages 79–93, 2009.

I. Steinwart and C. Scovel. Mercer's theorem on general domains: On the interaction between measures, kernels, and RKHSs. *Constr. Approx.*, 35(3):363-417, 2012.

Q. Wu. Bias corrected regularization kernel network and its applications. In *International Joint Conference on Neural Networks*, pages 1072–1079, 2017.

Y. Yao, L. Rosasco, and A. Caponnetto. On early stopping in gradient descent learning. *Constr. Approx.*, 26(2):289–315, 2007.

Y. Ying and D. X. Zhou. Unregularized online learning algorithms with general loss functions. *Appl. Comput. Harmonic Anal.*, 2(42):224–244, 2017.