Fast Learning With Polynomial Kernels

Shaobo Lin and Jinshan Zeng

***Abstract*—This paper proposes a new learning system of low computational cost, called fast polynomial kernel learning (FPL), based on regularized least squares with polynomial kernel and subsampling. The almost optimal learning rate as well as the feasibility verifications including the subsampling mechanism and solvability of FPL are provided in the framework of learning theory. Our theoretical assertions are verified by numerous toy simulations and real data applications. The studies in this paper show that FPL can reduce the computational burden of kernel methods without sacrificing its generalization ability very much.**

***Index Terms*—Kernel methods, learning systems, learning theory, polynomial kernel.**

I. Introduction

RAPID development in the data generation and acquisition technique brings data of unprecedented size, which abounds around our lives in terms of the high-frequency financial data, microarray, text, video, longitudinal data, Internet data, network data, among others. Understanding as well as handling massive data becomes a hot topic in the machine learning community and recently has triggered enormous research activities [4], [7], [14], [17], [30], [35], [36], [40]. The main challenge for this purpose is to reduce the computational burden and memory requirement of existing learning systems without degrading their generalization abilities.

Due to the prominent performance in applications [26] and theory [5], kernel approach [9] is one of the most popular tool to generate learning systems of high quality, such as the support vector machines (SVMs) [39] for classification and regularized least squares (RLSs) [14] for regression. Recall that learning systems generated by kernel approach require computational complexity at least $\mathcal{O}(m^2)$ [sometimes $\mathcal{O}(m^3)$] with m the number of samples. Moreover there are two parameters including the kernel and regularization parameters that need tuning in the learning process of these systems. Kernel approach are generally not suitable to tackle the massive data. To overcome these challenges, several scalable variants, such as the localized SVMs [17], [25], distributed learning [14], [15], [36] and learning with subsampling [16], [21], [22], have been developed to produce scalable learning systems based on kernel approach. All these variants are justified to maintain the prominent generalization abilities of the original learning systems based on kernel approach, but can significantly reduce its computational burden. The magic behind the success of these strategies is the use of an extra parameter, like the number of data blocks in distributed learning [36], the number of partition in localized SVMs [17] and the size of subsamples in learning with subsamples [22], to balance the generalization ability and computational burden. Existing theoretical results in [3], [14], [15], [17], [21], [22], and [36] show that the learning performance of these variants is not sensitive to the introduced parameters in the sense that the optimal learning rates can be derived for a great number of candidates of these parameters. However, the ranges of these candidates generally depend on some *a-priori* knowledge of the data like the smoothness of the regression function [5], which is commonly either unknown or not easy to verify.

In this paper, we attempt to propose a more efficient and practical learning system, of which the parameters can be determined easily and the generalization ability is not degraded so much. The basic idea is to incorporate the size of subsampling into the kernel parameter by using some intrinsic features of the polynomial kernel. To be detailed, let $K_s(x, x') := (1 + x \cdot x')^s$ be the polynomial kernel with kernel parameter s and $\mathcal{H}_s$ be the reproducing kernel Hilbert space (RKHS) associated with K_s endowed with norm $\|\cdot\|_s$. It is obvious that $\mathcal{H}_s$ is an n-dimensional linear space with $n = \binom{s+d}{s}$, where d denotes the dimension of the input space. Note that the regularization parameter and kernel parameter in kernel approach are used to reflect the bias-variance tradeoff via controlling the capacity of the hypothesis space, i.e., the set of all candidates of potential estimators in the learning process. An intuitive observation is: if we can show that the kernel parameter of the polynomial kernel, s, plays a leading role in the polynomial kernel regression, then it might be sufficient to produce a desired estimator via subsampling only n columns ($n \leq m$) of the kernel matrix, and remove the regularization term.

Motivated by the above observations, we develop a novel learning system, called fast polynomial kernel learning (FPL), mainly to reduce the computational burden of kernel approach. Two distinguishing features of FPL are: 1) the regularization

Manuscript received December 7, 2017; revised April 3, 2018 and May 26, 2018; accepted June 21, 2018. The work of S. Lin was supported in part by the National Natural Science Foundation of China under Grant 61502342 and Grant 11771012, and in part by the State Key Laboratory of Robotics (2018-O05). The work of J. Zeng was supported in part by NNSFC under Grant 61603162, Grant 11501440, Grant 61772246, and Grant 61603163, and in part by the Doctoral Start-Up Foundation of Jiangxi Normal University. This paper was recommended by Associate Editor Y. Tan. *(Corresponding author: Jinshan Zeng.)*

S. Lin is with the Department of Mathematics, Wenzhou University, Wenzhou 325035, China, and also with the State Key Laboratory of Robotics, Shenyang Institute of Automation, Chinese Academy of Sciences, Shenyang 110016, China.

J. Zeng is with the School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China (e-mail: jsh.zeng@gmail.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TCYB.2018.2850819

parameter used in the classical kernel approach is not required, thus there is only one parameter, i.e., the size of subsampling n and 2) n can be explicitly determined by $n = \binom{s+d}{s}$, where d is the dimension of the input space and s is the degree of the used polynomial kernel, which is discrete and generally much easier to be tuned from a deterministic range. Once the partial kernel matrix with n columns is selected, the estimator is derived directly by the pseudo-inverse technique. Moreover, we present some subsampling mechanisms with feasibility verifications to realize the optimal learning performance of FPL. According to the procedure of FPL, the computational complexity of FPL is $\mathcal{O}(n^2m)$, which is much smaller than $\mathcal{O}(m^3)$, especially when $n \ll m$. Since the parameter s is crucial for FPL, we suggest an upper bound of s, and show that it suffices to use the well-known *cross-validation* method [12, Ch. 8] or *hold-out* method [12, Ch. 7] to choose an s from this interval, with excellent performance of FPL.

The prominent performance of FPL is verified by both theoretical analysis and numerical experiments. Theoretically, the new learning system is proved to be an almost optimal strategy provided the regression function is smooth in the framework of learning theory [5]. Furthermore, it is also shown that the pseudo-inverse technique can realize the almost optimality. Numerically, both toy simulations and experiments of UCI datasets imply that FPL is more efficient than the classical kernel approaches, in the sense that it can significantly reduce the computational burden without degrading the generalization capability very much.

The rest of this paper is organized as follows. In the next section, we present the new learning system based on polynomial kernels. In Section III, we study the theoretical behaviors of the new method. In Section IV, we compare our results with some related work. In Section V, a series of numerical experiments are provided to show the effectiveness of the proposed method and also verify the developed theoretical results. Section VI presents the proofs of our main results, and we conclude this paper in Section VII.

II. Fast Learning With Polynomial Kernels

In this section, we develop the new learning system following the motivations and feasibility verifications.

A. Motivations

Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be the set of samples with $x_i \in \mathbf{B}^d$, the unit ball of $\mathbf{R}^d$ and $y_i \in [-M, M]$ for some positive number M. For a given Mercer kernel K, the RLSs based on kernel approach is defined by

$$f_{\mathbf{z},\lambda} = \arg\min_{f\in\mathcal{H}_K} \left\{ \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 + \lambda \|f\|_K^2 \right\} \tag{1}$$

where $(\mathcal{H}_K, \|\cdot\|_K)$ is the RKHS associated with K. The regularization term in (1) is introduced to avoid over-fitting in the sense that the derived estimator fits the training samples very well but fails to predict other points. If $\mathcal{H}_K$ is an infinite dimensional space, then the regularization term is necessary to guarantee the uniqueness of the solution to (1) via controlling its RKHS norm. However, if $\mathcal{H}_K$ is a finite dimensional space, one can control the dimension of $\mathcal{H}_K$ via the kernel parameter to reflect the tradeoff between bias and variance. In particular, for the polynomial kernel, [39] found that the regularization parameter in (1) should decrease exponentially fast with m. Furthermore, [27] and [12, Th. 11.3] showed that, as far as the learning rate is concerned, the regularization parameter in (1) can decrease arbitrarily fast for a suitable degree of polynomial kernel. An extreme case is that the regularization parameter in (1) associated with the polynomial kernel K_s can be vanished. Thus, the empirical risk minimization on $\mathcal{H}_s$

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_s} \left\{ \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 \right\} \tag{2}$$

can realize the optimal performance of algorithm (1) with $\mathcal{H}_K = \mathcal{H}_s$. Noting that for arbitrary s, $\mathcal{H}_s = \mathcal{P}_s^d$, we rewrite (2) with $\mathcal{H}_K = \mathcal{H}_s$ as

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{P}_s^d} \left\{ \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 \right\} \tag{3}$$

where $\mathcal{P}_s^d$ denotes the set of algebraic polynomials defined on $\mathbf{B}^d$ of degree at most s.

It is well-known that over-fitting is usually caused by two factors. The one is the algorithm-based factor, such as ill-condition of the kernel matrix and the other is the model-based factor like too large capacity of the hypothesis space. When the polynomial kernel is utilized, the main task of the regularization term is to assure that a simple matrix-inverse technique can finish the learning task, since the capacity of the hypothesis space can be controlled by tuning the kernel parameter s, which is essentially different from other kernels, such as the Sobolev kernel [5], [37] and Gaussian kernel [8], [38]. Since the dimension of $\mathcal{P}_s^d$ is $n = \binom{s+d}{s}$, if we select $\{\eta_i\}_{i=1}^n \subset \mathbf{B}^d$ such that $\{(1+\eta_i\cdot x)^s\}_{i=1}^n$ is a linear independent system, then

$$\mathcal{P}_s^d = \left\{ \sum_{i=1}^n c_i(1+\eta_i\cdot x)^s : c_i \in \mathbf{R} \right\} =: \mathcal{H}_{\eta,n}. \tag{4}$$

In this way, (3) can be converted to

$$f_{\mathbf{z}} = \arg\min_{f\in\mathcal{H}_{\eta,n}} \left\{ \frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 \right\}. \tag{5}$$

To verify the feasibility of (5), there are two things should be investigated. One is how to determine $\{\eta_i\}_{i=1}^n$ such that (4) holds, and the other is how to guarantee the nonsingularity of the selected matrix $A_{m,n} := ((1 + x_i\cdot\eta_j)^s)_{i,j=1}^{m,n}$.

B. Feasibility Verifications

To present the selection strategy of $\{\eta_i\}_{i=1}^n$, we introduce the conceptions of Haar space and fundamental system [34]. Let $C(\mathbf{B}^d)$ be the space of continuous functions defined on $\mathbf{B}^d$ endowed with the uniform norm, and $V \subset C(\mathbf{B}^d)$ be an N-dimensional linear space. V is called a Haar space of dimension N if for arbitrary distinct points $x_1, \ldots, x_N \in \mathbf{B}^d$ and arbitrary $f_1, \ldots, f_N \in \mathbf{R}$ there exists exactly one function $s \in V$ with $s(x_i) = f_i$, $1 \le i \le N$. The following Lemma 1 [34, Th. 2.2] shows some important properties of Haar space.

Lemma 1: The following statements are equivalent.

1) V is an N-dimensional Haar space.
2) Every $u \in V/\{0\}$ has at most $N-1$ zeros.
3) For any distinct points $x_1, \ldots, x_N \in \mathbf{B}^d$ and any basis $u_1, \ldots, u_N$ of V, the matrix $(u_j(x_i))_{i,j=1}^N$ is nonsingular.

Apparently, if we find a set of points in $\mathbf{B}^d$, $\{\eta_i\}_{i=1}^n$, such that $\mathcal{H}_{\eta,n}$ is the Haar space of dimension $n+1$, then it follows from Lemma 1 that all above problems can be resolved. However, for $d \geq 2$, this conjecture does not hold [34, Th. 2.3].

Lemma 2: Suppose $d \geq 2$. Then there does not exist Haar space on $\mathbf{B}^d$ of dimension $N \geq 2$.

Instead, for $d \geq 2$, we introduce the fundamental system with respect to the polynomial kernel K_s.

Definition 1: Let $\zeta := \{\zeta_i\}_{i=1}^n \subset \mathbf{B}^d$. ζ is called a K_s-fundamental system if

$$\dim \mathcal{H}_{\zeta,n} = \binom{s+d}{s}.$$

From the above definition, it is easy to see that an arbitrary K_s-fundamental system implies (4). The following proposition reveals that almost all set with $n = \binom{d+s}{s}$ points is a K_s-fundamental system.

Proposition 1: Let $s, n \in \mathbf{N}$ and $n = \binom{d+s}{s}$. Then the set

$$\left\{\zeta = (\zeta_i)_{i=1}^n : \dim \mathcal{H}_{\zeta,n} < n\right\}$$

has Lebesgue measure 0.

Based on Proposition 1, we can design simple strategies to choose the centers $\{\eta_j\}_{j=1}^n$. In particular, $\{\eta_j\}_{j=1}^n$ can be selected either deterministically on $\mathbf{B}^d$ or randomly Independent and identically distributed (i.i.d.) according to the uniform distribution, since the uniform distribution is continuous with respect to the Lebesgue measure. Then the following relation holds almost surely:

$$\mathcal{P}_s^d = \left\{\sum_{i=1}^n c_i (1 + \eta_i \cdot x)^s : c_i \in \mathbf{R}\right\}.$$

Based on these selection strategies of $\{\eta_j\}_{j=1}^n$, we prove the nonsingularity of the matrix $A_{m,n}$ in the following proposition.

Proposition 2: Let $s, m, n \in \mathbf{N}$. If $\{x_i\}_{i=1}^m$ are i.i.d. random variables drawn according to arbitrary distribution μ and $\{\eta_j\}_{j=1}^n$ is a K_s-fundamental system, then for arbitrary vector $c = (c_1, \ldots, c_n)$, there exists a set of positive number $\{a_i\}$ such that

$$\begin{aligned} C_1 \int_{\mathbf{B}^d} \left(\sum_{j=1}^n c_j K_s(\eta_j, x)\right)^2 \frac{dx}{\sqrt{1-|x|^2}} \\ \leq \sum_{i=1}^m a_i \left(\sum_{j=1}^n c_j K_s(\eta_j, x_i)\right)^2 \\ \leq C_2 \int_{\mathbf{B}^d} \left(\sum_{j=1}^n c_j K_s(\eta_j, x)\right)^2 \frac{dx}{\sqrt{1-|x|^2}} \end{aligned}$$

holds with probability at least $1 - (C_3 n/m)$, where C_3 is a constant depending only on d.

It can be easily deduced from Proposition 2 that with probability at least $1 - (C_3 n/m)$, the matrix $A_{m,n}$ is nonsingular.

Algorithm 1 FPL

Input: Let $\{(x_i, y_i)\}_{i=1}^m$ be a sample set, and $s \in \mathbf{N}$ be the degree of polynomial kernel $K_s(x, x') = (1 + x \cdot x')^s$.

Step 1: Let $n = \binom{s+d}{s}$ be the number of centers and $\{\eta_j\}_{j=1}^n$ be the set of centers, which is a K_s fundamental system. Set $A_{m,n} := (K_s(x_i, \eta_j))_{i,j=1}^{m,n}$, $\mathbf{y} = (y_1, \ldots, y_m)^T$.

Step 2: Set $\mathbf{c} = \text{pinv}(A_{m,n})\mathbf{y} = (c_1, \ldots, c_n)^T$, where $\text{pinv}(A_{m,n})$ denotes the pseudo-inverse operator in MATLAB.

Output: $f_{\mathbf{z},s}(x) = \sum_{j=1}^n c_j K_s(x, \eta_j)$.

Indeed, if $A_{m,n}$ is singular, then it follows from Proposition 2 that there exists a nontrivial set $\{c_j\}_{j=1}^n$ such that:

$$\int_{\mathbf{B}^d} \left(\sum_{j=1}^n c_j K_s(\eta_j, x)\right)^2 \frac{dx}{\sqrt{1-|x|^2}} = 0.$$

This implies

$$\sum_{j=1}^n c_j K_s(\eta_j, x) = 0, \ \ x \in \mathbf{B}^d$$

which is impossible since $\{\eta_j\}$ is a K_s-fundamental system. As n is generally much smaller than m, the probability to guarantee the nonsingularity of the matrix is close to 1. As a consequence, Propositions 1 and 2 verify the feasibility of (5).

C. Learning Algorithm

Inspired by the above two theorems, we present an efficient learning system based on (5), called FPL as shown in Algorithm 1.

According to Proposition 1, arbitrary n points on $\mathbf{B}^d$ build up a K_s-fundamental system almost surely. We suggest to draw $\{\eta_j\}_{j=1}^n$ independently according to the uniform distribution. Our theoretical analysis in Section III and numerical results in Section V will show that the learning performance of FPL is generally independent of the selection of $\{\eta_j\}_{j=1}^n$.

It can be found in Algorithm 1 that there is only one parameter s in FPL. To determine s, we use the *cross-validation* method [12, Ch. 8] or *hold-out* method [12, Ch. 7]. For the sake of brevity, we only provide theoretical assessments of the later one. It should be noted that for the *cross-validation* method, we can also derive a similar learning rate by using the similar method of this paper. Specifically, there are three steps to implement the *hold-out* strategy: 1) splitting the sample set into two independent subsets $\mathbf{z}_1$ with cardinality m_1 and $\mathbf{z}_2$ with cardinality m_2; 2) using $\mathbf{z}_1$ to build the sequence $\{f_{\mathbf{z},s}\}_{s=1}^{\lceil m^{1/d} \rceil}$; and 3) using $\mathbf{z}_2$ to select a proper value of s and thus yield the final estimator $f_{\mathbf{z}}$, where $\lceil a \rceil$ denotes the largest integer no more than $a \in \mathbf{R}$.

Now, let us explain how to determine the parameter s via $\mathbf{z}_2$. Given the set $\Xi = \{0, 1, 2, \ldots, \lceil m^{1/d} \rceil\}$, s is determined according to the following data-driven way:

$$s^* = \arg\min_{s \in \Xi} \frac{1}{m_2} \sum_{z_i \in \mathbf{z}_2} \left(\pi_M f_{\mathbf{z}_1,s}(x_i) - y_i\right)^2 \tag{6}$$

where $\pi_M t$ denotes the clipped value of t at $\pm M$, that is, $\pi_M t \coloneqq \min\{M, |t|\}\text{sgn}(t)$ and $\text{sgn}(t)$ is the sign of t.

III. Theoretical Behavior

We analyze the learning performance of FPL in the framework of learning theory [5]. Let $\mathbf{z} = \{(x_i, y_i)\}_{i=1}^m$ be drawn independently and identically according to an unknown distribution $\rho \coloneqq \rho_X(x)\rho(y|x)$ defined on $Z \coloneqq X \times Y$ with $X = \mathbf{B}^d$ and $Y = [-M, M]$. Define

$$\mathcal{E}(f) \coloneqq \int_Z (f(x) - y)^2 d\rho$$

be the generalization error of an estimator f. It is well known that $\mathcal{E}(f)$ is minimized by the regression function (see [5])

$$f_\rho(x) \coloneqq \int_Y y d\rho(y|x).$$

Let $L^2_{\rho_X}$ be the Hilbert space of ρ_X square integrable function on X, endowed with norm $\|\cdot\|_\rho$. With the assumption $f_\rho \in L^2_{\rho_X}$, it is easy to deduce

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2. \tag{7}$$

Therefore, the goal of learning is to find the best approximation of the regression function f_ρ within a hypothesis space $\mathcal{H}$. Let $f_{\mathbf{z}} \in \mathcal{H}$ be an estimator based on $\mathbf{z}$ and $f_{\mathcal{H}} \in \mathcal{H}$ be the best approximation of f_ρ, i.e., $f_{\mathcal{H}} \coloneqq \arg\min_{g\in\mathcal{H}} \|f - g\|_\rho$. A preferable way in learning theory to bound $\mathcal{E}(f) - \mathcal{E}(f_\rho)$ is to divide it into bias and variance

$$\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}\big(f_\rho\big) = \mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}\big(f_\rho\big) + \mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}}). \tag{8}$$

It is well known [5], [12] that a small $\mathcal{H}$ will derive a large bias $\mathcal{E}(f_{\mathcal{H}}) - \mathcal{E}(f_\rho)$, while a large $\mathcal{H}$ will deduce a large variance $\mathcal{E}(f_{\mathbf{z}}) - \mathcal{E}(f_{\mathcal{H}})$. The best hypothesis space $\mathcal{H}^*$ is obtained when the best comprise between the conflicting requirements of small bias and small variance is achieved. This is the well known "bias-variance" dilemma. Previous studies [8], [10], [13], [23], [24], [28], [31], [39] on (1) show that $\mathcal{H}^*$ can be obtained by tuning the regularization parameter as well as the kernel parameter and (almost) optimal learning rates can be obtained. In this section, we shall prove that such $\mathcal{H}^*$ can also be obtained via tuning s only for (5) in terms of establishing the same (almost) optimal learning rates.

To achieve this purpose, we introduce the measurement and *a-priori* information of f_ρ at first. It was shown in [6] and [27] that formulating the learning problem in terms of probability estimates may bring additional advantages, compared with the classical expectation estimates. For example, some phase-transition phenomenon can be observed in the former one. To this end, we present a formal way to measure the performance of learning schemes in probability. Let $\Theta \subset L^2_{\rho_X}$ and $\mathcal{M}(\Theta)$ be the class of all Borel measures ρ on Z such that $f_\rho \in \Theta$. For each $\varepsilon > 0$, we enter into a competition over all possible estimators based on m samples $\Phi_m : \mathbf{z} \mapsto f_{\mathbf{z}}$ by [6]

$$\mathbf{AC}_m(\Theta, \varepsilon) \coloneqq \inf_{f_{\mathbf{z}}\in\Phi_m} \sup_{\rho\in\mathcal{M}(\Theta)} \mathbf{P}\Big\{\mathbf{z} : \|f_\rho - f_{\mathbf{z}}\|_\rho^2 > \varepsilon\Big\}.$$

Let $\mathbf{k} = (k_1, k_2, \ldots, k_d)$, $k_i \in \mathbf{N}$, and define the derivative

$$D^{\mathbf{k}} f(x) \coloneqq \frac{\partial^{|\mathbf{k}|} f}{\partial^{k_1} x_1 \cdots \partial^{k_d} x_d}$$

where $|\mathbf{k}| \coloneqq k_1 + \cdots + k_d$. The classical Sobolev class is then defined for any $r \in \mathbf{N}$ by

$$W^r \coloneqq \left\{f : \max_{0\le|\mathbf{k}|\le r} \left\|D^{\mathbf{k}} f\right\|_\infty < \infty, r \in \mathbf{N}\right\}.$$

The following theorem is the first main result of this paper, showing that with appropriately selected s, FPL achieves the near optimal learning rate in probability.

Theorem 1: Let $r \in \mathbf{N}$ and $\varepsilon > 0$. If $s \sim m^{1/(2r+d)}$, then there exist positive constants C_i', $i = 1, \ldots, 4$, depending only on M, ρ, and d, $\varepsilon_0 > 0$ and $\varepsilon_m^-, \varepsilon_m^+$ satisfying

$$C_1' m^{-2r/(2r+d)} \le \varepsilon_m^- \le \varepsilon_m^+ \le C_2' (m/\log m)^{-2r/(2r+d)} \tag{9}$$

such that for any $\varepsilon < \varepsilon_m^-$

$$\sup_{f_\rho\in W^r} \mathbf{P}\Big\{\mathbf{z} : \|f_\rho - \pi_M f_{\mathbf{z},s}\|_\rho^2 > \varepsilon\Big\} \ge \mathbf{AC}_m\big(W^r, \varepsilon\big) \ge \varepsilon_0$$

and for any $\varepsilon \ge \varepsilon_m^+$

$$e^{-C_3' m\varepsilon} \le \mathbf{AC}_m\big(W^r, \varepsilon\big) \le \sup_{f_\rho\in W^r} \mathbf{P}\Big\{\mathbf{z} : \|f_\rho - \pi_M f_{\mathbf{z},s}\|_\rho^2 > \varepsilon\Big\} \le 3e^{-C_4' m\varepsilon}.$$

Different from the learning rates established in expectation [12], Theorem 1 exhibits a sharp phase-transition phenomenon of FPL in terms that the probability of success changes dramatically within the critical interval $[\varepsilon_m^-, \varepsilon_m^+]$. It drops from a constant ε_0 to an exponentially small quantity. Due to (9), up to a logarithmic factor, ε_m^- and ε_m^+ are asymptomatically identical, showing that the phase-transition interval is extremely narrow.

The values ε_m^- and ε_m^+ are critical for indicating learning rates of FPL. In particular, it can be found in [6] that ε_m^- implies a lower bound of learning rate, while ε_m^+ indicates an upper bound. Based on Theorem 1, we can derive almost optimal learning rates for FPL in the following corollary.

Corollary 1: If $f_\rho \in W^r$ and $s \sim m^{1/(2r+d)}$, then

$$C_5' m^{-2r/(2r+d)} \le \sup_{f_\rho\in W^r} \mathbf{E}\Big\{\|f_\rho - \pi_M f_{\mathbf{z},s}\|^2\Big\} \le C_6' (m/\log m)^{-2r/(2r+d)} \tag{10}$$

with constants C_5' and C_6' depending only on M, d, f_ρ, and r.

It was shown in [6] and [12] that the rate $\mathcal{O}(m^{-2r/(2r+d)})$ is the optimal learning rate for all learning algorithms based on m samples, provided $f_\rho \in W^r$ and $|y| \le M$. Theorem 1 and Corollary 1 show that (5) is almost optimal choice if the smoothness information of the regression function is known. In short, Theorem 1 and Corollary 1 demonstrate that FPL reduces the computational burden of kernel methods without degrading their generalization capability, provided the kernel parameter s is appropriately tuned. In the following theorem, we show that the *hold-out* strategy provides an efficient way to select s in terms of deducing the same almost optimal learning rate.

Theorem 2: Let $r \in \mathbf{N}$ and $\varepsilon > 0$. If $f_{\mathbf{z},s^*}$ is defined as in Algorithm 1 and (6) with $m_1 = \lceil m/2 \rceil$, then there exist positive constants $\tilde{C}_i$, $i = 1, \ldots, 4$, depending only on M, ρ, and d, $\varepsilon_0 > 0$ and ν_m^-, ν_m^+ satisfying

$$\tilde{C}_1 m^{-2r/(2r+d)} \le \nu_m^- \le \nu_m^+ \le \tilde{C}_2 (m/\log m)^{-2r/(2r+d)}$$

such that for any $\nu < \nu_m^-$

$$\sup_{f_\rho \in W^r} \mathbf{P}\left\{\mathbf{z} : \left\| f_\rho - \pi_M f_{\mathbf{z},s^*} \right\|_\rho^2 > \nu \right\} \ge \mathbf{AC}_m\left(W_2^r, \nu\right) \ge \nu_0$$

and for any $\nu \ge \nu_m^+$

$$e^{-\tilde{C}_3 m\nu} \le \mathbf{AC}_m\left(W^r, \nu\right) \le \sup_{f_\rho \in W^r} \mathbf{P}\left\{\mathbf{z} : \left\| f_\rho - \pi_M f_{\mathbf{z},s^*} \right\|_\rho^2 > \nu \right\} \le e^{-\tilde{C}_4 m\nu}.$$

IV. Comparison to the Existing Methods

In the era of big data, developing learning systems of high quality, i.e., good generalization ability and low computational burden, is a hot topic in the machine learning community. Up till now, several scalable variants of kernel methods have been proposed to reduce their computational burdens. Three typical examples are the localized learning [17], distributed learning [36], and learning with subsampling [11]. In particular, localized learning first divides the input space into ℓ disjoint partitions and then runs kernel methods on the samples whose inputs locate on the partition containing the query point. Distributed learning starts with partitioning the data set into ℓ disjoint subsets, then assigns each data subset to a local machine to produce a local estimator by using RLS and finally synthesizes a global estimator by (weighted) averaging all local estimators. Learning with subsampling devotes to randomly select centers of kernel with small size in a data dependent (or independent) way. The feasibility of these modifications has been verified in [17] and [25] for localized learning, [14], [15], [36] for distributed learning and [21], [22] for learning with subsampling in terms of providing the same optimal learning rates as the original kernel methods.

The advantage of the learning system proposed in this paper lies on the fact that there is no additional parameter involved, which is different from the aforementioned approaches. Utilizing the special features of the polynomial kernel, we even remove the regularization parameter of the classical kernel methods with polynomial kernel. To be detailed, there are three parameters in total: 1) kernel parameter; 2) regularization parameter; and 3) an additional parameter like the number of partition in localized learning, the number of data blocks in distributed learning and the size of subsamples in learning with subsampling, in the previous variants. However, there is only one parameter in FPL, i.e., the degree of polynomial kernel, need to be tuned in the learning process. Thus, although the computational complexity of FPL may be larger than these variants with fixed parameters, the total computational burden of FPL should be much lower. We also present the almost optimal learning rate in confidence to verify the feasibility and efficiency of FPL in the framework of learning theory. All the results show that FPL is a good candidate of learning system to handle massive data.

For polynomial kernel learning, it was deduced from [28] and [39] that the learning rate of algorithm (1) with the polynomial kernel behaves as $\mathcal{O}(m^{-(2r/2r+d+1)})$, which is improved by Theorem 1 in the perspective of three aspects. First, the learning rate analysis in Theorem 1 is based on distribution-free theory: we do not impose any assumptions on the marginal distribution ρ_X. Second, the optimal estimate is established for arbitrary W^r $(0 < r < \infty)$ rather than $0 < r \le 2$. Third, Theorem 1 states that the learning rate can be improved into the almost optimal one, $\mathcal{O}((m/\log m)^{-(2r/2r+d)})$. Another interesting paper concerning polynomial kernel learning is [29], where a rate of order $\mathcal{O}(m^{-1})$ was derived for SVMs with polynomial kernel.

It should be mentioned that for another popular kernel, Gaussian kernel, the generalization error analysis has been conducted in [8], [13], and [32]. It is valuable to compare the performance between Gaussian kernel learning and FPL, since Gaussian kernel learning is a baseline algorithm in the machine learning community for regression. In the former one, there are two parameters including the width of the Gaussian kernel and the regularization parameter that need tuning. However, both parameters are real numbers in some intervals. A common way to determine them is the *cross-validation* strategy, which causes tremendous computations if the size of samples is large. Distinguished with two real parameters in Gaussian kernel learning method, there is only one discrete parameter, i.e., the polynomial order s. Moreover, our developed theoretical result shows that $s = \lceil m^{(1/d+2r)} \rceil$ is an almost optimal choice for arbitrary $f_\rho \in W_p^r$. Although, the smoothness parameter r is usually unknown in practice, Theorem 1 gives us a potential effective way to determine s. Since s is discrete, and s may be smaller than $\lceil m^{1/d} \rceil$, there are only $\lceil m^{1/d} \rceil$ possible values of s. Noting that if d is large, no matter how large m is, $\lceil m^{1/d} \rceil$ is generally smaller than 10, and thus, it is easy to tune s through the *cross-validation* method.

V. Numerical Experiments

In this section, we present toy simulations and UCI data experiments to show the effectiveness of the proposed FPL as well as verify the developed theoretical assertions on FPL. All the numerical experiments are carried out in MATLAB R2013b environment running Windows 7, Intel Core i7-3770K CPU@ 3.50 GHz.

A. Toy Simulations

1) Experimental Setting: In this part, we introduce the settings of the toy simulations.

Methods: In toy simulations, there are four methods being employed. The first one is the RLSs (1) with Gaussian kernel (denoted by GKR), which is regarded as a baseline learning scheme for regression; the second one is the RLS algorithm (1) with polynomial kernel (denoted by PKR), which provides a reference for FPL; the third one is FPL whose centers $\{\eta_j\}_{j=1}^n$ are drawn independently and identically to the uniform distribution; the last one is the FPL (denoted by FPL1) whose centers $\{\eta_j\}_{j=1}^n$ are the first n points of the input samples.

Samples: In toy simulations, the training samples are generated as follows. Let $f(t) = (1-2t)_+^5(32t^2+10t+1)$, where $t \in [0,1]$ and $a_+ = \max\{a, 0\}$. Then it is easy to see that $f \in W^4([0,1])$ and $f \notin W^5([0,1])$. Let $\mathbf{x} = \{x_i\}_{i=1}^m$ be drawn independently and identically according to the uniform distribution with size m and $\mathbf{y} = \{y_i\}_{i=1}^m$ with $y_i = f(x_i) + \delta_i$, where the noise $\{\delta_i\}_{i=1}^m$ are drawn independently and identically according to the Gaussian distribution $\mathcal{N}(0, \sigma^2)$ with variance σ^2. The test samples are generated as follows. $\mathbf{x}' = \{x_i'\}_{i=1}^{m'}$ are drawn independently and identically according to the uniform distribution with size m' and $y_i' = f(x_i')$. In the numerical experiment, TestRMSE is the mean square root error (RMSE) of the testing data via ℓ times simulations with $\ell \in \mathbf{N}$. TrainRMSE is the mean RMSE of the training data via ℓ times simulations. TrainMT and TestMT denote the mean training time and the mean testing time via ℓ times simulations, respectively.

Targets: In these experiments, we implement six simulations to verify the theoretical assessments and show the effectiveness of FPL. The first one is to verify the motivation in Section II-A in terms of studying the role of regularization parameter in the RLSs (1) with polynomial kernel. The second one is to study the role of kernel parameter s in FPL via comparing it with PKR. The third one is to show the relation between the generalization ability and the center generation mechanism. In this simulation, we also verify the learning rates established in Corollary 1 via running FPL on the mentioned data with different sizes. The fourth one is to illustrate the relation between the generalization ability of FPL and the level of noise. The fifth one is to exhibit a phase-transition phenomenon of FPL to verify the assertions of Theorems 1 and 2. The last one is to compare FPL with RLS and show its advantages.

2) Simulation Results: In this part, we report the experimental results and present some discussions.

Simulation 1: In this simulation, we set $m = m' = 1000$, $\ell = 50$, and $\sigma^2 = 0.1$. As mentioned in Section II-A, the regularization parameter in RLS (1) with polynomial kernel is to conquer the singularity of the kernel matrix rather than reduce the capacity of the hypothesis space. To verify it, we draw $n = \binom{s+d}{s}$ samples i.i.d. from $\{x_i\}_{i=1}^m$ according to the uniform distribution to build up the center sets $\{\eta_j\}_{j=1}^n$. Since $\{x_i\}_{i=1}^m$ are i.i.d. drawn according to the uniform distribution, $\{\eta_j\}_{j=1}^n$ is almost surely a K_s-fundamental system. We then study the relations between TestRMSE and λ for the following two models: RLS (1) with polynomial kernel

$$f_{\mathbf{z},s,\lambda} = \arg\min_{f\in\mathcal{H}_s}\left\{\frac{1}{m}\sum_{i=1}^m (f(x_i)-y_i)^2 + \lambda\|f\|_s^2\right\} \qquad (11)$$

and

$$f'_{\mathbf{z},s,\lambda} = \arg\min_{f\in\mathcal{H}_{\eta,n}}\left\{\frac{1}{m}\sum_{i=1}^m (f(x_i)-y_i)^2 + \lambda\|f\|_s^2\right\}. \qquad (12)$$

It should be noted that the model (11) is almost the same as (12), since their hypothesis spaces are almost the same and the penalties are identical. To solve (11), we need to calculate the inverse of the matrix $\mathcal{K}+\lambda m I_m$, where $\mathcal{K} := (K(x_i, x_j))_{i,j=1}^m$ is the kernel matrix and I_m is an identity matrix of the size $m \times m$. Once $\binom{s+d}{s} < m$, the kernel matrix is singular and thus, the regularization is necessary. However, to solve (12), we only need to compute the *pseudo-inverse* of the matrix $(A_{m,n}+\lambda I_{m,n})$, where $A_{m,n}$ is defined in Algorithm 1 and $I_{m,n} \in \mathbf{R}^{m\times n}$ contains the first n columns of I_m. It can be derived from Proposition 2 that $A_{m,n}$ is nonsingular with high confidence. In a word, models (11) and (12) seem the same, while the approach to solve these models are different. For a fixed s (chosen to be the best according to the testing data directly), we consider TestRMSE as a function with respect to λ and depict it in Fig. 1.

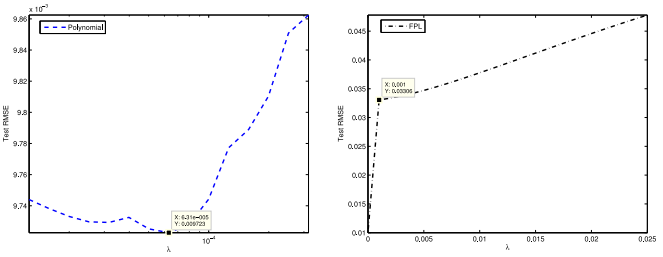

Fig. 1. Left figure shows the relation between test error and λ of (11). The right one illustrates the relation between test error and λ of (12). The details are better seen by zooming on a computer scene.

It is shown in Fig. 1 that there exists an optimal $\lambda > 0$ for (11) and the regularization term is not necessary for (12). In particular, TestRMSE of (12) increases linearly with λ when $\lambda < 10^{-3}$, while for $\lambda > 10^{-3}$, the slope becomes smaller. We give a simple explanation of this phenomenon. As discussed in (8), the generalization error can be decomposed into bias and variance. It is obvious that the bias is a linear function with respect to λ. However, the relation between the test error and λ is more sophisticated. The linearity when $\lambda < 10^{-3}$ thus implies that adding the penalty does not decrease the variance, showing that s plays the dominant role over the penalty in controlling the variance. On the other hand, when λ increases, the penalty plays more and more important role. The inflection of slope means that the penalty plays more important role than s to bound the variance. But it can be found in the left panel of Fig. 1 that the optimal λ of (11) is around 6×10^{-5}, which is much smaller than 10^{-3}, showing that the variance is independent of the penalty. Hence, the penalty in (11) is introduced to conquer the nonsingularity of the kernel matrix rather than reduce the capacity of $\mathcal{H}_s$. This verifies our motivation for designing fast learning strategy based on the polynomial kernel by noting models (11) and (12) are almost the same.

Simulation 2: In this simulation, we study the importance of s in both (5) and (11) to verify whether there are additional requirements imposed on FPL. The experimental settings of this simulation are the same as those of Simulation 1. We take TestRMSE as a function of s by selecting the optimal λ in (11), from 50 candidates drawn equally spaced in $[10^{-5}, 1]$, according to the test samples directly. The simulation results are reported in Fig. 2.

It can be found in the left panel of Fig. 2 that there exists an optimal s minimizing TestRMSE. Since $f \in W^4([0,1])$, it can be found in [28] that the optimal s may close to the value $\lceil m^{1/(2r+d)} \rceil = 3$. Our simulation result shows that the optimal s of (11) is 8, which is near 3. As far as FPL is concerned, it

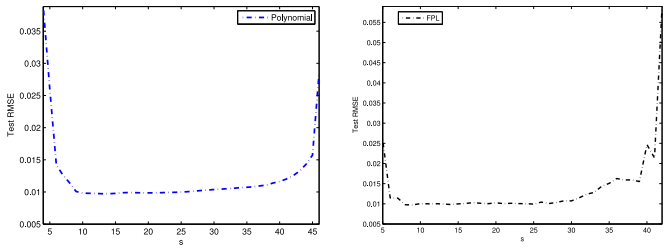

Fig. 2. Left figure shows the relation between the test error and s of (11), while the right figure illustrates the relation between the test error and s of FPL. The details are better seen by zooming on a computer scene.

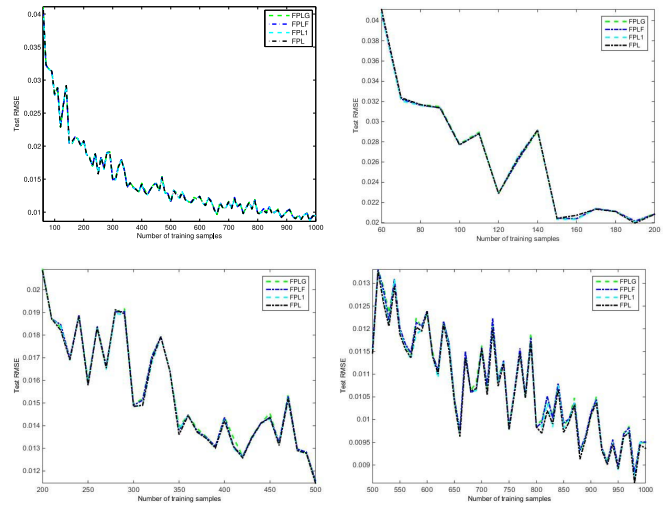

Fig. 3. Upper left figure shows the comparison of the learning capabilities of FPL, FPL1, FPLF, and FPLG. The others are details descriptions of it. The details are better seen by zooming on a computer scene.

is demonstrated in the right panel that the optimal value of s is 7, which is also near 3, verifying our theoretical assertion in Theorem 1. Comparing these two figures in Fig. 2, we find that the relations between test error and s are almost the same for the mentioned two learning schemes. This shows that the requirement of s in FPL is almost the same to that in PKR.

Simulation 3: In this simulation, we set $m' = 1000$, $\ell = 50$, and $\sigma^2 = 0.1$. Our aim is to study the action of the center generation mechanism in FPL. We compare the following four methods of choosing η in (5). FPL denotes that $\eta = \{\eta_i\}_{i=1}^n$ are drawn i.i.d. according to the uniform distribution. FPL1 denotes that $\{\eta_i\}_{i=1}^n$ are selected as the first n inputs of samples. FPLF denotes that $\{\eta_i\}_{i=1}^n$ are chosen as the n equal points in $[0, 1]$. FPLG denotes that $\{\eta_i\}_{i=1}^n$ are generated i.i.d. according to the Gaussian distribution $\mathcal{N}(1/2, 1)$. We figure out the TestRMSE of these four approaches with different sizes of samples (from 80 to 1000) and optimal s (selected according to the test samples). The experimental results are reported in Fig. 3.

It can be found in Fig. 3 that for a suitable s, the choice of η does not affect the learning capability of FPL. This verifies the theoretical results in Proposition 2 and Theorem 1. Furthermore, it is shown in the first panel of Fig. 3 that the learning rate decreases monotonously with respect to the size

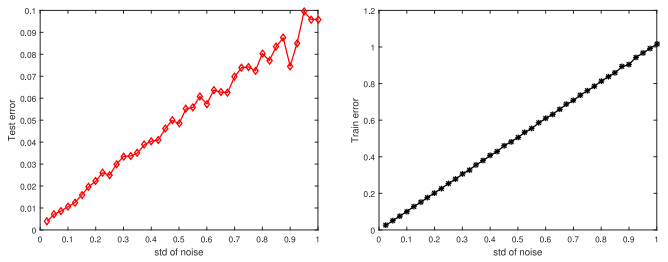

Fig. 4. Effect of the level of noise. The left and right ones are the trends of test and training errors, respectively. The details are better seen by zooming on a computer scene.

of samples. This also verifies the learning rate established in Corollary 1.

Simulation 4: In this simulation, we aim at studying the relation between generalization ability of FPL and the level of noise by drawing the curves of TestRMSE and TrainRMSE with respect to the standard deviation σ of the training noise. For this purpose, we set $m = m' = 1000$, and let the standard deviation vary from 0.025 to 1. For each setting, we repeat 50 times and use threefold cross validation to select the kernel parameter s. The simulation results are reported in Fig. 4

It is shown in Fig. 4 that both TestRMSE and TrainRMSE increase linearly with respect to the standard deviation. In particular, the slope of the line concerning TrainRMSE is almost 1, showing that there is no over-fitting for FPL. However, the slope of the line concerning TestRMSE is almost 0.1, implying the power of FPL in learning noisy data. In short, using the proposed parameter selection strategy, FPL can handle noisy data with different noise levels very well and avoid over-fitting. This coincides with our assertions in Section II and verifies Theorem 2.

Simulation 5: In this simulation, we focus on exhibiting the phase-transition phenomenon presented in Theorems 1 and 2. We set $\sigma^2 = 0.1$ and use threefold cross-validation to select the kernel parameter s. Given a series of tolerance values and sizes of samples, we repeat 50 times running of FPL at each point, and record its value as 0 (a successful case) if the test error is smaller than the tolerance and 1 (a failure case) otherwise, and then use the colors from red to blue to represent the failure probabilities from 1 to 0. We plot a 2-D figure to reflect the failure probability, where x and y-axes represent the sample size and tolerance, respectively. Since the main motivation of developing FPL is to tackle massive data. The sample sizes in this simulation are relatively large: $m = m' = 1000 : 1000 : 50000$. The simulation result is shown in Fig. 5.

Theorems 1 and 2 present a sharp phase-transition phenomenon of FPL in the sense that the probability of success changes dramatically when the tolerance value is in the critical interval $[\varepsilon_m^-, \varepsilon_m^+]$. Fig. 5 shows that such a phenomenon does exist, even when the size of sample are large. In the below part of Fig. 5, the colors of all points are red, which means that the failure probabilities are close to 1. Thus, if the size of samples is small, FPL cannot yield an estimator with very small tolerance. In the upper area, the colors of these points are blue, showing that the successful probabilities are close to 1. Between these two areas, there exists a band, in which the

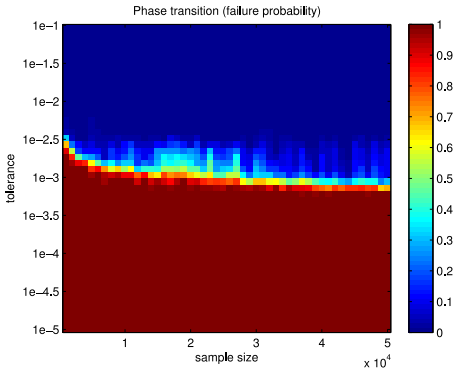

Fig. 5. Phase transition.

TABLE I
COMPARISONS ON THE SIMULATED DATA

Methods	TesRMSEt	TrainMT(second)	TestMT(second)
Gaussian	0.0090	172.31	7.965
Polynomial	0.0097	185.59	11.466
FPL	0.0097	0.214	0.0428
FPL1	0.0097	0.254	0.0383

colors of points vary from red to blue dramatically, exhibiting the phase-transition phenomenon. It is shown that the phase-transition interval is extremely narrow. All these coincide with the theoretical assertions of Theorems 1 and 2.

Simulation 6: In the last simulation, we compare the learning performance of FPL and FPL1 with GKR and PKR.[1] Here, we set $m = m' = 1000$, $\sigma^2 = 0.1$, and $\ell = 50$. Since there are two parameters in GKR: width of the Gaussian kernel and the regularization parameter λ_G, and two parameters in PKR: degree of polynomial kernel and the regularization parameter λ_P, we use the threefold cross-validation to choose these parameters from 50 candidates of λ_G and λ_P as $10^{-5+0.1i}$, where $i = 0, \ldots, 49$, 50 candidates of s as $\{1, 2, \ldots, 50\}$, 40 candidates of δ as $0.01 + 0.025j$, where $j = 0, \ldots, 39$. However, for FPL and FPL1, there is only the kernel parameter s, and we also use the threefold cross-validation method to choose the optimal s from $\{1, \ldots, 50\}$. First, we draw the original function as well as its noisy samples and also depict the estimators learned by the mentioned four learning schemes. The simulation results are shown in Fig. 6.

It can be found in Fig. 6 that the learned functions of all the mentioned methods are almost the same. Since both the Gaussian kernel and polynomial kernel are infinitely smooth function and the regression function is at most fourth smoothness and $m = 1000$, all of them cannot approximate the regression function within a very small tolerance value. This coincides with the lower bound of Theorem 1. We also exhibit the detailed comparisons in Table I.

It can also be found in Table I that all of the mentioned approaches possess the similar TestRMSE. However, the training time of FPL and FPL1 is much less than that of the other two methods. The main reasons of this phenomenon are based on the following assertions. On one hand, there is only one parameter that needs tuning in FPL. On the other hand, the computational complexity of FPL is $\mathcal{O}(mn^2)$, which is smaller than $\mathcal{O}(m^3)$ for a small s due to $n = \binom{s+d}{s}$. Noting that the deduced FPL (or FPL1) estimator is a linear combination of a few basis functions, while the number of adopted basis functions of GKR and PKR are 1000, the test time of FPL and FPL1 is also much less than that of the other two methods.

[1]We do not compare our result with other subsample approach, such as the Nyström regularization [21] and learning with random features [22], since we omit the regularization parameters which of course reduces the computational burden. So, our point in the experiments is to show the good generalization, in which GKR is regarded as the state-of-the-art.

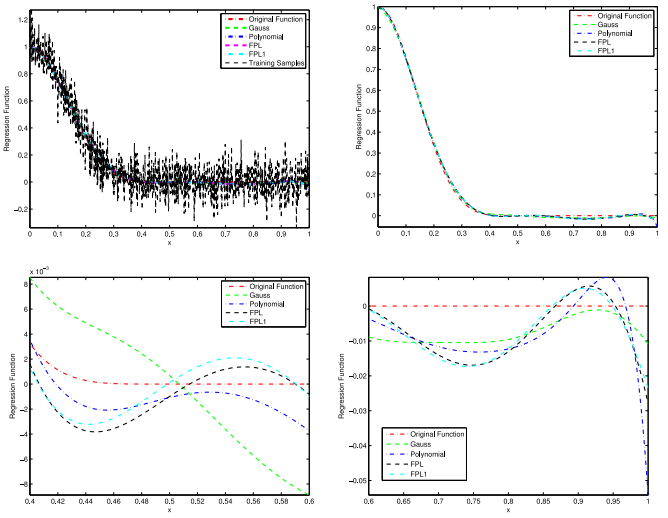

Fig. 6. Upper left figure shows the training samples and the functions learned from the aforementioned four learning strategy. The other three figures illustrate the details of them. The details are better seen by zooming on a computer scene.

TABLE II
SPECIFICATION OF REAL WORLD BENCHMARK DATA SETS

Data sets	Train Number	Test Number	#Attributes
Auto_price	106	53	15
Boston(housing)	337	169	13
Stock	633	317	9
Abalone	2785	1392	8
Bank8FM	2999	1500	8
Delta_ailerons	3565	3564	5
Computer activity	4096	4096	21
Delta_Elevators	4759	4758	6
California housing	10320	10320	8
Poker_hand	25010	1000000	11

B. UCI Data Experiments

1) Experimental Setting: In this part, we introduce the simulation settings of the UCI data experiments.

Method choices: In the UCI data experiment, we compare four methods containing SVMs [26], GKR, PKR, and FPL on ten real-world benchmark data sets covering various fields. We use threefold cross-validation to select parameters of the aforementioned methods among 40 candidates of the width of Gaussian kernel and 50 candidates of the regularization parameter λ. However, due to the theoretical analysis proposed in [39] and Theorem 1 in this paper, there are only $\{1, \ldots, \lceil m^{1/d} \rceil\}$ candidates of polynomial kernel parameter s. The centers of FPL are drawn i.i.d. according to the uniform distribution.

TABLE III
TRAINING ERROR AND TESTING ERROR

Data sets	TrainRMSE				TestRMSE			
	SVM	GKR	PKR	FPL	SVM	GKR	PKR	FPL
Auto_price	0.0674	0.019	0.0598	0.0713	0.0914	0.1132	**0.0894**	0.0968
Boston(housing)	0.0683	0.0114	0.0668	0.0881	0.0852	0.1201	**0.0748**	0.0998
Stock	0.0491	0.0141	0.023	0.0241	0.0503	**0.0286**	0.0327	0.0365
Abalone	0.0750	0.0716	0.0735	0.0744	0.0792	0.0759	**0.0748**	0.0753
Bank8FM	0.0446	0.0359	0.0367	0.0371	0.0458	**0.0422**	0.045	0.0475
Delta_airelons	0.0417	0.0369	0.037	0.0376	0.0422	**0.0388**	0.0392	0.039
Computer activity	0.0445	0.0221	0.0282	0.0259	0.0463	**0.0261**	0.03	0.0337
Delta_Elevators	0.0526	0.0526	0.0527	0.0532	0.0542	**0.0532**	**0.0532**	0.0534
California housing	0.0734	0.0575	0.0819	0.0611	0.072	**0.0625**	0.0832	0.0696
Poker_hand	−−	0.0827	0.0854	0.0853	−−	**0.0823**	0.0824	0.0824

TABLE IV
TRAINING TIME AND SPARSITY

Data sets	TrainMT				Mean sparsity			
	SVM	GKR	PKR	FPL	SVM	GKR	PKR	FPL
Auto_price	26.09272	2.624	0.1147	**0.0111**	22.9	71	71	**16**
Boston(housing)	11.3646	78.901	2.6658	**0.0178**	56.35	225	225	**41.3**
Stock	30.4451	25.057	0.9929	**0.0814**	**22.9**	422	422	154
Abalone	687.021	790.352	31.8664	**0.1693**	423.2	1857	1857	**45**
Bank8FM	660.301	974.147	39.1042	**0.187**	**84.25**	2000	2000	153
Delta_airelons	291.488	1421.4	114.9805	**1.2169**	99	2377	2377	**56**
Computer activity	723.985	2069.3	53.8626	**0.2684**	**80**	2731	2731	253
Delta_Elevators	882.53	2988.9	198.9652	**1.305**	292	3173	3173	**49.7**
California housing	8489.09	31031	1469.6	**3.0532**	924	7595	7595	**75**
Poker_hand	−−	136804.2	6280.7	**6.8836**	−−	16674	16674	**66**

Samples: All the data are cited from: http://www.niaad.liacc.up.pt/∼ltorgo/Regression/ds_menu.html. The sizes of training and test samples used are listed in Table II.

2) Experimental Results: The experiment results of UCI data are reported in Tables III and IV. As shown in Table III, TrainRMSE and TestRMSE of all the mentioned methods are similar. But as far as the TrainMT is concerned, it can be found in Table IV that FPL outperforms the others. It can also be found in Table IV that the TrainMT of PKR is smaller than GKR and SVM. This is because we use the theoretical result in [39] to select the kernel parameter s. It was shown in [39] (see also Theorem 1 in this paper) that it suffices to select s in the set $\{1, \ldots, \lceil m^{1/d} \rceil\}$. This degrades the difficulty of the parameter selection of PKR. Since TestMT depends on the sparsity of the estimator, we also give a comparison of the sparsity of the mentioned methods in Table IV. In a nutshell, as far as the generalization capability is concerned, all of these methods are of high quality. However, as far as the computational burden is concerned, FPL is superior to the others. Furthermore, different from PKR, FPL can deduce sparse estimators. Noting in Table IV that the training time of FPL is much less than the classical kernel approach, especially when the size of data exceeds ten thousands. From these experiment results, FPL provides a possibility to tackle massive data.

VI. PROOFS

Proof of Proposition 1: $\dim \mathcal{H}_{\zeta,n} < n$ means that there exists a nontrivial set $\{a_i\}_{j=1}^n$ such that

$$\sum_{j=1}^{n} a_j \big(1 + \zeta_j \cdot x\big)^s = 0.$$

That is, the system of equations

$$\sum_{j=1}^{n} a_j \big(1 + \zeta_j \cdot \zeta_k\big)^s = 0, \;\; k = 1, \ldots, n$$

is solvable. In other words, the matrix $((1 + \zeta_i \cdot \zeta_j)^s)_{i,j=1}^n$ is singular. Noting that

$$\big(1 + \zeta_j \cdot \zeta_i\big)^s = \sum_{k=0}^{s} \binom{s}{k} \big(\zeta_j \cdot \zeta_i\big)^k = \sum_{k=0}^{s} \binom{s}{k} \sum_{|\alpha|=k} C_\alpha^k \zeta_j^\alpha \zeta_i^\alpha$$

we obtain

$$\sum_{i,j=1}^{n} a_i a_j \big(1 + \zeta_i \cdot \zeta_j\big)^s = \sum_{k=0}^{s} \binom{s}{k} \sum_{|\alpha|=k} C_\alpha^k \left(\sum_{i=1}^{n} a_i \zeta_i^\alpha \right)^2$$

where

$$C_\alpha^k = \frac{d!}{\alpha_1! \cdots \alpha_d!}, \;\; \alpha := (\alpha_1, \ldots, \alpha_d).$$

Thus, the singularity of the matrix $((1 + \zeta_i \cdot \zeta_j)^s)_{i,j=1}^n$ implies

$$\sum_{i=1}^{n} a_i \zeta_i^\alpha = 0.$$

Let

$$P(\varsigma) := \sum_{i=1}^{n} a_i \varsigma^\alpha.$$

The above assertion shows that $\zeta_i, i = 1, \ldots, n$ are n distinct zero points of P. Noting that the degree of P is at most s, it

can be easily deduced from [1, Lemma 3.1] that the zero set of P

$$Z(p) := \Big\{x \in \mathbf{B}^d : P(x) = 0\Big\}$$

has Lebesgue measure 0. This completes the proof of Proposition 1. ■

To prove Proposition 2, we need the following two lemmas. The first one establishes a relation between the d-dimensional unit ball $\mathbf{B}^d$ and the $(d+1)$-dimensional unit sphere $\mathbf{S}^d$, which can be found in [33, Lemma 2.1].

Lemma 3: For any continuous function f defined on $\mathbf{S}^d$, there holds

$$\int_{\mathbf{S}^d} f(\xi) d\omega_d(\xi) = \int_{\mathbf{B}^d} \left[f\left(x, \sqrt{1-|x|^2}\right) + f\left(x, -\sqrt{1-|x|^2}\right)\right] \frac{dx}{\sqrt{1-|x|^2}}.$$

Let h_Λ be the mesh norm of a set of points $\Lambda = \{\xi_i\}_{i=1}^m \subset \mathbf{S}^d$ defined by

$$h_\Lambda := \max_{\xi \in \mathbf{S}^d} \min_j d(\xi, \xi_j)$$

where $d(\xi, \xi')$ is the geodesic (great circle) distance between the points ξ and ξ' on $\mathbf{S}^d$. The second one is the well-known cubature formula on the sphere, which can be found in [18].

Lemma 4: If there exists a constant c such that $h_\Lambda \le cn^{-1/d}$, then there exists a set of positive numbers $\{a_i\}_{i=1}^m$ satisfying

$$\sum_{i=1}^m a_i^p \le Cm^{1-p}$$

such that for any $Q \in \Pi_{2s}^d$

$$C_1 \int_{\mathbf{S}^d} |Q(\xi)|^p d\omega_d(\xi) \le \sum_{i=1}^m a_i |Q(x_i)|^p \le C_2 \int_{\mathbf{S}^d} |Q(\xi)|^p d\omega_d(\xi)$$

where Π_s^d denotes the set of algebraic polynomials of degree at most s defined on $\mathbf{S}^{d+1}$.

Proof of Proposition 2: Denote $\Lambda = \{x_i\}_{i=1}^m$. At first, we present an upper bound of h_Λ. Let $D(\xi, r)$ be the spherical cap with center ξ and radius r. Then for arbitrary $\varepsilon > 0$, due to the definition of the mesh norm, we obtain

$$\mathbf{P}\{h_\Lambda > \varepsilon\} = \mathbf{P}\left\{\max_{\xi \in \mathbf{S}^d} \min_j d\big(\xi, \xi_j\big) > \varepsilon\right\} \le \mathbf{E}\big\{(1 - \mu(D(\xi, \varepsilon)))^m\big\}.$$

Let $t_1, \ldots, t_N$ be the quasi-uniform points [34] on the sphere. Then it is easy to deduce

$$N \le \frac{c'}{\varepsilon^d}, \text{ and } \mathbf{S}^d \subset \bigcup_{j=1}^N D\big(t_j, \varepsilon/2\big)$$

for some $c' > 0$. If $\xi \in D(t_j, \varepsilon/2)$, then $D(t_j, \varepsilon/2) \subset D(\xi, \varepsilon)$. Therefore, we get

$$\begin{aligned}
&\mathbf{E}\big\{(1 - \mu(D(\xi, \varepsilon)))^m\big\} \\
&\quad \le \sum_{j=1}^N \int_{D(t_j, \varepsilon/2)} (1 - \mu(D(\xi, \varepsilon)))^m d\mu \\
&\quad \le \sum_{j=1}^N \int_{D(t_j, \varepsilon/2)} \big(1 - \mu\big(D\big(t_j, \varepsilon/2\big)\big)\big)^m d\mu \\
&\quad = \sum_{j=1}^N \mu\big(D\big(t_j, \varepsilon/2\big)\big)\big(1 - \mu\big(D\big(t_j, \varepsilon/2\big)\big)\big)^m \\
&\quad \le \sum_{j=1}^N \max_u u(1-u)^m \le \sum_{j=1}^N \max_u u e^{-mu} = \frac{eN}{m} \\
&\quad \le \frac{c'}{m\varepsilon^d}.
\end{aligned}$$

That is

$$\mathbf{P}\Big\{h_\Lambda > cn^{-1/d}\Big\} \le \frac{c^d c' n}{m} =: \frac{c'' n}{m}.$$

Thus, it follows from Lemma 4 that with confidence at least $1 - (c''n/m)$, there exists a set of numbers $\{a_i\}_{i=1}^m$ satisfying:

$$\sum_{i=1}^m |a_i|^2 \le Cm^{-1}$$

such that

$$C_1 \int_{\mathbf{S}^d} |Q_s(\xi)|^2 d\omega_d(\xi) \le \sum_{i=1}^m a_i |Q_s(x_i)|^2 \le C_2 \int_{\mathbf{S}^d} |Q_s(\xi)|^2 d\omega_d(\xi) \quad \text{for any} \quad Q_s \in \Pi_{2s}^d.$$

If we set $\tau = (x, x_{(d+1)}) \in \mathbf{S}^d$ with $x \in \mathbf{B}^d$ and $x_{(d+1)} = \sqrt{1-|x|^2}$. Then for every monomial $p_\alpha(x) = x^\alpha \in \mathcal{P}_k^d$ the function $q_\alpha(\tau) = g_\alpha(x)$ is a polynomial in $\mathcal{P}_s^{d+1}$, where $\alpha = (\alpha_1, \ldots, \alpha_d)$ and $k = \sum_{i=1}^d |\alpha_i| = k \le s$. Furthermore, if $\{x_i\}_{i=1}^m \subset \mathbf{B}^d$ is a set of i.i.d. random variables, then $\{\tau_i\}_{i=1}^m \subset \mathbf{B}^d$ is also a set of i.i.d. random variables. By Lemma 3, the above inequality is equivalent to

$$\begin{aligned}
&C_1 \int_{\mathbf{B}^d} \left| Q_s\left(x, \sqrt{1-|x|^2}\right) + Q_s\left(x, -\sqrt{1-|x|^2}\right)\right|^2 \frac{dx}{\sqrt{1-|x|^2}} \le \sum_{i=1}^m a_i |Q_s(\tau_i)|^2 \le C_2 \\
&\int_{\mathbf{B}^d} \left| Q_s\left(x, \sqrt{1-|x|^2}\right) + Q_s\left(x, -\sqrt{1-|x|^2}\right)\right| \frac{dx}{\sqrt{1-|x|^2}}.
\end{aligned}$$

Let $P_s(x) = Q_s(x, \sqrt{1-|x|^2})$, then we can obtain

$$C_1 \int_{\mathbf{B}^d} |P_s(x)|^2 \frac{dx}{\sqrt{1-|x|^2}} \le \sum_{i=1}^m a_i |P_s(x_i)|^2 \le C_2 \int_{\mathbf{B}^d} |P_s(x)|^2 \frac{dx}{\sqrt{1-|x|^2}}$$

holds with probability at least $1-(c''n/m)$. This means, for arbitrary $\mathbf{c}=(c_1,\ldots,c_n)$

$$C_1\int_{B^d}\left(\sum_{j=1}^{n}c_jK_s(\eta_j,x)\right)^2\frac{dx}{\sqrt{1-|x|^2}} \le \sum_{i=1}^{m}a_i\left(\sum_{j=1}^{n}c_jK_s(\eta_j,x_i)\right)^2 \le C_2\int_{\mathbf{B}^d}\left(\sum_{j=1}^{n}c_jK_s(\eta_j,x)\right)^2\frac{dx}{\sqrt{1-|x|^2}}$$

holds with probability at least $1-(c''n/m)$. This finishes the proof of Proposition 2. ∎

To prove Theorem 1, we need the following two lemmas, which can be found in [16] (see also [27]) and [19], respectively.

Lemma 5: Let $\mathcal{H}$ be a linear space with dimension k. Define

$$f_{\mathbf{z}}=\arg\min_{f\in\mathcal{H}}\mathcal{E}_{\mathbf{z}}(f).$$

Then, there holds for arbitrary $h\in\mathcal{H}$ that

$$\mathbf{P}\left\{\|\pi_M f_{\mathbf{z}}-f_\rho\|_\rho^2>\varepsilon\right\}\le\exp\left\{k\log\frac{\tilde{c}M}{\varepsilon}-\frac{3m\varepsilon}{128M^2}\right\}+2\exp\left\{\frac{-m\varepsilon^2}{(3M+\|h\|_\infty)^2\left(2\|h-f_\rho\|_\rho^2+\frac{1}{3}\varepsilon\right)}\right\}$$

where $\tilde{c}$ is constant independent of m or k.

To provide the second lemma, we should introduce the best approximation operator. A function η is said to be admissible [19] if $\eta\in C^\infty[0,\infty)$, $\eta(t)\ge 0$, and supp $\eta\subset[0,\ 2]$, $\eta(t)=1$ on $[0,\ 1]$, and $0\le\eta(t)\le 1$ on $[1,\ 2]$.

Let

$$h_k:=\frac{\pi^{1/2}\Gamma(d+k)\Gamma((d+1)/2)}{(k+d/2)k!\Gamma(d/2)}\Gamma(d).$$

Define

$$U_k:=(h_k)^{-1/2}G_k^{d/2},\ k=0,1,\ldots \tag{13}$$

where G_k^μ is the well known Gegenbauer polynomial with order μ [19], that is

$$G_k^\mu(t)=\sum_{j=0}^{[k/2]}(-1)^k\frac{\Gamma(k-j+\mu)}{\Gamma(\mu)j!(k-2j)!}(2t)^{k-2j},\ t\in[-1,1].$$

Then it is easy to see that $\{U_n\}_{n=0}^\infty$ is a complete orthonormal system for the weighted L^2 space $L^2(I,w)$, where $w(t):=(1-t^2)^{(d-1/2)}$. The best approximation kernel is defined by

$$L_s(x,y):=\sum_{k=0}^{\infty}\eta\left(\frac{k}{2s}\right)v_k^2\int_{\mathbf{S}^{d-1}}U_k(x\cdot\xi)U_k(y\cdot\xi)d\omega_{d-1}(\xi)$$

where $d\omega_{d-1}$ stands for the aero element of $\mathbf{S}^{d-1}$, the d-dimensional unit sphere and $v_k:=([(\Gamma(k+d))/(2(2\pi)^{d-1}\Gamma(k+1))])^{(1/2)}$. Let

$$E_s(f)_p:=\inf_{P\in\mathcal{P}_s^d}\|f-P\|_{L^p(\mathbf{B}^d)}$$

be the best approximation error of $\mathcal{P}_s^d$. Define

$$\mathcal{L}_sf(x):=\int_{\mathbf{B}^d}L_s(x,y)f(y)dy. \tag{14}$$

It is obvious that $\mathcal{L}_sf\in\mathcal{P}_s^d$.

Lemma 6: Let $1\le p\le\infty$, and $\mathcal{L}_s$ be defined in (14), then for arbitrary $f\in W^r$, there exists a constant $\tilde{c}$ depending only on d such that

$$\|f-\mathcal{L}_sf\|_\infty\le\tilde{c}s^{-r/d}$$

and

$$\|\mathcal{L}_sf\|_\infty\le\tilde{c}\|f\|_\infty.$$

With the help of the above two lemmas, we prove Theorem 1 as follows.

Proof of Theorem 1: The lower bound can be more easily deduced. Actually, it follows from [6, eq. (3.27)] that for any estimator $f_{\mathbf{z}}\in\Phi_m$, there holds:

$$\sup_{f_\rho\in W^r}P_m\left\{\mathbf{z}:\|f_{\mathbf{z}}-f_\rho\|_\rho^2\ge\varepsilon\right\}\ge\begin{cases}\varepsilon_0, & \varepsilon<\varepsilon^-\\ e^{-cm\varepsilon}, & \varepsilon\ge\varepsilon^-\end{cases}$$

where $\varepsilon_0=(1/2)$ and $\varepsilon^-=C_1'm^{-2r/(2r+d)}$.

From Lemma 6 and $\|\cdot\|_\rho\le\|\cdot\|_\infty$, we use Lemma 5 with $\mathcal{H}=\mathcal{H}_{\eta,n}$ and $h=\mathcal{L}_sf_\rho$. It is obvious that the dimension of $\mathcal{H}_{\eta,n}$ is $n\sim s^d$ and $\|h\|_\infty\le\tilde{c}M$. Then for arbitrary $s\in\mathbf{N}$

$$\mathbf{P}\left\{\|\pi_M f_{\mathbf{z}}-f_\rho\|_\rho^2>\varepsilon\right\}\le\exp\left\{n\log\frac{\tilde{c}M}{\varepsilon}-\frac{3m\varepsilon}{128M^2}\right\}+2\exp\left\{\frac{-m\varepsilon^2}{(3+\tilde{c})^2\left(2\tilde{c}^2s^{-2r/d}+\frac{1}{3}\varepsilon\right)}\right\}$$

where $\tilde{c_1}$ is a constant depending only on d. Since $s\sim m^{1/(2r+d)}$ and $n\sim s^d$, we have from $\varepsilon\ge\varepsilon_m^+$ that

$$\mathbf{P}\left\{\|\pi_M f_{\mathbf{z},s}-f_\rho\|_\rho^2>\varepsilon\right\}\le 3\exp\left\{-C_4'm\varepsilon\right\}$$

with C_4' depending only on $\tilde{c}$, M, d, and r. That is, for $\varepsilon\ge\varepsilon^+$

$$\mathcal{E}(\pi_M f_{\mathbf{z},\lambda,s})-\mathcal{E}(f_\rho)\le\varepsilon \tag{15}$$

holds with confidence at least $1-3\exp\{-C_4'm\varepsilon\}$. This finishes the proof of Theorem 1. ∎

Proof of Corollary 1: The lower bound of (10) can be found in [12, Th. 3.2]. It suffices to prove the upper bound of (10). we apply the formula

$$\mathbf{E}[\xi]=\int_0^\infty\text{Prob}[\xi>t]dt \tag{16}$$

with $\xi=\|\pi_M f_{\mathbf{z},n,\delta}-f_\rho\|_\rho^2$ to prove Corollary 1. Based on Theorem 1, we have

$$\mathbf{E}\left[\|\pi_M f_{\mathbf{z},s}-f_\rho\|_\rho^2\right]\le\varepsilon_m^+ +\int_0^\infty e^{-C_4'm\varepsilon}d\varepsilon \le C_6'(m/\log m)^{-2r/(2r+d)}$$

where $C_6'=C_2'+(C_4')^{-1}$. This completes the proof of Corollary 1. ∎

To prove Theorem 2, we need the following lemma, which can be found in [2, Proposition 11].

Lemma 7: Let $\{\xi_i\}_{i=1}^m$ be a set of real valued i.i.d. random variables with mean μ, $|\xi_i| \le B$ and $\mathbf{E}[(\xi_i-\mu)^2] \le \sigma^2$, for all $i \in \{1,2,\ldots,m\}$. Then, for arbitrary $a>0$, $\varepsilon>0$

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

and

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

Proof of Theorem 2: It suffices to prove the upper bound. Let

$$\hat{s} = \arg\min_{s\in\Xi}\int_Z \big(\pi_M f_{\mathbf{z_1},s}(x)-y\big)^2 d\rho.$$

We then get from (7) that

$$\hat{s} = \arg\min_{s\in\Xi}\big\|\pi_M f_{\mathbf{z_1},s} - f_\rho\big\|.$$

According to the definition of $f_{\mathbf{z_1},s}$ and Theorem 1, for $\varepsilon \ge \varepsilon^+$

$$\big\|\big(\pi_M f_{\mathbf{z_1},\hat{s}}\big) - f_\rho\big\|_\rho^2 \le \varepsilon \tag{17}$$

holds with confidence at least $1-3\exp\{-C_4'm\varepsilon\}$.

For $z_i \in \mathbf{z}_2$, let us define the random variables

$$\xi_i^s = \big(\pi_M f_{\mathbf{z_1},s}(x_i)-y_i\big)^2 - \big(f_\rho(x_i)-y_i\big)^2.$$

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

$$\mathbf{E}\big[\xi_i^s\big] = \big\|\big(\pi_M f_{\mathbf{z_1},s}\big) - f_\rho\big\|_\rho^2$$

and $\mathbf{E}[(\xi_i^s)^2]$ can be bounded by

$$\int_Z \big(\pi_M f_{\mathbf{z_1},s}(x)-f_\rho(x)\big)^2\big(\pi_M f_{\mathbf{z_1},s}(x)+f_\rho(x)-2y\big)^2 d\rho \le 16M^2\mathbf{E}\big[\xi_i^s\big].$$

Hence, using Lemma 7 with $a=1$, $\xi_i=\xi_i^s$, $\mu=\mathbf{E}[\xi_i^s]$, $B=4M^2$, and $\sigma^2 \le 16M^2\mu$, we obtain for all $s\in\Xi$, with probability greater than

$$1-|\Xi|\exp\{-\bar{c}m\varepsilon\} = 1-\exp\left\{-\bar{c}m\varepsilon+\frac{1}{d}\log m\right\}$$

there holds

$$\frac{1}{m_2}\sum_{i=1}^{m_2}\xi_i^s \le \bar{c}_1\mathbf{E}\big[\xi_i^s\big]+\varepsilon \tag{18}$$

and

$$\mathbf{E}\big[\xi_i^s\big] \le \bar{c}_2\frac{1}{m_2}\sum_{i=1}^{m_2}\xi_i^s + \varepsilon \tag{19}$$

where $\bar{c},\bar{c}_1,\bar{c}_2$ are constants depending only on d and M. Therefore, for arbitrary $\varepsilon \ge \varepsilon_m^+$, with confidence at least

$$1-5\exp\{-\bar{c}m\varepsilon\}$$

there holds

$$\begin{aligned}
&\big\|\big(\pi_M f_{\mathbf{z_1},s^*}\big)-f_\rho\big\|_\rho^2 \\
&= \mathbf{E}\big[\xi_i^{s^*}\big] \le \frac{\bar{c}_2}{m_2}\sum_{i=1}^{m_2}\Big(\big(\pi_M f_{\mathbf{z_1},s^*}(x_i)-y_i\big)^2-\big(f_\rho(x_i)-y_i\big)^2\Big)+\varepsilon \\
&\le \frac{\bar{c}_2}{m_2}\sum_{i=1}^{m_2}\Big(\big(\pi_M f_{\mathbf{z_1},\hat{s}}(x_i)-y_i\big)^2-\big(f_\rho(x_i)-y_i\big)^2\Big)+\varepsilon \\
&\le \bar{c}_1\bar{c}_2\mathbf{E}\big[\xi_i^{\hat{s}}\big]+(\bar{c}_2+1)\varepsilon \\
&= \bar{c}_1\bar{c}_2\big\|\big(\pi_M f_{\mathbf{z_1},\hat{s}}\big)-f_\rho\big\|_\rho^2+(\bar{c}_2+1)\varepsilon \le \bar{c}_3\varepsilon
\end{aligned}$$

where the first inequality is deduced by (19), the second inequality is according to the definition of s^*, the third one is based on (18), the last one follows from (17) and $\bar{c}_3 = \bar{c}_1\bar{c}_2+\bar{c}_2+1$. This completes the proof of Theorem 2. ■

VII. Conclusion

In this paper, we develop a new learning system based on special features of polynomial kernel and subsampling, called the FPL to reduce the computational burden for kernel methods. Both theoretical analysis and numerical experiments are conducted to show the effectiveness of FPL. Theoretically, we show that FPL can achieve the almost optimal learning rates for kernel methods. This demonstrates that FPL is at least not worse than the kernel methods. Numerically, we exhibited that FPL can significantly reduce the computational burden of kernel methods. Both theoretical and experimental results show that FPL is of high quality and is a good candidate of learning systems to handle massive data.

Acknowledgment

The authors would like to thank three anonymous reviewers and the Associate Editor for their constructive and helpful comments.

References

[1] R. F. Bass and K. Gröchenig, "Random sampling of multivariate trigonometric polynomials," *SIAM. J. Math. Anal.*, vol. 36, no. 3, pp. 773–795, 2005.

[2] A. Caponnetto and Y. Yao, "Cross-validation based adaptation for regularization operators in learning theory," *Anal. Appl.*, vol. 8, no. 2, pp. 161–183, 2010.

[3] X. Chang, S.-B. Lin, and D.-X. Zhou, "Distributed semi-supervised learning with kernel ridge regression," *J. Mach. Learn. Res.*, vol. 18, no. 46, pp. 1–22, 2017.

[4] M. Chen, S. Mao, and Y. Liu, "Big data: A survey," *Mobile Netw. Appl.*, vol. 19, no. 2, pp. 171–209, 2014.

[5] F. Cucker and D. X. Zhou, *Learning Theory: An Approximation Theory Viewpoint*. Cambridge, U.K.: Cambridge Univ. Press, 2007.

[6] R. A. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov, "Approximation methods for supervised learning," *Found. Comput. Math.*, vol. 6, no. 1, pp. 3–58, 2006.

[7] K. Dong, D. Eriksson, H. Nickisch, D. Bindel, and A. G. Wilson, "Scalable log determinants for Gaussian process kernel learning," in *Proc. NIPS*, 2017, pp. 1–11.

[8] M. Eberts and I. Steinwart, "Optimal regression rates for SVMs using Gaussian kernels," *Electron. J. Stat.*, vol. 7, pp. 1–42, Jan. 2013.

[9] T. Evgeniou, M. Pontil, and T. Poggio, "Regularization networks and support vector machines," *Adv. Comput. Math.*, vol. 13, pp. 1–50, Apr. 2000.

[10] Z.-C. Guo and L. Shi, "Learning with coefficient-based regularization and ℓ^1-penalty," *Adv. Comput. Math.*, vol. 39, nos. 3–4, pp. 493–510, 2013.
[11] A. Gittens and M. W. Mahoney, "Revisitng the Nyström method for improved large-scale machine learning," *J. Mach. Learn. Res.*, vol. 17, no. 1, pp. 1–65, 2016.
[12] L. Györfy, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Berlin, Germany: Springer, 2002.
[13] S. Lin, J. Zeng, J. Fang, and Z. Xu, "Learning rates of ℓ^q coefficient regularization learning with Gaussian kernel," *Neural Comput.*, vol. 26, pp. 2350–2378, Sep. 2014.
[14] S.-B. Lin, X. Guo, and D.-X. Zhou, "Distributed learning with regularized least squares," *J. Mach. Learn. Res.*, vol. 18, no. 92, pp. 1–31, 2017.
[15] S.-B. Lin and D.-X. Zhou, "Distributed kernel-based gradient descent algorithms," *Constr. Approx.*, vol. 47, no. 2, pp. 249–276, 2018.
[16] S. B. Lin, J. Fang, and X. Chang, "Learning with selected features," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.
[17] M. Meister and I. Steinwart, "Optimal learning rates for localized SVMs," *J. Mach. Learn. Res.*, vol. 17, no. 194, pp. 1–44, 2016.
[18] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, "Spherical Marcinkiewicz-Zygmund inequalities and positive quadrature," *Math. Comput.*, vol. 70, no. 235, pp. 1113–1130, 2000.
[19] P. P. Petrushev and Y. Xu, "Localized polynomial frames on the ball," *Constr. Approx.*, vol. 27, no. 2, pp. 121–148, 2008.
[20] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Application*. New York, NY, USA: Wiley, 1971.
[21] A. Rudi, R. Camoriano, and L. Rosasco, "Less is more: Nyström computational regularization," in *Proc. NIPS*, 2015, pp. 1657–1665.
[22] A. Rudi and L. Rosasco, "Generalization properties of learning with random features," in *Proc. NIPS*, 2017, pp. 1–11.
[23] L. Shi, Y.-L. Feng, and D.-X. Zhou, "Concentration estimates for learning with l^1-regularizer and data dependent hypothesis spaces," *Appl. Comput. Harmon. Anal.*, vol. 31, no. 2, pp. 286–302, 2011.
[24] L. Shi, "Learning theory estimates for coefficient-based regularized regression," *Appl. Comput. Harmon. Anal.*, vol. 34, no. 2, pp. 252–265, 2013.
[25] R. Tandon, S. Si, and P. Ravikumar, "Kernel ridge regression via partitioning," *arXiv preprint arXiv:1608.01976*, 2016.
[26] J. S. Taylor and N. Cristianini, *Kernel Methods for Pattern Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[27] V. N. Temlyakov, "Approximation in learning theory," *Constr. Approx.*, vol. 27, no. 1, pp. 33–74, 2008.
[28] H. Z. Tong, D.-R. Chen, and Z. P. Li, "Learning rates for regularized classifiers using multivariate polynomial kernels," *J. Complex.*, vol. 24, nos. 5–6, pp. 619–631, 2008.
[29] H. Z. Tong, "A note on support vector machines with polynomial kernels," *Neural Comput.*, vol. 28, no. 1, pp. 71–88, Jan. 2016.
[30] A. G. Wilson, E. Gilboa, A. Nehorai, and J. P. Cunningham, "Fast kernel learning for multidimensional pattern extrapolation," in *Proc. NIPS*, 2014, pp. 3626–3634.
[31] Q. Wu and D.-X. Zhou, "Learning with sample dependent hypothesis space," *Comput. Math. Appl.*, vol. 56, no. 11, pp. 2896–2907, 2008.
[32] D. H. Xiang and D. X. Zhou, "Classification with Gaussians and convex loss," *J. Mach. Learn. Res.*, vol. 10, pp. 1447–1468, Jul. 2009.
[33] Y. Xu, "Orthogonal polynomials and cubature formulae on spheres and on balls," *SIAM J. Math. Anal.*, vol. 29, pp. 779–793, Jul. 1998.
[34] H. Wendland, *Scattered Data Approximation*. Cambridge, U.K.: Cambridge Univ. Press, 2005.
[35] X. Wu, X. Zhu, G.-Q. Wu, and D. Wu, "Data mining with big data," *IEEE Trans. Knowl. Data Eng.*, vol. 26, no. 1, pp. 97–107, Jan. 2014.
[36] Y. C. Zhang, J. Duchi, and M. Wainwright, "Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates," *J. Mach. Learn. Res.*, vol. 16, no. 1, pp. 3299–3340, 2015.
[37] D.-X. Zhou, "The covering number in learning theory," *J. Complex.*, vol. 18, no. 3, pp. 739–767, 2002.
[38] D.-X. Zhou, "Capacity of reproducing kernel spaces in learning theory," *IEEE Trans. Inf. Theory*, vol. 49, no. 7, pp. 1743–1752, Jul. 2003.
[39] D.-X. Zhou and K. Jetter, "Approximation with polynomial kernels and SVM classifiers," *Adv. Comput. Math.*, vol. 25, nos. 1–3, pp. 323–344, 2006.
[40] Z.-H. Zhou, N. V. Chawla, Y. Jin, and G. J. Williams, "Big data opportunities and challenges: Discussions from data analytics perspectives," *IEEE Comput. Intell. Mag.*, vol. 9, no. 4, pp. 62–74, Nov. 2014.

Authors' photographs and biographies not available at the time of publication.