Pergamon

Neural Networks, Vol. 7, No. 4, pp. 609–628, 1994

0893-6080/94 $6.00 + .00

0893-6080(94)E0017-F

CONTRIBUTED ARTICLE

On Radial Basis Function Nets and Kernel Regression: Statistical Consistency, Convergence Rates, and Receptive Field Size

LEI XU,[1,2] ADAM KRZYŻAK,[3] AND ALAN YUILLE[1]

[1]Harvard University, [2]Peking University, and [3]Concordia University

(*Received* 28 *August* 1992; *revised and accepted* 12 *July* 1993)

Abstract—*Useful connections between radial basis function (RBF) nets and kernel regression estimators (KRE) are established. By using existing theoretical results obtained for KRE as tools, we obtain a number of interesting theoretical results for RBF nets. Upper bounds are presented for convergence rates of the approximation error with respect to the number of hidden units. The existence of a consistent estimator for RBF nets is proven constructively. Upper bounds are also provided for the pointwise and L_2 convergence rates of the best consistent estimator for RBF nets as the numbers of both the samples and the hidden units tend to infinity. Moreover, the problem of selecting the appropriate size of the receptive field of the radial basis function is theoretically investigated and the way this selection is influenced by various factors is elaborated. In addition, some results are also given for the convergence of the empirical error obtained by the least squares estimator for RBF nets.*

Keywords—Radial basis function networks, Kernel regression estimator, Universal approximation, Statistical consistency, Convergence rate, Receptive field size, Parzen window estimator.

1. INTRODUCTION

After several years' extensive study of multilayer perceptrons, many researchers have turned their attention to a number of other neural network models. Among these models, radial basis function (RBF) networks are perhaps the ones that have been studied most intensively (Poggio & Girosi, 1989; Moody & Darken, 1989; Broomhead & Lowe, 1988; Chen, Cowan, & Grant, 1991; Girosi & Poggio, 1989; Renals & Rohwer, 1989; Kraaijveld & Duin, 1991; Nowlan, 1990; Stokbro, Umberger, & Hertz, 1990; Xu, Krzyżak, & Oja, 1992; Xu, Klasa, & Yuille, 1992; Platt, 1991; Weymaere & Martens, 1991; Kardirkamanathan, Niranjan, & Fallside, 1991; Botros & Atkeson, 1991). There has not only been a lot of work on applications but also several theoretical results have been obtained. It has been shown that the RBF net can be naturally derived from the *regularization theory*, that is, the least squares fitting subject to a constraint term imposed by a differential operator (Poggio & Girosi, 1989; Yuille & Grzywacz, 1989). Like the multilayer perceptron, RBF nets have also been shown to have the universal approximation ability (Hartman, Keeler, & Kowalski, 1989; Park & Sandberg, 1991, 1993). Furthermore, RBF nets also have the so-called best approximation ability (Girosi & Poggio, 1989).[1] In contrast, Girosi & Poggio (1989) showed that multilayer perceptrons do not have the best approximation property for the class of continuous functions defined on a subset of R^d. In addition, RBF nets can be related to *Parzen window* estimators of probability density (it can be considered a special example of an RBF net) and *probabilistic neural networks* (Specht, 1990) that directly use the Parzen window estimator for estimating the class densities and then uses these estimators for classification by the Bayesian decision rule. It has long been known that, for any smooth density function, the Parzen window estimator

Acknowledgements: We would like to acknowledge support from DARPA and the Air Force with contracts AFOSR-89-0506 and F49620-92-J-0466, and from NSERC grant A0270, and from FCAR Grant EQ-2904. We also thank the Brown, Harvard, and MIT Center for Intelligent Control Systems for a United States Army Research Office grant number DAAL03-86-C-0171. The second author thanks the Alexander von Humboldt Foundation for support.

Requests for reprints should be sent to Dr. Lei Xu, Department of Computer Science, The Chinese University of Hong Kong, Shatin, NT, Hong Kong.

[1] An approximation scheme has the best approximation property if in the set $\mathcal{F}$ of approximating functions there is one that has minimum approximating error for any function to be approximated from a given set of functions.

is consistent in the quadratic sense (Specht, 1990), that is, the expected mean square error tends to zero as the number of the windows used tends to infinity.

In this paper, we establish connections between RBF nets and the *kernel regression estimator* (KRE), which is an extension of the Parzen window estimator from density estimation to statistical regression problems. We argue that KRE, which includes the Parzen window estimator as a special case, can also be regarded as a particular kind of an RBF net. By using the theoretical results obtained about KRE as tools, we get a number of interesting theoretical results for RBF nets. First, upper bounds are presented for the pointwise and L_2 convergence rates of the approximation error with respect to the number n of basis functions (i.e., hidden units); an example of such bounds is $O(n^{-2\alpha/(2\alpha+d)})$ for the L_2 convergence rate for approximating a function $f(x)$ in the function class that satisfies Lipschitz condition of order α, $0 < \alpha \leq 1$, or $O(n^{-2q/(2q+d)})$ for the L_2 convergence rate for approximating a function $f(x)$ in the class of functions that have order-q ($q \geq 1$) derivatives that are square integrable, where d is the dimensionality of x. Second, the learnability of RBF nets is proved by showing the existence of a consistent estimator for RBF nets constructively. Third, upper bounds are also provided for the pointwise and L_2 convergence rates of the best consistent estimator for RBF nets as n and N (the number of the learning samples, $N \geq n$) tend to ∞. An example of such bounds are $O(n^{-2\alpha/(2\alpha+d)})$, $N \geq n$ or $O(n^{-2q/(2q+d)})$, $N \geq n$ for the L_2 convergence rates for the two function classes described above. Fourth, the problem of selecting the appropriate size of the receptive field of the radial basis function is investigated theoretically and the ways in which some influential factors impact on this selection are qualitatively elaborated. In addition, some results are also given for the convergence of the empirical error obtained by the least squares estimator for RBF nets. We believe that these results are important both theoretically and practically, especially because no papers on these aspects of RBF nets have been published, to our best knowledge, in the current literature of neural networks. (While revising this paper we learned of technical reports by Girosi & Anzellotti, 1992, and Corradit & White, 1992, who have also got some results for the convergence rates of the approximation error. Their studies are significantly different from ours and we will discuss them in Section 3.)

Section 2 explores the connections between KRE and RBF nets. Section 3 gives heuristic descriptions of the main results of this paper. Section 4 describes theorems for various types of convergences and their rates. Section 5 is about the selection of the receptive fields of radial basis functions. The paper is concluded in Section 6. For clarity, the proofs of all the lemmas and theorems are placed in the appendix.

2. KRE, RBF NETS AND THEIR CONNECTIONS

2.1. RBF Networks

Many types of RBF nets can be summarized by the following general form:

$$f_n(x) = \sum_{i=1}^{n} w_i \phi([x - c_i]'\Sigma^{-1}[x - c_i]) \qquad (1)$$

where $\phi(r^2)$ is a prespecified basis function satisfying certain weak conditions. The most common choice is the Gaussian function, $\phi(r^2) = e^{-r^2}$ with $\Sigma = \sigma(n)^2 I$, but a number of alternatives can also be used, (e.g., several choices are listed in Poggio & Girosi, 1989). c_i is called the center vector and $w_i \in R^m$ is a weight vector. Σ is a $d \times d$ positive matrix that controls the receptive field of the basis functions $\phi([x - c_i]'\Sigma^{-1}[x - c_i])$.

The receptive field is defined as the support of the function $\Phi(x) = \phi([x - c_i]'\Sigma^{-1}[x - c_i]) - a_c$ with $a_c \geq 0$ being a constant. In other words, the receptive field is the subset of the domain of x such that $\phi([x - c_i]'\Sigma^{-1}[x - c_i])$ takes values larger than a previously specified number a_c. That is, the receptive field is the range for which an input x can cause a sufficiently large output. We will usually have different receptive fields for different kinds of basis functions $\phi(r^2)$. For a specific $\phi(r^2)$, that is, a Gaussian $\phi(r^2) = e^{-r^2}$, the size, shape, and orientation of the receptive field are determined by the matrix Σ. When $\Sigma = \sigma(n)^2 I$, the shape is a hyperspherical ball with its size (i.e., radius) given by the value of $\sigma(n)$. When $\Sigma = \text{diag}[\sigma(n)_1^2, \ldots, \sigma(n)_d^2]$, the shape is an elliptic ball with each axis coinciding with a coordinate axis, and the length of each axis being decided by $\sigma(n)_1, \ldots, \sigma(n)_d$, respectively. When Σ is a nondiagonal matrix, we have $\Sigma = R^T D R$ with D being a diagonal matrix that determines the shape and size of the receptive field, and with R being a rotation matrix that determines the orientation of the receptive field.

The model eqn (1) has also been further modified into the following normalized version that has often been used recently (Moody & Darken, 1989; Nowlan, 1990; Jones et al., 1991):

$$f_n(x) = \frac{\sum_{i=1}^{n} w_i \phi([x - c_i]'\Sigma^{-1}[x - c_i])}{\sum_{i=1}^{n} \phi([x - c_i]'\Sigma^{-1}[x - c_i])}, \qquad (2)$$

which reduces back to eqn (1) when $\sum_{i=1}^{n} \phi([x - c_i]'\Sigma^{-1}[x - c_i]) = 1$. In this paper, we will concentrate on this normalized model.

For a given fixed $\phi(r^2)$, in eqn (2) there are three sets of parameters: (i) the w_i, $i = 1, \ldots, n$, which are the weight vectors of the output layer of a RBF net, (ii) the center vectors c_i, $i = 1, \ldots, n$, and (iii) the matrix Σ. The last two sets constitute the weights of the hidden

layer of a RBF net. For convenience, we use Θ to denote the vector consisting of all these parameters. Each specific value of Θ specifies a function $f_n(x)$ in the set $\mathcal{F}_n$ of functions defined by eqn (2). In this case, we have a *specified* RBF net.

The problem of determining a specific value $\hat{\Theta}$ for Θ is called *learning* or *training*. In the literature of neural networks, the learning problem is solved based on a given sample set $\mathcal{D}_N = \{X_i, Y_i\}_1^N$. Usually, a value $\hat{\Theta}$ [and thus an $\hat{f}_{n,N}(x)$, which depends on $\mathcal{D}_N$] is decided by the following minimization:

$$\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N}) = \min_{\{f_{n,N} \in \mathcal{F}_n\}} \varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, f_{n,N}) = \min_{\Theta} \varepsilon^2_{\mathrm{RBF}}[\mathcal{D}_N, f_{n,N}(x, \Theta)],$$

$$\varepsilon^2_{\mathrm{RBF}}[\mathcal{D}_N, f_{n,N}(x, \Theta)] = \frac{1}{N} \sum_{i=1}^{N} |Y_i - f_{n,N}(X_i, \Theta)|^2 \quad (3)$$

where $|z|^2 = \sum_{i=1}^m |z^{(i)}|^2$ for a vector $z = [z^{(1)}, \ldots, z^{(m)}]^T$.

However, the minimization with respect to all the parameters simultaneously is usually a hard problem because the minimization with respect to Σ and c_i, $i = 1, \ldots n$ will lead to a problem of nonlinear optimization. Although it is theoretically possible to find a solution (maybe a local minimum) by using a gradient descent method, the iterative method has rarely been used due to its low efficiency. Instead, the existing studies (Broomhead & Lowe, 1988; Chen et al., 1991; Moody & Darken, 1989; Poggio & Girosi, 1989; Powell, 1987) usually assume that Σ takes some externally prespecified values such as $\Sigma = I$, $\Sigma = \sigma(n)^2 I$, or $\Sigma = \mathrm{diag}[\sigma(n)_1^2, \ldots, \sigma(n)_d^2]$ with known $\sigma(n)^2$, $\sigma(n)_1^2$, $\ldots$, $\sigma(n)_d^2$, and that the c_i, $i = 1, \ldots, n$ are determined directly based on the samples $\mathcal{D}_c^x = \{X_i\}_1^N$. Under such assumptions, the minimization of $\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, f_{n,N})$ can be simplified considerably because it is now made only with respect to w_i, $i = 1, \ldots, n$. The minimization of eqn (3) becomes linear with respect to w_i, $i = 1, \ldots, n$, and leads to a set of linear equations that can be solved by the least squares method, with the solution given by:

$$W = YM^T(MM^T)^{-1} \quad (4)$$

where $W = [w_1, \ldots w_n]$ is a $m \times n$ matrix, $Y = [Y_1, \ldots, Y_N]$ is a $m \times N$ matrix, and

$$M = [m_{ij}]_{n \times N}, \quad m_{ij} = \frac{\phi_{ij}}{\sum_{i=1}^n \phi_{ij}}, \quad \phi_{ij} = \phi([X_j - c_i]'\Sigma^{-1}[X_j - c_i]). \quad (5)$$

For determining c_i, $i = 1, \ldots, n$ from samples $\mathcal{D}_c^x = \{X_i\}_1^N$, there are two commonly used methods. In the first method, a clustering algorithm is used to let $\mathcal{D}_c^x$ be partitioned into n clusters, and the mean vectors of these clusters are used as center vectors c_i, $i = 1, \ldots$ n. The second way is much simpler:[2] a n-element subset is randomly selected from $\mathcal{D}_c^x$ and every selected sample is directly used as a center vector, that is, $c_i = X_i$, $i = 1, \ldots, n$. In this case, eqn (2) becomes

$$f_n(x) = \frac{\sum_{i=1}^n w_i \phi([x - X_i]'\Sigma^{-1}[x - X_i])}{\sum_{i=1}^n \phi([x - X_i]'\Sigma^{-1}[x - X_i])}. \quad (6)$$

For convenience, we call the specified RBF nets obtained by the minimization of all the parameters as ones of the *ideal type* or type-0; we call the specified RBF nets given by eqn (6) ones of the *basic type* or type-I, and we call those RBF nets with their center vectors determined by some clustering algorithm ones of the *clustering-aided type* or type-II. Both type-I and type-II nets have been used widely in the literature (Poggio & Girosi, 1989; Moody & Darken, 1989; Broomhead & Lowe, 1988; Chen et al., 1991; Girosi & Poggio, 1989; Renals & Rohwer, 1989; Kraaijveld & Duin, 1991; Nowlan, 1990; Stokbro et al., 1990; Xu et al., 1992; Platt, 1991; Botros & Atkeson, 1991; Weymaere & Martens, 1991; Powell, 1987).

2.2. KRE and Its Connections to RBF Nets

Let (X, Y) be a pair of random vectors in $R^d \times R^m$ and $f(x) = E\{Y|X = x\}$ be the corresponding regression function. Let μ denote the probability measure of X. Moreover, let $\mathcal{D}_n^g = \{X_i', Y_i'\}_1^n$ be a set of independent identically distributed samples drawn from (X, Y). The kernel regression estimate of $f(x)$ is defined as follows:

$$g_n(x) = g_n(x, \mathcal{D}_n^g) = \frac{\sum_{i=1}^n Y_i' K\left(\frac{x - X_i'}{h_n}\right)}{\sum_{i=1}^n K\left(\frac{x - X_i'}{h_n}\right)}, \quad (7)$$

which is the weighted average of Y_i' for approximating the conditional mean of Y under a given $X = x$ with weights depending nonlinearly on the $X_i's$. Here, h_n is usually called a *smoothness parameter* and is a positive number that depends on the number of samples n. $K \geq 0$ is a μ integrable kernel on R^d. The following condition will be imposed on K in a number of theorems proposed in subsequent sections:

$$c_1 H(\|x\|) \leq K(x) \leq c_2 H(\|x\|) \text{ and } cI_{\{\|x\| \leq r\}} \leq K(x) \quad (8)$$

where H is a nonincreasing bounded function with $t^d H(t) \to 0$ as $t \to \infty$ and c_1, c_2, c, r are positive constants. I is the indicator function.

---

[2] Any one of the $C_N^n = N!/n!(N-n)!$ n-sample subsets of $\{X_i, i = 1, \ldots N\}$ is chosen at random. However, without losing generality, we assume that the subset just consists of the first n samples of $\mathcal{D}_c^x$, because if this is not the case, we can reorder the indices of $\mathcal{D}_c^x$ to let it be true because these indices are originally specified arbitrarily.

The estimator (7) is closely related to the Parzen window estimator,

$$p_n(x) = p_n(x, \mathcal{D}_n^g) = \frac{1}{nh_n^d} \sum_{i=1}^{n} K\left(\frac{x - X_i'}{h_n}\right), \quad (9)$$

which is used to approximate a density function $p(x)$ on R^d. The so-called probabilistic neural network proposed in Specht (1990) is one possible direct extension of eqn (9): it uses eqn (9) to approximate the density function of each pattern class and uses these estimates to design a classifier based on the Bayesian decision rule.

Now let us explore the connections between the estimator (7) and various types of RBF nets introduced in the previous subsection.

First, let us start from observing the KRE equation, eqn (7). According to Krzyżak (1986), we know that the condition (8) is nearly as strong as assuming spherical symmetry of the kernel $K(x)$. Thus, we can rewrite eqn (7) as:

$$g_n(x) = \frac{\sum_{i=1}^{n} K\left(\frac{\|x - X_i'\|}{h_n}\right) Y_i'}{\sum_{i=1}^{n} K\left(\frac{\|x - X_i'\|}{h_n}\right)}. \quad (10)$$

Furthermore, we let

$$K(r^2) = \phi(r^2), \quad \Sigma = h_n^2 I, \quad \sigma(n)^2 = h_n^2,$$
$$\text{and} \quad w_i = Y_i' = Y_i, \quad X_i' = X_i, \quad i = 1, \ldots, n. \quad (11)$$

We see that eqn (10) is identical to eqn (6). That is, a spherically symmetrical kernel $K(r^2)$ is just a type of radial basis function where the smoothness parameter h_n represents the size of the basis function's receptive field (of a hyperspherical shape) and Y_i' acts as an approximate solution of w_i. Thus, we can consider the kernel regression estimator (7) as a particular case of RBF nets of type-I given by eqn (6), with a hyperspherically shaped receptive field specified by the matrix $\Sigma = h_n^2 I$, and with the weight vectors w_i, $i = 1, \ldots, n$ being given not by eqn (3), but simply assigned to the specified values Y_i', $i = 1, \ldots, n$. It is interesting to notice that the assumption of hyperspherically shaped receptive fields is commonly used in the existing studies of RBF nets (Broomhead & Lowe, 1988; Chen et al., 1991; Moody & Darken, 1989; Poggio & Girosi, 1989).

Second, let us start from observing the RBF net of the basic type given by eqn (6) and observe that, under the assumption of a hyperspherically shaped receptive field given by $\Sigma = h_n^2 I$, we have

$$M = [m_{ij}]_{n \times N}, \quad m_{ij} = \frac{\phi_{ij}}{\sum_{i=1}^{n} \phi_{ij}},$$
$$\phi_{ij} = \phi\left(\left\| \frac{X_j - X_i}{h} \right\|^2 \right). \quad (12)$$

For those commonly used $\phi(r^2)$ (e.g., Gaussians), we have $\phi(r^2) \approx 0$ for $r^2 \gg 0$. If, in addition, we impose the condition $h_n \ll \min\{\|X_j - X_i\|, i \neq j, i, j = 1, \ldots, n\}$, then it approximately holds that

$$m_{ij} = 1, \quad \text{when} \quad i = j;$$
$$= 0, \quad \text{when} \quad i \neq j. \quad (13)$$

It follows from eqn (6) that, approximately,

$$f_n(X_i) \approx w_i, \quad \text{for} \quad i \leq n;$$
$$= 0, \quad \text{for} \quad n < i \leq N,$$

and thus eqn (3) approximately becomes $M = [I_{n \times n} | 0_{n \times (N-n)}]$ and $(MM^T)^{-1} = I$. By putting them into eqn (4), we have

$$W = YM^T, \quad \text{or} \quad w_i = Y_i, \quad i = 1, \ldots, n. \quad (14)$$

Again, we see that the normalized RBF nets of type-I with hyperspherically shaped receptive field given by $\Sigma = h^2 I$ are approximately identical to the KRE given by eqn (7) when $X_i = X_i'$, $Y_i = Y_i'$, $i = 1, \ldots, n$ if the receptive field size h_n is appropriately chosen.

Third, let $\mathcal{D}_N = \{X_i, Y_i\}_1^N$ be the same as in eqn (3) and, in parallel to eqn (3), we denote the empirical error of KRE by

$$\varepsilon_{\text{KRE}}^2(\mathcal{D}_N, g_n) = \frac{1}{N} \sum_{i=1}^{N} |Y_i - g_n(X_i)|^2. \quad (15)$$

Note that $g_n(x)$ given by eqn (10) is specified by a set $\mathcal{D}_n^g$ and that $\mathcal{D}_N$ and $\mathcal{D}_n^g$ do not necessarily contain the same samples. Moreover, as given in eqn (3), $\varepsilon_{\text{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})$ denotes the minimum of $\varepsilon_{\text{RBF}}^2(\mathcal{D}_N, f_{n,N})$ obtained by minimizing all the parameters w_i, c_i, $i = 1, \ldots, n$ and Σ simultaneously, that is, $\varepsilon_{\text{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})$ is the empirical error obtainable by a specified RBF net of type-0 by the least squares estimator. Here, we further let $\varepsilon_{\text{RBF}}^2(\mathcal{D}_N, f_{n,N}^I)$ denote the value of $\varepsilon_{\text{RBF}}^2(\mathcal{D}_N, f_{n,N})$ for a specified RBF nets of type-I by the least squares estimator.

In the sequel, we propose the following lemma, which relates the errors obtained by KRE and RBF nets, respectively.

LEMMA 1. *Let $K(r^2) = \phi(r^2)$. We have:*

(A) $E\{\varepsilon_{\text{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})\} = E\{|Y_i - \hat{f}_{n,N}(X_i)|^2\}$, $E\{\varepsilon_{\text{KRE}}^2(\mathcal{D}_N, g_n) | \mathcal{D}_n^g\} = E\{|Y_i - g_n(X_i)|^2 | \mathcal{D}_n^g\}$, $(X_i, Y_i) \in \mathcal{D}_N$. *When $\mathcal{D}_N$, $\mathcal{D}_n^g$ are independent, we further have* $E\{|Y_i - g_n(X_i)|^2 | \mathcal{D}_n^g\} = e_0^2 + e_{\text{KRE}}^2(f, g_n)$, *where*

$$e_0^2 = E\{|Y_i - f(X_i)|^2\},$$
$$e_{\text{KRE}}^2(f, g_n) = E\{|g_n(X_i) - f(X_i)|^2 | \mathcal{D}_n^g\}$$
$$= \int_U |f(x) - g_n(x)|^2 \, d\mu(x), \quad (16)$$

where U is the support of the measure μ.

(*B*) $\varepsilon^2_{\text{RBF}}(\mathcal{D}_N, \hat{f}_{n,N}) \leq \varepsilon^2_{\text{KRE}}(\mathcal{D}_N, g_n)$; *when* $\mathcal{D}_N$, $\mathcal{D}^g_n$ *are independent, we also have*

$$E\{\varepsilon^2_{\text{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\} \leq e_0^2 + e^2_{\text{KRE}}(f, g_n);$$

(*C*) *Let* ${}^x\mathcal{D}^g_n$, ${}^y\mathcal{D}^g_n$ *denote the sets of X, Y samples in* $\mathcal{D}^g_n$, *respectively, that is,* $\mathcal{D}^g_n = \{X'_i, Y'_i\}_1^n$, ${}^x\mathcal{D}^g_n = \{X'_i\}_1^n$, ${}^y\mathcal{D}^g_n = \{Y'_i\}_1^n$. *When* ${}^x\mathcal{D}^g_n = \mathcal{D}^x_c$, *we have* $\varepsilon^2_{\text{RBF}}(\mathcal{D}_N, \hat{f}_{n,N}) \leq \varepsilon^2_{\text{RBF}}(\mathcal{D}_N, f^I_{n,N}) \leq \varepsilon^2_{\text{KRE}}(\mathcal{D}_N, g_n)$, *with the same receptive field* $\Sigma = h_n^2 I$ *for the specified* RBF *nets of type-I and* KRE; *furthermore, if* $\mathcal{D}_N$, ${}^y\mathcal{D}^g_n$ *are independent conditioning on* ${}^x\mathcal{D}^g_n$, *then*

$$E\{\varepsilon^2_{\text{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\} \leq E\{\varepsilon^2_{\text{RBF}}(\mathcal{D}_N, f^I_{n,N})\} \leq e_0^2 + E\{e^2_{\text{KRE}}(f, g_n)\}.$$

In the following sections, based on the connections discussed above, we will show that a number of existing theoretical results about kernel regression estimators developed in the statistical literature can be brought into the neural networks literature yielding important new results about RBF nets.

3. MAIN RESULTS: HEURISTIC DESCRIPTIONS

3.1. Mathematical Terms

3.1.1. *On Convergences and Rates.* Given a vector-valued function $f(x) = [f^{(1)}(x), \ldots, f^{(m)}(x)]^T$, and a sequence $\{f_n(x)\}_1^\infty$ of either deterministic or random functions, let

$$e_x(f, f_n) = |f(x) - f_n(x)|,$$

$$\Delta_U(f, f_n) = \sup_{x \in U} e_x(f, f_n) = \sup_{x \in U} |f(x) - f_n(x)|,$$

$$\rho^2_U(f, f_n) = \int_U |e_x(f, f_n)|^2 \, dx = \int_U |f(x) - f_n(x)|^2 \, d\mu(x), \tag{17}$$

where $|z(x)| = \sum_{i=1}^m |z^{(i)}(x)|$ for $z(x) = [z^{(1)}(x), \ldots, z^{(m)}(x)]^T$, U is the domain of x, and μ denotes the measure on x. For any $\varepsilon > 0$, (i) if there exists a specific $n_0(x)$ such that for each $x \in U$ we have $e_x(f, f_n) < \varepsilon$ for any $n > n_0(x)$, then f_n is said to *pointwisely* converge to f; (ii) if there exists a specific n_0 such that $\Delta_U(f, f_n) < \varepsilon$ for any $n > n_0$, then f_n is said to *uniformly* converge to f; and (iii) if there exists a specific n such that $\rho^2_U(f, f_n) < \varepsilon$ for any $n > n_0$, then f_n is said to converge in L_2 to f.

When $\{f_n(x)\}_1^\infty$ are random functions, we have three modes of convergence: *in probability, almost surely, and completely.* A random positive sequence ξ_n is said to converge to 0, (i) *in probability* if for every $\varepsilon > 0$, $\lim_{n\to\infty} P[\xi_n > \varepsilon] = 0$; (ii) *almost surely* if $P[\lim_{n\to\infty} \xi_n = 0] = 1$; and (iii) *completely* if $\sum_{n=1}^\infty P\{\xi_n > \varepsilon\} < \infty$, for every $\varepsilon > 0$. Using $e_x(f, f_n)$, $\Delta_U(f, f_n)$, and $\rho^2_U(f, f_n)$ to replace ξ_n, we can use the definitions of *pointwise, uniform, and* L_2 convergences for the three modes, respectively.

Given a random positive sequence ξ_n that tends to zero as $n \to \infty$, for any one of the three types of convergence (i.e., *pointwise, uniform, and* L_2), the convergence rate of ξ_n is said to be of $O[r(n)]$ (*in probability, almost surely,* and *completely,* respectively) if there is an explicit positive function $r(n)$ of n with $r(n) \to 0$ as $n \to \infty$ [e.g., $r(n) = n^{-q}$, $q > 0$] such that $a_n \xi_n / r(n) \to 0$ as $n \to \infty$ (*in probability, almost surely,* and *completely,* respectively) for any sequence of positive numbers $\{a_n\}$ that satisfies $a_n \to 0$ as $n \to \infty$.

3.1.2. *On Approximation Ability and Statistical Consistency.* A *function approximation scheme* is a device of a set $\mathcal{F}$ of functions supported on R^d. Usually, this device consists of a number of components so that the set $\mathcal{F}$ can be characterized by this number (say n), that is, we can denote it by $\mathcal{F}_n$. Examples of such devices are multilayer networks with n hidden sigmoid units and RBF nets with n radial basis functions given by eqns (1) or (2). Let $\mathcal{F}_U = \cup_{n=1}^\infty \mathcal{F}_n$, then the function approximation scheme is said to have the property of *universal approximation* (Hornik, Stinchocombe, & White, 1989) if $\mathcal{F}_U$ is *dense* in the space of the continuous functions $C[U]$ defined on some domain U of R^d; or in other words, if for any continuous function $f(x)$ supported on U, there exists a specific $\hat{f}_n \in \mathcal{F}_n$ such that $\hat{f}_n(x)$ converges to $f(x)$ *uniformly*. Similarly, for any function $f(x)$ of a given function class $\mathcal{F}_c(U)$ supported on U, if there exist a specific $\hat{f}_n \in \mathcal{F}_n$ such that $\hat{f}_n(x)$ converges to $f(x)$ in the L_2 sense, we say that the function approximation scheme has the property of L_2 approximation for the function class $\mathcal{F}_c(U)$.

These properties describe the *approximation ability* of one set of functions to another set of functions. For a given function $f(x)$, the properties *only* say that there exists, in the set $\mathcal{F}_n$ defined by the function approximation scheme, a function that can approximate $f(x)$ well as $n \to \infty$. They say nothing about how to find such a function. Usually, $\mathcal{F}_n$ is characterized by a set Θ of unspecified parameters. Each specified value $\hat{\Theta}$ of Θ determines a $\hat{f}_n(x)$ in $\mathcal{F}_n$. The value $\hat{\Theta}$ [thus $\hat{f}_n(x)$] is obtained based on a set of observed samples $\mathcal{D}_N = \{X_i, Y_i\}_1^N$ of a given function $f(x)$. Usually these observed samples $X_1, Y_1, \ldots, X_N, Y_N$ are identical and independent random variables with $f(x)$ being their regression function, that is, $f(X_i) = E(Y_i | X_i)$. Such a $\hat{f}_n(x)$ is called an estimator of $f(x)$. To explicitly indicate its dependence on $\mathcal{D}_N$, we denote it by $\hat{f}_{n,N}(x)$. Examples of such estimators include KRE, a specified RBF net obtained by eqn (3), as well as the specified RBF nets of type-I and type-II.

Because $\mathcal{D}_N$ are random samples, $\hat{f}_{n,N}(x)$ is also a random variable. Its convergence behavior is described

by a property called *statistica consistency* that describes how the estimator approaches the regression function $f(x) = E(Y|X = x)$ as the number of samples tends to infinity. An estimator $\hat{f}_{n,N}(x)$ is said to be *consistent* pointwisely, uniformly, or in L_2, respectively, if it converges to $f(x)$ pointwisely, uniformly, or in L_2, respectively, as $N \to \infty$ and $n \to \infty$. Each of the three consistency types can again have three different modes: *in probability, almost surely,* and *completely.* For example, we can say that $\hat{f}_{n,N}(x)$ is pointwisely consistent *in probability, almost surely,* or *completely* if $\hat{f}_{n,N}(x)$ converges pointwisely to $f(x)$ *in probability, almost surely,* or *completely,* respectively, as $N \to \infty$ and $n \to \infty$.

The class of functions $f(x)$ we consider in this paper is rather large [e.g., $\mathcal{F}_c(U) = C(U)$] and it cannot be defined using finite number of parameters. The estimation problem considered in this paper is fundamentally nonparametric.

3.2. Main Results

In recent years there have been many attempts to find mathematical justifications for the use of neural nets like multilayer perceptrons and varieties of RBF nets (Hornik et al., 1989; Hornik, 1991; Girosi & Poggio, 1989; Hartman et al., 1989; Park & Sandberg, 1991, 1993). Typical results are that both multilayer perceptrons and RBF nets possess the properties of *universal approximations* and L_2 approximation for continuous functions. Furthermore, they may also possess L_2 approximation for noncontinuous functions that satisfy some weak conditions.

However, these studies are far from complete in describing the convergence behaviors of neural networks. There remain many important open problems to be studied. We will now describe some of these problems and at the same time heuristically describe our main results (the more precise descriptions and the related proofs will be given in the later sections):

3.2.1. *Convergence Rate of Approximation Error.* The properties of *universal approximation* and L_2 approximation only say that for a given function $f(x)$, there exists an $\hat{f}_n(x)$ in $\mathcal{F}_n$ such that $\hat{f}_n(x)$ converges to $f(x)$ *uniformly* or in the L_2 sense. Actually, such a $\hat{f}_n(x)$ may not be unique. There may be a subset $\hat{\mathcal{F}}_n \subset \mathcal{F}_n$ such that for every $\hat{f}_n(x) \in \hat{\mathcal{F}}_n$, $\hat{f}_n(x)$ converges to $f(x)$ *uniformly* or in the L_2 sense. Let $\hat{f}_n^*(x)$ denote the one that approximates $f(x)$ best. That is, we have

$$e_x(f, \hat{f}_n^*) = \min_{\hat{f}_n \in \hat{\mathcal{F}}_n} e_x(f, \hat{f}_n), \quad \Delta_U(f, \hat{f}_n^*) = \min_{\hat{f}_n \in \hat{\mathcal{F}}_n} \Delta_U(f, \hat{f}_n),$$
$$\rho_U^2(f, \hat{f}_n^*) = \min_{\hat{f}_n \in \hat{\mathcal{F}}_n} \rho_U^2(f, \hat{f}_n) \quad (18)$$

where $e_x(f, f_n)$, $\Delta_U(f, f_n)$, $\rho_U^2(f, f_n)$ are defined by eqn (17). It is more important to know the rate at which $\hat{f}_n^*(x)$ converges with respect to the number n of hidden units. We call this rate the *convergence rate on approximation error* of the network. The faster the rate, the better. Recently, some results have been obtained on this issue for multilayer feedforward networks by Barron (1991), but scarcely any results have been reported for RBF networks.[3]

It follows from Section 2.2 that a KRE is a particular specified RBF net and its $g_n(x)$ belongs to the set $\mathcal{F}_n$ that is defined by the RBF net eqn (2). Theorems 1, 2, 3, and 4 in the next section provide the pointwise and L_2 convergence of $g_n(x)$ converges to regression function $f(x)$ *in probability* and *almost surely,* as n tends to ∞. In other words, for a given $f(x)$, we can construct a specific RBF net $f_n \in \mathcal{F}_n$ by simply letting the parameters Θ to assume the values provided by the samples $\{X_i, Y_i\}_1^n$ in the same way as it was done for the KRE eqn (10). As a result, this f_n will converge *pointwisely or in* L_2 to $f(x)$ *in probability* and *almost surely* with the same rates as given by Theorems 1, 2, 3, and 4. Moreover, such a specific f_n may not be the best $\hat{f}_n^*$. So the convergence rates of $\hat{f}_n^*$ will be not worse than the rates provided by Theorems 1, 2, 3, and 4. That is, we can get upper bounds for the convergence rates of $\hat{f}_n^*$. These bounds are described more precisely in Theorems 13, 14, 15, and 16.

It will be instructive to observe some special cases of the results given by Theorems 13, 14, 15, and 16. For a $f(x)$ belonging to the function class $\mathcal{F}_c$ that satisfies a Lipschitz condition of order α [i.e., $|f(x) - f(y)| \le C\|x - y\|^\alpha$, $0 < \alpha \le 1$, for all y in the neighborhood of x] and $\int |f|^s \, d\mu < \infty$, $s > 1$, we obtain in Theorems 13 and 14 the pointwise convergence rate $O(n^{-\alpha/(2\alpha+d)})$, and the L_2 convergence rate $O(n^{-2\alpha/(2\alpha+d)})$. Furthermore, if $f(x)$ is in the class of functions that exist order-q ($q \ge 1$) derivatives that are square integrable, from Theorems 15 and 16 we have the pointwise convergence rate $O(n^{-q/(2q+d)})$, and the L_2 convergence rate $O(n^{-2q/(2q+d)})$. It is interesting to observe that the rates are related to the ratio d/α or d/q. The higher the dimension d is, the slower the rate; the more smooth of the functions in the class $\mathcal{F}_c$ (i.e., the larger the α or q), the faster the rate is. Specifically, for functions of order-q ($q \ge 1$) derivatives that are square integrable and that have a constant ratio d/q,

[3] At about the same time as we finished the earlier version of the present paper—a technical report (Xu, Krzyżak & Yuille, 1992), Girosi and Anzellotti (1992) and Corradit and White (1992) also produced technical reports describing results for RBF net convergence rates of the approximation error. Their studies differ from ours in many aspects. First, they study the unnormalized RBF net eqn (1) instead of the normalized version eqn (2). Second, they use tools totally different from what we use here. Third, their results concern only convergence rate of the approximation error; our results are, as will be shown in the sequel, much broader. Fourth, even for this common case, the conditions assumed and the detailed results are also different, though their rates are consistent with ours.

the convergence rate does not depend on the dimensionality. In other words, for approximating the functions in this particular case, the problem of *curse of dimensionality* can be avoided. Furthermore, for the very smooth function class (q is quite large) the rate will approach the optimal value $O(1/\sqrt{n})$ for the pointwise convergence and $O(1/n)$ for the L_2 convergence.

Theorems 13, 14, 15, and 16 together give several rates under different conditions, which covers quite general cases.

In addition, one can also observe that our results also imply constructive proofs of the properties of *universal approximation* and L_2 approximation by the RBF nets (2).

3.2.2. *Statistical Consistency and Learnability.* The properties of *universal approximation* and L_2 approximation *only* say that there exists a specific $\hat{f}_n(x)$ in $\mathcal{F}_n$ that approximates $f(x)$ well. They say nothing about whether the network can learn such a specific $\hat{f}_n(x)$ from a set of observed random i.i.d. samples $\mathcal{D}_N = \{X_i, Y_i\}_1^N$ with $f(x_i) = E(Y_i \mid X_i = x_i)$. That is, for any $f(x)$ in a function class $\mathcal{F}_c$, can we find a consistent estimator $\hat{f}_{n,N}(x)$ that converges to $f(x)$ as $N \to \infty$ and $n \to \infty$? If so, then we say that the network can *learn* any functions in $\mathcal{F}_c$. This property is called *statistical consistency* or learnability. This issue has been studied by White (1990), Barron (1991), and Geman, Bienenstock, and Doursat (1992) for multilayer feedforward networks. However, to our knowledge, there is no reported study for RBF networks.

In this paper, we have shown that the RBF nets given by eqn (2) do have this property. From Theorems 1 and 2 we know that the KRE estimator $g_n(x)$ given by eqn (2) is both pointwise and L_2 consistent to $f(x)$ *in probability, almost surely, and completely* under very weak conditions on Y (e.g., $E\{|Y|^s\} < \infty$, $s \geq 1$ or $|Y| < M < \infty$ for pointwise consistency and $E\{|Y|^{2+s}\} < \infty$, $s > 0$ for L_2 consistency). It follows from Section 2.2 that we can construct a particular estimator $\hat{f}_{n,N}(x) \in \mathcal{F}_n$ for the RBF net (2) by simply setting $\hat{f}_{n,N}(x) = \hat{f}_{n,n}(x) = g_n(x)$. As both the number of samples and the number of radial basis functions $N = n \to \infty$, this estimator will converge pointwise and in L_2 to $f(x)$ *in probability, almost surely, and completely* under the weak conditions on Y specified above. That is, the RBF net given by eqn (2) has *learnability* or *statistical consistency* properties (Theorems 3 and 4). In addition, KRE has also provided a way to obtain this property.

3.2.3. *Convergence Rate of Estimation Error.* Let us modify eqn (17) slightly into

$$e_x(f, \hat{f}_{n,N}) = |f(x) - \hat{f}_{n,N}(x)|,$$

$$\Delta_U(f, \hat{f}_{n,N}) = \sup_{x \in U} e_x(f, \hat{f}_{n,N}) = \sup_{x \in U} |f(x) - \hat{f}_{n,N}(x)|,$$

$$\rho_U^2(f, \hat{f}_{n,N}) = \int_U |e_x(f, \hat{f}_{n,N})|^2 \, dx = \int_U |f(x) - \hat{f}_{n,N}(x)|^2 \, d\mu(x). \tag{19}$$

These are the estimation errors of the estimator $\hat{f}_{n,N}(x)$. The property of *learnability* only tells us that these errors will converge to zero as $N \to \infty$ and $n \to \infty$, but says nothing about how fast this convergence will be. Because it is not necessary to have a unique consistent estimator $\hat{f}_{n,N}(x)$, we let $\hat{\mathcal{F}}_{n,N}$ denote the subset consisting all these consistent estimators. Let $\hat{f}^*_{n,N}(x) \in \hat{\mathcal{F}}_{n,N}$ denote the one that approximates $f(x)$ best. That is, we have

$$e_x(f, \hat{f}^*_{n,N}) = \min_{\hat{f}_{n,N} \in \hat{\mathcal{F}}_{n,N}} e_x(f, \hat{f}_{n,N}),$$

$$\Delta_U(f, \hat{f}^*_{n,N}) = \min_{\hat{f}_{n,N} \in \hat{\mathcal{F}}_n} \Delta_U(f, \hat{f}_{n,N}),$$

$$\rho_U^2(f, \hat{f}^*_{n,N}) = \min_{\hat{f}_{n,N} \in \hat{\mathcal{F}}_{n,N}} \rho_U^2(f, \hat{f}_{n,N}). \tag{20}$$

More interestingly, we want to know the rate with which $\hat{f}^*_{n,N}(x)$ converges to $f(x)$ as N, $n \to \infty$. At present, there are some results available on this topic for multilayer feedforward networks (Barron, 1991; Lugosi & Zeger, 1993). However, once again there are no results reported on this issue for RBF networks (to our knowledge).

As pointed out a moment ago, we can get a consistent estimator in $\hat{\mathcal{F}}_{n,N}$ simply by letting $\hat{f}_{n,N}(x) = \hat{f}_{n,n}(x) = g_n(x)$. The convergence rates of $g_n(x)$ [and thus of the estimator $\hat{f}_{n,n}(x)$] are provided by Theorems 5, 6, 7, and 8. Furthermore, this specific $\hat{f}_{n,n}(x)$ may not necessarily be the best estimator $\hat{f}^*_{n,N}(x)$. Thus, the convergence rates of $\hat{f}^*_{n,N}(x)$ will not be worse than the rates of $g_n(x)$. In other words, the convergence rates provided by Theorems 5, 6, 7, and 8 for $g_n(x)$ provide upper bounds for the convergence rates of $\hat{f}^*_{n,N}(x)$. In Theorems 9, 10, 11, and 12, several rates are obtained under different conditions for several general cases. To get some flavor, we now informally give a special example of L_2 convergence. For an $f(x)$ belonging to the function class $\mathcal{F}_c$ that satisfies the above-mentioned Lipschitz condition of order α, $0 < \alpha \leq 1$, with the condition that Y is bounded and that the basis function ϕ has compact support, we find that the L_2 convergence rate is $O(n^{-2\alpha/(2\alpha+d)})$, $N \geq n$. Furthermore, if $f(x)$ is in the class of functions that have order-q ($q \geq 1$) derivatives that are square integrable, even when the basis function ϕ does not have compact support, we can also similarly get L_2 convergence rate of $O(n^{-2q/(2q+d)})$, $N \geq n$. The rate depends on the ratio of dimensionality to smoothness also. For a particular function class with a constant value for d/q, there will again be no curse of dimensionality. For a very smooth function class, the rate will approach the optimal value $O(1/n)$.

3.2.4. *Convergence Related to the Least Squares Estimator.* In the neural network literature, the least squares estimator is commonly used for the RBF net given by eqn (2). That is, $\hat{f}_{n,N}(x)$ is obtained by minimization of eqn (3). What properties does such an estimator have? Is it a consistent estimator? What is its rate of convergence? These questions are clearly important. If $\hat{f}_{n,N}(x)$ is the best estimator $\hat{f}^*_{n,N}(x)$, or if it is better than $\hat{f}_{n,n}(x) = g_n(x)$ in the sense that

$$e_x(f, \hat{f}_{n,N}) \le e_x(f, g_n), \quad \Delta_U(f, \hat{f}_{n,N}) \le \Delta_U(f, g_n), \quad \rho^2_U(f, \hat{f}_{n,N}) \le \rho^2_U(f, g_n)$$

then our results can be directly applied. Unfortunately, the answers to these questions remain open and will be dealt in future studies. Instead, it follows from Lemma 1 that $E\{\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\} \le e_0^2 + e^2_{\mathrm{KRE}}(f, g_n)$. Because we know that g_n is consistent with convergence rates given by Theorems 5, 6, 7, and 8, we obtain Theorems 17 and 18, which tell us that as $n \to \infty$ and $N \to \infty$, the expected empirical error of this estimator will reduce with the same rates and finally will drop below $e_0^2 = E\{|Y_i - f(X_i)|^2\}$—an error that is irrelevant to any estimator. Similarly, from $E\{\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\} \le E\{\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, f^I_{n,N})\} \le e_0^2 + E\{e^2_{\mathrm{KRE}}(f, g_n)\}$, we can also know how the expected empirical error of the estimator given by the specified RBF net of type-I reduces as $n \to \infty$. Our results described here for the least squares estimator are only preliminary. To complete this study, we need to know whether the empirical error $\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})$ can converge to its expected value $E\{\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\}$, and, if so, with what rate. Unfortunately the *large number law* does not automatically apply because the terms $|Y_i - f_{n,N}(X_i, \Theta)|^2$ in $\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})$ are not independent. Each term depends on all data in $\mathcal{D}_N$. In addition, we also need to know whether $E\{\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\}$ converges to e_0^2 instead of dropping below it. We leave these two questions out of this paper for further studies and hypothesize that the tool of *Vapnik–Cervonenkis dimensionality* may be useful in studying this issue.

3.2.5. *Selection of the Size of Receptive Field.* In practical applications, the specified RBF nets of type-I or type-II are mostly used because of their simplicity. That is, the size h_n of the receptive field for a RBF net is usually specified externally. Naturally, there rises a question: how can the scale parameter h_n be selected appropriately? At present very little is known about the appropriate selection of this parameter and in practice the selection is usually based on heuristic strategies (Mel & Omohundro, 1991). In Section 5, we will give results on this issue by studying the appropriate selection of the smoothness parameter h for KRE estimators—a particular RBF net of type-I. The key idea is to trade off between the two components of the estimation error $e_x(f, g_n)$ or $\rho^2_U(f, g_n)$ defined in eqn (17). This trade-off is closely related to the well-known trade-off between bias and variance. Theorems 19, 20, and 21 tell us that one component will increase monotonically with h at an order of $O(h^q)$, $q > 0$ pointwisely or $O(h^{2q})$ in the L_2 sense; the other will reduce at the order of $O(1/\sqrt{nh^d})$ pointwise or $O(1/nh^d)$ in L_2, depending on both n and h, where q is a parameter related to the smoothness of the regression function $f(x)$ (e.g., the function has q derivatives), and d is the dimension of x. To minimize $B(x)h_n^q + A(x)/\sqrt{nh_n^d}$ or $Dh_n^{2q} + C/nh_n^d$, we find that the appropriate size should be of the order $O\{[A(x)\ d/2qB(x)\sqrt{n}]^{2/(2q+d)}\}$ for pointwise or $O[(C\,d/2q\,Dn)^{1/(2q+d)}]$ for L_2. Usually, accurate calculation of the coefficients $A(x)$, $B(x)$, C, D is difficult. However, the results here can give us some theoretical understanding and qualitative guide for applications. The details about these coefficients are given in Section 5.

4. STATISTICAL CONSISTENCY AND CONVERGENCE RATES

4.1. Statistical Consistency

Theorems 1 and 2 are about the consistency of the KRE estimator $g_n(x)$ given by eqn (7).

THEOREM 1. (Pointwise consistency, KRE.) *Let K be a nonnegative kernel satisfying condition eqn (8), and H a bounded function nonincreasing in the interval $[0, \infty)$ and $t^d H(t) \to 0$ as $t \to \infty$. Let μ denote the probability measure of X, $f(x) = E\{Y|X = x\}$ denote the regression function.*

(A) Let $E\{|Y|^s\} < \infty$, $s \ge 1$. If $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d \to \infty$, then $g_n(x)$ is pointwisely consistent to $f(x)$ in probability for almost all $x(\mu) \in R^d$.

(B) Let $|Y| < M < \infty$ and $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d/\log n \to \infty$, then $g_n(x)$ is pointwisely consistent to $f(x)$ almost surely for almost all $x(\mu) \in R^d$.

(C) Let $E\{|Y|^s\} < \infty$, $s > 1$ and $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty}(n^{(s-1)/s}h_n^d/\log n) \to \infty$, then $g_n(x)$ is pointwisely consistent to $f(x)$ almost surely for almost all $x(\mu) \in R^d$.

(D) Let $|Y| \le M < \infty$. If $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n^d/\log n \to \infty$, then $g_n(x)$ is pointwisely consistent to $f(x)$ completely for almost all $x(\mu) \in R^d$.

(E) Let $E\{|Y|^s\} < \infty$, $s > 1$ and $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty}(n^{(s-1)/s}h_n^d/\log n) \to \infty$, then $g_n(x)$ is pointwisely consistent to $f(x)$ completely for almost all $x(\mu) \in R^d$.

THEOREM 2. (L_2 consistency, KRE.) *Let $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d \to \infty$. Assume that $K(x) \ge cI_{S_r}$, $S_r = \{x|\,\|x\| \le r\}$, $\int \sup_{y\in x+S_r} K(y)\,dx < +\infty$. Let μ denote the probability measure of X, and $f(x) = E\{Y|X = x\}$ denote the regression function. For all distributions of (X, Y) with $E\{|Y|^{2+s}\} < \infty$, $s > 0$, $g_n(x)$ is L_2 consistent to $f(x)$ almost surely.*

An example that satisfies the assumption $\int \sup_{y \in x+S_r} K(y)\,dx < \infty$ in Theorem 2 is the Riemann integrable monotonically decreasing kernel. In fact, the class of applicable kernels for the above theorems is very large. Not only does it includes various kernels with bounded support, for example,

$$K(r) = I_{\{\|r\| \le 1\}},$$

$$K(r) = (1 - r^2) I_{\{\|r\| \le 1\}}$$

(where $I_{\{\|r\| \le 1\}} = 1$ when $\|r\| \le 1$, and $I_{\{\|r\| \le 1\}} = 0$ otherwise), but it also includes a great many kernels with unbounded support, for example,

$$K(r) = e^{-|r|},$$

$$K(r) = e^{-r^2},$$

$$K(r) = \sin^2(r)/r^2,$$

$$K(r) = 1/(1 + |r|^{1+\delta}), \quad \delta > 0;$$

and even some nonintegrable kernels as well, for example,

$$K(r) = \begin{cases} 1/e & \text{if } |r| \le e \\ \ln(|x|)/|x| & \text{otherwise.} \end{cases}$$

Theorems 3 and 4 are about the consistency of the RBF nets (2).

THEOREM 3. (Pointwise consistency, RBF.) *Let $\mathcal{F}_n$ be the function set defined by the RBF nets* (6). *Let μ denote the probability measure of X, $f(x) = E\{Y|X = x\}$ denote the regression function, and assume that $E\{|Y|^s\} < \infty$, $s \ge 1$. Let $\phi(x)$ be a nonnegative radial basis function satisfying*

$$c_1 H(\|x\|) \le \phi(x) \le c_2 H(\|x\|) \text{ and } cI_{\{\|r\| \le r_0\}} \le \phi(x) \quad (21)$$

where H is a nonincreasing bounded function with $t^d H(t) \to 0$ as $t \to \infty$ and c_1, c_2, c, r_0 are positive constants. I is the indicator function. Given a set of i.i.d. random samples $\mathcal{D}_N = \{X_i, Y_i\}_1^N$, there exists (and also we can construct) an estimator $\hat{f}_{n,N}(x) \in \mathcal{F}_n$ such that as $n \to \infty$, $N \to \infty$, $\hat{f}_{n,N}(x)$ is pointwisely consistent to $f(x)$ in probability, almost surely, and completely, respectively, for almost all $x(\mu) \in R^d$.

THEOREM 4. (L_2 consistency, RBF.) *Let $\mathcal{F}_n$ be the function set defined by the RBF nets* (6). *Let μ denote the probability measure of X, $f(x) = E\{Y|X = x\}$ denote the regression function, and assume that $E\{|Y|^{2+s}\} < \infty$, $s > 0$. Assume that $\phi(x) \ge cI_{S_r}$, $S_r = \{x | \|x\| \le r\}$, $\int \sup_{y \in x+S_r} \phi(y)\,dx < +\infty$. Given a set of i.i.d. random samples $\mathcal{D}_N = \{X_i, Y_i\}_1^N$, there exists (and we can also construct) an estimator $\hat{f}_{n,N}(x) \in \mathcal{F}_n$ such that as $n \to \infty$, $N \to \infty$, $\hat{f}_{n,N}(x)$ is L_2 consistent to $f(x)$ almost surely.*

4.2. Convergence Rates of KRE

THEOREM 5. (Pointwise convergence rate, KRE.) *Let K be a nonnegative kernel satisfying condition* (8), *and $H(t)$ be a bounded function nonincreasing in the interval $[0, \infty)$ with $t^d H(t) \to 0$ as $t \to \infty$. Let μ denote the probability measure of X and $f(x) = E\{Y|X = x\}$ denote the regression function. Assume that $f(x)$ locally satisfies a Lipschitz condition of order α in the neighborhood of x, that is, $|f(x) - f(y)| \le C\|x - y\|^\alpha$, $0 < \alpha \le 1$ for all $y \in S_{x,\rho}$, where $S_{x,\rho}$ is a sphere of the radius ρ centered at x. Let b_n denote the smallest solution of the nonlinear equation*

$$[h_n H^+(b_n h_n^d)]^\alpha = b_n$$

where $H^+(z) = \sup\{t, H(t) > z\}$.

For the $g_n(x)$ *given by eqn* (7), *we have:*

(*A*) *Let $E\{|Y|^s\} < \infty$, $s > 1$. If $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d \to \infty$, then $g_n(x)$ converges pointwisely to $f(x)$ in probability with the rate $O(\max\{\theta_{s,n}, b_n\})$ for almost all $x(\mu) \in R^d$, where $\theta_{s,n} = \min\{(nh_n^d)^{-(s-1)/s}, (nh_n^d)^{-1/2}\}$.*

(*B*) *Let $\operatorname{esssup}_X E\{|Y|^s | X\} < \infty$, $s > 1$. If $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} (n^{(s-1)/s} h_n^d / \log n) \to \infty$, then $g_n(x)$ converges pointwisely to $f(x)$ almost surely with the rate $O(\beta_n \sqrt{\log n})$ for almost all $x(\mu) \in R^d$ where $\beta_n = \max\{(n^{(s-1)/s} h_n^d)^{-1/2}, b_n\}$.*

(*C*) *Let $|Y| \le M < \infty$. If $\lim_{n\to\infty} h_n = 0$ and $\lim_{n\to\infty} nh_n^d / \log n \to \infty$, then $g_n(x)$ converges pointwisely to $f(x)$ almost surely with the rate $O(\beta_n \sqrt{\log n})$ for almost all $x(\mu) \in R^d$, where $\beta_n = \max\{(nh_n^d)^{-1/2}, b_n\}$.*

It has also been shown in Krzyżak (1986) that the assumption that $f(x)$ locally satisfies a Lipschitz condition of order α in the neighborhood of x can be relaxed to include functions that do not require continuity of $f(x)$ almost everywhere. An example of such functions is the Dirichlet function defined on the closed interval [0, 1] as follows:

$$f(x) = \begin{cases} 1 & \text{if } x \text{ rational,} \\ 0 & \text{otherwise.} \end{cases}$$

It is easy to see that $f(x)$ is nowhere continuous.

THEOREM 6. (L_2 convergence rate, KRE.) *Let μ denote the probability measure of X and $f(x) = E\{Y|X = x\}$ denote the regression function. Assume that $f(x)$ satisfies the same conditions as in Theorems 5, and that $K(x) \ge cI_{S_r}$, $S_r = \{x | \|x\| \le r\}$, $\int \sup_{y \in x+S_r} K(y)\,dx < +\infty$ and $\|x\|^d K(x) \to 0$ as $\|x\| \to \infty$. Let b_n denote the smallest solution of the nonlinear equation*

$$[h_n K^+(b_n h_n^d)]^{2\alpha} = b_n$$

where $K^+(z) = \sup\{t, K^*(t) > z\}$, $K^*(t) = \sup_x\{K(x) I_{\{\|x\| > t\}}(x)\}$.

Let $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d \to \infty$, then for the $g_n(x)$ given by eqn (7) *we have:*

(*A*) *If $|Y| \le M < \infty$ (M is a positive constant), then $g_n(x)$ converges in L_2 to $f(x)$ in probability with the rate $O[\max\{(nh^d)^{-1}, b_n\}]$; moreover, if K has compact support, then this rate is $O[n^{-[2\alpha/(2\alpha+d)]}]$.*

(B) *If* $E\{|Y|^s\} < \infty$, $s > 2 + d/\alpha$ *and* K *has compact support, then* $g_n(x)$ *converges in* L_2 *to* $f(x)$ *in probability with the rate* $O[n^{(2/s)-[2\alpha/(2\alpha+d)]}]$.

In particular, when $f(x)$ has q-order continuous derivatives, then we also have:

THEOREM 7. (Pointwise convergence rate, KRE.) *Let* $p(x)$ *denote the probability density of* X. *Assume that* $E\{Y^2\} < \infty$, $p(x) \in C^q$, $f(x) \in C^q$, *and* $\int |f^{(q)}(x)|\, d\mu(x) < \infty$, *where* C^q *is the set of all functions that have q-order continuous derivatives. Also assume that* $K(x)$ *satisfies the following conditions:*

(a) $\int |K|\, dx < \infty$ *and* $\sup|K| < \infty$.

(b) $K(x) = \prod_{i=1}^d H(x^{(i)})$, $H(z)$ *is radially symmetric and* $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Let $h_n \to 0$, $nh_n^d \to \infty$. *Then* $g_n(x)$ *given by eqn* (7) *converges pointwisely to* $f(x)$ *in probability with the rate* $O(n^{-q/2q+d})$ *at all points of continuity of* $s^2(x) = E\{Y^2|X = x\}$.

THEOREM 8. (L_2 convergence rate, KRE.) *Let* $p(x)$ *denote the probability density of* X. *Assume* $|Y| \le M$ *and* $\inf_Q p(x) > 0$. *Assume that* $K(x)$ *satisfies the following conditions:*

(a) $\int |K|\, dx < \infty$ *and* $\sup|K| < \infty$.

(b) $K(x) = \prod_{i=1}^d H(x^{(i)})$, $H(z)$ *is radially symmetric and* $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Let $h_n \to 0$, $nh_n^d \to \infty$. *Then* $g_n(x)$ *given by eqn* (7) *converges in* L_2 *to* $f(x)$ *in probability with the rate* $O(n^{-2q/2q+d})$, *provided that all partial derivatives of* f *and* p *of orders* i, $1 \le i \le p$ *exist and are square integrable.*

4.3. Convergence Rates of RBF Nets

THEOREM 9. (Pointwise convergence rate, RBF nets.) *Let* ϕ *be a nonnegative kernel satisfying condition* (8), *and* $H(t)$ *be a bounded function nonincreasing in the interval* $[0, \infty)$ *with* $t^d H(t) \to 0$ *as* $t \to \infty$. *Let* μ *denote the probability measure of* X, $f(x) = E\{Y|X = x\}$ *denote the regression function which satisfies the same condition as in Theorem* 5, *and* b_n *denote the smallest solution of the nonlinear equation*

$$[h_n H^+(b_n h_n^d)]^\alpha = b_n$$

where $H^+(z) = \sup\{t, H(t) > z\}$.

Let $\hat{\mathcal{F}}_{n,N}$ *denote the set consisting of all the consistent estimators for the RBF nets* (6), *given a sample set* $\mathcal{D}_N$. *Let* $\hat{f}^*_{n,N}(x) \in \hat{\mathcal{F}}_{n,N}$ *denote the one that approximates* $f(x)$ *best in the sense of eqn* (20), *then as* $n \to \infty$, $N \to \infty$ *with* $N \ge n$, *we have:*

(A) *Let* $E\{|Y|^s\} < \infty$, $s > 1$, *then* $\hat{f}^*_{n,N}(x)$ *converges pointwisely to* $f(x)$ *in probability with a rate upper-bounded by* $O(\max\{\theta_{s,n}, b_n\})$ *for almost all* $x(\mu) \in R^d$, *where* $\theta_{s,n} = \min\{(nh_n^d)^{-(s-1)/s}, (nh_n^d)^{-1/2}\}$, and h_n *satisfies* $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d \to \infty$.

(B) *Let* $\operatorname{esssup}_X E\{|Y|^s|X\} < \infty$, $s > 1$, *then* $\hat{f}^*_{n,N}(x)$ *converges pointwisely to* $f(x)$ *almost surely with a rate upper-bounded by* $O(\beta_n\sqrt{\log n})$ *for almost all* $x(\mu) \in R^d$, *where* $\beta_n = \max\{(n^{(s-1)/s}h_n^d)^{-1/2}, b_n\}$, *and* h_n *satisfies* $\lim_{n\to\infty} h_n = 0$ *and* $\lim_{n\to\infty}(n^{(s-1)/s}h_n^d/\log n) \to \infty$.

(C) *Let* $|Y| \le M < \infty$, *then* $\hat{f}^*_{n,N}(x)$ *converges pointwisely to* $f(x)$ *almost surely with a rate upper-bounded by* $O(\beta_n\sqrt{\log n})$ *for almost all* $x(\mu) \in R^d$, *where* $\beta_n = \max\{(nh_n^d)^{-1/2}, b_n\}$, *and* h_n *that* $\lim_{n\to\infty} h_n = 0$ *and* $\lim_{n\to\infty} nh_n^d/\log n \to \infty$.

THEOREM 10. (L_2 convergence rate, RBF nets.) *Let* μ *denote the probability measure of* X *and* $f(x) = E\{Y|X = x\}$ *denote the regression function. Assume that* $f(x)$ *satisfies the same conditions as in Theorem* 5, *and that* $\phi(x) \ge cI_{S_r}$, $S_r = \{x| \|x\| \le r\}$, $\int \sup_{y\in x+S_r} \phi(y)\, dx < +\infty$ *and* $\|x\|^d \phi(x) \to 0$ *as* $\|x\| \to \infty$. *Let* b_n *denote the smallest solution of the nonlinear equation*

$$[h_n \phi^+(b_n h_n^d)]^{2\alpha} = b_n$$

where $\phi^+(z) = \sup\{t, \phi^*(t) > z\}$, $\phi^*(t) = \sup_x\{\phi(x) I_{\{\|x\|>t\}}(x)\}$.

Let $\hat{\mathcal{F}}_{n,N}$ *denote the set consisting of all the consistent estimators for the RBF nets* (6), *given a sample set* $\mathcal{D}_N$. *Let* $\hat{f}^*_{n,N}(x) \in \hat{\mathcal{F}}_{n,N}$ *denote the one that approximates* $f(x)$ *best in the sense of eqn* (20), *then as* $n \to \infty$, $N \to \infty$ *with* $N \ge n$, *we have:*

(A) *If* $|Y| \le M < \infty$ (M *is a positive constant*), *then* $\hat{f}^*_{n,N}(x)$ *converges in* L_2 *to* $f(x)$ *in probability with a rate upper-bounded by* $O(\max\{(nh^d)^{-1}, b_n\})$; *moreover, if* ϕ *has compact support, then this rate is* $O(n^{-[2\alpha/(2\alpha+d)]})$, *where* h_n *satisfies* $\lim_{n\to\infty} h_n = 0$, $lim_{n\to\infty} nh_n^d \to \infty$.

(B) *If* $E\{|Y|^s\} < \infty$, $s > 2 + d/\alpha$ *and* ϕ *has compact support, then* $\hat{f}^*_{n,N}(x)$ *converges in* L_2 *to* $f(x)$ *in probability with a rate upper-bounded by* $O(n^{(2/s)-[2\alpha/(2\alpha+d)]})$.

In particular, when $f(x)$ has q-order of continuous derivatives, then we also have

THEOREM 11. (Pointwise convergence rate, RBF nets.) *Let* $p(x)$ *denote the probability density of* X. *Assume that* $E\{Y^2\} < \infty$, $p(x) \in C^q$, $f(x) \in C^q$, *and* $\int |f^{(q)}(x)|\, d\mu(x) < \infty$, *where* C^q *is the set of all functions that have q-order continuous derivatives. Also assume that* $\phi(x)$ *satisfies the following conditions:*

(a) $\int |\phi|\, dx < \infty$ *and* $\sup|\phi| < \infty$.

(b) $\phi(x) = \prod_{i=1}^d H(x^{(i)})$, $H(z)$ *is radially symmetric and* $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Let $\hat{\mathcal{F}}_{n,N}$ *denote the set consisting of all the consistent estimators for the RBF nets* (6), *given a sample set* $\mathcal{D}_N$. *Let* $\hat{f}^*_{n,N}(x) \in \hat{\mathcal{F}}_{n,N}$ *denote the one that approximates* $f(x)$ *best in the sense of eqn* (20), *then as* $n \to \infty$, $N \to \infty$ *with* $N \ge n$, $\hat{f}^*_{n,N}(x)$ *converges pointwisely to* $f(x)$ *in probability with a rate upper-bounded by*

$O(n^{-q/2q+d})$ *for all points of continuity of* $\sigma^2(x) = E\{Y^2|X = x\}$.

THEOREM 12. (L_2 consistent convergence rate, RBF nets.) *Let* $p(x)$ *denote the probability density of* X. *Assume* $|Y| \le M$ *and* $\inf_Q p(x) > 0$. *Assume that* $\phi(x)$ *satisfies the following conditions:*

(*a*) $\int |\phi|\, dx < \infty$ *and* $\sup|\phi| < \infty$.

(*b*) $\phi(x) = \prod_{i=1}^{d} H(x^{(i)})$, $H(z)$ *is radially symmetric and* $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Let $\hat{\mathcal{F}}_{n,N}$ *denote the set consisting of all the consistent estimators for the RBF nets* (6), *given a sample set* $\mathcal{D}_N$. *Let* $\hat{f}^*_{n,N}(x) \in \hat{\mathcal{F}}_{n,N}$ *denote the one that approximates* $f(x)$ *best in the sense of eqn* (20), *then as* $n \to \infty$, $N \to \infty$ *with* $N \ge n$, $\hat{f}^*_{n,N}(x)$ *converges in* L_2 *to* $f(x)$ *in probability with a rate upper-bounded by* $O(n^{-2q/2q+d})$, *provided that all partial derivatives of* f *and* p *of orders* i, $1 \le i \le p$ *exist and are square integrable.*

4.4. RBF Net Convergence Rates in Approximation

THEOREM 13. (Pointwise convergence rate in approximation, RBF.) *Assume that* $\phi(x)$ *be a nonnegative radial basis function satisfying eqn* (21) *with* H *being a bounded function decreasing in the interval* $[0, \infty)$ *and* $t^d H(t) \to 0$ *as* $t \to \infty$. *Let* $\mathcal{F}_n$ *be the function set defined by the RBF nets* (6), *and* μ *denote the measure on* x. *Also assume that* $f(x)$ *satisfies the same conditions as in Theorem* 5. *Let* $\hat{f}^*_n(x)$ *denote the one in* $\mathcal{F}_n$ *that approximates* $f(x)$ *best in the sense of eqn* (18), *and* b_n *denote the smallest solution of nonlinear equation*

$$[h_n H^+(b_n h_n^d)]^\alpha = b_n$$

where $H^+(z) = \sup\{t, H(t) > z\}$. *Then, we have:*

(*A*) When $\int |f(x)|^s\, d\mu(x) < \infty$, $s > 1$, $\hat{f}^*_n(x)$ *converges to* $f(x)$ *pointwisely with a rate upper-bounded by* $O(\max\{\theta_{s,n}, b_n\})$ *for almost all* $x(\mu) \in R^d$, *where* $\theta_{s,n} = \min\{(nh_n^d)^{-(s-1)/s}, (nh_n^d)^{-1/2}\}$, *and* h_n *satisfies* $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d \to \infty$.

(*B*) *When* $|f(x)|$ *is bounded,* $f^*_n(x)$ *converges to* $f(x)$ *pointwisely with a rate upper-bounded by* $O(\beta_n\sqrt{\log n})$ *for almost all* $x(\mu) \in R^d$, *where* $\beta_n = \max\{(nh_n^d)^{-1/2}, b_n\}$, *and* h_n *satisfies* $\lim_{n\to\infty} h_n = 0$ *and* $\lim_{n\to\infty} nh_n^d/\log n \to \infty$.

THEOREM 14. (L_2 convergence rate in approximation, RBF.) *Let* $\mathcal{F}_n$ *be the function set defined by the RBF nets* (6), *and* μ *denote the measure on* x. *Assume that* $f(x)$ *satisfies the same conditions as in Theorem* 5, *and that* $\phi(x) \ge cI_{S_r}$, $S_r = \{x|\,\|x\| \le r\}$, $\int \sup_{y\in x+S_r} \phi(y)\, dx < +\infty$ *and* $\|x\|^d \phi(x) \to 0$ *as* $\|x\| \to \infty$. *Let* b_n *denote the smallest solution of the nonlinear equation*

$$[h_n \phi^+(b_n h_n^d)]^{2\alpha} = b_n$$

where $\phi^+(z) = \sup\{t, \phi^*(t) > z\}$, $\phi^*(t) = \sup_x\{\phi(x) I_{\{\|x\|>t\}}(x)\}$.

Let $\hat{f}^*_n(x)$ *denote the one in* $\mathcal{F}_n$ *that approximates* $f(x)$ *best in the sense of eqn* (18), *then*

(*A*) *When* $|f(x)|$ *is bounded,* $\hat{f}^*_n(x)$ *converges to* $f(x)$ *in* L_2 *with a rate upper-bounded by* $O(\max\{(nh^d)^{-1}, b_n\})$; *moreover, if* ϕ *has compact support, then this bound is* $O(n^{-2\alpha/(2\alpha+d)})$, *where* h_n *satisfies* $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d \to \infty$.

(*B*) *When* $\int |f(x)|^s\, d\mu(x) < \infty$, $s > 2 + d/\alpha$, $\hat{f}^*_n(x)$ *converges to* $f(x)$ *in* L_2 *with a rate upper-bounded by* $O(n^{(2/s)-[2\alpha/(2\alpha+d)]})$.

For the cases that $f(x)$ has q-order continuous derivatives, then we have

THEOREM 15. (Pointwise convergence rate in approximation, RBF.) *Let* $\mathcal{F}_n$ *be the function set defined by the RBF nets* (6), *and let* μ *denote the measure on* x *with* $\int_{-\infty}^{\infty} d\mu(x) = 1$. *Assume that* $\int f^2(x)\, d\mu(x) < \infty$, $\mu(x) \in C^{q+1}$, $f(x) \in C^q$ *and* $\int |f^{(q)}(x)|\, d\mu(x) < \infty$. *Also assume that* $\phi(x)$ *satisfies the following conditions:*

(*a*) $\int |\phi|\, dx < \infty$ *and* $\sup|\phi| < \infty$.

(*b*) $\phi(x) = \prod_{i=1}^{d} H(x^{(i)})$, $H(z)$ *is radially symmetric and* $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Let $\hat{f}^*_n(x)$ *denote the one in* $\mathcal{F}_n$ *that approximates* $f(x)$ *best in the sense of eqn* (18), *then* $\hat{f}^*_n(x)$ *converges to* $f(x)$ *pointwisely with a rate upper-bounded* $O(n^{-(q/2q+d)})$.

THEOREM 16. (L_2 convergence rate in approximation, RBF.) *Let* $\mathcal{F}_n$ *be the function set defined by the RBF nets* (6), *and let* μ *denote the measure on* x *with* $\int_{-\infty}^{\infty} d\mu(x) = 1$. *Assume* $|f(x)| \le M$ *and* $\inf_Q[d\mu(x)/dx] > 0$, *and assume that* $\phi(x)$ *satisfies the following conditions:*

(*a*) $\int |\phi|\, dx < \infty$ *and* $\sup|\phi| < \infty$.

(*b*) $\phi(x) = \prod_{i=1}^{d} H(x^{(i)})$, $H(z)$ *is radially symmetric and* $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Let $\hat{f}^*_n(x)$ *denote the one in* $\mathcal{F}_n$ *that approximates* $f(x)$ *best in the sense of eqn* (18), *then* $\hat{f}^*_n(x)$ *converges to* $f(x)$ *in* L_2 *with a rate upper-bounded* $O(n^{-(2q/2q+d)})$, *provided that all partial derivatives of* f *and* $d\mu(x)/dx$ *of orders* i, $1 \le i \le p$ *exist and are square integrable.*

4.5. Convergence Related to the Least Squares Estimator for RBF Nets

Here, we give some results about the least squares estimator used for the RBF net, that is, an estimator $\hat{f}_{n,N}(x)$ obtained by minimization of eqn (3). The results show how the expectation of the empirical error $\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})$ given by eqn (3) changes when $n \to \infty$, $N \to \infty$ with $N \ge n$.

THEOREM 17. *Let* μ *denote the probability measure of* X, $f(x) = E\{Y|X = x\}$ *denote the regression function. Assume that* $f(x)$ *satisfies the same conditions as in Theorem* 5, *and that* $\phi(x) \ge cI_{S_r}$, $S_r = \{x|\,\|x\| \le r\}$,

$\int \sup_{y \in x+S_r} \phi(y)\, dx < +\infty$ *and* $\|x\|^d \phi(x) \to 0$ *as* $\|x\| \to \infty$. *Let* b_n *denote the smallest solution of the nonlinear equation*

$$[h_n \phi^+(b_n h_n^d)]^{2\alpha} = b_n$$

where $\phi^+(z) = \sup\{t,\ \phi^*(t) > z\}$, $\phi^*(t) = \sup_x\{\phi(x) I_{\{\|x\|>t\}}(x)\}$.

Then, as $n \to \infty$, $N \to \infty$ *with* $N \geq n$, *we have*

(*A*) *If* $|Y| \leq M < \infty$ (*M is a positive constant*), *the expectation* $E\{\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\}$ *will drop below* $e_0^2 = E\{|Y_i - f(X_i)|^2\}$ *in probability with the rate* $O(\max\{(nh^d)^{-1}, b_n\})$; *moreover, if* ϕ *has compact support, then this rate is* $O(n^{-[2\alpha/(2\alpha+d)]})$, *where* h_n *satisfies* $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d \to \infty$.

(*B*) *If* $E\{|Y|^s\} < \infty$, $s > 2 + d/\alpha$ *and* ϕ *has compact support, the expectation* $E\{\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\}$ *will drop below* $e_0^2 = E\{|Y_i - f(X_i)|^2\}$ *in probability with the rate* $O(n^{[n(2/s)-[2\alpha/(2\alpha+d)]]})$.

In particular, when $f(x)$ has q-order of continuous derivatives, then we also have

THEOREM 18. *Let* $p(x)$ *denote the probability density of* X. *Assume* $|Y| \leq M$ *and* $\inf_Q p(x) > 0$. *Assume that* $\phi(x)$ *satisfies the following conditions:*

(*a*) $\int |\phi|\, dx < \infty$ *and* $\sup|\phi| < \infty$.

(*b*) $\phi(x) = \prod_{i=1}^d H(x^{(i)})$, $H(z)$ *is radially symmetric and* $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Then, as $n \to \infty$, $N \to \infty$ *with* $N \geq n$, $E\{\varepsilon^2_{\mathrm{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\}$ *will drop below* $e_0^2 = E\{|Y_i - f(X_i)|^2\}$ *in probability with the rate* $O(n^{-[2q/(2q+d)]})$, *provided that all partial derivatives of* f *and* p *of orders* i, $1 \leq i \leq p$ *exist and are square integrable.*

4.6. Examples of Convergence Rates for Some Specific Basis Functions

Let us now obtain the rates of pointwise convergences on some examples to see how the rates vary according to the types of the radial basis functions being used.

As shown in Krzyzak and Pawlak (1987), for certain types of functions, including basis functions with compact supports, Gaussian functions, and basis functions with polynomial tails, the key parameter b_n, defined in Theorems 5, 9, and 13, can be solved explicitly. Using these solutions, we can obtain a number of useful insights on the rates given in these theorems:

For all basis functions with compact support, we have $b_n = h_n^\alpha$. Let $s = 2$, then it follows that:

1. when $h_n = n^{-\tau}$, $0 < \tau < 1/d$, in Theorems 5(A), 9(A), and 13(A) the pointwise convergence rate given becomes $O(n^{-\alpha/(2\alpha+d)})$;
2. when $h_n = n^{-\tau}$, $0 < \tau < 1/2d$, in Theorems 5(B) and 9(B) the pointwise convergence rate given becomes $O(n^{-0.5\alpha/(2\alpha+d)}\sqrt{\log n})$;
3. when $h_n = n^{-\tau}$, $0 < \tau < 1/d$, in Theorems 5(C), 9(C), and 13(B) the pointwise convergence rate given becomes $O(n^{-\alpha/(2\alpha+d)}\sqrt{\log n})$.

For Gaussian basis functions, we have $b_n = h_n^\alpha |\ln h_n|$. Let $s = 2$, then it follows that:

1. when $h_n = n^{-\tau}$, $0 < \tau < 1/d$, in Theorems 5(A), 9(A), and 13(A) the pointwise convergence rate given becomes $O[n^{-\alpha\tau}\ln n]$;
2. when $h_n = n^{-\tau}$, $0 < \tau < 1/2d$, in Theorems 5(B) and 9(B) the pointwise convergence rate given becomes $O[n^{-\alpha\tau}(\ln n)^{3/2}]$;
3. when $h_n = n^{-\tau}$, $0 < \tau < 1/d$, in Theorems 5(C), 9(C), and 13(B) the pointwise convergence rate given becomes $O[n^{-\alpha\tau}(\ln n)^{3/2}]$.

For basis functions with polynomial tails [i.e., $H(t) = 1/t^{d+\eta}$, $\eta > 0$], we have $b_n = h_n^{\alpha\eta/(\alpha+\eta+d)}$. Let $s = 2$, it follows that:

1. when $h_n = n^{-\tau}$, $0 < \tau < 1/d$, in Theorems 5(A), 9(A), and 13(A) the pointwise convergence rate given becomes $O(n^{-\alpha\eta/[d(\alpha+\eta+d)+2\alpha\eta]})$;
2. when $h_n = n^{-\tau}$, $0 < \tau < 1/2d$, in Theorems 5(B) and 9(B) the pointwise convergence rate given becomes $O(n^{-0.5\alpha\eta/[d(\alpha+\eta+d)+2\alpha\eta]}\sqrt{\log n})$;
3. when $h_n = n^{-\tau}$, $0 < \tau < 1/d$, in Theorems 5(C), 9(C), and 13(B) the pointwise convergence rate given becomes $O(n^{-\alpha\eta/[d(\alpha+\eta+d)+2\alpha\eta]}\sqrt{\log n})$.

Next we consider the pointwise convergence rates given by these theorems for the special case that $\alpha = 1$, $d = 1$. It follows that: (i) for the basis functions with compact support, the rates become $O(n^{-1/3})$, $O(n^{-1/6}\sqrt{\log n})$, and $O(n^{-1/3}\sqrt{\log n})$, respectively; (ii) for Gaussian functions, the rates become $O(n^{-\tau}\ln n)$, $O[n^{-\tau}(\ln n)^{3/2}]$, and $O[n^{-\tau}(\ln n)^{3/2}]$, respectively; (iii) for basis functions with polynomial tails, when $\eta = 1$, the rates become $O(n^{-1/5})$, $O(n^{-1/10}\sqrt{\log n})$, and $O(n^{-1/5}\sqrt{\ln n})$, respectively, and when $\eta = \infty$, the rates become $O(n^{-1/3})$, $O(n^{-1/6}\sqrt{\log n})$, and $O(n^{-1/3}\sqrt{\ln n})$,respectively. We see that *the heavier the tail of the basis functions the slower the rate of convergence*. In fact, we can obtain a continuum of rates depending upon the rate of decrease of the tail of the basis functions. Therefore, *to obtain a good convergence rate we should prefer basis functions with compact support or light tails*.

5. SELECTION OF THE SIZE OF RECEPTIVE FIELD

In practical applications, RBF nets of type-I or type-II are commonly used because of their simplicity. That is, the size $\sigma_n = h_n$ of the receptive field for an RBF net is usually predefined externally. Naturally, there rises a question: how can the parameter h_n be selected appropriately? At present very little is known about the appropriate selection of this parameter and in practice the selection is usually based on heuristic strategies (Mel & Omohundro, 1991). Intuitions and experimental

experience have told us that either having h_n too large or too small will deteriorate the performances of RBF nets. This suggests that the estimation error $e_x(f, f_n)$ or $\rho_U^2(f, f_n)$ defined in eqn (17) may consist of two components. One decreases with h_n and the other increases with h_n. To select h_n, we need to trade off between the two components to make the overall estimation error minimized.

To study this issue more deeply, we concentrate on the KRE estimator eqn (7)—a particular RBF net of type-I. The following two theorems show that both $e_x(f, g_n)$ and $\rho_U^2(f, g_n)$ can be divided into such two parts.

THEOREM 19. (Pointwise error.) *Let $p(x)$ denote the probability density of X. Assume that $E\{Y^2\} < \infty$, $p(x) \in C^q$, $f(x) \in C^q$, and $\int |f^{(q)}(x)|\, d\mu(x) < \infty$, where C^q is the set of all functions that have q-order continuous derivatives. Also assume that $K(x)$ satisfies the following conditions:*

(a) $\int |K|\, dx < \infty$ and $\sup|K| < \infty$.

(b) $K(x) = \prod_{i=1}^d H(x^{(i)})$, $H(z)$ is radially symmetric and $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Then letting $h_n \to 0$ and $nh_n^d \to \infty$, we have

$$e_x(f, g_n) = |g_n(x) - f(x)| \overset{(P)}{=} O\left[\frac{A(x)}{\sqrt{nh_n^d}} + B(x)h_n^q\right] \quad (22)$$

for all points of continuity of $s^2(x) = E\{Y^2 | X = x\}$, where $\overset{(P)}{=}$ means being equal to . . . in probability, *and $A(x) = A_1(x) + A_2(x)$, $B(x) = B_1(x) + B_2(x)$ with*

$$A_1(x) = \left[\frac{E\{Y^2|X=x\}}{p(x)} \int K^2(t)\, dt\right]^{1/2},$$

$$A_2(x) = \left[\frac{\int K^2(t)\, dt}{p(x)}\right]^{1/2}$$

$$B_1(x) = \left\{\frac{1}{q!}\sum_{i=1}^d \left[f(x)\frac{\partial^q p(x)}{\partial x^{(i)q}} + p(x)\frac{\partial^q f(x)}{\partial x^{(i)q}}\right] + \sum_{i=1}^{q-1} \frac{1}{i!}\frac{1}{(q-i)!} \sum_{j=1}^d \frac{\partial^i f(x)}{\partial x^{(j)i}} \frac{\partial^{(q-i)} p(x)}{\partial x^{(j)(q-i)}}\right\} \int z^q H(z)\, dz,$$

$$B_2(x) = \sum_{i=1}^d \frac{1}{q!}\frac{\partial^q p(x)}{\partial x^{(i)q}} \int z^q H(z)\, dz$$

(*note:* $x = [x_1, \ldots, x_d]^t$).

THEOREM 20. (L_2 error.) *Let $p(x)$ denote the probability density of X. Assume $|Y| \le M$ and $\inf_Q p(x) > 0$. Assume that $K(x)$ satisfies the following conditions:*

(a) $\int |K|\, dx < \infty$ and $\sup|K| < \infty$.

(b) $K(x) = \prod_{i=1}^d H(x^{(i)})$, $H(z)$ is radially symmetric and $\int H(z)\, dz = 1$, $\int z^{(i)} H(z)\, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)|\, dz < \infty$.

Letting $h_n \to 0$ and $nh_n^d \to \infty$, we have

$$\rho_U^2(f, g_n) = \int_U |g_n(x) - f(x)|^2\, d\mu(x) \overset{(P)}{=} O\left(\frac{C}{nh_n^d} + Dh_n^{2q}\right) \quad (23)$$

where U is a compact subset of R^d provided that all partial derivatives of R and p of orders i, $1 \le i \le p$ exist and are square integrable. Here, $C = C_1 + C_2$ and $D = D_1 + D_2$ with

$$C_1 = \frac{32M^2}{c} \int K^2(t)\, dt,\ C_2 = \frac{16M^2}{c} \int K^2(t)\, dt,$$

$$D_1 = \int_Q \frac{D_1^2(x)}{p(x)}\, dx,$$

$$D_1(x) = \left\{\frac{1}{q!}\sum_{i=1}^d \left[f(x)\frac{\partial^q p(x)}{\partial x^{(i)q}} + p(x)\frac{\partial^q f(x)}{\partial x^{(i)q}}\right] + \sum_{i=1}^{q-1} \frac{1}{i!}\frac{1}{(q-i)!} \sum_{j=1}^d \frac{\partial^i f(x)}{\partial x^{(j)i}} \frac{\partial^{(q-i)} p(x)}{\partial x^{(j)(q-i)}}\right\} \int z^q H(z)\, dz,$$

$$D_2 = \int_Q \frac{D_2^2(x)}{p(x)}\, dx,\ D_2(x) = \left[\frac{1}{q!}\sum_{i=1}^d \frac{\partial^q p(x)}{\partial x^{(i)q}}\right] \int z^q H(z)\, dz.$$

The above theorems tell us that both the pointwise and the L_2 estimation errors consist of two components. One monotonically increases with h_n at the order of $O(h^q)$ pointwisely or $O(h^{2q})$ in L_2; the other component reduces with h_n at the order of $O(1/\sqrt{nh^d})$ pointwisely or $O(1/nh^d)$ in L_2. This is why the condition that $h_n \to 0$ and $nh_n^d \to \infty$ has appeared in many of the previous theorems: it is to force the estimation errors to tend to zero.

So, to minimize the overall estimation errors, we need to trade off the two components. That is, we should minimize the following functions with respect to h_n:

$$G(h_n) = \frac{A(x)}{\sqrt{nh_n^d}} + B(x)h_n^q, \quad \text{for pointwise error,}$$

$$G(h_n) = \frac{C}{nh_n^d} + Dh_n^{2q}, \quad \text{for } L_2 \text{ error.} \quad (24)$$

This minimization will give us that:

$$h_n = O\left\{\left[\frac{A(x)\, d}{2qB(x)\sqrt{n}}\right]^{2/(2q+d)}\right\}, \quad \text{for pointwise error,}$$

$$h_n = O\left[\left(\frac{Cd}{2qDn}\right)^{1/(2q+d)}\right], \quad \text{for } L_2 \text{ error.} \quad (25)$$

As shown in Theorems 19 and 20, the accurate calculation of the coefficients $A(x)$, $B(x)$, C, D requires knowledge of $f(x)$, $p(x)$ as well as some statistical properties of Y. Thus, it is usually difficult to use eqn (25) to accurately decide the optimal h_n. However, eqn (25) can still be used to qualitatively guide the selection of the receptive size of RBF nets of type-I and type-II, given by eqn (6). We know that h_n should be roughly

of the order given by eqn (25). We know that the larger the number of the basis functions used in a net the better we can reduce the size of receptive field; on the other hand, a receptive field of large size is best used for a net with a small number of basis functions. Moreover, we know that the size is also closely related to the smoothness q of the regression function and the dimension d of x, especially to their ratio. If the function to be estimated increases its smoothness proportionally to the dimensionality, that is, if we have a constant ratio d/q, then when n is large enough, we can use larger receptive field sizes for the smoother functions.

We will now point out that the above decomposition of the estimation errors into two components is closely related to the decomposition of the errors into its bias and variation parts. The trade-off between these two components is closely related to the well-known trade-off between bias and variation in the statistics literature.

The bias-variation decomposition is given as follows:

$$\begin{aligned} e_x(f, g_n) &= |g_n(x) - f(x)| \le B_x(h_n) + V_x(h_n), \\ B_x(h_n) &= |E\{g_n(x)\} - f(x)|, \\ V_x(h_n) &= |g_n(x) - E\{g_n(x)\}|, \end{aligned} \tag{26}$$

$$\rho_U^2(f, g_n) = \int_U |g_n(x) - f(x)|^2 \, d\mu(x) = B^2(h_n) + V^2(h_n),$$

$$B^2(h_n) = \int_U |E\{g_n(x)\} - f(x)|^2 \, d\mu(x),$$

$$V^2(h_n) = \int_U |g_n(x) - E\{g_n(x)\}|^2 \, d\mu(x), \tag{27}$$

where $B_x(\cdot)$, $B^2(\cdot)$ are called bias terms, and $V_x(\cdot)$, $V^2(\cdot)$ are called variation terms.

Roughly, the variations $V_x(\cdot)$, $V^2(\cdot)$ correspond to the components above that decrease with h_n with order $O(1/\sqrt{nh^d})$ pointwisely or $O(1/nh^d)$ in L_2; the bias corresponds to the components above that increases with h_n with order $O(h^q)$ pointwisely or $O(h^{2q})$ in L_2. Such a relation becomes precisely true for a special kind of kernel estimator—the Parzen window estimator (9), as shown in the following theorem.

THEOREM 21. (Parzen window.) *Let $E\{Y^2\} < \infty$. Assume that $p(x) \in C^p$ is the density function of X and $K(x)$ satisfies the following conditions:*

(a) $\int |K| \, dx < \infty$ *and* $\sup|K| < \infty$.

(b) $K(x) = \prod_{i=1}^d H(x^{(i)})$, *$H(z)$ is radially symmetric and* $\int H(z) \, dz = 1$, $\int z^{(i)} H(z) \, dz = 0$, $i = 1, \ldots, q-1$, $0 < \int |z|^q |H(z)| \, dz < \infty$.

Let $h_n \to 0$, $nh_n^d \to \infty$. Then for Parzen window estimator (9), we have:

(A) For the pointwise error,

$$V_x(h_n) = |p_n(x) - E\{p_n(x)\}| \overset{(P)}{=} O\left[\frac{A(x)}{\sqrt{nh_n^d}}\right]$$

$$B_x(h) = |E\{p_n(x)\} - p(x)| \overset{(P)}{=} O[B(x)h^q]$$

at $x \in C^q(p)$, where $C^q(f)$ is the set of continuity points of $f^{(q)}$, and

$$A(x) = \sqrt{p(x) \int K^2(t) \, dt},$$

$$B(x) = \frac{1}{q!} \sum_{i=1}^d \frac{\partial^q p(x)}{\partial x_i^{(p)}} \int z^q H(z) \, dz.$$

Furthermore, the convergence rate of $|p_n(x) - p(x)|$ is upper-bounded by $O\{[A(x)/\sqrt{nh_n^d}] + B(x)h_n^q\} = O(n^{-q/(2q+d)})$.

(B) Let $p(x)$ be square integrable. For the L_2 error,

$$V^2(h_n) = \int |p_n(x) - Ep_n(x)|^2 \, dx \overset{(P)}{=} O\left(\frac{A}{nh_n^d}\right)$$

$$B^2(h) = \int |E\{p_n(x)\} - p(x)|^2 \, dx \overset{(P)}{=} O(Bh_n^{2q}).$$

Furthermore, the convergence rate of $\int |p_n(x) - p(x)|^2 \, dx$ is upper-bounded by $O[(A/nh_n^d) + Bh_n^{2q}] = O(n^{-[2q/(2q+d)]})$. $A = \int K^2(t) \, dt$ and $B = \int B^2(x) \, dx$ with $B(x)$ being the same as the one given in (A) provided that $\partial^q p(x)/\partial x_i^{(q)}$, $i = 1, \ldots d$ are square integrable.

This theorem also justified the use of eqn (24) to qualitatively guide the selection of h_n for designing a special kind of KRE—Parzen window estimator, particularly for probabilistic neural networks (Specht, 1990).

6. CONCLUSIONS

By the connections we established between RBF nets and KRE, we showed that the theoretical results about KRE can be used as tools to obtain theoretical results for RBF nets. We have presented upper bounds for the convergence rates of the approximation error, proved constructively the existence of a consistent estimator for RBF nets, and also provided upper bounds for the pointwise and L_2 convergence rates of the best consistent estimator for RBF nets. Moreover, we have also studied the problem of selecting the appropriate size of the receptive field of the radial basis function and the convergence of the empirical error obtained by the least squares estimator for RBF nets. The results are useful for further theoretical analysis of RBF nets as well as for guiding the design of a RBF net in practice. The remaining open problems are whether the commonly used least squares estimators for RBF nets of type-I and type-II are consistent and whether it can get a better convergence rate than those given in this paper.

REFERENCES

Barron, A. R. (1991). Approximation and estimation bounds for artificial neural networks. *Proceedings of 4th Annual Workshop on Computational Learning Theory* (pp. 243–249). San Mateo, CA: Morgan Kaufmann.

Bennett, G. (1962). Probability inequalities for the sums of independent random variable. *Journal of American Statistical Association*, **57,** 33–45.

Botros, S. M., & Atkeson, C. G. (1991). Generalization properties of radial basis function. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing system* 3 (pp. 707–713). San Mateo: Morgan Kaufmann.

Broomhead, D. S., & Lowe, D. (1988). Multivariable functional interpolation and adaptive networks. *Complex Systems*, **2,** 321–323.

Chen, S., Cowan, C. F. N., & Grant, P. M. (1991). Orthogonal least squares learning algorithm for radial basis function networks. *IEEE Transactions on Neural Networks*, **2,** 302–309.

Chow, Y. S., & Teicher, H. (1978). *Probability theory, independence, interchangeability, martingale*. New York: Springer.

Corradit, V., & White, H. (1992). *Regularized neural networks: Some convergence rate results*. Unpublished manuscript, University of California, Department of Economics, San Diego.

Devroye, L. (1987). *A course in density estimation*. Boston: Birkhauser.

Devroye, L. (1981). On the almost everywhere convergence of nonparametric regression function estimates. *The Annals of Statistics*, **9,** 1310–1319.

Devroye, L., & Krzyzak, A. (1989). An equivalence theorem for L_1 convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **23,** 71–82.

Geman, S., Bienenstock, E., & Doursat, R. (1992). Neural networks and the bias/variance dilemma. *Neural Computation*, **4,** 1–58.

Girosi, F., & Anzellotti, G. (1992). Convergence rates of approximation by translates. MIT AI Memo, No. 1288, MIT, Cambridge.

Girosi, F., & Poggio, T. (1989). Networks and the best approximation property. MIT AI Memo, No. 1164, MIT, Cambridge.

Girosi, F., & Poggio, T. (1990). Networks and the best approximation property. *Biological Cybernetics,* **63,** 169–176.

Greblicki, W., Krzyzak, A., & Pawlak, M. (1984). Distribution-free consistency of kernel regression estimate. *The Annals of Statistics*, **12,** 1570–1575.

Hartman, E. J., Keeler, J. D., & Kowalski, J. M. (1990). Layered neural networks with Gaussian hidden units s universal approximations. *Neural Computation*, **2,** 210–215.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, **58,** 13–30.

Hornik, K. (1991). Approximation capabilities of multilayer feedforward networks. *Neural Networks*, **4,** 251–257.

Hornik, K., Stinchcombe, S., & White, H. (1989). Multilayer feedforward networks are universal approximators. *Neural Networks*, **2,** 359–366.

Jones, R. D., Lee, Y. C., Barnes, C. W., et al. (1991). Information theoretic derivation of network architecture and learning algorithms. *Proceedings of International Joint Conference on Neural Networks* 1991. Seattle (Vol. I, pp. 473–478).

Kardirkamanathan, V., Niranjan, M., & Fallside, F. (1991). Sequential adaptation of radial basis function neural networks and its application to time-series prediction. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing system* 3 (pp. 721–727). San Mateo: Morgan Kaufmann.

Kraaijveld, M. A., & Duin, R. P. W. (1991). Generalization capabilities of minimal kernel-based networks. *Proceedings of International Joint Conference on Neural Networks* 1991. Seattle (Vol. I, pp. 843–848).

Krzyżak, A. (1986). The rates of convergence of kernel regression estimates and classification rules. *IEEE Transactions on Information Theory*, **32,** 668–679.

Krzyżak, A. (1991). On exponential bounds on the Bayes risk of the kernal classification rule. *IEEE Transactions on Information Theory*, **37,** 490–499.

Krzyżak, A., & Pawlak, M. (1987). The pointwise rate of convergence of the kernel regression estimate. *Journal of Statistical Planning and Inference*, **16,** 159–166.

Lugosi, G., & Zeger, K. (1993). *Nonparametric estimation via empirical risk minimization*. Manuscript submitted for publication.

Mel, B. W., & Omohundro, S. M. (1991). How receptive field parameters affect neural learning. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky (Eds.), *Advances in neural information processing system* 3 (pp. 757–763). San Mateo: Morgan Kaufmann.

Moody, J., & Darken, J. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, **1,** 281–294.

Nowlan, S. J. (1990) *Max likelihood competition in RBF networks* (Tech. Rep. CRG-TR-90-2). University of Toronto, Department of Computer Science.

Park, J., & Sandberg, I. W. (1991). Universal approximation using radial-basis-function networks. *Neural Computation*, **3,** 246–257.

Park, J., & Sandberg, I. W. (1993). Universal approximation using radial-basis-function networks. *Neural Computation*, **5,** 305–316.

Platt, J. C. (1991). Learning by combining memorization and gradient descent. In R. P. Lippmann, J. E. Moody, & D. S. Touretzky (Eds.), *Advances in neural information processing system* 3 (pp. 714–720). San Mateo: Morgan Kaufmann.

Poggio, T., & Girosi, F. (1989). A theory of networks for approximation and learning. MIT AI Memo, No. 1140, MIT, Cambridge.

Poggio, T., & Girosi, F. (1990). Networks for approximation and learning. *Proceedings of the IEEE,* **78,** 1481–1497.

Powell, M. J. D. (1987). Radial basis functions for multivariable interpolation: A review. In J. C. Mason & M. G. Cox (Eds.), *Algorithms for approximation*. Oxford: Clarendon Press.

Renals, S., & Rohwer, R. (1989). Phoneme classification experiments using radial basis functions. *Proceedings of International Joint Conference on Neural Networks* 1989. Washington, DC (Vol. I, pp. 462–467).

Specht, D. F. (1990). Probabilistic neural networks. *Neural Networks*, **3,** 109–118.

Stokbro, K., Umberger, D. K., & Hertz, J. A. (1990). Exploiting neurons with localized receptive fields to learn chaos. Preprint 90/28 S, Nordita, Copenhagen, Denmark.

Weymaere, N., & Martens, J. (1991). A fast robust learning algorithm for feed-forward neural networks. *Neural Networks*, **4,** 361–369.

Wheeden, R. L., & Zygmund, A. (1977). *Measure and integral*. New York: Marcel Dekker.

White, H. (1990). Connectionist nonparametric regression: Multilayer feedforward networks that can learn arbitrary mappings. *Neural Networks*, **3,** 535–549.

Xu, L., Klasa, S., & Yuille, A. L. (1992). Recent advances on techniques static feedforward networks with supervised learning. *International Journal of Neural Systems*, **3**(3), 253–290.

Xu, L., Krzyżak, A., & Oja, E. (1993). Rival penalized competitive learning for clustering analysis, RBF net and curve detection. *IEEE Transactions on Neural Networks*, **4,** 636–649.

Xu, L., Krzyżak, A., & Yuille, A. L. (1992). On radial basis function nets and kernel regression: Approximation ability, convergence rate and receptive field size. (Tech. Rep. No. 92-4). Harvard University, Harvard Robotics Laboratory.

Yuille, A. L., & Grzywacz, N. M. (1989). A mathematical analysis of the motion coherence theory. *International Journal of Computer Vision*, **3,** 155–175.

APPENDIX: THE PROOFS OF THEOREMS

In the following proofs, some background knowledge of probability theory and nonparametric statistics are assumed to be known by readers (Chow & Teicher, 1987; Devroye, 1981, 1987; Hoeffding, 1963).

Proof of Lemma 1. (*A*) $E\{\varepsilon^2_{\mathrm{RBF}}(Y, f_{n,N})\} = E\{(1/N)\sum_{i=1}^{N} |Y_i - f_{n,N}(X_i)|^2\} = (1/N)\sum_{i=1}^{N} E\{|Y_i - f_{n,N}(X_i)|^2\} = E\{|Y_i - f_{n,N}(X_i)|^2\}$. Similarly, we have $E\{\varepsilon^2_{\mathrm{KRE}}(Y, g_n)|\mathcal{D}_n^g\} = E\{|Y_i - g_n(X_i)|^2|\mathcal{D}_n^g\}$.

Furthermore, we have

$$E\{|Y_i - g_n(X_i)|^2|\mathcal{D}_n^g\} = E\{|Y_i - f(X_i)|^2|\mathcal{D}_n^g\} + 2E\{[Y_i - f(X_i)][f(X_i) - g_n(X_i)]|\mathcal{D}_n^g\} + E\{|f(X_i) - g_n(X_i)|^2|\mathcal{D}_n^g\}.$$

When $\mathcal{D}_N$, $\mathcal{D}_n^g$ are independent, we have

$$E\{[Y_i - f(X_i)][f(X_i) - g_n(X_i)]|\mathcal{D}_n^g\} = E\{E\{[Y_i - f(X_i)][f(X_i) - g_n(X_i)]|\mathcal{D}_n^g, X_i\}\} = E\{0\} = 0$$

$$E\{|Y_i - f_n(X_i)|^2|\mathcal{D}_n^g\} = E\{|Y_i - f_n(X_i)|^2\} = e_0^2$$

$$E\{|f(X_i) - g_n(X_i)|^2|\mathcal{D}_n^g\} = \int_U |f(x) - g_n(x)|^2 \, d\mu(x) = e_{\mathrm{KRE}}^2(f, g_n).$$

(*B*) By the definitions, $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})$ is the minimum of $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N})$ with respect to all the parameters w_i, c_i, $i = 1, \ldots, n$ and Σ simultaneously, while the kernel regression estimator (7) is a particular specified RBF net with all the parameters specified directly by $c_i = X_i'$, $w_i = Y_i'$, $i = 1, \ldots, n$, $\Sigma = h_n^2 I$ without directly minimizing $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N})$. Thus, it is easy to see that $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N}) \le \varepsilon_{\mathrm{KRE}}^2(\mathcal{D}_N, g_n)$. When $\mathcal{D}_N$, $\mathcal{D}_n^g$ are independent, by taking expectations conditioning on $\mathcal{D}_n^g$, we have $E\{\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})\} \le E\{\varepsilon_{\mathrm{KRE}}^2(\mathcal{D}_N, g_n)|\mathcal{D}_n^g\}$. It further follows (*A*) that $E\{\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})\} \le e_0^2 + e_{\mathrm{KRE}}^2(f, g_n)$.

(*C*) Because $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})$ is the minimum of $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N})$ with respect to all the parameters, and $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N}^I)$ is the minimum of $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N})$ with respect to only w_i, $i = 1, \ldots, n$ with the previously determined specific values of Σ and $c_i = X_i$, $i = 1, \ldots, n$, we have $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N}) \le \varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N}^I)$. Moreover, when ${}^x\mathcal{D}_n^g = \mathcal{D}_c^x$ and when the same receptive field $\Sigma = h_n^2 I$ are used for the type-I RBF nets and KRE, the only difference between the specified RBF nets of type-I given by eqn (6) and the KRE given by eqn (7) is that for the former the weights w_i are solved for by minimizing $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N})$, and for the latter the weights w_i simply take the values given by Y_i', $i = 1, \ldots, n$. Therefore, we have $\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N}^I) \le \varepsilon_{\mathrm{KRE}}^2(\mathcal{D}_N, g_n)$. Furthermore, by taking expectations we have $E\{\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})\} \le E\{\varepsilon_{\mathrm{KRE}}^2(\mathcal{D}_N g_n)\} = E\{|Y_i - g_n(X_i)|^2\}$, where

$$E\{|Y_i - g_n(X_i)|^2\} = e_0^2 + 2E\{[Y_i - f(X_i)] \times [f(X_i) - g_n(X_i)]\} + E\{|f(X_i) - g_n(X_i)|^2\}.$$

When $\mathcal{D}_N$, ${}^y\mathcal{D}_n^g$ are independent conditioning on ${}^x\mathcal{D}_n^g$, we have

$$E\{[Y_i - f(X_i)][f(X_i) - g_n(X_i)]\} = E\{E\{[Y_i - f(X_i)][f(X_i) - g_n(X_i)]|{}^x\mathcal{D}_n^g, X_i\}\} = E\{0\} = 0$$

$$E\{|f(X_i) - g_n(X_i)|^2\} = E\{E\{|f(X_i) - g_n(X_i)|^2|\mathcal{D}_n^g\}\} = E\int_U |f(x) - g_n(x)|^2 \, d\mu(x) = E\{e_{\mathrm{KRE}}^2(f, g_n)\}.$$

In summary, we have $E\{\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, \hat{f}_{n,N})\} \le E\{\varepsilon_{\mathrm{RBF}}^2(\mathcal{D}_N, f_{n,N}^I)\} \le e_0^2 + E\{e_{\mathrm{KRE}}^2(f, g_n)\}$. ■

Proof of Theorem 1. For the special case that $m = 1$ (i.e., $Y \in R$), the proof of (*D*) is given in Theorem 2 of Greblicki, Krzyzak, and Pawlak (1984), (*E*) follows from the proof of (*C*), the statements (*A*)–(*C*) will be proven below.

(*A*) Using the inequality from Krzyzak and Pawlak (1987, p. 163) we have

$$P\left\{\left|\frac{N}{D} - m(x)\right| > t\right\} \le P\{|N - m(x)| > \varepsilon\} + P\{|D - 1| > \varepsilon\}$$

where $\varepsilon = t/(t + |m(x)| + 1)$ and

$$N = \sum_{i=1}^n [Y_i - f(x)]K_h(x - X_i)/nE\{K_h(x - X)\}$$

$$D = \sum_{i=1}^n K_h(x - X_i)/nE\{K_h(x - X)\} - 1.$$

Let us use the truncation approach. Denote $Y' = YI_{\{|Y| \le (nh^d)^{1/s}\}}$, $Y'' = Y - Y'$, $m'(x) = E\{Y'|X = x\}$, $m''(x) = m(x) - m'(x)$. Let N_n', N_n'' be N with Y replaced by Y' and Y'', respectively. We have

$$P\left\{\left|\frac{N}{D} - m(x)\right| > \varepsilon\right\} \le P\{|N' - m'(x)| > \varepsilon/2\} + P\{|N'' - m''(x)| > \varepsilon/2\} = A + B.$$

Clearly,

$$\varepsilon^2/4A \le \mathrm{var}(N') + [EN' - m'(x)]^2 = E\sigma'^2(X)K_h^2(x - X)/nE^2K_h(x - X) + [EY'K_h(x - X)/EK_h(x - X) - m'(x)] = A_1 + A_2$$

where $\sigma(x) = E\{Y^2|X = x\}$. Using the same argument as in Krzyzak and Pawlak (1987, p. 162) we get that A_1 is not greater than

$$c(x)\sigma_s(x)(nh^d)^{2(s-1)/s} \quad \text{for} \quad 1 < s < 2,$$

$$c(x)\sigma_2(x)(nh^d)^{-1} \quad \text{for} \quad 2 \le s,$$

where $\sigma_s(x) = E\sigma^s(X)K_h(x - X)/EK_h(x - X)$ and $c(x)$ is finite for almost all x mod μ. By Lemma 1 of Krzyżak and Pawlak (1987), $\sigma^s(x) \to 0$ as $h \to 0$ mod μ, we can easily see that $A_1 \to 0$ as $h \to 0$. To conclude the proof of (*A*) notice

$$B \le 2/\varepsilon E|N'' - m''(x)| \le 2/\varepsilon[\sigma_s(x) + E\{Y^s|X = x\}](nh^d)^{(s-1)/s} \to 0$$

as $h \to 0$ mod μ (Krzyzak & Pawlak, 1987, p. 163).

(*B*) We will generalize the proof of (*A*). Let us start with the inequality

$$P\{|N - m(x)| > \varepsilon\} \le P\{|N - EN| > \varepsilon/2\} + P\{|EN - m(x)| > \varepsilon/2\} = A + B.$$

We have by the result of Krzyzak (1986) $EN - m(x) = E[m(X) - m(x)]K_h(x - X)/EK_h(x - X) \to 0$ as $h \to 0$ for almost all x mod μ, so $B = 0$ for n large enough. On the other hand, by Bennett's inequality (Bennett, 1962) and

$$YK_h(x - X)/EK_h(x - X) \le c_1(x)h^{-d}$$

$$EY^2K_h^2(x - X)/E^2K_h(x - X) \le c_2(x)\sigma_1(x)h^{-d}$$

we have

$$A \le 2\exp[c_3(x)nh^d]$$

where c_1, c_2, and c_3 are finite for almost all x mod μ (Krzyzak & Pawlak, 1987, p. 164).

(*C*) The proof follows from Krzyzak and Pawlak (1987, p. 164) except for the term $|EN - m(x)|$. For this term we have

$$EN - m(x) = Em(X)K_h(x - X)/EK_h(x - X) - m(x) \to 0$$

as $h \to 0$ for almost all x mod μ.

These proofs can be easily extended to the general case that $m > 1$ (i.e., $Y \in R^m$). By noticing that for $q \ge 1$,

$$|f(x) - g_n(x)|^q = \sum_{i=1}^m |R^{(i)}(x) - g_n^{(i)}(x)|^q \qquad (A.1)$$

for $f(x) = [R^{(1)}(x), \ldots, R^{(m)}(x)]^T$ and $g_n(x) = [g_n^{(1)}(x), \ldots, g_n^{(m)}(x)]^T$. Because for each dimension i, $|R^{(i)}(x) - g_n^{(i)}(x)|^q \to 0$ as $n \to \infty$ *in probability/almost surely/completely, their sum* $|f(x) - g_n(x)|^q \to 0$ also as $n \to \infty$ *in probability/almost surely/completely.* ■

Proof of Theorem 2. Let $Y' = YI_{\{|Y| \le M\}}$, $Y'' = Y - Y'$. We have

$$|g_n(x) - m(x)| \le \left|\sum_{i=1}^n (Y_i - Y_i')K_h(x - X_i) \Big/ \sum_{i=1}^n K_h(x - X_i)\right| + \left|\sum_{i=1}^n [m'(x) - m(x)]K_h(x - X_i) \Big/ \sum_{i=1}^n K_h(x - X_i)\right| + \left|\sum_{i=1}^n [Y_i' - m'(x)]K_h(x - X_i) \Big/ \sum_{i=1}^n K_h(x - X_i)\right| = A + B + C$$

where m' corresponds to Y' and M is a finite constant. The rest of the proof is similar to the one given and is omitted. ■

Proof of Theorem 3. Let us randomly choose a subset $\mathcal{D}_n = \{X_i, Y_i\}_1^n$ among $\mathcal{D}_N$. For the RBF nets (2), we simply let the parameters c_i, w_i's of the estimator to be chosen be fixed at $c_i = X_i$, $w_i = Y_i$, $i = 1, \ldots, n$, and let the receptive field be specified hyperspherically with $\Sigma = h_n^2 I$. Then we have an estimator that is same as the KRE eqn (7). So for such an estimator with $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} nh_n^d(n) \to \infty$, it follows from Theorem 1(*A*) that it is pointwisely consistent to $f(x)$ *in probability* for almost all $x(\mu) \in R^d$. Moreover, for such an estimator with $\lim_{n\to\infty} h_n = 0$, $\lim_{n\to\infty} [n^{(s-1)/s}h_n^d(n)/\log n] \to \infty$, it follows from Theorem 1(*C*)(*E*) that it is pointwisely consistent to $f(x)$ *almost surely* and *completely, respectively, for almost all* $x(\mu) \in R^d$. ■

Proof of Theorem 4. The proof is similar to the above one. Now we use Theorem 2 instead of Theorem 1. ■

Proof of Theorem 5. For the special case that $m = 1$ (i.e., $Y \in R$), the proofs of (*A*)(*B*) are given by the proofs of Theorems 1 and 2 in Krzyzak and Pawlak (1987). The proof of (*C*) is given in the proof of Theorem 2 in Krzyzak (1986).

The proofs can also be extended to the general case with $m > 1$ (i.e., $Y \in R^m$). By noticing

$$e_x(f, g_n) = |f(x) - g_n(x)| = \sum_{i=1}^m |R^{(i)}(x) - g_n^{(i)}(x)| \qquad \text{(A.2)}$$

for $f(x) = [R^{(1)}(x), \ldots, R^{(m)}(x)]^T$ and $g_n(x) = [g_n^{(1)}(x), \ldots, g_n^{(m)}(x)]^T$ we can see that the convergence rate of $e_x(f, g_n)$ will be $\sum_{i=1}^m O[r(n)] = mO[r(n)]$, which is still the same order $O[r(n)]$ when the convergence rate of each $|R^{(i)}(x) - g_n^{(i)}(x)|$ is $O[r(n)]$. Because each of the rates given in the above statements (*A*), (*B*), and (*C*) holds for each $|R^{(i)}(x) - g_n^{(i)}(x)|$ (i.e., the special case that $m = 1$), by using the above arguments we can see that these statements also hold for the general case that $m > 1$ (i.e., $Y \in R^m$). ■

Proof of Theorem 6. Let us write

$$|g_n(x) - f(x)| = \left|\frac{N}{D+1}\right|$$

where

$$N = \sum_{i=1}^n [Y_i - f(x)]K_h(x - X_i)/nE\{K_h(x - X)\}$$

$$D = \sum_{i=1}^n K_h(x - X_i)/nE\{K_h(x - X)\} - 1.$$

Using similar arguments as those in Devroye and Krzyzak (1989, p. 74) and Krzyzak (1991, p. 495), we obtain

$$P\left\{\int_Q |g_n(x) - f(x)|^2 \, d\mu > \varepsilon\right\} = P\left\{\int_Q \left(\frac{N}{D+1}\right)^2 d\mu > \varepsilon\right\} \le P\left\{\int_Q N^2 \, d\mu > \varepsilon/8\right\} + P\left\{\int_Q D^2 \, d\mu > \varepsilon/8M\right\}.$$

(*A*) For the case that $|Y| \le M$. Let us first consider the term involving N. We have, by Chebyshev's inequality,

$$P\left\{\int_Q N^2 \, d\mu > \varepsilon/8\right\} \le \frac{8}{\varepsilon} E\left\{\int_Q N^2 \, d\mu\right\} = \frac{8}{n\varepsilon} E\left\{\int_Q [Y - f(x)]^2 K_h^2(x - X) \Big/ E^2\{K_h(x - X)\} \, d\mu(x)\right\} + \frac{8}{n^2\varepsilon} E\left\{\sum_{i\ne j} \int_Q [Y_i - f(x)]K_h(x - X_i) \times [Y_j - f(x)]K_h(x - X_j) \Big/ E^2\{K_h(x - X)\} \, d\mu(x)\right\} = \frac{8}{\varepsilon}(A + B).$$

Here,

$$A \le \frac{4M^2k^*}{n} \int_Q \frac{1}{E\{K_h(x - X)\}} \, d\mu \le \frac{4M^2k^*\gamma}{nh^d}$$

where $k^* = \sup|K| < \infty$ by the regularity of K and $\gamma < \infty$ by the covering argument of Devroye and Krzyzak (1989, p. 79). The constant γ depends on d only. Moreover,

$$B \le \int_Q [E\{[Y - f(x)]K_h(x - X)\}/E\{K_h(x - X)\}]^2 \, d\mu.$$

Mimicking the proof of Lemma 6 of Krzyzak (1991, p. 495) we obtain the following upper bound for B

$$s_{\delta,h}^{2\alpha} \int_Q G^2(x) \, d\mu + \int_Q \frac{[\int_0^\delta \int_{A_{t,h}} |f(x) - f(y)|\mu(dy) \, dt]^2}{E^2\{K_h(x - X)\}} \mu(dx)$$

where $s_{\delta,h} = hK^+(\delta)$, and $A_{t,h}(x) = \{y : K_h(x - y) > t\}$. By using Schwartz's inequality and following the treatment similar to that used in Krzyzak (1991, p. 495), we can see that the second summand in the formula above is bounded from above by

$$\int_Q \frac{\delta \int_0^\delta \{\int_{A_{t,h}(x)} |f(x) - f(y)|\mu(dy)\}^2 \, dt}{E^2\{K_h(x - X)\}} \, d\mu(x) \le \int_Q \frac{\delta \int_0^\infty \int_{A_{t,h}(x)} [f(x) - f(y)]^2 \mu(dy)\mu[A_{t,h}(x)] \, dt}{E^2\{K_h(x - X)\}} \, d\mu(x) \le \int_Q \frac{\delta E\{K_h(x - X)\}[f(x) - f(X)]^2}{E^2\{K_h(x - X)\}} \, d\mu(x) \le 4M^2\delta\gamma h^{-d}.$$

Taking $\delta = \varepsilon h^d$, we get

$$B \le [hK^+(\varepsilon h^d)]^{2\alpha} \int G^2 \, d\mu + 4M^2\gamma\varepsilon = \varepsilon\left(\int G^2 \, d\mu + 4M^2\gamma\right),$$

with $[hK^+(\varepsilon h^d)]^{2\alpha} = \varepsilon$.

Collecting the results about A, B, we see that the convergence rate of $\int_Q N^2 \, d\mu(x)$ is $O[\max\{(nh^d)^{-1}, b_n\}]$ in probability.

Similarly, we can also prove that the convergence rate of $\int_Q D^2 \, d\mu(x)$ is also $O[\max\{(nh^d)^{-1}, b_n\}]$ in probability.

More specifically, when K has compact support, hence $b_n = h_n^{2\alpha}$, then the rate of L_2 convergence of $e^2(f, g_n)$ becomes $O(n^{-[2\alpha/(2\alpha+d)]})$ *in probability*.

(*B*) For the case that $E\{|Y|^s\} < \infty$, $s > 1$. Let $\bar{Y} = YI_{(|Y|<n^{1/s})}$ and $\bar{R}(x) = E\{\bar{Y}|X = x\}$. We then use a truncation argument [noticing the inequality $(a+b+c)^2 \le 4(a^2+b^2+c^2)$]

$$P\left\{\int [R_n(x) - f(x)]^2\mu(dx) > \varepsilon\right\} \le P\left\{\int A^2\, d\mu > \varepsilon/12\right\} + P\left\{\int B^2\, d\mu > \varepsilon/12\right\} + P\left\{\int C^2\, d\mu > \varepsilon/12\right\}$$

where

$$A = \sum_{i=1}^{n} (Y_i - \bar{Y}_i)K_h(x - X_i) \Big/ \sum_{i=1}^{n} K_h(x - X_i)$$

$$B = \sum_{i=1}^{n} [f(x) - \bar{R}(x)]K_h(x - X_i) \Big/ \sum_{i=1}^{n} K_h(x - X_i)$$

$$C = \sum_{i=1}^{n} [\bar{Y}_i - \bar{R}(x)]K_h(x - X_i) \Big/ \sum_{i=1}^{n} K_h(x - X_i).$$

where $A^2 = 0$ a.s. for n large enough. Because $P\{A^2 > c_n\varepsilon\} = 0$ as $n \to \infty$ for any sequence $c_n \to 0$, A will not affect the rate and need not be considered further.

By Jensen's inequality we have

$$B^2 \le \sum_{i=1}^{n} [f(x) - \bar{R}(x)]^2 K_h(x - X_i) \Big/ \sum_{i=1}^{n} K_h(x - X_i) = [f(x) - \bar{R}(x)]^2.$$

Furthermore, by Schwartz's inequality we have

$$B^2 \le [f(x) - \bar{R}(x)]^2 = (E\{Y - \bar{Y}|X = x\})^2 \le E\{Y^2 I_{(|Y|>n^{1/s})}|X = x\}.$$

Using Hölder's inequality, we have

$$\int B^2\, d\mu \le EY^2 I_{(|Y|>n^{1/s})} \le (E|Y|^s)^{2/s}(P(|Y| > n^{1/s})^{(1-2/s)} = BB.$$

Moreover, by Chebyshev's inequality we further have

$$BB \le (E|Y|^s)^{2/s}(E|Y|^s)^{1-2/s}/n^{1-2/s} \le E\{|Y|^s\}/n^{(s-2)/s}.$$

Using the results for the case (*A*), we get

$$\int C^2\, d\mu(x) = O(n^{2/s}\max\{(nh_n^d)^{-1}, b_n\}) \quad \text{in probability.}$$

Combining the bounds we obtain the rate

$$\max\{n^{(2/s)-1}, n^{2/s}h_n^{-[2\alpha/(2\alpha+d)]}\} = n^{(2/s)-(2\alpha/2\alpha+d)}. \quad \blacksquare$$

Proof of Theorem 7. This proof is contained in the proof of Theorem 19.

Proof of Theorem 8. This proof is contained in the proof of Theorem 20.

Proof of Theorem 9. Similar to the proof of Theorem 3. Let us randomly choose a subset $\mathcal{D}_n = \{X_i, Y_i\}_1^n$ among $\mathcal{D}_N$. For the RBF nets (2), we simply let the parameters c_i, w_i's of the estimator to be chosen be fixed at $c_i = X_i$, $w_i = Y_i$, $i = 1, \ldots, n$, and let the receptive field be specified hyperspherically with $\Sigma = h_n^2 I$. Then, we have an estimator $f_{n,N}(x) \in \hat{\mathcal{F}}_{n,N}$ that is same as $g_n(x) \in \hat{\mathcal{F}}_{n,N}$—the one given by the KRE eqn (7). So, $f_{n,N}(x)$ converges to $f(x)$ with the same rate as $g_n(x)$. Moreover, such a specific $f_{n,N}(x)$ may not be the best $\hat{f}^*_{n,N}(x)$. So the convergence rate of $\hat{f}^*_{n,N}(x)$ will not be worse than the rate of $g_n(x)$. Therefore, from Theorem 5(*A*)(*B*)(*C*) we can get Theorem 9(*A*)(*B*)(*C*). ■

Proof of Theorem 10, *Theorem* 11, and *Theorem* 12. The proofs are similar to the one above. Now we need to use Theorem 6, Theorem 7, and Theorem 8, respectively, instead of using Theorem 5. ■

Proof of Theorem 13. Let us consider the special case that X is a random variable with probability measure $\mu(x)$ and $Y = f(X)$. We randomly choose a subset $\{X_i\}_1^n$ and form a set $\{X_i, Y_i | Y_i = f(X_I)\}_1^n$. For the RBF nets (2), we simply let the parameters c_i, w_i's of the estimator to be chosen to be fixed at $c_i = X_i$, $w_i = Y_i$, $i = 1, \ldots, n$, and let the receptive field be specified hyperspherically with $\Sigma = h_n^2 I$. Then we have an $f_n(x) \in \mathcal{F}_n$ that is same as $g_n(x) \in \mathcal{F}_n$—the one given by the KRE eqn (7). So, $f_n(x)$ converges to $f(x)$ with the same rate as $g_n(x)$. Moreover, such a specific f_n may not be the best $\hat{f}^*_n$. So the convergence rate of $\hat{f}^*_n$ will not be worse than the rate of $g_n(x)$. So, from Theorem 5(*A*) we have Theorem 13(*A*), and from Theorem 5(*C*) we have Theorem 13(*B*). ■

Proof of Theorem 14. The proof is similar to the one above. Now we use Theorem 6 instead of Theorem 5. ■

Proof of Theorem 15. Let $p(x) = [d\mu(x)/dx]$, then we use Theorem 7 and follow the similar line as in the proof of Theorem 13. ■

Proof of Theorem 16. The proof is similar to the one above. Now we use Theorem 8 instead of Theorem 7. ■

Proof of Theorem 17. It follows from Lemma 1 that $E\{\varepsilon^2_{\text{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\} \le e_0^2 + e^2_{\text{KRE}}(f, g_n)$. Comparing eqns (16) and (17), we see that $e^2_{\text{KRE}}(f, g_n)$ is actually just the L_2 error $\rho^2_U(f, g_n)$. Because $e_0^2 = E\{|Y_i - f(X_i)|^2\}$ is irrelevant to N and n, the change of $E\{\varepsilon^2_{\text{RBF}}(\mathcal{D}_N, \hat{f}_{n,N})\}$ is determined by $\rho^2_U(f, g_n)$. From Theorem 6, we know that as $n \to \infty$, $N \to \infty$ with $N \ge n$, (i) if $|Y| \le M < \infty$, $\rho^2_U(f, g_n)$ converges to 0 *in probability* with rate $O(\max\{(nh^d)^{-1}, b_n\})$ or $O(\max\{(nh^d)^{-1}, b_n\})$ if ϕ has compact support; (ii) $E\{|Y|^s\} < \infty$, $s > 2 + d/\alpha$ and ϕ has compact support, $\rho^2_U(f, g_n)$ converges to 0 *in probability* with rate $O(n^{(2/s)-[2\alpha/(2\alpha+d)]})$. ■

Proof of Theorem 18. The proof is similar to the one above. Now we use Theorem 8 instead of Theorem 6. ■

Proof of Theorem 19. In the proof we will use the following decomposition (Krzyzak & Pawlak, 1987):

$$P\{|g_n(x) - f(x)| > \varepsilon\} \le P\{|a_n(x) - f(x)| > \delta\} + P\{|b_n(x) - 1| > \delta\}, \quad \text{(A.3)}$$

where $\delta = [\varepsilon/(\varepsilon + |f(x)| + 1)]$

$$a_n = \frac{1}{nh_n^d p(x)} \sum_{i=1}^{n} Y_i K_{h_n}(x - X_i)$$

$$b_n = \frac{1}{nh_n^d p(x)} \sum_{i=1}^{n} K_{h_n}(x - X_i)$$

$$K_{h_n}(x) = K(x/h_n).$$

We have by Chebyshev's inequality

$$P\{|a_n(x) - f(x)| > \delta\} \le \frac{1}{\delta} E\{|a_n(x) - f(x)|\}$$

$$\le \frac{1}{\delta}(E\{|a_n(x) - E\{a_n(x)\}|\} + |E\{a_n(x)\} - f(x)|)$$

$$\le \frac{1}{\delta}\{\text{var}[a_n(x)]^{1/2} + \text{bias}\}.$$

We first consider the variance

$$\text{var}[a_n(x)] = \frac{1}{n^2 h_n^{2d} p^2(x)} \sum_{i=1}^{n} \text{var}\{Y_i K_{h_n}(x - X_i)\}$$

$$= \frac{1}{nh_n^{2d}p^2(x)} \operatorname{var}\{Y_1 K_{h_n}(x - X_1)\}$$

$$\leq \frac{1}{nh_n^{2d}p^2(x)} E\{Y^2 K_h^2(x - X)\}.$$

We will show that as $h_n \to 0$

$$\frac{1}{h_n^d} E\{Y^2 K_{h_n}^2(x - X)\} \to \sigma^2(x)p(x) \int K^2(t)\, dt \qquad \text{(A.4)}$$

at points of continuity of $\sigma^2(x)$ and $p(x)$, where $\sigma^2(x) = E\{Y^2 | X = x\}$.

We have

$$\left|\frac{1}{h_n^d}\left[E\{Y^2\}K_{h_n}^2(x - X) - \sigma^2(x)p(x)\int K^2(t)\, dt\right]\right|$$

$$= \frac{1}{h_n^d}\left|\int [\sigma^2(y)p(y) - \sigma^2(x)p(x)]K_{h_n}^2(x - y)\, dy\right|$$

$$= \left|\frac{1}{h_n^d}\int_{\|x-y\|\leq\delta}(\cdot) + \frac{1}{h_n^d}\int_{\|x-y\|>\delta}(\cdot)\right| = |I + II|$$

By continuity of σ^2 and p we have:

$$|I| \leq \sup_{\|x-y\|\leq\delta} |\sigma^2(y)p(y) - \sigma^2(x)p(x)| \times \int K^2(t)\, dt \to 0 \quad \text{as} \quad \delta \to 0.$$

By the properties of K we have:

$$II \leq \sigma^2(x)p(x)\int_{\|x\|>\delta/h_n} K^2(x)\, dx + \sup_{\|x\|>\delta/h_n} |K(x)|\, E\{Y^2\} \to 0 \quad \text{as} \quad h_n \to 0.$$

We will now consider the bias term:

$$E\{a_n(x)\} - f(x)$$

$$= \frac{1}{p(x)}\frac{1}{h_n^d} E\{YK_h(x - X)\} - f(x)$$

$$= \frac{1}{p(x)h_n^d}\int f(y)K_{h_n}(x - y)p(y)\, dy - f(x)$$

$$= \frac{1}{p(x)}\int f(x + h_n y)K(y)p(x + h_n y)\, dy - f(x)$$

$$= \frac{1}{p(x)}\int \left[f(x) + \sum_{i=1}^{q-1}\frac{h_n^i}{i!}(y^T\nabla_x)^i f(x) + \frac{h_n^q}{q!}(y^T\nabla_x)^q f(x) + o(h_n^q)\right]\left[p(x) + \sum_{i=1}^{q-1}\frac{h_n^i}{i!}(y^T\nabla_x)^i p(x) + \frac{h_n^q}{q!}(y^T\nabla_x)^q p(x) + o(h_n^q)\right]K(y)\, dy - f(x)$$

$$= \frac{h_n^q}{p(x)}B(x) + o(h_n^q)$$

where $B(x) = \{(1/q!)\sum_{i=1}^d (f(x)[\partial^q p(x)/\partial x^{(i)q}] + p(x)[\partial^q f(x)/\partial x^{(i)q}]) + \sum_{i=1}^{q-1}(1/i!)[1/(q-i)!]\sum_{j=1}^d [\partial^i f(x)/\partial x^{(j)i}][\partial^{(q-i)}p(x)/\partial x^{(j)(q-i)}]\}\int z^q H(z)\, dz$ with $x = [x^{(1)}, \ldots, x^{(d)}]^t$.

By combining all the above formulae, we see that the convergence rate of $|a_n(x) - f(x)|$ is $O([A_1(x)/\sqrt{nh_n^d}] + B_1(x)h_n^q)$ in probability with $A_1(x) = ([E\{Y^2|X = x\}/p(x)]\int K^2(t)\, dt)^{1/2}$. Next, we consider the second term in eqn (A.3), that is, $P\{|b_n(x) - 1| > \delta\}$.

Because $|b_n(x) - 1| = [1/p(x)]|p(x) - p_n(x)|$ with $p_n(x)$ being the Parzen window estimator given by eqn (9), we know, from Theorem 21(A), that the convergence rate of $1/p(x)|p(x) - p_n(x)|$ is $O\{[A_2(x)/\sqrt{nh_n^d}] + B_2(x)h_n^q\}$; therefore our conclusion is proved.

In addition, by minimizing $O\{[A(x)/\sqrt{nh_n^d}] + B(x)h_n^q\}$ with respect to h_n, we will find this becomes $O(n^{-q/2q+d})$. That is, we get the proof for Theorem 7. ■

Proof of Theorem 20. Let us use the decomposition

$$NT = \frac{1}{nh_n^d p(x)}\sum_{i=1}^n [Y_i - f(x)]K_{h_n}(x - X_i)$$

$$DT = \frac{1}{nh_n^d p(x)}\sum_{i=1}^n K_{h_n}(x - X_i) - 1.$$

Using the same approach as in the proof of Theorem 4, we have

$$P\left\{\int_Q |R_n(x) - f(x)|^2\, d\mu > \varepsilon\right\} = P\left\{\int_Q \left(\frac{NT}{DT + 1}\right)^2 d\mu > \varepsilon\right\} \leq P\left\{\int_Q NT^2\, d\mu > \varepsilon/8\right\} + P\left\{\int_Q DT^2\, d\mu > \varepsilon/8M\right\}.$$

Let us first consider the term involving N. We have, by Chebyshev's inequality:

$$P\left\{\int_Q NT^2\, d\mu(x) > \varepsilon/8\right\} \leq \frac{8}{\varepsilon}E\left\{\int_Q NT^2\, d\mu(x)\right\}$$

$$= \frac{8}{nh_n^{2d}\varepsilon}E\left\{\int_Q [Y - f(x)]^2 K_{h_n}^2(x - X)\Big/ p^2(x)\, d\mu\right\} + \frac{8}{n^2h_n^{2d}\varepsilon}\sum_{i\neq j}E\left\{\int_Q [Y_i - f(x)]K_{h_n}(x - X_i)[Y_j - f(x)] \times K_{h_n}(x - X_j)\Big/ p^2(x)\, d\mu\right\} \leq \frac{8}{nh_n^{2d}\varepsilon} \times E\left\{\int_Q [Y - f(x)]^2 K_h^2(x - X)\Big/ p^2(x)\, d\mu\right\} + \frac{8}{h_n^{2d}\varepsilon}\int_Q \frac{E^2\{[f(X) - f(x)]K_h(x - X)\}}{p^2(x)}\, d\mu$$

$$= \frac{8}{\varepsilon}(I + II).$$

Next we have

$$I \leq \frac{4M^2}{nh_n^{2d}}\int_Q \frac{K_{h_n}^2(x - y)p(y)}{p^2(x)}p(x)\, dx\, dy = \frac{4M^2}{nh_n^d}\int_Q \frac{K^2(y)p(x + h_n y)}{p(x)}\, dx\, dy \leq \frac{4M^2}{cnh_n^d}\int K^2(y)p(x + h_n y)\, dx\, dy = \frac{4M^2}{cnh_n^d}\int K^2(y)\, dy.$$

Applying the results of the proofs of the bias part in Theorem 19 to II, we have

$$II = \frac{1}{h_n^{2d}}\int_Q \frac{E^2\{[f(X) - f(x)]K_h(x - X)\}}{p(x)}\, dx \times \int_Q \frac{1}{p(x)}\left[\int [f(x + h_n y) - f(x)\right] \times K(y)p(x + h_n y)\, dy]^2\, dx = h^{2p}\int_Q \frac{D_1^2(x)}{p(x)}\, dx + o(h_n^{2q}).$$

Let us now consider the term involving DT, we have

$$P\left\{\int_Q DT^2\,d\mu > \varepsilon/8M\right\} \le \frac{8M}{\varepsilon} E\left\{\int_Q DT^2\,d\mu\right\}$$
$$= \frac{8M}{\varepsilon} E\left\{\int_Q \left[\frac{1}{n}\sum_{i=1}^{n}\left[\frac{1}{p(x)h^d}K_{h_n}(x-X_i)-1\right]\right]^2 d\mu\right\}$$
$$= \frac{8M}{n\varepsilon} E\left\{\int_Q \left[\frac{K_{h_n}(x-X_i)}{p(x)h_n^d}-1\right]^2 p(x)\,dx\right\}$$
$$+ \frac{8M}{n^2\varepsilon}\sum_{i\neq j} E\left\{\int_Q \left[\frac{K_{h_n}(x-X_i)}{p(x)h_n^d}-1\right]\left[\frac{K_{h_n}(x-X_j)}{p(x)h_n^d}-1\right]p(x)\,dx\right\}$$
$$\le \frac{8M}{n\varepsilon} E\left\{\int_Q \left[\frac{K_{h_n}(x-X_i)}{h_n^d}-p(x)\right]^2\frac{1}{p(x)}\,dx\right\}$$
$$+ \frac{8M}{\varepsilon}\int_Q \left[\frac{E\{K_{h_n}(x-X)\}}{h_n^d}-p(x)\right]^2\frac{1}{p(x)}\,dx$$
$$= \frac{8M}{\varepsilon}(I+II).$$

Next

$$I \le \frac{1}{nh_n^{2d}}\int\int_Q (K_{h_n}(x-y)-h_n^d p(x))^2\frac{p(y)}{p(x)}\,dx\,dy$$
$$= \frac{1}{nh_n^d}\int\int_Q (K(x)-h_n^d p(x)^2\frac{p(x+h_n y)}{p(x)}\,dx\,dy$$
$$\le \frac{q}{nh_n^d}\left\{\int\int_Q \frac{K^2(y)p(x+h_n y)}{p(x)}\,dx\,dy\right.$$
$$\left. + \int\int_Q h_n^{2d}p(x)p(x+h_n y)\,dx\,dy\right\} \to$$
$$\frac{q}{nh_n^d c}\left[\int_Q K^2(t)\,dt + h_n^{2d}\right]$$
$$= \frac{q}{nh_n^d c}\int_Q K^2(t)\,d.$$

We conclude the proof by considering II. From the results of proofs of the bias part in Theorem 21 we have

$$II = \int_Q \left[\frac{E\{K_h(x-X)\}}{h_n^d}-p(x)\right]^2\frac{1}{p(x)}\,dx$$
$$= \int_Q \left[\int K(y)p(x+h_n y)\,dy - p(x)\right]^2\frac{1}{p(x)}\,dx$$
$$= h_n^{2q}\int_Q \frac{D_2^2(x)}{p(x)}\,dx.$$

Therefore, our conclusion is proved.

In addition, by minimizing $O[A(x)/\sqrt{nh_n^d}] + B(x)h_n^q)$ with respect to h_n, we will find this become $O(n^{-[2q/(2q+d)]})$. That is, we get the proof for Theorem 8. ■

Proof of Theorem 21.

(1) Proof of (A). (i) for $V_x(h_n)$, by Chebyshev's inequality, for any $\delta > 0$ we have

$$P\{|p_n(x) - E\{p_n(x)\}| > \delta\} \le \frac{1}{\delta}\operatorname{var}[p_n(x)]^{1/2} \quad (A.5)$$

we further have

$$\operatorname{var}[p_n(x)] = \frac{1}{n^2h_n^{2d}}\sum_{i=1}^{n}\operatorname{var}\left\{K\left(\frac{x-X_i'}{h_n}\right)\right\} \le \frac{1}{nh_n^{2d}}E\left\{K^2\left(\frac{x-X}{h_n}\right)\right\}, \quad (A.6)$$

$$\frac{1}{h_n^d}E\left\{K^2\left(\frac{x-X}{h}\right)\right\} = \frac{1}{h_n^d}\int K^2\left(\frac{x-y}{h_n}\right)p(y)\,dy \quad (\text{by } x-y = h_n t)$$
$$= \int K^2(t)p(x-h_n t)\,dt \to$$
$$p(x)\int K^2(t)\,dt, \quad \text{as} \quad h_n \to 0 \quad (A.7)$$

at points of continuity of $p(x)$ (see the proof of Theorem 19. By combining all the above formulae, we have that the convergence rate of $|p_n(x) - E\{p_n(x)\}|$ is $O[A(x)/\sqrt{nh_n^d}]$ in probability.

(ii) For $B_x(h_n)$, we have

$$E\{p_n(x)\} - p(x) = \frac{1}{h^d}E\left\{K\left(\frac{x-X}{h}\right)\right\} - p(x) = \frac{1}{h_n^d}\int p(y)K\left(\frac{x-y}{h_n}\right)dy - p(x)$$
$$= \int p(x+h_n y)K(y)\,dy - p(x) \quad (A.8)$$
$$= \int\left[p(x) + \sum_{i=1}^{q-1}\frac{h_n^i}{i!}(y^T\nabla_x)^i p(x)\right.$$
$$\left. + \frac{h_n^q}{q!}(y^T\nabla_x)^q p(x) + o(h_n^q)\right]K(y)\,dy - p(x)$$
$$= h_n^q B(x) + o(h_n^q)$$
$$\{\text{by using the condition } (b) \text{ about } K(x)\}. \quad (A.9)$$

(iii) By minimizing $G(h_n) = B_x(h_n) + V_x(h_n)$ with respect to h_n, we get $h_n = n^{-[1/(2q+d)]}$. Then substituting into $G(h_n)$ and recalling eqn (26), we see that the rate of $|p_n(x) - p(x)|$ is upper-bounded by $O(n^{-[q/(2q+d)]})$.

(2) Proof of (B). (i) for $V^2(h_n)$,

$$P\left\{\int V^2(h_n)\,dx > \delta\right\} \le \frac{1}{\delta}E\int |p_n(x) - E\{p_n(x)\}|^2\,dx$$
$$= \frac{1}{\delta}\int \operatorname{var}(p_n(x))\,dx$$
$$\le \frac{1}{\delta}\frac{1}{nh_n^d}\frac{1}{h_n^d}\int E\left\{K^2\left(\frac{x-X}{h}\right)\right\}dx.$$

By the approximation of identity properties (Wheeden & Zygmund, 1977, Theorem 9.6, p. 148) we obtain that as $h_n \to 0$,

$$\frac{1}{h_n^d}\int E\left\{K^2\left(\frac{x-X}{h}\right)\right\}dx \to \int K^2(t)\,dt.$$

(ii) For $B^2(h_n)$, it follows from eqn (A.9) that

$$B^2(h_n) = \int (E\{p_n(x)\} - p(x))^2\,dx$$
$$= \int [h_n^q B(x) + o(h_n^q)]^2\,dx = [h_n^{2q}B + o(h_n^{2q})] \quad (A.10)$$

where $B = \int B^2(x)\,dx$.

(iii) By minimizing $G(h_n) = B^2(h_n) + V^2(h_n)$ with respect to h_n, we get $h_n = n^{-[1/(2q+d)]}$ again. Putting it into $G(h_n)$, and recalling eqn (27), we see that the rate of $\int |p_n(x) - p(x)|^2\,dx$ is upper-bounded by $O(n^{-[2q/(2q+d)]})$. ■