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

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

***Abstract*— An extreme learning machine (ELM) is a feedforward neural network (FNN) like learning system whose connections with output neurons are adjustable, while the connections with and within hidden neurons are randomly fixed. Numerous applications have demonstrated the feasibility and high efficiency of ELM-like systems. It has, however, been open if this is true for any general applications. In this two-part paper, we conduct a comprehensive feasibility analysis of ELM. In Part I, we provide an answer to the question by theoretically justifying the following: 1) for some suitable activation functions, such as polynomials, Nadaraya–Watson and sigmoid functions, the ELM-like systems can attain the theoretical generalization bound of the FNNs with all connections adjusted, i.e., they do not degrade the generalization capability of the FNNs even when the connections with and within hidden neurons are randomly fixed; 2) the number of hidden neurons needed for an ELM-like system to achieve the theoretical bound can be estimated; and 3) whenever the activation function is taken as polynomial, the deduced hidden layer output matrix is of full column–rank, therefore the generalized inverse technique can be efficiently applied to yield the solution of an ELM-like system, and, furthermore, for the nonpolynomial case, the Tikhonov regularization can be applied to guarantee the weak regularity while not sacrificing the generalization capability. In Part II, however, we reveal a different aspect of the feasibility of ELM: there also exists some activation functions, which makes the corresponding ELM degrade the generalization capability. The obtained results underlie the feasibility and efficiency of ELM-like systems, and yield various generalizations and improvements of the systems as well.**

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

I. Introduction

LEARNING abounds in the sciences and engineering. One of the main tasks of learning is to synthesize a function that can present an unknown but definite relation between the input and output. Given a finite number of input–output samples, a learning system is normally developed for defining the function and yielding an estimator. The learning system comprises a hypothesis space, a family of parameterized functions that regulate the forms and properties of the estimator to be found, an optimality criterion, in the sense that the estimator can be defined, and a learning strategy or an algorithm that numerically yields the parameters of the estimator. The performance of the learning system is then measured by its approximation capability, generalization capability and computational burden. It is known that there exists a dilemma in the approximation capability and generalization capability, and the generalization capability is usually one of the most important factors to be considered in many applications. Therefore, we focus on measuring a learning system in terms of generalization capability and computational burden in this paper.

There have been many types of learning systems. Artificial feedforward neural networks (FNNs) are, for instance, the well developed learning systems whose hypothesis spaces are functions represented as the multilayer neuron-like structured networks (the parameters are the synaptic connection weights among the neurons). FNNs can approximate any integrable functions provided the hidden neurons are sufficiently large [1], [2]. With appropriate training schemes, FNNs also have promising generalization capability [3]. Training a FNN is, however, by no means easy, especially when the activation function of neuron is discontinuous. Although many effective algorithms, such as the back-propagation [4], are available, training a FNN with all connection weights adjustable is usually time consuming, and is of high computational burden in general.

To overcome such difficulty, a useful learning scheme, called the extreme learning machine (ELM), was suggested in [5] and subsequently extended and applied in [6]–[16]. Different kinds of ELM variations have been studied in fields as metal temperature prediction [17], palmprint and handwitten character recognition [18], [19], and face recognition [20], [21]. Furthermore, ELM has been successfully applied to gene selection and cancer classification [22]. In essence, ELM is an FNN-like learning system whose connections with output neurons are adjustable, while the connections with and within hidden neurons are randomly fixed. With such settings, ELM then transforms the training of an FNN with all connections adjustable into a linear problem in which only connections with output neurons are adjusted. Thus, the well-known generalized inverse technique can be directly applied for the solution [23], [24]. The similar idea has been adopted earlier in [25]–[27] as the echo state network method and

Manuscript received July 4, 2013; revised May 12, 2014; accepted June 22, 2014. Date of publication July 21, 2014; date of current version December 16, 2014. This work was supported in part by the National 973 Program of China under Contract 2013CB329404 and in part by the National Science Foundation of China under Project 11131006 and Project 61075054.

X. Liu is with the School of Mathematics and Statistics, Institute for Information and System Sciences, Xi'an Jiaotong University, Xi'an 710049, China, and also with the Department of Mathematics and Statistics, Yulin University, Yulin 719000, China (e-mail: liuxia1232007@163.com).

S. Lin, 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: sblin1983@gmail.com; ender86@163.com; zbxu@mail.xjtu.edu.cn).

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/TNNLS.2014.2335212

in [28] as the random vector functional-link networks method. However, it is still open whether ELM is feasible and efficient for general applications. In particular, one would like to ask the following questions.

(Q1) FNN defines an estimator through adjusting all the possible connections among all neurons, while ELM searches for the estimator only by adjusting the connections with output neurons. Thus, does ELM degrade the generalization capability of FNN when the hidden connections are randomly fixed?

(Q2) If the answer to Q1 is yes, what is the number of hidden layer nodes needed to reach the optimal generalization capability?

(Q3) ELM finds its estimator by applying the generalized inverse technique, while this technique is efficient only when the induced hidden layer output matrix is weakly regular (i.e., is of full column rank or full row rank). Then, how can the weak regularity of the hidden layer output matrix be guaranteed?

All these problems are fundamental to the feasibility and effectiveness of ELM-like systems.

The aim of this paper is to provide answers to these questions. More precisely, we will theoretically justify the following: 1) for some good activation functions like polynomials, Nadaraya–Watson and sigmoid functions, the ELM-like systems can attain the generalization bound of the FNNs with all connections adjusted, i.e., they do not degrade the generalization capability of the FNNs even when the connections with and within hidden neurons are randomly fixed; 2) the number of hidden layer nodes needed for an ELM-like system to achieve the generalization bound is estimated under the assumption of the smoothness of the regression function; and 3) whenever the activation function is taken as polynomials, the induced hidden layer output matrix is of full column rank, therefore the generalized inverse technique can be efficiently applied to yield the solution. Moreover, for the nonpolynomial case, the Tikhonov regularization can be applied to guarantee the weak regularity without sacrificing the generalization capability. Thus, the obtained results not only underlie the feasibility and efficiency of ELM-like systems, but also yield some further generalizations and improvements of the systems as well.

The remainder of this paper is organized as follows. In Section II, we introduce the model and algorithm of ELM-like systems. In Section III, we answer questions (Q1) and (Q2) through developing a series of almost optimal generalization bound estimations of ELM systems, as compared with the known lower bound estimation of the FNNs when all connections adjusted. In Section IV, we answer question (Q3) by showing that whenever the activation functions are polynomials, the induced hidden layer output matrix is of full column rank, and therefore the generalized inverse technique can be efficiently applied. For the nonpolynomial activation function case, we further suggest the use of regularization technique, and show that the regularization scheme can assure the weak regularity of the hidden layer output matrix without sacrificing the generalization capability. We conclude this paper in Section V with some useful remarks.

II. ELM-Like Systems: Model and Algorithm

Let $\mathbf{N}$ be the the set of positive integers, $d, l, m, n \in \mathbf{N}$, $X \subseteq \mathbf{R}^d$ be the input space and $Y \subseteq \mathbf{R}$ be the output space. Suppose that the unknown probability measure ρ on $Z := X \times Y$ admits a decomposition $\rho(x, y) = \rho_X(x)\rho(y|x)$, here, $\rho(y|x)$ is the conditional probability measure at x. Let $\mathbf{z} = (x_i, y_i)_{i=1}^m$ be a family of random samples, drawn independently and identically according to ρ from Z. Without loss of generality, we assume $|y_i| \le M$ and $f : X \to Y$ is the function induced from ρ (i.e., the unknown but definite correspondence between X and Y). The difference between an estimator f and the output is measured by the generalization error

$$\mathcal{E}(f) = \int_Z (f(x) - y)^2 d\rho \tag{1}$$

which is known to be minimized by the regression function f_ρ defined by [29]

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

Because of the unknown feature of distribution ρ, the above regression function cannot be computed and applied directly. Learning then aims to find an approximation f^* of f_ρ in a given hypothesis space.

Let $L^2_{\rho_X}$ be the Hilbert space consisting of all ρ_X square integrable functions on X, with norm $\|\cdot\|_\rho$. It is well known that for every $f \in L^2_{\rho_X}$ [30]

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

In ELM framework, the hypothesis space is supposed to be

$$\mathcal{M}_n = \left\{ f_n(\alpha, \beta, x) = \sum_{i=1}^n \alpha_i \phi(\beta_i, x) : n \in \mathbf{N} \right\} \tag{3}$$

where $x \in X$, $\beta = (\beta_1, \beta_2, \ldots, \beta_n)^T \in \mathbf{R}^{n \times l}$ with $\beta_i \in \mathbf{R}^l$, $\alpha = (\alpha_1, \alpha_2, \ldots, \alpha_n)^T \in \mathbf{R}^n$, $\phi : \mathbf{R}^l \times \mathbf{R}^n \to \mathbf{R}$ is a nonlinear function (say, the activation function in FNN framework, the kernel function in kernel learning framework). $\mathcal{M}_n$ can be illustrated as the multilayer FNNs with variable hidden nodes and one output node whose hidden node parameters are β and connections with the output nodes are α. Under the least squares optimality criterion, FNNs define their estimator $f_{\text{FNN}} = f_{n^*}(\alpha^*, \beta^*, x)$ through

$$(n^*, \alpha^*, \beta^*) = \arg \min_{(n,\alpha,\beta)} \left\{ \frac{1}{m} \sum_{j=1}^m |f_n(\alpha, \beta, x_j) - y_j|^2 \right\}$$

or, whenever the number of hidden neurons n is preset, by

$$(\alpha^*, \beta^*) = \arg \min_{(\alpha,\beta)} \left\{ \frac{1}{m} \sum_{j=1}^m \left| \sum_{i=1}^n \alpha_i \phi(\beta_i, x_j) - y_j \right|^2 \right\}. \tag{4}$$

While, as compared, ELM define its estimator $f_{\text{ELM}} = f_{n^*}(\alpha^*, \widetilde{\beta}, x)$ through

$$(n^*, \alpha^*) = \arg \min_{(n,\alpha)} \left\{ \frac{1}{m} \sum_{j=1}^m |f_n(\alpha, \widetilde{\beta}, x_j) - y_j|^2 \right\}$$

or by

$$\alpha^* = \arg\min_{\alpha} \left\{ \frac{1}{m} \sum_{j=1}^{m} \left| \sum_{i=1}^{n} \alpha_i \phi(\widetilde{\beta}_i, x_j) - y_j \right|^2 \right\} \tag{5}$$

where $\widetilde{\beta}$ are randomly preselected according to a definite distribution μ in $\mathbf{R}^{n\times l}$.

In the traditional learning paradigm, problem (4) is solved by using a learning algorithm minimizing variables α and β simultaneously. In ELM methodology, however, problem (5) is usually solved directly using the generalized inverse technique. Denote by

$$H = \begin{bmatrix} \phi(\widetilde{\beta}_1, x_1) & \phi(\widetilde{\beta}_2, x_1) & \cdots & \phi(\widetilde{\beta}_n, x_1) \\ \phi(\widetilde{\beta}_1, x_2) & \phi(\widetilde{\beta}_2, x_2) & \cdots & \phi(\widetilde{\beta}_n, x_2) \\ \vdots & \vdots & \cdots & \vdots \\ \phi(\widetilde{\beta}_1, x_m) & \phi(\widetilde{\beta}_2, x_m) & \cdots & \phi(\widetilde{\beta}_n, x_m) \end{bmatrix} \tag{6}$$

the hidden layer output matrix, and $\mathbf{y} = (y_1, y_2, \ldots, y_m)^T$. Then, problem (5) is seen to be a standard least squares problem $\min_\alpha \|H\alpha - \mathbf{y}\|_2^2/m$ and its solution can be explicitly given by $\alpha^* = H^\dagger \mathbf{y}$, where $\|\cdot\|_2$ is the Euclidean norm of $\mathbf{R}^m$ and $H^\dagger$ is a Moore–Penrose generalized inverse of H. This leads to the following standard ELM algorithm [5].

A. ELM-Like Algorithm

Given the training samples $\mathbf{z} = (x_i, y_i)_{i=1}^m$, the nonlinear function ϕ, and the hidden neuron number n.

Step 1: Randomly assign $\widetilde{\beta}_i$ in $\mathbf{R}^l$, $i = 1, \ldots, n$.

Step 2: Calculate the hidden layer output matrix H.

Step 3: Calculate the output weight vector $\alpha^* = H^\dagger \mathbf{y}$ where $H^\dagger$ is a Moore–Penrose generalized inverse of H.

Finally, the estimator f_{ELM} is defined by

$$f_{\mathrm{ELM}}(x) = f_n(\alpha^*, \widetilde{\beta}, x) = \sum_{i=1}^{n} \alpha_i^* \phi(\widetilde{\beta}_i, x). \tag{7}$$

There are many available algorithms to calculate a Moore–Penrose generalized inverse of H, say, the orthogonal projection method, the orthogonalization method, and the singular value decomposition method [31]. In particular, whenever H is of full column rank or of full row rank, $H^\dagger$ can be explicitly expressed as

$$H^\dagger = (H^T H)^{-1} H^T \tag{8}$$

or

$$H^\dagger = H^T (H H^T)^{-1}.$$

Nevertheless, whenever H is neither of full column rank nor full row rank (that is, H is not weakly regular), calculating $H^\dagger$ is still time consuming. Many works have studied this issue [8], [32], [33]. We will provide new solutions to this problem in Section IV.

The ELM-like algorithm exhibits the most different points between FNNs and ELM: 1) unlike FNNs that require the hidden-layer parameters $\beta = \{\beta_i\}_{i=1}^n$ to be trained, ELM only randomly assigns the hidden-layer parameters, and thus saves a great amount of computation time and 2) training an FNN is a complicated nonlinear optimization problem, while training an ELM is a linear least squares problem whose solution can be directly generated by the generalized inverse of the hidden layer output matrix. These differences certainly make the computational complexity of ELM-like systems much lower than that of FNNs. The problem is, however, whether such significant complexity-deduction degrades other properties of FNN learning? This is rational of feasibility of ELM-like learning. Although wide range of applications of ELM have provided positive support to the question [32]–[34], there is still no solid theoretical assessment. We will provide such a theoretical assessment in the subsequent sections.

III. Does ELM Degrade the Generalization Capability?

In this section, we characterize the generalization capability of ELM algorithm from a theoretical point of view. The purpose is to show that ELM-like learning does not degrade the generalization capability of FNN provided the activation functions are appropriately chosen.

Let C_0 be a positive constant, and $q \in (0, 1]$, $r = k + q$ for some $k \in \mathbf{N}_0 := \{0\} \cup \mathbf{N}$. A function $f : \mathbf{R}^d \to \mathbf{R}$ is said to be (r, C_0)-smooth if for every $\gamma = (\gamma_1, \ldots, \gamma_d)$, $\gamma_i \in N_0$, $\sum_{j=1}^d \gamma_j = k$, the partial derivatives $\partial^k f / \partial {x_1}^{\gamma_1} \ldots \partial {x_d}^{\gamma_d}$ exist and satisfy

$$\left| \frac{\partial^k f}{\partial {x_1}^{\gamma_1} \ldots \partial {x_d}^{\gamma_d}}(x) - \frac{\partial^k f}{\partial {x_1}^{\gamma_1} \ldots \partial {x_d}^{\gamma_d}}(z) \right| \le C_0 \|x - z\|^q$$

for all $x, z \in \mathbf{R}^d$. Denote by $\mathcal{F}^{(r,C_0)}$ the set of all (r, C_0)-smooth functions, and by $\mathcal{D}^{(r,C_0)}$ the set of all distributions $\rho(x, y) = \rho_X(x)\rho(y|x)$, with ρ_X being uniformly distributed on X and $f_\rho \in \mathcal{F}^{(r,C_0)}$.

Given an estimator $f_{\mathbf{z}}$, by (2), its generalization capability can be measured by the difference between $f_{\mathbf{z}}$ and f_ρ, that is, by

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

This quantity depends on $\mathbf{z}$ and therefore has a stochastic nature. As a result, it is impossible to say something about (9) in general for a fixed $\mathbf{z}$. Instead, we look at its behavior in probability, say, as measured by the expected error

$$E_{\rho^m}\left(\|f_{\mathbf{z}} - f_\rho\|_\rho^2\right) = \int_{Z^m} \|f_{\mathbf{z}} - f_\rho\|_\rho^2 d\rho^m$$

where the expectation is taken over all realizations $\mathbf{z}$ obtained from a fixed m, and ρ^m is the m fold tensor product of ρ.

Let $\Theta \subset L^2_{\rho_X}$ and $\mathbf{U}(\Theta)$ be the class of all Borel measures ρ on Z such that $f_\rho \in \Theta$. Due to the unknown feature of ρ, we enter into a competition over all estimators $H_m : \mathbf{z} \to f_{\mathbf{z}}$ and define

$$e_m(\Theta) := \inf_{f_{\mathbf{z}} \in H_m} \sup_{\rho \in \mathbf{U}(\Theta)} E_{\rho^m}\left(\|f_\rho - f_{\mathbf{z}}\|_\rho^2\right).$$

It is easy to see that $e_m(\Theta)$ quantitively measures the generalization capability of $f_{\mathbf{z}}$.

We first provide a baseline of FNN estimators, which can be found in [35, Ch. 3] and [36]. Let $\Psi = \left\{f_\rho \in \mathcal{F}^{(r,C_0)} : \|f_\rho\|_\infty \le M\right\}$ and $\Phi = \left\{\rho(x, y) \in \mathcal{D}^{(r,C_0)} : |y| \le M\right\}$. It is

obvious that $\Phi \subseteq \Psi$ almost surely. Then, there holds the following lower bound estimation:

$$e_m(\Psi) \geq e_m(\Phi) \geq Cm^{-\frac{2r}{2r+d}}, \quad m = 1, 2, \ldots \tag{10}$$

where C is a constant independent of m. The above inequalities together with the upper bound estimation of FNN leaning [3] yield the following basic results on generalization capability of FNNs.

Proposition 1: Whenever the activation function of a FNN is of good property, there holds the following estimations:

$$\begin{aligned} C_1 m^{-\frac{2r}{2r+d}} &\leq e_m(\Phi) \leq e_m(\Psi) \\ &\leq \sup_{\rho \in \mathbf{U}(\Theta)} E_{\rho^m}\left(\|f_\rho - f_{\mathrm{FNN}}\|_\rho^2\right) \\ &\leq C_2 m^{-\frac{2r}{2r+d}} \log m \end{aligned} \tag{11}$$

where C_1 and C_2 are constants independent of m.

Proposition 1 means that the FNN estimators can almost realize the optimal generalization bound in term of (11) as long as the activation function possesses some good properties. That is, the activation functions should be exponential functions, rational functions, or logistic sigmoid functions whose Fourier transformations satisfy a certain smooth assumption. For more details on the assumption, we refer the readers to [3].

A. Main Results

We show, in this section, that whenever the activation functions are algebraic polynomials, Nadaraya–Watson functions, or sigmoid functions, the ELM-like systems does not degrade the generalization capabilities of FNNs.

We will repeatedly apply the following known result on the upper bound estimation of the FNN estimators [35, Th. 11.3].

Lemma 1: Given the sample set $\mathbf{z} = (x_i, y_i)_{i=1}^m$, let $\mathcal{K}_n$ be the linear space of functions $f : \mathbf{R}^d \to \mathbf{R}$, with dimension k, and $\pi_M : \mathbf{R} \to \mathbf{R}$ be the truncation operator defined by

$$\pi_M(x) = \min\{|x|, M\}\mathrm{sign}(x) \quad \forall x \in \mathbf{R}^d.$$

Then

$$\begin{aligned} E_{\rho^m}\|\pi_M f_{\mathbf{z}}^* - f_\rho\|_\rho^2 \leq CM^2 \frac{k(\log m + 1)}{m} \\ +8 \inf_{f \in \mathcal{K}_n} \int_X |f(x) - f_\rho(x)|^2 d\rho_X \end{aligned} \tag{12}$$

where

$$f_{\mathbf{z}}^* = \arg\min_{f \in \mathcal{K}_n} \frac{1}{m} \sum_{i=1}^m |f(x_i) - y_i|^2$$

and C is a constant independent of m and k.

1) Polynomial Activation Case: Considering the case $\phi(\beta_i, x) = (\omega_i x + b_i)^s$ with $\beta_i = (\omega_i, b_i) \in [\mathbf{0}, \mathbf{1}]^{d+1}$, both ω_i and b_i being drawn independently and identically according to the uniform distribution μ in $[0, 1]^d$ and $[0, 1]$, and $s \in \mathbf{N}$. In this case, the hypothesis space of ELM is defined by

$$\mathcal{M}_P = \left\{ f_n(\alpha, \omega, b, n, s, x) = \sum_{i=1}^n \alpha_i (\omega_i x + b_i)^s : n \in \mathbf{N} \right\}$$

and the ELM estimator is defined by

$$f_{\mathrm{ELM}_P}(x) = \sum_{i=1}^n \alpha_i{}^*(\omega_i x + b_i)^s \tag{13}$$

where, in terms of (6), $\alpha^* = (\alpha_1^*, \alpha_2^*, \ldots, \alpha_n^*)$ satisfies

$$\alpha^* = \arg\min_\alpha \left\{\|H\alpha - \mathbf{y}\|_2^2\right\}.$$

For the estimator f_{ELM_P}, we have the following conclusion.

Theorem 1: Assume $f_\rho \in \mathcal{F}^{(r,C_0)}$ with $0 < r \leq 1$ and $C_0 \geq 0$, and ρ_X is absolutely continuous with respect to Lebesgue measure on X. Let f_{ELM_P} be the ELM estimator defined as in (13). Then, whenever $s = [m^{1/d+2r}]$ and $n = [m^{d/d+2r}]$, there holds the estimations

$$\begin{aligned} C_1 m^{-\frac{2r}{d+2r}} &\leq E_{\rho^m} E_{\mu^n}(\|\pi_M(f_{ELM_P}) - f_\rho\|_\rho^2) \\ &\leq C_2 m^{-\frac{2r}{d+2r}} \log m \end{aligned} \tag{14}$$

where $[\cdot]$ denotes the integer part. C_1 and C_2 are positive constants depending only on M, C_0, r, and d.

Proof: We first prove that for any $n \geq \dim(\mathcal{P}_s^d) = \mathcal{O}(s^d)$, there holds $\mathcal{M}_p = \mathcal{P}_s^d$ almost surely, where $\mathcal{P}_s^d$ is the collection of polynomials defined on X with d variables and degree at most s. It is obvious that $\mathcal{M}_p \subseteq \mathcal{P}_s^d$. On the other hand, since ρ_X is absolutely continuous with respect to Lebesgue measure on X and $m > n$, it can be found in the proof of Theorem 4 below that matrix $H = ((\omega_i x + b_i)^s)_{m \times n}$ is of full column rank, thus there exists a set of functions $\{(w_1 \cdot x + b_1)^s, \ldots, (w_n \cdot x + b_n)^s\}$, which are linear independent almost surely (indeed, if $\{(w_1 \cdot x + b_1)^s, \ldots, (w_n \cdot x + b_n)^s\}$ is linear dependent, then there exists a set of real numbers (not all 0) $\{c_i\}_{i=1}^n$ such that for almost all $x \in X$

$$c_1(w_1 \cdot x + b_1)^s + \cdots + c_n(w_n \cdot x + b_n)^s = 0$$

which is impossible according to the proof of Theorem 4). Thus, the dimension of $\mathcal{M}_p$ is at least n, so $\mathcal{P}_s^d \subseteq \mathcal{M}_p$. Hence, $\mathcal{M}_p = \mathcal{P}_s^d$ almost surely. This shows that

$$\inf_{f \in \mathcal{M}_p} \int_X |f_\rho(x) - f(x)|^2 d\rho_X = \inf_{f \in \mathcal{P}_s^d} \int_X |f_\rho(x) - f(x)|^2 d\rho_X$$

holds almost surely. Consequently, the well-known Jensen's inequality [37] for algebraic polynomials [38] together with $f_\rho \in \mathcal{F}^{(r,C_0)}$ can imply

$$\inf_{f \in \mathcal{P}_s^d} \int_X |f_\rho(x) - f(x)|^2 d\rho_X \leq Cs^{-2r}.$$

Using Lemma 1 (12), we thus obtain

$$E_{\rho^m} E_{\mu^n} \|\pi_M(f_{\mathrm{ELM}_P}) - f_\rho\|_\rho^2 \leq C\left(M^2 \frac{n \log m}{m} + s^{-2r}\right).$$

Taking $n = [m^{d/(d+2r)}]$ and $s = [m^{1/(d+2r)}]$, we then arrive to the upper bound estimation of Theorem 1. The lower bound estimation of the theorem follows from (10) and the fact that the absolutely continuous distribution with respective to the Lebesgue measure is uniform distribution. With this, the proof of Theorem 1 is completed. ■

Theorem 1 provides bounds on the generalization error of the ELM estimator in terms of (14) when algebraic polynomials are taken as activation function. Our motivation to use the polynomial function as an activation function in ELM is from Zhou and Jetter [30], who gave an upper bound estimate for support vector machine with polynomial kernel. In fact, the polynomial kernel is a finite bandwidth kernel [39]. Therefore, the capability (dimension) of the hypothesis space of ELM with polynomial kernel is controlled by both the number of neurons and the degree of polynomial kernel. If the degree is fixed, no matter how many hidden neurons are employed, the capacity of the hypothesis space is fixed. Therefore, ELM with polynomial kernel of suitable degree does not suffer from the overfitting phenomenon. This is the essential feature of polynomial kernel.

At first glance, Theorem 1 seems contradict with the results in [40], which showed that the nonpolynomial assumption on the activation function is a necessary condition for the density of neural networks. However, it should be noted that the polynomial activation function in this paper is intrinsically different from that of [40]. Indeed, the polynomial stated in [40] is with the fixed degree, while in this paper the degree is variable. It can be found in Theorem 1 that the degree $s \to \infty$ as $m \to \infty$. Thus, the density is obvious in our setting.

Such an estimation can be generalized to the cases when the activation functions in ELM are Nadaraya–Watson or sigmoid function.

2) Nadaraya–Watson Activation Case: Let $D_n = \{t_i\}_{i=1}^n$ be a discrete subset of X, with t_i being drawn independently and identically according to the uniform distribution μ in X. The mesh norm of D_n and h_{D_n}, which measures the maximum distance from D_n, and for any points in X, be defined by

$$h_D := h_{D_n} := \max_{x\in X} \min_{t_i\in D_n} \|x - t_i\|.$$

It is known [35, Ch. 2] that the Nadaraya–Watson function is an important type of kernels in constructing local averaging estimators. Such type of functions is defined by

$$\phi_{\mathrm{NW}}(t_i, x) = \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}}$$

where A, a positive constant, is the width of the Nadaraya–Watson function. We consider the case of ELM with $\phi(\beta_i, x) = \phi_{\mathrm{NW}}(t_i, x)$, that is, the hypothesis space and the ELM estimator are, respectively, defined by

$$\mathcal{M}_{\mathrm{NW}} = \left\{ \begin{array}{r} f_n(\alpha, t, A, n, x) = \sum_{i=1}^n \alpha_i \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}} \\ : A > 0, n \in \mathbf{N} \end{array} \right\}$$

and by

$$f_{\mathrm{ELM_{NW}}}(x) = \sum_{i=1}^n {\alpha_i}^* \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}}. \tag{15}$$

Theorem 2: Let $0 < r \le 1$, $C_0 \ge 0$ and $f_{\mathrm{ELM_{NW}}}$ be the estimator defined as in (15). If $f_\rho \in \mathcal{F}^{(r,C_0)}$, $n = [m^{d/(d+2r)}]$, and $A \ge n^{\frac{1}{d}} \log n$, then there exit constants C_1 and C_2 depending only on M, r, C_0, and d, such that

$$\begin{aligned} C_1 m^{-\frac{2r}{2r+d}} &\le E_{\rho^m} E_{\mu^n}\left(\|\pi_M(f_{\mathrm{ELM_{NW}}}) - f_\rho\|_\rho^2\right) \\ &\le C_2 m^{-\frac{2r}{2r+d}} \log m. \end{aligned} \tag{16}$$

Proof: We denote by $B(x, \varepsilon)$, the closed ball centered at x with radius $\varepsilon > 0$. Then it is easy to see that

$$\begin{aligned} &\left| f_\rho(x) - \sum_{i=1}^n \frac{f_\rho(t_i) e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}} \right| \\ &\le \sum_{i=1}^n |f_\rho(x) - f_\rho(t_i)| \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}} \\ &= \sum_{t_i\in B(x,2h_{D_n})} |f_\rho(x) - f_\rho(t_i)| \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}} \\ &\quad + \sum_{t_i\notin B(x,2h_{D_n})} |f_\rho(x) - f_\rho(t_i)| \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}} \\ &\le C h_{D_n}^r + \sum_{t_i\notin B(x,2h_{D_n})} |f_\rho(x) - f_\rho(t_i)| \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}}. \end{aligned}$$

To bound the second term in the above inequality, we first notice that by definition of mesh norm of D_n, there is an t_j such that $t_j \in B(x, h_{D_n})$. Therefore, we have

$$\begin{aligned} &\sum_{t_i\notin B(x,2h_{D_n})} |f_\rho(x) - f_\rho(t_i)| \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}} \\ &\qquad \le 2\|f_\rho\| \sum_{t_i\notin B(x,2h_{D_n})} e^{-A(\|x-t_i\|-\|x-t_j\|)}. \end{aligned}$$

Since for arbitrary $t_i \notin B(x, 2h_{D_n})$ and $t_j \in B(x, h_{D_n})$, there hold, respectively, that $\|x - t_i\| > 2h_{D_n}$ and $\|x - t_j\| \le h_{D_n}$, we have $\|x - t_i\| - \|x - t_j\| > h_{D_n}$. This then implies

$$\sum_{t_i\notin B(x,2h_{D_n})} |f_\rho(x) - f_\rho(t_i)| \frac{e^{-A\|x-t_i\|}}{\sum_{j=1}^n e^{-A\|x-t_j\|}} \le 2\|f_\rho\| n e^{-Ah_{D_n}}.$$

In consequence, we obtain

$$\inf_{f\in\mathcal{M}_{\mathrm{NW}}} \|f_\rho - f\| \le C(h_{D_n}^r + n e^{-Ah_{D_n}}). \tag{17}$$

To prove (16), we first need to establish the following estimation:

$$E_{\mu^n}(h_{D_n}) \le C n^{-\frac{1}{d}}. \tag{18}$$

Indeed, since μ is a uniform distribution, it follows that $\vartheta(B(x, \delta)) = C\delta^d$, where ϑ is a probability measure. From the definition of h_{D_n}, we deduce that if there exists an $x \in X$ such that $B(x, \varepsilon) \cap D_n = \emptyset$, then $h_{D_n} > \varepsilon$. Thus

$$\begin{aligned} P\{h_{D_n} > \varepsilon\} &\le P\{B(x, \varepsilon) \cap D_n = \emptyset\} \\ &= (1 - C\varepsilon^d)^n \le e^{-Cn\varepsilon^d}. \end{aligned}$$

Hence

$$\begin{aligned} E_{\mu^n}\{h_{D_n}\} &= \int_0^\infty P\{h_{D_n} > \varepsilon\} d\varepsilon \le \int_0^\infty e^{-Cn\varepsilon^d} d\varepsilon \\ &= C n^{-\frac{1}{d}} \int_0^\infty e^{-t^d} dt \\ &\le C n^{-\frac{1}{d}} \left(\int_0^1 1 dt + \int_1^\infty e^{-t} dt \right) \le C n^{-\frac{1}{d}} \end{aligned} \tag{19}$$

hold as claimed. Now, noting that $f_\rho \in \mathcal{F}^{(r,C_0)}$ with $0<r\leq 1$, we can apply Jensen's inequality with (19) to get

$$E_{\mu^n}\left\{h_{D_n}^r\right\} \leq (E_{\mu^n}(h_{D_n}))^r \leq Cn^{-r/d}. \tag{20}$$

Since, by assumption, $A \geq n^{\frac{1}{d}} \log n$ and (17), it then follows that:

$$E_{\mu^n}(|f_\rho(x)-f(x)|) \leq E_{\mu^n}\left\{h_{D_n}^r\right\}+E_{\mu^n}\{ne^{-Ah_{D_n}}\} \leq Cn^{-r/d}. \tag{21}$$

From Lemma 1 and $n=[m^{d/2r+d}]$, it comes

$$\begin{aligned} &E_{\rho^m}E_{\mu^n}\|\pi_M f_{\mathrm{ELM_{NW}}}-f_\rho\|_\rho^2 \\ &\quad\leq E_{\rho^m}E_{\mu^n}(\tilde{C}M^2\frac{(\log m+1)n}{m} \\ &\qquad +8\inf_{f\in\mathcal{M}_{\mathrm{NW}}}\int_X |f(x)-f_\rho(x)|^2 d\rho) \\ &\quad\leq \tilde{C}M^2\frac{(\log m+1)n}{m}+\bar{C}n^{-2r/d} \\ &\quad\leq Cm^{-\frac{2r}{2r+d}}\log m. \end{aligned}$$

That is, the upper bound estimation of (16) is justified. The lower bound is obviously true from (10). This finishes the proof of Theorem 2. ■

3) Sigmoid Activation Case: In FNN applications, a most extensively used activation function is the sigmoid function σ that satisfies

$$\lim_{t\to-\infty}\sigma(t)=0 \quad \text{and} \quad \lim_{t\to\infty}\sigma(t)=1.$$

In this section, we consider the case of ELM with the bounded sigmoid activation functions, i.e., $\phi(\beta_i,x)=\sigma(b(x-t_i))$ with $\beta_i=(b,t_i)$, where σ is a bounded sigmoid function, b is a positive constant, and $D_n=\{t_i\}_{i=1}^n$ with t_i being chosen independently and identically according to the uniform distribution μ in $[0,1]$. By definition, for any sigmoid function σ, there exists a positive constant L such that

$$|\sigma(t)-1|<n^{-3r-1}, \quad \text{if } t\geq L$$

and

$$|\sigma(t)|<n^{-3r-1}, \quad \text{if } t\leq -L.$$

The hypothesis space is then defined by

$$\mathcal{M}_S=\left\{\begin{matrix} f_n(\alpha,t,b,a_0,n,x)=\sum_{i=1}^n \alpha_i\sigma(b(x-t_i))+a_0 \\ : x\in X, \alpha_i\in\mathbf{R}, a_0\in\mathbf{R}, n\in\mathbf{N} \end{matrix}\right\}$$

with

$$b\geq nh_D{}^{-1} \tag{22}$$

and then the estimator is defined by

$$f_{\mathrm{ELM}_S}(x)=\sum_{i=1}^n \alpha_i{}^*\sigma(b(x-t_i))+a_0. \tag{23}$$

Theorem 3: Let $0<r\leq 1$, $f_\rho\in\mathcal{F}^{(r,C_0)}$ with $C_0>0$. f_{ELM_S} be the ELM estimator defined as in (23). If $d=1$, b satisfies (22) and $n=[m^{1/1+2r}]$, then there holds

$$\begin{aligned} C_1m^{-\frac{2r}{2r+1}} &\leq E_{\rho^m}E_{\mu^n}(\|\pi_M(f_{\mathrm{ELM}_S})-f_\rho\|_\rho^2) \\ &\leq C_2m^{-\frac{2r}{2r+1}}\log m \end{aligned} \tag{24}$$

where C_1 and C_2 are constants depending only on M, r, and C_0.

Proof: For arbitrary $x\in[0,1]$ and any $k\in[1,n]$, we define

$$\begin{aligned} h_n(x) &= f_\rho(t_1)+\sum_{i=1}^n(f_\rho(t_{i+1})-f_\rho(t_i))\sigma(b(x-t_i)) \\ &= f_\rho(t_k)+\sum_{i=1}^{k-1}(f_\rho(t_{i+1})-f_\rho(t_i))(\sigma(b(x-t_i))-1) \\ &\quad +(f_\rho(t_{k+1})-f_\rho(t_k))\sigma(b(x-t_k)) \\ &\quad +\sum_{i=k+1}^n(f_\rho(t_{i+1})-f_\rho(t_i))\sigma(b(x-t_i)). \end{aligned}$$

Then, for arbitrary $k\in[1,n]$, we have

$$\begin{aligned} &h_n(x)-f_\rho(x) \\ &= f_\rho(t_k)-f_\rho(x)+\sum_{i=1}^{k-1}(f_\rho(t_{i+1})-f_\rho(t_i))(\sigma(b(x-t_i))-1) \\ &\quad +(f_\rho(t_{k+1})-f_\rho(t_k))\sigma(b(x-t_k)) \\ &\quad +\sum_{i=k+1}^n(f_\rho(t_{i+1})-f_\rho(t_i))\sigma(b(x-t_i)). \end{aligned} \tag{25}$$

Define

$$B:=n^{-3r-1},$$

by the definition of sigmoid function, there exists a positive constant $L>0$ such that

$$|\sigma(t)-1|<B, \quad \text{if } t\geq L$$

and

$$|\sigma(t)|<B, \quad \text{if } t\leq -L.$$

Define $\sigma^*(t):=\sigma(bt)$ with $b\geq Lh_D{}^{-1}$, then it comes

$$|\sigma^*(t)-1|<B, \quad \text{if } t\geq h_D \tag{26}$$

and

$$|\sigma^*(t)|<B, \quad \text{if } t\leq -h_D. \tag{27}$$

If $x\in[t_k-h_D,t_k+h_D]$, $k=1,\ldots,n$, then we have

$$x-t_i\geq h_D, \quad \text{if } i<k$$

and

$$x-t_i\leq -h_D, \quad \text{if } i>k.$$

Combine with (26) and $b\geq Lh_D{}^{-1}$

$$|\sigma^*(x-t_i)-1|<B, \quad \text{if } i<k$$

and

$$|\sigma^*(x-t_i)|<B, \quad \text{if } i>k.$$

Since $f_\rho \in \mathcal{F}^{(r,C_0)}$ and σ is bounded, it follows from (25) that:

$$\begin{aligned}|h_n(x) - f_\rho(x)| &\le {h_D}^r + (k-1){h_D}^r B + \|\sigma\|{h_D}^r + (n-k){h_D}^r B \\ &\le C_1\|h_D\|^r + n\|h_D\|^r B \end{aligned} \tag{28}$$

where $\|\sigma\| := \sup_{x\in\mathbf{R}}|\sigma(x)|$. Now we return to estimate $E_{\mu^n}\|h_{D_n}\|$. Since μ is the uniform distribution, similar to the proof of Theorem 2, we can find

$$E_{\mu^n}\{h_{D_n}\} = \int_0^\infty P\{h_{D_n} > \varepsilon\}d\varepsilon \le Cn^{-1} \tag{29}$$

for a constant C independent of n.

By the well-known Jensen's inequality, we thus deduce from (28) and (29) that

$$\begin{aligned}\inf_{f_\rho\in\mathcal{F}^{(r,C_0)}}\int |h_n(x) - f_\rho(x)|_\rho^2 d\rho &\le C_1 E_{\mu^n}\|h_{D_n}\|^{2r} \\ &\le C_1(E_{\mu^n}\|h_{D_n}\|^2)^r < \frac{C_2}{n^{2r}}. \end{aligned} \tag{30}$$

Consequently, combined with Lemma 1, it gives

$$E_{\rho^m}E_{\mu^n}(\|\pi_M(f_{ELM_S}) - f_\rho\|_\rho^2) \le C\left(\frac{n\log m}{m} + n^{-2r}\right)$$

where constant C is independent of m and n. Setting $n = [m^{1/2r+1}]$ in the above estimation, then it comes

$$E_{\rho^m}E_{\mu^n}\left(\|\pi_M(f_{\mathrm{ELM}_S}) - f_\rho\|_\rho^2\right) \le C_2 m^{-\frac{2r}{2r+1}}\log m. \tag{31}$$

This concludes the proof of the upper bound in (24). Together with (10), this implies Theorem 3. ∎

B. Remarks

Some remarks on Theorems 1–3 are as follows.

1) Theorems 1–3 say that whenever the unknown distribution ρ has the priors $f_\rho \in \mathcal{F}^{(r,C_0)}$ (roughly speaking, this amounts to that the unknown function f between the input and output is of up to r-order smoothness), the hidden parameters are randomly assigned according to a uniform distribution μ, and the nonlinear function ϕ in ELM is appropriately selected (say, polynomials, Nadaraya–Watson or sigmoid functions), the generalization error of ELM-like learning then obeys to

$$\begin{aligned}\mathcal{O}(m^{-\frac{2r}{2r+d}}) &\le \sup_{f_\rho\in\mathcal{F}^{(r,C_0)}} E_{\rho^m}E_{\mu^n}\left(\|\pi_M(f_{\mathrm{ELM}}) - f_\rho\|_\rho^2\right) \\ &\le \mathcal{O}\left(m^{-\frac{2r}{2r+d}}\log m\right) \end{aligned} \tag{32}$$

where m is the number of samples, d is the dimension of input space, and r is the order of the smoothness of the regression function. This upper bound estimation is asymptotically identical with the lower bound because of $\log m/m^\gamma \to 0$ for any $\gamma > 0$ as $m \to \infty$. Therefore, it turns out that the generalization error of the ELM learning can be essentially characterized by $\mathcal{O}(m^{-2r/2r+d})$. It is well known (10), however, that whenever the regression function is r smooth, any learning algorithm possesses the generalization bounds not less than $\mathcal{O}(m^{-2r/2r+d})$ (i.e., $\mathcal{O}(m^{-2r/2r+d})$ is an optimal generalization bound). Theorems 1–3 then show that averagely, ELM can attain the almost optimal generalization bound $\mathcal{O}(m^{-2r/2r+d})$ of FNNs learning. In consequence, we can conclude that with suitably selected activation functions, ELM does not degrade the generalization capability of FNNs.

2) When we say ELM does not degrade the generalization capability, it takes sense only in the understanding of estimations (32). In the FNN framework, an almost optimal generalization bound is expressed normally in terms of estimations [3], [35]

$$\begin{aligned}\mathcal{O}(m^{-\frac{2r}{2r+d}}) &\le \sup_{f_\rho\in\mathcal{F}^{(r,C_0)}} E_{\rho^m}\left(\|\pi_M(f_{\mathrm{FNN}}) - f_\rho\|_\rho^2\right) \\ &\le \mathcal{O}\left(m^{-\frac{2r}{2r+d}}\log m\right). \end{aligned} \tag{33}$$

As compared (32) with (33), we then can immediately find a difference: the double expectations E_{μ^n} and E_{ρ^m} are used in (32) while only one expectation E_{ρ^m} has been applied in measuring the generalization error in (33). The second expectation E_{μ^n} is inevitable since the random assignment of the hidden parameters in ELM brings another randomness, except for the sample randomness ρ. Thus, to be more precise, the estimations (32) (or equivalently, Theorems 1–3) just show that taking averagely (on all realizations of μ), the generalization performance of ELM can be as good as that of FNNs with all parameters adjusted. This is not, however, to say that for any single implementation of ELM, corresponding to one realization of random assignment of the hidden parameters, its performance is as good as FNNs.

3) Different from the existing literatures on ELM study, Theorems 1–3 reveal the close connection between the hidden nodes size and the number of training samples (e.g., $n = [m^{d/(2r+d)}]$). However, in the real word application, the smooth information of the regression function is usually unknown. Under this circumstance, the well-known cross validation can be employed to select the exact number of hidden neurons.

4) Theorems 1–3 have provided just partial but by no means a complete answer to questions (Q1) and (Q2). For instance, we have shown in Theorems 1–3 that the ELM does not degrade the generalization capability when the activation functions are polynomials, Nadaraya–Watson and sigmoid. On the other hand, we have found that for the Gaussian activation function, ELM does degrade the generalization capability of FNN, which can be found in a separate paper (Part II) [41]. To facilitate the use of ELM, the well-developed coefficient regularization technique as a remedy to this degradation is employed in ELM. These two papers (Parts I and II) give a comprehensive feasibility analysis of ELM, which reveal the essential characteristics of ELM.

The further question is then what more general types of activation functions, not only polynomials, Nadaraya–Watson or sigmoid, with which the ELM-like learning does or does

not degrade the generalization capability. Furthermore, since different random assignments lead to different performances of ELM, so the random assignment will be weighed into the consideration. We will return back to the research of such problems in our further work.

IV. How Can Weak Regularity of Hidden Layer Output Matrix H be Guaranteed?

In this section, we formalize a condition under which the weak regularity of the hidden layer output matrix H can be guaranteed, so that the generalized inverse technique can be efficiently applied in the ELM-like learning.

A. Polynomial Activation Implies Weak Regularity

We first show that whenever the activation functions are taken to be algebraic polynomials, the deduced hidden layer output matrix H is of full column rank. Therefore, the Moore–Penrose generalized inverse of H can be directly computed as in (8).

We need to establish the following lemma, which plays a key role in the proof of Theorem 4.

Lemma 2: Let $P \in \mathcal{P}_s^d$ be an algebraic polynomial with d variables and degree at most s. Then, its zero set

$$N(P) := \{u \in [0,1]^d : P(u) = 0\}$$

has Lebesgue measure 0.

Proof: Let $u = (u_{(1)}, \ldots, u_{(d)})$. For any fixed $k \in [1, d]$, we let

$$P_k(u_k) = P(u_{(1)}, \ldots, u_{(k)}, \ldots, u_{(d)})$$

which is an algebraic polynomial in univariate variable $u_{(k)}$ of degree at most s whenever

$$u_{(1)}, \ldots, u_{(k-1)}, u_{(k+1)}, \ldots, u_{(d)}$$

are fixed. Then, $P_k(u_k)$ has at most s zeros and its zero set

$$N(P_k) := \left\{ \begin{array}{c} u \in [0,1]^d : (u_{(1)}, \ldots, u_{(k-1)}, u_{(k)}, \\ u_{(k+1)}, \ldots, u_{(d)}) \in N(P) \end{array} \right\}$$

has Lebesgue measure 0, i.e., $\mathcal{L}(N(P_k)) = 0$ for every choice of $k \in [1, d]$. Therefore, by Fubini's Theorem [42], we have

$$\begin{aligned} &\mathcal{L}(N(P)) \\ &= \int_{[0,1]^{d-1}} \left(\int_{[0,1]} \mathcal{X}_{N(P)}(u_{(1)}, \ldots, u_{(k-1)}, u_{(k)}, u_{(k+1)}, \right. \\ &\qquad \left. \ldots, u_{(d)}) du_{(k)} \right) du_{(1)} \cdots du_{(k-1)} du_{(k+1)} \cdots du_{(d)} \\ &= \int_{[0,1]^{d-1}} \mathcal{L}(N(P_k)) du_{(1)} \cdots du_{(k-1)} du_{(k+1)} \cdots du_{(d)} = 0 \end{aligned}$$

where $\mathcal{X}_{N(P)}$ is the character function of $N(P)$. This implies Lemma 2. ∎

Theorem 4: Assume that the marginal distribution ρ_X is absolutely continuous with respect to Lebesgue measure on X, and the algebraic polynomial is used as the activation function in ELM, that is, $\phi(\beta_i, x) = (\omega_i x + b_i)^s$, with $\beta_i = (\omega_i, b_i) \in [0,1]^{d+1}$ randomly assigned. If $n \leq m$, then the hidden layer output matrix $H = (r_{ji})_{m \times n}$ is of full column rank almost surely, where $r_{ji} = (w_i x_j + b_i)^s$. In addition, $H^T H$ is invertible almost surely.

Proof: Since $n \leq m$, to prove Theorem 4, it suffices to prove that there exists an $n \times n$ invertible submatrix R_n of H with entries $r_{ji} = (w_i x_j + b_i)^s$, where $j = 1, \ldots, m$, $i = 1, \ldots, n$. For this purpose, we define

$$B_n := \{(x_1, \ldots, x_n) \in X^n : \det R_n = 0\}.$$

Then, we only need to prove that $\mathcal{L}(B_n) = 0$ for all $n \leq m$. We prove this by induction over n. It is obviously true for $n = 1$. Assume, this is true for $n = N$ (naturally we can assure $N \leq m - 1$), i.e., $\mathcal{L}(B_N) = 0$. Then, we prove that $\mathcal{L}(B_{N+1}) = 0$. In effect, since there is a submatrix, say, the first N row submatrix R_N that is invertible, so for arbitrary $a_{N+1} := (r_{N+1,1}, \ldots, r_{N+1,N})$, there exist coefficients $c_i := c_i(x_1, \ldots, x_N) \in \mathbf{R}$, not all 0, such that

$$a_{N+1} = c_1 a_1 + c_2 a_2 + \cdots + c_N a_N.$$

Here, we denote by $a_i = (r_{i1}, \ldots, r_{iN})$ the ith row of H. Now, we use this assertions to prove that there is a submatrix R_{N+1} such that $\mathcal{L}(B_{N+1}) = 0$. By looking at $(N+1)$st column of H, we find that R_{N+1} is invertible if and only if

$$r_{N+1,N+1} \neq c_1 r_{1,N+1} + \cdots + c_N r_{N,N+1}$$

or equivalently

$$\begin{aligned} (w_{N+1} x_{N+1} + b_{N+1})^s \neq\ & c_1 (w_1 x_{N+1} + b_1)^s \\ & + \cdots + c_N (w_N x_{N+1} + b_N)^s. \end{aligned}$$

In other words, R_{N+1} is invertible if and only if x_{N+1} does not appear in the set

$$D_N(x_1, \ldots, x_N) := \left\{ \begin{array}{c} x \in X : (w_{N+1} x + b_{N+1})^s \\ = c_1 (w_1 x + b_1)^s + \cdots + c_N (w_N x + b_N)^s \end{array} \right\}.$$

For any fixed $(x_1, \ldots, x_N) \in X^N$, D_N is clearly a zero set of some algebraic polynomials, so by Lemma 2, D_N has Lebesgue measure 0 in X. Since

$$B_{N+1} \subset \left\{ \begin{array}{c} (x_1, \ldots, x_N, x_{N+1}) \in X^{N+1} \\ : x_{N+1} \in D_N(x_1, \ldots, x_N) \end{array} \right\}$$

we see by Fubini's Theorem that

$$\begin{aligned} &\mathcal{L}(B_{N+1}) \\ &= \int_{X^{N+1}} \mathcal{X}_{B_{N+1}}(x_1, \ldots, x_N, x_{N+1}) dx_1 \cdots dx_N dx_{N+1} \\ &= \int_{X^N} \left(\int_X \mathcal{X}_{B_{N+1}}(x_1, \ldots, x_N, x_{N+1}) dx_{N+1} \right) dx_1 \cdots dx_N \\ &\leq \int_{X^N} \mathcal{L}(D_N(x_1, \ldots, x_N)) dx_1 \cdots dx_N = 0. \end{aligned}$$

That is, $\mathcal{L}(B_{N+1}) = 0$, the induction step is finished. Thus, we verified that $\mathcal{L}(B_n) = 0$ for all $n \leq m$. Note that $m \geq n$, the above equation implies that the $n \times n$ square matrix $H^T H$ is invertible for almost every choice of $x_1, \ldots, x_n$. Since the distribution ρ_X is absolutely continuous with respect to the Lebesgue measure $\mathcal{L}$, the set B_N also has measure 0 with respect to ρ_X. This finishes the proof of Theorem 4. ∎

Theorem 4 shows that the hidden layer output matrix H will be of full column rank provided the activation functions

are taken to be algebraic polynomials. Thus, in this case, the Moore–Penrose generalized inverse of H uniquely exists and can be directly computed via (8).

B. Use of Regularization

If the activation functions are not polynomials, then the hidden layer output matrices H are most likely not weakly regular. For such general cases, we then suggest the use of Tikhonov regularization scheme

$$\arg\min_{\alpha}\left\{\frac{1}{m}\sum_{j=1}^{m}\left|\sum_{i=1}^{n}\alpha_i\phi(\widetilde{\mathbf{t}}_i,x_j)-y_j\right|^2+\lambda\sum_{i=1}^{n}|\alpha_i|^2\right\} \quad (34)$$

where $\lambda:=\lambda(m,n)$ is a regularization parameter. With the matrix notations in (6), this amounts to saying that instead of f_{ELM} defined as in (7), we redefine the estimator of ELM by

$$f_{\mathrm{ELM}}^{(\lambda)}(x)=f_{n^*}(\alpha^*,t^*,x)=\sum_{i=1}^{n}\alpha_i^*\phi(\widetilde{t_i},x) \quad (35)$$

where

$$\alpha^*=\arg\min_{\alpha}\left\{\frac{1}{m}\|H\alpha-\mathbf{y}\|_2^2+\lambda\|\alpha\|_2^2\right\}. \quad (36)$$

The solution to (36) is unique and given by

$$\alpha^*=(H^TH+\lambda mI)^{-1}H^T\mathbf{y}. \quad (37)$$

Thus, whenever the Tikhonov regularization scheme is used, the weak regularity of H can be naturally guaranteed. The problem is then: does Tikhonov regularization maintain or improve on the generalization capability of ELM? Theorem 5 below provides an answer to this question.

The following two lemmas are needed, which can be found in [29].

Lemma 3: For any $R>0$ and $\eta>0$, we have

$$\log\mathcal{N}(B_R,\eta)\le n\log\left(\frac{4R}{\eta}\right)$$

where $\mathcal{N}(B_R,\eta)$ is the covering number of B_R, defined as the size of the smallest η covering of B_R, and B_R is the closed ball of l_2^n norm centered at origin

$$B_R=\left\{f\in\mathcal{M}_n:\|f\|_{l_2}^2:=\sum_{i=1}^{n}|a_i|^2\le R^2,\ f=\sum_{i=1}^{n}a_i\phi(t_i,\cdot)\right\}.$$

Lemma 4: Let $\mathcal{G}$ be the set of functions on $Z=X\times Y$ such that, for some $c\ge 0$, $|g-E(g)|\le B$ almost everywhere and $E(g^2)\le cE(g)$ for each $g\in\mathcal{G}$, where $E(g)=\int_Z g(x,y)d\rho$. Then, for every $\varepsilon>0$

$$\mathrm{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{g\in\mathcal{G}}\frac{E(g)-\frac{1}{m}\sum_{i=1}^{m}g(z_i)}{\sqrt{E(g)+\varepsilon}}\ge\sqrt{\varepsilon}\right\}\le\mathcal{N}(\mathcal{G},\varepsilon)\exp\left\{-\frac{m\varepsilon}{2c+\frac{2B}{3}}\right\}.$$

Theorem 5: Let $0<r\le 1$, $C_0>0$, $f_\rho\in\mathcal{F}^{(r,C_0)}$ and $f_{\mathrm{ELM}}^{(\lambda)}$ be defined as in (35). Then, there exist constants C and c independent of n and m such that

$$E_{\rho^m}E_{\mu^n}\left(\|\pi_M f_{\mathrm{ELM}}^{(\lambda)}-f_\rho\|_\rho^2\right)\le C\frac{n(\log m-\log\lambda)}{m}+\inf_{f\in\mathcal{M}_n}\left(\int_X(f(x)-f_\rho(x))^2d\rho+\lambda\|f\|_{l_2}^2\right). \quad (38)$$

Proof: From (2), it follows that:

$$\begin{aligned}&\left\|\pi_M f_{\mathrm{ELM}}^{(\lambda)}-f_\rho\right\|_\rho^2\\&\le\left\{\mathcal{E}\left(\pi_M f_{\mathrm{ELM}}^{(\lambda)}\right)-\mathcal{E}(f_\rho)-\left(\mathcal{E}_{\mathbf{z}}\left(\pi_M f_{\mathrm{ELM}}^{(\lambda)}\right)-\mathcal{E}_{\mathbf{z}}(f_\rho)\right)\right\}\\&\quad+\mathcal{E}_{\mathbf{z}}\left(\pi_M f_{\mathrm{ELM}}^{(\lambda)}\right)-\mathcal{E}_{\mathbf{z}}(f_\rho)+\lambda\left\|f_{\mathrm{ELM}}^{(\lambda)}\right\|_{l_2}^2:=S_1+S_2\end{aligned}$$

where $\mathcal{E}_{\mathbf{z}}(f):=1/m\sum_{i=1}^{m}(f(x_i)-y_i)^2$ is the empirical error of f (with respect to $\mathbf{z}$). Therefore

$$\begin{aligned}&E_{\rho^m}E_{\mu^n}\left(\|\pi_M f_{\mathrm{ELM}}^{(\lambda)}-f_\rho\|_\rho^2\right)\\&\le E_{\rho^m}E_{\mu^n}\left(\left\{\mathcal{E}\left(f_{\mathrm{ELM}}^{(\lambda)}\right)-\mathcal{E}(f_\rho)-\left(\mathcal{E}_{\mathbf{z}}\left(\pi_M f_{\mathrm{ELM}}^{(\lambda)}\right)-\mathcal{E}_{\mathbf{z}}(f_\rho)\right)\right\}\right)\\&\quad+E_{\rho^m}E_{\mu^n}\left(\mathcal{E}_{\mathbf{z}}\left(\pi_M f_{\mathrm{ELM}}^{(\lambda)}\right)-\mathcal{E}_{\mathbf{z}}(f_\rho)+\lambda\|f_{\mathrm{ELM}}^{(\lambda)}\|_{l_2}^2\right).\end{aligned}$$

Now, we use Lemmas 3 and 4 to estimate S_1. Set

$$\mathcal{F}_R:=\{(\pi_M f(x)-y)^2-(f_\rho(x)-y)^2:f\in B_R\}.$$

Then, for any fixed $g\in\mathcal{F}_R$, there exists $f\in B_R$ such that $g(\mathbf{z})=(\pi_M f(x)-y)^2-(f_\rho(x)-y)^2$. Therefore, we have

$$E_{\rho^m}(g)=\mathcal{E}(\pi_M f)-\mathcal{E}(f_\rho)\ge 0$$
$$\frac{1}{m}\sum_{i=1}^{m}g(z_i)=\mathcal{E}_{\mathbf{z}}(\pi_M f)-\mathcal{E}_{\mathbf{z}}(f_\rho).$$

Since $|\pi_M f(x)|\le M$ and $|f_\rho(x)|\le M$ hold almost everywhere, we deduce that

$$\begin{aligned}|g(\mathbf{z})|&=|(\pi_M f(x)-f_\rho(x))((\pi_M f(x)-y)\\&\quad+(f_\rho(x)-y))|\le 8M^2.\end{aligned}$$

It then follows that $|g(\mathbf{z})-E(g)|\le 16M^2$ almost everywhere and:

$$E_{\rho^m}(g^2)\le 16M^2\|\pi_M f-f_\rho\|_\rho^2=16M^2E_{\rho^m}(g).$$

Now, we apply Lemma 4 with $B=c=16M^2$ to the set of functions $\mathcal{F}_R$, yielding

$$\begin{aligned}\sup_{f\in B_R}\frac{\{\mathcal{E}(\pi_M f)-\mathcal{E}(f_\rho)\}-\{\mathcal{E}_{\mathbf{z}}(\pi_M f)-\mathcal{E}_{\mathbf{z}}(f_\rho)\}}{\sqrt{\{\mathcal{E}(\pi_M f)-\mathcal{E}(f_\rho)\}+\varepsilon}}&=\sup_{g\in\mathcal{F}_R}\frac{E_{\rho^m}(g)-\frac{1}{m}\sum_{i=1}^{m}g(z_i)}{\sqrt{E_{\rho^m}(g)+\varepsilon}}\le\sqrt{\varepsilon}\end{aligned}$$

which holds with probability at least

$$1-\mathcal{N}(\mathcal{F}_R,\varepsilon)\exp\left\{-\frac{3m\varepsilon}{128M^2}\right\}.$$

Observe that for any $g_1,g_2\in\mathcal{F}_R$, there exist $f_1,f_2\in B_R$ such that

$$g_j(z)=(\pi_M f_j(x)-y)^2-(f_\rho(x)-y)^2,\quad j=1,2.$$

It is obvious that

$$|g_1(\mathbf{z}) - g_2(\mathbf{z})| = |(\pi_M f_1(x) - y)^2 - (\pi_M f_2(x) - y)^2| \le 4M\|f_1 - f_2\|_\infty.$$

Therefore, for any $\varepsilon > 0$, an $(\varepsilon/4M)$-covering of B_R can provide an ε-covering of $\mathcal{F}_R$. Accordingly

$$\mathcal{N}(\mathcal{F}_R, \varepsilon) \le \mathcal{N}\left(B_R, \frac{\varepsilon}{4M}\right).$$

Thus, the probability is

$$1 - \mathcal{N}(\mathcal{F}_R, \varepsilon)\exp\left\{-\frac{3m\varepsilon}{128M^2}\right\} \ge 1 - \mathcal{N}\left(B_R, \varepsilon/(4M)\right)\exp\left\{-\frac{3m\varepsilon}{128M^2}\right\}.$$

It follows from Lemma 3 that:

$$\log \mathcal{N}(B_R, \varepsilon/(4M)) \le n \log\left(\frac{16MR}{\varepsilon}\right).$$

Consequently, it comes

$$\text{Prob}_{z\in Z^m}\left\{\sup_{f\in B_R}\frac{\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} - \{\mathcal{E}_{\mathbf{z}}(\pi_M f) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}}{\sqrt{\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} + \varepsilon}} \le \sqrt{\varepsilon}\right\} \ge 1 - \exp\left\{n\log\frac{16MR}{\varepsilon} - \frac{3m\varepsilon}{128M^2}\right\}.$$

Since

$$\sqrt{\varepsilon}\sqrt{\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} + \varepsilon} \le \frac{1}{2}\{\mathcal{E}(f) - \mathcal{E}(f_\rho)\} + \varepsilon$$

we conclude that with probability at least

$$1 - \exp\left\{n\log\frac{16MR}{\varepsilon} - \frac{3m\varepsilon}{128M^2}\right\}$$

there holds

$$\sup_{f\in B_R}\{(\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)) - (\mathcal{E}_{\mathbf{z}}(\pi_M f) - \mathcal{E}_{\mathbf{z}}(f_\rho))\} \le \frac{1}{2}(\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)) + \varepsilon.$$

Hence

$$\text{Prob}_{\mathbf{z}\in Z^m}\left\{\sup_{f\in B_R}(\{\mathcal{E}(\pi_M f) - \mathcal{E}(f_\rho)\} - 2\{\mathcal{E}_{\mathbf{z}}(\pi_M f) - \mathcal{E}_{\mathbf{z}}(f_\rho)\}) \le \varepsilon\right\} \ge 1 - \exp\left\{n\log\frac{32MR}{\varepsilon} - \frac{3m\varepsilon}{256M^2}\right\}.$$

On the other hand, it follows from (35) that:

$$\left\|f_{\text{ELM}}^{(\lambda)}\right\|_{l_2}^2 = \sum_{i=1}^n |\alpha_i|^2 \le \frac{M^2}{\lambda}.$$

Set

$$\mathcal{T} := \left\{\mathcal{E}\left(\pi_M f_{\text{ELM}}^{(\lambda)}\right) - \mathcal{E}(f_\rho)\right\} - 2\left\{\mathcal{E}_{\mathbf{z}}\left(\pi_M f_{\text{ELM}}^{(\lambda)}\right) - \mathcal{E}_{\mathbf{z}}(f_\rho)\right\}.$$

Then

$$\mathcal{E}\left(\pi_M f_{\text{ELM}}^{(\lambda)}\right) - \mathcal{E}(f_\rho) \le \mathcal{T} + 2S_2. \tag{39}$$

For arbitrary $t \ge 32M^2/m$, there holds

$$\begin{aligned}
E_{\rho^m}(\mathcal{T}) &= \int_0^\infty \text{Prob}_{z\in Z^m}\left\{\{\mathcal{E}(\pi_M f_{\text{ELM}}^{(\lambda)}) - \mathcal{E}(f_\rho)\} - 2\{\mathcal{E}_{\mathbf{z}}(\pi_M f_{\text{ELM}}^{(\lambda)}) - \mathcal{E}_{\mathbf{z}}(f_\rho)\} > \varepsilon\right\} d\varepsilon \\
&\le t + \int_t^\infty \exp\left\{n\log\frac{32M^2}{\varepsilon\lambda} - \frac{3m\varepsilon}{256M^2}\right\} d\varepsilon \\
&\le t + \exp\left\{-\frac{3mt}{256M^2}\right\}\int_t^\infty \left(\frac{32M^2}{\varepsilon\lambda}\right)^n d\varepsilon \\
&\le t + \lambda^{-n}\exp\left\{-\frac{3mt}{256M^2}\right\}\left(\frac{32M^2}{t}\right)^n t \\
&\le t + \lambda^{-n}\exp\left\{-\frac{3mt}{256M^2}\right\} m^n t.
\end{aligned}$$

Setting $t = (256M^2 n(\log m - \log\lambda))/3m$, we then obtain

$$E_{\rho^m}(\mathcal{T}) \le 2t = \frac{512M^2 n(\log m - \log\lambda)}{3m}. \tag{40}$$

Now, we turn to estimate $E_{\rho^m}(S_2)$

$$\begin{aligned}
E_{\rho^m}(S_2) &= E_{\rho^m}\left(\mathcal{E}_{\mathbf{z}}(f_{\text{ELM}}^{(\lambda)}) - \mathcal{E}_{\mathbf{z}}(f_\rho) + \lambda\|f_{\text{ELM}}^{(\lambda)}\|_{l_2}^2\right) \\
&= E_{\rho^m}\left(\frac{1}{m}\sum_{i=1}^m (f_{\text{ELM}}^{(\lambda)}(x_i) - y_i)^2 - \frac{1}{m}\sum_{i=1}^m (f_\rho(x_i) - y_i)^2 + \lambda\|f_{\text{ELM}}^{(\lambda)}\|_{l_2}^2\right) \\
&= E_{\rho^m}\left(\inf_{f\in\mathcal{M}_n}\left(\frac{1}{m}\sum_{i=1}^m (f(x_i) - y_i)^2 - \frac{1}{m}\sum_{i=1}^m (f_\rho(x_i) - y_i)^2 + \lambda\|f_{\text{ELM}}^{(\lambda)}\|_{l_2}^2\right)\right) \\
&\le \inf_{f\in\mathcal{M}_n}\left(E_{\rho^m}((f(x) - y)^2) - E_{\rho^m}((f_\rho(x) - y)^2) + \lambda\|f_{\text{ELM}}^{(\lambda)}\|_{l_2}^2\right) \\
&= \inf_{f\in\mathcal{M}_n}\left(\int_X (f(x) - f_\rho(x))^2 d\rho + \lambda\|f_{\text{ELM}}^{(\lambda)}\|_{l_2}^2\right).
\end{aligned} \tag{41}$$

Then, (39), (40), and (41) imply that there exists a constant C independent on m and n such that

$$E_{\rho^m}E_{\mu^n}\left(\left\|\pi_M f_{\text{ELM}}^{(\lambda)} - f_\rho\right\|_\rho^2\right) \le C\frac{n(\log m - \log\lambda)}{m} + \inf_{f\in\mathcal{M}_n}\int_X (f(x) - f_\rho(x))^2 d\rho + \lambda\|f_{\text{ELM}}^{(\lambda)}\|_{l_2}^2. \tag{42}$$

This arrives to Theorem 5. ∎

Theorem 5 shows that the ELM-like learning with Tikhonov regularization [(35) and (37)] can be effectively used to overcome the weak regularity problem of ELM, and maintain a promising generalization capability.

Some direct consequences of Theorem 5 are as follows.

Corollary 1: If the activation function used in (34) is Nadaraya–Watson, $\lambda = [m^{-1}]$ and $n = [m^{d/(d+2r)}]$, then there

holds

$$C_1 m^{-\frac{2r}{2r+d}} \le E_{\rho^m} E_{\mu^n}\left(\left\|\pi_M\left(f^{(\lambda)}_{\mathrm{ELM_{NW}}}\right) - f_\rho\right\|^2_\rho\right) \le C_2 m^{-\frac{2r}{2r+d}} \log m. \quad (43)$$

Proof: From (21), we know that

$$E_{\mu^n}(|f_\rho(x) - f(x)|) \le C n^{-r/d}$$

where $f(x) = \sum_{i=1}^n (f_\rho(t_i) e^{-A\|x-t_i\|}) / \sum_{j=1}^n e^{-A\|x-t_j\|}$. Since $\lambda = [m^{-1}]$, we get

$$\lambda \|f\|^2_{l_2} = \lambda \sum_{i=1}^n |f_\rho|^2 \le \lambda n M^2 \le C M^2 m^{-\frac{2r}{2r+d}}.$$

Thus from (38), we obtain

$$E_{\rho^m} E_{\mu^n}\left(\left\|\pi_M f^{(\lambda)}_{ELM_{\mathrm{NW}}} - f_\rho\right\|^2_\rho\right) \le C \frac{n(\log m - \log \lambda)}{m} + C n^{-2r/d} + M^2 m^{-\frac{2r}{2r+d}} \le C_2 m^{-\frac{2r}{2r+d}} \log m$$

which arrives to the upper bound of (43). The lower bound of (43) is ready in (10). ■

Corollary 2: If the activation function used in (34) is sigmoid function, $\lambda = [m^{-1}]$ and $n = [m^{1/(1+2r)}]$, then there holds

$$C_1 m^{-\frac{2r}{2r+1}} \le E_{\rho^m} E_{\mu^n}\left(\left\|\pi_M\left(f^{(\lambda)}_{ELM_S}\right) - f_\rho\right\|^2_\rho\right) \le C_2 m^{-\frac{2r}{2r+1}} \log m. \quad (44)$$

Proof: From (30), it follows that:

$$\inf_{f_\rho \in \mathcal{F}^{(r,C_0)}} \int |h_n(x) - f_\rho(x)|^2_\rho d\rho \le \frac{C_2}{n^{2r}}$$

where $h_n(x) = f_\rho(t_1) + \sum_{i=1}^n (f_\rho(t_{i+1}) - f_\rho(t_i)) \sigma(b(x - t_i))$. Since $\lambda = [m^{-1}]$, we get

$$\lambda \|h_n\|^2_{l_2} = \lambda \sum_{i=1}^n |f_\rho|^2 \le \lambda n M^2 \le C M^2 m^{-\frac{2r}{2r+1}}.$$

Thus, from (38), it yields

$$E_{\rho^m} E_{\mu^n}\left(\left\|\pi_M f^{(\lambda)}_{ELM_S} - f_\rho\right\|^2_\rho\right) \le C \frac{n(\log m - \log \lambda)}{m} + C n^{-2r} + C M^2 m^{-\frac{2r}{2r+1}} \le C_2 m^{-\frac{2r}{2r+1}} \log m$$

this comes to the result of the upper bound of (44). The lower bound of (44) can be deduced from (10). ■

Corollaries 1 and 2 show that when regularization scheme is applied to ELM with Nadaraya–Watson function and sigmoid function not only can the weak regularity of ELM be guaranteed but also its generalization capability is maintained.

C. Application Support

To verify the theoretical results presented above, we present an application demonstration. In the experiments, we compared the generalization performance of ELM with sigmoid activation function (denoted by ELM), the regularized version of ELM (34) with sigmoid activation function (denoted by L_2ELM) and ELM with polynomial activation function (denoted by PolyELM) under the same number of hidden neurons when applied to Abalone data, Machine CPU data and Census (house8L) data in UCI Database, respectively.[1] The input data were normalized into the unit cube, while the output data normalized into the range [0, 1]. Figs. 1, 2, 4, 5, 7, and 8 show the experimental results, particularly the training and testing root mean square error (RMSE) curves of the three ELM algorithms with increasing number of hidden neurons. Figs. 3, 6, and 9 show the testing results in detail by cutting the number of neurons to [0, 200] or [0, 500]. It can be observed from Figs. 1, 4, and 7 that L_2ELM and PolyELM possess good approximation performance with the increasing of n, but

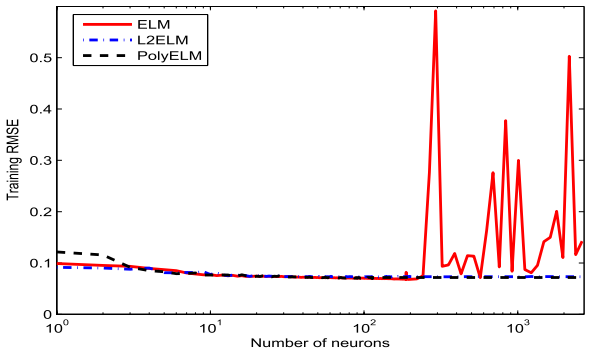


Fig. 1. Performance comparison (training RMSE) with ELM, L_2ELM, and PolyELM for Abalone data set.

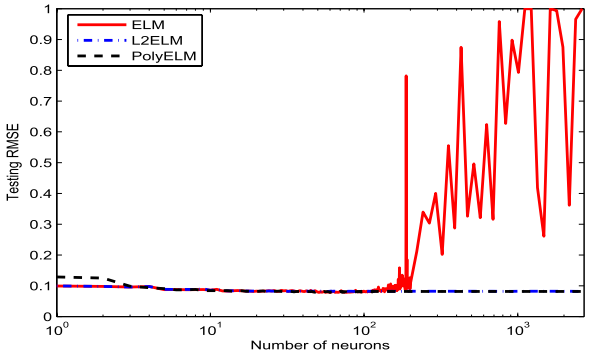


Fig. 2. Performance comparison (testing RMSE) with ELM, L_2ELM, and PolyELM for Abalone data set.

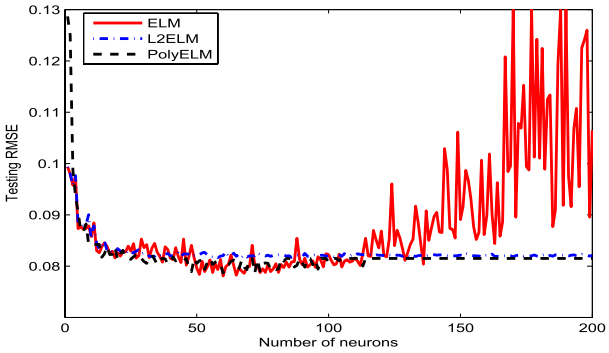


Fig. 3. Performance comparison (testing RMSE) with ELM, L_2ELM, and PolyELM for Abalone data set in detail.

[1] http://www.archive.ics.uci.edu/ml/datasets.html

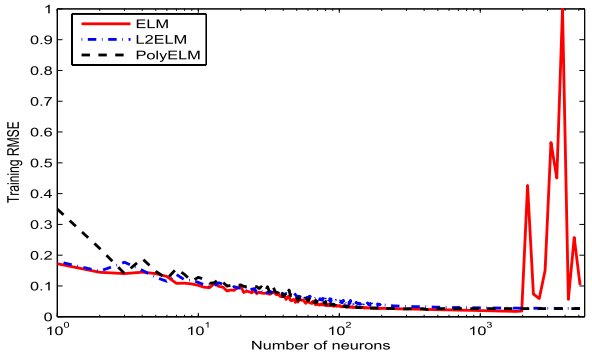


Fig. 4. Performance comparison (training RMSE) with ELM, L_2ELM, and PolyELM for Machine CPU data set.

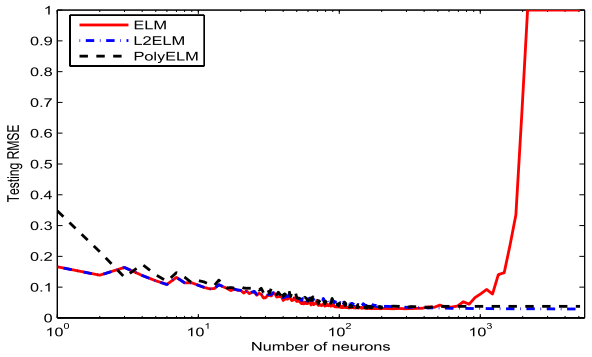


Fig. 5. Performance comparison (testing RMSE) with ELM, L_2ELM, and PolyELM for Machine CPU data set.

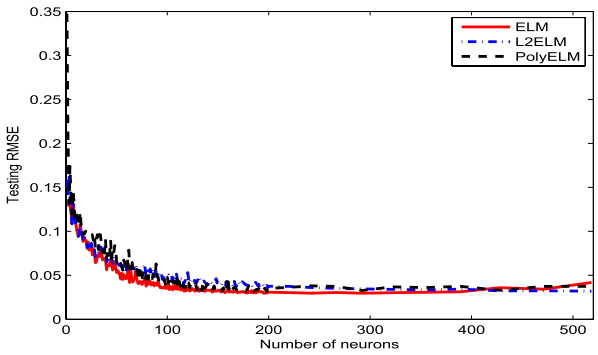


Fig. 6. Performance comparison (testing RMSE) with ELM, L_2ELM, and PolyELM for Machine CPU data set in detail.

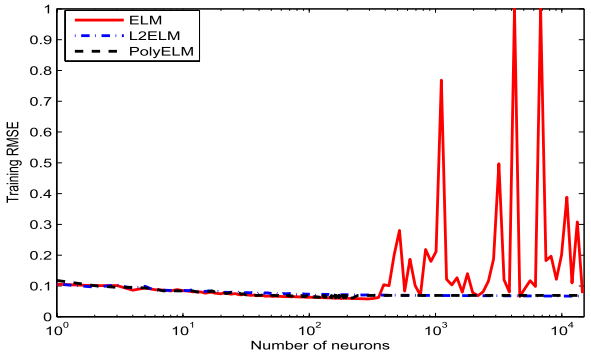


Fig. 7. Performance comparison (training RMSE) with ELM, L_2ELM, and PolyELM for Census (house8L) data set.

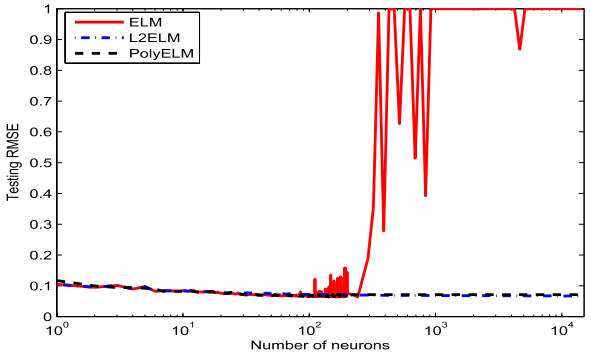


Fig. 8. Performance comparison (testing RMSE) with ELM, L_2ELM, and PolyELM for Census (house8L) data set.

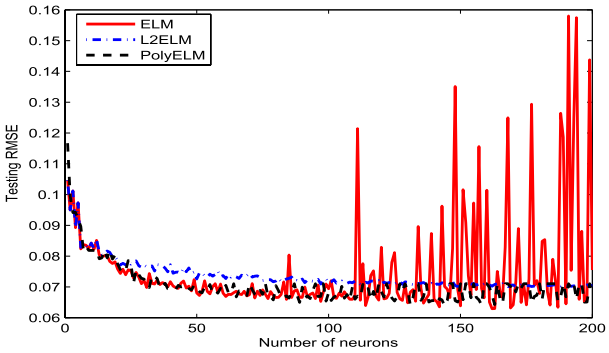


Fig. 9. Performance comparison (testing RMSE) with ELM, L_2ELM, and PolyELM for Census (house8L) data set in detail.

this is not true for ELM with sigmoid case. As far as the generalization capability is concerned, as shown in Figs. 2, 3, 5, 6, 8, and 9, L_2ELM and PolyELM both exhibit very good generalization capability and very stable with respect to n. However, ELM with sigmoid activation function (ELM) performs a little unstable, it keeps relative lower generalization error when n is within a suitable interval and then get larger and larger error when n is beyond this interval. The unstable performance in training and testing caused by the fact that when the sigmoid activation is applied, the induced hidden layer output matrix H is not necessarily weak regular. This supports the fact that either taking a good activation function (say, polynomials) or using regularization scheme can assuredly yield promising performance of ELM learning, as proved by Theorems 1 and 5.

V. Conclusion

The ELM-like learning provides a powerful complexity-reduction learning paradigm that adjusts the output layer connections only while randomly fixing the hidden parameters. Numerous experiments and applications have supported the effectiveness and efficiency of ELM. The feasibility or the theoretical foundation of the ELM-like systems has, however, been open. In this paper, we justified such feasibility through systematically answering the related three questions:

1) we showed that even ELM adjusts partial connections in an FNN, it does not degrade the generalization capability provided the activation functions used are carefully selected. Especially, we verified that for the polynomial, Nadaraya–Watson and sigmoid activation functions, ELM can realize the almost optimal generalization error bound; 2) to realize the almost optimal generalization error bound (i.e., attains the almost optimal generalization capability), a close connection between the number of hidden layer nodes and the number of training samples is also achieved; and 3) we proved that whenever the nonlinear function is algebraic polynomial, the induced hidden layer output matrix in the ELM-like systems is of full column rank, hence the well-known generalized inverse technique can be efficiently applied. For the nonpolynomial case, we further showed that Tikhonov regularization can be applied to guarantee the weak regularity without sacrificing the generalization capability. The obtained results underlie the feasibility and effectiveness of ELM-like systems from a theoretical point of view.

Several problems still open. For example, do different random assignments of the hidden parameters affect the performance of an ELM-like system? If yes, how it affects? what is the best random assignment? For other activation functions, does ELM still not degrade the generalization capability? All these problems are under our current investigation.

ACKNOWLEDGMENT

The authors would like to thank the associate editor and the anonymous reviewers for their insightful comments and suggestions on this paper.

REFERENCES

[1] G. Cybenko, "Approximation by superpositions of a sigmoidal function," *Math. Control, Signals, Syst.*, vol. 2, no. 4, pp. 303–314, 1989.

[2] T. Chen and H. Chen, "Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems," *IEEE Trans. Neural Netw.*, vol. 6, no. 4, pp. 911–917, Apr. 1995.

[3] V. Maiorov, "Approximation by neural networks and learning theory," *J. Complex.*, vol. 22, no. 1, pp. 102–117, 2006.

[4] D. E. Rumelhart, G. E. Hintont, and R. J. Williams, "Learning representations by back-propagating errors," *Nature*, vol. 323, no. 6088, pp. 533–536, 1986.

[5] G.-B. Huang, Q.-Y. Zhu, and C.-K. Siew, "Extreme learning machine: Theory and applications," *Neurocomputing*, vol. 70, no. 1, pp. 489–501, 2006.

[6] 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, Apr. 2006.

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

[8] 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.

[9] 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, Feb. 2012.

[10] H.-J. Rong, G.-B. Huang, and Y.-S. Ong, "Extreme learning machine for multi-categories classification applications," in *Proc. IEEE Int. Joint Conf. Neural Netw. (IJCNN), World Congr. Comput. Intell.*, Jun. 2008, pp. 1709–1713.

[11] Y. Miche, A. Sorjamaa, and A. Lendasse, "OP-ELM: Theory, experiments and a toolbox," in *Artificial Neural Networks-ICANN*. New York, NY, USA: Springer-Verlag, 2008, pp. 145–154.

[12] Q. Liu, Q. He, and Z. Shi, "Extreme support vector machine classifier," in *Advances in Knowledge Discovery and Data Mining*. Berlin, Germany: Springer-Verlag, 2008, pp. 222–233.

[13] B. Frénay and M. Verleysen, "Using SVMs with randomised feature spaces: An extreme learning approach," in *Proc. Eur. Symp. Artif. Neural Netw. (ESANN)*, 2010.

[14] G.-B. Huang, X. Ding, and H. Zhou, "Optimization method based extreme learning machine for classification," *Neurocomputing*, vol. 74, no. 1, pp. 155–163, 2010.

[15] E. Parviainen, J. Riihimäki, Y. Miche, and A. Lendasse, "Interpreting extreme learning machine as an approximation to an infinite neural network," in *Proc. Knowl. Discovery Inform. Retr. (KDIR)*, 2010, pp. 65–73.

[16] B. Frénay and M. Verleysen, "Parameter-insensitive kernel in extreme learning for non-linear support vector regression," *Neurocomputing*, vol. 74, no. 16, pp. 2526–2531, 2011.

[17] H.-X. Tian and Z.-Z. Mao, "An ensemble ELM based on modified AdaBoost.RT algorithm for predicting the temperature of molten steel in ladle furnace," *IEEE Trans. Autom. Sci. Eng.*, vol. 7, no. 1, pp. 73–80, Jan. 2010.

[18] J. Lu, Y. Zhao, Y. Xue, and J. Hu, "Palmprint recognition via locality preserving projections and extreme learning machine neural network," in *Proc. 9th Int. Conf. Signal Process. (ICSP)*, Beijing, China, Oct. 2008, pp. 2096–2099.

[19] B. P. Chacko, V. V. Krishnan, G. Raju, and P. B. Anto, "Handwritten character recognition using wavelet energy and extreme learning machine," *Int. J. Mach. Learn. Cybern.*, vol. 3, no. 2, pp. 149–161, 2012.

[20] A. Mohammed, R. Minhas, Q. Jonathan Wu, and M. Sid-Ahmed, "Human face recognition based on multidimensional PCA and extreme learning machine," *Pattern Recognition*, vol. 44, no. 10, pp. 2588–2597, 2011.

[21] 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.

[22] S. Saraswathi, S. Sundaram, N. Sundararajan, M. Zimmermann, and M. Nilsen-Hamilton, "ICGA-PSO-ELM approach for accurate multiclass cancer classification resulting in reduced gene sets in which genes encoding secreted proteins are highly represented," *IEEE/ACM Trans. Comput. Biol. Bioinform.*, vol. 8, no. 2, pp. 452–463, Feb. 2011.

[23] C. R. Rao and S. K. Mitra, *Generalized Inverse of Matrices and its Applications*. New York, NY, USA: Wiley, 1971.

[24] D. Serre, *Matrices: Theory and Applications*. New York, NY, USA: Springer-Verlag, 2010.

[25] H. Jaeger, "The 'echo state' approach to analysing and training recurrent neural networks-with an erratum note," German Nat. Res. Center Inform. Technol., Berlin, Germany, GMD Rep. 148, Jan. 2010.

[26] H. Jaeger, *Tutorial on Training Recurrent Neural Networks, Covering BPPT, RTRL, EKF and the Echo State Network Approach*. Sankt Augustin, Germany: GMD-Forschungszentrum Informationstechnik, 2002.

[27] H. Jaeger and H. Haas, "Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication," *Science*, vol. 304, no. 5667, pp. 78–80, 2004.

[28] Y. Pao, *Adaptive Pattern Recognition and Neural Networks*. Boston, MA, USA: Addison-Wesley, 1989.

[29] F. Cucker and S. Smale, "On the mathematical foundations of learning," *Amer. Math. Soc.*, vol. 39, no. 1, pp. 1–49, 2002.

[30] 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.

[31] G. H. Golub and C. F. Van Loan, *Matrix Computations*. Baltimore, MD, USA: JHU Press, 2012.

[32] 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.

[33] H. T. Huynh, Y. Won, and J.-J. Kim, "An improvement of extreme learning machine for compact single-hidden-layer feedforward neural networks," *Int. J. Neural Syst.*, vol. 18, no. 05, pp. 433–441, 2008.

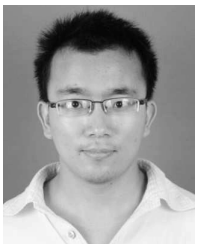
[34] H. T. Huynh and Y. Won, "Regularized online sequential learning algorithm for single-hidden layer feedforward neural networks," *Pattern Recognit. Lett.*, vol. 32, no. 14, pp. 1930–1935, 2011.

[35] L. Györfi, M. Kohler, A. Krzyzak, and H. Walk, *A Distribution-Free Theory of Nonparametric Regression*. New York, NY, USA: Springer-Verlag, 2002.

[36] R. DeVore, G. Kerkyacharian, D. Picard, and V. Temlyakov, "Approximation methods for supervised learning," *Found. Comput. Math.*, vol. 6, no. 1, pp. 3–58, 2006.
[37] S. P. Boyd and L. Vandenberghe, *Convex Optimization*. Cambridge, U.K.: Cambridge Univ. Press, 2004.
[38] R. DeVore and G. Lorentz, *Constructive Approximation*. New York, NY, USA: Springer-Verlag, 1993.
[39] A. Pinkus, *n-Widths in Approximation Theory*. Berlin, Germany: Springer-Verlag, 1985.
[40] M. Leshno, V. Y. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with a nonpolynomial activation function can approximate any function," *Neural Netw.*, vol. 6, no. 6, pp. 861–867, 1993.
[41] S. Lin, X. Liu, J. Fang, and Z. Xu, "Is extreme learning machine feasible? A theoretical assessment (part II)," *IEEE Trans. Neural Netw. Learn. Syst.*, to be published.
[42] J. L. Kelley and M. Stone, *General Topology*, vol. 233. New York, NY, USA: Van Nostrand, 1955.

Xia Liu was born in Yulin, China, in 1984. She received the M.S. degree in basic mathematics from Yan'an University, Yan'an, China, in 2010. She is currently pursuing the Ph.D. degree with the School of Mathematics and Statistics, Xi'an Jiaotong University, Xi'an, China.

She has been with the Department of Mathematics, Yulin University, Yulin, since 2010. Her current research interests include learning theory and nonlinear functional analysis.



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

His current research interests include machine learning and scattered data fitting.

Jian Fang was born in Jiangsu, China, in 1986. He received the B.Sc. 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 the National Basic Research Program of China (973 Project), and the Director of the Institute for Information and System Sciences, Xi'an Jiaotong University, where he is a member of the Chinese Academy of Sciences. His current research interests include intelligent information processing and applied mathematics.

Prof. Xu was a recipient of the National Natural Science Award of China in 2007, and the CSIAM Su Buchin Applied Mathematics Prize in 2008. He delivered a 45 minute talk on the International Congress of Mathematicians in 2010.