JOURNAL OF COMPLEXITY **13,** 489–508 (1997)
ARTICLE NO. CM970458

Nonlinear Approximation in Finite-Dimensional Spaces*

R. A. DeVore† and V. N. Temlyakov

Department of Mathematics, University of South Carolina, Columbia, South Carolina 29208

Received December 4, 1996

1. INTRODUCTION

Nonlinear approximation is utilized in many numerical algorithms. In this paper, we shall consider certain problems of nonlinear approximation which arise in image processing. This includes approximation using m terms from a dictionary of functions and greedy algorithms for approximation from such a dictionary.

Let X be a Banach space and let $\mathcal{D} \subset X$ be a subset of X whose linear span is dense in X. We shall call such a set $\mathcal{D}$ a *dictionary*. Let $\Sigma_m(\mathcal{D})$ be the set of all functions $g \in X$ such that g is a linear combination of at most m elements of $\mathcal{D}$. For $f \in X$, we introduce the error of *m-term approximation*,

$$\sigma_m(f, \mathcal{D})_X := \inf_{g \in \Sigma_m(\mathcal{D})} \|f - g\|_X.$$

One of the central problems of nonlinear approximation is to characterize, for a fixed dictionary $\mathcal{D}$, the functions which have a specific order of approximation, e.g., to characterize the functions which have approximation order $O(m^{-\alpha})$ for some fixed $\alpha > 0$. Results of this type are known and easy to prove (see [DT2]) when X is a Hilbert space and $\mathcal{D}$ is an orthogonal system in X. For general dictionaries, there are sufficient conditions on $f \in X$ which guarantee certain rates of decrease for $\sigma_m(f, \mathcal{D})_X$ (see, again, [DT2]).

More generally, if $F \subset X$ is a class of functions, we define

$$\sigma_m(F, \mathcal{D})_X := \sup_{f \in F} \sigma_m(f, \mathcal{D})_X.$$

*This research was supported by the Office of Naval Research Contract N0014-91-J1343 and the National Science Foundation Grant DMS 96-22925.

†E-mail: devore@math.sc.edu.

0885-064X/97 $25.00

The asymptotic behavior of $\sigma_m(F, \mathcal{D})_X$ has been studied for particular dictionaries $\mathcal{D}$ and classes F. For example, the dictionary $\mathcal{D}$ of all multivariate trigonometric polynomials and various classes F of functions have been studied in [DT1]. There are also some results in the general setting (see [P1; J; B; DDGS; DMA; DT2]).

It is clear that larger dictionaries $\mathcal{D}$ provide better approximation. On the other hand, the numerical implementation is more costly as the dictionary gets larger. It is, therefore, important to understand the trade-off between the accuracy of approximation and the size of the dictionary. Yet, it is not obvious how to even define the size of a dictionary when dealing with infinite dimensional spaces X and infinite dictionaries $\mathcal{D}$, as are usually encountered in analysis. In order to understand better the relationship between the size of a dictionary and its approximation power, we shall consider in this paper approximation in finite-dimensional Euclidean spaces equipped with various norms. Many problems of approximation by functions reduce to this type of finite-dimensional approximation (see, e.g., [DT1]).

Let $\mathbb{R}^n$ denote the n-dimensional space of real vectors $x = (x_1, \dots, x_n)$ and let

$$\|x\|_p := \begin{cases} \left(\sum_{j=1}^n |x_j|^p \right)^{1/p}, & 0 < p < \infty, \\ \max_j |x_j|, & p = \infty, \end{cases}$$

be the ℓ_p norms on $\mathbb{R}^n$. We use B_p^n to denote the unit ℓ_p-ball of $\mathbb{R}^n$ and $B_p^n(y; r)$ to denote the ℓ_p-ball of radius r with center y.

A sample of the results we shall prove in this paper are the following estimates for approximation of $F = B_2^n$ in the ℓ_2 norm. We use the abbreviated notation,

$$\sigma_m(F, \mathcal{D})_p := \sigma_m(F, \mathcal{D})_{\ell_p}.$$

If $\mathcal{D}$ is any dictionary with $|\mathcal{D}| = N$ elements, then Corollary 2.2 shows that

$$\sigma_m(B_2^n, \mathcal{D})_2 \ge C N^{-m/(n-m)}, \qquad m \le n/2. \tag{1.1}$$

In particular, (1.1) shows that dictionaries with n^k elements, k fixed, are not effective in approximating the class B_2^n.

If we take $|\mathcal{D}| = a^n$ for some $a > 1$, then (1.1) shows that $\sigma_m(B_2^n, \mathcal{D}) \ge C a^{-2m}$, $m \le n/2$. We shall show in Theorem 2.2 the existence of a dictionary $\mathcal{D}$ with $|\mathcal{D}| = b^n$, $b = 2a + 1$, $a > 1$, which provides the upper bound,

$$\sigma_m(B_2^n, \mathcal{D})_2 \le C a^{-m}. \tag{1.2}$$

There is a gap between (1.1) and (1.2) when one compares the size of the dictionary with the approximation rate. At our present level of understanding, we are not able to remove such gaps. We give in Sections 2 and 3 estimates similar to (1.1) and (1.2) for $\sigma_m(B_p^n, \mathcal{D})_q$ for all $1 \le p, q \le \infty$.

These results should be compared with lower estimates for optimal basis selection derived recently by Kashin and Temlyakov (see [KT]). They show that for each K there exists a positive $C(K)$ such that for any set of $k \le K^n$ bases $\mathcal{B}^j$, $j = 1, \ldots, k$, in $\mathbb{R}^n$, we have

$$\sup_{f \in B_2^n} \inf_j \sigma_m(f, \mathcal{B}^j)_2 \ge C(K), \qquad m < n/2.$$

Of course from a dictionary with C^n elements we can generally form many more than C^n bases.

The second part of this paper turns to the question of how to find good m-term approximations. One common method for generating such approximations is to use a greedy algorithm (sometimes called adaptive pursuit). The greedy idea for generating an m-term approximation to a function $f \in X$ is to take as the first approximation to f its best one-term approximation (which can usually be implemented numerically). Then one iterates this procedure m times on the residual error (see Section 3 for more details). In Sections 4–6, we consider greedy algorithms for various dictionaries $\mathcal{D}$ in $\mathbb{R}^n$. We show how to appropriately choose dictionaries for which the greedy algorithms achieve estimates similar to those of best m-term approximation (such as that given in (1.2)).

For the purpose of orientation and for further use in this paper, we mention three simple examples of nonlinear approximation in $\mathbb{R}^n$.

I. Let $\mathcal{B}$ be the canonical basis for $\mathbb{R}^n$. The following two estimates are well known:

$$\sigma_m(B_1^n, \mathcal{B})_2 \le (m+1)^{-1/2}, \qquad m = 0, 1, \ldots, n-1; \tag{1.3}$$

$$\sigma_m(B_2^n, \mathcal{B})_2 \ge 2^{-1/2}, \qquad m \le n/2. \tag{1.4}$$

To prove (1.3) for $x \in B_1^n$, it is sufficient to approximate x by the vector y which agrees with x in the m largest coordinates of x and is zero otherwise. The vector $x \in B_2^n$ with all coordinates equal to $n^{-1/2}$ provides the lower estimate (1.4).

II. In this example, we want to bring out the connection between approximation from a dictionary and covering numbers. These covering numbers play an important role in many problems of approximation including entropy and widths. We recall the definition of the covering numbers $N_\epsilon(F, \ell_p)$ for a set $F \subset \mathbb{R}^n$. For each $\epsilon > 0$,

$$N_\epsilon(F, \ell_p) := \min\left\{N: F \subset \bigcup_{j=1}^{N} B_p^n(y^j, \epsilon)\right\}$$

with the minimum taken over all sets $\{y_j\}_{j=1}^N$ of points from $\mathbb{R}^n$. By considering dictionaries $\mathcal{D}$ consisting of the points y^j, we find

$$\inf_{|\mathcal{D}|=N_\epsilon(F, \ell_p)} \sigma_1(F, \mathcal{D})_{\ell_p} \le \epsilon. \tag{1.5}$$

In other words, the covering numbers immediately give estimates for 1-term approximation.

III. We can extend the example **II** to m-term approximation by using the concept of metric entropy. Let X be a linear metric space and for a set $\mathcal{D} \subset X$, let $\mathcal{L}_m(\mathcal{D})$ denote the collection of all linear spaces spanned by m elements of $\mathcal{D}$. For a linear space $L \subset X$, the ϵ-neighborhood $U_\epsilon(L)$ of L is the set of all $x \in X$ which are at a distance not exceeding ϵ from L (i.e., those $x \in X$ which can be approximated to an error not exceeding ϵ by the elements of L). For any compact set $F \subset X$ and any integers N, $m \ge 1$, we define the *metric entropy*,

$$\epsilon_{N,m}(F, X) := \inf_{|\mathcal{D}|=N} \inf\{\epsilon: F \subset \cup_{L\in\mathcal{L}_m(\mathcal{D})} U_\epsilon(L)\}.$$

We can express $\sigma_m(F, \mathcal{D})$ as

$$\sigma_m(F, \mathcal{D}) = \inf\{\epsilon: F \subset \cup_{L\in\mathcal{L}_m(\mathcal{D})} U_\epsilon(L)\}.$$

It follows, therefore, that

$$\inf_{|\mathcal{D}|=N} \sigma_m(F, D) = \epsilon_{N,m}(F, X).$$

In other words, finding the best dictionaries for m-term approximation of F is the same as finding sets $\mathcal{D}$ which attain the generalized metric entropy $\epsilon_{N,m}(F, X)$.

We conclude this Introduction by giving some remarks on the implementation of greedy algorithms in image processing which may be beneficial to the reader not familiar with such applications. Two common applications of greedy algorithms and approximation from dictionaries are to compression and feature extraction. One can view the image as a vector in $\mathbb{R}^N$ with N the number of pixel values of the image (typically 0.25 to 10 Mb; the latter occurring, for example, in digital mammography). Lossy compression is interested in approximating the image by a simpler image which can be stored with fewer bits. One frequently transforms the pixel values to a more sparse vector such as the vector of wavelet coefficients. The compressed image is used for storage or transmission of the

image. Compression is also used as a preprocessor for other image processing tasks such as feature extraction.

Greedy algorithms are also used for feature extraction (see, e.g., [DMA]). The idea is to hopefully have a dictionary which can represent the feature to be extracted from the image as a linear combination of a few dictionary elements. In this case, a highly compressed image given by an m-term approximation from the dictionary with m small (for example, $m \le 100$) will extract the feature. One should note that in these applications, in contrast to other applications such as numerical integration or solving boundary value problems, one has readily all information about the target function (image) but wants an approximation with reduced complexity.

The present paper is concerned with approximation from dictionaries of functions. We present various results about the possibility of approximating with a certain efficiency. However, we do not address the important question of how to efficiently numerically implement the approximation.

2. LOWER ESTIMATES FOR m-TERM APPROXIMATION IN $\mathbb{R}^n$

In this section, we shall consider m-term approximation in the ℓ_p norm of certain sets $F \subset \mathbb{R}^n$. In Theorem 2.1, we use ideas from [KT] to give a lower estimate for m-term approximation in the ℓ_1 norm from a general dictionary to general sets $F \subset \mathbb{R}^n$. Lower estimates in the ℓ_1 norm automatically provide lower estimates in the other ℓ_q norms, $q > 1$ (see Corollary 2.1).

We let $\mathrm{Vol}_n(S)$ denote the Euclidean n-dimensional volume of the set $S \subset \mathbb{R}^n$. We recall that the volume of the unit ball B_p^n, $1 \le p \le \infty$, in $\mathbb{R}^n$ can be estimated by

$$C_1^n n^{-n/p} \le \mathrm{Vol}_n(B_p^n) \le C_2^n n^{-n/p}, \tag{2.1}$$

with C_1, $C_2 > 0$ absolute constants.

THEOREM 2.1. *If $F \subset B_2^n$ satisfies*

$$\mathrm{Vol}_n F \ge K^n \mathrm{Vol}_n B_2^n$$

for some $0 < K \le 1$, then for any dictionary $\mathcal{D}$, $|\mathcal{D}| = N$, we have

$$\sigma_m(F, \mathcal{D})_1 \ge CK^2 n^{1/2} N^{-m/(n-m)}, \qquad m \le n/2,$$

With $C > 0$ an absolute constant.

Proof. Let F and $\mathcal{D}$ be as stated in the theorem and let $\rho := \sigma_m(F, \mathcal{D})_1$. We use $\mathcal{L}_m(\mathcal{D})$ to denote the set of all subspaces Y of dimension m which are spanned by m elements from $\mathcal{D}$. Note that the span of any m elements of $\mathcal{D}$ is contained in such a space Y. Denote by $Y + Z$ the direct sum of two sets Y and Z. From the definition of ρ we get

$$F \subset \bigcup_{Y \in \mathcal{L}_m(\mathcal{D})} (Y + B_1^n(0, \rho)), \tag{2.2}$$

where $B_1^n(0, \rho)$ is the ℓ_1-ball of radius ρ centered at the origin. It follows from (2.2) that

$$\mathrm{Vol}_n F \le \sum_{Y \in \mathcal{L}_m(\mathcal{D})} \mathrm{Vol}_n(F \cap (Y + B_1^n(0, \rho)). \tag{2.3}$$

We now fix an arbitrary $Y \in \mathcal{L}_m(\mathcal{D})$ and estimate $\mathrm{Vol}_n(F \cap (Y + B_1^n(0, \rho))$. Let $Y^\perp$ denote the orthogonal complement of Y in $\mathbb{R}^n$. By the definition of the direct sum each $f \in F \cap (Y + B_1^n(0, \rho))$ has the representation

$$f = y + z, \qquad y \in Y, \qquad z \in B_1^n(0, \rho). \tag{2.4}$$

Consider the orthogonal projectors P_Y and $P_{Y^\perp}$ mapping $\mathbb{R}^n$ onto Y and $Y^\perp$, respectively. We have

$$z = P_Y z + P_{Y^\perp} z,$$

which means that (2.4) can be rewritten

$$f = (y + P_Y z) + P_{Y^\perp} z = P_Y f + P_{Y^\perp} f. \tag{2.5}$$

Since $P_Y f \in Y$ and $\|P_Y f\|_2 \le \|f\|_2 \le 1$, we find that $P_Y f \in Y \cap B_2^n$. Also,

$$P_{Y^\perp} f = P_{Y^\perp} z \in P_{Y^\perp}(B_1^n(0, \rho)).$$

Thus, we obtain the volume estimate,

$$\begin{aligned} \mathrm{Vol}_n(F \cap (Y + B_1^n(0, \rho))) &\le \mathrm{Vol}_m(Y \cap B_2^n)\, \mathrm{Vol}_{n-m}(P_{Y^\perp}(B_1^n(0, \rho))) \\ &\le \mathrm{Vol}_m(B_2^m)\, \mathrm{Vol}_{n-m}(P_{Y^\perp}(B_1^n(0, \rho))) \\ &\le C_2^m m^{-m/2}\, \mathrm{Vol}_{n-m}(P_{Y^\perp}(B_1^n(0, \rho))), \end{aligned} \tag{2.6}$$

where the last inequality used (2.1).

We recall (see, for example, [KT, (13)]) that for any subspace Z of $\mathbb{R}^n$ of dimension $r \geq n/2$, we have

$$\mathrm{Vol}_r(P_Z(B_1^n)) \leq C^r r^{-r}, \tag{2.7}$$

with C here and, later in this proof, an absolute positive constant (which may change from line to line). Combining the inequalities (2.6) and (2.7) results in

$$\begin{aligned}\mathrm{Vol}_n(F \cap (Y + B_1^n(0,\ \rho))) &\leq C^n m^{-m/2}(n-m)^{m-n}\rho^{n-m} \\ &\leq C^n m^{-m/2} n^{m-n}\rho^{n-m}.\end{aligned} \tag{2.8}$$

Now there are at most N^m subspaces in $\mathcal{L}_m(\mathcal{D})$. Therefore, by the assumptions of the theorem,

$$\begin{aligned}K^n C_1^n n^{-n/2} &\leq K^n \mathrm{Vol}_n B_2^n \\ &\leq \mathrm{Vol}_n(F) \\ &\leq C^n N^m m^{-m/2} n^{m-n}\rho^{n-m}.\end{aligned} \tag{2.9}$$

From this, we obtain for $m \leq n/2$,

$$\begin{aligned}\rho &\geq C K^{n/(n-m)} N^{-m/(n-m)} n^{(n-2m)/(2n-2m)} m^{m/(2n-2m)} \\ &\geq C K^2 n^{1/2} N^{-m/(n-m)},\end{aligned} \tag{2.10}$$

where we used the fact that

$$n^{(n-2m)/(2n-2m)} m^{m/(2n-2m)} \geq n^{1/2}\left(\frac{m}{n}\right)^{m/(2n-2m)} \geq Cn^{1/2}, \qquad m \leq n/2.$$

Inequality (2.10) is the desired estimate. ■

COROLLARY 2.1. *Let F and $\mathcal{D}$ be as in Theorem* 2.1. *For any $1 \leq q \leq \infty$, we have*

$$\sigma_m(F,\ \mathcal{D})_q \geq C K^2 n^{1/q-1/2} N^{-m/(n-m)}, \qquad m \leq n/2$$

with C an absolute constant.

Proof. Let $x \in F$ be such that $\sigma_m(x,\ \mathcal{D})_1 \geq 2^{-1}\sigma_m(F,\ \mathcal{D})_1$ and let g be any element of $\mathbb{R}^n$ which can be written as a linear combination of at most m elements of $\mathcal{D}$. The inequality

$$\|x - g\|_1 \leq n^{1-1/q}\|x - g\|_q, \qquad 1 \leq q \leq \infty, \tag{2.11}$$

implies that

$$\sigma_m(F, \mathcal{D})_q \geq n^{-1+1/q}\sigma_m(F, \mathcal{D})_1$$

and, therefore, the corollary follows from Theorem 2.1. ■

COROLLARY 2.2. *Let $\mathcal{D}$ be as in Theorem* 2.1. *For any $1 \leq p, q \leq \infty$, we have*

$$\sigma_m(B_p^n, \mathcal{D})_q \geq Cn^{1/q-1/p}N^{-m/(n-m)}, \qquad m \leq n/2. \tag{2.12}$$

with C an absolute constant.

Proof. The set $F = n^{-1/2}B_\infty^n \subset B_2^n$ has volume

$$\mathrm{Vol}_n(F) = 2^n n^{-n/2} \geq C^n \mathrm{Vol}_n(B_2^n).$$

Hence, from Corollary 2.1, we have

$$\begin{aligned}\sigma_m(B_\infty^n, \mathcal{D})_q &= n^{1/2}\sigma_m(F, \mathcal{D})_q \\ &\geq Cn^{1/q}N^{-m/(n-m)}, \qquad m \leq n/2.\end{aligned}$$

This proves the case $p = \infty$ in the corollary. For any $1 \leq p < \infty$, the set $n^{-1/p}B_\infty^n \subset B_p^n$. Hence,

$$\sigma_m(B_p^n, \mathcal{D})_q \geq \sigma_m(n^{-1/p}B_\infty^n, \mathcal{D})_q = n^{-1/p}\sigma_m(B_\infty^n, \mathcal{D})_q,$$

and therefore the general case in (2.12) follows from the case $p = \infty$. ■

Remark 2.1. In the case $N = a^n$ and $p = q$, the lower bound in Corollary 2.2 can be replaced by Ca^{-2m}.

3. UPPER ESTIMATES FOR m-TERM APPROXIMATION USING COVERING NUMBERS

We shall next consider upper estimates for $\sigma_m(F, \mathcal{D})_p$. We begin with the following simple theorem.

THEOREM 3.1. *Let X be any n-dimensional Banach space and let B be its unit ball. For any N there exists a system $\mathcal{D} \subset X$, $|\mathcal{D}| = N$, such that*

$$\sigma_m(B, \mathcal{D})_X \leq \min(1, \epsilon_N^m), \qquad \epsilon_N := \frac{2}{N^{1/n} - 1}. \tag{3.1}$$

Proof. We shall use simple results about covering numbers of the ball B. The following estimate for $N_\epsilon := N_\epsilon(B, X)$ can be found in [P2, p. 63]:

$$N_\epsilon \le (1 + 2/\epsilon)^n.$$

This estimate implies that for a given N we can cover the unit ball B by N balls $B_2^n(y^j, \epsilon_N)$. We define $\mathcal{D} = \{y^j\}_{j=1}^N$.

For each $x \in X$, let $G(x)$ be any best one-term approximation using the elements of $\mathcal{D}$; i.e., $G(x)$ is a best approximation to x by multiples of the elements of $\mathcal{D}$. Then,

$$\|x - G(x)\|_X \le \epsilon_N \|x\|_X.$$

We repeat this argument with x replaced by $x^1 := x - G(x)$ and obtain

$$\|x - G(x) - G(x_1)\|_X \le \epsilon_N \|x - G(x)\|_X \le \epsilon_N^2 \|x\|_X.$$

Repeating this argument m times, we derive the upper estimate

$$\sigma_m(B, \mathcal{D})_X \le \epsilon_N^m.$$

The estimate $\sigma_m(B, \mathcal{D})_X \le 1$ is trivial. This completes the proof. ■

Remark 3.1. The algorithm used in the above proof to approximate x by a linear combination of m-terms of $\mathcal{D}$ is an example of the greedy algorithms discussed in the next section.

Remark 3.2. In the case $N = b^n$ with $b = 2a + 1$, the right side of the estimate (3.1) becomes simply a^{-m}. This is a companion to the lower bound in Remark 2.1.

4. GREEDY ALGORITHMS IN $\mathbb{R}^n$

In the last two sections, we proved upper and lower estimates for $\sigma_m(F, \mathcal{D})_p$ for certain sets F and dictionaries $\mathcal{D}$ in $\mathbb{R}^n$. However, the question exists how to construct natural dictionaries and approximants which achieve this error of approximation. One of the most common numerical methods for generating m-term approximants are greedy algorithms. We consider in this section ℓ_p-greedy algorithms, $1 \le p \le \infty$. In the case $p = 2$, the ℓ_p-greedy algorithm defined below coincides with the *pure greedy algorithm* of [DT2].

Let $1 \le p \le \infty$ and let $\mathcal{D} \subset \mathbb{R}^n$ be a dictionary. If $x \in \mathbb{R}^n$, we let

$$G^p(x) := G^p(x, \mathcal{D})$$

denote a best one-term approximation using $\mathcal{D}$. That is, $G^p(x) = \alpha(x)g(x)$, where $\alpha(x) \in \mathbb{R}$ and $g(x) \in \mathcal{D}$ satisfy

$$\min_{\alpha\in\mathbb{R},\, g\in\mathcal{D}} \|x - \alpha g\|_p = \|x - \alpha(x)g(x)\|_p.$$

Let us also define the residual of approximation

$$R^p(x) := R^p(x, \mathcal{D}) := x - G^p(x).$$

THE ℓ_p-GREEDY ALGORITHM. For $x \in \mathbb{R}^n$, we define $R_0^p(x) := x$ and $G_0^p(x) := 0$. Then for each $m \geq 1$, we inductively define

$$\begin{aligned} G_m^p(x) &:= G_m^p(x, \mathcal{D}) := G_{m-1}^p(x) + G^p(R_{m-1}^p(x)), \\ R_m^p(x) &:= R_m^p(x, \mathcal{D}) := R^p(R_{m-1}^p(x)). \end{aligned}$$

Then, $G_m^p(x)$ is an m-term approximation to x from $\mathcal{D}$ which we call the mth greedy approximant. The error of the mth greedy approximation is

$$\|x - G_m^p(x)\|_p = \|R_m^p(x)\|_p.$$

We note that the best approximation to $x \in \mathbb{R}^n$ from $\mathcal{D}$ is not necessarily unique and, therefore, $G_m^p(x)$ and $R_p^m(x)$ are not necessarily unique. We define

$$\overline{\gamma}_m^p(x, \mathcal{D})_q := \sup \|x - G^p(x, \mathcal{D})\|_q,$$

where the supremum is taken over all possible resulting $G_m^p(x, \mathcal{D})$. Similarly, we define

$$\underline{\gamma}_m^p(x, \mathcal{D})_q := \inf \|x - G^p(x, \mathcal{D})\|_q,$$

where the infimum is taken over all possible resulting $G_m^p(x, \mathcal{D})$. Thus, $\overline{\gamma}$ measures the worst possible error over all possible choices of best approximations in the greedy algorithm and $\underline{\gamma}$ represents the best possible error.

More generally, for a class $F \subset \mathbb{R}^n$ we define

$$\overline{\gamma}_m^p(F, \mathcal{D})_q := \sup_{f\in F} \overline{\gamma}_m^p(f, \mathcal{D})_q$$

with a similar definition for $\underline{\gamma}_m^p(F, \mathcal{D})_q$. In upper estimates for greedy approximation we would like to use $\overline{\gamma}$ and for lower estimates $\underline{\gamma}$.

Theorem 3.1 shows that for $p = q$ and for each $a > 1$ there exists a dictionary $\mathcal{D}$, $|\mathcal{D}| = b^n$, $b = 2a + 1$, such that

$$\overline{\gamma}_m^p(B_p^n, \mathcal{D})_p \le a^{-m}.$$

However, the dictionary $\mathcal{D}$ in that theorem is not very natural or easy to describe. This estimate and Remark 2.1 to Corollary 2.2 indicate that dictionaries $\mathcal{D}$ with $|\mathcal{D}|$ of order C^n play an important role in m-term approximation in $\mathbb{R}^n$. We proceed now to study a natural family of such dictionaries.

Let $M \ge 3$ be an integer and consider the partition of $[-1, 1]$ into M disjoint intervals I_i of equal length: $|I_i| = 2/M$, $i = 1, \dots, M$. We let ξ_i denote the midpoint of the interval I_i, $i = 1, \dots, M$, and $\Xi := \{\xi_i\}_{i=1}^M$. We introduce the dictionary

$$\mathcal{V}_M := \{x \in \mathbb{R}^n : x_j \in \Xi,\ j = 1, \dots, n\}.$$

Clearly $|\mathcal{V}_M| = M^n$. We shall study in this section the ℓ_∞-greedy algorithm for the dictionaries $\mathcal{V}_M$.

THEOREM 4.1. *For any $1 \le q \le \infty$ we have*

$$\overline{\gamma}_m^\infty(B_\infty^n, \mathcal{V}_M)_q \le n^{1/q} M^{-m}, \qquad m = 1, 2, \dots. \tag{4.1}$$

Proof. We prove the relation (4.1) for $q = \infty$. The general case $1 \le q \le \infty$ follows from the case $q = \infty$ by virtue of the inequality

$$\|x\|_q \le n^{1/q} \|x\|_\infty. \tag{4.2}$$

Let $x \in \mathbb{R}^n$ and apply the ℓ_∞-greedy algorithm to obtain any best approximation $G^\infty(x) := G^\infty(x, \mathcal{V}_M)$. It is easy to estimate $\mathbb{R}^\infty(x, \mathcal{V}_M)$. Namely, for each $i = 1, \dots, M$, let Λ_i be the set of j such that $x_j/\|x\|_\infty \in I_i$ and let e_{Λ_i} denote the vector in $\mathbb{R}^n$ which is one for all coordinates $j \in \Lambda_i$ and 0 otherwise. Then,

$$y := \|x\|_\infty \sum_{i=1}^M \xi_i e_{\Lambda_i}$$

is in $\Sigma_1(\mathcal{V}_M)$. Hence,

$$\|R^\infty(x)\|_\infty = \|x - G^\infty(x)\|_\infty \le \|x - y\|_\infty \le \frac{1}{M} \|x\|_\infty.$$

Iterating this m times we obtain for any realization of $G_m^\infty(x)$,

$$\|x - G_m^\infty(x)\|_\infty = \|R_m^\infty(x)\|_\infty \le M^{-m}\|x\|_\infty \tag{4.3}$$

as desired. ■

5. AN UPPER ESTIMATE FOR $\sigma_m(B_p^n, \mathcal{D})_q$, $1 \le p, q \le \infty$

By choosing a suitable dictionary $\mathcal{D}$ and using the result of the previous section, we can derive estimates for $\sigma_m(B_p^n, \mathcal{D})_q$, $1 \le p, q \le \infty$, which will serve as companions to the lower estimates of Corollary 2.2.

We shall use the following known estimates (see [S]) for the covering numbers for the set B_p^n in ℓ_∞^n: for any $M \ge 2$, there exists M^n balls $B_\infty^n(y_j, \epsilon)$, $j = 1, \dots, M^n$, with $\epsilon = Cn^{-1/p}M^{-1}$ and $C > 0$ an absolute constant, such that

$$B_p^n \subset \bigcup_{j=1}^{M^n} B_\infty^n(y_j, \epsilon). \tag{5.1}$$

Let $\mathcal{D}_0 := \{y_j\}_{j=1}^{M^n}$ and let $\mathcal{V}_M$ be the dictionary of the previous section. We let $\mathcal{D} := \mathcal{D}_0 \cup \mathcal{V}_M$. Then, clearly $|\mathcal{D}| \le 2M^n$.

THEOREM 5.1. *For the dictionary $\mathcal{D}$ defined above, we have*

$$\sigma_m(B_p^n, \mathcal{D})_q \le Cn^{1/q-1/p}M^{-m}, \qquad m = 1, 2, \dots, \tag{5.2}$$

with $C > 0$ an absolute constant.

Proof. If $x \in B_p^n$, then there is a $y \in \mathcal{D}_0$ such that $\|x-y\|_\infty \le Cn^{-1/p}M^{-1}$. From Theorem 4.1, there is a $z := G_m^\infty(x-y)$ which satisfies

$$\|x - y - z\|_q \le n^{1/q}M^{-m}\|x - y\|_\infty \le Cn^{1/q-1/p}M^{-(m+1)}.$$

Since $y + z$ is a linear combination of at most $m + 1$ elements of the dictionary $\mathcal{D}$, we have proved (5.2). ■

Remark 5.1. Corollary 2.2 gives that for any dictionary $\mathcal{D}$ with $2M^n$ elements

$$\begin{aligned}\sigma_m(B_p^n, \mathcal{D})_q &\ge Cn^{1/q-1/p}(2M)^{-mn/(n-m)}\\ &\ge Cn^{1/q-1/p}M^{-2m}, \qquad m \le n/2.\end{aligned}$$

Thus, (5.2) is a companion to the lower estimate of Corollary 2.2.

6. THE ℓ_1 GREEDY ALGORITHM

In this section, we shall prove results about the ℓ_1 greedy algorithm for the dictionary $\mathcal{V}_3$ of Section 4. We consider this dictionary in detail for the following reasons. It is a simple dictionary which is easy to describe geometrically. Also, it is fairly easy to analyze the approximation properties of this dictionary. Moreover, it turns out that this dictionary gives geometric order of approximation (see, for example, Theorem 6.2 and Theorem 7.1) which we know is the best we can expect for general dictionaries (see Corollary 2.2).

For each $x \in \mathbb{R}^n$, it is easy to describe a best approximation from $\Sigma_1(\mathcal{D})$ and compute the error

$$\sigma_1(x, \mathcal{V}_3) = \|x - G_1^1(x, \mathcal{V}_3)\|_1 = \overline{\gamma}_1(x, \mathcal{V}_3) = \underline{\gamma}_1(x, \mathcal{V}_3). \tag{6.1}$$

We begin with a few simple remarks about one-term approximation from $\mathcal{V}_3$. Each element $g \in \Sigma_1(\mathcal{V}_3)$ is of the form

$$g = ce_{\Lambda_+} - ce_{\Lambda_-} \tag{6.2}$$

with $c > 0$ and Λ_+ and Λ_- disjoint subsets of $\{1, \ldots, n\}$. Let $\mathcal{G}_c$ be the collection of all g of the form (6.2).

Remark 6.1. For each $x \in \mathbb{R}^n$ and $1 \le p < \infty$, for any best ℓ_p^n, approximation g_c to x from $\mathcal{G}_c$ the set $\Lambda_\pm$ contains all $i \in \{1, \ldots, n\}$ such that $|x_i \mp c| < |x_i|$. Thus, $g_i = c \operatorname{sgn} x_i$ if $|x_i| > c/2$ and $g_i = 0$ if $|x_i| < c/2$. Moreover, $|x_i - g_i| \le |x_i|$, for all $i \in \{1, \ldots, n\}$.

If we use Remark 6.1 for $p = 1$, we see that

$$\|x - g_c\|_1 = \sum_{I_0} |x_i| + \sum_{I_+} |x_i| - \sum_{I_-} |x_i| - c(|I_+| - |I_-|) \tag{6.3}$$

with $I_0 := \{i: |x_i| \le c/2\}$, $I_+ := \{i: |x_i| > c\}$, and $I_- := \{i: c/2 < |x_i| < c\}$.

Given $x \in \mathbb{R}^n$, we denote by x^* its decreasing arrangement. That is x_i^* is the ith largest of the numbers $|x_j|$, $j = 1, \ldots, n$. Then, there is a one to one rearrangement sequence $j^*(i)$ such that $x_i^* = |x_{j^*(i)}|$. We let j^* be any such rearrangement (which is not unique because of possible tie values).

LEMMA 6.1. *For any $x \in \mathbb{R}^n$, we have that*

$$\sigma_1(x, \mathcal{V}_3)_1 = \|x\|_1 - \delta, \tag{6.4}$$

where

$$\delta := \max\{s_j, t_k: 1 \le j \le (n+1)/2;\ 1 \le k \le n/2\},$$

$$s_j := x_j^* + 2 \sum_{i=j+1}^{2j-1} x_i^*, \qquad 1 \le j \le (n+1)/2,$$

and

$$t_j := 2 \sum_{i=j+1}^{2j} x_i^*, \qquad 1 \le j \le n/2.$$

If $\delta = s_j$ for some $1 \le j \le (n+1)/2$, then a best approximation g to x from $\Sigma_1(\mathcal{V}_3)$ is obtained by taking $g_{j^*(i)} = |x_j^*| \operatorname{sign} x_{j^*(i)}$, $i = 1, \dots, 2j-1$, and $g_i = 0$ otherwise. If $\delta = t_j$ for some $1 \le j \le n/2$, then a best approximation g to x from $\Sigma_1(\mathcal{V}_3)$ is obtained by taking $g_{j^*(i)} := c \operatorname{sign} x_{j^*(i)}$, $i = 1, \dots, 2j$, $c = (x_j^* + x_{j+1}^*)/2$ and $g_i := 0$ otherwise.

Proof. We first observe that the left side of (6.4) does not exceed the right. Fix j and define a sequence $g \in \Sigma_1(\mathcal{V}_3)$ as follows. We take $c = (x_j^* + x_{j+1}^*)/2$ and we define g by $g_{j^*(i)} = c \operatorname{sign} x_{j^*(i)}$, whenever $1 \le j^*(i) \le 2j$ and $g_i := 0$ otherwise. Then,

$$\|x - g\|_1 = \sum_{i=1}^{j} (x_i^* - c) + \sum_{i=j+1}^{2j} (c - x_i^*) + \sum_{i=2j+1}^{n} x_i^* = \|x\|_1 - t_j. \quad (6.5)$$

Similarly, if we take $c := x_j^*$ and define g by $g_{j^*(i)} = c \operatorname{sign} x_{j^*(i)}$, whenever $1 \le j^*(i) \le 2j-1$ and $g_i := 0$ otherwise, we obtain

$$\|x - g\|_1 = \sum_{i=1}^{j-1} (x_i^* - c) + \sum_{i=j+1}^{2j-1} (c - x_i^*) + \sum_{i=2j}^{n} x_i^* = \|x\|_1 - s_j. \quad (6.6)$$

The equalities (6.5) and (6.6) imply that the left side of (6.4) does not exceed the right.

We now prove that the right side of (6.4) does not exceed the left. Let c and $g_c \in \mathcal{G}_c$ be such that $\|x - g_c\|_1 = \sigma_1(x, \mathcal{V}_3)$. Clearly, $x_n^* \le c \le x_1^*$. Suppose that $x_j^* > c > x_{j+1}^*$, for some $1 \le j \le n-1$ and consider (6.3). If $|I_+| > |I_-|$, then increasing c slightly does not change I_+ and I_- but gives a smaller $\|x - g_c\|_1$. Similarly, if $|I_-| > |I_+|$, then decreasing c slightly does not change I_+, may change I_-, but always gives a smaller $\|x - g_c\|$. Hence, $|I_+| = |I_-|$ and therefore $I_+ = \{1, \dots, j\}$ and $I_- = \{j+1, \dots, 2j\}$ and, in particular, $j \le n/2$. This shows that

$$\sigma_m(x, \mathcal{V}_3) = \|x - g_c\|_1 = \|x\|_1 - t_j$$

and proves that the right side of (6.4) does not exceed the left side in this case.

Finally, if $c = x_k^*$ for some $1 \le k \le n$, then let $Z := \{i : x_i^* = c\}$. If $|I_+| > |I_-| + |Z|$, then with a view toward (6.3), it is easy to see that increasing c slightly will decrease $\|x - g_c\|_1$ and provide a contradiction to the minimality

of c. Similarly, if $|I_-| > |I_+| + |Z|$, then decreasing c slightly will decrease $\|x-g_c\|_1$ and again give a contradiction. Let $\ell := |I_+|+|I_-|+|Z|$. If $\ell = 2j-1$ for some integer j then $j \le (n+1)/2$ and $c = x_j^*$ and $\|x - g_c\|_1 = \|x\|_1 - s_j$. If $\ell = 2j$ for some integer j, then, with a view toward (6.3), it is easy to see that $c = x_j^* = x_{j+1}^* = (x_j^* + x_{j+1}^*)/2$, and we have

$$\|x - g_c\|_1 = \|x\|_1 - t_j.$$

In either case, we have shown that the right side of (6.4) does not exceed the left side.

The claim about the form of a best approximation also follows from what we have just proved. ■

Remark 6.2. In the case $\delta = t_j$ of Theorem 6.1, any c with $x_j^* \le c \le x_{j+1}^*$ also yields a best approximation g_c to x from $\mathcal{V}_3$. Thus, in all cases, a best approximation to x is of the form g_c with $c = x_j^*$ for some j.

Remark 6.3. We can define numbers s_j, $(n+1)/2 < j \le n$, and t_j, $j < n/2 \le n$, as in Theorem 6.1 by setting $x_j^* := 0$, $j \ge n$. Then Theorem 6.1 remains valid with $\delta := \max\{s_j, t_j\colon 1 \le j \le n\}$. None of the newly defined numbers assume the max, however.

THEOREM 6.2. *We have the estimate*

$$\sigma_m(B_1^n, \mathcal{V}_3)_1 \le \overline{\gamma}_m^1(B_1^n, \mathcal{V}_3)_1 \le \left(1 - \frac{1}{k+1}\right)^m, \tag{6.7}$$

where $k := [\log_2(n+1)]$.

Proof. If $x \in \mathbb{R}^n$, then for the numbers s_j of Lemma 6.1 and Remark 6.3, we have

$$s_1 + s_2 + s_4 + \cdots + s_{2^k} \ge \|x\|_1.$$

Hence for one of these values of j, we have $s_j \ge \|x\|_1/(k+1)$. From Lemma 6.1 and Remark 6.3, we find

$$\gamma_1(x, \mathcal{V}_3) \le \|x\|_1 \left(1 - \frac{1}{k+1}\right).$$

If we iterate this inequality m times, we obtain the lemma. ■

Finally, we want to close this section by showing that in a certain sense, the estimate (6.7) cannot be improved.

THEOREM 6.2. *Let* $n = 2^{2k}$, *with* k *a positive integer. For any* $1 \le m \le k/2$, *we have*

$$\overline{\gamma}_m(B_1^n,\ \mathcal{V}_3)_1 \geq 1/2 \tag{6.8}$$

Proof. Let $x \in \mathbb{R}^n$ be the sequence with $x_i := 2^{-j}$, $2^{j-1} < i \leq 2^j$, $j = 0, 1, \ldots, 2k$. We let $x^m := R_1^m(x)$ denote a residual after m steps of the ℓ_1 greedy algorithm for $\mathcal{V}_3$ when applied to $x^0 := x$. We prove by induction that for each $m = 0, 1, \ldots, k$, the residuals can be chosen to satisfy

$$|x_i^m| = \begin{cases} 2^{-2m}, & 1 \leq i \leq 2^{2m} \\ x_i^0, & i > 2^{2m}. \end{cases} \tag{6.9}$$

For $m = 0$ this is obvious from the definition of x^0. Let us see how to advance the induction to $m = 1$ which will give the general idea of the proof. For the sequence x^0, we have $s_1 = t_1 = s_2 = \ldots$. Therefore, taking $c = (x_1 + x_2)/2 = 3/4$ and defining $g_i = c$, $i = 1, 2$, and $g_i = 0$ gives a best ℓ_1 approximation to x^0 and its residual x^1 satisfies (6.9) for $m = 1$.

Assume that we have proven (6.9) for some $m \geq 1$ and let $y_i = |x_i^m|$. For the numbers s_j and t_j of Lemma 6.1, we have

$$t_j - s_j = 2y_{2j} - y_j \geq 0, \qquad j = 1, 2, \ldots,$$

and

$$s_{j+1} - t_j = 2y_{2j+1} - y_{j+1} \geq 0, \qquad j = 1, 2, \ldots.$$

Hence,

$$s_1 \leq t_1 \leq s_2 \leq t_2 \leq \ldots. \tag{6.10}$$

Moreover,

$$2y_{2j+1} - y_{j+1} = 0 = 2y_{2j} - y_j, \qquad j \geq 2^{2m}.$$

Therefore $t_{2^{2m}}$ is a maximum for the sequence (6.9). Lemma 6.1 says that for $c := (2^{-2m} + 2^{-2m-1})/2$, the sequence g defined by $g_i := c \operatorname{sign} x_i^m$, $1 \leq i \leq 2^{2m+1}$, and $g_i := 0$ otherwise gives a best ℓ_1 approximation to x^m from $\Sigma_1(\mathcal{V}_3)$. One easily checks that the residual $x^{m+1} := x^m - g$ satisfies (6.9) and, thus, the induction hypothesis is advanced and we have proven (6.9).

From (6.9), we have

$$\|x^m\|_1 = \|x^{m-1}\|_1 - 1, \qquad m = 1, 2, \ldots, k.$$

Hence $\|x^m\|_1 \geq \|x\|_1 - m = k - m$. The sequence $x/\|x\|_1$ which is in B_1^n establishes (6.8). ■

Theorem 6.2 can be improved by replacing $\overline{\gamma}$ by $\underline{\gamma}$. For this we shall need the following remark.

Remark 6.4. Suppose that b_k is a sequence of real numbers satisfying $b_1 > 2b_2 > \cdots > 2b_{k-1} > 0$ and $N_0, \ldots, N_k$ are natural numbers satisfying $1 = N_0 < 2^{-1}N_1 < \cdots < 2^{-k}N_k$. If the vector $a \in \mathbb{R}^{N_k-1}$ has a rearrangement which satisfies

$$a_i^* = b_j, \qquad i = N_{j-1}, \ldots, N_j - 1, \qquad j = 1, \ldots, k,$$

then the best ℓ_1-approximation $g = G^1(a, \mathcal{V}_3)$ to a from $\Sigma_1(\mathcal{V}_3)$ is unique and of the form

$$g_i = \begin{cases} b_j, & i = 1, \ldots, N_j, \\ 0, & i = N_j + 1, \ldots, N_k - 1. \end{cases}$$

We leave the proof of this remark (which is along the lines of Theorem 6.1) to the reader.

THEOREM 6.3. *Let $n = 4^k - 1$, with k a positive integer. For any $m \le 3k/8$, we have*

$$\underline{\gamma}_m^1(B_1^n, \mathcal{V}_3)_1 \ge 1/2.$$

Proof. The proof is somewhat along the lines of Theorem 6.1, but more involved. We shall state the main steps and leave the details to the reader. We define the vector x by

$$x_i = 4^{-j-1}, \qquad i = 4^j, \ldots, 4^{j+1} - 1, \qquad j = 0, \ldots, k-1.$$

Then $\|x\| = 3k/4$. We can monitor the performance of the ℓ_1-greedy algorithm on x by using Remark 6.4. At each step j of the greedy algorithm $G^j(x)$ and $x^j := R^j(x)$ are uniquely defined and x^j satisfies the assumptions of Remark 6.4. Using this one can show that $\|x^j\|_1 \ge \|x^{j-1}\|_1 - 1$. The proof is then completed as for Theorem 6.2. ■

7. GREEDY APPROXIMATION IN ℓ_2^n FOR THE DICTIONARY $\mathcal{V}_3$

In this section, we want to carry out an analysis similar to that of Section 6 for the ℓ_2 greedy algorithm and the dictionary $\mathcal{V}_3$.

THEOREM 7.1. *Let $k := [\log_2 n]$. Then,*

$$\overline{\gamma}_m^2(B_2^n, \mathcal{V}_3)_2 \le \left(1 - \frac{1}{k+1}\right)^{m/2}, \qquad m = 1, 2, \ldots. \tag{7.1}$$

Proof. Let $x \in \mathbb{R}^n$ and let x_i^* be its decreasing rearrangement as defined in the previous section. For any $1 \le j \le n$, let $c := x_j^*$ and consider $g^j \in \Sigma_1(\mathcal{V}_3)$ defined by $g_i^j := c \operatorname{sign} x_i$ if $|x_i| \ge c$; $g_i^j := 0$, otherwise. Then,

$$\begin{aligned}\|x - g^j\|_2^2 &= \sum_{i=1}^{j} (x_i^* - x_j^*)^2 + \sum_{i=j+1}^{n} (x_i^*)^2 \\ &= \|x\|_2^2 - \sum_{i=1}^{j} (2x_i^* x_j^* - (x_j^*)^2) \\ &\le \|x\|_2^2 - j(x_j^*)^2 \\ &\le \|x\|_2^2 - u_j. \end{aligned} \tag{7.2}$$

with

$$u_j := \sum_{i=j}^{2j-1} (x_i^*)^2.$$

For the purposes of defining u_j, we define $x_i^* := 0$, $i > n$. Since $\sum_{j=0}^{k} u_{2^j} = \|x\|_2^2$ for one of these values of j, we have $u_{2^j} \ge \|x\|_2^2/(k+1)$. Hence, returning to (7.2), we get for this value of j:

$$\|x - G^2(x, \mathcal{V}_3)\|_2^2 \le \|x - g^{2^j}\|_2^2 \le \left(1 - \frac{1}{k+1}\right) \|x\|_2^2. \tag{7.3}$$

Iterating (7.3), we arrive at (7.1). ■

The following theorem shows that in a certain sense the estimates of Theorem 7.1 cannot be improved.

THEOREM 7.2. *Let $n = 2^k$ for some positive integer k. For any $m \le k/2$, we have*

$$\underline{\gamma}_m^2(B_2^n, \mathcal{V}_3)_2 \ge 1/2.$$

Proof. Consider the vector $z \in \mathbb{R}^n$ with $z_1 := 1$, $z_i := i^{1/2} - (i-1)^{1/2}$, $i = 2, 3, \ldots, n$. Then, we have

$$\begin{aligned}\|z\|_2^2 &= 1 + \sum_{i=2}^{n} (i^{1/2} - (i-1)^{1/2})^2 \ge 1 + \sum_{i=2}^{n} \left(\frac{1}{2i^{1/2}}\right)^2 \\ &= 1 + \frac{1}{4} \sum_{i=2}^{n} \frac{1}{i} \ge 1 + \frac{1}{4} \int_2^{n+1} \frac{dx}{x} \ge \frac{1}{4}(1 + \ln n). \end{aligned} \tag{7.4}$$

Also, for each $\ell \le n$ one has

$$\sum_{i=1}^{l} z_i = \ell^{1/2},$$

which implies that for each $g \in \mathcal{V}_3$, $\|g\|_2 = 1$, we have

$$|\langle z, g\rangle| \le 1. \tag{7.5}$$

From Remark 6.1, we have

$$|\langle R_j^2(z), g\rangle| \le 1 \tag{7.6}$$

for all natural numbers j and elements $g \in \mathcal{V}_3$. Further, for each $x \in \mathbb{R}^n$, the ℓ_2-greedy algorithm satisfies

$$G^2(x, \mathcal{V}_3) = \langle x, g(x)\rangle g(x), \tag{7.7}$$

where $g(x)$ maximizes the inner product $\langle x, g\rangle$ over all $g \in \Sigma_1(\mathcal{V}_3)$ of unit length. This implies

$$\|x\|_2^2 = \|G^2(x, \mathcal{V}_3)\|_2^2 + \|R^2(x, \mathcal{V}_3)\|_2^2.$$

Applying this to the vector z, we obtain

$$\|z\|_2^2 = \|R_m^2(z, \mathcal{V})\|_2^2 + \sum_{k=1}^{m} \|G^2(R_{k-1}^2(z), \mathcal{V})\|_2^2.$$

Using (7.6) and (7.7) we find

$$\|z\|_2^2 \le \|R_m^2(z, \mathcal{V})\|_2^2 + m$$

and by (7.4),

$$\gamma_m^2(B_2^n, \mathcal{V}_3)_2^2 \ge \|R_m^2(z, \mathcal{V}_3)\|_2^2/\|z\|_2^2 \ge 1 - m/\|z\|_2^2 \ge 1/4.$$

This proves Theorem 7.2. ■

REFERENCES

[DMA] Davis, G., Mallat, S., and Avallaneda, M. (1997), Adaptive greedy approximations, *Constr. Approx.* **13,** 57–98.

[DDGS] Donahue, M. J., Gurvits, L., Darken, C., and Sontag, E. (1997), Rates of convex approximation in non-Hilbert spaces, *Constr. Approx.* **13,** 187–220.

[B] Barron, A. (1993), Universal approximation bounds for superposition of a sigmoidal function, *IEEE Trans. Inform. Theory* **39,** 930–945.

[DT1] DeVore, R. A., and Temlyakov, V. N. (1995), Nonlinear approximation by trigonometric sums, *J. Fourier Anal. Appl.* **2,** 29–48.

[DT2] DeVore, R. A., and Temlyakov, V. N. (1996), Some remarks on greedy algorithms, *Adv. Comput. Math.* **5,** 173–187.

[J] Jones, L. (1992), A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Ann. Stat.* **20,** 608–613.

[KT] Kashin, B. S., and Temlyakov, V. N. (1994), On best *m*-term approximations and the entropy of sets in the space L^1, *Math. Notes* **56,** 1137–1157.

[P1] Pisier, G. Remarques sur un résultat non publié de B. Maurey, Séminaire de'analyse fonctionelle 1980–1981, École Polytechnique, Centre de Mathématiques, Palaiseau.

[P2] Pisier, G. (1989), The Volume of Convex Bodies and Banach Space Geometry, Cambridge Univ. Press, Cambridge.

[S] Schütt, C. (1984), Entropy numbers of diagonal operators between symmetric Banach spaces, *J. Approx. Theory* **40,** 121–128.