[12] P. Jacquet, "Random infinite trees and supercritical behavior of collision resolution algorithms," *IEEE Trans. Inform. Theory*, vol. 39, pp. 1460–1465, July 1993.
[13] P. Mathys, "Analysis of random-access algorithms," Ph.D. dissertation, Swiss Federal Inst. Technol., Zurich, 1984.
[14] N. Mehravari and T. Berger, "Poisson multiple-access contention with binary feedback," *IEEE Trans. Inform. Theory*, vol. IT-30, pp. 745–751, Sept. 1984.
[15] J. Mosely and P. A. Humblet, "A class of efficient contention resolution algorithms for multiple access channels," *IEEE Trans. Commun.*, vol. COM-33, Feb. 1985.
[16] B. S. Tsybakov, "Survey of USSR contribution to random multiple-access communication," *IEEE Trans. Inform. Theory*, vol. IT-31, pp. 143–165, Mar. 1985.
[17] ——, "On a random process in multiple access problems," *Probl. Inform. Transm.*, vol. 29, no. 2, pp. 163–175, 1993.
[18] B. S. Tsybakov and N. D. Vvedenskaya, "Random multiple access stack algorithm," *Probl. Inform. Transm.*, vol. 16, no. 2, pp. 230–243, 1980.
[19] B. S. Tsybakov and V. A. Mikhailov, "Random multiple packet access: Part-and-try algorithm," *Probl. Inform. Transm.*, vol. 16, no. 4, pp. 305–317, 1980.
[20] S. Verdú, "Computation of the efficiency of the Mosely–Humblet contention resolution algorithm: A simple method," *Proc. IEEE*, vol. 74, pp. 613–614, Apr. 1986.
[21] N. D. Vvedenskaya and B. S. Tsybakov, "Packet delay in the case of a multiple access stack algorithm," *Probl. Inform. Transm.*, vol. 20, no. 2, 1984.
[22] N. D. Vvedenskaya, P. Jacquet, and B. S. Tsybakov, "Packet delay caused by stack algorithm for overcritical income flow," *Probl. Inform. Transm.*, vol. 30, no. 4, pp. 357–369, 1994.
[23] N. D. Vvedenskaya and M. S. Pinsker, "Bounds on the capacity of FCFS multiple-access algorithms," *Probl. Inform. Transm.*, vol. 26, no. 1, pp. 274–279, 1990.
[24] Z. Zhang and T. Berger, "Improved upper bound to capacity of RMA system" (in Russian), *Probl. Pered. Inform.*, vol. 21, no. 4, pp. 83–87, 1985.

Lower Bounds for Multivariate Approximation by Affine-Invariant Dictionaries

Vitaly Maiorov and Ron Meir, *Member, IEEE*

***Abstract*—The problem of approximating locally smooth multivariate functions by linear combinations of elements from an affine-invariant redundant dictionary is considered. Augmenting recent upper bound results for approximation, we establish lower bounds on the performance of such schemes. The lower bounds are tight to within a logarithmic factor in the number of elements used in the approximation. Using a recently introduced notion of nonlinear approximation, we show that the approximation ability may be completely characterized by the pseudodimension of the approximation space with respect to a finite set of points. This result establishes a useful link between the problems of approximation and estimation, or learning, the latter often being conveniently characterized, at least in terms of upper bounds, by the pseudodimension.**

***Index Terms*—Affine invariance, approximation error, dictionaries, pseudodimension.**

I. Introduction

One of the most interesting outcomes of the research concerning statistical learning in recent years has been the confluence of ideas from the rather disparate fields of empirical process theory and approximation theory. In the former context, it turns out that a quantity of major importance in characterizing the performance of empirical estimators, at least in terms of upper bounds, is the so-called pseudodimension. Its relevance to approximation, as well as estimation, has been recently demonstrated in [21], enabling the formulation of precise performance bounds in terms of this quantity.

A great deal of work has been devoted over the past few years to the problem of nonlinear approximation (for a broad outlook see the recent review by DeVore [7]). Although the optimality of certain nonlinear approaches such as free-knot splines [4] has been known for some years, their computational intractability has rendered them of limited practical use, especially for high-dimensional problems. Recently, Donoho, Johnstone, and co-workers [11] have shown that computationally simpler methods, based on wavelet thresholding, yield similar near-optimal performance at a greatly reduced computational cost. However, most of these results and algorithms are pertinent only to one-dimensional (1-D) problems, leaving the problem of effective multivariate approximation very much an open problem.

Concerning the problem of multivariate approximation by wavelet-based dictionaries, some recent progress has been made. Following the work of Barron [2], Delyon *et al.* [6] have considered Monte Carlo based methods for constructing wavelet networks, and upper bounds have been established on their performance. Further work relating to greedy approximation by wavelet dictionaries is discussed in [12], while algorithms and bounds for greedy approximation by neural networks are given in [24]. A recent survey of some of these results may be found in [27].

Manuscript received September 29, 1999; revised July 3, 2000. This work was supported in part by the Technion V.P.R. Fund for the Promotion of Sponsored Research and by the Ollendorff Center of the Department of Electrical Engineering at the Technion.

V. Maiorov is with the Department of Mathematics, Technion–Israel Institute of Technology, Haifa 32000, Israel.

R. Meir is with the Department of Electrical Engineering, Technion–Israel Institute of Technology, Haifa 32000, Israel (e-mail: rmeir@ee.technion.ac.il).

Communicated by G. Lugosi, Associate Editor for Nonparametric Estimation, Classification, and Neural Networks.

Publisher Item Identifier S 0018-9448(01)02708-0.

The problem of establishing lower bounds for the performance of nonlinear approximation is somewhat delicate. A classic approach to the problem considers the selection of the best n-term approximation from a given dictionary. Along these lines there have been several results providing lower bounds for various types of *discrete* dictionaries, e.g., the work of DeVore, Kashin and Temlyakov, [10], [14], [15], [30]. The problem was entirely solved in [8] for the case of trigonometric polynomials. Another direction involves the so-called Alexandroff n-width connected with the *continuous* approximation of classes of smooth functions. The work by Tichomirov [31] and DeVore *et al.* [28] provides several results in this setting. Finally, a different approach was taken recently by Maiorov and Ratsaby [21] who devised a new nonlinear measure of approximation, relating it to the pseudodimension (see Section II).

The methods established in this work enable us to obtain lower bounds on the approximation error with respect to the Sobolev class, for any dictionary-based method for which the pseudodimension, with respect to a *finite* set of points, can be computed. In particular, we focus in this work on affine-invariant dictionary based approximations, and establish lower bounds for several types of activation functions. Standard neural networks and radial basis functions [13], as well as wavelet networks [6] fall within this class.

In particular, we consider approximation by linear combinations of n nonlinear functions as in (3), where the functions ϕ are chosen from some large redundant set of functions, often referred to as a dictionary (e.g., [7]). We consider several types of functions, namely, rational functions, spline functions, and exponential polynomial functions, which cover a large fraction of the functions used in applications. For these types of functions we compute lower bounds on the error incurred in approximating Sobolev functions (see (2)) in the L_q norm. Upper bounds are available for these n-term approximations in [25], [26], [24], [5]; see [27] for a review. All these upper bounds are of the form $c_1 n^{-r/d}$, where r is the degree of smoothness of the Sobolev space, d is the Euclidean dimension, and c is a constant that is independent of n. The lower bounds derived here, for a very large class of dictionaries, are of the form $c_2(n \log n)^{-r/d}$, which match the upper bounds up to logarithmic terms. It should be noted that there exists a specific function ϕ^*, for which an upper bound of the order $n^{-r/(d-1)}$ exists [17]. However, the function ϕ^* is rather intricate and of little practical use. Moreover, it was recently proved in [18] that this upper bound cannot be improved (up to constants), since it is matched by a corresponding lower bound. However, it can be shown that the linear superposition of just three functions of this form leads to a family of functions with an infinite Vapnik–Chervonenkis (VC) dimension, thus rendering them of little use for learning purposes.

Finally, we comment that in this correspondence the symbols $c, c_1, c_2, \ldots$ represent constants which are independent of n, but may depend on other relevant parameters (such as d, r, and others). We retain the subscripts on the constants to distinguish between constants arising from different bounds, as several mathematical manipulations make use of the different origins of these constants. In any event, no attempt is made here to provide optimal values for these constants.

II. Preliminaries

We first recall some results from the theory of 1-D orthogonal wavelets (e.g., [22]). One begins with a 1-D wavelet ψ, such that the set $\{\psi_{jk}(\cdot)\}$, $j, k \in \mathbb{Z}$, where $\psi_{jk}(x) = 2^{k/2}\psi(2^k x - j)$, is an orthonormal basis of $L_2(\mathbb{R})$, i.e., for any $f \in L_2(\mathbb{R})$

$$f = \sum_{j,\,k\in\mathbb{Z}} c_{jk}\psi_{jk}(x) \qquad \|f\|^2 = \sum_{j,\,k\in\mathbb{Z}} |c_{jk}|^2$$

where $c_{jk} = \int_{\mathbb{R}} f(x)\psi_{jk}(x)\,dx$. In order to obtain an optimal n-term approximation, the largest (in absolute value) coefficients c_{jk} are retained. For various reasons (e.g., lack of translation invariance—see [22] for details) one often considers expansions in terms of functions g selected from some *dictionary* $\mathcal{D}$ (see below). When moving to higher dimensions, the construction of orthogonal bases respecting higher dimensional symmetries becomes much more complex. Two possible solutions are the construction of bases formed by tensor products of univariate wavelets or the use of radial wavelets (e.g., [6]). Alternatively, multidimensional frames can be constructed [16]. However, there does not seem to be at present a comprehensive theory for the effective construction and performance assessment of multivariate wavelets. Note that the problem of optimally approximating a function with a linear expansion over a redundant dictionary is known to be NP-hard even in the 1-D case [12], leading to several approximate procedures, such as the matching pursuit of Mallat and Zhang [23].

We consider the problem of approximating functions based on superpositions of functions belonging to some *dictionary*. As far as we are concerned, a dictionary is a rather arbitrary subset of some functional space (such as L_2), which can be conveniently parameterized. For example, neural networks are constructed from dictionaries of the form $\{\phi(\boldsymbol{a}^T\boldsymbol{x} + b)\colon \boldsymbol{a} \in \mathbb{R}^d,\ b \in \mathbb{R}\}$, and for radial basis functions we have $\{\phi(\|\boldsymbol{x} - \boldsymbol{a}\|/b) : \boldsymbol{a} \in \mathbb{R}^d,\ b \in \mathbb{R}\}$. A major feature distinguishing this work from classical approaches in approximation theory is that the family of functions used in the approximation process is highly redundant. One is then interested in constructing a good approximation based on a linear combination of n terms from the dictionary. In this work, we consider dictionaries based on affine-invariant classes of functions, which take the form

$$H^\ell(\phi) = \{\phi(A\boldsymbol{x} + \boldsymbol{b})) : A \in M^{\ell,\,d},\ \boldsymbol{b} \in \mathbb{R}^\ell\} \tag{1}$$

where $M^{\ell,\,d}$ is the space of real-valued $\ell \times d$ matrices. Note that for $\ell = 1$ we obtain the standard ridge function used in neural networks. For $\ell = d$ and $A = aI$, where I is the unit matrix and a is a scalar, and assuming $\phi(\boldsymbol{x}) = \phi(\|\boldsymbol{x}\|)$, we obtain the widely used radial basis function (e.g., [13]).

In this work, we make use of an important characteristic of any functional class H, the so-called pseudodimension, which has proved to be essential in the theory of learning. We start with the more familiar VC dimension.

Definition 1 (VC Dimension): Let H be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and let $\Omega \subseteq \mathcal{X}$. The VC dimension of H with respect to the set Ω, denoted by $\mathrm{VC}\dim(H, \Omega)$, is the largest value of n for which there exist points $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \Omega$ such that

$$|\{\mathrm{sgn}(h(\boldsymbol{x}_1)), \ldots, \mathrm{sgn}(h(\boldsymbol{x}_n))\} : h \in H| = 2^n.$$

If no such finite value exists, $\mathrm{VC}\dim(H, \Omega) = \infty$.

A slightly more refined concept is the so-called pseudodimension of a class of functions H, defined as follows.

Definition 2 (Pseudodimension): Let H be a class of functions from $\mathcal{X}$ to $\mathbb{R}$ and let $\Omega \subseteq \mathcal{X}$. The pseudodimension of H, denoted by $\mathrm{P}\dim(H, \Omega)$, is the largest value of n for which there exist $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \Omega$ and constants $\{c_1, \ldots, c_n\} \in \mathbb{R}$, such that

$$|\mathrm{sgn}(h(\boldsymbol{x}_1) - c_1), \ldots, \mathrm{sgn}(h(\boldsymbol{x}_n) - c_n)\} : h \in H| = 2^n.$$

If no such finite value exists, $\mathrm{P}\dim(H, \Omega) = \infty$.

From the definition it is clear that $\mathrm{P}\dim(H) \geq \mathrm{VC}\dim(H)$.

As mentioned in Section II, the task of defining and analyzing an appropriate nonlinear distance measure between two functional spaces has not yet been satisfactorily addressed (see, for example, [7, Sec. 9]). In this work, we adopt the proposal of [21], which provides a very natural definition of such a distance; see Definition 3 below.

First, however, we need to characterize the space of functions we wish to approximate. We present below the standard definition of the

Sobolev space, for which nontrivial lower bounds can be established. Its relation to the currently popular Besov space will be briefly alluded to below. Let

$$\mathrm{dist}(F,\, H,\, L_q) = \sup_{f\in F} \inf_{h\in H} \|f-h\|_{L_q}$$

denote the L_q distance between two functional spaces F and H. We refer to $\mathrm{dist}(F,\, H,\, L_q)$ as the *approximation error* in approximating F by H.

Definition 3 (Nonlinear ρ-Width): Let F and H be two sets consisting of functions from $\mathcal{X}$ to $\mathbb{R}$. Then

$$\rho_n(F,\, L_q) = \inf_{H\in\mathcal{H}_n} \mathrm{dist}(F,\, H,\, L_q)$$

where $\mathcal{H}_n$ is the set of all functional classes with $\mathrm{P\dim}(H) = n$.

Let $K \subset \mathbb{R}^d$ be a compact domain, and denote by $L_p(K)$ the L_p norm computed over the domain K, namely,

$$\|f\|_{L_p(K)} = \left(\int_K |f|^p\right)^{1/p}.$$

Definition 4 (Sobolev Space): Let $\vec{k} = (k_1,\, k_2,\, \ldots,\, k_d)$, $k_i \in \mathbb{N}$, and define the derivative

$$D^{\vec{k}} f(x) = \frac{\partial^{|\vec{k}|} f}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}}$$

where $|\vec{k}| = k_1 + \cdots + k_d$. The classic Sobolev class is then defined for $r \in \mathbb{N}$ as

$$W_p^{r,\,d} = W_p^{r,\,d}(K) = \left\{ f : \max_{0\le|\vec{k}|\le r} \|D^{\vec{k}} f\|_{L_p(K)} < \infty,\ r\in\mathbb{N} \right\}. \tag{2}$$

This definition has been extended to the case of real-valued r. We omit the details here (see, for example, [24, Sec. 2.2]).

We note that a great deal of recent work has been devoted to the so-called Besov spaces $B_{p,\,q}^{r,\,d}(K)$, which can be thought of as extensions of Sobolev spaces (e.g., [32]). These spaces are attractive for practical applications since they seem to include many naturally occurring signals. However, as established in [32, Sec. 2.3.2], these spaces are very closely related to Sobolev spaces with nonintegral smoothness parameter r. Since all our results apply to any value of r, there is no loss of generality in our analysis. In particular, we have from [32, Sec. 2.3] the useful relation

$$W_p^{r+\epsilon,\,d}(K) \subseteq B_{p,\,q}^{r,\,d}(K) \subseteq W_p^{r-\epsilon,\,d}(K)$$

which holds for every $q \ge 1$, $\epsilon > 0$, and any compact domain K. Since ϵ is arbitrary, very little is lost by working with the Sobolev space rather than the Besov space. In fact, any bounds for Sobolev spaces can be turned into bounds for Besov spaces, at the cost of adding a logarithmic factor in the number of approximating terms (see, for example, [24] for a more extensive discussion).

The final result we quote is related to the ρ-width of the Sobolev class [21]. It is an immediate consequence of [21, Theorem 1]. Let x be a real number, then we define $(x)_+ = \max(0,\, x)$. We then have the following result.

Lemma 2.1 ([21, Theorem 1]): Let H be a class of measurable real-valued functions over a compact domain, characterized by pseudodimension $\mathrm{P\dim}(H)$. For any r and $1 \le p,\, q \le \infty$ satisfying $\frac{r}{d} > (\frac{1}{p} - \frac{1}{q})_+$, and integers $1 \le n,\, d \le \infty$, we have

$$\mathrm{dist}(W_p^{r,\,d},\, H,\, L_q) \ge \frac{c}{\{\mathrm{P\dim}(H)\}^{r/d}}$$

where $c = c(r,\, d,\, p,\, q)$.

In other words, for any space H of pseudodimension n, $\mathrm{dist}(W_p^{r,\,d},\, H,\, L_q) \ge cn^{-r/d}$. Note that if the pseudodimension of H is infinite, as may occur in some cases (see, for example, [29]), this bound becomes useless. In this correspondence, we show that nonzero lower bounds may be established even in cases where the standard pseudodimension is infinite, which motivates the specific approach taken in this correspondence. It should also be commented, that as far as we are aware, Lemma 2.1 provides the first nontrivial lower bound for nonlinear approximation in terms of the pseudodimension of the approximating class of functions.

Before presenting the technical details, we outline the basic procedure used to establish the results of this work. This methodology can be used in many similar problems. In this work, we apply it to the problem of nonlinear dictionary-based approximation. Let F be a set of functions from Ω to $\mathbb{R}$. The goal is to approximate F by n-term expansions from some dictionary $\mathcal{D}$, which in his work will be assumed to take the form given in (1). Denote by $H_n^\ell(\phi)$ the collection of functions that can be expressed as linear combinations of at most n elements from $H^\ell(\phi)$. We wish to obtain a lower bound on $\mathrm{dist}(F,\, H_n^\ell(\phi),\, L_q)$. We proceed as follows.

- Let $S_m = \{\boldsymbol{\xi}_1,\, \ldots,\, \boldsymbol{\xi}_m\}$ be a finite set of points in Ω, and denote by $H_{nm}^\ell(\phi)$ the restriction of the class $H_n^\ell(\phi)$ to S_m, where $H_n^\ell(\phi)$ is the set of all linear combinations of at most n terms from $H^\ell(\phi)$. Assuming $F \subseteq L_p(\Omega)$, we relate $\mathrm{dist}(F,\, H_n^\ell(\phi),\, L_q)$ to the distance between the two finite-dimensional sets B_p^m and $H_{nm}^\ell(\phi)$, where B_p^m is the unit ball in $\mathbb{R}^m$ with respect to the l_p norm.
- The quantity $\mathrm{dist}(B_p^m,\, H_{nm}^\ell(\phi),\, l_q^m)$ is then lower-bounded by using bounds on the pseudodimension of sets in Euclidean space, and recent results from [20].
- Choose an appropriate value for m in order to achieve the best lower bound.

III. Dictionary-Based Approximations

We consider a dictionary based on the affine function $\phi(\cdot)$, as given in (1). This leads to an n-term approximation of the form

$$H_n^\ell(\phi) = \left\{ h(\boldsymbol{x}) = \sum_{k=1}^n c_k \phi(\boldsymbol{A}_k \boldsymbol{x} + \boldsymbol{b}_k) : \boldsymbol{A}_k \in M^{\ell,\,d},\, \boldsymbol{b}_k \in \mathbb{R}^d,\, c_k \in \mathbb{R} \right\}. \tag{3}$$

Observe that the class of functions $H_n^\ell(\phi)$ is invariant with respect to affine transformations, $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$. This type of approximation is often used as a basis for wavelet expansions, where typically $\ell = 1$. In the standard applications of wavelets, however, the translation vectors $\boldsymbol{b}_k$ and the scalar scaling parameters $\boldsymbol{A}_k$ are prescribed on a predetermined infinite grid of points [22]. The nonlinear approximation problem then consists of selecting an optimal subset of n terms which yields the best approximation in $H_n^\ell(\phi)$. Here we allow all the parameters to be free, yielding a more general representation. Obviously, the lower bounds derived in this work apply to the more restricted case where the parameters are constrained to take values on a grid.

We consider the problem of approximation over the d-dimensional cube $I^d = [0,\, 1]^d$. Other compact domains may be treated similarly, with the specific region only affecting the constants appearing in the bounds. Set $m = \tilde{m}^d$, and let

$$S_m = \left\{ \left(\frac{i_1}{\tilde{m}},\, \ldots,\, \frac{i_d}{\tilde{m}} \right) : 0 \le i_j \le \tilde{m}-1,\, j = 1,\, \ldots,\, d \right\} \equiv \{\boldsymbol{\xi}_i\}_{i=1}^m \tag{4}$$

be a finite set of points defined on a grid in $[0,\, 1]^d$. Note that we assume here for simplicity that $\tilde{m}$ is an integer. The general case can be treated by taking integer parts of real numbers, but will only affect the constants which we ignore anyway. Furthermore, let

$$H_{nm}^\ell(\phi) = \{(h(\boldsymbol{\xi}_1),\, \ldots,\, h(\boldsymbol{\xi}_m)) : h \in H_n^\ell(\phi)\}$$

be the restriction of the class $H_n^\ell(\phi)$ to the finite set of points S_m. Observe that $H_{nm}^\ell(\phi) \subseteq \mathbb{R}^m$.

We first quote a result from [19], which relates the approximation error to the distance between two finite-dimensional sets.

Lemma 3.1: If $1 \le p, q \le \infty$ and $r/d > (1/p - 1/q)_+$, then for any $m \ge 1$

$$\operatorname{dist}(W_p^{r,d}, H_n^\ell(\phi), L_q) \ge \frac{c_1}{m^{\frac{r}{d}-\frac{1}{p}+\frac{1}{q}}} \operatorname{dist}(B_p^m, H_{nm}^\ell(\phi), l_q^m)$$

where $B_p^m = \{x \in \mathbb{R}^m : \|x\|_{l_p^m} \le 1\}$, and $\|x\|_{l_p^m} = \left(\sum_{i=1}^m |x_i|^p\right)^{1/p}$.

Proof: The proof of the lemma relies on [19, Lemma 7], which was, however, restricted to the case of neural networks, namely, $\ell = 1$. However, that proof only relied on the affine invariance of the family $H_n^1(\phi)$, which holds for general ℓ. □

Next, we need a result which relates the distance between the unit ball B_p^m and the set $H_{nm}^\ell(\phi)$ through the pseudodimension of the latter class. From [21] we have the following.

Lemma 3.2 ([21, Theorem 1]): If $N = \operatorname{Pdim}(H_n^\ell(\phi), S_m)$, $m \ge \eta N$, $\eta = \lceil 16 \log_2(8e) \rceil$ then

$$\operatorname{dist}(B_p^m, H_{nm}^\ell(\phi), l_q^m) \ge \begin{cases} c_{p,q}(m-N)^{1/q-1/p}, & \text{if } 1 \le q \le p \le \infty \\ c_{p,q} N^{1/q-1/p}, & \text{if } 1 \le p < q \le \infty \end{cases}$$

where $c_{p,q} = 1/16$ if $1 \le q \le p \le \infty$ and $c_{p,q} = \eta^{1/q-1/p}/16$ if $1 \le p < q \le \infty$.

An immediate consequence of Lemmas 3.1 and 3.2 is given by the following.

Corollary 3.1: Let m be an integer and set

$$N = \operatorname{Pdim}(H_n^\ell(\phi), S_m), \qquad m \ge \eta N, \ \eta = \lceil 16 \log_2(8e) \rceil.$$

Then

$$\operatorname{dist}(W_p^{r,d}, H_n^\ell(\phi), L_q) \ge \begin{cases} c\,\frac{(m-N)^{1/q-1/p}}{m^{r/d-1/p+1/q}}, & \text{if } 1 \le q \le p \le \infty \\ c\,\frac{N^{1/q-1/p}}{m^{r/d-1/p+1/q}}, & \text{if } 1 \le p < q \le \infty \end{cases}$$

where $c = c_1 c_{p,q}$.

It should be noted that Corollary 3.1 was established in [21] for *any* functional class with pseudodimension N. However, we require only a more restricted version for our purposes.

Using Corollary 3.1 we observe that the result will be established if an *upper* bound is found for $\operatorname{Pdim}(H_n^\ell(\phi), S_m)$, the pseudodimension of set of functions $H_n^\ell(\phi)$, restricted to the finite set S_m. Two comments are in order at this point. First, the pseudodimension is only needed with respect to a *finite* set of points. Second, the pseudodimension depends critically on ϕ, the type of wavelet used. In fact, simple examples are known [29] for which the pseudodimension is infinite.

We proceed to discuss several specific wavelet functions and their respective approximation bounds. The major conclusion of all the special cases is that, under appropriate conditions on the function ϕ

$$\boxed{\operatorname{dist}(W_p^{r,d}, H_n^\ell(\phi), L_q) \ge c(n \log n)^{-r/d}}$$

where the constant c does not depend on n.

A. Rational Functions

Let $\phi(\boldsymbol{x}) = p(\boldsymbol{x})/q(\boldsymbol{x})$, where $p(\cdot), q(\cdot) \in \mathcal{P}_s^d$, the space of degree s polynomials over $\mathbb{R}^d$. We establish an upper bound on the pseudodimension of the class $H_n^\ell(\phi)$.

Before proceeding we quote a result from [34]. The formulation we use is [33, Lemma 10.3].

Lemma 3.3: If $p_1(\boldsymbol{\gamma}), \ldots, p_M(\boldsymbol{\gamma})$ are algebraic polynomials of degree at most r in $N \le M$ variables, $\boldsymbol{\gamma} = (\gamma_1, \ldots, \gamma_N) \in \mathbb{R}^N$, then

$$|\{(\operatorname{sgn}(p_1(\boldsymbol{\gamma})), \ldots, \operatorname{sgn}(p_M(\boldsymbol{\gamma}))) : \boldsymbol{\gamma} \in \mathbb{R}^N\}| \le \left(\frac{4eMr}{N}\right)^N.$$

Note that $|\{(\operatorname{sgn}(p_1(\boldsymbol{\gamma})), \ldots, \operatorname{sgn}(p_M(\boldsymbol{\gamma}))) : \boldsymbol{\gamma} \in \mathbb{R}^N\}|$ is the number of distinct sign assignments that can be obtained by varying $\boldsymbol{\gamma}$ over $\mathbb{R}^N$. We comment that a slightly better bound may be obtained by using [1, Theorem 8.3]. However, since the latter result only affects the constants, we retain the bound of Lemma 3.3. From Lemma 3.3 we conclude as follows.

Lemma 3.4: Let $\phi \in \mathcal{R}_s^d$, the class of rational functions of degree s over $\mathbb{R}^d$. Then

$$\operatorname{Pdim}(H_n^\ell(\phi), S_m) \le 2n(\ell d + \ell + 1) \log \frac{4em(sn+1)}{(\ell d + \ell + 1)n} \le c_3 n \log \frac{c_1 m(sn+1)}{n}.$$

Proof: Let h be any function from $H_n^\ell(\phi)$, namely,

$$h(\boldsymbol{x}) = \sum_{i=1}^n c_i \frac{p(\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i)}{q(\boldsymbol{A}_i\boldsymbol{x} + \boldsymbol{b}_i)} = \frac{P(\boldsymbol{x};\, \boldsymbol{A},\, \boldsymbol{b},\, \boldsymbol{c})}{Q(\boldsymbol{x};\, \boldsymbol{A},\, \boldsymbol{b},\, \boldsymbol{c})},$$

where $\boldsymbol{A} = \{\boldsymbol{A}_1, \ldots, \boldsymbol{A}_n\}$, $\boldsymbol{b} = \{\boldsymbol{b}_1, \ldots, \boldsymbol{b}_n\}$, and $\boldsymbol{c} = \{c_1, \ldots, c_n\}$. Here, $\boldsymbol{A}_i$ are $\ell \times d$ real-valued matrices, and P and Q are polynomials of degree at most $sn + 1$ with respect to the variables $\boldsymbol{x}$, $\boldsymbol{A}$, $\boldsymbol{b}$, $\boldsymbol{c}$. Set $\boldsymbol{\gamma} = (\boldsymbol{A}, \boldsymbol{c}, \boldsymbol{b}) \in \mathbb{R}^{n(\ell d+\ell+1)}$, and let $\{\boldsymbol{\xi}_1, \ldots, \boldsymbol{\xi}_m\} \in S_m$, $\Gamma = \mathbb{R}^{n(\ell d+\ell+1)}$. Then for any $r_1, \ldots, r_m$

$$\begin{aligned} &|\{(\operatorname{sgn}(h(\boldsymbol{\xi}_1) - r_1), \ldots, \operatorname{sgn}(h(\boldsymbol{\xi}_m) - r_m)) : h \in H\}| \\ &= \left|\left\{\left(\operatorname{sgn}\left(\frac{P(\boldsymbol{\xi}_1;\, \boldsymbol{\gamma})}{Q(\boldsymbol{\xi}_1;\, \boldsymbol{\gamma})} - r_1\right), \ldots, \operatorname{sgn}\left(\frac{P(\boldsymbol{\xi}_m;\, \boldsymbol{\gamma})}{Q(\boldsymbol{\xi}_m;\, \boldsymbol{\gamma})} - r_m\right)\right) : \boldsymbol{\gamma} \in \Gamma\right\}\right| \\ &\le |\{(\operatorname{sgn}(P(\boldsymbol{\xi}_1;\, \boldsymbol{\gamma}) - r_1 Q(\boldsymbol{\xi}_1;\, \boldsymbol{\gamma})), \ldots, \operatorname{sgn}(P(\boldsymbol{\xi}_m;\, \boldsymbol{\gamma}) - r_m Q(\boldsymbol{\xi}_m;\, \boldsymbol{\gamma}))) : \boldsymbol{\gamma} \in \Gamma\}| \\ &\quad \times |\{(\operatorname{sgn} Q(\boldsymbol{\xi}_1;\, \boldsymbol{\gamma}), \ldots, \operatorname{sgn} Q(\boldsymbol{\xi}_m;\, \boldsymbol{\gamma})) : \boldsymbol{\gamma} \in \Gamma\}| \\ &\le \left(\frac{4em(sn+1)}{(\ell d + \ell + 1)n}\right)^{2n(\ell d+\ell+1)}. \end{aligned}$$

The final step follows from Lemma 3.3. An upper bound on the pseudodimension of $H_n^\ell(\phi)$ with respect to the finite set S_m is obtained by looking for the smallest value of t for which

$$[(4em(sn+1)/((\ell d + \ell + 1)n)]^{2n(\ell d+\ell+1)}$$

is smaller than 2^t. This yields the desired result. □

Remark 1: We comment that for the rational functions considered in this section, the lower bound may be obtained directly from Lemma 2.1, since, in this case, a similar argument to the one in Lemma 3.4 shows that $\operatorname{Pdim}(H_n^\ell(\phi), I^d) < vn \log(sn)$, i.e., the pseudodimension over the entire cube $[0, 1]^d$ is upper-bounded by a similar term to that obtained by the restriction to a finite grid. A similar argument applies to the case of spline functions considered in Section III-C. However, for the case exponential functions studied in Section III-B, the restriction to a finite grid is essential, as the current upper bounds on the pseudodimension of such networks are prohibitively large.

In order to establish a lower bound on the approximation error, we first observe that

$$\begin{aligned} \operatorname{dist}(W_p^{r,d}, H_n^\ell(\phi), L_q) &\ge \operatorname{dist}(W_\infty^{r,d}, H_n^\ell(\phi), L_q) \\ &\ge \operatorname{dist}(W_\infty^{r,d}, H_n^\ell(\phi), L_1) \end{aligned}$$

where the first inequality follows from $W_\infty^{r,d} \subseteq W_p^{r,d}$, $1 \le p \le \infty$, and the second inequality uses $\|f\|_{L_q} \ge c\|f\|_{L_1}$, $q \ge 1$, which holds over compacta. We then conclude that for some positive constant c_2,

upon setting $m = c_2 n \log n$, we have $m > \eta \mathrm{Pdim}(H_n^\ell(\phi), S_m)$. Then from Corollary 3.1 we conclude

$$\begin{aligned}
&\mathrm{dist}(W_p^{r,d}, H_n^\ell(\phi), L_q) \\
&\quad \geq \mathrm{dist}(W_\infty^{r,d}, H_n^\ell(\phi), L_1) \\
&\quad \geq c_{p,q} \frac{m - \mathrm{Pdim}(H_n^\ell(\phi), S_m)}{m^{r/d+1}} \\
&\quad \geq c \frac{c_2 n \log n - c_3 n \log \frac{c_1 m(sn+1)}{n}}{(n \log n)^{r/d+1}} \\
&\quad = \frac{c}{(n \log n)^{r/d}} \left(c_2 - \frac{c_3 \log(c_1 c_2 (sn+1) \log n)}{\log n} \right).
\end{aligned}$$

Considering the terms in the parentheses we have

$$\begin{aligned}
c_2 - \frac{c_3 \log(c_1 c_2 (sn+1) \log n)}{\log n} &\geq c_2 - c_3 \frac{\log 2 c_1 c_2 n^2}{\log n} \\
&= c_2 - c_3 \left(2 + \frac{c_4}{\log n} \right).
\end{aligned}$$

Choose $c_2 = 3c_3 + c_3 c_4$ and $n \geq 2$. Then $c_2 - c_3(2 + \frac{c_4}{\log n}) \geq c_3$. Hence we conclude that

$$\mathrm{dist}(W_p^{r,d}, H_n^\ell(\phi), L_q) \geq c(n \log n)^{-r/d}.$$

B. Exponential Functions

A large class of wavelet functions used in applications consists of exponential functions. As a specific example we consider the case of the Gaussian function, namely, $\phi(\boldsymbol{x}) = \exp(-\|\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}\|^2)$, $\boldsymbol{x} \in \mathbb{R}^d$, $\boldsymbol{A} \in \mathbb{R}^{\ell d}$, $\boldsymbol{b} \in \mathbb{R}^d$, and $\|\boldsymbol{x}\|^2 = \sum_{i=1}^d x_i^2$. The extension to other cases will be mentioned at the end of this section. We first estimate the pseudodimension $\mathrm{Pdim}(H_n^\ell(\phi), S_m)$ where S_m is given in (4). The approach we use is based on the method introduced by Bartlett and Williamson in [3].

Let $\boldsymbol{\xi} = (\frac{\ell_1}{\tilde{m}}, \ldots, \frac{\ell_m}{\tilde{m}})$, $1 \leq \ell_j \leq \tilde{m}$, be any point in S_m. We have for any $a \in \mathbb{R}$ and $\boldsymbol{b} \in \mathbb{R}^d$

$$\begin{aligned}
\phi(\boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{b}) &= \exp\{-\|\boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{b}\|^2\} \\
&= \exp\{-[\boldsymbol{\xi}^T \boldsymbol{A}^T \boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{\xi}^T \boldsymbol{A}^T \boldsymbol{b} \\
&\qquad + \boldsymbol{b}^T \boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{b}^T \boldsymbol{b}]\} \\
&= \exp\left\{ - \left[\sum_{i,j} (\boldsymbol{A}^T \boldsymbol{A})_{ij} \xi_i \xi_j + \sum_i \right.\right. \\
&\qquad \left.\left. \cdot [(\boldsymbol{A}^T \boldsymbol{b})_i + (\boldsymbol{b}^T \boldsymbol{A})_i] \xi_i + \sum_i b_i^2 \right] \right\} \\
&= \exp\left\{ - \left[\frac{1}{\tilde{m}^2} \sum_{i,j} (\boldsymbol{A}^T \boldsymbol{A})_{ij} \ell_i \ell_j + \frac{1}{\tilde{m}} \sum_i \right.\right. \\
&\qquad \left.\left. \cdot [(\boldsymbol{A}^T \boldsymbol{b})_i + (\boldsymbol{b}^T \boldsymbol{A})_i] \ell_i + \sum_i b_i^2 \right] \right\}
\end{aligned}$$

where for $1 \leq i, j \leq d$, $(\boldsymbol{A}^T \boldsymbol{A})_{ij} = \sum_{k=1}^\ell a_{ki} a_{kj}$ and $a_{ki} = [\boldsymbol{A}]_{ki}$. For $1 \leq i, j \leq d$, introduce new variables

$$\begin{aligned}
y_{ij} &= \exp\left\{ - \frac{(\boldsymbol{A}^T \boldsymbol{A})_{ij}}{\tilde{m}^2} \right\} \\
y_i &= \exp\left\{ - \frac{(\boldsymbol{A}^T \boldsymbol{b} + \boldsymbol{b}^T \boldsymbol{A})_i}{\tilde{m}} \right\} \\
y_{d+i} &= e^{-b_i^2}.
\end{aligned}$$

For fixed $\boldsymbol{\xi}$, the function $(\boldsymbol{A}, \boldsymbol{b}) \mapsto \phi(\boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{b})$ is a polynomial in the $d^2 + 2d$ variables $y_{1,1}, \ldots, y_{d,d}, y_1, \ldots, y_{2d}$, namely,

$$\phi(\boldsymbol{A}\boldsymbol{\xi} + \boldsymbol{b}) = \left(\prod_{i,j=1}^d y_{ij}^{\ell_i \ell_j} \right) \left(\prod_{i=1}^d y_i^{\ell_i} \right) \left(\prod_{i=1}^d y_{d+i} \right).$$

Hence, any function in $H_n^\ell(\phi)$ may be expressed as

$$h(\boldsymbol{\xi}, \tilde{\boldsymbol{\gamma}}) = \sum_{k=1}^n c_k \left(\prod_{i,j=1}^d y_{k,i,j}^{\ell_i \ell_j} \right) \left(\prod_{i=1}^d y_{k,i}^{\ell_i} \right) \left(\prod_{i=1}^d y_{k,d+i} \right)$$

where $\tilde{\boldsymbol{\gamma}} \in \mathbb{R}^{n(d+1)^2}$. It follows, therefore, that $h(\boldsymbol{\xi}, \tilde{\boldsymbol{\gamma}})$ is a polynomial of degree $\tilde{m}^2$ in the $n(d+1)^2$ variables

$$\boldsymbol{y}_k = \{c_k, y_{k,1,1}, \ldots, y_{k,d,d}, y_{k,1}, \ldots, y_{k,2d}\}_{k=1}^n.$$

Using the same arguments as in Lemma 3.4, we conclude that

$$\begin{aligned}
N &\equiv \mathrm{P\,dim} \left\{ (h(\boldsymbol{\xi}_1, \tilde{\boldsymbol{\gamma}}), \ldots, (\boldsymbol{\xi}_m, \tilde{\boldsymbol{\gamma}})) : \tilde{\boldsymbol{\gamma}} \in \mathbb{R}^{n(d+1)^2} \right\} \\
&\leq \mathrm{Pdim} \left\{ (P(\boldsymbol{\xi}_1, \tilde{\boldsymbol{\gamma}}), \ldots, P(\boldsymbol{\xi}_m, \tilde{\boldsymbol{\gamma}})) : \tilde{\boldsymbol{\gamma}} \in \mathbb{R}^{n(d+1)^2} \right\} \\
&\leq \log \left(\frac{4em\tilde{m}^2)}{2n(d+1)^2} \right)^{2n(d+1)^2} \\
&\leq cn \log \left(\frac{c_1 m^{1+2/d}}{n} \right)
\end{aligned}$$

where we used $\tilde{m} = m^{1/d}$. Thus, from Corollary 3.1, arguing as in Section III-A, we obtain that

$$\begin{aligned}
\mathrm{dist}(W_p^{r,d}, H_n^\ell(\phi), L_q) &\geq \mathrm{dist}(W_p^{r,d}, H_n^\ell(\phi), L_1) \\
&\geq c \frac{m - N}{m^{r/d+1}} \\
&\geq c(n \log n)^{-r/d}
\end{aligned}$$

where, similarly to Section III-A, we have chosen $m = c_2 n \log n$.

The same type of analysis may be performed for any function constructed from exponential functions of the variable $\boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$. In fact, using the results of Section III-A, rational functions of exponentials may be analyzed as well. For example, the generalized standard function $(1 + e^{-\|\boldsymbol{u}\|})^{-1}$, $\boldsymbol{u} = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$, used in neural networks, may be expressed as a rational function of polynomials in variables similar to the y's studied in the Gaussian case. In fact, any product of exponential and polynomial functions falls within this framework.

We conclude this section with a widely used class, namely, the class of Gabor functions, which consists of a Gaussian multiplied by a sinusoidal function. For simplicity, we focus on the univariate case, although the results extend naturally to any dimension. In this case, the class of functions considered is given in (3), with $\ell = 1$ and $\phi(x) = e^{-x^2 + ix}$, $x \in \mathbb{R}$. For any x we have

$$e^{-(ax+b)^2 + i(ax+b)} = e^{-a^2 x^2 - 2abx - b^2} e^{i(ax+b)}.$$

As before, let S_m be the uniform grid of m points in $[0, 1]$,

$$S_m = \left\{ 0, \frac{1}{m}, \ldots, \frac{m-1}{m} \right\} \equiv \{\xi_i\}_{i=1}^m.$$

Given the class of functions $H_{nm}(\phi)$ defined in (3), namely,

$$H_{nm}(\phi) = \left\{ \left\{ \sum_{k=1}^n c_k e^{-(a_k \xi_l + b_k)^2 + i(a_k \xi_l + b_k)} \right\}_{l=0}^{m-1} : \right.$$
$$\left. a_k, b_k, c_k \in \mathbb{R} \right\}$$

and using $\xi_l = l/m$, we define the new class

$$\hat{H}_{nm}(\phi) = \left\{ \left\{ \sum_{k=1}^{n} c_k e^{-a_k^2 \xi_l^2 - 2a_k b_k \xi_l - b_k^2} z_k^l w_k \right\}_{l=0}^{m-1} : a_k,\ b_k,\ c_k \in \mathbb{R},\ z_k,\ w_k \in \mathbb{C} \right\}$$

where $\mathbb{C}$ is the complex plane. It is clear that $H_{nm}(\phi) \subseteq \hat{H}_{nm}(\phi)$. From Lemma 3.1

$$\operatorname{dist}(W_p^{r,d},\ H_n(\phi),\ L_q) \geq \frac{c_1}{m^{\frac{r}{d}-\frac{1}{p}+\frac{1}{q}}} \operatorname{dist}(B_p^m,\ H_{nm}(\phi),\ l_q^m).$$

Since the vectors of B_p^m are real

$$\begin{aligned}\operatorname{dist}(B_p^m,\ H_{nm}(\phi),\ l_q^m) &\geq \operatorname{dist}(B_p^m,\ \hat{H}_{nm}(\phi),\ l_q^m) \\ &\geq \operatorname{dist}(B_p^m,\ \operatorname{Re}(H_{nm}(\phi)),\ l_q^m)\end{aligned}$$

where the last inequality follows from the fact that $|x - z| \geq |x - \operatorname{Re}(z)|$ for $x \in \mathbb{R}$ and $z \in \mathbb{C}$. Now, the exponential functions may be transformed into polynomials as was done for the exponential case studied at the beginning of the section. All the resulting polynomials in the set $\operatorname{Re}\{\hat{H}_{nm}(\phi)\}$ are real, and of degree at most m^2, and thus we can proceed as before to obtain similar results for the pseudodimension and the approximation error.

C. Spline Functions

The final example we present involves spline functions, i.e., piecewise polynomial functions. The regions over which the functions are polynomial are very general, their boundaries being defined by zeros of general sets of polynomials. In particular, they include standard applications where the regions are defined to be axis-parallel or diagonal linear splits. Let $P_1, \ldots, P_t$ be polynomials over $\mathbb{R}^d$ of degree at most s. The *zero set* of a polynomial P is defined by

$$Z(P) = \{\boldsymbol{x} \in \mathbb{R}^d : P(\boldsymbol{x}) = 0\}.$$

Consider now the set

$$G(P_1, \ldots, P_t) = \mathbb{R}^d \setminus \cup_{i=1}^{t} Z(P_i)$$

consisting of a finite number of connected components $D_1, \ldots, D_L$, over which none of the polynomials vanish. From the work of Warren [34] (see also [33]), we have

$$L \leq \left(\frac{4ets}{d}\right)^d.$$

Let $q_1, \ldots, q_L$ be any polynomials of degree s over $\mathbb{R}^d$, and construct the following *piecewise polynomial* function:

$$\phi(\boldsymbol{x}) = \begin{cases} q_i(\boldsymbol{x}), & \text{if } \boldsymbol{x} \in D_i \\ 0, & \text{if } \boldsymbol{x} \in \cup_{i=1}^{\ell} Z(P_i). \end{cases} \tag{5}$$

Since the D_i, $i = 1, \ldots, L$, partition $\mathbb{R}^d$, this defines a piecewise polynomial function over all $\mathbb{R}^d$. We then consider the n-term manifold of the form (3), and study the pseudodimension of this set of functions when ϕ is constructed as in (5).

In the space of the parameters

$$\boldsymbol{\gamma} = \{(\boldsymbol{A}_k,\ \boldsymbol{b}_k,\ c_k)\}_{k=1}^{n} \in \mathbb{R}^{(\ell d+\ell+1)n}$$

consider the set

$$\Gamma_m = \mathbb{R}^{(\ell d+\ell+1)n} \setminus \bigcup_{\boldsymbol{\xi} \in S_m} \left\{ \{(\boldsymbol{A}_k,\ \boldsymbol{b}_k,\ c_k)\}_{k=1}^{n} \in \mathbb{R}^{(\ell d+\ell+1)n} : \sum_{k=1}^{n} P(\boldsymbol{A}_k \boldsymbol{\xi} + b_k) = 0 \right\}$$

consisting of the connected components of the system of polynomials.

The set Γ_m consists of connected components $Q_1, \ldots, Q_N$ where by [33, Lemma 10.2]

$$N \leq \left(\frac{4emns}{(\ell d+\ell+1)n}\right)^{n(\ell d+\ell+1)} \leq (cms)^{n(\ell d+\ell+1)}. \tag{6}$$

An upper bound on the cardinality of the class of functions $H_{nm}^{\ell}(\phi)$, restricted to the set S_m, can then be easily established as follows. Note that for each $i \in \{1, 2, \ldots, m\}$, $h(\boldsymbol{\xi}_i, \boldsymbol{\gamma})$ retains a fixed sign on each of the regions Q_j, $j = 1, 2, \ldots, N$. Let $\Gamma = \mathbb{R}^{(\ell d+\ell+1)n}$, then

$$\begin{aligned}&|\{(\operatorname{sgn}(h(\boldsymbol{\xi}_1, \boldsymbol{\gamma})), \ldots, \operatorname{sgn}(h(\boldsymbol{\xi}_m, \boldsymbol{\gamma})) : \boldsymbol{\gamma} \in \Gamma\}| \\ &\qquad = \sum_{i=1}^{N} |\{(\operatorname{sgn}(h(\boldsymbol{\xi}_1, \boldsymbol{\gamma})), \ldots, \operatorname{sgn}(h(\boldsymbol{\xi}_m, \boldsymbol{\gamma})) : \boldsymbol{\gamma} \in Q_i\}|.\end{aligned}$$

Using Lemma 3.3 once more, together with (6), and noting that for every $\boldsymbol{\gamma} \in Q_i$ the function $h(\boldsymbol{\xi}_i, \boldsymbol{\gamma})$ is a polynomial of degree s, we have

$$\begin{aligned}&\sum_{i=1}^{N} |\{(\operatorname{sgn}(h(\boldsymbol{\xi}_1, \boldsymbol{\gamma})), \ldots, \operatorname{sgn}(h(\boldsymbol{\xi}_m, \boldsymbol{\gamma})) : \boldsymbol{\gamma} \in Q_i\}| \\ &\quad \leq N \max_{1 \leq i \leq N} |\{(\operatorname{sgn}(h(\boldsymbol{\xi}_1, \boldsymbol{\gamma})), \ldots, \operatorname{sgn}(h(\boldsymbol{\xi}_m, \boldsymbol{\gamma})) : \boldsymbol{\gamma} \in Q_i\}| \\ &\quad \leq N \left(\frac{4ems}{(\ell d+\ell+1)n}\right)^{n(\ell d+\ell+1)} \\ &\quad \leq \left(\frac{cm^2 s^2 n}{n}\right)^{n(\ell d+\ell+1)}\end{aligned} \tag{7}$$

where c depends on s and d. Set

$$m = cn \log n$$

then $\operatorname{Pdim}(H_n^{\ell}(\phi),\ S_m)$ is upper-bounded by the logarithm of the final term appearing in (7). Thus, we find that

$$\operatorname{Pdim}(H_n^{\ell}(\phi),\ S_m) \leq n(\ell d+\ell+1) \log(cm^2 n) \leq \tilde{c} n \log n.$$

From Corollary 3.1 we then immediately obtain, using the same arguments as in Section III-A, the following result:

$$\begin{aligned}\operatorname{dist}(W_p^{r,d},\ H_n^{\ell}(\phi),\ L_q) &\geq \frac{m - \tilde{c} n \log n}{m^{r/d+1}} \\ &\geq \frac{c}{(n \log n)^{r/d}}.\end{aligned}$$

IV. Upper Bounds and Algorithms

In this work, we were concerned solely with establishing lower bounds on approximation by affine-invariant dictionaries. In fact, recent results demonstrate that the bounds we obtain are in fact tight, up to logarithmic factors. For example, the results in [24] established $O((\log n/n)^{r/d})$ rates of convergence of the approximation error in the case of neural networks (namely, (3) with $\ell = 1$). Similar upper bounds are known to hold for wavelet networks, as shown in [6].

As mentioned in Section II, the actual construction, assessment, and algorithmic implementation of multivariate dictionary-based approximations is still under vigorous research. Moreover, the problem of optimally approximating a function with a linear expansion over a redundant dictionary is known to be NP-hard even in the 1-D case [12]. One possible approach, proposed by Delyon *et al.* [6], attempts to construct wavelet networks by Monte Carlo sampling based on an integral representation of general functions in L_2 using the continuous wavelet transform. While this approach leads to attractive rates of convergence in terms of the approximation error, the computational burden may be rather severe, due to the combinatorial problem of sampling in high-dimensional space (the dimension here is that of the parameter set $\boldsymbol{\gamma}$,

which is of the order of dn). Additionally, for high values of r (the degree of smoothness of the Sobolev space), these authors provide an upper bound of order $n^{-r/d}$, which (up to logarithmic factors) equals the lower bound provided here. A similar approach was proposed recently in [19], where a related procedure was used for neural networks, and similar rates of convergence were established under weaker conditions (in particular, the degree of smoothness r was arbitrary). Both these procedures, while establishing upper bound on the approximation error, were not overly concerned with computational issues. Along similar lines, greedy algorithms, which greedily add functions to a pre-existing subdictionary have been proposed by Mallat and coworkers (e.g., [23], [12]). For the special case of the Gabor functions, discretized algorithms were suggested which lead to fast and efficient implementation and excellent practical performance. However, no approximation bounds were provided. More recently, the present authors [24] as well as Temlyakov and coworkers [9], [30] have considered greedy algorithms, with particular emphasis on establishing upper bounds on the error incurred. For example, in the special case of neural networks (namely, (3) with $\ell = 1$), [24] established $O((\log n/n)^{r/d})$ rate of convergence of the approximation error, which again matches the lower bound up to logarithmic factors.

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their very helpful comments.

REFERENCES

[1] M. Anthony and P. L. Bartlett, *Neural Network Learning; Theoretical Foundations*. Cambridge, U.K.: Cambridge Univ. Press, 1999.
[2] A. R. Barron, "Universal approximation bound for superpositions of a sigmoidal function," *IEEE Trans. Inform. Theory*, vol. 39, pp. 930–945, May 1993.
[3] P. L. Bartlett and R. C. Williamson, "The VC dimension and pseudodimension of two-layer neural networks with discrete inputs," *Neural Comput.*, vol. 8, pp. 625–628, 1996.
[4] M. Birman and M. Solomyak, "Piecewise polynomial approximation of functions of the class w_p^α," *Mat. Sbornik*, vol. 2, pp. 295–317, 1967.
[5] E. J. Candés, "Ridgelets: Theory and applications," Ph.D. dissertation, Standard Univ., Stanford, CA, Aug. 1998.
[6] B. Delyon, A. Juditsky, and A. Benveniste, "Accuracy analysis for wavelet approximations," *IEEE Trans. Neural Networks*, vol. 6, pp. 332–348, Mar. 1995.
[7] R. DeVore, "Nonlinear approximation," *Acta Numer.*, vol. 7, pp. 51–151, 1998.
[8] R. A. DeVore and V. N. Temlyakov, "Nonlinear approximation by trigonometric sums," *J. Fourier Anal. Applic.*, vol. 2, pp. 29–48, 1995.
[9] ——, "Some remarks on greedy algorithms," *Adv. Comput. Math.*, vol. 5, pp. 173–187, 1996.
[10] ——, "Nonlinear approximation in finite-dimensional spaces," *J. Complexity*, vol. 13, pp. 489–508, 1997.
[11] D. L. Donoho and I. M. Johnstone, "Wavelet shrinkage: Asymptopia," *J. Roy. Statist. Soc. B*, vol. 57, no. 2, pp. 301–639, 1995.
[12] G. Davis, S. Mallat, and M. Avellaneda, "Adaptive greedy approximations," *J. Const. Approx.*, vol. 13, pp. 57–98, 1997.
[13] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. Englewood Cliffs, NJ: Prentice-Hall, 1998.
[14] B. S. Kashin, "On approximation properties of complete orthonormal systems," *Trudy Mat. Inst. Steklov*, vol. 172, pp. 187–191, 1985.
[15] B. S. Kashin and V. N. Temlyakov, "On best m-term approximation and the entropy of sets in the space l^1," *Math. Notes*, vol. 56, pp. 1137–1157, 1994.
[16] T. Kugarajah and Q. Zhang, "Multidimensional wavelet frames," *IEEE Trans. Neural Networks*, vol. 6, pp. 1552–1556, Nov. 1995.
[17] V. Maiorov and A. Pinkus, "Lower bounds for approximation by mlp neural networks," *Neurocomputing*, vol. 25, pp. 81–91, 1998.
[18] V. E. Maiorov, "On best approximation by ridge functions," *J. Approx. Theory*, vol. 99, pp. 68–94, 1999.
[19] V. E. Maiorov and R. Meir, "On the near optimality of the stochastic approximation of smooth functions by neural network," *Adv. Comput. Math.*, vol. 13, no. 1, pp. 79–103, 2000.
[20] V. E. Maiorov and J. Ratsaby, "The degree of approximation of sets in Euclidean space using sets with bounded Vapnik–Chervonenkis dimension," *J. Discr. Appl. Math.*, vol. 86, pp. 81–93, 1998.
[21] ——, "On the degree of approximation using manifolds of finite pseudo-dimension," *J. Constr. Approx.*, vol. 15, pp. 291–300, 1999.
[22] S. Mallat, *A Wavelet Tour of Signal Processing*. San Diego, CA: Academic, 1998.
[23] S. Mallat and Z. Zhang, "Matching pursuit with time-frequency dictionaries," *IEEE Trans. Signal Processing*, vol. 41, pp. 3397–3415, Dec. 1993.
[24] R. Meir and V. Maiorov, "On the optimality of neural network approximation using incremental algorithms," *IEEE Trans. Neural Networks*, vol. 11, pp. 323–337, Mar. 2000.
[25] H. Mhaskar, "Neural networks for optimal approximation of smooth and analytic functions," *Neural Comput.*, vol. 8, no. 1, pp. 164–177, 1996.
[26] P. P. Petrushev, "Approximation by ridge functions and neural networks," *SIAM J. Math. Anal.*, vol. 30, pp. 155–189, 1998.
[27] A. Pinkus, "Approximation theory of the MLP model in neural networks," *Acta Numerica*, vol. 8, pp. 143–195, 1999.
[28] R. Howard, R. A. DeVore, and C. Michelli, "Optimal nonlinear approximation," *Manuscripta Math.*, vol. 63, pp. 460–478, 1989.
[29] E. D. Sontag, "Feedforward nets for interpolation and classification," *J. Comp. Syst. Sci.*, vol. 45, pp. 20–48, 1992.
[30] V. N. Temlyakov, "The best m-term approximation and greedy algorithms," *Adv. Comput. Math.*, vol. 8, no. 3, pp. 249–265, 1998.
[31] V. M. Tichomirov, *Some Problems in the Theory of Approximation*. Moscow, U.S.S.R.: Nauka, 1976.
[32] H. Triebel, *Interpolation Theory Function Spaces and Differential Operators*. Berlin, Germany: Veb Deutcher Verlag, 1978.
[33] M. Vidyasagar, *A Theory of Learning and Generalization*. New York: Springer-Verlag, 1996.
[34] H. E. Warren, "Lower bounds for approximation by nonlinear manifolds," *Trans. Amer. Math. Soc.*, vol. 133, pp. 167–178, 1968.

New Self-Dual Codes over GF (4) with the Highest Known Minimum Weights

Jon-Lark Kim

***Abstract*—The purpose of this correspondence is to construct new Hermitian self-dual codes over GF (4) of lengths 22, 24, 26, 32, and 34 which have the highest known minimum weights. In particular, for length 22, we construct eight new extremal self-dual [22, 11, 8] codes over GF (4) which do not have a nontrivial automorphism of odd order. The existence of such codes has been left open since 1991 by Huffman [9].**

***Index Terms*—Hermitian self-dual codes, weight enumerators.**

I. INTRODUCTION

A *linear* $[n, k]$ *code* $\mathcal{C}$ over GF (4) is a k-dimensional vector subspace of GF $(4)^n$, where GF (4) is the Galois field with four elements. In this correspondence, GF$(4) = \{0, 1, 2, 3\}$, where $2 = \omega$, $3 = \omega^2 = \overline{\omega}$, and $\overline{\omega} = 1 + \omega$. The *weight* wt $(\boldsymbol{c})$ of a codeword $\boldsymbol{c} \in \mathcal{C}$ is the number of nonzero components of $\boldsymbol{c}$. The minimum nonzero weight d of all codewords in $\mathcal{C}$ is called the *minimum weight* of $\mathcal{C}$.

Manuscript received June 20, 2000; revised November 15, 2000.

The author is with the Department of Mathematics, Statistics, and Computer Science, 322 SEO(M/C 249), University of Illinois–Chicago, Chicago, IL 60607-7045 USA (e-mail: jlkim@math.uic.edu).

Communicated by P. Solé, Associate Editor for Coding Theory.
Publisher Item Identifier S 0018-9448(01)02837-1.