Stochastic Choice of Basis Functions in Adaptive Function Approximation and the Functional-Link Net

Boris Igelnik and Yoh-Han Pao, *Fellow, IEEE*

***Abstract*— A theoretical justification for the random vector version of the functional-link (RVFL) net is presented in this paper, based on a general approach to adaptive function approximation. The approach consists of formulating a limit-integral representation of the function to be approximated and subsequently evaluating that integral with the Monte-Carlo method. Two main results are: 1) the RVFL is a universal approximator for continuous functions on bounded finite dimensional sets, and 2) the RVFL is an efficient universal approximator with the rate of approximation error convergence to zero of order $O(C/\sqrt{n})$, where n is number of basis functions and with C independent of n. Similar results are also obtained for neural nets with hidden nodes implemented as products of univariate functions or radial basis functions. Some possible ways of enhancing the accuracy of multivariate function approximations are discussed.**

I. Introduction

The primary purpose of this paper is to give a theoretical justification for the functional-link net which was proposed by one of the authors [1] and which has been shown to be capable of excellent performance in various areas of applications [2]–[4].

In the process of writing the paper, however, we concluded that a method developed for that purpose can be applied not only in network-based neurocomputing but can also serve as a more general method for adaptive function approximation. We address both issues in this paper.

The essence of the method is in the use of a limit-integral representation of the function to be approximated and subsequent evaluation of the integral by the Monte-Carlo approach [5], [6]. The integration is made over the space of the parameters which specify basis functions in the approximation of the function.

Let $f \in C(I^d)$ be a continuous function, defined on the standard hypercube $I^d = [0; 1]^d \subset R^d$, and consider a limit-integral representation of the function f

$$f(x) = \lim_{L \to a} \int_V T[f(\lambda)] G_{\lambda, L}(x)\, d\lambda$$

where L and λ are low-dimensional (one or two) and multidimensional parameters, respectively, G is an activation function, T is an operator, defined on $C(I^d)$, a is a finite or infinite real number, and V is the domain of the parameter λ.

We attain an approximation of the function f with use of two stages of approximation. The first stage consists of approximating the limiting value of the integral by the integral

$$f(x) \approx \int_V T[f(\lambda)] G_{\lambda, l}(x)\, d\lambda$$

where $l \approx a$. The second stage consists of obtaining an estimate of the multiple integral with use of the Monte-Carlo method

$$\int_V T[f(\lambda)] G_{\lambda, l}(x)\, d\lambda \approx \frac{|V|}{n} \sum_{k=1}^{n} T[f(\lambda_k)] G_{\lambda_k, l}(x)$$

where $\lambda = (\lambda_1, \cdots, \lambda_n)$ is a sample of size n drawn from the uniform distribution on V, that is, a set of n random variables, independent and uniformly distributed on V. Information about the function is incorporated in the coefficients $a_k = (|V|/n) T[f(\lambda_k)]$, $k = 1, \cdots, n$, available in the "learning" phase of function approximation task and so we arrive at a representation

$$f(x) \approx f_{n, \lambda, l}(x)$$

where

$$f_{n, \lambda, l}(x) = \sum_{k=1}^{n} a_k G_{\lambda_k, l}(x), \quad \lambda = (\lambda_1 \cdots, \lambda_n, l).$$

$G_{\lambda_k, l}$ can be regarded as basis functions in this representation. They are parameterized by random variables λ_k, which are not to be learned, and a deterministic (but low-dimensional!) variable l. Thus the learning in this approach is a linear one and, therefore, simple and fast. This is one evident advantage of the approach. In practice we use the conjugate gradient method of optimization [7] for learning. The second advantage of this stochastic approach is that the error of the Monte-Carlo approximation tends to zero as $n \to \infty$ in the manner of $C/\sqrt{n}$, where C is independent of n (but, generally speaking, not of d) and is determined by the variance of the integrand. Thus the Monte-Carlo approach is an efficient one in the approximation of multiple integrals. The question about efficiency of the overall approximation is more complex and subtle. Indeed the overall error of approximation can be bounded by sum of the error of approximating the limiting value of the integral by the integral and the error of approximating the integral by the finite sum using the Monte-Carlo method. Both errors depend on l (the error of the Monte-Carlo method through C). C may depend on l in such a way that it tends to infinity when $l \to a$. Nevertheless

Manuscript received October 2, 1993; revised February 15, 1995 and April 20, 1995. This work was supported in part by the Air Force Office of Scientific Research (F33615-90-C-5944) and Material Directorate, Wright Laboratory (F33615-87-C-5250).

The authors are with the Electrical Engineering and Applied Physics Department, Case Western Reserve University, Cleveland, OH 44106 USA.

IEEE Log Number 9413443.

1045–9227/95$04.00 © 1995 IEEE

in some cases, including that of our paper, C can be bounded and the overall error of approximation is of the order of $1/\sqrt{n}$.

To fix ideas, we cite, as illustration, the Poisson representation [8] of a continuous function, defined on the interval $[-\pi, \pi]$, namely

$$f(x) = \lim_{r \to 1-0} \frac{1}{2\pi} \int_{-\pi}^{\pi} f(u) \frac{1-r^2}{1-2r\cos(u-x)+r^2}\, du.$$

Following our procedure, this limit-integral representation would lead to the following adaptive approximation

$$f(x) \approx \sum_{k=1}^{n} a_k \frac{1-l^2}{1-2l\cos(u_k - x)+l^2}, \quad l < 1$$

where the random parameters u_k are drawn from $[-\pi, \pi]$ and l is the one additional parameter to be adapted for minimal error of approximation.

In this paper we consider a special case of this general method, the random vector version of the functional-link (RVFL) net. For every $f \in C(I^d)$ we define an RVFL net as an one hidden layer feedforward neural net of the form

$$f_n(x) = \sum_{k=1}^{n} a_k g(w_k x + b_k) \tag{1}$$

where $a_k, b_k \in R$, $w_k \in R^d$, $x \in I^d$. RVFL has the same form as the general nonlinear perceptron (GNP), except that in the RVFL the parameters w_k, b_k of the hidden layer are selected randomly and independently in advance and parameters a_k of the output layer are learned using simple quadratic optimization, while in the GNP all parameters a_k, w_k, b_k need to be learned using complex nonquadratic optimization. Recently the universal approximation capability of the GNP was proved for very general choice of activation function g [9]–[14]. There exists a number of papers as well [15]–[20], where the efficiency of the approximation by the GNP was investigated. In particular Barron [16] has proved that the approximation error for the GNP tends to zero with the rate not worse than that of the order $O(1/\sqrt{n})$ (The approximation error is defined as a distance between a function to be approximated and the best approximation in a given class of approximations). At the same time Barron [16] showed that in case of GNP with fixed basis functions (that is with w_k, b_k defined deterministically in advance) there is no chance of avoiding exponential growth, in d, the number of basis functions. We also came to the same conclusion in our earlier discussions of the functional-link net [21]. The RVFL turns out to be a practical compromise between the full GNP and those GNP's with fixed basis functions, combining both simplicity of learning and efficiency of representation since, as we show, the rate of error convergence for the RVFL is of the same order as for the GNP.

An intuitive argument explaining the universal approximation capability of the RVFL can be given in the form of the following proposition.

Proposition—RVFL Networks Are Universal Approximators: Suppose a continuous function f is to be approximated on the bounded set in R^d. There exists a single hidden layer feedforward net N_1 with certain weights which approximate f within $\varepsilon/2$ [9]–[11]. A partly random selection of weights will eventually produce an RVFL net N_2, the effect of whose weights are very close to those of N_1, close enough so that N_2 approximates N_1 within $\varepsilon/2$. Then N_2 approximates f within ε.

Recently we showed that combined use of an ensemble of RVFL networks can result in the minimization of generalization error. We give here a brief discussion of that matter.

Explanation: In the RVFL the generalization error $E_{n,N}(\theta)$ is a function of a parameter vector $\theta = (w_1, \cdots, w_n, b_1, \cdots, b_n)$. Suppose the function $E_{n,N}(\theta)$ is defined on the bounded set Θ and $\theta_* = (\theta_1, \cdots, \theta_{2n}) = \arg\min_{\theta \in \Theta} E_{n,N}(\theta)$. Using the Pincus formula for estimating the point of minimum of a continuous function [28], we have

$$\theta_{*j} = \lim_{\lambda \to \infty} \frac{\int_{\Theta} \theta_j \exp[-\lambda E_{n,N}(\theta)]\, d\theta}{\int_{\Theta} \exp[-\lambda E_{n,N}(\theta)]\, d\theta}, \quad j = 1, \cdots, 2n.$$

Approximating in the right-hand side of this equation the limit (using finite but large enough λ) and the integrals (using the Monte-Carlo method), we obtain

$$\theta_{*j} \approx \frac{\sum_{k=1}^{K} \theta_{kj} \exp[-\lambda E_{n,N}(\theta_k)]}{\sum_{k=1}^{K} \exp[-\lambda E_{n,N}(\theta_k)]}, \quad j = 1, \cdots, 2n$$

where $\theta_k = (\theta_{k1}, \cdots, \theta_{k,2n})$, $k = 1, \cdots, K$ is a sample of parameter θ from the uniform distribution in Θ. From this approximation the following construction stems. Generate K RVFL nets, and find an estimate of the best parameter θ by the last formula. Determine parameters $a_{*1}, \cdots, a_{*n}$ by training using fast quadratic optimization. The net with the parameter values $w_{*1}, \cdots, w_{*n}, b_{*1}, \cdots, b_{*n}, a_{*1}, \cdots, a_{*n}$ should be close to the optimal one. The training procedure consists of $(K+1)$ quadratic optimization steps.

Thus it is possible to generate K RVFL nets and determine an estimate of the optimal parameter θ using the approximate Pincus construction and the fast quadratic optimization characteristic of the RVFL.

This explains why the RVFL construct is a powerful one and easy to understand intuitively.

As a digression but also to provide additional insight, we ask and answer the following question, namely, how does it happen that the same order of accuracy of approximation can be achieved by both the full GNP and its simplified (in terms of number of learning parameters) version, the RVFL?

We get an answer by looking over two corresponding proofs, one by Barron and another by ourselves. In the case of GNP, to get a bound for the approximation error Barron [16] approximates some integral $\int T[f(\lambda)]G_\lambda(x)\, d\lambda$, using the Monte-Carlo method implicitly, but in a way different from ours. He transforms $T[f(\lambda)]G(\lambda)\,d\lambda$ to a new form $\overline{T}[f(\lambda)]\overline{G}(\lambda)\,d\lambda$ so that $\overline{T}[f(\lambda)]\,d\lambda$ can be represented as a probability measure $\mu(d\lambda)$. Then the integral is approximated as follows

$$\int T[f(\lambda)]G_{\lambda}(x)\,d\lambda \approx \frac{1}{n}\sum_{k=1}^{n}\overline{G}_{\lambda_k}(x)$$

where λ_k are drawn randomly from the domain V, supplied by the measure $\mu(d\lambda)$. Therefore all information about a function to be approximated is included in the mechanism of random selection (which is not convenient in practice), and there are no coefficients in the linear combination of the basis functions to be learned. In case of the RVFL we use a simple mechanism for random selection of the parameters (uniform distribution) and include all information about unknown function in the parameters to be learned.

To show clearly the difference between these two approaches, consider briefly the basic ideas of the Monte-Carlo method for multiple integral evaluation. Suppose that our task is the evaluation of the multiple integral $J = \int_{I^d} F(x)\,dx$ and that the task cannot be fulfilled in the explicit form because of the complex analytic nature of the function F. The simple Monte-Carlo method is based on the fact that the integral J can be also represented as an expectation of the random variable $F(\xi)$, where ξ is the random variable uniformly distributed in I^d. The evaluation of the integral, therefore, can be done by generating a sample $(\xi_1, \cdots, \xi_n)$ from this uniform distribution and taking the estimate of the integral in the form

$$J \approx \frac{F(\xi_1) + \cdots + F(\xi_n)}{n}.$$

The right-hand side of this equation is a random variable with the expected value equal J. The error of the Monte-Carlo approximation can be evaluated as

$$\frac{k\sqrt{\operatorname{var} F(\xi)}}{\sqrt{n}}$$

where $\operatorname{var} F(\xi)$ is a variance of the random variable $F(\xi)$ and k is a constant which depends on the confidence level. The essence of the method is that the error of approximation does not depend on the dimension d of the integration space in terms of the size n of the sample and may depend on d only through the variance $\operatorname{var} F(\xi)$. In more complicated versions of the Monte-Carlo method (known as variance reduction methods [5], [6]), the integrand is transformed so to decrease the variance. One of the general variance reduction methods is the method of importance sampling. In that method we use a function p which satisfies the conditions

$$p(x) \geq 0, \quad x \in I^d$$
$$\int_{I^d} p(x) = 1.$$

Then we have

$$\begin{aligned} J &= \int_{I^d} F(x)\,dx \\ &= \int_{I^d} \frac{F(x)}{p(x)}\,p(x)\,dx \\ &= E_{\mu}\,\frac{F(\varsigma)}{p(\varsigma)} \end{aligned}$$

where expectation E_μ of the random variable $F(\varsigma)/p(\varsigma)$, is taken with respect to the probability measure $\mu(dx) = p(x)\,dx$. Thus the evaluation of the integral in this case can be made as follows

$$J \approx \frac{F(\varsigma_1)/p(\varsigma_1) + \cdots F(\varsigma_n)/p(\varsigma_n)}{n}$$

where $(\varsigma_1, \cdots, \varsigma_n)$ is a sample from the nonuniform distribution of random variable ς having density function p. Therefore generating the sample $(\varsigma_1, \cdots, \varsigma_n)$ is not as simple as in the case of simple Monte-Carlo method. Moreover to get a real variance reduction we need to include in p some information about the unknown integral. Indeed the variance $\operatorname{var}[F(\varsigma)/p(\varsigma)]$ can be evaluated as

$$\operatorname{var}\left[\frac{F(\varsigma)}{p(\varsigma)}\right] = \int_{I^d}\left[\frac{F(x)}{p(x)} - J\right]^2 p(x)\,dx.$$

If $F(x) \geq 0$ for $x \in I^d$, then taking $p(x) = F(x)/J$ we evidently obtain $\operatorname{var}[F(\varsigma)/p(\varsigma)] = 0$. The error of approximation is reduced to zero but we need to know not only the integral J but even F! The conclusion is that to reduce the approximation error, we would need to increase the complexity of the sample generating, which equivalent to increasing information about the unknown quantities.

Returning to the previous discussion we can say that our approach and Barron's approach differ in that we use these two extreme version of the Monte-Carlo method (simple Monte-Carlo and maximum variance reduction).

The results obtained in the paper are given in the form of three theorems. In Theorem 1 we prove that the RVFL is a universal approximator of continuous functions. In Theorem 2 we prove that radial basis functions with random parameters also can serve as a universal approximators for the same class of functions. In Theorem 3 we prove that the approximation error for RVFL tends to zero with the rate $O(C/\sqrt{n})$ for a given class of functions to be approximated, that is this kind of neural networks is efficient for multivariate function approximation.

The organization of the paper is as follows. We state our results in Section II and give a sketch of the proof in Section III. Section IV is devoted to citing some examples of use of the RVFL. Conclusions and recommendations are contained in Section V. We give proofs of the theorems in the Appendix.

Briefly, results of the paper give theoretical justification for the RVFL and open ways both for enhancing efficiency of the RVFL and for the investigation of other perspective adaptive approximations using stochastic approaches. Qualitative discussion of the present results have been reported at the conferences and published in papers [22] and [23].

II. Main Results

We consider a continuous function $f \in C(I^d)$ and an RVFL, which we denote as

$$f_{\omega_n}(x) = \sum_{k=1}^{n} a_k g(w_k \cdot x + b_k) \tag{2}$$

where $\omega_n = (n, a_1, \cdots, a_n, b_1, \cdots, b_n, w_1, \cdots, w_n)$ is an overall parameter of the net $w_k \cdot x$ is an inner product of the vectors w_k, x. The random part of ω_n is denoted as $\lambda_n = (w_1, \cdots, w_n, b_1, \cdots, b_n)$. Suppose that λ_n is defined on the probabilistic space $S_n(\Omega, \alpha)$ with probability measure $\mu_{n, \Omega, \alpha}$ and E is a symbol of expectation with respect to $S_n(\Omega, \alpha)$. We assume that $S_n(\Omega, \alpha)$ and $\mu_{n, \Omega, \alpha}$ depend on a deterministic parameter (Ω, α) which should be determined in the learning stage. The distance between f and f_{ω_n} on any compact set K, $K \subset I^d$ can be defined as

$$\rho_K(f, f_{\omega_n}) = \sqrt{E \int_K [f(x) - f_{\omega_n}(x)]^2 \, dx}. \tag{3}$$

Then our first result is the following.

Theorem 1: For any compact K, $K \subset I^d$, $K \neq I^d$ and any absolutely integrable activation function g such, that

$$\int_R g^2(x) \, dx < \infty \tag{4}$$

there exist a sequence of RVFL $\{f_{\omega_n}\}$ and a sequence of probability measures $\{\mu_{n, \Omega, \alpha}\}$ such that

$$\rho_K(f, f_{\omega_n}) \underset{n \to \infty}{\longrightarrow} 0. \tag{5}$$

The probability measures $\mu_{n, \Omega, \alpha}$ can be specified as follows. Let $\hat{w}_0 = (\hat{w}_{01}, \cdots, \hat{w}_{0d})$, $y_0 = (y_{01}, \cdots, y_{0d})$ and u_0 be independent and uniformly distributed in $V^d = [0; \Omega] \times \cdots \times [-\Omega; \Omega]$, I^d and $[-2\Omega; 2\Omega]$, respectively, $w_0 = \alpha \hat{w}_0$, $b_0 = -w_0 \cdot y_0 - u_0$. Then $(w_1, \cdots, w_n)$ and $(b_1, \cdots, b_n)$ are two samples from the distributions of w_0 and b_0, respectively.

Thus Theorem 1 states that RVFL is a universal approximator for any continuous function on any compact set which is inside I^d. The distribution of the random parameter λ_n is simple in generating. Examples of activation functions, which satisfy conditions of the theorem, are Gaussian, subsequent derivatives of Gaussian, any integrable with square functions with finite support. To include as an activation function sigmoidal functions (with some, not important in practice, restrictions) we prove a corollary.

Corollary 1: For any compact K, $K \subset I^d$, $K \neq I^d$ and any differentiable activation function g, such that

$$\int_R [g'(x)]^2 \, dx < \infty \tag{6}$$

there exist a sequence of RVFL $\{f_{\omega_n}\}$ and a sequence of probability measures $\{\mu_{n, \Omega, \alpha}\}$ such that

$$\rho_K(f, f_{\omega_n}) \underset{n \to \infty}{\longrightarrow} 0.$$

The probability measures $\mu_{n, \Omega, \alpha}$ are determined in Theorem 1.

Thus any activation function considered in the practice of neural net computing can be used in RVFL as well.

It is interesting to note that in the process of proving Theorem 1 we obtained as an intermediate result, that adaptive universal approximation for continuous function can be also taken in the form

$$f_{\omega_n}(x) = \sum_{k=1}^{n} a_k \prod_{i=1}^{d} g(w_{ki} x_i + b_{ki}) \tag{7}$$

where $x = (x_1, \cdots, x_d)$, $w_k = (w_{k1}, \cdots, w_{kd})$, $b_k = (b_{k1}, \cdots, b_{kd})$. We state the result in Corollary 2. If $f_{\omega_n}(x)$ is defined by the formula (7), $\rho_K(f, f_{\omega_n})$ is defined by the formula (3), then

$$\rho_K(f, f_{\omega_n}) \underset{n \to \infty}{\longrightarrow} 0.$$

The probability measures $\mu_{n, \Omega}$ can be specified as follows. Let $\hat{w}_0$, y_0 are determined in Theorem 1, $w_0 = \hat{w}_0$, $b_0 = -w_0 \circ y_0$, $w_0 \circ y_0 = (w_{01} y_{01}, \cdots, w_{0d} y_{0d})$ is an outer product of two vectors $w_0 = (w_{01}, \cdots, w_{0d})$, $b_0 = (b_{01}, \cdots, b_{0d})$. Then $(w_1, \cdots, w_n)$ and $(b_1, \cdots, b_n)$ are two samples from the distributions of w_0 and b_0, respectively.

Under random choice of parameters w_k, b_k we do not see an advantage of representation of the basis function in the form $g(\sum_{i=1}^d w_{ki} x_i)$ compared with the form $\prod_{i=1}^d g(w_{ki} x_i + b_{ki})$ (at least we cannot prove it).

We proved as well the universal approximation capability for adaptive approximation using radial basis functions with random parameters. The corresponding result is stated in the following theorem.

Theorem 2: Let an adaptive approximation be taken in the form

$$f_{\omega_n}(x) = \sum_{k=1}^{n} a_k g\{w_k \cdot [(x - y_k) \circ (x - y_k)]\} \tag{8a}$$

where $\omega_n = (n, a_1, \cdots, a_n, y, \cdots, y_n, w_1, \cdots, w_n)$ is the overall parameter of approximation, the distance between f and f_{ω_n} is defined by (3), and the distribution of random parameters w_k, y_k is defined in Corollary 2, except that now $b_0 = -w_0 \cdot y_0$. Then for any compact K, $K \subset I^d$, $K \neq I^d$ and any absolutely integrable activation function g satisfying (4) or any differentiable activation function g, satisfying (6), and additionally the condition

$$\int_{R^d} g\left(\sum_{i=1}^{d} z_i^2\right) dz < \infty \tag{8b}$$

we have

$$\rho_K(f, f_{\omega_n}) \underset{n \to \infty}{\longrightarrow} 0.$$

We need condition (8b) to normalize g by equation

$$\int_{R^d} g\left(\sum_{i=1}^{d} z_i^2\right) dz = 1.$$

Now we state our result concerning the efficiency of the RVFL. First of all we introduce some restrictions on a continuous function f to be approximated. We suppose that this function satisfies the Lipshitz condition [24], that is, there exists a constant $\kappa > 0$ such, that for any x, $y \in I^d$

$$|f(x) - f(y)| \le \kappa ||x - y|| \tag{9}$$

where $||x - y|| = \sum_{i=1}^{d} |x_i - y_i|$. Thus we narrow the class of continuous functions to the class of smoother functions satisfying (9). Second, we apply some tuning of the activation function g, namely we suppose that support of the function g is in $\prod_{i=1}^{d} [-\beta w_i; \beta w_i]$ and denote it as g_β. Then we have Theorem 3.

Theorem 3: For any $f \in C(I^d)$, satisfying (9), any compact K, $K \subset I^d$, $K \neq I^d$, any activation function g_β, satisfying conditions (4) or (6), there exist a sequence of RVFL $\{f_{\omega_n}\}$, a sequence of probability measures $\{\mu_{n,\Omega,\alpha}\}$ and a constant $C_{f,g,\Omega,\alpha,\beta,d}$ such, that

$$\rho_K(f, f_{\omega_n}) \le \frac{C_{f,g,\Omega,\alpha,\beta,d}}{\sqrt{n}}. \tag{10}$$

Probability measures $\mu_{n,\Omega,\alpha}$ are defined by Theorem 1.

In the last theorem, we prove that RVFL is not only an universal but an efficient universal approximator. Indeed the constant $C_{f,g,\Omega,\alpha,\beta,d}$ does not depend on n and the approximation error is of the order of $O(C/\sqrt{n})$, while approximation by linear combination of fixed basis functions in the given smoothness class of the functions to be approximated gives an approximation error on the order of $O(1/n^{1/d})$ [25].

III. Sketch of the Proofs

We discuss our basic ideas and then present an outline of our proofs. The proofs of corollaries from Theorem 1 are presented in this section because of their simplicity.

We now consider a sketch of the proof of Theorem 1. We represent a function to be approximated as the limiting value of a multidimensional integral over parameter space. The integrand of the integral is constructed so that it represents a window transformation of the function in a neighborhood of a given point x. The window transformation is made by a function

$$h_x(y, w) = \prod_{i=1}^{d} w_i g[w_i(x_i - y_i)]$$

where vectors $x = (x_1, \cdots, x_d)$, $w = (w_1, \cdots, w_d)$ determine, respectively, the location and the shape of the window. Without loss of generality we can assume that g satisfies the condition

$$\int_R g(z)\, dz = 1.$$

This condition can be dropped in the final stage of this proof. Then the function $h_x(y, w)$ satisfies the equation

$$\int_{R^d} h_x(y, w)\, dy = 1$$

for any $x, w \in R^d$. The function $h_x(y, w)$ approaches the multidimensional delta function $\delta_x(y)$ as $w_1, \cdots, w_d \to \infty$. Since $\delta_x(y)$ satisfy the equations

$$\delta_x(y) = 0, \quad \text{if} \quad y \neq x$$

$$\int_{R^d} \delta_x(y)\, dy = 1$$

we come to the limit-integral representation

$$f(x) = \lim_{w_1 \to \infty} \cdots \lim_{w_d \to \infty} \int_{I^d} f(y) h_x(y, w)\, dy. \tag{11}$$

Considering $F(w) = \int_{I^d} f(y) h_x(y, w)\, dy$ as a function of w and applying the l'Hospital rule [24] to this function, we obtain the limit-integral representation of the function f in the form

$$f(x) = \lim_{\Omega_1 \to \infty} \cdots \lim_{\Omega_d \to \infty} \frac{1}{\prod_{i=1}^{d} \Omega_i} \int_{I^d \times \Omega^d} \cdot f(y) h_x(y, w)\, dy\, dw \tag{12}$$

where $\Omega^d = [0; \Omega_1] \times \cdots \times [0; \Omega_d]$. Henceforth we consider $\Omega_1 = \cdots = \Omega_d = \Omega$.

The second step is replacing the function $h_x(y, w)$ by the function $g[w \cdot x + b(y)]$, where g is an activation function satisfying the conditions of Theorem 1. We do it in the two-stage procedure. First replace, temporarily, the general activation function g by the special one, $\sin_\Omega$, which coincides with sine-function on the interval $[-2\Omega; 2\Omega]$ and equals zero outside this interval. Then applying the identity relationship $\sin_\Omega a\, sin_\Omega b = [\cos_\Omega (a - b) - \cos_\Omega (a + b)]/2$, repeatedly, $(d - 1)$ times, we transform the integral in (12) to the form

$$\frac{1}{2^{d-1}} \sum \pm \int_{I^d \times \Omega^d} f(y) \cos_\Omega \cdot [w_1(x_1 - y_1) \pm w_2(x_2 - y_2) \pm w_d(x_d - y_d)]\, dy\, dw$$

where summation is made over all combinations of + and −. Replacing all variables $-w_i$ by w_i we come to the formula

$$f(x) = \lim_{\Omega \to \infty} \frac{1}{\Omega^d 2^{d-1}} \int_{I^d \times V^d} \cdot f(y) \cos_\Omega [w \cdot (x - y)] \prod_{i=1}^{d} w_i\, dy\, dw \tag{13}$$

where $V^d = [0; \Omega] \times [-\Omega; \Omega]^{d-1}$.

The second stage in this step is replacing $\cos_\Omega$ by the general activation function g in (13). Using (11) we can represent $\cos_\Omega$ on any compact $K \subset [-2\Omega; 2\Omega]$ uniformly by formula

$$\cos_\Omega (z) = \lim_{\alpha \to \infty} \int_{-2\Omega}^{2\Omega} \cos_\Omega (u) g[\alpha(z - u)] \alpha\, du. \tag{14}$$

Substituting (14) in (13) we obtain

$$f(x) = \lim_{\alpha \to \infty} \lim_{\Omega \to \infty} \frac{1}{\Omega^d 2^{d-1}} \int_{-2\Omega}^{2\Omega} du \int_{I^d \times V^d} f(y) \cos_\Omega(u) \cdot g\{\alpha[w \cdot (x - y) - u]\} \alpha \prod_{i=1}^{d} w_i\, dy\, dw. \tag{15}$$

If we denote

$$\begin{aligned}\omega &= (y, w, u),\\ d\omega &= dy\,dw\,du,\\ W^d &= [-2\Omega; 2\Omega] \times I^d \times V^d,\\ b &= -(\alpha w \cdot y + u)\end{aligned}$$

and

$$F_{\alpha,\Omega}(y, w, u) = \frac{\alpha \prod_{i=1}^{d} w_i}{\Omega^d 2^{d-1}} f(y) \cos_\Omega(u)$$

then formula (15) can be rewritten in the form

$$f(x) = \lim_{\alpha\to\infty} \lim_{\Omega\to\infty} \int_{W^d} F_{\alpha,\Omega}(\omega) g(\alpha w \cdot x + b)\, d\omega. \tag{16}$$

Formula (16) gives us the basis limit-integral representation of the function f.

Our next step in proving Theorem 1 is to approximate the integral on the right side of (16) by the Monte-Carlo method. Let $(u_1, \cdots, u_n)$, $(y_1, \cdots, y_n)$, $(w_1, \cdots, w_n)$ be independent samples of size n, drawn from the uniform distributions in $[-2\Omega; 2\Omega]$, I^d, V^d, respectively, and $(\omega_1, \cdots, \omega_n)$ the corresponding sample drawn from W^d. Then we have

$$\int_{W^d} F_{\alpha,\Omega}(\omega) g(\alpha w \cdot x + b)\, d\omega \sim \frac{4\Omega^d}{n} \sum_{k=1}^{n} F_{\alpha,\Omega}(\omega_k) g(\alpha w_k \cdot x + b_k). \tag{17}$$

This notation means that

$$E \int_K dx \left[\int_{W^d} F_{\alpha,\Omega}(\omega) g(\alpha w \cdot x + b)\, d\omega - \frac{4\Omega^d}{n} \sum_{k=1}^{n} F_{\alpha,\Omega}(\omega_k) g(\alpha w_k \cdot x + b_k) \right]^2 \xrightarrow[n\to\infty]{} 0$$

according to the theory of the Monte-Carlo method. If we denote $a_k = (4\Omega^d/n)\, F_{\alpha_i,\Omega_i}(\omega_k)$, $k = 1, \cdots, n$ and consider random variable αw as a random variable w drawn from $V_\alpha^d = [0; \alpha\Omega] \times [-\alpha\Omega; \alpha\Omega]^{d-1}$, then we can rewrite (17) in the form

$$\int_{W^d} F_{\alpha,\Omega}(\omega) g(\alpha w \cdot x + b)\, d\omega \sim \sum_{k=1}^{n} a_k g(w_{ki} \cdot x + b_k). \tag{18}$$

Combining (16) and (18), we complete the proof of Theorem 1.

Proof of Corollary 1: It is sufficient to note that, under conditions of Corollary 1, g' satisfies the conditions of Theorem 1 and therefore

$$\cos_\Omega(z) = \lim_{\alpha\to\infty} \int_{-2\Omega}^{2\Omega} \cos_\Omega(u) g'[\alpha(z - u)]\alpha\, du. \tag{19}$$

Integrating (19) by parts, we obtain

$$\cos_\Omega(z) = \lim_{\alpha\to\infty} \cos_\Omega(u) g[\alpha(z-u)]|_{u=2\Omega}^{-2\Omega} - \lim_{\alpha\to\infty} \int_{-2\Omega}^{2\Omega} \sin_\Omega(u) g[\alpha(z-u)]\, du.$$

Since Ω can be chosen so that $\cos_\Omega z|_{u=2\Omega}^{-2\Omega} = 0$, we obtain

$$\cos_\Omega(z) = -\lim_{\alpha\to\infty} \int_{-2\Omega}^{2\Omega} \sin_\Omega(u) g[\alpha(z-u)]\, du. \tag{20}$$

Substituting (20) in (13) and repeating the previous argument, we receive the formula

$$\int_{W^d} F_{\alpha,\Omega}(\omega) g'(\alpha w \cdot x + b)\, d\omega \sim \sum_{k=1}^{n} a_k g(w_k \cdot x + b_k) \tag{21}$$

completing the proof.

Proof of Corollary 2: It is sufficient to apply the Monte-Carlo method to the evaluation of the integral in (12). Then we obtain

$$\int_{I^d\times\Omega^d} f(y) h_x(y, w)\, dy\, dw \sim \frac{1}{n} \sum_{k=1}^{n} f(y_k) \prod_{i=1}^{d} w_{ki} g[w_{ki}(x_i - y_{ki})]. \tag{22}$$

Denoting

$$a_k = \frac{f(y_k) \prod_{i=1}^{d} w_{ki}}{n \prod_{i=1}^{d} \Omega_i}$$

come to the representation

$$\frac{1}{\prod_{i=1}^{d} \Omega_i} \int_{I^d\times\Omega^d} f(y) h_x(y, w)\, dy\, dw \sim \sum_{k=1}^{n} a_k \prod_{i=1} g[w_{ki}(x_i - y_{ki})]$$

that, combined with (12), completes the proof.

The proof of Theorem 2 does not contain any new ideas compared with Theorem 1, but rather underlines the general principle: making limit-integral representation of the function to be approximated, we need to construct the kernel of the integrand in such a way that it concentrates a function to be approximated at the points or surfaces of its domain. For example, the kernel $h_x(y, w) = \prod_{i=1}^{d} w_i g[w_i(x_i - y_i)]$ concentrates the function f at the points x, while the kernel $\hat{h}_x(y, w) = g(w \cdot x + b)$ concentrates the function at the hyperplanes ("ridges") $w \cdot x + b = \text{const}$. In Theorem 2 we use the kernel $\tilde{h}_x(y, w) = g[w \cdot (x - y) \circ (x - y)]$ which concentrates the function again at the points x, but in a way different from that of $h_x(y, w)$ does. The comparison of these different kernels remains an open, interesting problem.

Now we proceed to the basic ideas of Theorem 3. First we note that $\rho_K(f, f_{\omega_n})$ can be bounded by

$$\rho_K(f, f_{\omega_n}) \le \sup_{x\in I^d} |f(x) - f_{\alpha,\Omega}(x)| + \rho_K(f_{\alpha,\Omega}, f_{\omega_n}) \tag{23}$$

TABLE I
COMPARISON OF LEARNING EFFICIENCIES OF THE FUNCTIONAL-LINK (RVFL) AND BACKPROPAGATION (BP) METHODS

Task	**RVFLN**				**BP**			
	variables	*training patterns*	*time*	*iterations*	*variables*	*training patterns*	*times**	*iterations*
Ellipsometry	7	30,000	17 Hrs (sparc II)	5,000	7	30,000	terminated after 24 Hrs	> 10^4
Character Recognition	5	2,000	6 Hrs (486PC)	250	5	2,000	terminated after 96 Hrs	> 10^6
Chemical Product Formulation	16	2,628	1 Hr (sparc II)	125	16	2,628	terminated after 12 Hrs	> 10^4
Underwater Acoustics	31	385	5 Hrs (486 PC)	1000	31	385	terminated after 50 Hrs	> 10^5

* *Ended because of failure to achieve comparable accuracy*

where

$$f_{\alpha,\Omega}(x) = \int_{W^d} F_{\alpha,\Omega}(\omega) g(\alpha w \cdot x + b)\, d\omega. \tag{24}$$

The first term can be made arbitrarily small by choosing big enough values of the parameters α, Ω. But in this case, the variance

$$\begin{aligned}\mathrm{var}_{\alpha,\Omega} = E \int_K \cdot\, dx \frac{1}{|W^d|} \int_{W^d} \\ \cdot\, [|W^d| F_{\alpha,\Omega}(\omega) g(\alpha w \cdot x + b)]^2\, d\omega - f^2(x)\end{aligned}$$

tends to infinity with α, $\Omega \to \infty$. If we can guarantee that the rate of convergence $\sup_{x \in I^d} |f(x) - f_{\alpha,\Omega}(x)|$ to zero is less than the rate of convergence $\mathrm{var}_{\alpha,\Omega}$ to infinity, then we can obtain an estimate

$$\rho_K(f, f_{\omega_n}) = \frac{o(\sqrt{n})}{\sqrt{n}}.$$

But in this case the restrictions imposed on the function f would be too severe. Instead of this, we restrict support of the activation function g to the interval with length $\beta(n)\Omega$, where $\beta(n) \underset{n\to\infty}{\longrightarrow} 0$. Then we can keep α, Ω and, therefore, $\mathrm{var}_{\alpha,\Omega}$ bounded, to obtain an estimate

$$\rho_K(f, f_{\omega_n}) = O\left(\frac{C}{\sqrt{n}}\right)$$

with C independent of n, that completes the proof.

IV. EXPLANATION OF THE USE IN APPLICATIONS

The RVFL method has been used in several tasks by the present authors and their research collaborators and by other researchers, all with favorable results.

As an example, in principle the thickness of film created by molecular beam epitaxy can be monitored through optical ellipsometry. Given the (complex) refractive index of the substrate and the film thickness, it is possible to calculate the values of the ellipsometry measurements. But even when given the requisite ellipsometry measurements, it is very difficult to obtain accurate estimation of the (complex) refractive index of the deposited film and of the film thickness. The traditional methods of numerical solution of systems of nonlinear equations, such as different versions of the Newton method or secant methods [26], failed because of their localized nature. This task of inversion of a complex functional relationship $(\psi_0, \Delta_0, \psi_1, \Delta_1) \to (n, k, d_0, d_1)$ was carried out using a system of RVFL nets with good results. In this task the input variables are two pairs of ψ and Δ angular measurements and there are four outputs, n and k (real and imaginary parts of the refractive index, respectively) of the film, and two film thicknesses d_0 and d_1. Each of the nets were trained with 10^4 training patterns (constituting a very sparse set of training patterns in four-dimensional (4-D) space) and excellent generalization was achieved. Particularly the refractive index n was evaluated with the error less than 0.1%.

Training for 10^4 training set patterns with a consultation system error of less than 10^{-4} could usually be achieved in about six hours on a SPARC 2 workstation, whereas training was never satisfactorily achieved with backpropagation [27]. The number of basis functions varied in the range 50–200.

Our experience with the use of the RVFL approach in dealing with a number of other tasks confirms our judgment that this approach is indeed of high efficiency and of reasonable accuracy. Some features of those tasks and our experiences with those learning tasks are summarized in Table I.

V. CONCLUSIONS AND RECOMMENDATIONS

We have presented a theoretical justification for the random vector version of the functional-link net; this is presented as a particular case of a more general stochastic approach for adaptive function approximation. We proved that RVFL is an universal approximator and that the rate of convergence to zero of the approximation error is of the order of $O(C/\sqrt{n})$ with C independent of n. Similar results were also proved for networks with functional units in the form $\prod_{i=1}^d g[w_{ki}(x_i - y_{ki})]$ or $g[w_k \cdot (x - y_k) \circ (x - y_k)]$ with random parameters y_k, w_k. Thus we demonstrated that a stochastic approach based on an limit-integral representation of the function to be approximated with subsequent evaluation of the integral by the Monte-Carlo method leads to efficient approximation of multivariate functions.

We used a simple Monte Carlo because, in the general case, we did not want to require any specific information about the function. Of course the availability of such information would allow us to use variance reduction methods. This topic supposed to be the subject of future investigations.

The use of Monte Carlo in neural computing or other adaptive (and nonlinear) function approximation opens new possibilities compared with traditional Monte Carlo for multiple integral calculation. Consider a simple example. Suppose, in evaluating a multiple integral, we try to decrease the variance, making two independent samples each of size n, and then estimate the integral as an average of the estimates made over each sample. Clearly the effect of variance reduction is the same as that achieved by using one sample of size $2n$. Suppose now that we evaluate the function with the RVFL in the same manner, with the final estimate taken as an average of two estimates, with $\tilde{n}$ basis functions in each case, where $\tilde{n}$ is the optimal value of the number of basis functions for a given number of training patterns N. This optimal value, which minimizes the total statistical risk, exists, as was shown by Barron [17] for a definite class of functions to be approximated. Existence of such an optimal value of n is a well-known fact for practitioners, including the authors of this paper. Therefore, it is impossible to decrease the variance by simply increasing $\tilde{n}$, but it is possible to do so by taking the average of two independent estimates with $\tilde{n}$ basis functions each. Why? Because we use two stages of learning instead of one.

Appendix

Proof of Theorem 1: We divide the proof in several steps.

Step 1: Representation (11) is true uniformly on any compact $K \subset I^d$, $K \neq I^d$.

Proof: For any $\delta > 0$ and any compact $K \subset I^d$, $K \neq I^d$ can be found a hypercube $I_\delta^d = [\delta; 1-\delta]^d$ such that $K \subset I_\delta^d \subset I^d$, $I_\delta^d \neq I^d$. Using properties of $h_x(y, w)$ from Section III, we obtain

$$\begin{aligned}\left|\int_{I^d} f(y)h_x(y, w)\,dy - f(x)\right| &\le \left|\int_{I^d} f(y)h_x(y, w)\,dy - \int_{I^d} f(x)h_x(y, w)\,dy\right| \\ &\quad + \int_{R^d\setminus I^d} |f(x)|\,|h_x(y, w)|\,dy \\ &\le \int_{I^d} |f(y) - f(x)|\,|h_x(y, w)|\,dy \\ &\quad + \int_{R^d\setminus I^d} |f(x)|\,|h_x(y, w)|\,dy.\end{aligned}$$

Replacing variable y by $z = w \circ (x - y)$ and denoting

$$\begin{aligned}V^d(x, w) &= [(x_1-1)w_1; x_1w_1] \times \cdots \times [(x_d-1)w_d; x_dw_d],\\ I_\delta^d(w) &= [-\delta w_1; (1-\delta)w_1] \times \cdots \times [-\delta w_d; (1-\delta)w_d],\\ w^{-1} &= (w_1^{-1}, \cdots, w_d^{-1})\end{aligned}$$

we go on the evaluation of the integrals. We have

$$\begin{aligned}\int_{I^d} |f(y) - f(x)|\,|h_x(y, w)|\,dy &= \int_{V^d(x,w)} |f(x - z\circ w^{-1}) - f(x)| \left|\prod_{i=1}^d g(z_i)\right| dz \\ &\le \int_{I_\delta^d(w)} |f(x - z\circ w^{-1}) - f(x)| \left|\prod_{i=1}^d g(z_i)\right| dz\end{aligned}$$

uniformly on $x \in K$. $S(w) = [-\sqrt{w_1}; \sqrt{w_1}] \times \cdots \times [-\sqrt{w_d}; \sqrt{w_d}]$ is contained in $I_\delta^d(w)$ for sufficiently large w. Therefore

$$\begin{aligned}&\int_{I_\delta^d(w)} |f(x - z\circ w^{-1}) - f(x)| \left|\prod_{i=1}^d g(z_i)\right| dz \\ &\le \int_{S(w)} |f(x - z\circ w^{-1}) - f(x)| \left|\prod_{i=1}^d g(z_i)\right| dz \\ &\quad + 2M \int_{R^d/S(w)} \left|\prod_{i=1}^d g(z_i)\right| dz \\ &\le \sup_{|y_i|\le 1/\sqrt{w},\, x\in I^d} |f(x-y) - f(x)| \\ &\quad + 2M \int_{R^d\setminus S(w)} \left|\prod_{i=1}^d g(z_i)\right| \\ &\qquad \text{where} \quad M = \sup_{x\in I^d} f(x).\end{aligned}$$

Both last terms tend to zero when w tends to infinity (we assume that $w \to \infty$ amounts for $w_1 \to \infty, \cdots, w_d \to \infty$), the first term because of continuity of the function f in I^d and the second one because of integrability of $|g|$. We complete the proof, noting that

$$\begin{aligned}&\int_{R^d\setminus I^d} |f(x)|\,|h_x(y, w)|\,dy \\ &\le M \int_{R^d\setminus \Pi_{i=1}^d [-\delta w_i; \delta w_i]} \left|\prod_{i=1}^d g(z_i)\,dz\right| \xrightarrow[w\to\infty]{} 0\end{aligned}$$

and collecting all estimates of the integrals together.

Step 2: Representation (12) is true uniformly on any compact $K \subset I^d$, $K \neq I^d$.

Proof: Applying L'Hospital rule d times we obtain

$$\begin{aligned}&\lim_{\Omega_1\to\infty} \cdots \lim_{\Omega_d\to\infty} \frac{1}{\prod_{i=1}^d \Omega_i} \int_{I^d\times\Omega^d} f(y)h_{x,d}(y, w)\,dy\,dw_{(d)} \\ &= \lim_{\Omega_1\to\infty} \cdots \lim_{\Omega_d\to\infty} \frac{1}{\prod_{i=1}^{d-1} \Omega_i} \int_{I^d\times\Omega^{d-1}} \\ &\quad \cdot f(y)h_{x,\,d-1}(y, w)\Omega_d g[\Omega_d(x-y)]\,dy\,dw_{(d-1)} \\ &= \cdots = \lim_{\Omega_1\to\infty} \cdots \lim_{\Omega_d\to\infty} \int_{I^d} f(y)h_x(y, \Omega)\,dy = f(x)\end{aligned}$$

where we temporarily denote $h_{x,d}(y, w) = \prod_{i=1}^d w_i g[w_i(x_i - y_i)]$, $dw_{(d)} = \prod_{i=1}^d dw_i$.

Step 3: Representation (16) is true uniformly on any compact $K \subset I^d$, $K \neq I^d$. Proof is given in Section III.

Step 4: The following estimate of the error of the Monte-Carlo approximation holds

$$E \int_K dx \left[\int_{W^d} F_{\alpha,\Omega}(\omega)g(\alpha w\cdot x + b)\,d\omega - \frac{4\Omega^d}{n} \sum_{k=1}^n F_{\alpha,\Omega}(\omega_k)g(\alpha w_k\cdot x + b_k)\right]^2 \xrightarrow[n\to\infty]{} 0. \quad (25)$$

Proof: Denoting the left side of (25) as ε_{MC}^2 and using Fubini's theorem, we obtain

$$\begin{aligned}\varepsilon_{MC}^2 &= \int_K dx\, E\Bigg[\bigg[\int_{W^d} F_{\alpha,\Omega}(\omega) g(\alpha w\cdot x + b)\, d\omega \\ &\quad - \frac{4\Omega^d}{n}\sum_{k=1}^{n} F_{\alpha,\Omega}(\omega_k) g(\alpha w_k\cdot x + b_k)\bigg]^2\Bigg] \\ &= \frac{1}{n}\int_K dx\, E\bigg[\int_{W^d} F_{\alpha,\Omega}(\omega) g(\alpha w\cdot x + b)\, d\omega \\ &\quad - 4\Omega^d F_{\alpha,\Omega}(\omega) g(\alpha w\cdot x + b)\bigg]^2 \\ &= \frac{C_{f,g,\Omega,\alpha,d}}{n}. \end{aligned} \tag{26}$$

Therefore

$$\varepsilon_{MC} \underset{n\to\infty}{\longrightarrow} 0$$

that completes the proof.

Step 5: Combining (16) and (25), we complete the proof of Theorem 1.

Proof of Theorem 2: Using the kernel $\tilde{h}_x(y, w) = g[w\cdot(x-y)\circ(x-y)]\prod_{i=1}^{d}\sqrt{w_i}$, normalized so that

$$\int_{R^d} g\left(\sum_{i=1}^{d} z_i^2\right) dz = 1$$

instead of the kernel $h_x(y, w)$ and repeating the argument of Step 1 and Step 2 of the previous proof, we obtain the limit-integral representation

$$\begin{aligned} f(x) = \lim_{\Omega_1\to\infty}\cdots\lim_{\Omega_d\to\infty} \frac{1}{\prod_{i=1}^{d}\Omega_i}\int_{I^d\times\Omega^d} \\ \cdot f(y)\tilde{h}_x(y, w)\, dy\, dw. \end{aligned} \tag{27}$$

Estimating the integral in (27) with the Monte-Carlo method, as we did in Step 4, and collecting both estimate of approximation

$$f(x) \sim \frac{1}{\prod_{i=1}^{d}\Omega_i}\int_{I^d\times\Omega^d} f(y)\tilde{h}_x(y, w)\, dy\, dw$$

and the estimate of approximation

$$\begin{aligned}&\frac{1}{\prod_{i=1}^{d}\Omega_i}\int_{I^d\times\Omega^d} f(y)\tilde{h}_x(y, w)\, dy\, dw \\ &\sim \sum_{k=1}^{d} a_k g\{w_k\cdot[(x-y_k)\circ(x-y_k)]\}\end{aligned}$$

we complete the proof.

Proof of Theorem 3: The proof of the theorem is based on the simple lemma.

Lemma: Let g_β be an activation function with support on $\prod_{i=1}^{d}[-\beta w_i; \beta w_i]$, where $\beta, w_1, \cdots, w_d$ are any positive numbers, $h_{x,\beta}(y, w)$ is defined by the formula

$$h_{x,\beta}(y, w) = \prod_{i=1}^{d} w_i g_\beta[w_i(x_i - y_i)].$$

Then uniformly on any compact $K \subset I^d$, $K \neq I^d$

$$f(x) = \lim_{\beta\to 0}\int_{I^d} f(y) h_{x,\beta}(y, w)\, dy.$$

Proof: Indeed we have

$$\begin{aligned}&\left|\int_{I^d} f(y) h_{x,\beta}(y, w)\, dy - f(x)\right| \\ &\leq \int_{supp\, g_\beta} |f(x - z\circ w^{-1}) - f(x)| \left|\prod_{i=1}^{d} g_\beta(z_i)\right| dz \\ &\leq \sup_{z\in supp\, g_\beta} |f(x - z\circ w^{-1}) - f(x)| \\ &\leq \kappa \sup_{z\in supp\, g_\beta} |z\circ w| \leq \kappa\beta\end{aligned}$$

where $supp\, g_\beta$ is support of g_β.

Thus we can approximate $f(x)$ by $\int_{I^d} f(y) h_{x,\beta}(y, w)\, dy$ with bounded w, and, therefore, approximate $f(x)$ by $\int_{W^d} F_{\alpha,\Omega}(\omega) g(\alpha w\cdot x + b)\, d\omega$ with bounded α and Ω. Then from (26) we obtain that $C_{f,g,\Omega,\alpha,d}$ is bounded, completing the proof.

REFERENCES

[1] Y.-H. Pao, *Adaptive Pattern Recognition and Neural Networks.* Reading, MA: Addison-Wesley, 1989.

[2] Y.-H. Pao and Y. Takefuji, "Functional-link computing: Theory, system architecture, and functionalities," *Comput. Mag.*, vol. 3, pp. 76–79, 1991.

[3] Y.-H. Pao, S. Phillips, and D. J. Sobajic, "Neural-net computing and the intelligent control of systems," *Int. J. Contr.*, vol. 56, pp. 263–290, 1992.

[4] Y.-H. Pao, G. H. Park, and D. J. Sobajic, "Learning and generalization characteristics of the random vector functional-link net," *Neurocomputing,* vol. 6, no. 2, pp. 163–180, 1994.

[5] J. M. Hammersley and D. C. Handscomb, *Monte Carlo Method.* London: Methuen, 1964.

[6] A. H. Stroud, *Approximate Calculation of Multiple Integrals.* Englewood Cliffs, NJ: Prentice-Hall, 1971.

[7] R. Fletcher and C. M. Reeves, "Function minimization by conjugate gradients," *Comput. J.*, vol. 7, pp. 149–154, 1964.

[8] W. Rudin, *Principles of Mathematical Analysis.* New York: McGraw-Hill, 1966.

[9] K. Funahashi, "On the approximate realization of continuous mappings by neural networks," *Neural Networks*, vol. 2, pp. 183–192, 1989.

[10] G. Cybenco, "Approximation by superposition of a sigmoidal function," *Math. Contr., Signals, Syst.*, vol. 2, pp. 303–314, 1989.

[11] K. Hornik, M. Stinchcombe, and H. White, "Multilayer feedforward networks are universal approximators," *Neural Networks,* vol. 2, pp. 359–366, 1989.

[12] K. Hornik, "Approximation capabilities of multilayer perceptrons," *Neural Networks,* vol. 4, pp. 251–257, 1991.

[13] ——, "Some new results on neural network approximation," *Neural Networks,* vol. 6, pp. 1069–1072, 1993.

[14] M. Leshno, V. Ya. Lin, A. Pinkus, and S. Schocken, "Multilayer feedforward networks with nonpolynomial activation function can approximate any function," *Neural Networks,* vol. 6, pp. 861–867, 1993.

[15] L. K. Jones, "A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training," *Annals Statist.,* vol. 20, pp. 608–613, Mar. 1992.

[16] A. R. Barron, "Universal approximation bounds for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory,* vol. 39, pp. 930–945, 1993.

[17] ——, "Approximation and estimation bounds for artificial neural networks," *Machine Learning,* vol. 14, pp. 115–133, 1994.

[18] A. R. Barron and J. Thomas, "Minimum complexity density estimation," *IEEE Trans. Inform. Theory,* vol. 37, pp. 1034–1054, 1991.

[19] D. Haussler, "Decision theoretic generalization of the PAAC model for neural net and other learning applications," *Inform. Computa.,* vol. 100, pp. 78–150, 1992.

[20] H. N. Mhaskar and C. A. Michelli, "Approximation by superposition of a sigmoidal function," *Advances Applied Probability,* vol. 13, pp. 350–373, 1992.

[21] B. Igelnik and Y.-H. Pao, "Additional perspectives on feedforward neural nets and the functional-link net," in *Proc. IJCNN'93,* Nagoya, Japan, 1993, vol. 3, pp. 2284–2287.

[22] ——, "Random vector version of the functional-link net," in *Proc. 28th Annu. Conf. Inform. Sci. Syst.,* Princeton, NJ, 1994, vol. 2, pp. 976–980.

[23] Y.-H. Pao and B. Igelnik, "Mathematical concepts underlying the functional-link approach," in *Proc. World Congr. Neural Networks,* San Diego, CA, 1994, vol. 1, pp. 236–246.

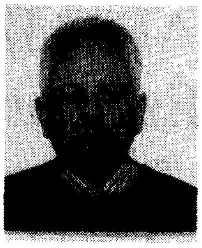
[24] R. A. Adams, *Sobolev Spaces.* New York: Academic, 1975.

[25] P. J. Davis and P. Rabinovitz, *Methods of Numerical Integration.* New York: Academic, 1975.

[26] J. M. Ortega and W. C. Rheinboldt, *Iterative Solution of Nonlinear Equations in Several Variables.* New York: Academic, 1960.

[27] D. E. Rumelhart, G. E. Hinton, and R. J. Williams, "Learning internal representations by error propagation," in *Parallel Distributed Processing: Explorations in the Microstructure of Cognition,* D. E. Rumelhart and J. L. McClelland, Eds. Cambridge, MA: MIT Press, 1986, pp. 318–362.

[28] M. Pincus, "A closed form solution of certain types of constrained optimization problems," *Operations Res.,* vol. 16, pp. 690–694, 1968.



Boris Igelnik received the M.S. degree in the Department of Radio-Telecommunications and Broadcasting in 1962 and the Ph.D. degree in the Department of Automatic Electrical Telecommunications in 1969 from Moscow Institute of Telecommunications, Moscow, USSR. He also received the M.S. degree in the Department of Mathematics and Mechanics in 1971 from Moscow State University, Moscow, USSR.

He is currently with Department of Electrical Engineering and Applied Physics, Case Western Reserve University, Cleveland, OH and Technical Management Concepts, Inc., Dayton, OH. His current research interests include neural-net computing, fuzzy systems, stochastic guided search, optimization, performance analysis of computer-communication networks, and stochastic fluid models for ATM multiplexing.

Yoh-Han Pao (SM'70–F'78) received the B.S. degree from the Henry Lester Institute for Technical Education, Shanghai, China, in 1945, and the Ph.D. degree in applied physics from Pennsylvania State University, State College, in 1952.

He is the Emeritus George S. Dively Distinguished Professor and Director of the Center for Automation and Intelligent Systems Research at Case Western Reserve University, Cleveland, OH, and Emeritus Professor of Electrical Engineering and Computer Science. He also served as the Chairman of Case Western's Electrical Engineering Department (1969–1977), as the Director of the Division of Electrical, Computer and Systems Engineering (1978–1980), National Science Foundation, and as a Visiting Professor at MIT's AI Laboratory, Cambridge, MA (1980). His research interests are adaptive pattern recognition, neural networks, computational intelligence, and signal and image processing.

Dr. Pao is on the editorial board of several international journals. He is also President of AI Ware, Inc.