Is Extreme Learning Machine Feasible? A Theoretical Assessment (Part II)

Shaobo Lin, Xia Liu, Jian Fang, and Zongben Xu

***Abstract*— An extreme learning machine (ELM) can be regarded as a two-stage feed-forward neural network (FNN) learning system that randomly assigns the connections with and within hidden neurons in the first stage and tunes the connections with output neurons in the second stage. Therefore, ELM training is essentially a linear learning problem, which significantly reduces the computational burden. Numerous applications show that such a computation burden reduction does not degrade the generalization capability. It has, however, been open that whether this is true in theory. The aim of this paper is to study the theoretical feasibility of ELM by analyzing the pros and cons of ELM. In the previous part of this topic, we pointed out that via appropriately selected activation functions, ELM does not degrade the generalization capability in the sense of expectation. In this paper, we launch the study in a different direction and show that the randomness of ELM also leads to certain negative consequences. On one hand, we find that the randomness causes an additional uncertainty problem of ELM, both in approximation and learning. On the other hand, we theoretically justify that there also exist activation functions such that the corresponding ELM degrades the generalization capability. In particular, we prove that the generalization capability of ELM with Gaussian kernel is essentially worse than that of FNN with Gaussian kernel. To facilitate the use of ELM, we also provide a remedy to such a degradation. We find that the well-developed coefficient regularization technique can essentially improve the generalization capability. The obtained results reveal the essential characteristic of ELM in a certain sense and give theoretical guidance concerning how to use ELM.**

***Index Terms*— Extreme learning machine (ELM), Gaussian kernel, generalization capability, neural networks.**

I. Introduction

An EXTREME learning machine (ELM) is a feed-forward neural network (FNN) like learning system whose connections with output neurons are adjustable, while the connections with and within hidden neurons are randomly fixed. ELM then transforms the training of an FNN into a linear problem in which only connections with output neurons need adjusting. Thus, the well-known generalized inverse technique [24], [25] can be applied for the solution directly. Due to the fast implementation, ELM has been widely used in regression [11], classification [15], fast object recognition [34], illuminance prediction [8], mill load prediction [30], face recognition [23] and so on.

Compared with the enormous emergences of applications, the theoretical feasibility of ELM is, however, almost vacuum. Up till now, only the universal approximation property of ELM is analyzed [11]–[13], [35]. It is obvious that one of the main reasons of the low computational burden of ELM is that only a few neurons are utilized to synthesize the estimator. Without such an attribution, ELM cannot outperform other learning strategies in implementation. For example, as a special case of ELM, learning in the sample-dependent hypothesis space (the number of neurons equals to the number of samples) [29], [31], [32] cannot essentially reduce the computational complexity. Thus, the universal approximation property of ELM is too weak and cannot capture the essential characteristics of ELM. Therefore, the generalization capability and approximation property of ELM should be investigated. The generalization capability focuses on the relationship between the prediction accuracy and the number of samples, while the approximation property discusses the dependency between the prediction accuracy and the number of hidden neurons.

The aim of this paper is to theoretically verify the feasibility of ELM by analyzing the pros and cons of ELM. In the first part of this topic [20], we casted the analysis of ELM into the framework of statistical learning theory and concluded that with appropriately selected activation functions (polynomial, Nadaraya–Watson, and sigmoid), ELM did not degrade the generalization capability in the expectation sense. This means that, ELM reduces the computational burden without sacrificing the prediction accuracy by selecting appropriate activation function, which can be regarded as the main advantage of ELM. To give a comprehensive feasibility analysis of ELM, we should also study the disadvantage of ELM and, consequently, reveal the essential characteristics of ELM.

Compared with the classical FNN learning [10], this paper shows that there are mainly two disadvantages of ELM. One is that the randomness of ELM causes an additional uncertainty problem, both in approximation and in learning. The other one is that there also exists a generalization degradation phenomenon for ELM with inappropriate activation function. The uncertainty problem of ELM means that there exists an uncertainty phenomenon between the small approximation error (or generalization error) and high confidence of ELM estimator. As a result, it is difficult to judge whether a single time trail of ELM succeeds or not. Concerning the generalization

Manuscript received January 16, 2014; revised June 23, 2014; accepted June 23, 2014. Date of publication July 23, 2014; date of current version December 16, 2014. This work was supported in part by the National 973 Program under Grant 2013CB329404 and in part by the Key Program of National Natural Science Foundation of China under Grant 11131006.

S. Lin is with the College of Mathematics and Information Science, Wenzhou University, Wenzhou 325035, China, and also with the School of Mathematics and Statistics, Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: sblin1983@gmail.com).

X. Liu, J. Fang, and Z. Xu are with the School of Mathematics and Statistics, Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an 710049, China (e-mail: liuxia1232007@163.com; ender86@163.com; zbxu@mail.xjtu.edu.cn).

Digital Object Identifier 10.1109/TNNLS.2014.2336665

degradation phenomenon, we find that with the widely used Gaussian-type activation function (or Gaussian kernel for the sake of brevity), ELM degrades the generalization capability of FNN.

To facilitate the use of ELM, we provide certain remedies to circumvent the aforementioned drawbacks. On one hand, we find that multiple times training can overcome the uncertainty problem of ELM. On the other hand, we show that, by adding neurons and implementing l^2 coefficient regularization simultaneously, the generalization degradation phenomenon of ELM can be avoided. In particular, using l^2 coefficient regularization to determine the connections with output neurons, ELM with Gaussian kernel can reach almost the optimal learning rate of FNN in the sense of expectation, provided the regularization parameter is appropriately tuned.

The study of this paper together with the conclusions in [20] provides a comprehensive feasibility analysis of ELM. To be detailed, the performance of ELM depends heavily on the activation function and the random mechanism. With appropriately selected activation function and random mechanism, ELM does not degrade the generalization capability of FNN learning in the sense of expectation. However, there also exist some activation functions, with which ELM degrades the generalization capability for arbitrary random mechanism. Moreover, due to the randomness, ELM suffers from an uncertainty problem, both in approximation and learning. This paper also shows that both the uncertainty problem and degradation phenomenon are remediable. All these results lay a solid fundamental for ELM and give a guidance of how to use ELM more efficiently.

The rest of this paper is organized as follows. After giving a fast review of ELM, we present an uncertainty problem of ELM approximation in the next section. In Section III, we first introduce the conception of statistical learning theory and then study the generalization capability of ELM with Gaussian kernel. We find that the deduced generalization error bound is larger than that of FNN with Gaussian kernel. This means that ELM with Gaussian kernel may degrade the generalization capability. In Section IV, we provide a remedy to such a degradation. Using the empirical covering number technique, we prove that implementing l_2 coefficient regularization can essentially improve the generalization capability of ELM with Gaussian kernel. In Section V, we give proofs of the main results. We conclude this paper in the final section with some useful remarks.

II. Uncertainty Problem of ELM Approximation

In this section, we study the approximation property of ELM. We find that with the widely used Gaussian-type activation function, ELM suffers from an uncertainty problem.

A. ELM

The ELM, introduced by Huang *et al.* [11] can be regarded as a two-stage FNN learning system that randomly assigns the connections with and within hidden neurons in the first stage and tunes the connections with output neurons in the second stage. Since then, various variants of ELM such as evolutionary ELM [37], Bayesian ELM [27], incremental ELM [14], and regularized ELM [4] were proposed. We refer the readers to a fruitful survey [16] for more information about ELM.

As a two-stage learning scheme, ELM comprises a choice of hypothesis space and a selection of optimization strategy (or learning algorithm) in the first and second stages, respectively. To be precise, in the first stage, ELM picks hidden parameters with and within the hidden neurons randomly to build up the hypothesis space. This makes the hypothesis space of ELM form as

$$\mathcal{H}_{\phi,n} = \left\{ \sum_{j=1}^{n} a_j \phi(w_j, x) : \ a_j \in \mathbf{R} \right\}$$

where w_j's are drawn independently and identically distributed (i.i.d.) according to a specified distribution μ and $x \in \mathbf{R}^d$. It is easy to see that the hypothesis space of ELM is essentially a linear space. In the second stage, ELM tunes the output weights using the well developed linear optimization technique. In this paper, we study the generalization capability of the classical ELM [11] rather than its variants. That is, the linear optimization technique employed in the second stage of ELM is the least square

$$f_{\mathbf{z},\phi,n} = \arg \min_{f \in \mathcal{H}_{\phi,n}} \sum_{i=1}^{m} |f(x_i) - y_i|^2 \tag{1}$$

where $(x_i, y_i)_{i=1}^m$ are the given samples.

B. Uncertainty Problem for ELM Approximation

The randomness of ELM leads to a reduction of computational burden. However, there also exists a certain defect caused by the randomness. The main purpose of this section is to quantify such a defect by studying the approximation capability of ELM with Gaussian kernel.

For this purpose, we introduce a quantity called the modulus of smoothness [5] to measure the approximation capability. The rth modulus of smoothness [5] on $A \subseteq \mathbf{R}^d$ is defined by

$$\omega_{r,A}(f,t) = \sup_{|\mathbf{h}|_2 \le t} \|\Delta^r_{\mathbf{h},A}(f,\cdot)\|_A$$

where $|\cdot|_2$ denotes the Euclidean norm, $\|\cdot\|_A$ denotes the uniform norm on $C(A)$, and the rth difference $\Delta_{\mathbf{h},A}(f,\cdot)$ is defined by

$$\Delta^r_{\mathbf{h},A}(f,x) = \begin{cases} \sum_{j=0}^{r} \binom{r}{j} (-1)^{r-j} f(x+j\mathbf{h}) & \text{if } x \in A_{r,\mathbf{h}} \\ 0 & \text{if } x \notin A_{r,\mathbf{h}} \end{cases}$$

for $\mathbf{h} = (h_1, \ldots, h_d) \in \mathbf{R}^d$ and $A_{r,\mathbf{h}} := \{x \in A : x + s\mathbf{h} \in A,$ for all $s \in [0, r]\}$. It is well known [5] that

$$\omega_{r,A}(f,t) \le \left(1 + \frac{t}{u}\right)^r \omega_{r,A}(f,u) \tag{2}$$

for all $f \in C(A)$ and all $u > 0$.

Let $s \in \mathbf{N}$, we focus on the following Gaussian-type activation function (or Gaussian kernel):

$$K_{\sigma,s}(t) = \sum_{j=1}^{s} \binom{s}{j} (-1)^{1-j} \frac{1}{j^d} \left(\frac{2}{\sigma^2 \pi}\right)^{\frac{d}{2}} \exp\left(-\frac{2t^2}{j^2\sigma^2}\right). \tag{3}$$

Then, the corresponding ELM estimator is defined by

$$f_{\mathbf{z},\sigma,s,n} = \arg\min_{f\in\mathcal{H}_{\sigma,s,n}} \sum_{i=1}^{m} |f(x_i) - y_i|^2 \tag{4}$$

where

$$\mathcal{H}_{\sigma,s,n} = \left\{ \sum_{j=1}^{n} a_j K_{\sigma,s}(\theta_j, x),\ \ x \in I^d \right\}$$
$$K_{\sigma,s}(\theta_j, x) : = K_{\sigma,s}((\theta_j - x)^2) := K_{\sigma,s}(|\theta_j - x|_2^2)$$

where $I^d := [0,1]^d$ and $\{\theta\}_{j=1}^n$ are drawn i.i.d. according to arbitrary fixed distribution μ on the interval $[-a, 1+a]^d$ with $a > 0$.

The following Theorem 1 shows that there exists an uncertainty problem of ELM approximation.

Theorem 1: Let $d, s, n \in \mathbf{N}$. If $f \in C(I^d)$, then with confidence at least $1 - 2\exp\{-cn\sigma^{2d}\}$ (with respect to μ^n), there holds

$$\inf_{g_n\in\mathcal{H}_{\sigma,s,n}} \|f - g_n\|_{I^d} \leq C(\omega_{s,I^d}(f,\sigma) + \|f\|_{I^d}\sigma^d)$$

where c and C are constants depending only on a, d and s.

It follows from Theorem 1 that the approximation capability of ELM with Gaussian kernel depends on the kernel parameters, s, σ, and the number of hidden neurons, n. Furthermore, Theorem 1 shows that, compared with the classical FNN approximation, there exists an additional uncertainty problem of ELM approximation. That is, both the approximation error and the confidence monotonously increase with respect to σ. Therefore, it is impossible to deduce a small approximation error with extremely high confidence. In other words, it is difficult to judge whether the approximation error of ELM is smaller than arbitrary specified approximation accuracy, which does not appear in the classical Gaussian-FNN approximation [33].

We find further in Theorem 1 that the best choice of the kernel parameter, σ, is a tradeoff between the confidence and the approximation error. An advisable way to determine σ is to set $\sigma^{2d} = n^{\varepsilon-1}$ for arbitrary small $\varepsilon \in \mathbf{R}_+$. Under this circumstance, we can deduce that the approximation error of $\mathcal{H}_{\sigma,s,n}$ asymptotically equals to $\omega_{s,I^d}(f, n^{(-1+\varepsilon)/(2d)}) + n^{(-1+\varepsilon)/2}$ with confidence at least $1 - 2\exp(-cn^{\varepsilon})$. Finally, we should verify the optimality of the above approximation bound and therefore justify the optimality of the selected σ. To this end, we introduce the set of rth smoothness functions.

Let $u \in \mathbf{N}_0 := \{0\} \cup \mathbf{N}$, $v \in (0,1]$, and $r = u + v$. A function $f : I^d \to \mathbf{R}$ is said to be rth smooth if for every $\alpha = (\alpha_1, \ldots, \alpha_d), \alpha_i \in N_0$, $\sum_{j=1}^d \alpha_j = u$, the partial derivatives $(\partial^u f)/(\partial x_1{}^{\alpha_1} \cdots \partial x_d{}^{\alpha_d})$ exist and satisfy

$$\left| \frac{\partial^u f}{\partial x_1{}^{\alpha_1} \cdots \partial x_d{}^{\alpha_d}}(x) - \frac{\partial^u f}{\partial x_1{}^{\alpha_1} \cdots \partial x_d{}^{\alpha_d}}(z) \right| \leq c_0 |x - z|_2^v$$

where c_0 is an absolute constant. Denote the set of all rth smooth functions by $\mathcal{F}^r$. Furthermore, for arbitrary $f \in \mathcal{F}^r$, it is easy to deduce [5] that

$$\omega_{s,I^d}(f,t) \leq Ct^r \tag{5}$$

if $s \geq r$. According to Theorem 1 and (5), we obtain that

$$\inf_{g_n\in\mathcal{H}_{\sigma,s,n}} \|f - g_n\|_{I^d} \leq Cn^{\frac{-r+\varepsilon}{2d}} \tag{6}$$

holds with confidence at least $1 - 2\exp\{-cn^{\varepsilon}\}$ for arbitrary $\varepsilon \in \mathbf{R}_+$, provided $f \in \mathcal{F}^r$, $s \geq r$, and $r \leq d$. In the following Proposition 1, we show that the approximation rate (6) cannot be essentially improved, at least for the univariate case.

Proposition 1: Let $d = s = 1, n \in \mathbf{N}$, $\beta > 0$, $0 < \varepsilon < 1$, and $r = 1 - \varepsilon$. If $f_\rho \in \mathcal{F}^r$ and $\sigma = n^{(-1+\varepsilon)/2}$, then with confidence at least $1 - 2\exp\{-cn^{\varepsilon}\}$ (with respect to μ^n), there holds

$$C_1 n^{-\frac{r}{2}-\varepsilon} \leq \sup_{f\in\mathcal{F}^r} \inf_{g_n\in\mathcal{H}_{\sigma,r,n}} \|f - g_n\|_{I^d} \leq C_2 n^{\frac{-r+\varepsilon}{2}}. \tag{7}$$

III. Generalization Degradation Problem of ELM With Gaussian Kernel

Along the flavor of [20], we also analyze the feasibility of ELM in the framework of statistical learning theory [3]. We find in this section that there exists a generalization degradation phenomenon of ELM. In particular, unlike [20], the result in this section shows that ELM with Gaussian kernel degrades the generalization capability of FNN.

A. Fast Review of Statistical Learning Theory

Let $M > 0$, $X = I^d$, $Y \subseteq [-M, M]$ be the input and output spaces, respectively. Suppose that $\mathbf{z} = (x_i, y_i)_{i=1}^m$ is a finite set of random samples drawing i.i.d. according to an unknown but definite distribution ρ, where ρ is assumed to admit the decomposition

$$\rho(x, y) = \rho_X(x)\rho(y|x).$$

Suppose further that $f : X \to Y$ is a function that one uses to model the correspondence between X and Y, as induced by ρ. One natural measurement of the error incurred by using f of this purpose is the generalization error, defined by

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

which is minimized by the regression function [3], defined by

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

We do not know this ideal minimizer f_ρ, since ρ is unknown, but we have access to random examples from $X \times Y$ sampled according to ρ. Let $L^2_{\rho_X}$ be the Hilbert space of ρ_X square integrable function on X, with norm denoted by $\|\cdot\|_\rho$. Then for arbitrary $f \in L^2_{\rho_X}$, there holds

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|_\rho^2 \tag{8}$$

with the assumption $f_\rho \in L^2_{\rho_X}$.

B. Generalization Capability of ELM With Gaussian Kernel

Let $\pi_M f(x) = \min\{M, |f(x)|\}\mathrm{sgn}(f(x))$ be the truncation operator on $f(x)$ at level M. As $y \in [-M, M]$, it is easy to check [36] that

$$\|\pi_M f_{\mathbf{z},\sigma,s,n} - f_\rho\|_\rho \leq \|f_{\mathbf{z},\sigma,s,n} - f_\rho\|_\rho.$$

Thus, the aim of this section is to bound

$$\mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s,n}) - \mathcal{E}(f_\rho) = \|\pi_M f_{\mathbf{z},\sigma,s,n} - f_\rho\|_\rho^2. \tag{9}$$

The error (9), called the generalization error or learning rate, clearly depends on $\mathbf{z}$ and therefore has a stochastic nature. As a result, it is impossible to say anything about (9) in general for a fixed $\mathbf{z}$. Instead, we can look at its behavior in statistics as measured by the expected error

$$\mathbf{E}_{\rho^m}(\|\pi_M f_{\mathbf{z},\sigma,s,n} - f_\rho\|_\rho) := \int_{Z^m} \|\pi_M f_{\mathbf{z},\sigma,s,n} - f_\rho\| d\rho^m$$

where the expectation is taken over all realizations $\mathbf{z}$ obtained for a fixed m, and ρ^m is the m fold tensor product of ρ. In following Theorem 2, we give an upper bound estimate for (9) in the sense of expectation.

Theorem 2: Let $d, s, n, m \in \mathbf{N}$, $\varepsilon > 0$, $r \in \mathbf{R}$, and $f_{\mathbf{z},\sigma,s,n}$ be defined as in (4). If $f_\rho \in \mathcal{F}^r$ with $r \leq s$, $\sigma = m^{(-1+\varepsilon)/(2r+2d)}$ and $n = [m^{(d)/(r+d)}]$, then with probability at least $1 - 2\exp\{-cm^{(\varepsilon d)/(d+r)}\}$ (with respect to μ^n), there holds

$$\mathbf{E}_{\rho^m}(\|\pi_M f_{\mathbf{z},\sigma,s,n} - f_\rho\|_\rho^2) \leq C\left(m^{-\frac{(1-\varepsilon)r}{r+d}} \log m + m^{-\frac{d(1-\varepsilon)}{r+d}}\right) \tag{10}$$

where $[t]$ denotes the integer part of the real number t, c, and C are constants depending only on M, s, r, and d.

It can be found in Theorem 2 that a new quantity ε is introduced to quantify the randomness of ELM. It follows from (10) that ε describes the uncertainty between the confidence and generalization capability. That is, we cannot obtain both extremely small generalization error and high confidence. This means that there also exists an uncertainty problem for ELM learning. Accordingly, Theorem 2 shows that it is reasonable to choose a very small ε, under which circumstance, we can deduce a learning rate close to $m^{(-r)/(r+d)} \log m$ with a tolerable confidence, provided $r \leq d$.

Before drawing the conclusion that ELM with Gaussian kernel degrades the generalization capability, we should verify the optimality of both the established learning rate (10) and the selected parameters such as σ and n. We begin the analysis by illustrating the optimality of the learning rate deduced in (10). For this purpose, we give the following Proposition 2.

Proposition 2: Let $d = s = 1, n, m \in \mathbf{N}$, $\beta > 0$, $0 < \varepsilon < 1$, $r = 1 - \varepsilon$, and $f_{\mathbf{z},\sigma,s,n}$ be defined as in (4). If $f_\rho \in \mathcal{F}^r$, $\sigma = m^{(-1+\varepsilon)/(2r+2)}$, and $n = \left[m^{(1)/(1+r)}\right]$, then with probability at least $1 - 2\exp\{-cm^{(\varepsilon)/(1+r)}\}$ (with respect to μ^n), there holds

$$C_1 m^{-(r)/(1+r)} \leq \mathbf{E}(\|\pi_M f_{\mathbf{z},\sigma,s,n} - f_\rho\|_\rho^2) \leq C_2 m^{-\frac{r(1-\varepsilon)}{1+r}} \log m \tag{11}$$

where c, C_1, and C_2 are constants depending only on r and M.

Modulo an arbitrary small number ε and the logarithmic factor, the upper and lower bounds of (11), are asymptomatically identical. Therefore, the established learning rate in Theorem 2 is almost essential. This means that the established learning rate (10) cannot be essentially improved, at least for the univariate case.

Now, we turn to justify the optimality of the selections of σ and n in Theorem 2. The optimality of σ can be directly derived from the uncertainty problem of ELM. To be detailed, according to Theorem 1 and Proposition 1, the optimal selection of σ is to set $\sigma = n^{(\varepsilon-1)/(2d)}$. Noting that $n = [m^{(d)/(d+r)}]$, it is easy to deduce that the optimal selection of σ is $m^{(-1+\varepsilon)/(2r+2d)}$. Finally, we show the optimality of the parameter n. The main principle to qualify it is the known bias and variance dilemma [3], which declares that a small n may derive a large bias (approximation error), while a large n deduces a large variance (sample error). The best n is thus obtained when the best compromise between the conflicting requirements of small bias and small variance is achieved. In the proof of Theorem 2, we can find that the quantity $n = [m^{d/(r+d)}]$ is selected to balance the approximation and sample errors. Therefore, we can conclude that n is optimal in the sense of bias and variance balance.

Based on the above assertions, we compare Theorem 2 with some related work and propose then the main viewpoint of this section. Imposing the same smooth assumption on the regression function, the optimal learning rate of the FNN with Gaussian kernel was established, where Lin *et al.* [19] deduced that FNNs can achieve the learning rate as $m^{-2r/(2r+d)} \log m$. They also showed that there are $[m^{d/(2r+d)}]$ neurons needed to deduce the almost optimal learning rate. Similarly, Eberts and Steinwart [6] have also built an almost optimal learning rate analysis for the support vector machine (SVM) with the Gaussian kernel. They showed that, modulo an arbitrary small number, both the upper and lower bounds of learning rate of SVM with Gaussian can also attain the optimal learning rate, $m^{-2r/(2r+d)}$. However, Theorem 2 and Proposition 2 imply that the learning rate of ELM with Gaussian kernel cannot be faster than $m^{-r/(r+d)}$. Noting $m^{-2r/(2r+d)} < m^{-r/(r+d)}$ and $m^{d/(2r+d)} < m^{d/(r+d)}$, we find that the prediction accuracy of ELM with Gaussian kernel is much larger than that of FNN even though more neurons are used in ELM. Furthermore, it should be pointed out that if the numbers of utilized neurons in ELM and FNN are identical, then the learning rate of ELM is even worse. If $n = [m^{d/(2r+d)}]$, then the learning rate of ELM with Gaussian kernel cannot be faster than $m^{-r/(2r+d)}$.[1] Therefore, we can draw the conclusion that ELM with Gaussian kernel degrades the generalization capability.

IV. Remedy of the Degradation

As is shown in the previous section, ELM with inappropriately selected activation function suffers from the uncertainty problem and generalization degradation phenomenon. To circumvent the former one, we can employ a multiple training

[1] The proof of this conclusion is the same as that of Theorem 2, we omit it for the sake of brevity.

strategy that has already been proposed in [20]. The main focus of this section is to tackle the generalization capability degradation phenomenon. For this purpose, we use the l^2 coefficient regularization strategy [32] in the second stage of ELM. That is, we implement the following strategy to build up the ELM estimator:

$$f_{\mathbf{z},\sigma,s,\lambda,n} = \arg\min_{f\in\mathcal{H}_{\sigma,s,n}} \left\{\frac{1}{m}\sum_{i=1}^{m}(f(x_i)-y_i)^2 + \lambda\Omega(f)\right\} \quad (12)$$

where $\lambda = \lambda(m) > 0$ is a regularization parameter and

$$\Omega(f) = \inf_{a_i\in\mathbf{R}} \sum_{i=1}^{m} |a_i|^2, \text{ for } f = \sum_{i=1}^{n} a_i K_{\sigma,s}(\theta_i, x) \in \mathcal{H}_{\sigma,s,n}.$$

The following theorem shows that the generalization capability of ELM with Gaussian kernel can be essentially improved using the regularization technique, provided the number of neurons is appropriately adjusted.

Theorem 3: Let $d, s, n, m \in \mathbf{N}$, $\varepsilon > 0$, and $f_{\mathbf{z},\sigma,s,\lambda,n}$ be defined in (12). If $f_\rho \in \mathcal{F}^r$ with $d/2 \le r \le d$, $\sigma = m^{-(1/2r+d)+\varepsilon}$, $n = [m^{2d/2r+d}]$, $s \ge r$, and $\lambda = m^{-2r-d/4r+2d}$, then with confidence at least $1 - 2\exp\{-cm^{\varepsilon d/d+r}\}$ (with respect to μ^n), there holds

$$C_1 m^{\frac{-2r}{2r+d}} \le \mathbf{E}_{\rho^m}\|\pi_M f_{\mathbf{z},\sigma,s,\lambda,n} - f_\rho\|_\rho^2 \le C_2 m^{-\frac{2r}{2r+d}+\varepsilon}\log m \quad (13)$$

where C_1 and C_2 are constants depending only on d, r, s and M.

Theorem 3 shows that, up to an arbitrary small real number ε and the logarithmic factor, the regularized ELM estimator (12) can achieve a learning rate as fast as $m^{-2r/(2r+d)}$ with high probability. Noting that $m^{-2r/(2r+d)} < m^{-r/(r+d)}$ we can draw the conclusion that l^2 coefficient regularization technique can essentially improve the generalization capability of ELM with Gaussian kernel. Furthermore, as is shown above, the best learning rates of both SVM and FNN with Gaussian kernel asymptomatically equal to $m^{-2r/(2r+d)}$. Thus, Theorem 3 shows that the regularization technique not only improves the generalization capability of ELM with Gaussian kernel but also optimizes its generalization capability. In other words, implementing l^2 coefficient regularization in the second stage, ELM with Gaussian kernel can be regarded as an almost optimal FNN learning strategy.

However, it should also be pointed out that the utilized neurons of regularized ELM is much larger than that of the FNN. To obtain the same optimal learning rate, $m^{-2r/(2r+d)}$, there are $[m^{2d/(2r+d)}]$ neurons required in ELM with Gaussian kernel, while the number of utilized neurons in the traditional FNN learning is $[m^{d/(2r+d)}]$. Therefore, although regularized ELM can attain almost the optimal learning rate with high probability, the price to obtain such a rate is higher than that of FNN.

V. Proofs

In this section, we give proofs of the main results presented in the previous two sections.

A. Proof of Theorem 1

To prove Theorem 1, we need the following nine lemmas. The first one can be found in [18], which is an extension of Lemma 2.1 in [33].

Lemma 1: Let $f \in C(I^d)$. There exists an $F \in C(\mathbf{R}^d)$ satisfying

$$F(x) = f(x),\ x \in I^d$$

such that for arbitrary $x \in I^d$, $\|\mathbf{h}\| < \delta \le 1$, there holds

$$\|F\|_\infty := \sup_{x\in\mathbf{R}^d} |F(x)| \le \|f\| = \sup_{x\in I^d} |f(x)|$$

and

$$\omega_{r,\mathbf{R}^d}(F,\delta) \le \omega_{r,I^d}(f,\delta). \quad (14)$$

To state the next lemma, we should introduce a convolution operator concerning the kernel $K_{\sigma,s}$. Denote

$$K_{\sigma,s} * F(x) := \int_{\mathbf{R}^d} F(y)K_{\sigma,s}(x-y)dy.$$

The following Lemma 2 gives an error estimate for the deviation of continuous function and its Gaussian convolution, which can be deduced from [6, Th. 2.2].

Lemma 2: Let $F \in C(\mathbf{R}^d)$ be a bounded and uniformly continuous function defined on $\mathbf{R}^d$. Then

$$\|F - K_{\sigma,s} * F\|_\infty \le C_s\omega_{s,\mathbf{R}^d}(F,\sigma). \quad (15)$$

Let J be arbitrary compact subset of $\mathbf{R}^d$. For $l \ge 0$, denote the set of trigonometric polynomials defined on J with degree at most l by $\mathcal{T}_l^d$. The following Nikol'skii inequality can be found in [2].

Lemma 3: Let $1 \le p < q \le \infty$, $l \ge 1$ be an integer, and $T_l \in \mathcal{T}_l^d$. Then

$$\|T_l\|_{L^q(J)} \le Cl^{\frac{d}{p}-\frac{d}{q}}\|T_l\|_{L^p(J)}$$

where the constant C depends only on d.

For further use, we also should introduce the following probabilistic Bernstein inequality for random variables, which can be found in [3].

Lemma 4: Let ξ be a random variable on a probability space Z with mean $\mathbf{E}(\xi)$ and variance $\gamma^2(\xi) = \gamma_\xi^2$. If $|\xi(z) - \mathbf{E}(\xi)| \le M_\xi$ for almost all $\mathbf{z} \in Z$, then, for all $\varepsilon > 0$

$$\mathbf{P}\left\{\left|\frac{1}{n}\sum_{i=1}^{n}\xi(z_i) - E(\xi)\right| \ge \varepsilon\right\} \le 2\exp\left(-\frac{n\varepsilon^2}{2(\gamma_\xi^2 + \frac{1}{3}M_\xi\varepsilon)}\right)$$

By the help of Lemmas 3 and 4, we are in a position to give the following probabilistic Marcinkiewicz–Zygmund inequality for trigonometric polynomials.

Lemma 5: Let J be a compact subset of $\mathbf{R}^d$ and $0 < p \le \infty$. If $\Xi = \{\theta_i\}_{i=1}^n$ is a set of i.i.d. random variables drawn on J according to arbitrary distribution μ, then

$$\frac{1}{2}\|T_l\|_p^p \le \frac{1}{n}\sum_{i=1}^{n}|T_l(\theta_i)|^p \le \frac{3}{2}\|T_l\|_p^p \quad \forall T_l \in \mathcal{T}_l^d \quad (16)$$

holds with probability at least

$$1 - 2\exp\left(-\frac{C_p n}{l^d}\right)$$

where C_p is a constant depending only on d and p.

Proof: Since we model the sampling set Ξ as a sequence of i.i.d. random variables in J, the sampling points are a sequence of functions $\theta_j = \theta_j(\omega)$ on some probability space $(\mathbf{\Omega}, \mathbf{P})$. Without loss of generality, we assume $\|T_l\|_p = 1$ for arbitrary fixed p. If we set $\xi_j^p(T_l) = |T_l(\theta_j)|^p$, then we have

$$\frac{1}{n}\sum_{i=1}^{n}|T_l(\theta_i)|^p - E\xi_j^p = \frac{1}{n}\sum_{i=1}^{n}|T_l(\theta_i)|^p - \|T_l\|_p^p$$

where we use the equality

$$E\xi_j^p = \int_{\Omega}|T_l(\eta(\omega_j))|^p d\omega_j = \int_J |T_l(\theta)|^p d\theta = \|T_l\|_p^p = 1.$$

Furthermore

$$|\xi_j^p - E\xi_j^p| \le \sup_{\omega\in\Omega}\left||T_l(\theta(\omega))|^p - \|T_l\|_p^p\right| \le \|T_l\|_\infty^p - \|T_l\|_p^p.$$

It follows from Lemma 3 that

$$\|T_l\|_\infty \le Cl^{\frac{d}{p}}\|T_l\|_p = Cl^{\frac{d}{p}}.$$

Hence

$$|\xi_j^p - E\xi_j^p| \le (Cl^d - 1).$$

On the other hand, we have

$$\begin{aligned}\gamma_\xi^2 &= E((\xi_j^p)^2) - (E(\xi_j^p))^2\\ &= \int_\Omega |T_l(\theta(\omega))|^{2p}d\omega - \left(\int_\Omega |T_l(\theta(\omega))|^p d\omega\right)^2\\ &= \|T_l\|_{2p}^{2p} - \|T_l\|_p^{2p}.\end{aligned}$$

Then using Lemma 3 again, there holds

$$\gamma_\xi^2 \le Cl^{2dp(\frac{1}{p}-\frac{1}{2p})}\|T_l\|_p^{2p} - \|T_l\|_p^{2p} = (Cl^d - 1).$$

Thus, it follows from Lemma 4 that with confidence at least

$$\begin{aligned}1 - 2\exp&\left(-\frac{n\varepsilon^2}{2\left(\gamma^2 + \frac{1}{3}M_\xi\varepsilon\right)}\right)\\ &\ge 1 - 2\exp\left(-\frac{n\varepsilon^2}{2\left((Cl^d - 1) + \frac{1}{3}(Cl^d - 1)\varepsilon\right)}\right)\end{aligned}$$

there holds

$$\left|\frac{1}{n}\sum_{i=1}^{n}|T_l(\theta_i)|^p - \|T_l\|_p^p\right| \le \varepsilon.$$

This means that if X is a sequence of i.i.d. random variables, then the Marcinkiewicz–Zygmund inequality

$$(1-\varepsilon)\|T_l\|_p^p \le \frac{1}{n}\sum_{i=1}^{n}|T_l(\theta_i)|^p \le (1+\varepsilon)\|T_l\|_p^p$$

holds with probability at least

$$1 - 2\exp\left(-\frac{cn\varepsilon^2}{l^d(1+\varepsilon)}\right)$$

where c is a constant depending only on d. Then almost the same metric entropy argument as the proof of [1, Theorem 5.1] or the proof of [17, Lemma 7] yields (16) by setting $\varepsilon = 1/2$. ■

To state the next lemma, we need introduce the following definitions. Let $\mathcal{X}$ be a finite-dimensional vector space with norm $\|\cdot\|_{\mathcal{X}}$, and $\mathcal{Z} \subset \mathcal{X}^*$ be a finite set. We say that $\mathcal{Z}$ is a norm generating set for $\mathcal{X}$ if the mapping $T_{\mathcal{Z}} : \mathcal{X} \to \mathbf{R}^{\mathrm{Card}(\mathcal{Z})}$ defined by $T_{\mathcal{Z}}(x) = (z(x))_{z\in\mathcal{Z}}$ is injective, where $\mathrm{Card}(\mathcal{Z})$ is the cardinality of the set $\mathcal{Z}$ and $T_{\mathcal{Z}}$ is named as the sampling operator. Let $W := T_{\mathcal{Z}}(\mathcal{X})$ be the range of $T_{\mathcal{Z}}$, then the injectivity of $T_{\mathcal{Z}}$ implies that $T_{\mathcal{Z}}^{-1} : W \to \mathcal{X}$ exists. Let $\mathbf{R}^{\mathrm{Card}(\mathcal{Z})}$ have a norm $\|\cdot\|_{\mathbf{R}^{\mathrm{Card}(\mathcal{Z})}}$, with $\|\cdot\|_{\mathbf{R}^{\mathrm{Card}(\mathcal{Z})^*}}$ being its dual norm on $\mathbf{R}^{\mathrm{Card}(\mathcal{Z})^*}$. Equipping W with the induced norm, and let $\|T_{\mathcal{Z}}^{-1}\| := \|T_{\mathcal{Z}}^{-1}\|_{W\to\mathcal{X}}$. In addition, let $\mathcal{K}_+$ be the positive cone of $\mathbf{R}^{\mathrm{Card}(\mathcal{Z})}$: that is, all $(r_z) \in \mathbf{R}^{\mathrm{Card}(\mathcal{Z})}$ for which $r_z \ge 0$. Then the following Lemma 6 can be found in [22].

Lemma 6: Let $\mathcal{Z}$ be a norm generating set for $\mathcal{X}$, with $T_{\mathcal{Z}}$ being the corresponding sampling operator. If $y \in \mathcal{X}^*$ with $\|y\|_{\mathcal{X}^*} \le A$, then there exist real numbers $\{a_z\}_{z\in\mathcal{Z}}$, depending only on y such that for every $x \in \mathcal{X}$

$$y(x) = \sum_{z\in\mathcal{Z}} a_z z(x)$$

and

$$\|(a_z)\|_{\mathbf{R}^{\mathrm{Card}(\mathcal{Z})^*}} \le A\|T_{\mathcal{Z}}^{-1}\|.$$

Also, if W contains an interior point $v_0 \in \mathcal{K}_+$ and if $y(T_{\mathcal{Z}}^{-1}v) \ge 0$ when $v \in V \cap \mathcal{K}_+$, then we may choose $a_z \ge 0$.

Using Lemmas 5 and 6, we can deduce the following probabilistic numerical integral formula for trigonometric polynomials.

Lemma 7: Let J be a compact subset of $\mathbf{R}^d$. If $\Xi = \{\theta_i\}_{i=1}^n$ are i.i.d. random variables drawn according to arbitrary distribution μ, then there exists a set of real numbers $\{c_i\}_{i=1}^n$ such that

$$\int_J T_l(x)dx = \sum_{i=1}^{n} c_i T_l(\theta_i) \quad \forall T_l \in \mathcal{T}_l^d$$

holds with confidence at least

$$1 - 2\exp\left(-\frac{C_1 n}{l^d}\right)$$

subject to

$$\sum_{i=1}^{n}|c_i|^2 \le \frac{C}{n}$$

where C_1 and C are constants depending only on d.

Proof: In Lemma 6, we take $\mathcal{X} = \mathcal{T}_l^d$, $\|T_l\|_{\mathcal{X}} = \|T_l\|_p$, and $\mathcal{Z}$ to be the set of point evaluation functionals $\{\delta_{\theta_i}\}_{i=1}^n$. The operator $T_{\mathcal{Z}}$ is then the restriction map $T_l \mapsto T_l|_\Xi$, with

$$\|f\|_{\Xi,p}^p := \begin{cases}\left(\frac{1}{n}\sum_{i=1}^{n}|f(\theta_i)|^p\right)^{\frac{1}{p}} & 0 < p < \infty\\ \sup_{1\le i\le n}\{|f(\theta_i)|\}, & p = \infty.\end{cases}$$

It follows from Lemma 5 with $p = 2$ that with confidence at least

$$1 - 2\exp\left(-\frac{Cn}{l^d}\right)$$

there holds $\|T_{\mathcal{Z}}^{-1}\| \leq 2$. We now take y to be the functional

$$y : T_l \mapsto \int_J T_l(x)dx.$$

By Hölder inequality, $\|y\|_{\mathcal{X}^*} \leq |J|$, where $|J|$ denotes the volume of J. Therefore, Lemma 6 shows that

$$\int_I T_l(x)dx = \sum_{i=1}^n c_i T_l(\theta_i)$$

holds with confidence at least

$$1 - 2\exp\left(-\frac{C_p n}{l^d}\right)$$

subject to

$$\frac{1}{n}\sum_{i=1}^n \left(\frac{|c_i|}{1/n}\right)^2 \leq 2|J|.$$

Therefore, we obtain that $\sum_{i=1}^n |c_i|^2 \leq C/n$, where C is a constant depending only on d. ■

Let $B = [-a, 1+a]^d$ and $\mathcal{P}_l^d$ be the class of algebraic polynomials defined on B with degree at most l. By the help of the above lemma, we can get the following probabilistic numerical integral formula for algebraic polynomials.

Lemma 8: If $\Xi = \{\eta_i\}_{i=1}^n$ are i.i.d. random variables drawn according to arbitrary distribution μ, then there exists a set of real numbers $\{a_i\}_{i=1}^n$ such that

$$\int_B P_l(x)dx = \sum_{i=1}^n a_i P_l(\eta_i) \quad \forall P_l \in \mathcal{P}_l^d$$

holds with confidence at least

$$1 - 2\exp\left(-\frac{C_1 n}{l^d}\right)$$

subject to

$$\sum_{i=1}^m |a_i|^2 \leq \frac{C}{n}$$

where C_1 and C are constants depending only on d.

Proof: Since $x = (x_{(1)}, \dots, x_{(d)})$, we have

$$\int_B f(x)dx = \int_{-a}^{1+a} \cdots \int_{-a}^{1+a} f(x_{(1)}, \dots, x_{(d)})dx_{(1)} \cdots dx_{(d)}.$$

Set $x_{(i)} = (1+|a|)\cos v_i$, $i = 1, \dots, d$, then we have

$$\begin{aligned}\int_B P_l(x)dx &= \int_{-a}^{1+a} \cdots \int_{-a}^{1+a} P_l((1+|a|)\cos v_1, \dots, \\ &\times (1+|a|)\cos v_d)d(1+|a|)\cos v_1 \cdots \\ &\times d(1+|a|)\cos v_d = \int_{J_a} T_{l+d}(v)dv\end{aligned}$$

where J_a is a compact subset of $\mathbf{R}^d$ and

$$\begin{aligned}T_{l+d}(v) = &(-(1+|a|))^d P_l((1+|a|)\cos v_1, \dots, (1+|a|) \\ &\times \cos v_d)\sin v_1 \cdots \sin v_d.\end{aligned}$$

Hence, $T_{l+d} \in \mathcal{T}_{l+d}^d$ and then Lemma 8 can be directly deduced from Lemma 7. ■

By using Lemma 8, we can deduce the following error estimator.

Lemma 9: Let $a > 0$, $u, l \in \mathbf{N}$. If $\Xi := \{\eta_i\}_{i=1}^n$ is a random variable drawing identically and independently according to μ on $[-a, 1+a]$, then with confidence at least $1 - 2\exp\{-cn/(u+l)^d\}$, there holds

$$\begin{aligned}&\inf_{g_n \in \mathcal{H}_{\sigma,s,n}} \|K_{\sigma,s} * F - g_n\| \\ &\quad \leq C_r \left(\omega_{s,I^d}(f, 1/l) + a\|f\|\sigma^d + \sigma^{-d}\frac{2^u}{u!\sigma^2}\right)\end{aligned}$$

where C_s is a constant depending only on d and s.

Proof: For arbitrary $f \in C(I^d)$, let F and $K_{\sigma,s}*F$ defined as in Lemmas 1 and 2, respectively. Then

$$\begin{aligned}K_{\sigma,s} * F &= \int_{\mathbf{R}^d} K_{\sigma,s}(x-y)F(y)dy = \int_B K_{\sigma,s}(x-y)F(y)dy \\ &+ \int_{\mathbf{R}^d - B} K_{\sigma,s}(x-y)F(y)dy.\end{aligned}$$

At first, we give an upper bound estimate for $\int_{\mathbf{R}^d - B} K_{\sigma,s}(x-y)F(y)dy$. It follows from Lemma 1 and the definition of $K_{\sigma,s}$ that

$$\begin{aligned}&\left|\int_{\mathbf{R}^d - B} K_{\sigma,s}(x-y)F(y)dy\right| \\ &\leq \|f\|_{I^d} \sum_{j=1}^s \binom{s}{j}\frac{1}{j^d}\left(\frac{2}{\sigma^2\pi}\right)^{\frac{d}{2}} \times \int_{\mathbf{R}^d - B} \exp\left(-\frac{2\|x-y\|_2^2}{j^2\sigma^2}\right)dy \\ &\leq \|f\|_{I^d} \sum_{j=1}^s \binom{s}{j}\frac{1}{j^d}\left(\frac{2}{\sigma^2\pi}\right)^{\frac{d}{2}} \\ &\quad \times \left(\left(\int_{-\infty}^{-a} + \int_a^{\infty}\right)\exp\left(-\frac{2t^2}{j^2\sigma^2}\right)dt\right)^d \\ &\leq 2\|f\|_{I^d} \sum_{j=1}^s \binom{s}{j}\frac{1}{j^d}\left(\frac{2}{\sigma^2\pi}\right)^{\frac{d}{2}} \times \left(\int_a^{\infty}\exp\left(-\frac{2at}{j^2\sigma^2}\right)dt\right)^d \\ &\leq C_s\|f\|_{I^d} a^{-1}\sigma^d\end{aligned}$$

where C_s is a constant depending only on d and r.

On the other hand, for $F \in C(B)$ and $s \in \mathbf{N}$, it is well known [5] that there exists a $P_l \in \mathcal{P}_l^d$ and absolute constants C_1 and C_2 such that

$$\|F - P_l\| \leq C_1 \inf_{P \in \mathcal{P}_l^d} \|F - P\|_B =: C_1 E_l(F) \tag{17}$$

and

$$\|P_l\|_B \leq C_2\|F\|_B \leq C_2\|f\|_{I^d}. \tag{18}$$

Then, for arbitrary $\{b_i\}_{i=1}^n \subset \mathbf{R}$, there holds

$$\begin{aligned}&\int_B F(y)K_{\sigma,s}(x-y)dy - \sum_{i=1}^n b_i K_{\sigma,s}(x-\eta_i) \\ &\quad = \int_B (F(y) - P_l(y))K_{\sigma,s}(x-y)dy \\ &\qquad + \int_B P_l(y)K_{\sigma,s}(x-y)dy - \sum_{i=1}^n b_i K_{\sigma,s}(x-\eta_i).\end{aligned} \tag{19}$$

Let $u \in \mathbf{N}$. Then, for arbitrary univariate algebraic polynomial q of degree not larger than u, we obtain

$$\begin{aligned}&\int_B P_l(y)K_{\sigma,s}(x-y)dy - \sum_{i=1}^n b_i K_{\sigma,s}(x-\eta_i)\\ &= \int_B P_l(y)(K_{\sigma,s}(x-y) - q(x-y))dt\\ &\quad + \int_B P_l(y)q(x-y)dy - \sum_{i=1}^n b_i(K_{\sigma,s}(x-y) - q(x-\eta_i))\\ &\quad - \sum_{i=1}^n b_i q(x-\eta_i).\end{aligned}$$

Since $P_l(y)q(x-y) \in \mathcal{P}^d_{l+u}(B)$ for fixed x, it follows from Lemma 8 that with confidence at least $1-2\exp\{-cn/(u+l)^d\}$, there exists a set of real numbers $\{w_i\}_{i=1}^n \subset \mathbf{R}$ such that

$$\int_B P_l(y)q(x-y)dy = \sum_{i=1}^n w_i P_l(\eta_i)q(x-\eta_i).$$

If we set $a_i = w_i P_l(\eta_i)$, then

$$\begin{aligned}&\int_B P_l(y)K_{\sigma,s}(x-y)dy - \sum_{i=1}^n a_i K_{\sigma,s}(x-\eta_i)\\ &\qquad = \int_B P_l(y)(K_{\sigma,s}(x-y) - q(x-y))dy\\ &\qquad\quad - \sum_{i=1}^n w_i P_l(\eta_i)(K_{\sigma,s}(x-\eta_i) - q(x-\eta_i))\end{aligned}$$

holds with confidence at least $1-2\exp\{-cn/(u+l)^d\}$. Under this circumstance

$$\begin{aligned}&\left\|\int_B P_l(y)K_{\sigma,s}(\cdot-y)dy - \sum_{i=1}^n a_i K_{\sigma,s}(\cdot-\eta_i)\right\|_{I^d}\\ &\qquad \le \left\|\int_B P_l(y)(K_{\sigma,s}(\cdot-y) - q(\cdot-y))dy\right\|_{I^d}\\ &\qquad\quad + \left\|\sum_{i=1}^n w_i P_l(\eta_i)(K_{\sigma,s}(\cdot-\eta_i) - q(\cdot-\eta_i))\right\|_{I^d}.\end{aligned}$$

To bound the above quantities, denote

$$\mathcal{L}_j(v) := \exp -\frac{2v}{j^2\sigma^2}.$$

Let $\mathcal{T}_u^1([0,(1+a)^2])$ be the set of univariate algebraic polynomials of degrees not larger than u defined on $[0,(1+a)^2]$, and set $q_u^j = \arg\min_{q\in\mathcal{T}_u^1([0,(1+a)^2]^d} \|\mathcal{L}_j - q\|$ and

$$q_u(v) := \sum_{j=1}^s \binom{s}{j}\frac{1}{j^d}\left(\frac{2}{\sigma^2\pi}\right)^{\frac{d}{2}} q_u^j(v).$$

Then, it follows from (18) that

$$\begin{aligned}&\left\|\int_B P_l(y)(K_{\sigma,s}(\cdot-y) - q_u((\cdot-y)^2))dy\right\|_{I^d}\\ &\quad \le C\|f\|_{I^d}\left\|K_{\sigma,s}(\cdot-y) - q_u((\cdot-y)^2)\right\|_{I^d}\\ &\quad \le C\|f\|_{I^d}\sum_{j=1}^r\binom{r}{j}\frac{1}{j^d}\left(\frac{2}{\sigma^2\pi}\right)^{\frac{d}{2}}\\ &\qquad \times \inf_{q\in\mathcal{T}_u^1([0,(1+a)^2])}\|\mathcal{L}_j - q\|.\end{aligned}$$

On the other hand, since

$$\sum_{i=1}^n|w_i| \le \sqrt{n\sum_{i=1}^n|w_i|^2} \le C$$

we also obtain

$$\begin{aligned}&\left\|\sum_{i=1}^n w_i P_l(\eta_i)(K_{\sigma,s}(\cdot-\eta_i) - q_u((\cdot-\eta_i)^2))\right\|\\ &\quad \le C\|f\|_{I_d}\sum_{j=1}^s\binom{s}{j}\frac{1}{j^d}\left(\frac{2}{\sigma^2\pi}\right)^{\frac{d}{2}}\\ &\qquad \times \inf_{q\in\mathcal{T}_u^1([0,(1+a)^2])}\|\mathcal{L}_j - q\|.\end{aligned}$$

Thus, the only thing remainder is to bound $\int_B(F(y) - P_l(y))K_{\sigma,s}(x-y)dy$. It follows from (17) that

$$\begin{aligned}&\left\|\int_B(F(y) - P_l(y))K_{\sigma,s}(x-y)dy\right\|\\ &\quad \le E_l(F)\times\int_B K_{\sigma,s}(x-y)d\mathbf{y} \le C_s\omega_{s,\mathbf{R}^d}(F,1/l)\end{aligned}$$

where we use the fact [6]

$$\int_B K_{\sigma,s}(x-y)d\mathbf{y} \le 1$$

and the known Jackson inequality [5] in the last inequality. All above together with Lemma 1 yields that

$$\begin{aligned}&\inf_{g_n\in\mathcal{G}_n}\|K_{\sigma,s}*F - g_n\| \le C_s\omega_{s,I^d}(f,1/l)\\ &\qquad + C_s a\|f\|\sigma^d + C\|f\|\sum_{j=1}^s\binom{s}{j}\frac{1}{j^d}\left(\frac{2}{\sigma^2\pi}\right)^{\frac{d}{2}}\\ &\qquad \times \inf_{q\in\mathcal{T}_u^1([0,(1+a)^2])}\|\mathcal{L}_j - q\|\end{aligned}$$

holds with confidence at least $1-2\exp\{-cn/(u+l)^d\}$. Furthermore, it is straightforward to check, using the power series [21, p. 136] for $\exp\{-2v/j^2\sigma^2\}$ that

$$\begin{aligned}&\sum_{j=1}^s\binom{s}{j}\frac{1}{j^d}\left(\frac{2}{\sigma^2\pi}\right)^{\frac{d}{2}}\inf_{q\in\mathcal{T}_u^1([0,(1+a)^2])}\|\mathcal{L}_j - q\|\\ &\quad \le C_s\sigma^{-d}\frac{2^u}{u!\sigma^2}.\end{aligned}$$

Thus, the proof of Lemma 9 is completed. ■

By the help of the above nine lemmas, we can proceed the proof of Theorem 1 as follows.

Proof of Theorem 1: Since

$$\inf_{g_n\in\mathcal{H}_{\sigma,s,n}}\|f-g_n\|_{I^d}\le\|f-K_{\sigma,s}*F\|_{I^d}+\|K_{\sigma,s}*F-g_n\|_{I^d}$$

setting $\sigma=l^{-1/2}$, it follows from Lemmas 2 and 9 that

$$\inf_{g_n\in\mathcal{H}_{\sigma,s,n}}\|f-g_n\|_{I^d}\le C_s\left(\omega_{s,I^d}(f,l^{-1/2})+a\|f\|\sigma^d+\sigma^{-d}\frac{(s^2\sigma^2)^u}{2^u u!}\right)$$

holds with confidence at least $1-2\exp(-cn/(u+l)^d)$, where c is a constant depending only on d. By the Stirling's formula, it is easy to check that

$$\sigma^{-d}\frac{(s^2\sigma^2)^u}{2^u u!}\le Cu^d\frac{(u/2)^u}{2^u u!}\le C\frac{u^d}{(2d)^u}\le Cl^{\frac{-d}{2}}$$

with $u=2dl$. Therefore, we obtain

$$\inf_{g_n\in\mathcal{H}_{\sigma,s,n}}\|f-g_n\|\le C\left(\omega_{s,I^d}(f,l^{-1/2})+a\|f\|l^{\frac{-d}{2}}\right)$$

with confidence at least $1-2\exp\{-cn/l^d\}$, where C is a constant depending only on d, s, and a. Therefore, Theorem 1 follows by noticing $\sigma=1/\sqrt{l}$.

B. Proof of Proposition 1

To prove Proposition 1, we need the following two lemmas, the first one concerning Bernstein inequality for $\mathcal{H}_{\sigma,s,n}$ can be easily deduced from [7, eq. (3.1)].

Lemma 10: Let $d=1$, $s=1$, and $\sigma\ge n^{-1/2}$. Then, for arbitrary $g_n\in\mathcal{H}_{\sigma,s,n}$, there holds

$$\|g_n'\|_{[0,1]}\le Cn^{1/2}\|g_n\|_{[0,1]}$$

where C is an absolute constant.

By the help of the Bernstein inequality, the standard method in approximation theory [5, Chap. 7] yields the following Lemma 11.

Lemma 11: Let $d=1$, $s=1$, $r\in\mathbf{N}$, $\sigma\ge n^{-1/2}$ and $f\in C(I^1)$. If

$$\sum_{n=1}^{\infty}n^{r/2-1}\mathrm{dist}(f,\mathcal{H}_{\sigma,1,n})<\infty$$

then $f\in\mathcal{F}^r$, where $\mathrm{dist}(f,\mathcal{H}_{\sigma,1,n})=\inf_{g\in\mathcal{H}_{\sigma,1,n}}\|f-g\|_{I^1}$.

Proof: Let $g_n:=\arg\inf_{g\in\mathcal{H}_{\sigma,1,n}}\|f-g\|_{I^1}$. For arbitrary $n\in\mathbf{N}$, set n_0 such that

$$2^{n_0}\le n\le 2^{n_0+1}.$$

It is easy to see that

$$\sum_{n=1}^{\infty}n^{r/2-1}\mathrm{dist}(f,\mathcal{H}_{\sigma,1,n})<\infty$$

implies $\mathrm{dist}(f,\mathcal{H}_{\sigma,1,n})\to 0$ in $C(I^1)$. If it does not hold, then there exists an absolute constant C such that $\mathrm{dist}(f,\mathcal{H}_{\sigma,1,n})\ge C>0$. Therefore

$$C\sum_{n=1}^{\infty}n^{-1}<\sum_{n=1}^{\infty}n^{\frac{r}{2}-1}\mathrm{dist}(f,\mathcal{H}_{\sigma,1,n})<\infty$$

which is impossible. So we have

$$f-g_{2^{n_0}}=\sum_{j=n_0}^{\infty}g_{2^{j+1}}-g_{2^j}. \tag{20}$$

By Lemma 10, we then have

$$\|g_{2^{j+1}}'-g_{2^j}'\|_{I^1}\le C2^{(j+1)r/2}\mathrm{dist}(f,\mathcal{H}_{\sigma,1,2^j}).$$

Then direct computation yields that

$$\|g_{2^{j+1}}'-g_{2^j}'\|_{I^1}\le C\sum_{j=1}^{\infty}\sum_{k=2^{j-1}+1}^{2^j}k^{r/2-1}\mathrm{dist}(f,\mathcal{H}_{\sigma,1,k})\le C\sum_{k=1}^{\infty}k^{r/2-1}\mathrm{dist}(f,\mathcal{H}_{\sigma,1,k})<\infty.$$

Therefore, $\{g_{2^j}\}$ is the Cauchy sequence of $\mathcal{F}^r$. Differentiating (20), we have

$$f'-g_{2^{n_0}}'=\sum_{j=n_0}^{\infty}g_{2^{j+1}}-g_{2^j}.$$

Since $\{g_{2^j}\}$ is the Cauchy sequence of $\mathcal{F}^r$, we have $f'-g_{2^{n_0}}'\to 0$ when $n_0\to\infty$, which implies $f\in\mathcal{F}^r$. ■

Now we continue the proof of Proposition 1.

Proof of Proposition 1: Let $\varepsilon\in(0,1)$, and $r=1-\varepsilon$. It is obvious that there exists a function h_r satisfying $h_r\in\mathcal{F}^r$ and $h_r\notin F^{r'}$ with $r'>r$. Assume

$$\inf_{g\in\mathcal{H}_{\sigma,1,n}}\|f-g\|\le Cn^{-r/2-\varepsilon}$$

holds for all $f\in\mathcal{F}^r$, where C is a constant independent of n. Then

$$\inf_{g\in\mathcal{H}_{\sigma,1,n}}\|h_r-g\|\le Cn^{-r/2-\varepsilon}.$$

Then

$$\sum_{n=1}^{\infty}n^{1/2-1}\mathrm{dist}(h_r,\mathcal{H}_{\sigma,1,n})=\sum_{n=1}^{\infty}n^{-1-\varepsilon/2}<\infty.$$

Therefore, it follows from Lemma 11 that $h_r\in\mathcal{F}^1$, which is impossible. Hence

$$\sup_{f\in\mathcal{F}^r}\inf_{g\in\mathcal{H}_{\sigma,1,n}}\|f-g\|\ge Cn^{-r/2-\varepsilon}.$$

This together with Theorem 1 finishes the proof of Proposition 1.

C. Proof of Theorem 2

The main tool to prove Theorem 2 is the following Lemma 12, which can be found in [9, Ch. 11].

Lemma 12: Let $f_{\mathbf{z},\sigma,s,n}$ be defined as in (4). Then

$$E_{\rho^m}\|\pi_M f_{\mathbf{z},\sigma,s,n}-f_\rho\|_\rho^2\le CM^2\frac{(\log m+1)n}{m}+8\inf_{f\in\mathcal{H}_{\sigma,s,n}}\int_X|f(x)-f_\rho(x)|^2d\rho_X \tag{21}$$

for some universal constant C.

Now, we use Proposition 1 and Lemma 12 to prove Theorem 2.

Proof of Theorem 2: Since $\mathcal{H}_{\sigma,s,n}$ is a n-dimensional linear space, then Lemma 12 yields that

$$E_{\rho^m}\|\pi_M f_{\mathbf{z},\sigma,s,n} - f_\rho\|_\rho^2 \le CM^2\frac{(\log m+1)n}{m} + 8\inf_{f\in\mathcal{H}_{\sigma,s,n}}\int_X |f(x)-f_\rho(x)|^2 d\rho.$$

Therefore, it suffices to bound

$$\inf_{f\in\mathcal{H}_{\sigma,s,n}}\int_X |f(x)-f_\rho(x)|^2 \le \inf_{f\in\mathcal{H}_{\sigma,s,n}}\|f-f_\rho\|_X^2.$$

From Theorem 1, it follows that

$$\inf_{g\in\mathcal{H}_{\sigma,s,n}}\|g-f_\rho\|_X \le C\left(\omega_{s,I^d}(f_\rho,\sigma)+\|f_\rho\|\sigma^d\right)$$

holds with probability at least $1-2\exp\{-cn\sigma^{2d}\}$. Noting $r\le s$ and $f_\rho\in\mathcal{F}^r$, with probability at least $1-2\exp\{-cn\sigma^{2d}\}$, there holds

$$\inf_{f\in\mathcal{H}_{\sigma,s,n}}\|f-f_\rho\|_X^2 \le C(\sigma^{2r}+\sigma^{2d}).$$

Setting $\sigma = n^{(-1+\varepsilon)/(2d)}$, we observe that with probability at least $1-2\exp\{-n^\varepsilon\}$, there holds

$$\inf_{f\in\mathcal{H}_{\sigma,s,n}}\|f-f_\rho\|_X^2 \le C\left(n^{-r/d+r\varepsilon/d}+n^{-1+\varepsilon}\right).$$

Finally, choosing $n=[m^{(d)(r+d)}]$, we obtain that with probability at least $1-2\exp\{-n^\varepsilon\}$, there holds

$$E_{\rho^m}\|\pi_M f_{\mathbf{z},\sigma,s,n}-f_\rho\|_\rho^2 \le C\left(m^{-\frac{(1-\varepsilon)r}{r+d}}\log m + m^{-\frac{d(1-\varepsilon)}{r+d}}\right).$$

This finishes the proof of Theorem 2.

D. Proof of Proposition 2

To prove Proposition 2, we need the following three lemmas. The first one is the interpolation theorem of linear functionals, which can be found in [2, p. 385].

Lemma 13: Let $C(Q)$ be the set of real valued continuous functions on the compact Hausdorff space Q. Let S be an n-dimensional linear subspace of $C(Q)$ over $\mathbf{R}$. Let $L\neq 0$ be a real-valued linear functional on S. Then there exist points $x_1,x_2,\ldots,x_r$ in Q and nonzero real numbers $a_1,a_2,\ldots,a_r$, where $1\le r\le n$, such that

$$L(s)=\sum_{i=1}^r a_i s(x_i),\quad s\in S$$

and

$$\|L\| = \sup\{|L(s)| : s\in S,\ \|s\|_Q\le 1\} = \sum_{i=1}^r |a_i|.$$

By using Lemmas 13 and 10, we can obtain the following Bernstein inequality for ELM with Gaussian kernel in the metric of $L^2_{\rho_X}$.

Lemma 14: Let $d=1$, $s=1$, and $\sigma\ge n^{-1/2}$. Then, for arbitrary $g_n\in\mathcal{H}_{\sigma,s,n}$, there holds

$$\|g_n'\|_\rho \le Cn^{1/2}\|g_n\|_\rho$$

where C is an absolute constant.

Proof: We apply Lemma 13 with $Q=[1/2,1]$, $S=\mathcal{H}_{\sigma,s,n}$, and $L(s)=s'(1)$. It follows from Lemma 10 that

$$\|L\| = |s'(1)| \le Cn^{1/2}|s(1)| = C_1 n^{1/2}. \tag{22}$$

We deduce that there are $v_1,v_2,\ldots,v_r$ in $[1/2,1]$ and $a_1,a_2,\ldots,a_r\in I^1$ so that for every $s\in\mathcal{H}_{\sigma,s,n}$

$$\frac{|s'(1)|}{C_1n^{1/2}} = \frac{|\sum_{i=1}^r a_i s(v_i)|}{C_1 n^{1/2}} \le \sum_{i=1}^r\left|\frac{a_i}{C_1n^{1/2}}\right||s(v_i)|$$

with $1\le r\le n$. By (22) we have

$$\sum_{i=1}^r\left|\frac{a_i}{C_1n^{1/2}}\right| \le 1.$$

Therefore, there is a sequence of numbers $\{c_i\}$ with $\sum_{i=1}^r |c_i|=1$ such that

$$\frac{|s'(1)|}{C_1n^{1/2}} \le \sum_{i=1}^r |c_i||s(v_i)|.$$

Now let $\phi:[0,\infty)\to[0,\infty)$ be a nondecreasing convex function. Using monotonicity and convexity, we have

$$\phi\left(\frac{|s'(1)|}{C_1n^{1/2}}\right) \le \phi(\sum_{i=1}^r |c_i s(v_i)|) \le \sum_{i=1}^r |c_i|\phi(|s(v_i)|).$$

Applying this inequality with $s(t)=g_n(t+u-1)\in\mathcal{H}_{\sigma,s,n}$, we get

$$\phi\left(\frac{|g_n'(u)|}{C_1n^{1/2}}\right) \le \sum_{i=1}^r |c_i|\phi(|P(v_i+u-b)|)$$

for every $P\in\mathcal{H}_{\sigma,s,n}$ and $u\in[1/2,1]$. Since $x_i\in[1/2,1]$ and $u\in[1/2,1]$, then $v_i+u-1\in[0,1]$ for each $i=1,2,\ldots,r$. Integrating on the interval $[1/2,1]$ with respect to u, we obtain

$$\begin{aligned}&\int_{1/2}^1 \phi\left(\frac{|g_n'(u)|}{C_1n^{1/2}}\right)d\rho_X(u)\\ &\le \sum_{i=1}^r\int_{1/2}^1 |c_i|\phi(|g_n(v_i+u-1)|)d\rho_X(u)\\ &\le \sum_{i=1}^r\int_0^1 |c_i|\phi(|g_n(t)|)d\rho_X(t) \le \int_0^1\phi(|g_n(t)|)dt\end{aligned}$$

in which $\sum_{i=1}^r |c_i|=1$ has been used.

It can be shown exactly in the same way that

$$\int_0^{1/2}\phi\left(\frac{|g_n'(u)|}{C_1\lambda_n}\right)d\rho_X(u) \le \int_0^1 \phi(|g_n(t)|)d\rho_X(t).$$

Combining the last two inequalities and choosing $\phi(x)=x^2$, we finish the proof of Lemma 14. ∎

Using almost the same method as that in the proof of Lemma 11, the following Lemma 15 can be deduced directly from Lemma 14.

Lemma 15: Let $d=1$, $s=1$, $r\in\mathbf{N}$, $\sigma\ge n^{-1/2}$, and $f\in C(I^1)$. If

$$\sum_{n=1}^\infty n^{r/2-1}\mathrm{dist}(f,\mathcal{H}_{\sigma,1,n})_\rho < \infty$$

then $f \in \mathcal{F}^r$, where $\mathrm{dist}(f, \mathcal{H}_{\sigma,1,n})_\rho = \inf_{g\in\mathcal{H}_{\sigma,1,n}} \|f-g\|_\rho$.

Now, we proceed the proof of Proposition 2.

Proof of Proposition 2: With the help of the above lemmas, we can use the almost same method as that in the proof of Proposition 1 to obtain

$$\sup_{f\in\mathcal{F}^r}\inf_{g\in\mathcal{H}_{\sigma,1,n}} \|f-g\|_\rho \geq Cn^{-r/2-\varepsilon}.$$

Then, Proposition 2 can be deduced from the above inequality using the conditions, $\sigma = m^{(-1+\varepsilon)(2+2r)}$ and $n = [m^{(1)(1+r)}]$.

E. Proof of Theorem 3

To prove Theorem 3, we need the following concepts and lemmas. Let $(\mathcal{M}, \tilde{d})$ be a pseudometric space and $T \subset \mathcal{M}$ a subset. For every $\varepsilon > 0$, the covering number $\mathcal{N}(T, \varepsilon, \tilde{d})$ of T with respect to ε and $\tilde{d}$ is defined as the minimal number of balls of radius ε whose union covers T, that is

$$\mathcal{N}(T,\varepsilon,\tilde{d}) := \min\left\{l\in\mathbf{N} : T\subset\bigcup_{j=1}^{l} B(t_j,\varepsilon)\right\}$$

for some $\{t_j\}_{j=1}^l \subset \mathcal{M}$, where $B(t_j,\varepsilon) = \{t\in\mathcal{M} : \tilde{d}(t,t_j)\leq\varepsilon\}$. The l^2-empirical covering number [29] of a function set is defined by means of the normalized l^2-metric $\tilde{d}_2$ on the Euclidean space $\mathbf{R}^d$ given in with $\tilde{d}_2(\mathbf{a},\mathbf{b}) = (1/m\sum_{i=1}^m |a_i - b_i|^2)^{1/2}$ for $\mathbf{a} = (a_i)_{i=1}^m, \mathbf{b} = (b_i)_{i=1}^m \in \mathbf{R}^m$.

Definition 1: Let $\mathcal{G}$ be a set of functions on X, $\mathbf{x} = (x_i)_{i=1}^m$, and

$$\mathcal{G}|_{\mathbf{x}} := \{(f(x_i))_{i=1}^m : f\in\mathcal{G}\}\subset R^m.$$

Set $\mathcal{N}_{2,\mathbf{x}}(\mathcal{G},\varepsilon) = \mathcal{N}(\mathcal{G}|_{\mathbf{x}},\varepsilon,\tilde{d}_2)$. The l^2-empirical covering number of $\mathcal{G}$ is defined by

$$\mathcal{N}_2(\mathcal{F},\varepsilon) := \sup_{m\in\mathbf{N}}\sup_{\mathbf{x}\in S^m}\mathcal{N}_{2,\mathbf{x}}(\mathcal{G},\varepsilon),\quad \varepsilon>0.$$

Let H_σ be the reproducing kernel Hilbert space of $K_{\sigma,s}$ [28] and B_{H_σ} be the unit ball in H_σ. The following Lemmas 16 and 17 can be easily deduced from [28, Th. 2.1] and [29], respectively.

Lemma 16: Let $0 < \sigma \leq 1$, $X \subset \mathbf{R}^d$ be a compact subset with nonempty interior. Then for all $0 < p \leq 2$ and all $\nu > 0$, there exists a constant $C_{p,\nu,d,s} > 0$ independent of σ such that for all $\varepsilon > 0$, we have

$$\log\mathcal{N}_2(B_{H_\sigma},\varepsilon)\leq C_{p,\mu,d,s}\sigma^{(p/2-1)(1+\nu)d}\varepsilon^{-p}.$$

Lemma 17: Let $\mathcal{F}$ be a class of measurable functions on Z. Assume that there are constants B, $c > 0$, and $\alpha\in[0,1]$ such that $\|f\|_\infty \leq B$ and $\mathbf{E}f^2 \leq c(\mathbf{E}f)^\alpha$ for every $f\in\mathcal{F}$. If for some $a > 0$ and $p\in(0,2)$

$$\log\mathcal{N}_2(\mathcal{F},\varepsilon)\leq a\varepsilon^{-p}\quad\forall\varepsilon>0 \tag{23}$$

then there exists a constant c'_p depending only on p such that for any $t > 0$, with probability at least $1 - e^{-t}$, there holds

$$\mathbf{E}f - \frac{1}{m}\sum_{i=1}^m f(z_i) \leq 4\frac{1}{2}\eta^{1-\alpha}(\mathbf{E}f)^\alpha + c'_p\eta + 2\left(\frac{ct}{m}\right)^{\frac{1}{2-\alpha}} + \frac{18Bt}{m}\quad\forall f\in\mathcal{F} \tag{24}$$

where

$$\eta := \max\left\{c^{\frac{2-p}{4-2\alpha+p\alpha}}\left(\frac{a}{m}\right)^{\frac{2}{4-2\alpha+p\alpha}}, B^{\frac{2-p}{2+p}}\left(\frac{a}{m}\right)^{\frac{2}{2+p}}\right\}.$$

The next lemma states a variant of Lemma 4, which can be found in [26].

Lemma 18: Let ξ be a random variable on a probability space Z with variance γ^2 satisfying $|\xi - \mathbf{E}\xi| \leq M_\xi$ for some constant M_ξ. Then for any $0 < \delta < 1$, with confidence $1-\delta$, we have

$$\frac{1}{m}\sum_{i=1}^m\xi(z_i) - \mathbf{E}\xi \leq \frac{2M_\xi\log\frac{1}{\delta}}{3m} + \sqrt{\frac{2\sigma^2\log\frac{1}{\delta}}{m}}.$$

From the proof of Lemma 9, we can also deduce the following Lemma 19.

Lemma 19: Let $d, s, n\in\mathbf{N}$. Then, with confidence at least $1 - 2\exp\{-cn\sigma^{2d}\}$, there exists a $f_0\in\mathcal{H}_{\sigma,s,n}$ such that

$$\|f_\rho - f_0\|_\rho^2 + \lambda\Omega(f_0) \leq C\left(\omega_{s,I^d}(f_\rho,\sigma)^2 + \sigma^{2d} + \frac{\lambda}{n}\right)$$

where C is a constant depending only on d, s and M.

Proof: Let

$$f_0 = \sum_{i=1}^n a_iK_{\sigma,s}(x-\eta_i) = \sum_{i=1}^n w_iP_l(\eta_i)K_{\sigma,s}(x-\eta_i)$$

where $\{w_i\}_{i=1}^n$ and P_l are the same as those in the proof of Lemma 9. Then, it has already been proved that

$$\|f_\rho - f_0\|_\rho \leq C(\omega_{s,I^d}(f_\rho,\sigma) + \sigma^d).$$

Furthermore, it can be deduced from Lemma 8 and (18) by taking $f = f_\rho$ that

$$\Omega(f_0) = \sum_{i=1}^n |w_i|^2|P_l(\eta_i)|^2 \leq \|f_\rho\|^2\sum_{i=1}^n|w_i|^2 \leq \frac{C}{n}.$$

This finishes the proof of Lemma 19. ■

Now we proceed the proof of Theorem 3.

Proof of Theorem 3: Let $f_{\mathbf{z},\sigma,s,\lambda,n}$ and f_0 be defined as in (12) and Lemma 19, respectively. Define

$$\mathcal{D} := \mathcal{E}(f_0) - \mathcal{E}(f_\rho) + \lambda\Omega(f_0)$$

and

$$\mathcal{S} := \mathcal{E}_{\mathbf{z}}(f_0) - \mathcal{E}(f_0) + \mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s,\lambda,n}) - \mathcal{E}_{\mathbf{z}}(\pi_M f_{\mathbf{z},\sigma,s,\lambda,n})$$

where $\mathcal{E}_{\mathbf{z}}(f) = 1/m\sum_{i=1}^m(y_i - f(x_i))^2$. Then, it is easy to check that

$$\mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s,\lambda,n}) - \mathcal{E}(f_\rho) \leq \mathcal{D} + \mathcal{S}. \tag{25}$$

As $f_\rho\in\mathcal{F}^r$, it follows from Lemma 19 that with confidence at least $1 - 2\exp\{-cn\sigma^{2d}\}$ (with respect to μ^n), there holds

$$\mathcal{D} \leq C\left(\sigma^{2r} + \sigma^{2d} + \frac{\lambda}{n}\right). \tag{26}$$

Upon using the short hand notations

$$\mathcal{S}_1 := \{\mathcal{E}_{\mathbf{z}}(f_0) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} - \{\mathcal{E}(f_0) - \mathcal{E}(f_\rho)\}$$

and

$$\mathcal{S}_2 := \{\mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s,\lambda,n}) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(\pi_M f_{\mathbf{z},\sigma,s,\lambda,n}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}$$

we have

$$\mathcal{S} = \mathcal{S}_1 + \mathcal{S}_2. \tag{27}$$

We first turn to bound $\mathcal{S}_1$. Let the random variable ξ on Z be defined by

$$\xi(\mathbf{z}) = (y - f_0(x))^2 - (y - f_\rho(x))^2 \quad \mathbf{z} = (x, y) \in Z.$$

Since $|f_\rho(x)| \le M$ and

$$|f_0| \le \sum_{i=1}^n |w_i||P_l(\eta_i)||K_{\sigma,s}(\eta_i, x)| \le \|f_\rho\| \sum_{i=1}^n |w_i| \le CM$$

hold almost everywhere, we have

$$\begin{aligned}|\xi(\mathbf{z})| &= (f_\rho(x) - f_0(x))(2y - f_0(x) - f_\rho(x)) \\ &\le (M + CM)(3M + CM) \le M_\xi := (3M + CM)^2\end{aligned}$$

and almost surely

$$|\xi - \mathbf{E}\xi| \le 2M_\xi.$$

Moreover, we have

$$\begin{aligned}E(\xi^2) &= \int_Z (f_0(x) + f_\rho(x) - 2y)^2 (f_0(x) - f_\rho(x))^2 d\rho \\ &\le M_\xi \|f_\rho - f_0\|_\rho^2\end{aligned}$$

which implies that the variance γ^2 of ξ can be bounded as $\gamma^2 \le E(\xi^2) \le M_\xi \mathcal{D}$. Now applying Lemma 18, we obtain

$$\begin{aligned}\mathcal{S}_1 &\le \frac{4M_\xi \log\frac{2}{\delta}}{3m} + \sqrt{\frac{2M_\xi \mathcal{D} \log\frac{2}{\delta}}{m}} \\ &\le \frac{7(3M + CM)^2 \log\frac{2}{\delta}}{3m} + \frac{1}{2}\mathcal{D}\end{aligned} \tag{28}$$

holds with confidence $1 - (\delta/2)$ (with respect to ρ^m).

To bound $\mathcal{S}_2$, we need apply Lemma 17 to the set $\mathcal{G}_R$, where

$$\mathcal{G}_R := \left\{(y - \pi_M f(x))^2 - (y - f_\rho(x))^2 : f \in \mathcal{B}_R\right\}$$

and

$$\mathcal{B}_R := \left\{ f = \sum_{i=1}^n b_i K_{\sigma,s}(\eta_i, x) : \sum_{i=1}^n |b_i|^2 \le R \right\}.$$

Each function $g \in \mathcal{G}_R$ has the form

$$g(z) = (y - \pi_M f(x))^2 - (y - f_\rho(x))^2, \quad f \in \mathcal{B}_R$$

and is automatically a function on Z. Hence

$$\mathbf{E}g = \mathcal{E}(f) - \mathcal{E}(f_\rho) = \|\pi_M f - f_\rho\|_\rho^2$$

and

$$\frac{1}{m}\sum_{i=1}^m g(z_i) = \mathcal{E}_{\mathbf{z}}(\pi_M f) - \mathcal{E}_{\mathbf{z}}(f_\rho)$$

where $z_i := (x_i, y_i)$. Observe that

$$g(z) = (\pi_M f(x) - f_\rho(x))((\pi_M f(x) - y) + (f_\rho(x) - y)).$$

Therefore

$$|g(z)| \le 8M^2$$

and

$$\begin{aligned}\mathbf{E}g^2 &= \int_Z (2y - \pi_M f(x) - f_\rho(x))^2 (\pi_M f(x) - f_\rho(x))^2 d\rho \\ &\le 16M^2 \mathbf{E}g.\end{aligned}$$

For $g_1, g_2 \in \mathcal{F}_{R_q}$ and arbitrary $m \in \mathbf{N}$, we have

$$\begin{aligned}&\left(\frac{1}{m}\sum_{i=1}^m (g_1(z_i) - g_2(z_i))^2\right)^{1/2} \\ &\le \left(\frac{4M}{m}\sum_{i=1}^m (f_1(x_i) - f_2(x_i))^2\right)^{1/2}.\end{aligned}$$

It follows that

$$\mathcal{N}_{2,\mathbf{z}}(\mathcal{G}_R, \varepsilon) \le \mathcal{N}_{2,\mathbf{x}}\left(\mathcal{B}_R, \frac{\varepsilon}{4M}\right) \le \mathcal{N}_{2,\mathbf{x}}\left(\mathcal{B}_1, \frac{\varepsilon}{4MR}\right)$$

which together with Lemma 16 implies

$$\log \mathcal{N}_{2,\mathbf{z}}(\mathcal{G}_R, \varepsilon) \le C_{p,\mu,d} \sigma^{\frac{p-2}{2}(1+\nu)d} (4MR)^p \varepsilon^{-p}.$$

By Lemma 17 with $B = c = 16M^2$, $\alpha = 1$ and $a = C_{p,\mu,d}\sigma^{(p-2/2)(1+\nu)d}(4MR)^p$, we know that for any $\delta \in (0, 1)$, with confidence $1 - \delta/2$, there exists a constant C depending only on d such that for all $g \in \mathcal{G}_R$

$$\mathbf{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i) \le \frac{1}{2}\mathbf{E}g + C\eta + C(M+1)^2 \frac{\log(4/\delta)}{m}.$$

Here

$$\eta = \{16M^2\}^{\frac{2-p}{2+p}} C_{p,\nu,d}^{\frac{2}{2+p}} m^{-\frac{2}{2+p}} \sigma^{\frac{p-2}{2}(1+\nu)d\frac{2}{2+p}} R^{\frac{2p}{2+p}}.$$

Hence, we obtain

$$\begin{aligned}\mathbf{E}g - \frac{1}{m}\sum_{i=1}^m g(z_i) &\le \frac{1}{2}\mathbf{E}g + \{16(M+1)^2\}^{\frac{2-p}{2+p}} C_{p,\nu,d}^{\frac{2}{2+p}} \\ &\quad \times m^{-\frac{2}{2+p}} \sigma^{\frac{p-2}{2}(1+\nu)d\frac{2}{2+p}} R^{\frac{2p}{2+p}} \log\frac{4}{\delta}.\end{aligned}$$

Now we turn to estimate R. It follows form the definition of $f_{\mathbf{z},\sigma,s,\lambda,n}$ that

$$\lambda\Omega(f_{\mathbf{z},\sigma,s,\lambda,n}) \le \mathcal{E}_{\mathbf{z}}(0) + \lambda \cdot 0 \le M^2.$$

Thus, we obtain that for arbitrary $0 < p \le 2$ and arbitrary $\nu > 0$, there exists a constant C depending only on d, ν, p, and M such that

$$\begin{aligned}\mathcal{S}_2 &\le \frac{1}{2}\{\mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s\lambda,n}) - \mathcal{E}(f_\rho)\} \\ &\quad + C\log\frac{4}{\delta} m^{-\frac{2}{2+p}} \sigma^{\frac{(p-2)(1+\nu)d}{2+p}} \lambda^{\frac{-2p}{2+p}}\end{aligned} \tag{29}$$

with confidence at least $1 - (\delta)/(2)$ (with respect to ρ^m).

From (25) to (29), we obtain

$$\begin{aligned}\mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s\lambda,n}) - \mathcal{E}(f_\rho) &\\ \leq C\Bigg(\sigma^{2r} + \sigma^{2d} + \lambda/n + \frac{\log\frac{4}{\delta}}{3m} &\\ + \frac{1}{2}\{\mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s\lambda,n}) - \mathcal{E}(f_\rho)\} &\\ + \log\frac{4}{\delta} m^{-\frac{2}{2+p}} \sigma^{\frac{(p-2)(1+\nu)d}{2+p}} \lambda^{\frac{-2p}{2+p}}\Bigg)&\end{aligned}$$

holds with confidence at least $(1-\delta)\times(1-2\exp\{-cn\sigma^{2d}\})$ (with respect to $\rho^m \times \mu^n$).

Set $\sigma = m^{-1/2r+d+\varepsilon}$, $n = m^{2d}/2r+d$, $\lambda = m^{-a} := m^{-2r-d/4r+2d}$, $\nu = \varepsilon/2d(2r+d)$, and

$$p = \frac{2d + 2\varepsilon(2r+d) - 2(1+\nu) + 2(2r+d)\varepsilon(1+\nu)d}{(2r+d)(2a+d\varepsilon+\nu d\varepsilon-\varepsilon)+2r-(1+\nu)d}.$$

Since $r \geq d/2$, it is easy to check that $\nu > 0$, and $0 < p \leq 2$. Then, we get

$$\begin{aligned}\mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s\lambda,n}) - \mathcal{E}(f_\rho) \leq C m^{-\frac{2r}{2r+d}+\varepsilon}\log 4\delta &\\ + m^{-\frac{2d}{2r+d}+\varepsilon} + \log 4\delta m^{-\frac{2r+3d}{4r+d}}.&\end{aligned}$$

Noting further that $r \leq d$, we obtain

$$\mathcal{E}(\pi_M f_{\mathbf{z},\sigma,s\lambda,n}) - \mathcal{E}(f_\rho) \leq C m^{-\frac{2r}{2r+d}+\varepsilon}\log 4\delta.$$

Noticing the identity

$$E_{\rho^m}(\mathcal{E}(f_\rho) - \mathcal{E}(f_{\mathbf{z},\lambda,q})) = \int_0^\infty P^m\{\mathcal{E}(f_\rho) - \mathcal{E}(f_{\mathbf{z},\lambda,q}) > \varepsilon\}d\varepsilon$$

direct computation yields the upper bound of (13). The lower bound can be found in [9, Chap.3]. This finishes the proof of Theorem 3.

VI. Conclusion

The ELM-like learning provides a powerful computational burden reduction technique that adjusts only the output connections. Numerous experiments and applications have demonstrated the effectiveness and efficiency of ELM. The aim of this paper is to provide theoretical fundamentals of it. After analyzing the pros and cons of ELM, we found that the theoretical performance of ELM depends heavily on the activation function and randomness mechanism. In the previous cousin paper [20], we have provided the advantages of ELM in theory, that is, with appropriately selected activation function, ELM reduces the computation burden without sacrificing the generalization capability in the sense of expectation. In this paper, we discussed certain disadvantages of ELM. Via rigorous proof, we found that ELM suffered from both the uncertainty and generalization degradation problem. Indeed, we proved that for the widely used Gaussian-type activation function, ELM degraded the generalization capability. To facilitate the use of ELM, some remedies of the aforementioned two problems are also recommended. That is, multiple time trials can avoid the uncertainty problem and the l^2 coefficient regularization technique can essentially improve the generalization capability of ELM. All these results reveal the essential characteristics of ELM learning and give a feasible guidance concerning how to use ELM .

We conclude this paper with a crucial question about ELM learning.

Question 1: As is shown in [20] and the current paper, the performance of ELM depends heavily on the activation function. For appropriately selected activation function, ELM does not degrade the generalization capability, while there also exists an activation function such that the degradation exists. As it is impossible to enumerate all the activation functions and study the generalization capabilities of the corresponding ELM, we are asked for a general condition on the activation function, under which the corresponding ELM degrade (or does not degrade) the generalization capability. In other words, we are interested in a criterion to classify the activation functions into two classes. With the first class, ELM degrades the generalization capability and with the other class, ELM does not degrade the generalization capability. We will keep working on this interesting project and report our progress in a future publication.

Acknowledgment

Three anonymous referees have carefully read this paper and have provided to us numerous constructive suggestions. As a result, the overall quality of the paper has been noticeably enhanced, to which we feel much indebted and are grateful.

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] P. B. Borwein and T. Erdélyi, *Polynomials and Polynomial Inequalities*. New York, NY, USA: Springer-Verlag, 1995.

[3] F. Cucker and S. Smale, “On the mathematical foundations of learning,” *Bull. Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2001.

[4] W. Deng, Q. Zheng, and L. Chen, “Regularized extreme learning machine,” in *Proc. IEEE Symp. Comput. Inell. Data Mining (CIDM)*, Mar./Apr. 2009, pp. 389–395.

[5] R. DeVore and G. Lorentz, *Constructive Approximation*. Berlin, Germany: Springer-Verlag, 1993.

[6] M. Eberts and I. Steinwart, “Optimal learning rates for least squares SVMs using Gaussian kernels,” in *Advances in Neural Information Processing Systems*, vol. 24. Dec. 2011, pp. 1539–1547.

[7] T. Erdélyi, “Bernstein-type inequalities for linear combinations of shifted Gaussian,” *Bull. London Math. Soc.*, vol. 38, no. 1, pp. 124–138, 2006.

[8] S. Ferrari *et al.*, “Illuminance prediction through extreme learning machines,” in *Proc. IEEE Workshop Environ. Energy Struct. Monitor. Syst. (EESMS)*, Sep. 2012, pp. 97–103.

[9] L. Györfy, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. Berlin, Germany: Springer-Verlag, 2002.

[10] M. Hagan, M. Beale, and H. Demuth, *Neural Network Design*. Boston, MA, USA: PWS-Kent, 1996.

[11] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, “Extreme learning machine: Theory and applications,” *Neurocomputing*, vol. 70, nos. 1–3, pp. 489–501, 2006.

[12] G.-B. Huang, L. Chen, and C.-K. Siew, “Universal approximation using incremental constructive feedforward networks with random hidden nodes,” *IEEE Trans. Neural Netw.*, vol. 17, no. 4, pp. 879–892, Jul. 2006.

[13] G.-B. Huang, Q.-Y. Zhu, K. Z. Mao, C.-K. Siew, P. Saratchandran, and N. Sundararajan, “Can threshold networks be trained directly?” *IEEE Trans. Circuits Syst. II, Exp. Briefs*, vol. 53, no. 3, pp. 187–191, Mar. 2006.

[14] G.-B. Huang and L. Chen, “Convex incremental extreme learning machine,” *Neurocomputing*, vol. 70, nos. 16–18, pp. 3056–3062, 2007.

[15] G.-B. Huang, H. Zhou, X. Ding, and R. Zhang, "Extreme learning machine for regression and multiclass classification," *IEEE Trans. Syst., Man, Cybern. B, Cybern.*, vol. 42, no. 2, pp. 513–529, Apr. 2012.

[16] G.-B. Huang, D. H. Wang, and Y. Lan, "Extreme learning machines: A survey," *Int. J. Mach. Learn. Cybern.*, vol. 2, no. 2, pp. 107–122, 2011.

[17] S. Lin, C. Xu, J. Zeng, and J. Fang, "Does generalization performance of l^q regularization learning depend on q? A negative example," ArXiv preprint, arXiv:1307.6616, 2013.

[18] S. Lin, J. Zeng, J. Fang, and Z. Xu, "Learning rate of l^q coefficient regularization learning with Gaussian kernel," *Neural Comput.*, doi: 10.1162/NECO_a_00641.

[19] S. Lin, X. Liu, Y. Rong, and Z. Xu, "Almost optimal estimates for approximation and learning by radial basis function networks," *Mach. Learn.*, vol. 95, no. 2, pp. 147–164, 2014.

[20] X. Liu, S. B. Lin, J. Fang, and Z. B. Xu, "Is extreme learning machine feasible? A theoretical assessment (Part I)," *IEEE Trans. Neural Netw. Learn. Syst.*, doi: 10.1109/TNNLS.2014.2335212.

[21] H. N. Mhaskar, F. J. Narcowich, and J. D. Ward, "Approximation properties of zonal function networks using scattered data on the sphere," *Adv. Comput. Math.*, vol. 11, nos. 2–3, pp. 121–137, 1999.

[22] 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.

[23] I. Marques and M. Graña, "Face recognition with lattice independent component analysis and extreme learning machines," *Soft Comput.*, vol. 16, no. 9, pp. 1525–1537, 2012.

[24] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and Its Application*. New York, NY, USA: Wiley, 1971.

[25] D. Serre, *Matrices: Theory and Applications*. New York, NY, USA: Springer-Verlag, 2002.

[26] 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.

[27] E. Soria-Olivas *et al.*, "BELM: Bayesian extreme learning machine," *IEEE Trans. Neural Netw.*, vol. 22, no. 3, pp. 505–509, Mar. 2011.

[28] I. Steinwart and C. Scovel, "Fast rates for support vector machines using Gaussian kernels," *Ann. Statist.*, vol. 35, no. 2, pp. 575–607, 2007.

[29] H. Sun and Q. Wu, "Least square regression with indefinite kernels and coefficient regularization," *Appl. Comput. Harmon. Anal.*, vol. 30, no. 1, pp. 96–109, 2011.

[30] J. Tang, D. Wang, and T. Chai, "Predicting mill load using partial least squares and extreme learning machines," *Soft Comput.*, vol. 16, no. 9, pp. 1585–1594, 2012.

[31] H. Tong, D.-R. Chen, and F. Yang, "Least square regression with l^p-coefficient regularization," *Neural Comput.*, vol. 22, no. 12, pp. 3221–3235, 2010.

[32] Q. Wu and D.-X. Zhou, "Learning with sample dependent hypothesis spaces," *Comput. Math. Appl.*, vol. 56, no. 11, pp. 2896–2907, 2008.

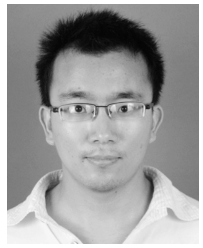
[33] T. F. Xie and F. L. Cao, "The rate of approximation of Gaussian radial basis neural networks in continuous function space," *Acta Math. Sinica, English Ser.*, vol. 29, no. 2, pp. 295–302, 2013.

[34] J. Xu, H. Zhou, and G.-B. Huang, "Extreme learning machine based fast object recognition," in *Proc. 15th Int. Conf. Inform. Fusion (FUSION)*, Jul. 2012, pp. 1490–1496.

[35] R. Zhang, Y. Lan, G.-B. Huang, and Z.-B. Xu, "Universal approximation of extreme learning machine with adaptive growth of hidden nodes," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 23, no. 2, pp. 365–371, Feb. 2012.

[36] 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.

[37] Q.-Y. Zhu, A. K. Qin, P. N. Suganthan, and G.-B. Huang, "Evolutionary extreme learning machine," *Pattern Recognit.*, vol. 38, no. 10, pp. 1759–1763, 2005.



Shaobo Lin received the B.S. degree in mathematics and the M.S. degree in basic mathematics from Hangzhou Normal University, Hangzhou, China, and the Ph.D. degree from Xi'an Jiaotong University, Xi'an, China.

He is currently with Wenzhou University, Wenzhou, China. His current research interests include statistical learning theory and scattered data fitting.

Xia Liu received the M.S. degree in basic mathematics from Yan'an University, Yan'an, China, in 2010. Since then, she has been with the Department of Mathematics, Yulin University, Yulin, China. She is currently pursuing the Ph.D. degree with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China.

Her current research interests include learning theory and nonlinear functional analysis.

Jian Fang received the B.S. degree in mathematics from Nanjing Normal University, Nanjing, China, in 2008. He is currently pursuing the Ph.D. degree with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China.

His current research interests include sparse modeling and synthetic aperture radar imaging.

Zongben Xu received the Ph.D. degree in mathematics from Xi'an Jiaotong University, Xi'an, China, in 1987.

He currently serves as the Vice President of Xi'an Jiaotong University, the Chief Scientist of National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences, Xi'an Jiaotong University. His current research interests include intelligent information processing and applied mathematics.

Prof. Xu is a member of the Chinese Academy of Sciences, Beijing, China. He delivered a 45-min talk on the International Congress of Mathematicians in 2010. He was a recipient of the National Natural Science Award of China in 2007 and the CSIAM Su Buchin Applied Mathematics Prize in 2008.