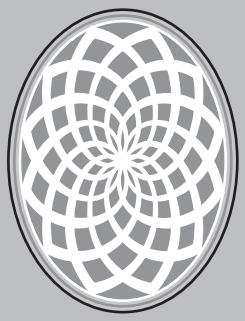

Industrial Mathematics Institute

2002:10

Vector greedy algorithms

A. Lutoborski and V.N. Temlyakov

IMI

Preprint Series

Department of Mathematics
University of South Carolina

Vector Greedy Algorithms

Adam Lutoborski
Department of Mathematics
Syracuse University
Syracuse, NY 13244-1150

Vladimir N. Temlyakov
Department of Mathematics
University of South Carolina
Columbia, SC 29208

Abstract

Our objective is to study nonlinear approximation with regard to redundant systems. 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 problems. Greedy type approximations proved to be convenient and efficient ways of constructing m-term approximants. We introduce and study vector greedy algorithms that are designed with aim of constructing mth greedy approximants simultaneously for a given finite number of elements. We prove convergence theorems and obtain some estimates for the rate of convergence of vector greedy algorithms when elements come from certain classes.

1 Introduction

Our objective is to study nonlinear approximation of vector functions. The most basic concept of nonlinear approximation is to use in it the elements from a set depending on the function being approximated rather than from a fixed vector space.

Assume that X is a Banach space with a basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$ so that each function $f \in X$ has a unique representation

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

It has been established that the algorithm which forms a sum of m terms with the largest $\|c_k(f)\psi_k\|_X$ out of expansion (1), and is hence called greedy, realizes in many cases almost the best m-term approximation for function classes ([6]) and even for individual functions ([21]). 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]).

Another more complicated form of nonlinear approximation can be called highly nonlinear approximation. In it, the basis is replaced by a larger system

of functions $\mathcal{D}$ that is usually redundant and called a dictionary. 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 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 approximation involving dictionaries was considered by E. Schmidt in 1907, [19] who considered the approximation of functions $f(x,y)$ of two variables in $L_2([0,1]^2)$ by bilinear forms

$$B_m(x,y) = \sum_{j=1}^{m} c_j u_j(x) v_j(y).$$

This approximation problem can be seen as an m-term approximation with regard to the dictionary

$$\Pi = \{g \,:\, g(x,y) = u(x)v(y); \quad u, v \in L_2([0,1]), \|u\|_{L_2} = \|v\|_{L_2} = 1\}.$$

The above problem is closely connected with the properties of the integral operator

$$J_f(g)(x) := \int_0^1 f(x,y) g(y)\,\mathrm{d}y,$$

with kernel $f(x,y)$. Schmidt gave an expansion (known under his name)

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

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

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

It was understood later that the above best bilinear approximation can be realized by the following greedy algorithm. Assume that 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. At the m-th step of the algorithm 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 in the next section).

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

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

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. At the m-th step choose a unit vector ω_m and a function r_m to minimize the error

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

This is the second example of Pure Greedy Algorithm. The Pure Greedy Algorithm and some other versions of greedy type algorithms have been recently intensively studied (see [1], [4], [7], [9], [10], [11], [13], [14], [15], [24]).

2 Greedy Algorithms. Basic Notions

In this paper we will 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 and Vector Weak Greedy Algorithms. We will study only theoretical aspects of the efficiency of m-term approximation and possible ways to realize this efficiency. The 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 approximation optimizing process". Keeping this 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 (see for instance [5], [9]).

In order 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 $\|x\| := \langle x, x \rangle^{1/2}$. 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. We say a set $\mathcal{D}$ of elements from H is a dictionary if each $g \in \mathcal{D}$ has norm one ($\|g\| = 1$) and $\overline{\text{span}}\ \mathcal{D} = H$.

The objective of greedy algorithms is to construct a sequence $\{g_k\}_{k \geq 1}$, $g_k \in \mathcal{D}$ and a sequence of approximants $G_k \in \mathcal{D}_k = \text{span}\{g_1, \dots, g_k\}$ such that

$$\lim_{k \to \infty} \|f - G_k\| = 0.$$

The most important step of a greedy algorithm is to choose a new dictionary element g_{n+1} to add to the existing set of $g_1, \dots, g_n$ obtained after n steps. The specific optimization criteria for constructing G_k will define specific greedy algorithms.

For a given dictionary $\mathcal{D}$ we can introduce a norm associated with $\mathcal{D}$ by the formula

$$\|f\|_{\mathcal{D}} := \sup_{g \in \mathcal{D}} |\langle f, g \rangle|.$$

It is clear that in the case of general dictionary $\mathcal{D}$ we can not guarantee that for each $f \in H$ there exists $g^* \in \mathcal{D}$ such that

$$\|f\|_{\mathcal{D}} = |\langle f, g^* \rangle|.$$

In order to overcome this difficulty we use two ways. In the first way, when we define the Pure Greedy Algorithm and the Orthogonal Greedy Algorithm (see Algorithms 1 and 2 below), we make the following additional assumption. We assume we can define a selection operator $S = S_{\mathcal{D}}$, $S : H \to \mathcal{D}$ so that it satisfies

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

We define

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

In the second way we do not impose an extra assumption on $\mathcal{D}$ and instead we weaken the condition

$$|\langle f, S(f) \rangle| = \|f\|_{\mathcal{D}},$$

for selection of an element from the dictionary $\mathcal{D}$ to the condition

$$|\langle f, \varphi \rangle| \geq t \|f\|_{\mathcal{D}}, \quad \varphi \in \mathcal{D},$$

with $t \in (0, 1)$. This way is realized in all weak type greedy algorithms (see Algorithms 3–6 below). The following greedy algorithms have been studied in [9].

Algorithm 1 (Pure Greedy Algorithm, PGA) *We define $R_0(f) = f$ and $G_0(f) = 0$. Then, for each $m \geq 1$, we inductively define:*

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

2.
$$R_m(f) := R_m(f, \mathcal{D}) := f - G_m(f) = R(R_{m-1}(f)).$$

Algorithm 2 (Orthogonal Greedy Algorithm, OGA) *We define* $R_0^o(f) = f$ *and* $G_0^o(f) = 0$. *Then for each* $m \geq 1$, *we inductively define*

1. $$H_m(f) := \mathrm{span}\{G(R_0^o(f)), \dots, G(R_{m-1}^o(f))\},$$

2. $$G_m^o(f) := G_m^o(f, \mathcal{D}) := P_{H_m}(f),$$

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

We remark that for each f we have

$$\|f - G_m^o(f, \mathcal{D})\| \leq \|R_{m-1}^o(f) - G_1(R_{m-1}^o(f), \mathcal{D})\|. \tag{2}$$

In [25] we studied some modifications of the Pure Greedy Algorithm and the Orthogonal Greedy Algorithm which we called respectively Weak Greedy Algorithm (WGA) and Weak Orthogonal Greedy Algorithm (WOGA). We give now the corresponding definitions. Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \leq t_k \leq 1$, be given.

Algorithm 3 (Weak Greedy Algorithm, WGA) *We define* $f_0^\tau := f$. *Then for each* $m \geq 1$, *we inductively define:*

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

$$|\langle f_{m-1}^\tau, \varphi_m^\tau \rangle| \geq t_m \|f_{m-1}^\tau\|_{\mathcal{D}},$$

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]. In our paper we modify WGA and WOGA in a way that allows us to build simultaneous approximation for a given vector of elements $f^1, \dots, f^N$ each of its components in H. We call this new modification "vector" type greedy algorithms VGA.

Let us consider a particular example that gives a motivation for studying "vector" type greedy algorithms. Let N functions $f^i(x, y) \in L_2([0, 1]^2)$, $i = 1, \dots, N$ be given. We want to approximate different kernels $f(x, y)$ of the form

$$f(x, y) = \sum_{j=1}^{N} a_j f^j(x, y),$$

by bilinear forms

$$B_m(x,y) = \sum_{j=1}^{m} c_j u_j(x) v_j(y).$$

As we mentioned in the Introduction the truncated Schmidt expansion

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

provides the best (in the $L_2([0,1]^2)$ with regard to Π) approximation of f. However, this way has a disadvantage. Changing the coefficients $a_1,\dots,a_N$ we change everything in the Schmidt expansion ($s_j(J_f)$, $\{\phi_j\}$, $\{\psi_j\}$). We want to have a more efficient way of constructing good approximants of $f(x,y)$ for different sets of coefficients $a_1,\dots,a_N$. We suggest to build a kind of "simultaneous" Schmidt expansion for the functions

$$f^i(x,y) = \sum_{j=1}^{\infty} b_j^i \phi_j(x)\psi_j(y), \qquad i = 1,\dots,N,$$

with systems $\{\phi_j\}$ and $\{\psi_j\}$ independent of i. Then the corresponding expansion for $f(x,y)$ with any given coefficients $a_1,\dots,a_N$ can be easily obtained as

$$f(x,y) = \sum_{j=1}^{\infty}(\sum_{i=1}^{N} a_i b_j^i)\phi_j(x)\psi_j(y).$$

Algorithm 4 (Vector Weak Greedy Algorithm, VWGA) *Let a vector of elements $f^i \in H$, $i = 1,\dots,N$ be given. We define $f_0^{i,v,\tau} := f^i$. Then for each $m \ge 1$, we inductively define:*

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

$$\max_i |\langle f_{m-1}^{i,v,\tau}, \varphi_m^{v,\tau}\rangle| \ge t_m \max_i \|f_{m-1}^{i,v,\tau}\|_{\mathcal{D}},$$

2.

$$f_m^{i,v,\tau} := f_{m-1}^{i,v,\tau} - \langle f_{m-1}^{i,v,\tau}, \varphi_m^{v,\tau}\rangle \varphi_m^{v,\tau}, \quad i = 1,\dots,N,$$

3.

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

We will prove (see Theorem 3 below) that under certain conditions on τ the VWGA converges. This implies that the VWGA provides the convergent expansions

$$f^i = \sum_{j=1}^{\infty} b_j^i g_j, \qquad g_j \in \mathcal{D},$$

with the property

$$\|f^i\|^2 = \sum_{j=1}^{\infty} |b_j^i|^2, \qquad i = 1, \dots, N.$$

Algorithm 5 (Weak Orthogonal Greedy Algorithm, WOGA) *We define* $f_0^{\tau,o} := f$ *and* $f_1^{\tau,o} := f_1^\tau$; $\varphi_1^{\tau,o} := \varphi_1^\tau$ *where* f_1^τ, φ_1^τ *are from the above definition of WGA. Then for each* $m \geq 2$ *we inductively define:*
1. $\varphi_m^{\tau,o} \in \mathcal{D}$ *is any element satisfying*

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

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

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

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

Algorithm 6 (Vector Weak Orthogonal Greedy Algorithm, VWOGA) *Let a set vector* f^i, $i = 1, \dots, N$ *be given. We define* $f_0^{i,v,\tau,o} := f^i$, $i = 1, \dots, N$. *Then for each* $m \geq 1$ *we inductively define:*
1. i_m *is such that*

$$\|f_{m-1}^{i_m,v,\tau,o}\| \geq \|f_{m-1}^{i,v,\tau,o}\|, \qquad i = 1, \dots, N,$$

2. $\varphi_m^{v,o,\tau} \in \mathcal{D}$ *is any element satisfying*

$$|\langle f_{m-1}^{i_m,v,\tau,o}, \varphi_m^{v,\tau,o} \rangle| \geq t_m \|f_{m-1}^{i_m,v,\tau,o}\|_{\mathcal{D}},$$

ıt3. $$G_m^{v,o,\tau}(f^i, \mathcal{D}) := P_{H_m^{v,\tau}}(f^i), \quad \text{where} \quad H_m^{v,\tau} := \operatorname{span}\{\varphi_1^{v,\tau,o}, \dots, \varphi_m^{v,\tau,o}\},$$

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

We turn first to formulate some theorems on convergence of WGA and WOGA. We begin with 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]. The convergence of WGA and WOGA was studied in [25] including the following theorems:

Theorem 1 *Assume*

$$\sum_{k=1}^{\infty} \frac{t_k}{k} = \infty. \tag{3}$$

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*

$$\sum_{k=1}^{\infty} t_k^2 = \infty. \tag{4}$$

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

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

We will prove here the following generalizations of these theorems to the case of vector approximation.

Theorem 3 *Assume $\sum_{k=1}^{\infty} \frac{t_k}{k} = \infty$ is satisfied. Then for any dictionary $\mathcal{D}$ and any set of functions $f^i \in H$, $i = 1,\dots,N$ we have*

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

Theorem 4 *Assume $\sum_{k=1}^{\infty} t_k^2 = \infty$ is satisfied. Then for any dictionary $\mathcal{D}$ and any set of functions $f^i \in H$, $i = 1,\dots,N$ we have*

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

The following criterion on τ for convergence of WGA has been established in [26].

Let us introduce some notation. We define by $\mathcal{V}$ the class of sequences of real numbers $x = \{x_k\}_{k=1}^{\infty}$, $x_k \ge 0$, $k = 1,2,\dots$, with the following property: there exists a sequence $0 = q_0 < q_1 < \dots$ which may depend on x such that

$$\sum_{s=1}^{\infty} \frac{2^s}{\Delta q_s} < \infty,$$

and

$$\sum_{s=1}^{\infty} 2^{-s} \sum_{k=1}^{q_s} x_k^2 < \infty,$$

where $\Delta q_s := q_s - q_{s-1}$.

Theorem 5 *The condition $\tau \notin \mathcal{V}$ is necessary and sufficient for convergence of Weak Greedy Algorithm with weakness sequence τ for each f and all Hilbert spaces H and dictionaries $\mathcal{D}$.*

It is clear that the condition $\tau \notin \mathcal{V}$ is also necessary for convergence of VWGA with the weakness sequence τ. We will prove that this condition ($\tau \notin \mathcal{V}$) is also sufficient for convergence of VWGA.

Theorem 6 *The condition $\tau \notin \mathcal{V}$ is necessary and sufficient for convergence of Vector Weak Greedy Algorithm with weakness sequence τ for all vectors $f^1, \dots, f^N$ and all Hilbert spaces H and dictionaries $\mathcal{D}$.*

3 Convergence of the VWGA

The following two lemmas imply Theorem 3.

Lemma 1 *Assume that (4) is satisfied. Then if all component sequences $\{f_m^{i,v,\tau}\}_{m=1}^{\infty}$, $i = 1, \dots, N$ converge then they converge to zero.*

Proof of Lemma 1 is the same as proof of Lemma 2.1 from [25].

Lemma 2 *Assume (3) is satisfied. Then each component sequence $\{f_m^{i,v,\tau}\}_{m=1}^{\infty}$, $i = 1, \dots, N$ converges.*

Proof of Lemma 2. For simplicity in notations we drop the superscripts v and τ. It is easy to derive from the definition of VWGA the following two relations for $i = 1, \dots, N$

$$f_m^i = f^i - \sum_{j=1}^{m} \langle f_{j-1}^i, \varphi_j \rangle \varphi_j, \tag{5}$$

$$\|f_m^i\|^2 = \|f^i\|^2 - \sum_{j=1}^{m} |\langle f_{j-1}^i, \varphi_j \rangle|^2. \tag{6}$$

Denote

$$a_j^i := |\langle f_{j-1}^i, \varphi_j \rangle|, \qquad a_j := \sum_{i=1}^{N} a_j^i.$$

We get from (6) that

$$\sum_{j=1}^{\infty} (a_j^i)^2 \le \|f^i\|^2. \tag{7}$$

We take any two indices $n < m$ and consider for $i = 1, \dots, N$

$$\|f_n^i - f_m^i\|^2 = \|f_n^i\|^2 - \|f_m^i\|^2 - 2\langle f_n^i - f_m^i, f_m^i \rangle.$$

Denote and estimate the quantities

$$\theta_{n,m}^i := |\langle f_n^i - f_m^i, f_m^i \rangle|.$$

By Lemma 2.4 from [25] (see below) for convergence of $\{f_m^i\}$ it is sufficient to prove that

$$\liminf_{m\to\infty} \max_{n<m} \theta_{n,m}^i = 0.$$

Using (5) and the definition of the VWGA we get for all $n < m$ that

$$\max_{n<m} \theta_{n,m}^i \le \max_{n<m} \sum_{j=n+1}^{m} |\langle f_{j-1}^i, \varphi_j\rangle||\langle f_m^i, \varphi_j\rangle| \le \sum_{j=1}^{m+1} a_j^i |\langle f_m^i, \varphi_j\rangle|. \tag{8}$$

¿From the definition of φ_{m+1} we get

$$\max_i |\langle f_m^i, \varphi_{m+1}\rangle| \ge t_{m+1} \max_i \sup_{g\in\mathcal{D}} |\langle f_m^i, g\rangle|.$$

This implies for all j

$$|\langle f_m^i, \varphi_j\rangle| \le \frac{1}{t_{m+1}} \max_i |\langle f_m^i, \varphi_{m+1}\rangle| \le \frac{1}{t_{m+1}} \sum_{i=1}^{N} a_{m+1}^i. \tag{9}$$

We get from (8) and (9) that for $i = 1, \dots, N$

$$\max_{n<m} \theta_{n,m}^i \le \frac{a_{m+1}}{t_{m+1}} \sum_{j=1}^{m} a_j.$$

It remains to use (3) and (7) in the following lemma from [25].

Lemma 3 *(Temlyakov and 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

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

Lemma 4 *([25]) Let a sequence $\{x_n\}_{n=1}^{\infty}$ be given in a Banach space X. Assume that for any m, n we have*

$$\|x_n - x_m\|^2 = y_m - y_n + \theta_{m,n},$$

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

$$\liminf_{n\to\infty} \max_{m<n} |\theta_{m,n}| = 0.$$

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

We now proceed to the proof of Theorem 6. As it was mentioned above we need to prove only the sufficiency part. Taking into account Lemma 1 we can claim that in the proof of Lemma 2 we actually proved the following statement.

Lemma 5 *Let τ be such that for any $\{a_j\}_{j=1}^{\infty} \in l_2$, $a_j \geq 0$, $j = 1, 2, \ldots$ we have*

$$\liminf_{n\to\infty} a_n \sum_{j=1}^{n} a_j/t_n = 0.$$

Then for any H, $\mathcal{D}$, and $f^i \in H$, $i = 1, \ldots, N$ we have

$$\lim_{m\to\infty} \|f_m^{i,v,\tau}\| = 0, \qquad i = 1, \ldots, N.$$

We now use the following theorem from [26].

Theorem 7 *The following two conditions are equivalent*

$$\tau \notin \mathcal{V},$$

$$\forall \{a_j\}_{j=1}^{\infty} \in l_2, \quad a_j \geq 0, \quad \liminf_{n\to\infty} a_n \sum_{j=1}^{n} a_j/t_n = 0.$$

Combining this theorem and Lemma 5 we complete the proof of Theorem 6.

We proceed now to the rate of convergence of VWGA. For a general dictionary $\mathcal{D}$, we define the class of functions

$$\mathcal{A}_1^o(\mathcal{D}, M) := \{f \in H : f = \sum_{k=1}^{n} c_k w_k, \quad w_k \in \mathcal{D}, \quad \sum_{k=1}^{n} |c_k| \leq M\},$$

where n is any integer. Denote $\mathcal{A}_1(\mathcal{D}, M)$ the closure of $\mathcal{A}_1^o(\mathcal{D}, M)$. For $M = 1$ we denote $\mathcal{A}_1(\mathcal{D}) := \mathcal{A}_1(\mathcal{D}, 1)$. We prove the following theorem.

Theorem 8 *Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k = t$, $k = 1, \ldots$, $0 < t < 1$. Then for any vector $f^1, \ldots, f^N$, $f^i \in \mathcal{A}_1(\mathcal{D}, M)$, $i = 1, \ldots, N$ we have*

$$\sum_{i=1}^{N} \|f_m^{i,v,\tau}\|^2 \leq M^2 \left(N + mt^2\right)^{-t/(2N+t)} N^{\frac{2N+3t}{2N+t}}. \tag{10}$$

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

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

and consider the sequences $\{b_n^i\}$ defined as follows

$$b_0^i := 1, \qquad b_m^i := b_{m-1}^i + y_m^i, \qquad m = 1, 2, \ldots.$$

Consider also the sequences

$$a_m := \sum_{i=1}^{N} a_m^i; \quad y_m := \sum_{i=1}^{N} y_m^i; \quad b_m := \sum_{i=1}^{N} b_m^i.$$

Note that by the Cauchy inequality we have

$$\sum_{i=1}^{N}(y_m^i)^2 \geq y_m^2/N. \tag{11}$$

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

$$\sup_{g\in\mathcal{D}} |\langle f_{m-1}^i, g\rangle| \geq \|f_{m-1}^i\|^2/b_{m-1}^i. \tag{12}$$

¿From (12) and from the equality (see (6))

$$\|f_m^i\|^2 = \|f_{m-1}^i\|^2 - |\langle f_{m-1}^i, \varphi_m\rangle|^2 \tag{13}$$

we obtain the following relations

$$a_m^i = a_{m-1}^i - (y_m^i)^2, \tag{14}$$

$$b_m^i = b_{m-1}^i + y_m^i, \tag{15}$$

$$\max_i y_m^i \geq t \max_i a_{m-1}^i/b_{m-1}^i. \tag{16}$$

This implies that

$$y_m \geq \max_i y_m^i \geq t\frac{a_{m-1}}{b_{m-1}}. \tag{17}$$

We have used the following simple property of fractions that the mediant of two fractions always lies between them in value. If a, b, c, d be nonnegative and $c/d \leq a/b$ then

$$\frac{c}{d} \leq \frac{a+c}{b+d} \leq \frac{a}{b}.$$

Thus from (11), (14), (15), and (17) we get

$$a_m \leq a_{m-1} - y_m^2/N, \tag{18}$$

$$b_m = b_{m-1} + y_m, \tag{19}$$

$$y_m \geq t\,\frac{a_{m-1}}{b_{m-1}}. \tag{20}$$

We get from (18) and (20) that

$$a_m \leq a_{m-1}\left(1 - \frac{t^2}{N}\frac{a_{m-1}}{b_{m-1}^2}\right),$$

and due to $b_m^i \geq b_{m-1}^i$

$$\frac{a_m}{b_m^2} \le \frac{a_{m-1}}{b_{m-1}^2}\left(1 - \frac{t^2}{N}\frac{a_{m-1}}{b_{m-1}^2}\right).$$

By Lemma 3.1 from [25] with $t_m^2 := t^2/N$ and $A = 1$ we obtain

$$\frac{a_m}{b_m^2} \le \left(1 + \frac{mt^2}{N}\right)^{-1}. \tag{21}$$

On the other hand by (18) and (20) we have

$$a_m \le a_{m-1}\left(1 - \frac{t}{N}\frac{y_m}{b_{m-1}}\right),$$

and by (19)

$$b_m = b_{m-1}\left(1 + \frac{y_m}{b_{m-1}}\right).$$

Similarly to the case of WGA (see [25], Section 5) we get from here that

$$a_m b_m^{t/N} \le N^{1+t/N}. \tag{22}$$

Combining (21) and (22) we get

$$a_m^{2+t/N} \le \left(1 + \frac{mt^2}{N}\right)^{-t/N} N^{2(1+t/N)}.$$

This completes the proof of Theorem 8.

4 Rate of Approximation of VWOGA

The following theorem has been proved in [25] for $\tau = \{t_k\}$, $0 \le t_k \le 1$, $k \ge 1$.

Theorem 9 *Let $\mathcal{D}$ be an arbitrary dictionary in H. Then for each $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have*

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

We will use this theorem in this section to establish the convergence rate of VWOGA.

Theorem 10 *Let $\mathcal{D}$ be an arbitrary dictionary in H and $\tau = \{t\}$, $0 < t \le 1$. Then for each $f^i \in \mathcal{A}_1(\mathcal{D}, M)$ we have*

$$\|f^i - G_m^{v,o,\tau}(f^i, \mathcal{D})\| \le M \min(1, \left(\frac{N}{mt^2}\right)^{1/2}), \qquad i = 1, \dots, N.$$

Proof. We will carry out the proof for $M = 1$. The inequality $\|f_m^i\| \le 1$ follows from the assumption $f^i \in A_1(\mathcal{D})$ and from the obvious remark that the sequences $\{\|f_m^i\|\}_{m\ge 1}$ are decreasing. Let us prove the estimate

$$\|f_m^i\| \le \left(\frac{N}{mt^2}\right)^{1/2}.$$

Take $m \ge N$ and define i_0 to be the one with (i_j, $j = 1, 2, \dots$ are defined in the definition of VWOGA)

$$\#\{j \ : \ i_j = i_0,\ 1 \le j \le m\} \ge m/N. \tag{23}$$

Let

$$n := \max\{j,\ j \in [1, m],\ :\ i_j = i_0\}.$$

Then we have

$$\|f_m^i\| \le \|f_{n-1}^i\| \le \|f_{n-1}^{i_0}\|. \tag{24}$$

The VWOGA can be seen as a realization of WOGA for each f^i with appropriately chosen τ^i, $i = 1, \dots, N$. For instance for f^{i_0} we get $f_1^{i_0}, \dots, f_m^{i_0}$ as a realization of WOGA with $\tau^{i_0} = \{t_k^{i_0}\}$, $t_k^{i_0} = t$ if $i_k = i_0$ and $t_k^{i_0} = 0$ otherwise. Then by Theorem 9 we get

$$\|f_{n-1}^{i_0}\| \le (1 + (m/N - 1)t^2)^{-1/2} \le \left(\frac{N}{mt^2}\right)^{1/2}.$$

Using (24) we complete the proof.

Comparing Theorem 9 with $\tau = \{t\}$ and Theorem 10 we see that in approximation of a vector with N components by the VWOGA the number $[m/N]$ plays the same role as the number m in the case of WOGA. This means that in essence the VWOGA has the same guaranteed upper estimate for the error as the following N-fold WOGA. For a given $t \in (0, 1]$ we apply $n := [m/N]$ steps of the WOGA to each f^i, $i = 1, \dots, N$. For each f^i we get a subspace $H_n^t(f^i)$, $i = 1, \dots, N$ (see the definition of Algorithm 5). Denoting

$$H_m^t := \bigoplus_{i=1}^N H_n^t(f^i),$$

we obtain by Theorem 9 the estimate

$$\begin{aligned}\|f^i - P_{H_m^t}(f^i)\| \le& \|f^i - P_{H_n^t(f^i)}(f^i)\| \\ \le& M(1 + nt^2)^{-1/2} \\ \le& M \min(1, ((m/N)t^2)^{-1/2}), \qquad i = 1, \dots, N.\end{aligned}$$

However the VWOGA is more adaptive than the above N-fold WOGA. For instance if we have one "bad" element f^1 and all "good" elements $f^2, \dots, f^N$ then the VWOGA will work all m steps with f^1 while the N-fold WOGA will

use only $[m/N]$ steps for working with f^1. This adaptivity makes the VWOGA more suitable for applications than the N-fold WOGA.

We also note that the replacement of m by $[m/N]$ when we switch from the WOGA to the VWOGA (with a vector of N components) is natural. In order to understand this let us consider the following general example. Let H and $\mathcal{D}$ be given. For a fixed N consider the direct sums of N copies of H and $\mathcal{D}$

$$H^N := H + \cdots + H, \qquad \mathcal{D}^N := \mathcal{D} + \cdots + \mathcal{D}.$$

Then it is clear that the behavior of the WOGA with $\tau = \{t\}$ applied to a $f \in H$ at the $[m/N]$th step is equivalent to the behavior of the VWOGA in H^N with $\mathcal{D}^N$ applied to the vector $f, \dots, f$ (of N components) at the mth step.

References

[1] A.R. Barron, *Universal approximation bounds 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*, The Annals of Statistics **25** (1997), 1870–1911.

[4] M. Donahue, L. Gurvits, C. Darken, and 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 Applications **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 Approximation **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 **98** (1999), 117–145.

[25] V.N. Temlyakov, *Weak greedy algorithms*, Advances in Comp. Math. **12** (2000), 213–227.

[26] V.N. Temlyakov, *A criterion for convergence of Weak Greedy Algorithm*, IMI-Preprint series **21** (2000), 1–10.