CONF-971067--1

PAC Learning Using Nadaraya-Watson Estimator Based on Orthonormal Systems

Hongzhu Qiao
Department of Mathematics and Physics
Fort Valley State College
Fort Valley, GA 31030

Nageswara S.V. Rao, V. Protopopescu
Center for Engineering Systems Advanced Research
Oak Ridge National Laboratory
Oak Ridge, Tennessee 37831-6364
raons@ornl.gov

Paper submitted to the *Eighth International Workshop on Algorithmic Learning Theory*, Sendai, Japan, October 6-8, 1997.

†Research sponsored by the Engineering Research Program of the Office of Basic Energy Sciences, of the U.S. Department of Energy, under Contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.

PAC Learning Using Nadaraya-Watson Estimator Based on Orthonormal Systems*

Hongzhu Qiao[1], Nageswara S.V. Rao[2], V. Protopopescu[2]

[1] Department of Mathematics and Physics, Fort Valley State College, Fort Valley, GA 31030

[2] Oak Ridge National Laboratory, Oak Ridge, TN 37831-6364,
raons@ornl.gov,vvp@ornl.gov

Abstract. Regression or function classes of Euclidean type with compact support and certain smoothness properties are shown to be PAC learnable by the Nadaraya-Watson estimator based on complete orthonormal systems. While requiring more smoothness properties than typical PAC formulations, this estimator is computationally efficient, easy to implement, and known to perform well in a number of practical applications. The sample sizes necessary for PAC learning of regressions or functions under sup norm cost are derived for a general orthonormal system. The result covers the widely used estimators based on Haar wavelets, trignometric functions, and Daubechies wavelets.

1 Introduction

The problem of learning regressions or functions in the Probably Approximately Correct (PAC) framework of Valiant [32] continues to generate significant interest and activity [1, 3, 4, 2]. The ability to obtain sample sizes that ensure specified levels of precision and confidence is one of the main strengths of this paradigm. Recent results establish that a function which achieves small empirical error on an independently and identically distributed (iid) sample yields a PAC approximation under the finiteness of combinatorial parameters such as the fat-shattering index [1, 5], Euclidean parameters [31, 33], pseudo-dimension [14, 23], and capacity [34]. Smoothness properties such as piecewise differentiability [16], nth order continuous differentiability [21], and bounded variation [24] have also been used for obtaining PAC results.

The function estimation is a special case of the well-known non-linear regression problem studied in classical statistics [13, 25]. Typical results for regression estimators are asymptotic [30, 17] and are warranted by smoothness properties [22]. The appeal of such estimators stems from the ease of implementation and good performance in practical applications [7].

Recently, by combining smoothness and combinatorial (capacity) conditions, several specific statistical estimators based on Haar kernels have been shown

* Research sponsored by the Engineering Research Program of the Office of Basic Energy Sciences, of the U.S. Department of Energy, under Contract No. DE-AC05-96OR22464 with Lockheed Martin Energy Research Corp.

to provide PAC solutions for function estimation [26]. In this paper, we obtain PAC-style sample size estimates for the regression problem using the Nadaraya-Watson estimator [19] based on general orthogonal systems when: (a) the regression class is Euclidean [20, 33], and (b) the expansion coefficients of the marginal density and the product of regression and marginal density functions with respect to the orthonormal system satisfy mild decay conditions. The Euclidean class includes several well-known function classes such as VC graph class [11] and functions with finite pseudo-dimension [23]. Our approach is also applicable to more general regression classes with bounded scale-sensitive dimension [1].

Let (X, Y) be a random vector on $B \times \Re$, for compact $B \subset \Re$. Generalization of our results to higher dimensions can be done using existing methods (see [27, 12]). We denote random and deterministic variables by X and x, respectively. The *regression function* is $g(x) = E(Y|X = x)$. Let $\mathcal{L}^2(D)$ denote the Hilbert space of real square integrable functions defined on the set D, and let $h(.,.) \in \mathcal{L}^2(B \times \Re)$ and $f(.) \in \mathcal{L}^2(B)$ denote the density of X and Y, and the marginal density of X, respectively. Let $m(x) = \int yh(x, y)dy$ exist and be square integrable on B. Note that the regression is given by $g(x) = m(x)/f(x)$. Given the iid sample $(X_1, Y_1), (X_2, Y_2), \cdots, (X_n, Y_n)$, the regression problem in *random design* setting deals with estimating $g(x)$ from the sample. Such problems have been extensively studied in statistics, and more recently in machine learning [35, 1]. In this paper, we consider the classical Nadaraya-Watson estimator, based on a measurable orthonormal system $\{\phi_i | i = 1, 2, \ldots\}$ defined on $A \subseteq \Re$; the regression estimator is defined by

$$g_n(x) = \begin{cases} \dfrac{\sum_{k=1}^{s_n} \sum_{i=1}^{n} Y_i \phi_k(x) \phi_k(X_i)}{\sum_{k=1}^{s_n} \sum_{i=1}^{n} \phi_k(x) \phi_k(X_i)} & \text{if } \sum_{k=1}^{s_n} \sum_{i=1}^{n} \phi_k(x) \phi_k(X_i) \neq 0 \\ 0 & \text{elsewhere,} \end{cases}$$

where $s_n = \lceil n^{w_0} \rceil$, $w_0 \leq 1/2$. These estimators have been extensively studied [19, 12], and are known to perform well in practice. Rigorous results for these estimators, however, are in terms of asymptotic consistency [30, 10] or convergence rates [17, 12]. In fact, the same is true for most nonparametric regression estimators, with the possible exception of [27, 28], whose results can be used to derive samples sizes under certain smoothness conditions. Here we obtain sample size n that ensures

$$P\left(\sup_{x \in B} |g_n(x) - g(x)| > \epsilon\right) < \delta,$$

where P denotes the distribution of the sample $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$. The sample size is a function of ϵ, δ, and certain parameters of regressions and marginal densities. Due to the compactness of B, the above condition also implies

$$P\left(\int_{x \in B} |g_n(x) - g(x)| f(x) dx > \epsilon\right) < \delta$$

for the same sample size. This condition is often used in the PAC formulations of the function learning problem.

Additional motivation for our work stems from the computational complexity. In general, PAC results are associated with high computational complexity. For instance, when feedforward Heaviside networks are used as estimators, the computational problem is NP-complete [6]. In our case, however, the estimated function value, $g_n(x)$, at any x, can be computed in $O(n^{1+w_0})$ evaluations of $\phi_k(.)$; for some orthogonal systems (e.g. Haar wavelets) each evaluation can be done in $O(1)$ time. These computational properties of $g_n(.)$ are achieved at the expense of the following trade-offs: (i) the results are based on smoothness conditions for densities and regressions, and (ii) sample size estimates are less "compact" compared to usual PAC results. However, this makes the results more transparent since smoothness conditions are sometimes easy to visualize and quantify. The interpretation of the bounds is also easier since their dependence on various smoothness and combinatorial factors is more explicit. Furthermore, our results provide sample sizes for the estimator based on familiar orthonormal systems such as Haar wavelets, trignometric functions, and Daubechies wavelets.

In section 3, we present a result valid when the regression is chosen from a Euclidean class and satisfies certain smoothness conditions. Then we consider some interesting variations of this result in Section 4, where the orthonormal system itself is a Euclidean class, as is the case with trignometric system, Daubechies wavelets, and Chebyshev polynomials. Euclidean classes of Lipschitz functions are considered in Section 5.

2 Preliminaries

Let $\mathcal{A}$ be a collection of subsets of $\Re^d$. The trace $tr(S,\mathcal{A})$ of a set $S \subset \Re^d$ with respect to $\mathcal{A} \subset 2^{\Re^d}$ is defined as $tr(S,\mathcal{A}) = \{S \cap A | A \in \mathcal{A}\}$. For $|S| = n$ (here $|\cdot|$ denotes cardinality), we have $|tr(S,\mathcal{A})| \leq 2^n$. The growth function is defined by $\Pi_n(\mathcal{A}) = \max_{S \subset \Re,\ |S|=n} |tr(S,\mathcal{A})|$. Then $\mathcal{A}$ is called *VC class of dimension k* if k is the largest j such that $\Pi_j(\mathcal{A}) = 2^j$.

Let $\mathcal{C}(S)$ and $\mathcal{L}^\infty(S)$ denote the classes of continuous and essentially bounded functions defined on $S \subseteq \Re^d$, respectively. For $f \in \mathcal{L}^\infty(S)$, we have $\| f \|_\infty =$ ess sup$\{|f(x)| : x \in S\}$. The modulus of smoothness of $f \in \mathcal{L}^\infty(S)$ is defined as $\omega_\infty(f;r) = \sup_{|h|_\infty < r} (\text{ess sup}_{S(h)} |f(x+h) - f(x)|)$ where $S(h) = \{x \in S : x+h \in S\}$ and $|h|_\infty = \max(|h_1|, \ldots, |h_d|)$. A function $f \in \mathcal{C}(S)$ is called *Lipschitz* if there exist $0 < C < \infty$, $0 < \alpha \leq 1$ (called Lipschitz constant and exponent, respectively) such that $\omega_\infty(f,r) \leq Cr^\alpha$. We denote the class of such Lipschitz functions by $\mathcal{C}^\alpha(S)$.

The *graph* of a function f is defined as $graph(f) = \{(x,t) \in \Re^d \times \Re \mid 0 \leq t \leq f(x) \text{ or } f(x) \leq t \leq 0\}$. $\mathcal{F}$ is called a *VC graph class* if $\{graph(f) \mid f \in \mathcal{F}\}$ has finite VC dimension. If Q is a measure, we will use $Q(F)$ or simply QF to denote $\int F dQ$. For $1 \leq p < \infty$ and P a probability measure, the *covering number* of $\mathcal{F}$

is defined by

$$N_p(\epsilon, \mathcal{F}, P) = \min\left\{ m \mid \sup_{f\in\mathcal{F}} \min_{1\le i\le m} P|f - f_i|^p < \epsilon^p, \quad \{f_1, f_2, \cdots, f_m\} \subset \mathcal{F}^{\Re} \right\}.$$

For two functions $f, g : S \mapsto \Re$, we say $f \geq g$ if $f(x) \geq g(x)$ for all $x \in S$. The *envelope* of a function class $\mathcal{F}$ is a function satisfying $F \geq |f|$, for any $f \in \mathcal{F}$. Then $\mathcal{F}$ is defined as *Euclidean class* [20] with envelope F if there exist constants $C_{\mathcal{F}}$ and $V_{\mathcal{F}}$ (called Euclidean parameters) such that for any measure Q of finite support, we have $N_1(\epsilon QF, \mathcal{F}, Q) \leq C_{\mathcal{F}}\epsilon^{-V_{\mathcal{F}}}$. Each VC graph class is Euclidean with envelope $\sup_{f\in\mathcal{F}} |f|$, and each class of bounded functions with finite pseudo-dimension is also Euclidean [23, 20]. Like the VC-dimension, the Euclidean property is not immediately appealing to intuition. Metaphorically speaking, a class of functions is Euclidean if it contains elements that are sufficiently "well-behaved" and thus – in some sense – predictable.

Let $\mathcal{F} \cdot \mathcal{G} = \{fg | f \in \mathcal{F}, g \in \mathcal{G}\}$, and $f \cdot \mathcal{G} = \{fg | g \in \mathcal{G}\}$ for a given function $f \in \mathcal{F}$. The following Lemma is based on ideas from [23, 20].

Lemma 1. *(i) Assume $\mathcal{F}, \mathcal{G}$ are Euclidean with envelopes F, G, respectively. Then $\mathcal{F} \cdot \mathcal{G}$ has an envelope FG with parameters $C_{\mathcal{F}\cdot\mathcal{G}} = 2^{V_{\mathcal{F}}+V_{\mathcal{G}}} C_{\mathcal{F}} C_{\mathcal{G}}$ and $V_{\mathcal{F}\cdot\mathcal{G}} = V_{\mathcal{F}} + V_{\mathcal{G}}$.*

(ii) If $\mathcal{F}$ is Euclidean with envelope F, we have $N_2(\epsilon, \mathcal{F}, Q) \leq N_1\left(\frac{\epsilon^2}{2}, F \cdot \mathcal{F}, Q\right)$. Moreover, if $\max_x F(x) \leq \gamma_{\mathcal{F}}$, then $N_2(\epsilon, \mathcal{F}, Q) \leq C_{\mathcal{F}}\left(\frac{2\gamma_{\mathcal{F}}}{\epsilon^2}\right)^{V_{\mathcal{F}}}$.

Proof: Consider part (i). Let Q be a measure with finite support, and let λ, μ denote measures of densities F and G, respectively, with respect to Q. Let $m = N_1(\epsilon QF, \mathcal{F}, Q)$ and $n = N_1(\epsilon QG, \mathcal{G}, Q)$. Then for any $\epsilon > 0$ there exist $\{f_1, \cdots, f_m\}$ and $\{g_1, g_2, \cdots, g_n\}$ such that for any $f \in \mathcal{F}$, $g \in \mathcal{G}$, and for some i and j, we have $\lambda|f - f_i| < \epsilon\lambda F$, and $\mu|g - g_j| < \epsilon\mu G$, respectively. Observe that

$$\begin{aligned} Q(|fg - f_i g_j|) &\leq Q|f_i(g - g_j)| + Q|g(f - f_i)| \leq \lambda|g - g_j| + \mu|f - f_i| \\ &\leq \epsilon\lambda G + \epsilon\mu F = 2\epsilon QFG. \end{aligned}$$

There are at most mn different $f_i g_j$ in $\mathcal{F} \cdot \mathcal{G}$, and hence we have

$$N_1(2\epsilon QFG, \mathcal{F} \cdot \mathcal{G}, Q) \leq N_1(\epsilon QF, \mathcal{F}, Q) N_1(\epsilon QG, \mathcal{G}, Q),$$

which proves Part (i). Part (ii) follows from the inequalities $Q|f - f_i|^2 \leq 2QF|f - f_i| \leq 2\frac{\epsilon^2}{2} = \epsilon^2$, where $\{Ff_i\}$ is the cover for $F \cdot C_{\mathcal{F}}$ with covering number $N_1(\frac{\epsilon^2}{2}, F \cdot \mathcal{F}, Q)$. □

The following result follows from Talagrand [31] (also see van der Vaart and Wellner [33]).

Lemma 2. *Consider a class $\mathcal{F}$ of functions f such that $0 \leq f \leq 1$. Assume that for any given $\epsilon > 0$, and any probability Q on Ω that is supported on a compact*

set, we have $N_2(\epsilon, \mathcal{F}, Q) \leq \left(\frac{V}{\epsilon}\right)^v$, *where* V, v *are constants independent of* ϵ. *Then, for all* $M > 0$, *we have*

$$P\left(\sup_{f\in\mathcal{F}} |\frac{1}{n}\sum_{i=1}^{n} f(X_i) - Ef| \geq \epsilon\right) \leq K_{\mathcal{F}}\epsilon^v n^{v/2} e^{-2\epsilon^2 n},$$

where $K_{\mathcal{F}}(V, v) = \left(\frac{K(V)}{\sqrt{v}}\right)^v$ *with* $K(V)$ *specified in Talagrand [31].*

Proof: From [31], we have

$$P\left(\sup_{f\in\mathcal{F}} |\sum_{i=1}^{n} f(X_i) - nEf| \geq M\sqrt{n}\right) \leq \left(K(V)\frac{M}{\sqrt{v}}\right)^v e^{-2M^2},$$

from which the lemma follows. □

Lemma 3. *Suppose* a, b, c, d, *and* δ *are positive finite constants and* n *is a positive integer. Then the inequality* $an^b e^{-cn^d} \leq \delta$ *is satisfied for* $n \geq \omega(a, b, c, d, \delta)$, *where*

$$\omega(a, b, c, d, \delta) = \left[\max\left(1, \frac{2\ln\frac{a}{\delta}}{c}, \frac{(2b - cd)4b}{c^2d^2}\right)\right]^{1/d}.$$

Proof: If $n \geq \omega(a, b, c, \delta)$, then

$$\frac{cn^d}{2} \geq \ln\frac{a}{\delta}. \tag{2.1}$$

Moreover, since $n^d \geq \frac{(2b-cd)4b}{c^2d^2}$, by letting $t = \frac{cd}{2b}$, we have $n^d \geq \frac{(2b-2bt)4b}{4b^2t^2} = \frac{2(1-t)}{t^2}$ and $\frac{t^2n^{2d}}{2} \geq (1-t)n^d$. It follows that $e^{tn^d} \geq tn^d + \frac{t^2n^{2d}}{2} \geq n^d$. Therefore $tn^d \geq \ln n^d$, implying $\frac{cdn^d}{2b} \geq \ln n^d$, or

$$\frac{cn^d}{2} \geq \frac{b}{d}\ln n^d = b\ln n. \tag{2.2}$$

Combining (2.1) and (2.2), we have $cn^d = \frac{cn^d}{2} + \frac{cn^d}{2} \geq b\ln n + \ln\frac{a}{\delta}$. Thus, $\ln\delta \geq \ln a + b\ln n - cn^d$, yielding $an^b e^{-cn^d} \leq \delta$. □

3 Main Result

Let $\{\phi_k : k = 1, 2, \ldots\}$ be an orthonormal system defined on $A \subseteq \Re$ such that:
I. $\max_{x\in A} |\phi_k(x)| \leq u_2 k^{w_2}$ for all k, and some finite $w_2 \in \Re$, $u_2 > 0$.

Let $\mathcal{F} = \{f\}$ and $\mathcal{M} = \{m\}$ denote sets of functions in $\mathcal{L}^2(A)$ with compact support $B \subseteq A$, and $\mathcal{G} = \{g = m/f : f \in \mathcal{F}, m \in \mathcal{M}\}$ satisfy the following conditions:

IIa $\mathcal{G}$ is Euclidean with L^1-integrable envelope $G \leq 1$ and parameters $(C_{\mathcal{G}}, V_{\mathcal{G}})$.

IIb $\min_{x \in B} |f(x)| \geq u > 0$, for $f \in \mathcal{F}$, where u is a constant.

IIc The functions $f \in \mathcal{F}$ and $m \in \mathcal{M}$ satisfy, for some $\eta_1, \eta_2, C_1, C_2 > 0$

$$\left\| \sum_{k=n+1}^{\infty} a_k \phi_k(x) \right\|_{\infty} \leq C_1 (\ln n)^{-\eta_1}, \quad \text{and} \quad \left\| \sum_{k=n+1}^{\infty} b_k \phi_k(x) \right\|_{\infty} \leq C_2 (\ln n)^{-\eta_2}$$

where $a_k = \int f(t)\phi_k(t)dt = E\phi_k$ and $b_k = \int m(t)\phi_k(t)dt = E(Y\phi_k) = E(g\phi_k)$.

The condition **I** specifies that the magnitude of the elements of the orthonormal system must not increase faster than a polynomial in the index variable. The condition **IIa** specifies that the regression class be Euclidean; in spirit, this condition is similar to specifying the finiteness of capacity or graph dimension used in PAC paradigms. Euclidean class is not the weakest function class that is learnable, but our approach can be applied to more general classes (see Reamrk 4.1). The condition **IIb** specifies that the marginal density be bounded away from zero. The condition **IIc** relates the function classes $\mathcal{F}$ and $\mathcal{M}$ to the orthonormal system in that each function must be expressible in terms of the orthonormal system with decaying coefficients. Essentially, the conditions **I** and **IIa**-c guarantee that the regressions to be estimated and the orthonormal systems used to represent the regressions are reasonable enough both in terms of smoothness and combinatorial parameters.

Compared to the distribution-free results typical in the PAC paradigm, additional smoothness is required here both on marginal densities (which are assumed to exist) and regressions. Conditions such as **IIa** (or weaker forms, see Remark 4.1) are usual for the PAC paradigm [5, 2], while **I**, **IIb**-c are typical for the statistical paradigm [28, 17].

Theorem 4. *Let $\{\phi_k\}$ be an orthonormal system satisfying condition* **I**. *If function classes $\mathcal{F}$ and $\mathcal{G}$ satisfy conditions* **IIa** *through* **IIc**, *then for any $\delta > 0$ and $\epsilon > 0$ we have*

$$P\left(\sup_{x \in B} |g_n(x) - g(x)| > \epsilon \right) < \delta,$$

for sample size $n \geq \max(N_{11}, N_{12}, N_{21}, N_{22}, N_{31}, N_{33})$ with N_{j2} of form $e^{(a/b)^{1/c}}$ and N_{j1} of form $\omega(a,b,c,d,e) = \left[\max\left(1, 2/c \ln \frac{a}{e}, \frac{(2b-cd)4b}{c^2 d^2}\right)\right]^{1/d}$ with the following parameters

	a	b	c	d	e		a	b	c
N_{11}	$K_G(\sqrt{2}C_G^{\frac{1}{2V_G}}, 2V_G)\epsilon_1^{2V_G}$	$V_G w$	$2\epsilon_1^2$	w	$\delta/3$	N_{12}	C_2	ϵ_1	$\eta_2 w_0$
N_{21}	18	$1+2w_0$	$\epsilon_1^2/4$	$1-w_0(2+3w_2)$	$\delta/3$	N_{22}	C_1	ϵ_1	$\eta_1 w_0$
N_{31}	18	$1+2w_0$	$\epsilon^2/4$	$1-w_0(2+3w_2)$	$\delta/3$	N_{32}	$2C_1$	ϵ	$\eta_1 w_0$

where $s_n = n^{w_0}$, $0 < w_0 \leq 1/2$, $w = 1 - 2w_0(1+w_2)$, and $\epsilon_1 = \frac{\epsilon(u-\epsilon)}{4}$.

Proof: Let $m_n(x) = \frac{1}{n}\sum_{k=1}^{s_n}\sum_{i=1}^{n} Y_i\phi_k(x)\phi_k(X_i)$ and $f_n(x) = \frac{1}{n}\sum_{k=1}^{s_n}\sum_{i=1}^{n} \phi_k(x)\phi_k(X_i)$ such that $g_n(x) = m_n(x)/f_n(x)$. Nadaraya's decomposition inequality yields [18]:

$$
\begin{aligned}
&P\left(\sup_{x\in B}|g_n(x)-g(x)|\geq\epsilon\right)\\
&=P\left(\sup_{x\in B}|m_n(x)-m(x)|\geq\frac{\epsilon(u-\epsilon)}{2}\right)+P\left(\sup_{x\in B}|f_n(x)-f(x)|\geq\frac{\epsilon(u-\epsilon)}{2}\right)\\
&\quad+P(\sup_{x\in B}|f_n(x)-f(x)|>\epsilon)=I_1+I_2+I_3\,.
\end{aligned}
$$

Writing

$$
\begin{aligned}
m_n(x)-m(x) &= m_n(x)-Em_n(x)+Em_n(x)-m(x)\\
&=\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\phi_k(X_i)-E(Y\phi_k)\right)\phi_k(x)+\sum_{k=s_n+1}^{\infty}b_k\phi_k(x),
\end{aligned}
$$

we estimate the first term, I_1, as

$$
\begin{aligned}
&P\left(\sup_{x\in B}|m_n(x)-m(x)|>2\epsilon_1\right)\\
&\leq P\left(\sup_{x\in B}\left|\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\phi_k(X_i)-E(Y\phi_k)\right)\phi_k(x)\right|>\epsilon_1\right)\\
&\quad+P\left(\sup_{x\in B}\left|\sum_{k=s_n+1}^{\infty}b_k\phi_k(x)>\right|\epsilon_1\right)\\
&=I_{11}+I_{12}.
\end{aligned}
$$

The term I_{12} can be made zero when n is large enough such that $s_n = n^{w_0} \geq e^{\left(\frac{c_2}{\epsilon_1}\right)^{\frac{1}{\eta_2}}}$ which yields the expression for N_{12}. Now, for $0<\varepsilon<\epsilon_1$, we have

$$
\begin{aligned}
&P\left(\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\phi_k(X_i)-E(Y\phi_k)\right)\phi_k(x)>\epsilon_1\right)\\
&\leq P\left(\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n}[Y_i-g(X_i)]\phi_k(X_i)\right)\phi_k(x)>\varepsilon\right)\\
&\quad+P\left(\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n}g(X_i)\phi_k(X_i)-E(Y\phi_k)\right)\phi_k(x)>\epsilon_1-\varepsilon\right).
\end{aligned}
$$

For any $\varepsilon>0$, the first term is upperbounded by

$$
P\left(\left|\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n}[Y_i-g(X_i)]\phi_k(X_i)\right)\phi_k(x)\right|>\varepsilon\right)
$$

which is in turn upperbounded by

$$\sum_{k=1}^{s_n} P\left(\left|\left(\frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]\phi_k(X_i)\right)\phi_k(x)\right| > \varepsilon/s_n\right)$$
$$\leq \frac{s_n}{\varepsilon}\sum_{k=1}^{s_n} E\left[\left|\left(\frac{1}{n}\sum_{i=1}^{n}[Y_i - g(X_i)]\phi_k(X_i)\right)\phi_k(x)\right|\right],$$

where the last step is due to Chebyschev's inequality. Now each term under the expectation is zero, and hence the sum is zero. Thus for $n > N_{12}$ and $\Phi_{s_n} = \{\phi_1, \phi_2, \ldots, \phi_{s_n}\}$, we have

$$I_1 = I_{11} \leq P\left(\sup_{x\in B}\left|\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n} Y_i\phi_k(X_i) - E(Y\phi_k)\right)\phi_k(x)\right| > \epsilon_1\right)$$
$$\leq P\left(\sup_{x\in B}\left|\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\phi_k(X_i) - E(Y\phi_k)\right)\phi_k(x)\right| > \epsilon_1\right)$$
$$\leq P\left(\left|\sum_{k=1}^{s_n}\left(\frac{1}{n}\sum_{i=1}^{n} g(X_i)\phi_k(X_i) - E(Y\phi_k)\right)\right| > \epsilon_1/A_{s_n}\right)$$
$$\leq s_n P\left(\sup_{\phi\in\Phi_{s_n}}\sup_{g\in\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i)\phi(X_i) - E(g\phi)\right| > \epsilon_1/s_n^{1+w_2}\right)$$
$$\leq s_n^2 \sup_{\phi\in\Phi_{s_n}} P\left(\sup_{g\in\phi\mathcal{G}}\left|\frac{1}{n}\sum_{i=1}^{n} g(X_i)\phi(X_i) - E(g\phi)\right| > \epsilon_1/s_n^{1+2w_2}\right).$$

From Lemma 1, $\phi\mathcal{G}$ is Euclidean with parameters $(2^{V_G}C_G, V_G)$ with envelope $\phi\mathcal{G}$, which yields $N_2(\epsilon, \phi\mathcal{G}, Q) \leq \left(2C_G^{\frac{1}{2V_G}}/\epsilon\right)^{2V_G}$. Thus by Lemmas 2 and 3, we have, for $n \geq N_{11}$,

$$I_1 \leq K_G(2C_G^{\frac{1}{2V_G}}, 2V_G)\epsilon_1^{2V_G} n^{V_G(1-2w_0(1+w_2))} e^{-2\epsilon_1^2 n^{1-2w_0(1+w_2)}} \leq \delta/3.$$

The treatment of the terms I_2 and I_3 is similar, and we consider I_3. For $n \geq N_{32} = e^{\left(\frac{c_1}{\epsilon}\right)^{\frac{1}{w_1 w_0}}}$, we have

$$\begin{aligned} I_3 &\leq P\left(\sup_{x\in B}|f_n(x) - Ef_n(x)| > \epsilon\right) \\ &\leq P\left(\sup_{x\in B}\left|\sum_{i=1}^{s_n}\frac{1}{n}\sum_{i=1}^{n}\phi_k(X_i)\phi_k(x) - E(\phi_k)\phi_k(x)\right| > \epsilon\right) \\ &\leq P\left(\sum_{i=1}^{s_n}\left|\frac{1}{n}\sum_{i=1}^{n}\phi_k(X_i) - E(\phi_k)\right| > \epsilon/s_n^{w_2}\right) \\ &\leq s_n P\left(\sup_{\phi\in\Phi_{s_n}}\left|\frac{1}{n}\sum_{i=1}^{n}\phi(X_i) - E(\phi)\right| > \epsilon/s_n^{1+w_2}\right). \end{aligned}$$

The last term is upperbounded by $18n^{1+2w_0}e^{-\epsilon^2/4n^{1-w_0(2+3w_2)}}$ since the supremum is taken over a finite set of functions uniformly bounded by $s_n^{w_2} = n^{w_0 w_2}$ [34]. Then Lemma 3 yields the expression for N_{31}. □

We now describe well-known examples from harmonic analysis, where conditions **I** and **II**c are extensively investigated. Note that the additional conditions **II**a-b are needed for sample estimates based on Theorem 4.

Example 3.1: When the trignometric system is used for $\{\phi_k\}$, $w_2 = 0$, $A = \Re$, and condition **II**c is satisfied for Lipschitz functions ([36], p. 61). Since $w_2 = 0$, a simpler formulae can be obtained for the sample sizes of Theorem 4. By choosing $w_0 = 1/4$, we have $w = 1/2$. For simplicity assume that $\mu < 1$ and $\epsilon < 1$, which implies that $\epsilon_1 \leq \epsilon$. Let $L_G = \max\left(18, K_G(\sqrt{2}C_G^{\frac{1}{2V_G}}, 2V_G)\epsilon_1^{2V_G}\right)$. Then we have the following simpler form for the sample size

$$\frac{1}{\epsilon_1^4}\left[\ln(3/\delta) + \ln L_G\right]^2$$

when s_n is chosen to be $e^{\frac{2\max(C_1,C_2)}{\epsilon_1}\max(\eta_1,\eta_2)}$. Compared to typical PAC estimates, this sample size is higher since it is proportional to: (a) $1/\epsilon_1^4$ as compared to the usual $1/\epsilon^2$, and (b) the square of $\ln L_G$ as compared to the linear dependence on a similar term (for example, based on capacity or graph dimension). On the other hand, the estimated function value can be computed in $O(n^{3/2})$ time. Note that the computational problem of minimizing empirical error required by PAC methods could be intractable.

Example 3.2: For Haar wavelets, we have $w_2 = 1/2$, $A = B = [0,1]$, [12], and condition **II**c holds for any function f with $\omega_\infty(f,r) = O(r^\alpha)$, $0 < \alpha \leq 1$ [8]. The specific properties of the Haar system have been utilized in [26] for sample size estimates, whereas Theorem 4 is more general.

Example 3.3: For Legendre polynomials, we have $w_2 = 1/2$ [28]. Let $h(x)$ be integrable on $[-1,1]$ with bounded variation. Then functions of the form $f(x) = f(-1) + \int_{-1}^{x} h(x)dx$ satisfy condition **II**c (Jackson's Theorem [29]).

4 Variations

Consider the following conditions:

Ia $\max_x |\phi_k(x)| = A_k$, where $u_1 k^{w_1} \leq A_k \leq u_2 k^{w_2}$, and $u_1 > 0$, $u_2 > 0$, $w_1 \leq 0$, $w_2 \geq w_1$.

Ib $\Phi = \{\phi_k/A_k, k = 1, 2, \cdots\}$ is Euclidean with $\mathcal{L}^1$-integrable envelope H and parameters (C_Φ, V_Φ).

IIa' $\mathcal{F}, \mathcal{G}$ are Euclidean with $\mathcal{L}^1$-integrable envelopes F, G and parameters $(C_\mathcal{F}, V_\mathcal{F})$, $(C_\mathcal{G}, V_\mathcal{G})$, respectively.

Theorem 5. *Let $\{\phi_k\}$ be an orthonormal system satisfying conditions* **I**a *and* **I**b. *If function classes $\mathcal{F}, \mathcal{G}$ satisfies conditions* **II**a' *and* **II**b-c, *then for any $\delta > 0$, $\epsilon > 0$ we have*

$$P\left(\sup_{x\in B} |g_n(x) - g(x)| > \epsilon\right) < \delta,$$

for sample size $n \geq \max(N_{11}, N_{12}, N_{21}, N_{22}, N_{31}, N_{33})$, *with* N_{j2} *of form* $e^{(a/b)^{1/c}}$ *and* N_{j1} *of form* $\omega(a,b,c,d,e) = \left[\max\left(1, 2/c \ln \frac{a}{e}, \frac{(2b-cd)4b}{c^2 d^2}\right)\right]^{1/d}$ *with the following parameters*

	a	b	c	d	e
N_{11}	$\frac{K_{\mathcal{G}\cdot\Phi}\epsilon_2^{2V_{\mathcal{G}\cdot\Phi}}}{u_1^{2V_{\mathcal{G}\cdot\Phi}}}$	$V_{\mathcal{G}\cdot\Phi} - 2w_0w_1V_{\mathcal{G}\cdot\Phi} + w_0$	$\frac{2\epsilon_2^2}{u_2^2}$	$1-2w_0(w_0+w_2)$	$\delta/3$
N_{21}	$\frac{K_{\Phi}\epsilon_2^{2V_{\Phi}}}{u_1^{2V_{\Phi}}}$	$V_{\Phi} - 2w_0w_1V_{\Phi} + w_0$	$\frac{2\epsilon_2^2}{u_2^2}$	$1-2w_0(w_0+w_2)$	$\delta/3$
N_{31}	$\frac{K_{\Phi}\epsilon_2^{2V_{\Phi}}}{u_1^{2V_{\Phi}}}$	$V_{\Phi} - 2w_0w_1V_{\Phi} + w_0$	$\frac{2\epsilon^2}{u_2^2}$	$1-2w_0(w_0+w_2)$	$\delta/3$

	a	b	c
N_{12}	$2C_2$	ϵ_2	$\eta_2 w_0$
N_{22}	$2C_1$	ϵ_2	$\eta_1 w_0$
N_{32}	$2C_1$	ϵ	$\eta_1 w_0$

where $\epsilon_2 = \frac{\epsilon(u-\epsilon)}{2}$, *and* $s_n = n^{w_0}$ *such that* $0 < w_0 < \sqrt{\frac{1}{2} + \frac{1}{4}w_2^2} - \frac{w_2}{2}$.

To prove Theorem 5, we need the following lemma.

Lemma 6. *Let* $\mathcal{F}$ *denote a Euclidean class of function with envelope* F *bounded by 1. For* $f_k \in \mathcal{F}$, *we have*

$$P\left(\sum_{k=1}^{s_n}\left|\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - Ef_k\right| A_k > \epsilon\right) \leq K_{\mathcal{F}}\, n^{V_{\mathcal{F}}}\epsilon^{2V_{\mathcal{F}}}\sum_{k=1}^{s_n} A_k^{-2V_{\mathcal{F}}} e^{-2n\left(\frac{\epsilon}{s_k A_k}\right)^2}.$$

Proof: Noting

$$P\left(\sum_{k=1}^{s_n}\left|\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - Ef_k\right| A_k > \epsilon\right) \leq \sum_{k=1}^{s_n} P\left(\left|\frac{1}{n}\sum_{i=1}^{n} f_k(X_i) - Ef_k\right| > \frac{\epsilon}{s_k A_k}\right),$$

the lemma follows from Lemma 2. □

Proof of Theorem 5: The proof is similar to Theorem 4 except for details of the bounds for N_{11}, N_{21} and N_{31}. For $n > N_{12}$, by using Lemma 1, we have

$$\begin{aligned}
I_1 = I_{11} &\leq K_{\mathcal{G}\cdot\Phi}\, n^{V_{\mathcal{G}\cdot\Phi}}\epsilon_2^{2V_{\mathcal{G}\cdot\Phi}}\sum_{k=1}^{s_n} A_k^{-2V_{\mathcal{G}\cdot\Phi}} e^{-2n\left(\frac{\epsilon_2}{s_k A_k}\right)^2} \\
&\leq K_{\mathcal{G}\cdot\Phi}\epsilon_2^{2V_{\mathcal{G}\cdot\Phi}} n^{V_{\mathcal{G}\cdot\Phi}}\sum_{k=1}^{n^{w_0}}\left(\frac{1}{u_1 k^{w_1}}\right)^{2V_{\mathcal{G}\cdot\Phi}} e^{-2n\frac{\epsilon_2^2}{u_2^2 k^{2w_0+2w_2}}} \\
&\leq \frac{K_{\mathcal{G}\cdot\Phi}\epsilon_2^{2V_{\mathcal{G}\cdot\Phi}}}{u_1^{2V_{\mathcal{G}\cdot\Phi}}} n^{V_{\mathcal{G}\cdot\Phi}}\sum_{k=1}^{n^{w_0}} k^{-2V_{\mathcal{G}\cdot\Phi}w_1} e^{-\frac{2n\epsilon_2^2}{u_2^2 k^{2w_0+2w_2}}} \\
&\leq \frac{K_{\mathcal{G}\cdot\Phi}\epsilon_2^{2V_{\mathcal{G}\cdot\Phi}}}{u_1^{2V_{\mathcal{G}\cdot\Phi}}} n^{V_{\mathcal{G}\cdot\Phi}-2V_{\mathcal{G}\cdot\Phi}w_0w_1+w_0} e^{-\frac{2\epsilon_2^2}{u_2^2}\cdot n^{1-2w_0(w_0+w_2)}}.
\end{aligned}$$

By Lemma 3, $I_1 < \frac{\delta}{3}$, for $n > \max(N_{11}, N_{12})$, where $N_{12} = \omega(a, b, c, d, \delta/3)$, and

$$a = \frac{K_{\mathcal{G}\cdot\Phi}\epsilon_2^{2V_{\mathcal{G}\cdot\Phi}}}{u_1^{2V_{\mathcal{G}\cdot\Phi}}}, \quad b = V_{\mathcal{G}\cdot\Phi} - 2w_0w_1V_{\mathcal{G}\cdot\Phi} + w_0, \quad c = \frac{2\epsilon_2^2}{u_2^2}, \quad d = 1 - 2w_0(w_0 + w_2).$$

For I_3 (I_2 can be similarly handled), we have

$$I_3 \leq K_\Phi \; n^{V_\Phi}\epsilon^{2V_\Phi} \sum_{k=1}^{s_n} A_k^{-2V_\Phi} e^{-2n\left(\frac{\epsilon}{s_k A_k}\right)^2} < \frac{\delta}{3}$$

for $n = \max(N_{31}, N_{32})$, where $N_{31} = \exp\left(\left(\frac{2C_1}{\epsilon}\right)^{\frac{1}{\eta_1 w_0}}\right)$ and $N_{32} = \omega\left(a, b, c, d, \delta/3\right)$ with parameters specified in the statement of the theorem. □

Remark 4.1: Condition **II**a can be relaxed in Theorem 4, namely: $\mathcal{G}$ with envelope $G \leq 1$ has finite P_ρ-dimension [1]. A different expression for N_{11} must be derived in this case by using the sample size estimate of [1].

Remark 4.2: A generalization of Theorem 5 can be obtained by eliminating condition **I**b and replacing **II**a' by **II**a, along the lines of Theorem 4. Conditions **I**a and **I**b, however, are satisfied by a number of orthonormal systems, which results in the above compact form for the sample size estimates.

Lemma 7. *The following orthonormal systems are Euclidean with parameters (C, 4):*
(a) trignometric system $\{\sin nx, \cos nx\}$ *on* $[-\pi, \pi]$*;*
(b) Daubechies wavelets on $\Re$*; and*
(c) Chebyshev polynomials, $T_n(x)$*, on* $[-1, 1]$.

Proof: Noting that $T_n(x) = \cos(n \arccos x))$, (a)-(c) follow from Lemma 22 of [20] because $\sin x$, $\cos x$ and Daubechies' mother wavelet [9] are of bounded variation. Furthermore, we can obtain $N_1(\epsilon, \mathcal{F}, Q) \leq C\epsilon^{-4}$, for function classes (a)-(c). □

5 Lipschitz Functions

In this section, we show that condition **II**c is satisfied for Lipschitz functions for several orthonormal systems. Recall that for trignometric system and Haar wavelets condition **II**c holds for Lipschitz functions, when A is $[-\pi, \pi]$ and $[0, 1]$, respectively (Examples 3.1 and 3.2).

For Lipschitz functions, we now show that condition **II**c is satisfied by the Daubechies wavelets $\{\phi_{j,k}\}$, generated by the scaling function ϕ (details can be found in [9]).

Lemma 8. *For any* $f \in \mathcal{C}^\alpha(\Re)$*, there exists* $C_3 > 0$ *such that*

$$\| f(x) - \sum_k c_{j,k}\phi_{j,k}(x) \|_\infty = \| f(x) - \sum_k c_{j,k}2^{j/2}\phi(2^j x - k) \|_\infty \leq C_3 2^{-j\alpha}$$

where $c_{j,k} = 2^{j/2} \int f(x)\phi(2^j x - k)dx$.

Proof: Let $b_{j,k} = 2^j \int_{2^{-j}k}^{2^{-j}(k+1)} f(x)dx$. From [15] for any $f \in C^\alpha(\Re)$, there exists $C_1 > 0$ such that $\| f(x) - \sum_{k\in Z} b_{j,k}\phi(2^j x - k) \|_\infty \leq C_1 2^{-j\alpha}$. Then we have

$$|f(x)\phi(2^j x - k)dx - \int f(k2^{-j})\phi(2^j x - k)dx|$$
$$\leq |C \int |x - 2^{-j}k|^\alpha \phi(2^j x - k)dx| = |C \int 2^{-\alpha j - j}|y|^\alpha \phi(y)dy \leq C_0 2^{-(\alpha+1)j}.$$

Notice that $\int f(k2^{-j})\phi(2^j x - k)dx = f(k2^{-j})2^{-j}$ and

$$\left| \int_{2^{-j}k}^{2^{-j}(k+1)} f(x)dx - f(k2^{-j})2^{-j} \right| \leq \int_{2^{-j}k}^{2^{-j}(k+1)} C|x - k2^{-j}|dx \leq C2^{-2j}.$$

Therefore $|\int f(x)\phi(2^j x - k)dx - \int_{2^{-j}k}^{2^{-j}(k+1)} f(x)dx| \leq C_2 2^{-j(1+\alpha)}$ i.e. $|2^{j/2}c_{j,k} - b_{j,k}| \leq C_2 2^{-j\alpha}$.

Thus, $\| f(x) - \sum_k c_{i,k} 2^{j/2}\phi(2^j x - k) \|_\infty$ is upper bounded by

$$\| f(x) - \sum_k b_{j,k}\phi(2^j x - k) \|_\infty + \| \sum_k (b_{j,k} - c_{j,k}2^{j/2})\phi(2^j x - k) \|_\infty$$
$$\leq C_1 2^{-j\alpha} + C_2 2^{-j\alpha} \| \sum_k |\phi(2^j x - k)| \|_\infty \leq C_3 2^{-j\alpha}. \quad \square$$

For functions with compact support, it is convenient to replace the two indices j, k by a single index, n. For each $j \in Z_+ \cup \{0\}$, $k \in Z$, let us define $t_0 = 0$, $t_k = 2|k| - \frac{|k|}{2k} - 1/2$, $k \geq 1$ and $n = \frac{(j+t_k)(j+t_k+1)}{2} + t_k + 1$ (see the table below). It is easy to prove that these relationships establish a one-to-one correspondence between $(Z_+ \cup \{0\}) \times Z$ and Z_+.

$j\backslash k$	0	1	-1	2	-2	3	-3	...
0	1	3	6	10	15	...		
1	2	5	9	14	...			
2	4	8	13	...				
3	7	12	...					
4	11	...						
5	...							
t_k	0	1	2	3	4	5	6	...

Lemma 9. *With the definitions above, if f is Lipschitz with support in $[-1, 1]$, we have*

$$\| f(x) - (P_n f)(x) \|_\infty \leq C_4 2^{-j\alpha} \leq C_5 n^{-\alpha/2},$$

for some $C_4, C_5 > 0$, where P_n is the wavelet approximation.

Proof: Since f has support in $[-1,1]$, the case $|k| > 2^j$ is uninteresting. If k is cut off at 2^j, then $t_k \approx 2|k|$ and $n \approx 2k^2 \approx 2^{2j}$. Thus, using the lemmas above $\| f - P_n f \|_\infty \approx 2^{-j\alpha} \approx n^{-\alpha/2}$ $\square$.

Remark 5.1: In this paper, we consider batch PAC formulation under smooth densities with sup norm cost, whereas [16] considered distribution-free on-line formulation under $\mathcal{L}^2$-norm for piecewise twice-differentiable continuous functions.

6 Conclusions

Euclidean classes of functions and regression with compact support and certain smoothness properties are shown to be PAC learnable. The Nadaraya-Watson estimator based on complete orthonormal systems is employed to learn the regressions or functions. The results require more smoothness properties than typical PAC formulations, but, offer computationally efficiency. Furthermore, this estimator is known to perform well in a number of practical applications. Although well-studied in statistics, the available results on Nadaraya-Watson estimator only specify asymptotic consistency or convergence rates. By combining the traditional analysis methods with PAC-style results, we derived sample sizes necessary for learning regressions or functions under sup norm cost. Furthermore, by restricting the estimator to an orthonormal system, low computational complexity is achieved. Our results also provide finite sample results for widely used estimators based on Haar wavelets, trignometric functions, and Daubechies wavelets.

There are several open issues and futher research directions. It will be interesting to see lower bounds for the sample sizes under the conditions considered in this paper. Also, a more direct comparison with existing function learning methods will be useful in juding the performance of the proposed method. It is expected that larger sample sizes are needed for our method, but, at a lower computational cost. Finally, it will be useful to investigate other estimators known in statistics, such as Kernel estimators, regressograms, and delta estimators, for solving function or regression learning problems.

References

1. N. Alon, S. Ben-David, N. Cesa-Bianchi, and D. Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. In *Proc. of 1993 IEEE Symp. on Foundations of Computer Science*, 1993.
2. M. Anthony and P. Bartlett. Function learning from interpolation. NeuroCOLT Technical Report Series NC-TR-94-013, Royal Holloway, University of London, 1994.
3. K. Apsitis, R. Freivalds, and C. H. Smith. On the inductive inference of real valued functions. In *Proc. of 8th Ann. ACM Conf. on Computational Learning Theory*, 1995.

4. P. Auer, P. M. Long, W. Mass, and G. J. Woeginger. On the complexity of function learning. *Machine Learning*, 18:187–230, 1995.
5. P. L. Bartlett, P. M. Long, and R. C. Williamson. Fat-shattering and the learnability of real-valued functions. *Journal of Computer and Systems Sciences*, 52:434–452, 1996.
6. A. L. Blum and R. L. Rivest. Training a 3-node neural network is NP-complete. *Neural Networks*, 5:117–127, 1992.
7. L. Brieman, J. H. Friedman, R. A. Olshen, and C. J. Stone. *Classification and Regression Trees.* Wadsworth and Brooks, 1984.
8. Z. Ciesielski. Haar system and nonparametric density estimation in several variables. *Probability and Mathematical Statistics*, 9:1–11, 1988.
9. I. Daubechies. Orthonormal bases of compactly supported wavelets. *Comm. Pure and Appl. Math.*, 41:909–996, 1988.
10. L. Devroye. The uniform convergence of the Nadaraya-Watson regression function estimate. *The Canadian Journal of Statistics*, 6(2):179–191, 1978.
11. R. Dudley. *École d'Été de Probabilités de St. Flour 1982*, volume 1097 of *Lecture Notes in Mathematics*, chapter A Course on Empirical Processes, pages 2–142. Springer-Verlag, New York, 1984.
12. J. Engel. A simple wavelet approach to nonparametric regression from recursive partitioning schemes. *Journal of Multivariate Analysis*, 49:242–254, 1994.
13. W. Hardle. *Applied Nonparametric Regression.* Cambridge University Press, New York, 1990.
14. D. Haussler. Decision theoretic generalizations of the PAC model for neural net and other learning applications. *Information and Computation*, 100:78–150, 1992.
15. R. Q. Jia. Subdivision schemes in l_p spaces. *Advances in Comput. Math.*, 3:309–341, 1995.
16. D. Kimber and P. M. Long. The learning complexity of smooth functions of a single variable. In *Proc. of the 1992 Workshop on Computational Learning*, pages 153–159, 1992.
17. H. Liero. Strong uniform consistency of nonparametric regression function estimates. *Probability Theory and Related Fields*, 82:587–614, 1989.
18. E. A. Nadaraya. Remarks on non-parametric estimates for density functions and regression curves. *Theory of Probability and Applications*, 15:134–137, 1970.
19. E. A. Nadaraya. *Nonparametric Estimation of Probability Densities and Regression Curves.* Kluwer Academic Publishers, Dordrecht, 1989.
20. D. Nolan and D. Pollard. U-Processes: Rates of convergence. *Annals of Statistics*, 15(2):780–799, 1987.
21. D. N. Osherson, M. Stob, and S. Weinstein. A note on PAC-inference of real functions, 1989. manuscript.
22. E. Parzen. On estimation of probability density and mode. *Annals of Mathematical Statistics*, 35:1065–1076, 1962.
23. D. Pollard. *Empirical Processes: Theory and Applications.* Institute of Mathematical Statistics, Haywood, California, 1990.
24. S. E. Posner and S. R. Kulkarni. On the complexity of function learning. In *Proc. of 6th Ann. ACM Conf. on Computational Learning Theory*, pages 439–445, 1993.
25. B. L. S. Prakasa Rao. *Nonparametric Functional Estimation.* Academic Press, New York, 1983.
26. N. S. V. Rao and V. Protopopescu. On PAC learning of functions with smoothness properties using feedforward sigmoidal networks. *Proceedings of the IEEE*, 84(10):1562–1569, 1996.

27. P. Revesz. How to apply the method of stochastic approximation in the nonparametric estimation of a regression function. *Math. Operationsforsch. Statist., Ser. Statistics*, 8(1):119–126, 1977.
28. L. Rutkowski. Sequential estimates of a regression function by orthogonal series with applications in discrimination. In *Lecture Notes in Statistics*, volume 8, pages 236–244. Springer-Verlag, New York, 1981.
29. G. Sansone. *Orthogonal Functions.* Dover Publications, Inc., New York, 1991.
30. C. J. Stone. Consistent nonparametric regression (with discussion). *Annals of Statistics*, 5:595–645, 1977.
31. M. Talagrand. Sharper bounds for Gaussian and empirical processes. *Annals of Probability*, 22(1):28–76, 1994.
32. L. G. Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.
33. A. W. van der Vaart and J. A. Wellner. *Weak Convergences and Empirical Processes.* Springer-Verlag, New York, 1996.
34. V. Vapnik. *Estimation of Dependences Based on Empirical Data.* Springer-Verlag, New York, 1982.
35. V. N. Vapnik. *The Nature of Statistical Learning Theory.* Springer-Verlag, New York, 1995.
36. A. Zygmund. *Trignometric Series: Volume I.* Cambridge University Press, Cambridge, UK, 1959.

This article was processed using the LaTeX macro package with LLNCS style