IMI

Preprint Series

Industrial Mathematics Institute

1999:03

Weak greedy algorithms

V.N. Temlyakov

Department of Mathematics
University of South Carolina

WEAK GREEDY ALGORITHMS[1]

V.N.Temlyakov

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

Abstract. Theoretical greedy type algorithms are studied: a Weak Greedy Algorithm, a Weak Orthogonal Greedy Algorithm, and a Weak Relaxed Greedy Algorithm. These algorithms are defined by weaker assumptions than their analogs the Pure Greedy Algorithm, an Orthogonal Greedy Algorithm, and a Relaxed Greedy Algorithm. The weaker assumptions make these new algorithms more ready for practical implementation. We prove the convergence theorems and also give estimates for the rate of approximation by means of these algorithms. The convergence and the estimates apply to approximation from an arbitrary dictionary in a Hilbert space.

1. Introduction

We study nonlinear approximation in this paper. The basic idea behind nonlinear approximation is that the elements used in the approximation do not come from a fixed linear space but are allowed to depend on the function being approximated. The standard problem in this ragard is the problem of m-term approximation where one fixes a basis and looks to approximate a target function f by a linear combination of m terms of the basis. When the basis is a wavelet basis or a basis of other waveforms, then this type of approximation is the starting point for compression algorithms. An important feature of approximation using a basis $\Psi := \{\psi_k\}_{k=1}^\infty$ of a Banach space X is that each function $f \in X$ has a unique representation

$$f = \sum_{k=1}^{\infty} c_k(f)\psi_k \tag{1.1}$$

and we can identify f with the set of its Fourier coefficients $\{c_k(f)\}_{k=1}^\infty$. The problem of m-term approximation with regard to a basis has been studied thoroughly and rather complete results have been established (see [2], [3], [5], [6], [8], [16], [17], [20], [21], [22], [23]). In particular, it was established that the greedy type algorithm which forms a sum of m terms with the largest $\|c_k(f)\psi_k\|_X$ out of expansion (1.1) realizes in many cases almost the best m-term approximation for function classes ([6]) and even for individual functions ([21]).

Recently, there has emerged another more complicated form of nonlinear approximation which we call highly nonlinear approximation. It takes many forms but has the basic ingredient that a basis is replaced by a larger system of functions that is usually redundant. We call such systems dictionaries. Redundancy on the one hand offers much promise for greater efficiency in terms of approximation rate, but on the other hand gives rise to highly nontrivial theoretical and practical

[1]This research was supported by the National Science Foundation Grant DMS 9622925 and by ONR Grant N00014-96-1-1003

problems. The problem of characterizing approximation rate for a given function or function class is now much more substantial and results are quite fragmentary. However, such results are very important for understanding what this new type of approximation offers. Perhaps the first example of this type was considered by E. Schmidt in 1907 [19] 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

$$J_f(g) := \int_0^1 f(x,y)g(y)dy$$

with kernel $f(x,y)$. E. Schmidt [19] gave an expansion (known as the Schmidt expansion)

$$f(x,y) = \sum_{j=1}^{\infty} s_j(J_f)\phi_j(x)\psi_j(y)$$

where $\{s_j(J_f)\}$ is a nonincreasing sequence of singular numbers of J_f, i.e. $s_j(J_f) := \lambda_j(J_f^* J_f)^{1/2}$, $\{\lambda_j(A)\}$ is a sequence of eigenvalues of an operator A, J_f^* is the adjoint operator to J_f. The two sequences $\{\phi_j(x)\}$ and $\{\psi_j(y)\}$ form orthonormal sequences of eigenfunctions of the operators $J_f J_f^*$ and $J_f^* J_f$ respectively. He also proved that

$$\|f(x,y)-\sum_{j=1}^{m} s_j(J_f)\phi_j(x)\psi_j(y)\|_{L_2} = \inf_{u_j,v_j\in L_2,\quad j=1,\dots,m} \|f(x,y)-\sum_{j=1}^{m} u_j(x)v_j(y)\|_{L_2}.$$

It was understood later that the above best bilinear approximation can be realized by the following greedy algorithm. Assume c_j, $u_j(x)$, $v_j(y)$, $\|u_j\|_{L_2} = \|v_j\|_{L_2} = 1$, $j = 1,\dots,m-1$, have been constructed after $m-1$ steps of algorithm. At the m-th step we choose c_m, $u_m(x)$, $v_m(y)$, $\|u_m\|_{L_2} = \|v_m\|_{L_2} = 1$, to minimize

$$\|f(x,y) - \sum_{j=1}^{m} c_j u_j(x)v_j(y)\|_{L_2}.$$

We call this type of algorithm the Pure Greedy Algorithm (see the general definition below).

Remark 1. In this paper, we study only theoretical aspects of the efficiency of m-term approximation and possible ways to realize this efficiency. The above defined "greedy algorithm" gives a procedure to construct an approximant which turns out to be a good approximant. The procedure of constructing a greedy approximant is not a numerical algorithm ready for computational implementation. Therefore it would be more precise to call this procedure a "theoretical greedy algorithm" or "stepwise optimizing process". Keeping this remark in mind we, however, use term "greedy algorithm" in this paper because it has been used in previous papers and has become a standard name for procedures like the above and for more general procedures of this type (see for instance [5], [9]).

Another problem of this type which is well known in statistics is the projection pursuit regression problem. We formulate the related results in the function theory language. The problem is to approximate in L_2 a given function $f \in L_2$ by a sum of ridge functions, i.e. by

$$\sum_{j=1}^{m} r_j(\omega_j \cdot x), \quad x, \omega_j \in \mathbb{R}^d, \quad j = 1, \dots, m,$$

where r_j, $j = 1, \dots, m$, are univariate functions. The following greedy type algorithm (projection pursuit) was proposed in [12] to solve this problem. Assume functions $r_1, \dots, r_{m-1}$ and vectors $\omega_1, \dots, \omega_{m-1}$ have been determined after $m-1$ steps of algorithm. Choose at m-th step a unit vector ω_m and a function r_m to minimize the error

$$\|f(x) - \sum_{j=1}^{m} r_j(\omega_j \cdot x)\|_{L_2}.$$

This is one more example of Pure Greedy Algorithm. The Pure Greedy Algorithm and some other versions of greedy type algorithms have been intensively studied recently (see [1], [4], [7], [9], [10], [11], [13], [14], [15], [24]). In this paper we propose to study a modification of greedy type algorithms which makes them more ready for implementation. We call this new type of greedy algorithms Weak Greedy Algorithms.

In oder to orient the reader we remind some notations and definitions from the theory of greedy algorithms. Let H be a real Hilbert space with an inner product $\langle \cdot, \cdot \rangle$ and the norm $\|x\| := \langle x, x \rangle^{1/2}$. We say a set $\mathcal{D}$ of functions (elements) from H is a dictionary if each $g \in \mathcal{D}$ has norm one ($\|g\| = 1$) and $\overline{\operatorname{span}}\mathcal{D} = H$. We studied in [9] the following two greedy algorithms. If $f \in H$, we let $g = g(f) \in \mathcal{D}$ be the element from $\mathcal{D}$ which maximizes $|\langle f, g \rangle|$ (we make an additional assumption that a miximizer exists) and define

$$G(f) := G(f, \mathcal{D}) := \langle f, g \rangle g \tag{1.2}$$

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* $m \geq 1$, *we inductively define*

$$G_m(f) := G_m(f, \mathcal{D}) := G_{m-1}(f) + G(R_{m-1}(f))$$
$$R_m(f) := R_m(f, \mathcal{D}) := f - G_m(f) = R(R_{m-1}(f)).$$

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*

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

We remark that for each f we have

$$\|f - G_m^o(f,\mathcal{D})\| \leq \|R_{m-1}^o(f) - G_1(R_{m-1}^o(f),\mathcal{D})\|. \tag{1.3}$$

In sections 2,3,5 we study some modifications of the Pure Greedy Algorithm (PGA) and the Orthogonal Greedy Algorithm (OGA) which we call respectively Weak Greedy Algorithm (WGA) and Weak Orthogonal Greedy Algorithm (WOGA). We give now the corresponding definitions. Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 < t_k < 1$, be given.

Weak Greedy Algorithm. *We define* $f_0^\tau := f$*. Then for each* $m \geq 1$*, we inductively define:*

1). $\varphi_m^\tau \in \mathcal{D}$ *is any satisfying*

$$|\langle f_{m-1}^\tau, \varphi_m^\tau\rangle| \geq t_m \sup_{g\in\mathcal{D}} |\langle f_{m-1}^\tau, g\rangle|;$$

2).

$$f_m^\tau := f_{m-1}^\tau - \langle f_{m-1}^\tau, \varphi_m^\tau\rangle \varphi_m^\tau;$$

3).

$$G_m^\tau(f,\mathcal{D}) := \sum_{j=1}^m \langle f_{j-1}^\tau, \varphi_j^\tau\rangle \varphi_j^\tau.$$

We note that in a particular case $t_k = t$, $k = 1,2,\dots,$ this algorithm was considered in [14].

Weak Orthogonal Greedy Algorithm. *We define* $f_0^{o,\tau} := f$ *and* $f_1^{o,\tau} := f_1^\tau$*;* $\varphi_1^{o,\tau} := \varphi_1^\tau$ *where* f_1^τ, φ_1^τ *are from the above definition of WGA. Then for each* $m \geq 2$ *we inductively define:*

1). $\varphi_m^{o,\tau} \in \mathcal{D}$ *is any satisfying*

$$|\langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau}\rangle| \geq t_m \sup_{g\in\mathcal{D}} |\langle f_{m-1}^{o,\tau}, g\rangle|;$$

2).

$$G_m^{o,\tau}(f,\mathcal{D}) := P_{H_m^\tau}(f), \quad \textit{where} \quad H_m^\tau := \operatorname{span}(\varphi_1^{o,\tau},\dots,\varphi_m^{o,\tau});$$

3).

$$f_m^{o,\tau} := f - G_m^{o,\tau}(f,\mathcal{D}).$$

It is clear that G_m^τ and $G_m^{o,\tau}$ in the case $t_k = 1$, $k = 1,2,\dots,$ coincide with PGA G_m and OGA G_m^o respectively. It is also clear that WGA and WOGA are more ready for implementaion than PGA and OGA.

We turn first to formulate some theorems on convergence of WGA and WOGA. We make first some historical remarks. The weak L_2-convergence of projection pursuit was established in [13] and the strong L_2-convergence of it was proved in [14]. The proof from [14] also works in the general problem of convergence of PGA (see [18], [11]). For convergence of OGA see [11].

Theorem 1. *Assume*

$$\sum_{k=1}^{\infty} \frac{t_k}{k} = \infty. \tag{1.4}$$

Then for any dictionary $\mathcal{D}$ and any $f \in H$ we have

$$\lim_{m\to\infty} \|f - G_m^{\tau}(f,\mathcal{D})\| = 0.$$

Theorem 2. *Assume*

$$\sum_{k=1}^{\infty} t_k^2 = \infty. \tag{1.5}$$

Then for any dictionary $\mathcal{D}$ and any $f \in H$ we have

$$\lim_{m\to\infty} \|f - G_m^{o,\tau}(f,\mathcal{D})\| = 0. \tag{1.6}$$

Remark 2. It is easy to see that in the case $\mathcal{D} = \mathcal{B}$ - orthonormal basis the assumption (1.5) is also necessary for convergence (1.6) for all f.

There is one more greedy type algorithm which works well for functions from the convex hull of $0 \cup \mathcal{D}^{\pm}$, where $\mathcal{D}^{\pm} := \{\pm g, \quad g \in \mathcal{D}\}$.

For a general dictionary $\mathcal{D}$, we define the class of functions

$$\mathcal{A}_1(\mathcal{D}, M) := \{f \in H : f = \sum_{k=1}^{\infty} c_k w_k, \quad w_k \in \mathcal{D}, \quad \text{and} \quad \sum_{k=1}^{\infty} |c_k| \le M\}$$

and for $M = 1$ we denote $\mathcal{A}_1(\mathcal{D}) := \mathcal{A}_1(\mathcal{D}, 1)$.

There are several modifications of Relaxed Greedy Algorithm (see for instance [1], [9]). Before giving the definition of Weak Relaxed Greedy Algorithm (WRGA) we make one remark which helps to motivate the corresponding definition. Assume $G_{m-1} \in \mathcal{A}_1(\mathcal{D})$ is an approximant to $f \in \mathcal{A}_1(\mathcal{D})$ obtained at the $(m-1)$-th step. The major idea of relaxation in greedy algorithms is to look for an approximant at the m-th step of the form $G_m := (1-a)G_{m-1} + ag$, $g \in \mathcal{D}^{\pm}$, $0 \le a \le 1$. This form guarantees that $G_m \in \mathcal{A}_1(\mathcal{D})$. Thus we are looking for co-convex approximant. The best we can do at the m-th step is to achieve

$$\delta_m := \inf_{g\in\mathcal{D}^{\pm}, 0\le a\le 1} \|f - ((1-a)G_{m-1} + ag)\|.$$

Denote $f_n := f - G_n$, $n = 1, \dots, m$. It is clear that for a given $g \in \mathcal{D}^{\pm}$ we have

$$\inf_a \|f_{m-1} - a(g - G_{m-1})\|^2 = \|f_{m-1}\|^2 - \langle f_{m-1}, g - G_{m-1}\rangle^2 \|g - G_{m-1}\|^{-2},$$

and this inf is attained for

$$a(g) = \langle f_{m-1}, g - G_{m-1}\rangle \|g - G_{m-1}\|^{-2}.$$

Next, it is not difficult to derive from the definition of $\mathcal{A}_1(\mathcal{D})$ that for any $h \in H$ and $u \in \mathcal{A}_1(\mathcal{D})$ there exists $g \in \mathcal{D}^{\pm}$ such that

$$\langle h, g\rangle \geq \langle h, u\rangle. \tag{1.8}$$

Taking $h = f_{m-1}$ and $u = f$ we get from (1.8) that there exists $g_m \in \mathcal{D}^{\pm}$ such that

$$\langle f_{m-1}, g_m - G_{m-1}\rangle \geq \langle f_{m-1}, f - G_{m-1}\rangle = \|f_{m-1}\|^2. \tag{1.9}$$

This implies in particular that we get for g_m

$$\|g_m - G_{m-1}\| \geq \|f_{m-1}\| \tag{1.10}$$

and $0 \leq a(g_m) \leq 1$. Thus,

$$\delta_m^2 \leq \|f_{m-1}\|^2 - \frac{1}{4}\sup_{g\in\mathcal{D}^{\pm}} \langle f_{m-1}, g - G_{m-1}\rangle^2.$$

We give now the definition of two versions of WRGA.

Weak Relaxed Greedy Algorithms. *We define* $f_0^{\tau,i} := f$ *and* $G_0^{\tau,i} := 0$ *for* $i = 1, 2$. *Then for each* $m \geq 1$ *we inductively define*

1). $\varphi_m^{\tau,1} \in \mathcal{D}^{\pm}$ *is any satisfying*

$$\langle f_{m-1}^{\tau,1}, \varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\rangle \geq t_m \|f_{m-1}^{\tau,1}\|^2 \tag{1.11}$$

and

$$\|\varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\| \geq \|f_{m-1}^{\tau,1}\|; \tag{1.12}$$

$\varphi_m^{\tau,2} \in \mathcal{D}^{\pm}$ *is any satisfying*

$$\langle f_{m-1}^{\tau,2}, \varphi_m^{\tau,2} - G_{m-1}^{\tau,2}\rangle \geq t_m \|f_{m-1}^{\tau,2}\|^2. \tag{1.13}$$

2).

$$G_m^{\tau,1} := G_m^{\tau,1}(f, \mathcal{D}) := (1 - \alpha_m) G_{m-1}^{\tau,1} + \alpha_m \varphi_m^{\tau,1},$$

$$\alpha_m := \langle f_{m-1}^{\tau,1}, \varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\rangle \|\varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\|^{-2};$$

$$G_m^{\tau,2} := G_m^{\tau,2}(f, \mathcal{D}) := (1 - \beta_m) G_{m-1}^{\tau,2} + \beta_m \varphi_m^{\tau,2},$$

$$\beta_m := t_m (1 + \sum_{k=1}^{m} t_k^2)^{-1} \quad \textit{for} \quad m \geq 1.$$

3).

$$f_m^{\tau,i} := f - G_m^{\tau,i}, \quad i = 1, 2.$$

We formulate now some theorems on convergence rates of greedy type algorithms WOGA and WRGA for functions from $\mathcal{A}_1(\mathcal{D}, M)$.

Theorem 3. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Then for each $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have*

$$\|f - G_m^{o,\tau}(f, \mathcal{D})\| \le M(1 + \sum_{k=1}^{m} t_k^2)^{-1/2}.$$

Theorem 4. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Then for each $f \in \mathcal{A}_1(\mathcal{D})$ we have*

$$\|f - G_m^{\tau,1}(f, \mathcal{D})\| \le 2(1 + \sum_{k=1}^{m} t_k^2)^{-1/2}; \tag{1.14}$$

$$\|f - G_m^{\tau,2}(f, \mathcal{D})\| \le 2(1 + \sum_{k=1}^{m} t_k^2)^{-1/2}. \tag{1.15}$$

2. Proof of Theorem 1

The following two lemmas imply Theorem 1.

Lemma 2.1. *Assume that (1.5) is satisfied. Then if $\{f_m^\tau\}_{m=1}^\infty$ converges it converges to zero.*

Lemma 2.2. *Assume (1.4) is satisfied. Then $\{f_m^\tau\}_{m=1}^\infty$ converges.*

Proof of Lemma 2.1. We prove this lemma by contradiction. Assume $f_m^\tau \to u \neq 0$ as $m \to \infty$. It is clear that

$$\sup_{g \in \mathcal{D}} |\langle u, g\rangle| \ge 2\delta$$

with some $\delta > 0$. Therefore, there exists N such that for all $m \ge N$ we have

$$\sup_{g \in \mathcal{D}} |\langle f_m^\tau, g\rangle| \ge \delta.$$

From the definition of WGA we get for all $m > N$

$$\|f_m^\tau\|^2 = \|f_{m-1}^\tau\|^2 - |\langle f_{m-1}^\tau, \varphi_m^\tau\rangle|^2 \le \|f_N^\tau\|^2 - \delta^2 \sum_{k=N+1}^{m} t_k^2,$$

what contradicts (1.5). □

Proof of Lemma 2.2. It is easy to derive from the definition of WGA the following two relations

$$f_m^\tau = f - \sum_{j=1}^{m} \langle f_{j-1}^\tau, \varphi_j^\tau\rangle \varphi_j^\tau, \tag{2.1}$$

$$\|f_m^\tau\|^2 = \|f\|^2 - \sum_{j=1}^{m} |\langle f_{j-1}^\tau, \varphi_j^\tau\rangle|^2. \tag{2.2}$$

Denote $a_j := |\langle f^\tau_{j-1}, \varphi^\tau_j\rangle|$. We get from (2.2) that

$$\sum_{j=1}^\infty a_j^2 \le \|f\|^2. \tag{2.3}$$

We take any two indecies $n < m$ and consider

$$\|f^\tau_n - f^\tau_m\|^2 = \|f^\tau_n\|^2 - \|f^\tau_m\|^2 - 2\langle f^\tau_n - f^\tau_m, f^\tau_m\rangle.$$

Denote

$$\theta^\tau_{n,m} := |\langle f^\tau_n - f^\tau_m, f^\tau_m\rangle|.$$

Using (2.1) and the definition of the WGA we get for all $n < m$ that

$$\theta^\tau_{n,m} \le \sum_{j=n+1}^m |\langle f^\tau_{j-1}, \varphi^\tau_j\rangle||\langle f^\tau_m, \varphi^\tau_j\rangle| \le \frac{a_{m+1}}{t_{m+1}} \sum_{j=1}^{m+1} a_j. \tag{2.4}$$

We need now a property of the l_2-sequences.

Lemma 2.3 (V.T. and S.V. Konyagin). *Assume $y_j \ge 0$, $j = 1, 2, \dots$, and*

$$\sum_{k=1}^\infty \frac{t_k}{k} = \infty, \qquad \sum_{j=1}^\infty y_j^2 < \infty.$$

Then

$$\varliminf_{n\to\infty} \frac{y_n}{t_n} \sum_{j=1}^n y_j = 0.$$

Proof. Consider a series

$$\sum_{n=1}^\infty \frac{t_n}{n} \frac{y_n}{t_n} \sum_{j=1}^n y_j. \tag{2.5}$$

We shall prove that this series converges. It is clear that convergence of this series together with the assumption $\sum_{k=1}^\infty t_k/k = \infty$ imply the statement of Lemma 2.3.

We use the following known fact. If $\{y_j\}_{j=1}^\infty \in l_2$ then $\{n^{-1}\sum_{j=1}^n y_j\}_{n=1}^\infty \in l_2$ (see [25,Ch.1,S.9]). We have by Cauchy inequality

$$\sum_{n=1}^\infty \frac{t_n}{n} \frac{y_n}{t_n} \sum_{j=1}^n y_j \le (\sum_{n=1}^\infty y_n^2)^{1/2} (\sum_{n=1}^\infty (n^{-1} \sum_{j=1}^n y_j)^2)^{1/2} < \infty.$$

This completes the proof of Lemma 2.3. $\square$

The relation (2.4) and Lemma 2.3 imply that

$$\varliminf_{m\to\infty} \max_{n<m} \theta^\tau_{n,m} = 0.$$

It remains to use the following simple lemma.

Lemma 2.4. *Let in a Banach space X a sequence $\{x_n\}_{n=1}^\infty$ be given. Assume that for any k,l we have*

$$\|x_k - x_l\|^2 = y_k - y_l + \theta_{k,l},$$

with $\{y_n\}_{n=1}^\infty$ is a convergent sequence of real numbers and $\theta_{k,l}$ satisfying the property

$$\varliminf_{l\to\infty} \max_{k<l} \theta_{k,l} = 0.$$

Then $\{x_n\}_{n=1}^\infty$ converges.

3. Convergence and rate of approximation of WOGA

We begin this section with the proof of Theorem 2. Let $f \in H$ and $\varphi_1^{o,\tau}, \varphi_2^{o,\tau}, \dots$ are from the definition of WOGA. Denote

$$H_n := H_n^\tau = \operatorname{span}(\varphi_1^{o,\tau}, \dots, \varphi_n^{o,\tau}).$$

It is clear that $H_n \subset H_{n+1}$ and therefore $\{P_{H_n}(f)\}$ converges to some function v. We prove that $v = f$. Assume the contrary $v \neq f$. Denote $u := f - v$. Then similarly to the proof of Lemma 2.1 there exist $\delta > 0$ and N such that for all $m \geq N$ we have

$$\sup_{g\in\mathcal{D}} |\langle f_m^{o,\tau}, g\rangle| \geq \delta.$$

Next, alike (1.3) we have

$$\|f_m^{o,\tau}\|^2 \leq \|f_{m-1}^{o,\tau}\|^2 - t_m^2 (\sup_{g\in\mathcal{D}} |\langle f_{m-1}^{o,\tau}, g\rangle|)^2 \leq \|f_N^{o,\tau}\|^2 - \delta^2 \sum_{k=N+1}^{m} t_k^2$$

what contradicts divergence of $\sum_k t_k^2$. $\square$

We proceed now to the proof of Theorem 3.

Proof of Theorem 3. We assume for simplicity $M = 1$. From the definition of WOGA we have

$$\begin{aligned} \|f_m^{o,\tau}\|^2 &\leq \|f_{m-1}^{o,\tau} - \langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau}\rangle \varphi_m^{o,\tau}\|^2 = \\ &\|f_{m-1}^{o,\tau}\|^2 - \langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau}\rangle^2 \leq \|f_{m-1}^{o,\tau}\|^2 - t_m^2 \sup_{g\in\mathcal{D}} |\langle f_{m-1}^{o,\tau}, g\rangle|^2. \end{aligned} \tag{3.1}$$

Using (1.8) we get

$$\sup_{g\in\mathcal{D}} |\langle f_{m-1}^{o,\tau}, g\rangle| \geq \langle f_{m-1}^{o,\tau}, f\rangle = \|f_{m-1}^{o,\tau}\|^2. \tag{3.2}$$

Combining (3.1) with (3.2) we obtain

$$\|f_m^{o,\tau}\|^2 \leq \|f_{m-1}^{o,\tau}\|^2 (1 - t_m^2 \|f_{m-1}^{o,\tau}\|^2). \tag{3.3}$$

It remains to use the following lemma.

Lemma 3.1. *Let $\{a_m\}_{m=0}^\infty$ be a sequence of nonnegative numbers satisfying the inequalities*

$$a_0 \leq A, \qquad a_m \leq a_{m-1}(1 - t_m^2 a_{m-1}/A), \quad m = 1, 2, \dots.$$

Then we have for each m

$$a_m \leq A(1 + \sum_{k=1}^m t_k^2)^{-1}.$$

Proof of this lemma repeats the proof of Lemma 3.4 from [9].

4. Proof of Theorem 4 and some comments

We begin with consideration of the first type of WRGA. From its definition we obtain

$$\|f_m^{\tau,1}\|^2 = \|f_{m-1}^{\tau,1}\|^2 - \langle f_{m-1}^{\tau,1}, \varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\rangle^2 \|\varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\|^{-2} \le$$

$$\le \|f_{m-1}^{\tau,1}\|^2 - \frac{1}{4}\langle f_{m-1}^{\tau,1}, \varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\rangle^2 \le$$

$$\le \|f_{m-1}^{\tau,1}\|^2 (1 - \frac{t_m^2}{4}\|f_{m-1}^{\tau,1}\|^2).$$

Using Lemma 3.1 we get from here

$$\|f_m^{\tau,1}\|^2 \le 4(1 + \sum_{k=1}^m t_k^2)^{-1}.$$

The property (1.12) guarantees that $0 \le \alpha_m \le 1$ and $G_m^{\tau,1} \in \mathcal{A}_1(\mathcal{D})$.

We proceed now to the second variant of WRGA. First of all it is clear from the definition of β_m that $0 \le \beta_m \le 1$ and hence $G_m^{\tau,2} \in \mathcal{A}_1(\mathcal{D})$ for all m. We estimate now $\|f_m^{\tau,2}\|^2$. We use the abbreviated notations:

$$f_m := f_m^{\tau,2}; \quad G_m := G_m^{\tau,2}(f, \mathcal{D}); \quad \varphi_m := \varphi_m^{\tau,2}.$$

Then we have

$$\|f_m\|^2 = \|f_{m-1} - \beta_m(\varphi_m - G_{m-1})\|^2 = \tag{4.1}$$

$$= \|f_{m-1}\|^2 - 2\beta_m \langle f_{m-1}, \varphi_m - G_{m-1}\rangle + \beta_m^2 \|\varphi_m - G_{m-1}\|^2 \le$$

$$\le \|f_{m-1}\|^2 - 2\beta_m t_m \|f_{m-1}\|^2 + 4\beta_m^2.$$

We complete the proof of (1.15) by induction. For $m = 0$ it is trivial. For $m = 1$ it can be checked directly similarly to (4.1) that $\|f_1\| \le 1$. Denote $T_m := 1 + \sum_{k=1}^m t_k^2$ and assume for $m \ge 2$

$$\|f_{m-1}\|^2 \le 4T_{m-1}^{-1}.$$

Taking into account that $\beta_m = t_m/T_m$ and $1 - 2\beta_m t_m \ge 0$ we get from (4.1)

$$\|f_m\|^2 T_m \le 4(1 - 2\beta_m t_m)T_m/T_{m-1} + 4\beta_m^2 T_m =$$

$$= 4(1 - t_m^2/T_{m-1} + t_m^2/T_m) \le 4.$$

This completes the proof of Theorem 4. $\Box$

We note that the generality of a sequence $\tau := \{t_k\}_{k=1}^\infty$ in WRGA allows us to design a simple algorithm, which does not require solving optimization problems and gives a bound for its error of approximation.

Co-convex Algorithm. *Take $f \in \mathcal{A}_1(\mathcal{D})$.*

Step 1. *Find $\varphi_1 \in \mathcal{D}^\pm$ such that $\langle f, \varphi_1\rangle > 0$ and define*

$$t_1 := \min\{1, \langle f, \varphi_1\rangle \|f\|^{-2}\}; \quad G_1 := \beta_1\varphi_1; \quad f_1 := f - G_1; \quad \beta_1 = t_1(1+t_1^2)^{-1}.$$

Step 2. *Find $\varphi_2 \in \mathcal{D}^\pm$ such that $\langle f_1, \varphi_2 - G_1\rangle > 0$ and define*

$$t_2 := \min\{1, \langle f_1, \varphi_2 - G_1\rangle \|f_1\|^{-2}\}; \quad \beta_2 := t_2(1+t_1^2+t_2^2)^{-1};$$

$$G_2 := (1-\beta_2)G_1 + \beta_2\varphi_2; \quad f_2 := f - G_2.$$

$$\dots$$

Step m. *Find $\varphi_m \in \mathcal{D}^\pm$ such that $\langle f_{m-1}, \varphi_m - G_{m-1}\rangle > 0$ and define*

$$t_m := \min\{1, \langle f_{m-1}, \varphi_m - G_{m-1}\rangle \|f_{m-1}\|^{-2}\}; \quad \beta_m := t_m(1+\sum_{k=1}^m t_k^2)^{-1};$$

$$G_m := (1-\beta_m)G_{m-1} + \beta_m\varphi_m; \quad f_m := f - G_m.$$

After m steps we have the following error bound

$$\|f_m\| \le 2(1+\sum_{k=1}^m t_k^2)^{-1/2}. \tag{4.2}$$

The estimate (4.2) follows from the observation that the Co-convex Algorithm is WRGA of the second type with $\tau = \{t_k\}_{k=1}^\infty$ and from Theorem 4. The above defined algorithm does not contain an optimization part. However, it is clear that the bigger $\langle f_{j-1}, \varphi_j - G_{j-1}\rangle$, $j = 1, \dots, m$, the better the error estimate.

5. Rate of approximation for WGA

We prove in this section the following theorem.

Theorem 5.1. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume $\tau := \{t_k\}_{k=1}^\infty$ is a nonincreasing sequence. Then for $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have*

$$\|f - G_m^\tau(f,\mathcal{D})\| \le M(1+\sum_{k=1}^m t_k^2)^{-t_m/2(2+t_m)}. \tag{5.1}$$

Proof. It is clear from rescaling argument that it is sufficient to prove the theorem for $M = 1$. We introduce new notations:

$$a_m := \|f_m^\tau\|^2, \quad y_m := |\langle f_{m-1}^\tau, \varphi_m^\tau\rangle|, \quad m = 1, 2, \dots, \quad y_0 := 0,$$

and consider the sequence $\{b_n\}$ defined as follows

$$b_0 := 1, \quad b_m := b_{m-1} + y_m, \quad m = 1, 2, \dots.$$

It is clear that $f^\tau_n \in \mathcal{A}_1(\mathcal{D}, b_n)$. By Lemma 3.5 from [9] we get

$$\sup_{g\in\mathcal{D}} |\langle f^\tau_{m-1}, g\rangle| \ge \|f^\tau_{m-1}\|^2/b_{m-1}. \tag{5.2}$$

From here and from the equality (see (2.2))

$$\|f^\tau_m\|^2 = \|f^\tau_{m-1}\|^2 - \langle f^\tau_{m-1}, \varphi^\tau_m\rangle^2$$

we obtain the following relations

$$a_m = a_{m-1} - y_m^2, \tag{5.3}$$

$$b_m = b_{m-1} + y_m, \tag{5.4}$$

$$y_m \ge t_m a_{m-1}/b_{m-1}. \tag{5.5}$$

From (5.3) and (5.5) we get

$$a_m \le a_{m-1}(1 - t_m^2 a_{m-1} b_{m-1}^{-2}).$$

Using that $b_{m-1} \le b_m$ we derive from here

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

By Lemma 3.1 with $A = 1$ we obtain

$$a_m b_m^{-2} \le (1 + \sum_{k=1}^m t_k^2)^{-1}. \tag{5.6}$$

The relations (5.3) and (5.5) imply

$$a_m \le a_{m-1} - y_m t_m a_{m-1}/b_{m-1} = a_{m-1}(1 - t_m y_m/b_{m-1}). \tag{5.7}$$

Rewriting (5.4) in the form

$$b_m = b_{m-1}(1 + y_m/b_{m-1}), \tag{5.8}$$

and using the inequality

$$(1+x)^\alpha \le 1 + \alpha x, \quad 0 \le \alpha \le 1, \quad x \ge 0,$$

we get from (5.7) and (5.8) that

$$a_m b_m^{t_m} \le a_{m-1} b_{m-1}^{t_m}.$$

Next, $b_{m-1} \geq 1$ and $t_m \leq t_{m-1}$. Therefore

$$b_{m-1}^{t_m} \leq b_{m-1}^{t_{m-1}}$$

and

$$a_m b_m^{t_m} \leq a_{m-1} b_{m-1}^{t_{m-1}} \leq \cdots \leq a_0 \leq 1. \tag{5.9}$$

Combining (5.6) and (5.9) we obtain

$$a_m^{2+t_m} \leq (1 + \sum_{k=1}^{m} t_k^2)^{-t_m},$$

what completes the proof. □

Example. In a particular case $t_1 = 1$, $t_k = (\log_2 k)^{-s}$, $k = 2, 3, \ldots$, $0 < s < 1$, we have

$$1 + \sum_{k=1}^{m} t_k^2 \geq m(\log_2 m)^{-2s}$$

and

$$\|f_m\| \leq 2^{-C(\log_2 m)^{1-s}}$$

with an absolute constant $C > 0$.

Acknoledgments. The author would like to thank professors R. DeVore, S. Konyagin, and K. Oskolkov for interesting and helful discussions. Also, the author is grateful to Dr. N. Nagase for carefully reading the manuscript and his helpful remarks.

References

[1] Andrew R. Barron, *Universal approximation byounds for superposition of n sigmoidal functions*, IEEE Transactions on Information Theory **39** (1993), 930– 945.

[2] D.L. Donoho, *Unconditional bases are optimal bases for data compression and for statistical estimation*, Appl. Comput. Harmon. Anal. **1** (1993), 100–115.

[3] D.L. Donoho, *CART and Best-Ortho-Basis: A Connection*, Preprint (1995), 1–45.

[4] C. Darken, M. Donahue, L. Gurvits, E. Sontag, *Rate of convex approximation in non-Hilbert spaces*, Constructive Approximation **13** (1997), 187–220.

[5] R.A. DeVore, *Nonlinear Approximation*, Acta Numerica (1998), 51–150.

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

[7] G. Davis, S. Mallat, and M. Avellaneda, *Adaptive greedy approximations*, Constructive Approximation **13** (1997), 57–98.

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

[9] R.A. DeVore and V.N. Temlyakov, *Some remarks on Greedy Algorithms*, Advances in computational Mathematics **5** (1996), 173–187.

[10] R.A. DeVore and V.N. Temlyakov, *Nonlinear approximation in finite-dimensional spaces*, J. Complexity **13** (1997), 489–508.

[11] V.V. Dubinin, *Greedy Algorithms and Applications*, Ph.D. Thesis, University of South Carolina, 1997.

[12] J.H. Friedman and W. Stuetzle, *Projection pursuit regression*, J. Amer. Statist. Assoc. **76** (1981), 817–823.

[13] P.J. Huber, *Projection Pursuit*, The Annals of Statistics **13** (1985), 435–475.

[14] L. Jones, *On a conjecture of Huber concerning the convergence of projection pursuit regression*, The Annals of Statistics **15** (1987), 880–882.

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

[16] B. S. Kashin and V. N. Temlyakov, *On best m-terms approximations and the entropy of sets in the space L^1*, Math. Notes **56** (1994), 57–86.

[17] B.S. Kashin and V.N. Temlyakov, *On estimating approximative characteristics of classes of functions with bounded mixed derivative*, Math. Notes **58** (1995), 922–925.

[18] L. Rejtö and G.G. Walter, *Remarks on projection pursuit regression and density estimation*, Stochastic Analysis and Application **10** (1992), 213–222.

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

[20] V.N. Temlyakov, *Greedy algorithm and m-term trigonometric approximation*, Constructive Approx. **14** (1998), 569–587.

[21] V.N. Temlyakov, *The best m-term approximation and Greedy Algorithms*, Advances in Comp. Math. **8** (1998), 249–265.

[22] V.N. Temlyakov, *Nonlinear m-term approximation with regard to the multivariate Haar system*, East J. Approx. **4** (1998), 87–106.

[23] V.N. Temlyakov, *Greedy algorithms with regard to the multivariate systems with a special structure*, Preprint (1998), 1–26.

[24] V.N. Temlyakov, *Greedy algorithms and m-term approximation with regard to redundant dictionaries*, J. Approx. Theory (to appear).

[25] A. Zygmund, *Trigonometric series*, University Press, Cambridge, 1959.