Advances in Computational Mathematics 5(1996)173–187

Some remarks on greedy algorithms*

R.A. DeVore and V.N. Temlyakov

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

Estimates are given for the rate of approximation of a function by means of greedy algorithms. The estimates apply to approximation from an arbitrary dictionary of functions. Three greedy algorithms are discussed: the Pure Greedy Algorithm, an Orthogonal Greedy Algorithm, and a Relaxed Greedy Algorithm.

1. Introduction

There has recently been much interest in approximation by linear combinations of functions taken from a redundant set $\mathcal{D}$. That is, the elements of $\mathcal{D}$ are not linearly independent. Perhaps the first example of this type was considered by Schmidt in 1907 [6] who considered the approximation of functions $f(x,y)$ of two variables by bilinear forms $\sum_{i=1}^{m} u_i(x)v_i(y)$ in $L_2([0,1]^2)$. This problem is closely connected with properties of the integral operator with kernel $f(x,y)$.

We mention two other prominent examples of this type of approximation.

In neural networks, one approximates functions of d-variables by linear combinations of functions from the set

$$\{\sigma(a\cdot x+b):\ a\in\mathbb{R}^d, b\in\mathbb{R}\},$$

where σ is a fixed univariate function. The functions $\sigma(a\cdot x+b)$ are planar waves. Usually, σ is required to have additional properties. For example, the sigmoidal functions, which are used in neural networks, are monotone non-decreasing, tend to 0 as $x\to-\infty$, and tend to 1 as $x\to\infty$.

Another example, from signal processing, uses the Gabor functions

$$g_{a,b}(x) := e^{iax}e^{-bx^2}$$

and approximates a univariate function by linear combinations of the elements

$$\{g_{a,b}(x-c):\ a,b,c\in\mathbb{R}\}.$$

The common feature of these examples is that the family of functions used in the approximation process is redundant. The redundancy leads to interesting new questions not treated in the classical approximation theory. In this note, we shall

* This research was supported by the Office of Naval Research Contract N0014-91-J1343.

be interested in the rate of approximation possible by such redundant families. We shall derive estimates for the rate of approximation for certain classes of functions and study certain greedy algorithms which are commonly used for constructing approximants in this setting. Some of our results are preliminary. Our main desire is to focus attention on this type of approximation.

We shall restrict our discussion in this paper to the case where approximation takes place in a real, separable Hilbert space H equipped with an inner product $\langle \cdot, \cdot \rangle$ and the norm $\|x\| := \langle x, x \rangle^{1/2}$. We can formulate our approximation problem in the following general way. We say a set of functions $\mathcal{D}$ from H is a *dictionary* if each $g \in \mathcal{D}$ has norm one ($\|g\| = 1$) and

$$g \in \mathcal{D} \quad \text{implies} \quad -g \in \mathcal{D}.$$

1.1. *Best approximation using at most m dictionary elements*

We let $\Sigma_m(\mathcal{D})$ denote the collection of all functions in H which can be expressed as a linear combination of at most m elements of $\mathcal{D}$. Thus each function $s \in \Sigma_m := \Sigma_m(\mathcal{D})$ can be written in the form

$$s = \sum_{g \in \Lambda} c_g g, \quad \Lambda \subset \mathcal{D}, \quad |\Lambda| \leq m, \tag{1.1}$$

with the $c_g \in \mathbb{R}$. In some cases, it may be possible to write an element from $\Sigma_m(\mathcal{D})$ in the form (1.1) in more than one way. The space Σ_m is not linear: the sum of two functions from Σ_m is generally not in Σ_m.

For a function $f \in H$, we define its approximation error

$$\sigma_m(f) := \sigma_m(f, \mathcal{D}) := \inf_{s \in \Sigma_m} \|f - s\|. \tag{1.2}$$

We shall be interested in estimates for σ_m (from above and below).

In order to orient our discussion, we begin with the case when $\mathcal{D} = \mathcal{B} = \{\pm h_k\}_{k=1}^{\infty}$ where $\{h_k\}_{k=1}^{\infty}$ is an orthonormal basis for H. (We say $\mathcal{D}$ is given by an orthonormal basis; the inclusion of both $\pm h_k$ in $\mathcal{D}$ is for later notational convenience.) In this case, much is already known about the above approximation problem. Any element $f \in H$ has the orthogonal expansion

$$f = \sum_j \langle f, h_j \rangle h_j$$

and

$$\|f\|^2 = \sum_j |\langle f, h_j \rangle|^2.$$

It follows that a best approximation s_m to f from $\Sigma_m(\mathcal{B})$ is obtained as follows. We order the coefficients $\langle f, h_j \rangle$ according to the absolute value of their size and we choose $\Lambda := \Lambda_m$ as a set of m indices j for which $|\langle f, h_j \rangle|$ is largest. Then,

$$s_m = \sum_{j \in \Lambda} \langle f, h_j \rangle h_j$$

is a best approximation to f from $\Sigma_m(\mathcal{B})$ and

$$\sigma_m(f)^2 = \|f - s_m\|^2 = \sum_{j \notin \Lambda} |\langle f, h_j \rangle|^2.$$

The set Λ_m and the approximant s_m are not unique because of possible ties in the ordering of the coefficients. Of course $\sigma_m(f)$ is always uniquely defined.

We turn now to the question of what can be said about the rate of decrease of $\sigma_m(f)$. Of course to say anything about the decrease of $\sigma_m(f)$ we need to assume something about f. Typical results in approximation theory of this type relate the rate of approximation to the smoothness of a function.

For a general dictionary $\mathcal{D}$, and for any $\tau > 0$, we define the class of functions

$$\mathcal{A}_\tau^o(\mathcal{D}, M) := \{f \in H : f = \sum_{k \in \Lambda} c_k w_k \quad w_k \in \mathcal{D}, \ |\Lambda| < \infty \text{ and } \sum_{k \in \Lambda} |c_k|^\tau \le M^\tau\}, \tag{1.3}$$

and we define $A_\tau(\mathcal{D}, M)$ as the closure (in H) of $A_\tau^o(\mathcal{D}, M)$. Furthermore, we define $A_\tau(\mathcal{D})$ as the union of the classes $A_\tau(\mathcal{D}, M)$ over all $M > 0$. For $f \in A_\tau(\mathcal{D})$, we define the "semi-norm"

$$|f|_{\mathcal{A}_\tau(\mathcal{D})}$$

as the smallest M such that $f \in A_\tau(\mathcal{D}, M)$. In the case that $\mathcal{D} = \mathcal{B}$ is given by an orthonormal basis $\{h_k\}_{k=1}^\infty$, then $f \in \mathcal{A}_\tau(\mathcal{B})$ if and only if

$$\sum_k |\langle f, h_k \rangle|^\tau$$

is finite and this last expression equals $|f|^\tau_{\mathcal{A}_\tau(\mathcal{B})}$.

In this case, we can characterize certain approximation orders by the spaces $\mathcal{A}_\tau$. The first result of this type was the result of Stechkin [7]

$$f \in \mathcal{A}_1(\mathcal{B}) \quad \leftrightarrow \quad \sum_{m=1}^\infty m^{1/2} \sigma_m(f, \mathcal{B}) \frac{1}{m} < \infty.$$

A slight modification of Stechkin's proof gives the following generalization (see section 2).

Theorem 1.1

In the case $\mathcal{D} = \mathcal{B}$ is given by an orthonormal basis $\{h_k\}_{k=1}^\infty$ for H, for each $\alpha > 0$, and $\tau := (\alpha + 1/2)^{-1}$, we have

$$\sum_{m=1}^\infty [m^\alpha \sigma_m(f, \mathcal{B})]^\tau \frac{1}{m} < \infty \quad \longleftrightarrow \quad f \in A_\tau(\mathcal{B}). \tag{1.4}$$

Thus, theorem 1.1 provides a characterization of functions with an approximation order like $O(m^{-\alpha})$. We mention two interesting examples of this theorem. If $\mathcal{B}$ is given by an orthonormal wavelet basis in $L_2(\mathbb{R})$, then theorem 1.1 is a

special case of (1.3) in [3] and the class $\mathcal{A}_\tau(\mathcal{B})$ coincides with the Besov space $B^\alpha_\tau(L_\tau(\mathbb{R}))$. If $\mathcal{B}$ is given by the Fourier basis e^{ikx}, $k \in \mathbb{Z}$, in $L_2(\mathbb{T})$, then theorem 1.1 characterizes approximation order for a function f by absolute summability of a power of its Fourier coefficients. For example, when $\alpha = 1/2$, $\tau = 1$, the class $\mathcal{A}_1(\mathcal{B})$ is the class of functions whose Fourier series converges absolutely.

As a special case of theorem 1.1, we have for $\mathcal{D} = \mathcal{B}$ the estimate

$$\sigma_m(f,\mathcal{D}) \le C|f|_{\mathcal{A}_\tau(\mathcal{D})} m^{-\alpha}, \quad \tau := (\alpha + 1/2)^{-1}. \tag{1.5}$$

It is very interesting to note that the estimate (1.5) actually holds for a general dictionary provided $\alpha \ge 1/2$ (equivalently $\tau \le 1$). In the case $\alpha = 1/2$ ($\tau = 1$) this was proved by Maurey (see [8]) and an iterative algorithm was given by Jones [4]. The case $\alpha > 1/2$ is easily derived from the case $\alpha = 1/2$ (see section 3). For $\alpha < 1/2$ ($1 \le \tau \le 2$) there seems to be no obvious analogue of (1.5).

1.2. *Greedy algorithms*

In the case that $\mathcal{D} = \mathcal{B}$, a best m-term approximation s_m is generated by the following "greedy" algorithm. We describe this algorithm for a general dictionary $\mathcal{D}$ (in which case it does not generally generate a best approximation). If $f \in H$, we let $g = g(f) \in \mathcal{D}$ be an element from $\mathcal{D}$ which maximizes $\langle f, g\rangle$. We shall assume for simplicity that such a maximizer exists; if not, suitable modifications are necessary in the algorithms that follow. We define

$$G(f) := G(f,\mathcal{D}) := \langle f, g\rangle g \tag{1.6}$$

and

$$R(f) := R(f,\mathcal{D}) := f - G(f).$$

Pure Greedy Algorithm

We define $R_0(f) := R_0(f,\mathcal{D}) := f$ and $G_0(f) := 0$. Then, for each $m \ge 1$, we inductively define

$$\begin{aligned} G_m(f) &:= G_m(f,\mathcal{D}) := G_{m-1}(f) + G(R_{m-1}(f)), \\ R_m(f) &:= R_m(f,\mathcal{D}) := f - G_m(f) = R(R_{m-1}(f)). \end{aligned} \tag{1.7}$$

It is clear that in the case of the orthonormal basis $\mathcal{D} = \mathcal{B}$ then $G_m(f) = s_m(f)$ and $R_m(f) = f - s_m(f)$ and

$$\sigma_m(f,\mathcal{B}) = \|f - G_m(f)\| = \|R_m(f)\|.$$

The above algorithm is greedy in the sense that at each iteration it approximates the residual $R_m(f)$ as best possible by a single function from $\mathcal{D}$. Of course, for a general dictionary $\mathcal{D}$ (i.e. when $\mathcal{D}$ is not given by an orthonormal basis), the function $G_m(f)$ will generally not be the best m-term approximaton from $\Sigma_m(\mathcal{D})$. We refer the reader to the paper of Davis et al. [5] for an interesting study of the Pure Greedy Algorithm.

It is interesting therefore to ask what rate of approximation is achievable by specific numerical algorithms such as the Greedy Algorithm. We shall prove in section 3 that for a general dictionary $\mathcal{D}$, the Pure Greedy Algorithm provides the following estimate:

$$\|f - G_m(f)\| \le |f|_{A_1(\mathcal{D})} m^{-1/6}. \tag{1.8}$$

We see here that the estimate (1.8) is not as good as the estimate (1.5) and it remains an open question whether (1.8) can be improved.

On the other hand, we shall show that the Greedy Algorithm will not provide the estimate (1.5) when $\alpha > 1/2$. We show for example in section 4, that there is a dictionary $\mathcal{D} = \mathcal{B} \cup \{\pm g\}$ with $\mathcal{B}$ given by an orthonormal basis and g one additional function from H, such that the function $f = ah_1 + bh_2$ (with appropriately chosen a, b) will not be approximated with error better than $O(m^{-1/2})$.

There are modifications of the Greedy Algorithm with favorable approximation properties. We mention two of these: the Greedy Algorithm with Relaxation, and the Orthogonal Greedy Algorithm.

There are several variants of relaxed greedy algorithms. One can find some variants of relaxed greedy algorithms and their application for different dictionaries in [1] and [2]. We shall consider the following.

Relaxed Greedy Algorithm

We define $R_0^r(f) := R_0^r(f, D) := f$ and $G_0^r(f) := G_0^r(f, \mathcal{D}) := 0$. For $m = 1$, we define $G_1^r(f) := G_1^r(f, \mathcal{D}) := G_1(f)$ and $R_1^r(f) := R_1^r(f, \mathcal{D}) := R_1(f)$. Let, as before, for a function $h \in H$, $g = g(h)$ denote a function from D which maximizes $\langle h, g\rangle$. Then, for each $m \ge 2$, we inductively define

$$G_m^r(f) := G_m^r(f, \mathcal{D}) := \left(1 - \frac{1}{m}\right) G_{m-1}^r(f) + \frac{1}{m} g(R_{m-1}^r(f)),$$

$$R_m^r(f) := R_m^r(f, \mathcal{D}) := f - G_m^r(f).$$

As was pointed out to us by Andrew Barron, it was shown in [2] that for any function $f \in A_1(\mathcal{D})$, the Relaxed Greedy Algorithm provides the approximation order

$$\|f - G_m^r(f)\| \le Cm^{-1/2}, \quad m = 1, 2, \ldots.$$

For a proof of this see section 3. Thus, the Relaxed Greedy Algorithm gives a constructive proof of the estimate (1.5) in the case $\alpha = 1/2$. We should also mention that an estimate of the form $O(m^{-\min\{1/2, 1-1/p\}})$ for approximation in L_p was given in [2].

The Pure Greedy Algorithm chooses functions $g_j := G(R_j(f))$, $j = 1, \ldots, m$ to use in approximating f. One of the deficiencies of the Greedy Algorithm is that it does not provide the best approximation from the span of $g_1, \ldots, g_m$. We can modify the Greedy Algorithm as follows to remove this deficiency.

If H_0 is a finite-dimensional subspace of H, we let P_{H_0} be the orthogonal projector from H onto H_0. That is, $P_{H_0}(f)$ is the best approximation to f from H_0.

Orthogonal Greedy Algorithm
We define $R_0^o(f) := R_0^o(f, \mathcal{D}) := f$ and $G_0^o(f) := G_0^o(f, \mathcal{D}) := 0$. Then for each $m \geq 1$, we inductively define

$$\begin{aligned} H_m &:= H_m(f) := \operatorname{span}\{g(R_0^o(f)), \dots, g(R_{m-1}^o(f))\}, \\ G_m^o(f) &:= G_m^o(f, \mathcal{D}) := P_{H_m}(f), \\ R_m^o(f) &:= R_m^o(f, \mathcal{D}) := f - G_m^o(f). \end{aligned}$$

Thus, the distinction between the Orthogonal Greedy Algorithm and the Greedy Algorithm is that the Orthogonal Greedy Algorithm takes the best approximation from the functions $G(R_0^o(f)), \dots, G(R_{m-1}^o(f))$ generated at each iteration. The first step of the Orthogonal Greedy Algorithm is the same as the Pure Greedy Algorithm.

We prove in section 3 that the Orthogonal Greedy Algorithm satisfies the estimate

$$\|f - G_m^o(f, D)\| \leq |f|_{A_1(\mathcal{D})} m^{-1/2}, \tag{1.9}$$

which is the same as (1.5) for the case $\alpha = 1/2$.

2. Proof of theorem 1.1

For the sake of completeness of our discussion in this paper, we begin with the case when $\mathcal{D} = \mathcal{B}$ is given by an orthonormal basis $\{h_k\}_{k=1}^{\infty}$ for H and prove theorem 1.1. We shall use the following lemma about numerical sequences.

Lemma 2.1
If $(a_k)_{k=1}^{\infty}$ is a non-increasing sequence of nonnegative numbers and $\sigma_m^2 := \sum_{k=m}^{\infty} a_k^2$, then for any $0 < \tau < 2$ and $\alpha := 1/\tau - 1/2$, we have

$$c_1 \left(\sum_{m=1}^{\infty} a_m^{\tau} \right)^{1/\tau} \leq \left(\sum_{m=1}^{\infty} [m^{\alpha} \sigma_m]^{\tau} \frac{1}{m} \right)^{1/\tau} \leq c_2 \left(\sum_{m=1}^{\infty} a_m^{\tau} \right)^{1/\tau} \tag{2.1}$$

with the constants $c_1, c_2 > 0$ depending only on τ.

Proof
We have

$$a_{2m} \leq a_{2m-1} \leq m^{-1/2} \left(\sum_{k=m}^{2m-1} a_k^2 \right)^{1/2} \leq m^{-1/2} \sigma_m.$$

Raising this inequality to the power τ, summing, and using the fact that $-\tau/2 = \alpha\tau - 1$, we derive the left inequality in (2.1). In the other direction, we have

$$\sigma_{2^m} = \left(\sum_{k=2^m}^{\infty} a_k^2\right)^{1/2} \le \left(\sum_{k=m}^{\infty} 2^k a_{2^k}^2\right)^{1/2} \le \left(\sum_{k=m}^{\infty} 2^{k\tau/2} a_{2^k}^{\tau}\right)^{1/\tau}$$

because an ℓ_2-norm does not exceed an ℓ_τ-norm. It follows that

$$\sum_{m=1}^{\infty} 2^{m\alpha\tau}\sigma_{2^m}^{\tau} \le \sum_{m=1}^{\infty} 2^{m\alpha\tau} \sum_{k=m}^{\infty} 2^{k\tau/2} a_{2^k}^{\tau} \le c \sum_{k=1}^{\infty} 2^{k\alpha\tau} 2^{k\tau/2} a_{2^k}^{\tau} = c \sum_{k=1}^{\infty} 2^k a_{2^k}^{\tau},$$

where for the last inequality we reversed the order of summation and for the last equality we used the relation $\alpha\tau + \tau/2 = 1$. The right inequality in (2.1) now follows from the monotonicity of the sequences (σ_m) and (a_m). □

Proof of theorem 1.1
Let $f \in H$ and let $\langle f, h_k\rangle$ be its coefficients with respect to the orthonormal basis $\mathcal{B} = \{h_k\}_{k=1}^{\infty}$. We let (a_k) be the decreasing rearrangement of the sequence $(|\langle f, h_k\rangle|)$. The numbers σ_m as defined in lemma 2.1 are the same as $\sigma_{m-1}(f, \mathcal{B})$, $m = 1, 2 \ldots$. Thus theorem 1.1 follows from lemma 2.1. □

3. Upper bounds for approximation by general dictionaries

We shall next discuss approximation from a general dictionary $\mathcal{D}$. We begin with a discussion of the approximation properties of the Relaxed Greedy Algorithm. The result we give below in theorem 3.2 is known and can be found for example in the papers of Jones [4] in a different form. We begin with the following elementary lemma about numerical sequences.

Lemma 3.1
If $A > 0$ and (a_m) is a sequence of nonnegative numbers satisfying $a_1 \le A$ and

$$a_m \le a_{m-1} - \frac{2}{m} a_{m-1} + \frac{A}{m^2}, \quad m = 2, 3, \ldots, \tag{3.1}$$

then

$$a_m \le \frac{A}{m}. \tag{3.2}$$

Proof
The proof is by induction. Suppose we have

$$a_{m-1} \le \frac{A}{m-1}$$

for some $m \geq 2$. Then from our assumption (3.1), we have

$$a_m \leq \frac{A}{m-1}\left(1-\frac{2}{m}\right)+\frac{A}{m^2} = A\left(\frac{1}{m}-\frac{1}{(m-1)m}+\frac{1}{m^2}\right) \leq \frac{A}{m}.$$

If $f \in \mathcal{A}_1^o(\mathcal{D},1)$, then $f = \sum_j c_j g_j$, for some $g_j \in \mathcal{D}$ and with $\sum_j |c_j| \leq 1$. Since the functions g_j all have norm one, it follows that

$$\|f\| \leq \sum_j |c_j| \|g_j\| \leq 1.$$

Since the functions $g \in \mathcal{D}$ have norm one, it follows that $G_1^r(f) = G_1(f)$ also has norm at most one. By induction, we find that $\|G_m^r(f)\| \leq 1$, $m \geq 1$. □

Theorem 3.2

For the Relaxed Greedy Algorithm we have for each $f \in \mathcal{A}_1(\mathcal{D},1)$, the estimate

$$\|f - G_m^r(f)\| \leq \frac{2}{\sqrt{m}}, \quad m \geq 1. \tag{3.3}$$

Proof

We use the abbreviated notation $r_m := G_m^r(f)$ and $g_m := g(R_{m-1}^r(f))$. From the definition of r_m, we have

$$\|f - r_m\|^2 = \|f - r_{m-1}\|^2 + \frac{2}{m}\langle f - r_{m-1}, r_{m-1} - g_m\rangle + \frac{1}{m^2}\|r_{m-1} - g_m\|^2. \tag{3.4}$$

The last term on the right hand side of (3.4) does not exceed $4/m^2$. For the middle term, we have

$$\begin{aligned}\langle f - r_{m-1}, r_{m-1} - g_m\rangle &= \inf_{g\in\mathcal{D}}\langle f - r_{m-1}, r_{m-1} - g\rangle \\ &= \inf_{\phi\in\mathcal{A}_1(\mathcal{D},1)}\langle f - r_{m-1}, r_{m-1} - \phi\rangle \\ &\leq \langle f - r_{m-1}, r_{m-1} - f\rangle \\ &= -\|f - r_{m-1}\|^2.\end{aligned}$$

Here, to derive the second inequality we use the fact that $\phi = \sum c_k g_k$ with $g_k \in \mathcal{D}$, $c_k \geq 0$, and $\sum c_k = 1$. Returning to (3.4), we obtain

$$\|f - r_m\|^2 \leq \left(1-\frac{2}{m}\right)\|f - r_{m-1}\|^2 + \frac{4}{m^2}.$$

Thus the theorem follows from lemma 3.1 with $A = 4$ and $a_m := \|f - r_m\|^2$. □

Remark

The Relaxed Greedy Algorithm is not homogeneous. We can make it homogeneous by changing the definition for $m \geq 2$ as follows:

$$G_m^r(f) := \left(1-\frac{1}{m}\right)G_{m-1}^r(f) + \frac{\|f\|_{\mathcal{A}_1(\mathcal{D})}}{m} g(R_{m-1}^r(f)).$$

For this modification, we can prove (3.3) with the right hand side multiplied by $\|f\|_{\mathcal{A}_1(\mathcal{D})}$.

With this last theorem, we can prove now a general estimate for the error in approximation of functions $f \in A_\tau(\mathcal{D})$, $\tau \leq 1$.

Theorem 3.3

If $f \in \mathcal{A}_\tau(\mathcal{D})$, $\tau \leq 1$, then for $\alpha := 1/\tau - 1/2$, we have

$$\sigma_m(f, \mathcal{D}) \leq c|f|_{\mathcal{A}_\tau(\mathcal{D})} m^{-\alpha}, \quad m = 1, 2, \ldots, \tag{3.5}$$

where c depends on τ if τ is small.

Proof

It is enough to prove (3.5) for functions f which are a finite sum $f = \sum_j c_j g_j$, $g_j \in \mathcal{D}$, with $\sum_j |c_j|^\tau \leq M^\tau$. Indeed, these functions are dense in $\mathcal{A}_\tau(\mathcal{D}, M)$ (by the definition of $\mathcal{A}_\tau(\mathcal{D}, M)$). It is also enough to prove (3.5) for $m = 2n$ even. Without loss of generality we can assume that the c_j are positive and nonincreasing. We let $s_1 := \sum_{j=1}^n c_j g_j$ and $R_1 := f - s_1 = \sum_{j>n} c_j g_j$. Now,

$$c_n^\tau \leq \frac{1}{n} \sum_{j=1}^n |c_j|^\tau \leq \frac{M^\tau}{n}.$$

Hence $c_j \leq M n^{-1/\tau}$, $j > n$. It follows that

$$\sum_{j>n} c_j = \sum_{j>n} c_j^{1-\tau} c_j^\tau \leq M^{1-\tau} n^{1-1/\tau} \sum_{j>n} c_j^\tau \leq M n^{1-1/\tau}.$$

Hence, R_1 is in $\mathcal{A}_1(\mathcal{D}, M n^{1-1/\tau})$. According to theorem 3.2, there is a function s_2 which is a linear combination of at most n of the $g \in \mathcal{D}$ such that

$$\|f - (s_1 + s_2)\| = \|R_1 - s_2\| \leq 2Mn^{1-1/\tau} n^{-1/2} = 2Mn^{-\alpha},$$

and (3.5) follows. □

We now turn our analysis to the approximation properties of the Pure Greedy Algorithm and the Orthogonal Greedy Algorithm.

We shall need the following simple known lemma.

Lemma 3.4

Let $\{a_m\}_{m=1}^\infty$ be a sequence of non-negative numbers satisfying the inequalities

$$a_1 \leq A, \quad a_{m+1} \leq a_m(1 - a_m/A), \quad m = 1, 2, \ldots.$$

Then we have for each m

$$a_m \leq A/m.$$

Proof

The proof is by induction on m. For $m = 1$ the statement is true by assumption. We

assume $a_m \le A/m$ and prove that $a_{m+1} \le A/(m+1)$. If $a_{m+1} = 0$ this statement is obvious. Assume therefore that $a_{m+1} > 0$. Then, we have

$$a_{m+1}^{-1} \ge a_m^{-1}(1 - a_m/A)^{-1} \ge a_m^{-1}(1 + a_m/A) = a_m^{-1} + A^{-1} \ge (m+1)A^{-1},$$

which implies $a_{m+1} \le A/(m+1)$. □

We want next to estimate the decrease in error provided by one step of the Pure Greedy Algorithm. Let $\mathcal{D}$ be an arbitrary dictionary. If $f \in H$ and

$$\rho(f) := \langle f, g(f)\rangle / \|f\|, \tag{3.6}$$

where as before $g(f) \in \mathcal{D}$ satisfies

$$\langle f, g(f)\rangle = \sup_{g \in \mathcal{D}} \langle f, g\rangle.$$

Then,

$$R(f)^2 = \|f - G(f)\|^2 = \|f\|^2(1 - \rho(f)^2). \tag{3.7}$$

The larger $\rho(f)$ is, the better the decrease of the error in the step of the Pure Greedy Algorithm. The following lemma estimates $\rho(f)$ from below.

Lemma 3.5
If $f \in \mathcal{A}_1(\mathcal{D}, M)$, then

$$\rho(f) \ge \|f\|/M. \tag{3.8}$$

Proof
It is sufficient to prove (3.8) for $f \in A_1^o(\mathcal{D}, M)$ since the general result follows from this by taking limits. We can write $f = \sum c_k g_k$ where this sum has a finite number of terms and $g_k \in \mathcal{D}$ and $\sum |c_k| \le M$. Hence,

$$\|f\|^2 = \langle f, f\rangle = \langle f, \sum c_k g_k\rangle = \sum c_k \langle f, g_k\rangle \le M\rho(f)\|f\|,$$

and (3.8) follows. □

Theorem 3.6
Let $\mathcal{D}$ be an arbitrary dictionary in H. Then for each $f \in A_1(\mathcal{D})$ we have

$$\|f - G_m(f, \mathcal{D})\| \le |f|_{A_1(\mathcal{D})} m^{-1/6}.$$

Proof
It is enough to prove the theorem for $f \in \mathcal{A}_1(\mathcal{D}, 1)$; the general result then follows by rescaling. We shall use the abbreviated notation $f_m := R_m(f)$ for the residual. Let

$$a_m := \|f_m\|^2 = \|f - G_m(f, \mathcal{D})\|^2, \quad m = 0, 1, \ldots, \quad f_0 := f$$

and define the sequence (b_m) by

$$b_0 := 1, \quad b_{m+1} := b_m + \rho(f_m)\|f_m\|, \quad m = 0, 1, \dots.$$

Since $f_{m+1} := f_m - \rho(f_m)\|f_m\|g(f_m)$, we obtain by induction that

$$f_m \in \mathcal{A}_1(\mathcal{D}, b_m), \quad m = 0, 1, \dots,$$

and consequently we have the following relations for $m = 0, 1, \dots$

$$a_{m+1} = a_m(1 - \rho(f_m)^2), \tag{3.9}$$

$$b_{m+1} = b_m + \rho(f_m)a_m^{1/2}, \tag{3.10}$$

$$\rho(f_m) \geq a_m^{1/2}b_m^{-1}. \tag{3.11}$$

The last two relations give

$$b_{m+1} = b_m(1 + \rho(f_m)a_m^{1/2}b_m^{-1}) \leq b_m(1 + \rho(f_m)^2). \tag{3.12}$$

Combining this inequality with (3.9) we find

$$a_{m+1}b_{m+1} \leq a_m b_m(1 - \rho(f_m)^4),$$

which in turn implies for all m

$$a_m b_m \leq a_0 b_0 = \|f\|^2 \leq 1. \tag{3.13}$$

Further, using (3.9) and (3.11) we get

$$a_{m+1} = a_m(1 - \rho(f_m)^2) \leq a_m(1 - a_m/b_m^2).$$

Since $b_m \leq b_{m+1}$, this gives

$$a_{m+1}b_{m+1}^{-2} \leq a_m b_m^{-2}(1 - a_m b_m^{-2}).$$

Applying lemma 3.4 to the sequence $(a_m b_m^{-2})$ we obtain

$$a_m b_m^{-2} \leq m^{-1}. \tag{3.14}$$

The relations (3.13) and (3.14) imply

$$a_m^3 = (a_m b_m)^2 a_m b_m^{-2} \leq m^{-1}.$$

In other words,

$$\|f_m\| = a_m^{1/2} \leq m^{-1/6},$$

which proves the theorem. □

The next theorem estimates the error in approximation by the Orthogonal Greedy Algorithm.

Theorem 3.7
Let $\mathcal{D}$ be an arbitrary dictionary in H. Then for each $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have

$$\|f - G_m^o(f, \mathcal{D})\| \leq Mm^{-1/2}.$$

Proof
The proof of this theorem is similar to the proof of theorem 3.6 but technically even simpler. We can again assume that $M = 1$. We let $f_m^o := R_m^o(f)$ be the residual in the Orthogonal Greedy Algorithm. Then, from the definition of Orthogonal Greedy Algorithm, we have

$$\|f_{m+1}^o\| \leq \|f_m^o - G_1(f_m^o, \mathcal{D})\|. \tag{3.15}$$

From (3.7), we obtain

$$\|f_{m+1}^o\|^2 \leq \|f_m^o\|^2(1 - \rho(f_m^o)^2). \tag{3.16}$$

By the definition of the Orthogonal Greedy Algorithm, $G_m^o(f) = P_{H_m} f$ and hence $f_m^o = f - G_m^o(f)$ is orthogonal to $G_m^o(f)$. Using this as in the proof of lemma 3.5, we obtain

$$\|f_m^o\|^2 = \langle f_m^o, f \rangle \leq \rho(f_m^o)\|f_m^o\|.$$

Hence,

$$\rho(f_m^o) \geq \|f_m^o\|.$$

Using this inequality in (3.16), we find

$$\|f_{m+1}^o\|^2 \leq \|f_m^o\|^2(1 - \|f_m^o\|^2).$$

In order to complete the proof it remains to apply lemma 3.4 with $A = 1$ and $a_m = \|f_m^o\|^2$. □

4. A lower estimate for the Pure Greedy Algorithm

In this section we shall give an example which shows that replacing a dictionary $\mathcal{B}$ given by an orthogonal basis by a nonorthogonal redundant dictionary $\mathcal{D}$ may damage the efficiency of the Pure Greedy Algorithm. The dictionary $\mathcal{D}$ in our example differs from dictionary $\mathcal{B}$ by the addition of the two elements $\pm g$ for a certain suitably chosen g.

Let $\{h_k\}_{k=1}^{\infty}$ be an orthonormal basis in a Hilbert space H and let $\mathcal{B} = \{\pm h_k\}_{k=1}^{\infty}$ be the corresponding dictionary. Consider the following element

$$g := Ah_1 + Ah_2 + aA \sum_{k \geq 3} (k(k+1))^{-1/2} h_k$$

with

$$A := (33/89)^{1/2} \quad \text{and} \quad a := (23/11)^{1/2}.$$

Then, $\|g\| = 1$. We define the dictionary $\mathcal{D} = \mathcal{B} \cup \{\pm g\}$.

Theorem 4.1
For the function

$$f = h_1 + h_2,$$

which is in each space $A_\tau(D)$, $0 < \tau \leq 2$, we have

$$\|f - G_m(f)\| \geq m^{-1/2}, \quad m \geq 4. \tag{4.1}$$

Proof
We shall examine the steps of the Pure Greedy Algorithm applied to the function $f = h_1 + h_2$. We shall use the abbreviated notation $f_m := R_m(f) := f - G_m(f)$ for the residual at step m.

The first step. We have

$$\langle f, g\rangle = 2A > 1, \quad |\langle f, h_k\rangle| \leq 1, \quad k = 1, 2, \ldots.$$

This implies

$$G_1(f, \mathcal{D}) = \langle f, g\rangle g,$$

and

$$f_1 = f - \langle f, g\rangle g = (1 - 2A^2)(h_1 + h_2) - 2aA^2 \sum_{k \geq 3} (k(k+1))^{-1/2} h_k.$$

The second step. We have

$$\langle f_1, g\rangle = 0, \quad \langle f_1, h_k\rangle = (1 - 2A^2), \quad k = 1, 2, \quad \langle f_1, h_3\rangle = -aA^2 3^{-1/2}.$$

Comparing $\langle f_1, h_1\rangle$ and $|\langle f_1, h_3\rangle|$ we get

$$|\langle f_1, h_3\rangle| = (23/89)(33/23)^{1/2} > 23/89 = 1 - 2A^2 = \langle f_1, h_1\rangle.$$

This implies that the second approximation $G_1(f_1, \mathcal{D})$ is $\langle f_1, h_3\rangle h_3$ and

$$f_2 = f_1 - \langle f_1, h_3\rangle h_3 = (1 - 2A^2)(h_1 + h_2) - 2aA^2 \sum_{k \geq 4} (k(k+1))^{-1/2} h_k.$$

The third step. We have

$$\langle f_2, g\rangle = -\langle f_1, h_3\rangle\langle h_3, g\rangle = (A/2)(23/89),$$
$$\langle f_2, h_1\rangle = \langle f_2, h_2\rangle = 1 - 2A^2 = 23/89,$$
$$\langle f_2, h_4\rangle = -aA^2 5^{-1/2} = -(23/89)(99/115)^{1/2}.$$

Therefore, the third approximation should be $\langle f_2, h_1\rangle h_1$ or $\langle f_2, h_2\rangle h_2$. Let us take the first of these so that

$$f_3 = f_2 - \langle f_2, h_1\rangle h_1.$$

The fourth step. It is clear that for all $k \neq 1$ we have

$$\langle f_3, h_k\rangle = \langle f_2, h_k\rangle.$$

This equality and the calculations from step 3 show that it is sufficient to compare $\langle f_3, h_2\rangle$ and $\langle f_3, g\rangle$. We have

$$\langle f_3, g\rangle = \langle f_2, g\rangle - \langle f_2, h_1\rangle\langle h_1, g\rangle = -(23/89)(A/2).$$

This means that

$$f_4 = f_3 - \langle f_3, h_2\rangle h_2 = -2aA^2 \sum_{k\geq 4} (k(k+1))^{-1/2} h_k. \tag{4.2}$$

The *m*-th step ($m > 4$). We prove by induction that for all $m \geq 4$ we have

$$f_m = -2aA^2 \sum_{k\geq m} (k(k+1))^{-1/2} h_k. \tag{4.3}$$

For $m = 4$ this relation follows from (4.2). We assume we have proved (4.3) for some m and derive that (4.3) also holds true for $m+1$. To find f_{m+1}, we have only to compare the two inner products: $\langle f_m, h_m\rangle$ and $\langle f_m, g\rangle$. We have

$$|\langle f_m, h_m\rangle| = 2aA^2(m(m+1))^{-1/2}$$

and

$$|\langle f_m, g\rangle| = 2a^2A^3 \sum_{k\geq m} (k(k+1))^{-1} = 2a^2A^3 m^{-1}.$$

Since

$$(|\langle f_m, g\rangle|/|\langle f_m, h_m\rangle|)^2 = (aA)^2(1 + 1/m) \leq 345/356 < 1,$$

we have that

$$|\langle f_m, g\rangle| < |\langle f_m, h_m\rangle|, \quad m \geq 4.$$

This proves (4.3) with m replaced by $m+1$.

From (4.3), we obtain

$$\|f - G_m(f, \mathcal{D})\| = \|f_m\| = 2aA^2 m^{-1/2} > m^{-1/2}, \quad m \geq 4.$$

□

References

[1] A. Barron, Universal approximation bounds for superposition of a sigmoidal function, IEEE Transactions on Information Theory 39 (1993) 930–945.

[2] C. Darken, M. Donahue, L. Gurvits and E. Sontag, Rate of approximation results motivated by robust neural network learning, *6th ACM Conference on Computer Learning Theory*, ACM, 1993, pp. 303–309.

[3] R. DeVore, B. Jawerth and V. Popov, Compression of wavelet decompositions, American Journal of Mathematics 114 (1992) 737–785.

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

[5] G. Davis, S. Mallat and M. Avellaneda, Adaptive nonlinear approximations, Constructure Approximation, to appear.

[6] E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen. I, Math. Annalen 63 (1906–1907) 433–476.

[7] S.B. Stechkin, On absolute convergence of orthogonal series, Dok. Akad. Nauk SSSR 102 (1955) 37–40.

[8] G. Pisier, Remarques sur un resultat non publié de B. Maurey, *Seminaire d'Analyse Fonctionelle*, 1980–81, Ecole Polytechnique Centre de Mathematiques, Palaiseau.