# Greedy approximation in convex optimization

V.N. Temlyakov *

June 2, 2012

**Abstract**

We study sparse approximate solutions to convex optimization problems. It is known that in many engineering applications researchers are interested in an approximate solution of an optimization problem as a linear combination of elements from a given system of elements. There is an increasing interest in building such sparse approximate solutions using different greedy-type algorithms. The problem of approximation of a given element of a Banach space by linear combinations of elements from a given system (dictionary) is well studied in nonlinear approximation theory. At a first glance the settings of approximation and optimization problems are very different. In the approximation problem an element is given and our task is to find a sparse approximation of it. In optimization theory an energy function is given and we should find an approximate sparse solution to the minimization problem. It turns out that the same technique can be used for solving both problems. We show how the technique developed in nonlinear approximation theory, in particular, the greedy approximation technique can be adjusted for finding a sparse solution of an optimization problem.

## 1 Introduction

We study sparse approximate solutions to convex optimization problems. We apply the technique developed in nonlinear approximation known under the

*University of South Carolina. Research was supported by NSF grant DMS-0906260

name of *greedy approximation*. A typical problem of convex optimization is to find an approximate solution to the problem

$$\inf_x E(x) \tag{1.1}$$

under assumption that $E$ is a convex function. Usually, in convex optimization function $E$ is defined on a finite dimensional space $\mathbb{R}^n$ (see [3], [10]). Recent needs of numerical analysis call for consideration of the above optimization problem on an infinite dimensional space, for instance, a space of continuous functions. One more important argument that motivates us to study this problem in the infinite dimensional space setting is the following. In many contemporary numerical applications the dimension $n$ of the ambient space $\mathbb{R}^n$ is large and we would like to obtain bounds on the convergence rate independent of the dimension $n$. Our results for infinite dimensional spaces provide such bounds on the convergence rate. Thus, we consider a convex function $E$ defined on a Banach space $X$. It is pointed out in [21] that in many engineering applications researchers are interested in an approximate solution of problem (1.1) as a linear combination of elements from a given system $\mathcal{D}$ of elements. There is an increasing interest in building such sparse approximate solutions using different greedy-type algorithms (see, for instance, [21], [12], [5], and [20]). The problem of approximation of a given element $f \in X$ by linear combinations of elements from $\mathcal{D}$ is well studied in nonlinear approximation theory (see, for instance [6], [16], [17]). In order to address the contemporary needs of approximation theory and computational mathematics, a very general model of approximation with regard to a redundant system (dictionary) has been considered in many recent papers. As such a model, we choose a Banach space $X$ with elements as target functions and an arbitrary system $\mathcal{D}$ of elements of this space such that the closure of $\operatorname{span}\mathcal{D}$ coincides with $X$ as an approximating system.

The fundamental question is how to construct good methods (algorithms) of approximation. Recent results have established that greedy type algorithms are suitable methods of nonlinear approximation in both sparse approximation with regard to bases and sparse approximation with regard to redundant systems. It turns out that there is one fundamental principal that allows us to build good algorithms both for arbitrary redundant systems and for very simple well structured bases like the Haar basis. This principal is the use of a greedy step in searching for a new element to be added to a given sparse approximant. By a *greedy step*, we mean one which maximizes

a certain functional determined by information from the previous steps of the algorithm. We obtain different types of greedy algorithms by varying the above mentioned functional and also by using different ways of constructing (choosing coefficients of the linear combination) the $m$-term approximant from the already found $m$ elements of the dictionary.

We point out that at a first glance the settings of approximation and optimization problems are very different. In the approximation problem an element $f \in X$ is given and our task is to find a sparse approximation of it. In optimization theory an energy function $E(x)$ is given and we should find an approximate sparse solution to the minimization problem. It turns out that the same technique can be used for solving both problems.

We show how the technique developed in nonlinear approximation theory, in particular, the greedy approximation technique can be adjusted for finding a sparse with respect to $\mathcal{D}$ solution of problem (1.1).

We begin with a brief description of greedy approximation methods in Banach spaces. The reader can find a detailed discussion of greedy approximation in the book [17]. Let $X$ be a Banach space with norm $\|\cdot\|$. We say that a set of elements (functions) $\mathcal{D}$ from $X$ is a dictionary, respectively, symmetric dictionary, if each $g \in \mathcal{D}$ has norm bounded by one ($\|g\| \le 1$),

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

and the closure of $\operatorname{span} \mathcal{D}$ is $X$. In this paper symmetric dictionaries are considered. We denote the closure (in $X$) of the convex hull of $\mathcal{D}$ by $A_1(\mathcal{D})$. For a nonzero element $f \in X$ we let $F_f$ denote a norming (peak) functional for $f$:

$$\|F_f\| = 1, \qquad F_f(f) = \|f\|.$$

The existence of such a functional is guaranteed by Hahn-Banach theorem. We describe a typical greedy algorithm from a family of *dual greedy algorithms*. Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given weakness sequence of nonnegative numbers $t_k \le 1$, $k = 1, \dots$. We define first the Weak Chebyshev Greedy Algorithm (WCGA) (see [14]) that is a generalization for Banach spaces of the Weak Orthogonal Greedy Algorithm.

**Weak Chebyshev Greedy Algorithm (WCGA).** We define $f_0^c := f_0^{c,\tau} := f$. Then for each $m \ge 1$ we have the following inductive definition.

(1) $\varphi_m^c := \varphi_m^{c,\tau} \in \mathcal{D}$ is any element satisfying

$$F_{f_{m-1}^c}(\varphi_m^c) \ge t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}^c}(g).$$

(2) Define

$$\Phi_m := \Phi_m^\tau := \operatorname{span}\{\varphi_j^c\}_{j=1}^m,$$

and define $G_m^c := G_m^{c,\tau}$ to be the best approximant to $f$ from $\Phi_m$.

(3) Let

$$f_m^c := f_m^{c,\tau} := f - G_m^c.$$

Let us make a remark that justifies the idea of the dual greedy algorithms in terms of real analysis. We consider here approximation in uniformly smooth Banach spaces. For a Banach space $X$ we define the modulus of smoothness

$$\rho(u) := \sup_{\|x\|=\|y\|=1} (\frac{1}{2}(\|x+uy\| + \|x-uy\|) - 1).$$

The uniformly smooth Banach space is the one with the property

$$\lim_{u\to 0} \rho(u)/u = 0.$$

We note that from the definition of modulus of smoothness we get the following inequality.

$$0 \le \|x+uy\| - \|x\| - uF_x(y) \le 2\|x\|\rho(u\|y\|/\|x\|). \tag{1.2}$$

This inequality implies the proposition.

**Proposition 1.1.** *Let $X$ be a uniformly smooth Banach space. Then, for any $x \neq 0$ and $y$ we have*

$$F_x(y) = \left(\frac{d}{du}\|x+uy\|\right)(0) = \lim_{u\to 0}(\|x+uy\| - \|x\|)/u. \tag{1.3}$$

Proposition 1.1 shows that in the WCGA we are looking for an element $\varphi_m \in \mathcal{D}$ that provides a big derivative of the quantity $\|f_{m-1} + ug\|$. Here is one more important greedy algorithm.

**Weak Greedy Algorithm with Free Relaxation (WGAFR).** Let $\tau := \{t_m\}_{m=1}^\infty$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$ and $G_0 := 0$. Then for each $m \ge 1$ we have the following inductive definition.

(1) $\varphi_m \in \mathcal{D}$ is any element satisfying

$$F_{f_{m-1}}(\varphi_m) \ge t_m \sup_{g\in\mathcal{D}} F_{f_{m-1}}(g).$$

(2) Find $w_m$ and $\lambda_m$ such that

$$\|f - ((1-w_m)G_{m-1} + \lambda_m \varphi_m)\| = \inf_{\lambda,w} \|f - ((1-w)G_{m-1} + \lambda \varphi_m)\|$$

and define

$$G_m := (1-w_m)G_{m-1} + \lambda_m \varphi_m.$$

(3) Let

$$f_m := f - G_m.$$

It is known that both algorithms WCGA and WGAFR converge in any uniformly smooth Banach space under mild conditions on the weakness sequence $\{t_k\}$, for instance, $t_k = t$, $k = 1, 2, \dots$, $t > 0$, guarantees such convergence. The following theorem provides rate of convergence (see [17], pp. 347, 353).

**Theorem 1.1.** *Let $X$ be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$. Take a number $\epsilon \ge 0$ and two elements $f$, $f^\epsilon$ from $X$ such that*

$$\|f - f^\epsilon\| \le \epsilon, \quad f^\epsilon / A(\epsilon) \in A_1(\mathcal{D}),$$

*with some number $A(\epsilon) > 0$. Then, for both algorithms WCGA and WGAFR we have ($p := q/(q-1)$)*

$$\|f_m^{c,\tau}\| \le \max\left(2\epsilon, C(q,\gamma)(A(\epsilon)+\epsilon)(1+\sum_{k=1}^m t_k^p)^{-1/p}\right).$$

The above Theorem 1.1 simultaneously takes care of two issues: noisy data and approximation in an interpolation space. In order to apply it for noisy data we interpret $f$ as a noisy version of a signal and $f^\epsilon$ as a noisless version of a signal. Then, assumption $f^\epsilon/A(\epsilon) \in A_1(\mathcal{D})$ describes our smoothness assumption on the noisless signal. Theorem 1.1 can be applied for approximation of $f$ under assumption that $f$ belongs to one of interpolation spaces between $X$ and the space generated by the $A_1(\mathcal{D})$-norm (atomic norm). We now make a remark showing that the $A_1(\mathcal{D})$-norm (in other words, the assumption $f/A \in A_1(\mathcal{D})$) appears naturally in convex optimization problems.

It is pointed out in [7] that there has been considerable interest in solving the convex unconstrained optimization problem

$$\min_x \frac{1}{2}\|y - \Phi x\|_2^2 + \lambda \|x\|_1 \tag{1.4}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^k$, $\Phi$ is an $k \times n$ matrix, $\lambda$ is a nonnegative parameter, $\|v\|_2$ denotes the Euclidian norm of $v$, and $\|v\|_1$ is the $\ell_1$ norm of $v$. Problems of the form (1.4) have become familiar over the past three decades, particularly in statistical and signal processing contexts. Problem (1.4) is closely related to the following convex constrained optimization problem

$$\min_x \frac{1}{2}\|y - \Phi x\|_2^2 \quad \text{subject to} \quad \|x\|_1 \le A. \tag{1.5}$$

The above convex optimization problem can be recast as an approximation problem of $y$ with respect to a dictionary $\mathcal{D} := \{\pm\varphi_i\}_{i=1}^n$ which is associated with a $k \times n$ matrix $\Phi = [\varphi_1 \dots \varphi_n]$ with $\varphi_j \in \mathbb{R}^k$ being the column vectors of $\Phi$. The condition $y \in A_1(\mathcal{D})$ is equivalent to existence of $x \in \mathbb{R}^m$ such that $y = \Phi x$ and

$$\|x\|_1 := |x_1| + \dots + |x_m| \le 1. \tag{1.6}$$

As a direct corollary of Theorem 1.1, we get for any $y \in A_1(\mathcal{D})$ that the WCGA and the WGAFR with $\tau = \{t\}$ guarantee the following upper bound for the error

$$\|y_k\|_2 \le Ck^{-1/2}. \tag{1.7}$$

The bound (1.7) holds for any $\mathcal{D}$ (any $\Phi$).

We note that in the study of greedy-type algorithms in approximation theory (see [17]) emphasis are put on the theory of approximation with respect to arbitrary dictionary $\mathcal{D}$. The reader can find examples of specific dictionaries of interest in [17] and [20]. We present some results on sparse solutions for convex optimization problems in the setting with an arbitrary dictionary $\mathcal{D}$.

We generalize the algorithms WCGA and WGAFR to the case of convex optimization and prove an analog of Theorem 1.1 for the new algorithms. Let us illustrate this on the generalization of the WGAFR.

We assume that the set

$$D := \{x : E(x) \le E(0)\}$$

is bounded. For a bounded set $D$ define the modulus of smoothness of $E$ on $D$ as follows

$$\rho(E, u) := \frac{1}{2} \sup_{x \in D, \|y\|=1} |E(x + uy) + E(x - uy) - 2E(x)|. \tag{1.8}$$

We assume that $E$ is Fréchet differentiable. Then convexity of $E$ implies that for any $x, y$

$$E(y) \geq E(x) + \langle E'(x), y - x\rangle \tag{1.9}$$

or, in other words,

$$E(x) - E(y) \leq \langle E'(x), x - y\rangle = \langle -E'(x), y - x\rangle. \tag{1.10}$$

We will often use the following simple lemma.

**Lemma 1.1.** *Let $E$ be Fréchet differentiable convex function. Then the following inequality holds for $x \in D$*

$$0 \leq E(x + uy) - E(x) - u\langle E'(x), y\rangle \leq 2\rho(E, u\|y\|). \tag{1.11}$$

*Proof.* The left inequality follows directly from (1.9). Next, from the definition of modulus of smoothness it follows that

$$E(x + uy) + E(x - uy) \leq 2(E(x) + \rho(E, u\|y\|)). \tag{1.12}$$

Inequality (1.9) gives

$$E(x - uy) \geq E(x) + \langle E'(x), -uy\rangle = E(x) - u\langle E'(x), y\rangle. \tag{1.13}$$

Combining (1.12) and (1.13), we obtain

$$E(x + uy) \leq E(x) + u\langle E'(x), y\rangle + 2\rho(E, u\|y\|).$$

This proves the second inequality. □

**Weak Greedy Algorithm with Free Relaxation (WGAFR(co)).** Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0, 1]$, be a weakness sequence. We define $G_0 := 0$. Then for each $m \geq 1$ we have the following inductive definition.

(1) $\varphi_m \in \mathcal{D}$ is any element satisfying

$$\langle -E'(G_{m-1}), \varphi_m\rangle \geq t_m \sup_{g\in\mathcal{D}} \langle -E'(G_{m-1}), g\rangle.$$

(2) Find $w_m$ and $\lambda_m$ such that

$$E((1 - w_m)G_{m-1} + \lambda_m\varphi_m) = \inf_{\lambda,w} E((1 - w)G_{m-1} + \lambda\varphi_m)$$

and define

$$G_m := (1 - w_m)G_{m-1} + \lambda_m\varphi_m.$$

In Section 4 we prove the following rate of convergence result.

**Theorem 1.2.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E,u) \le \gamma u^q$, $1 < q \le 2$. Take a number $\epsilon \ge 0$ and an element $f^\epsilon$ from $D$ such that*

$$E(f^\epsilon) \le \inf_{x\in D} E(x) + \epsilon, \quad f^\epsilon / A(\epsilon) \in A_1(\mathcal{D}),$$

*with some number $A(\epsilon) \ge 1$. Then we have for WGAFR(co) ($p := q/(q-1)$)*

$$E(G_m) - \inf_{x\in D} E(x) \le \max\left(2\epsilon, C(q,\gamma)A(\epsilon)\left(C(E,q,\gamma) + \sum_{k=1}^m t_k^p\right)^{1-q}\right).$$

We note that in all algorithms studied in this paper the sequence $\{G_m\}_{m=0}^\infty$ of approximants satisfies the conditions

$$G_0 = 0, \quad E(G_0) \ge E(G_1) \ge E(G_2) \ge \dots.$$

This guarantees that $G_m \in D$ for all $m$.

This paper is the first author's paper on greedy-type methods in convex optimization. It is a slight modification of the paper [18]. For the reader's convenience we now give a brief general description and classification of greedy-type algorithms for convex optimization. The most difficult part of an algorithm is to find an element $\varphi_m \in \mathcal{D}$ to be used in approximation process. We consider greedy methods for finding $\varphi_m \in \mathcal{D}$. We have two types of greedy steps to find $\varphi_m \in \mathcal{D}$.

**I. Gradient greedy step.** At this step we look for an element $\varphi_m \in \mathcal{D}$ such that

$$\langle -E'(G_{m-1}), \varphi_m\rangle \ge t_m \sup_{g\in\mathcal{D}} \langle -E'(G_{m-1}), g\rangle.$$

**II. $E$-greedy step.** At this step we look for an element $\varphi_m \in \mathcal{D}$ which satisfies (we assume existence):

$$\inf_{c\in\mathbb{R}} E(G_{m-1} + c\varphi_m) = \inf_{g\in\mathcal{D}, c\in\mathbb{R}} E(G_{m-1} + cg).$$

The above WGAFR(co) uses the greedy step of type **I**. In this paper we only discuss algorithms based on the greedy step of type **I**. These algorithms fall into a category of the first order methods. The greedy step of type **II** uses only the function values $E(x)$. We discussed some of the algorithms of this type in [19] and plan to study them in our future work.

After we found $\varphi_m \in \mathcal{D}$ we can proceed in different ways. We now list some typical steps that are motivated by the corresponding steps in greedy approximation theory (see [17]). These steps or their variants are used in optimization algorithms like *gradient method*, *reduced gradient method*, *conjugate gradients*, *gradient pursuits* (see, for instance, [8], [10], [9], [11], [1] and [2]).

(A) Best step in the direction $\varphi_m \in \mathcal{D}$. We choose $c_m$ such that

$$E(G_{m-1} + c_m\varphi_m) = \inf_{c\in\mathbb{R}} E(G_{m-1} + c\varphi_m)$$

and define

$$G_m := G_{m-1} + c_m\varphi_m.$$

(B) Reduced best step in the direction $\varphi_m \in \mathcal{D}$. We choose $c_m$ as in (A) and for a given parameter $b > 0$ define

$$G_m^b := G_{m-1}^b + bc_m\varphi_m.$$

Usually, $b \in (0, 1)$. This is why we call it *reduced*.

(C) Chebyshev-type methods. We choose $G_m \in \operatorname{span}(\varphi_1, \dots, \varphi_m)$ which satisfies

$$E(G_m) = \inf_{c_j, j=1,\dots,m} E(c_1\varphi_1 + \cdots + c_m\varphi_m).$$

(D) Fixed relaxation. For a given sequence $\{r_k\}_{k=1}^\infty$ of relaxation parameters $r_k \in [0, 1)$ we choose $G_m := (1 - r_m)G_{m-1} + c_m\varphi_m$ with $c_m$ from

$$E((1 - r_m)G_{m-1} + c_m\varphi_m) = \inf_{c\in\mathbb{R}} E((1 - r_m)G_{m-1} + c\varphi_m).$$

(F) Free relaxation. We choose $G_m \in \operatorname{span}(G_{m-1}, \varphi_m)$ which satisfies

$$E(G_m) = \inf_{c_1,c_2} E(c_1G_{m-1} + c_2\varphi_m).$$

(G) Prescribed coefficients. For a given sequence $\{c_k\}_{k=1}^\infty$ of positive coefficients in the case of greedy step **I** we define

$$G_m := G_{m-1} + c_m\varphi_m. \tag{1.14}$$

In the case of greedy step **II** we define $G_m$ by formula (1.14) with the greedy step **II** modified as follows: $\varphi_m \in \mathcal{D}$ is an element satisfying

$$E(G_{m-1} + c_m\varphi_m) = \inf_{g\in\mathcal{D}} E(G_{m-1} + c_m g).$$

We prove convergence and rate of convergence results here. Our setting in an infinite dimensional Banach space makes the convergence results nontrivial. The rate of convergence results are of interest in both finite dimensional and infinite dimensional settings. In these results we make assumptions on the element minimizing $E(x)$ (in other words we look for $\inf_{x\in S} E(x)$ for a special domain $S$). A typical assumption in this regard is formulated in terms of the convex hull $A_1(\mathcal{D})$ of the dictionary $\mathcal{D}$.

We have already mentioned above (see (1.5) and below) an example which is of interest in applications in compressed sensing. We now mention another example that attracted a lot of attention in the recent literature. In this example $X$ is a Hilbert space of all real matrices of size $n \times n$ equipped with the Frobenius norm $\|\cdot\|_F$. A dictionary $\mathcal{D}$ is the set of all matrices of rank one normalized in the Frobenius norm. In this case $A_1(\mathcal{D})$ is the set of matrices with nuclear norm not exceeding 1. We are interested in sparse minimization of $E(x) := \|f - x\|_F^2$ (sparse approximation of $f$) with respect to $\mathcal{D}$.

# 2 The Weak Chebyshev Greedy Algorithm

We begin with the following two simple and well-known lemmas.

**Lemma 2.1.** *Let $E$ be a uniformly smooth convex function on a Banach space $X$ and $L$ be a finite-dimensional subspace of $X$. Let $x_L$ denote the point from $L$ at which $E$ attains the minimum:*

$$E(x_L) = \inf_{x\in L} E(x).$$

*Then we have*

$$\langle E'(x_L), \phi\rangle = 0$$

*for any $\phi \in L$.*

*Proof.* Let us assume the contrary: there is a $\phi \in L$ such that $\|\phi\| = 1$ and

$$\langle E'(x_L), \phi\rangle = \beta > 0.$$

It is clear that $x_L \in L \cap D$. For any $\lambda$ we have from the definition of $\rho(E, \lambda)$ that

$$E(x_L - \lambda\phi) + E(x_L + \lambda\phi) \le 2(E(x_L) + \rho(E, \lambda)). \tag{2.1}$$

Next by (1.9)

$$E(x_L + \lambda\phi) \geq E(x_L) + \langle E'(x_L), \lambda\phi\rangle = E(x_L) + \lambda\beta. \tag{2.2}$$

Combining (2.1) and (2.2) we get

$$E(x_L - \lambda\phi) \leq E(x_L) - \lambda\beta + 2\rho(E, \lambda). \tag{2.3}$$

Taking into account that $\rho(E, u) = o(u)$, we find $\lambda' > 0$ such that

$$-\lambda'\beta + 2\rho(E, \lambda') < 0.$$

Then (2.3) gives

$$E(x_L - \lambda'\phi) < E(x_L),$$

which contradicts the assumption that $x_L \in L$ is the point of minimum of $E$. □

**Lemma 2.2.** *For any bounded linear functional $F$ and any dictionary $\mathcal{D}$, we have*

$$\sup_{g\in\mathcal{D}}\langle F, g\rangle = \sup_{f\in A_1(\mathcal{D})}\langle F, f\rangle.$$

*Proof.* The inequality

$$\sup_{g\in\mathcal{D}}\langle F, g\rangle \leq \sup_{f\in A_1(\mathcal{D})}\langle F, f\rangle$$

is obvious. We prove the opposite inequality. Take any $f \in A_1(\mathcal{D})$. Then for any $\epsilon > 0$ there exist $g_1^\epsilon, \dots, g_N^\epsilon \in \mathcal{D}$ and numbers $a_1^\epsilon, \dots, a_N^\epsilon$ such that $a_i^\epsilon > 0$, $a_1^\epsilon + \cdots + a_N^\epsilon \leq 1$ and

$$\|f - \sum_{i=1}^N a_i^\epsilon g_i^\epsilon\| \leq \epsilon.$$

Thus

$$\langle F, f\rangle \leq \|F\|\epsilon + \langle F, \sum_{i=1}^N a_i^\epsilon g_i^\epsilon\rangle \leq \epsilon\|F\| + \sup_{g\in\mathcal{D}}\langle F, g\rangle$$

which proves Lemma 2.2. □

We define the following generalization of the WCGA for convex optimization.

**Weak Chebyshev Greedy Algorithm (WCGA(co)).** We define $G_0 := 0$. Then for each $m \geq 1$ we have the following inductive definition.

(1) $\varphi_m := \varphi_m^{c,\tau} \in \mathcal{D}$ is any element satisfying

$$\langle -E'(G_{m-1}), \varphi_m \rangle \geq t_m \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g \rangle.$$

(2) Define

$$\Phi_m := \Phi_m^{\tau} := \operatorname{span}\{\varphi_j\}_{j=1}^m,$$

and define $G_m := G_m^{c,\tau}$ to be the point from $\Phi_m$ at which $E$ attains the minimum:

$$E(G_m) = \inf_{x \in \Phi_m} E(x).$$

The following lemma is a key lemma in studying convergence and rate of convergence of WCGA(co).

**Lemma 2.3.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E,u)$. Take a number $\epsilon \geq 0$ and an element $f^\epsilon$ from $D$ such that*

$$E(f^\epsilon) \leq \inf_{x \in X} E(x) + \epsilon, \quad f^\epsilon / A(\epsilon) \in A_1(\mathcal{D}),$$

*with some number $A(\epsilon) \geq 1$. Then we have for the WCGA(co)*

$$E(G_m) - E(f^\epsilon) \leq E(G_{m-1}) - E(f^\epsilon)$$

$$+ \inf_{\lambda \geq 0} (-\lambda t_m A(\epsilon)^{-1} (E(G_{m-1}) - E(f^\epsilon)) + 2\rho(E,\lambda)),$$

*for $m = 1, 2, \dots$ .*

*Proof.* It follows from the definition of WCGA(co) that $E(0) \geq E(G_1) \geq E(G_2) \dots$. Therefore, if $E(G_{m-1}) - E(f^\epsilon) \leq 0$ then the claim of Lemma 2.3 is trivial. Assume $E(G_{m-1}) - E(f^\epsilon) > 0$. By Lemma 1.1 we have for any $\lambda$

$$E(G_{m-1} + \lambda \varphi_m) \leq E(G_{m-1}) - \lambda \langle -E'(G_{m-1}), \varphi_m \rangle + 2\rho(E,\lambda) \tag{2.4}$$

and by (1) from the definition of the WCGA(co) and Lemma 2.2 we get

$$\langle -E'(G_{m-1}), \varphi_m \rangle \geq t_m \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g \rangle =$$

$$t_m \sup_{\phi \in A_1(\mathcal{D})} \langle -E'(G_{m-1}), \phi \rangle \geq t_m A(\epsilon)^{-1} \langle -E'(G_{m-1}), f^\epsilon \rangle.$$

By Lemma 2.1 and (1.10) we obtain

$$\langle -E'(G_{m-1}), f^\epsilon \rangle = \langle -E'(G_{m-1}), f^\epsilon - G_{m-1} \rangle \geq E(G_{m-1}) - E(f^\epsilon).$$

Thus,

$$E(G_m) \leq \inf_{\lambda \geq 0} E(G_{m-1} + \lambda \varphi_m)$$

$$\leq E(G_{m-1}) + \inf_{\lambda \geq 0} (-\lambda t_m A(\epsilon)^{-1} (E(G_{m-1}) - E(f^\epsilon)) + 2\rho(E, \lambda), \qquad (2.5)$$

which proves the lemma. □

We proceed to a theorem on convergence of the WCGA. In the formulation of this theorem we need a special sequence which is defined for a given modulus of smoothness $\rho(u)$ and a given $\tau = \{t_k\}_{k=1}^\infty$.

**Definition 2.1.** *Let $\rho(E, u)$ be an even convex function on $(-\infty, \infty)$ with the property:*

$$\lim_{u \to 0} \rho(E, u)/u = 0.$$

*For any $\tau = \{t_k\}_{k=1}^\infty$, $0 < t_k \leq 1$, and $\theta > 0$ we define $\xi_m := \xi_m(\rho, \tau, \theta)$ as a number $u$ satisfying the equation*

$$\rho(E, u) = \theta t_m u. \qquad (2.6)$$

**Remark 2.1.** *Assumptions on $\rho(E, u)$ imply that the function*

$$s(u) := \rho(E, u)/u, \quad u \neq 0, \quad s(0) = 0,$$

*is a continuous increasing function on $[0, \infty)$. Thus 2.6 has a unique solution $\xi_m = s^{-1}(\theta t_m)$ such that $\xi_m > 0$ for $\theta \leq \theta_0 := s(2)$. In this case we have $\xi_m(\rho, \tau, \theta) \leq 2$.*

**Theorem 2.1.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E, u)$. Assume that a sequence $\tau := \{t_k\}_{k=1}^\infty$ satisfies the condition: for any $\theta \in (0, \theta_0]$ we have*

$$\sum_{m=1}^\infty t_m \xi_m(\rho, \tau, \theta) = \infty.$$

*Then*

$$\lim_{m \to \infty} E(G_m) = \inf_{x \in D} E(x).$$

**Corollary 2.1.** *Let a convex function $E$ have modulus of smoothness $\rho(E,u)$ of power type $1 < q \le 2$, that is, $\rho(E,u) \le \gamma u^q$. Assume that*

$$\sum_{m=1}^{\infty} t_m^p = \infty, \quad p = \frac{q}{q-1}. \tag{2.7}$$

*Then*

$$\lim_{m\to\infty} E(G_m) = \inf_{x\in D} E(x).$$

*Proof.* The definition of the WCGA(co) implies that $\{E(G_m)\}$ is a non-increasing sequence. Therefore we have

$$\lim_{m\to\infty} E(G_m) = a.$$

Denote

$$b := \inf_{x\in D} E(x), \quad \alpha := a - b.$$

We prove that $\alpha = 0$ by contradiction. Assume to the contrary that $\alpha > 0$. Then, for any $m$ we have

$$E(G_m) - b \ge \alpha.$$

We set $\epsilon = \alpha/2$ and find $f^\epsilon$ such that

$$E(f^\epsilon) \le b + \epsilon \quad \text{and} \quad f^\epsilon/A(\epsilon) \in A_1(\mathcal{D}),$$

with some $A(\epsilon) \ge 1$. Then, by Lemma 2.3 we get

$$E(G_m) - E(f^\epsilon) \le E(G_{m-1}) - E(f^\epsilon) + \inf_{\lambda\ge 0}(-\lambda t_m A(\epsilon)^{-1}\alpha/2 + 2\rho(E,\lambda)).$$

Let us specify $\theta := \min\left(\theta_0, \frac{\alpha}{8A(\epsilon)}\right)$ and take $\lambda = \xi_m(\rho,\tau,\theta)$. Then we obtain

$$E(G_m) \le E(G_{m-1}) - 2\theta t_m \xi_m.$$

The assumption

$$\sum_{m=1}^{\infty} t_m \xi_m = \infty$$

brings a contradiction, which proves the theorem. □

**Theorem 2.2.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E,u) \le \gamma u^q$, $1 < q \le 2$. Take a number $\epsilon \ge 0$ and an element $f^\epsilon$ from $D$ such that*

$$E(f^\epsilon) \le \inf_{x\in D} E(x) + \epsilon, \quad f^\epsilon/A(\epsilon) \in A_1(\mathcal{D}),$$

*with some number $A(\epsilon) \ge 1$. Then we have for the WCGA(co) ($p := q/(q-1)$)*

$$E(G_m) - \inf_{x\in D} E(x) \le \max\left(2\epsilon, C(q,\gamma)A(\epsilon)^q \left(C(E,q,\gamma) + \sum_{k=1}^m t_k^p\right)^{1-q}\right). \tag{2.8}$$

*Proof.* Denote

$$a_n := E(G_n) - E(f^\epsilon).$$

The sequence $\{a_n\}$ is non-increasing. If $a_n \le 0$ for some $n \le m$ then $E(G_m) - E(f^\epsilon) \le 0$ and $E(G_m) - \inf_{x\in D} E(x) \le \epsilon$ which implies (2.8). Thus we assume that $a_n > 0$ for $n \le m$.

By Lemma 2.3 we have

$$a_m \le a_{m-1} + \inf_{\lambda \ge 0}\left(-\frac{\lambda t_m a_{m-1}}{A(\epsilon)} + 2\gamma\lambda^q\right). \tag{2.9}$$

Choose $\lambda$ from the equation

$$\frac{\lambda t_m a_{m-1}}{A(\epsilon)} = 4\gamma\lambda^q$$

which implies that

$$\lambda = \left(\frac{t_m a_{m-1}}{4\gamma A(\epsilon)}\right)^{\frac{1}{q-1}}.$$

Let

$$A_q := 2(4\gamma)^{\frac{1}{q-1}}.$$

Using the notation $p := \frac{q}{q-1}$ we get from (2.9)

$$a_m \le a_{m-1}\left(1 - \frac{\lambda t_m}{2A(\epsilon)}\right) = a_{m-1}(1 - t_m^p a_{m-1}^{\frac{1}{q-1}}/(A_q A(\epsilon)^p)).$$

Raising both sides of this inequality to the power $\frac{1}{q-1}$ and taking into account the inequality $x^r \le x$ for $r \ge 1$, $0 \le x \le 1$, we obtain

$$a_m^{\frac{1}{q-1}} \le a_{m-1}^{\frac{1}{q-1}}(1 - t_m^p a_{m-1}^{\frac{1}{q-1}}/(A_q A(\epsilon)^p)).$$

We now need a simple known lemma (see [13]).

**Lemma 2.4.** *Suppose that a sequence $y_1 \ge y_2 \ge \cdots \ge 0$ satisfies inequalities*

$$y_k \le y_{k-1}(1 - w_k y_{k-1}), \quad w_k \ge 0,$$

*for $k > n$. Then for $m > n$ we have*

$$\frac{1}{y_m} \ge \frac{1}{y_n} + \sum_{k=n+1}^{m} w_k.$$

*Proof.* It follows from the chain of inequalities

$$\frac{1}{y_k} \ge \frac{1}{y_{k-1}}(1 - w_k y_{k-1})^{-1} \ge \frac{1}{y_{k-1}}(1 + w_k y_{k-1}) = \frac{1}{y_{k-1}} + w_k.$$

□

By Lemma 2.4 with $y_k := a_k^{\frac{1}{q-1}}$, $n = 0$, $w_k = t_m^p/(A_q A(\epsilon)^p)$ we get

$$a_m^{\frac{1}{q-1}} \le C_1(q,\gamma)A(\epsilon)^p \left( C(E,q,\gamma) + \sum_{n=1}^{m} t_n^p \right)^{-1}$$

which implies

$$a_m \le C(q,\gamma)A(\epsilon)^q \left( C(E,q,\gamma) + \sum_{n=1}^{m} t_n^p \right)^{1-q}.$$

Theorem 2.2 is now proved. □

# 3 Relaxation. Co-convex approximation

In this section we study a generalization for optimization problem of relaxed greedy algorithms in Banach spaces considered in [14]. Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given weakness sequence of numbers $t_k \in [0,1]$, $k = 1, \ldots$.

**Weak Relaxed Greedy Algorithm (WRGA(co)).** We define $G_0 := G_0^{r,\tau} := 0$. Then, for each $m \geq 1$ we have the following inductive definition.

(1) $\varphi_m := \varphi_m^{r,\tau} \in \mathcal{D}$ is any element satisfying

$$\langle -E'(G_{m-1}), \varphi_m - G_{m-1} \rangle \geq t_m \sup_{g \in \mathcal{D}} \langle -E'(G_{m-1}), g - G_{m-1} \rangle.$$

(2) Find $0 \leq \lambda_m \leq 1$ such that

$$E((1-\lambda_m)G_{m-1} + \lambda_m \varphi_m) = \inf_{0 \leq \lambda \leq 1} E((1-\lambda)G_{m-1} + \lambda \varphi_m)$$

and define

$$G_m := G_m^{r,\tau} := (1-\lambda_m)G_{m-1} + \lambda_m \varphi_m.$$

**Remark 3.1.** *It follows from the definition of the WRGA that the sequence* $\{E(G_m)\}$ *is a non-increasing sequence.*

We call the WRGA(co) *relaxed* because at the $m$th step of the algorithm we use a linear combination (convex combination) of the previous approximant $G_{m-1}$ and a new element $\varphi_m$. The relaxation parameter $\lambda_m$ in the WRGA(co) is chosen at the $m$th step depending on $E$. We prove here the analogs of Theorems 2.1 and 2.2 for the Weak Relaxed Greedy Algorithm.

**Theorem 3.1.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E,u)$. Assume that a sequence $\tau := \{t_k\}_{k=1}^{\infty}$ satisfies the condition: for any $\theta \in (0, \theta_0]$ we have*

$$\sum_{m=1}^{\infty} t_m \xi_m(\rho, \tau, \theta) = \infty.$$

*Then, for the WRGA(co) we have*

$$\lim_{m \to \infty} E(G_m) = \inf_{x \in A_1(\mathcal{D})} E(x).$$

**Theorem 3.2.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E,u) \leq \gamma u^q$, $1 < q \leq 2$. Then, for a sequence $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \leq 1$, $k = 1, 2, \dots$, we have for any $f \in A_1(\mathcal{D})$ that*

$$E(G_m) - E(f) \leq \left(1 + C_1(q,\gamma) \sum_{k=1}^{m} t_k^p\right)^{1-q}, \quad p := \frac{q}{q-1},$$

*with a positive constant $C_1(q,\gamma)$ which may depend only on $q$ and $\gamma$.*

*Proof.* This proof is similar to the proof of Theorems 2.1 and 2.2. Instead of Lemma 2.3 we use the following lemma.

**Lemma 3.1.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E,u)$. Then, for any $f \in A_1(\mathcal{D})$ we have*

$$E(G_m) \le E(G_{m-1}) + \inf_{0\le\lambda\le 1}(-\lambda t_m(E(G_{m-1})-E(f))+2\rho(E,2\lambda)), \quad m=1,2,\dots.$$

*Proof.* We have

$$G_m := (1-\lambda_m)G_{m-1} + \lambda_m\varphi_m = G_{m-1} + \lambda_m(\varphi_m - G_{m-1})$$

and

$$E(G_m) = \inf_{0\le\lambda\le 1} E(G_{m-1} + \lambda(\varphi_m - G_{m-1})).$$

As for (2.4) we have for any $\lambda$

$$E(G_{m-1} + \lambda(\varphi_m - G_{m-1}))$$

$$\le E(G_{m-1}) - \lambda\langle -E'(G_{m-1}), \varphi_m - G_{m-1}\rangle + 2\rho(E,2\lambda) \tag{3.1}$$

and by (1) from the definition of the WRGA(co) and Lemma 2.2 we get

$$\langle -E'(G_{m-1}), \varphi_m - G_{m-1}\rangle \ge t_m \sup_{g\in\mathcal{D}}\langle -E'(G_{m-1}), g - G_{m-1}\rangle =$$

$$t_m \sup_{\phi\in A_1(\mathcal{D})}\langle -E'(G_{m-1}), \phi - G_{m-1}\rangle \ge t_m\langle -E'(G_{m-1}), f - G_{m-1}\rangle.$$

By (1.10) we obtain

$$\langle -E'(G_{m-1}), f - G_{m-1}\rangle \ge E(G_{m-1}) - E(f).$$

Thus,

$$E(G_m) \le \inf_{0\le\lambda\le 1} E(G_{m-1} + \lambda(\varphi_m - G_{m-1}))$$

$$\le E(G_{m-1}) + \inf_{0\le\lambda\le 1}(-\lambda t_m(E(G_{m-1}) - E(f)) + 2\rho(E,2\lambda), \tag{3.2}$$

which proves the lemma. □

The remaining part of the proof uses the inequality (3.2) in the same way relation (2.5) was used in the proof of Theorems 2.1 and 2.2. The only additional difficulty here is that we are optimizing over $0 \le \lambda \le 1$. In the proof of Theorem 3.1 we choose $\theta = \alpha/8$, assuming that $\alpha$ is small enough to guarantee that $\theta \le \theta_0$ and $\lambda = \xi_m(\rho, \tau, \theta)/2$.

We proceed to the proof of Theorem 3.2. Denote

$$a_n := E(G_n) - E(f).$$

The sequence $\{a_n\}$ is non-increasing. If $a_n \le 0$ for some $n \le m$ then $E(G_m) - E(f) \le 0$ which implies Theorem 3.2. Thus we assume that $a_n > 0$ for $n \le m$. We obtain from Lemma 3.1

$$a_m \le a_{m-1} + \inf_{0 \le \lambda \le 1} (-\lambda t_m a_{m-1} + 2\gamma(2\lambda)^q).$$

We choose $\lambda$ from the equation

$$\lambda t_m a_{m-1} = 4\gamma(2\lambda)^q \tag{3.3}$$

if it is not greater than 1 and choose $\lambda = 1$ otherwise. The sequence $\{a_k\}$ is monotone decreasing and therefore we may choose $\lambda = 1$ only at first $n$ steps and then choose $\lambda$ from (3.3). Then we get for $k \le n$

$$a_k \le a_{k-1}(1 - t_k/2)$$

and

$$a_n \le a_0 \prod_{k=1}^{n} (1 - t_k/2). \tag{3.4}$$

For $k > n$ we have

$$a_k \le a_{k-1}(1 - \lambda t_k/2), \quad \lambda = \left( \frac{t_m a_{m-1}}{2^{2+q}\gamma} \right)^{\frac{1}{q-1}}. \tag{3.5}$$

As in the proof of Theorem 2.2 we obtain using Lemma 2.4

$$\frac{1}{y_m} \ge \frac{1}{y_n} + \sum_{k=n+1}^{m} w_k, \quad y_k := a_k^{\frac{1}{q-1}}, \quad w_k := \frac{t_k^p}{2(2^{2+q}\gamma)^{\frac{1}{q-1}}}.$$

By (3.4) we get

$$\frac{1}{y_n} \ge \frac{1}{y_0} \prod_{k=1}^{n} (1 - t_k/2)^{\frac{1}{1-q}}.$$

Next,

$$\prod_{k=1}^{n}(1-t_k/2)^{\frac{1}{1-q}} \ge \prod_{k=1}^{n}(1+t_k/2)^{\frac{1}{q-1}} \ge \prod_{k=1}^{n}(1+t_k/2)$$

$$\ge 1+\frac{1}{2}\sum_{k=1}^{n} t_k \ge 1+\frac{1}{2}\sum_{k=1}^{n} t_k^p.$$

Combining the above inequalities we complete the proof. □

# 4 Free relaxation

Both of the above algorithms, the WCGA(co) and the WRGA(co), use the functional $E'(G_{m-1})$ in a search for the $m$th element $\varphi_m$ from the dictionary to be used in optimization. The construction of the approximant in the WRGA(co) is different from the construction in the WCGA(co). In the WCGA(co) we build the approximant $G_m$ so as to maximally use the minimization power of the elements $\varphi_1, \dots, \varphi_m$. The WRGA(co) by its definition is designed for working with functions from $A_1(\mathcal{D})$. In building the approximant in the WRGA(co) we keep the property $G_m \in A_1(\mathcal{D})$. As we mentioned in Section 3 the relaxation parameter $\lambda_m$ in the WRGA(co) is chosen at the $m$th step depending on $E$. The following modification of the above idea of relaxation in greedy approximation will be studied in this section (see [15]).

**Weak Greedy Algorithm with Free Relaxation (WGAFR(co)).** Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $G_0 := 0$. Then for each $m \ge 1$ we have the following inductive definition.

(1) $\varphi_m \in \mathcal{D}$ is any element satisfying

$$\langle -E'(G_{m-1}), \varphi_m \rangle \ge t_m \sup_{g\in\mathcal{D}} \langle -E'(G_{m-1}), g\rangle.$$

(2) Find $w_m$ and $\lambda_m$ such that

$$E((1-w_m)G_{m-1}+\lambda_m\varphi_m) = \inf_{\lambda,w} E((1-w)G_{m-1}+\lambda\varphi_m)$$

and define

$$G_m := (1-w_m)G_{m-1}+\lambda_m\varphi_m.$$

**Remark 4.1.** *It follows from the definition of the WGAFR(co) that the sequence $\{E(G_m)\}$ is a non-icreasing sequence.*

We begin with an analog of Lemma 2.3.

**Lemma 4.1.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E,u)$. Take a number $\epsilon \geq 0$ and an element $f^\epsilon$ from $D$ such that*

$$E(f^\epsilon) \leq \inf_{x\in D} E(x) + \epsilon, \quad f^\epsilon/A(\epsilon) \in A_1(\mathcal{D}),$$

*with some number $A(\epsilon) \geq 1$. Then we have for the WGAFR(co)*

$$E(G_m) - E(f^\epsilon) \leq E(G_{m-1}) - E(f^\epsilon)$$

$$+ \inf_{\lambda\geq 0}(-\lambda t_m A(\epsilon)^{-1}(E(G_{m-1}) - E(f^\epsilon)) + 2\rho(E, C_0\lambda)),$$

*for $m = 1, 2, \dots$ .*

*Proof.* By the definition of $G_m$

$$E(G_m) \leq \inf_{\lambda\geq 0, w} E(G_{m-1} - wG_{m-1} + \lambda\varphi_m).$$

As in the arguments in the proof of Lemma 2.3 we use Lemma 1.1

$$E(G_{m-1} + \lambda\varphi_m - wG_{m-1}) \leq E(G_{m-1})$$

$$-\lambda\langle -E'(G_{m-1}), \varphi_m\rangle - w\langle E'(G_{m-1}), G_{m-1}\rangle + 2\rho(E, \|\lambda\varphi_m - wG_{m-1}\|) \quad (4.1)$$

and estimate

$$\langle -E'(G_{m-1}), \varphi_m\rangle \geq t_m \sup_{g\in\mathcal{D}}\langle -E'(G_{m-1}), g\rangle =$$

$$t_m \sup_{\phi\in A_1(\mathcal{D})} \langle -E'(G_{m-1}), \phi\rangle \geq t_m A(\epsilon)^{-1}\langle -E'(G_{m-1}), f^\epsilon\rangle.$$

We set $w^* := \lambda t_m A(\epsilon)^{-1}$ and obtain

$$E(G_{m-1} - w^* G_{m-1} + \lambda\varphi_m)$$

$$\leq E(G_{m-1}) - \lambda t_m A(\epsilon)^{-1}\langle -E'(G_{m-1}), f^\epsilon - G_{m-1}\rangle. \quad (4.2)$$

By (1.10) we obtain

$$\langle -E'(G_{m-1}), f^\epsilon - G_{m-1}\rangle \geq E(G_{m-1}) - E(f^\epsilon).$$

Thus,

$$E(G_m) \leq E(G_{m-1})$$

$$+\inf_{\lambda\geq 0}(-\lambda t_m A(\epsilon)^{-1}(E(G_{m-1})-E(f^\epsilon))+2\rho(E,\|\lambda\varphi_m-w^*G_{m-1}\|). \quad (4.3)$$

We now estimate

$$\|w^*G_{m-1}-\lambda\varphi_m\|\leq w^*\|G_{m-1}\|+\lambda.$$

Next, $E(G_{m-1})\leq E(0)$ and, therefore, $G_{m-1}\in D$. Our assumption on boundedness of $D$ implies that $\|G_{m-1}\|\leq C_1$. Thus, under assumption $A(\epsilon)\geq 1$ we get

$$w^*\|G_{m-1}\|\leq C_1\lambda t_m\leq C_1\lambda.$$

Finally,

$$\|w^*G_{m-1}-\lambda\varphi_m\|\leq C_0\lambda.$$

This completes the proof of Lemma 4.1. □

We now prove a convergence theorem for an arbitrary uniformly smooth convex function. Modulus of smoothness $\rho(E,u)$ of a uniformly smooth convex function is an even convex function such that $\rho(E,0)=0$ and

$$\lim_{u\to 0}\rho(E,u)/u=0.$$

**Theorem 4.1.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E,u)$. Assume that a sequence $\tau:=\{t_k\}_{k=1}^\infty$ satisfies the following condition. For any $\theta\in(0,\theta_0]$ we have*

$$\sum_{m=1}^\infty t_m\xi_m(\rho,\tau,\theta)=\infty. \quad (4.4)$$

*Then, for the WGAFR(co) we have*

$$\lim_{m\to\infty}E(G_m)=\inf_{x\in D}E(x).$$

*Proof.* By Remark 4.1, $\{E(G_m)\}$ is a non-increasing sequence. Therefore we have

$$\lim_{m\to\infty}E(G_m)=a.$$

Denote

$$b:=\inf_{x\in D}E(x),\quad \alpha:=a-b.$$

We prove that $\alpha = 0$ by contradiction. Assume to the contrary that $\alpha > 0$. Then, for any $m$ we have

$$E(G_m) - b \geq \alpha.$$

We set $\epsilon = \alpha/2$ and find $f^\epsilon$ such that

$$E(f^\epsilon) \leq b + \epsilon \quad \text{and} \quad f^\epsilon / A(\epsilon) \in A_1(\mathcal{D}),$$

with some $A(\epsilon) \geq 1$. Then, by Lemma 4.1 we get

$$E(G_m) - E(f^\epsilon) \leq E(G_{m-1}) - E(f^\epsilon) + \inf_{\lambda \geq 0}(-\lambda t_m A(\epsilon)^{-1}\alpha/2 + 2\rho(E, C_0\lambda)).$$

Let us specify $\theta := \min\left(\theta_0, \frac{\alpha}{8A(\epsilon)}\right)$ and take $\lambda = C_0\xi_m(\rho, \tau, \theta)$. Then we obtain

$$E(G_m) \leq E(G_{m-1}) - 2\theta t_m \xi_m.$$

The assumption

$$\sum_{m=1}^{\infty} t_m \xi_m = \infty$$

brings a contradiction, which proves the theorem. □

**Theorem 4.2.** *Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E, u) \leq \gamma u^q$, $1 < q \leq 2$. Take a number $\epsilon \geq 0$ and an element $f^\epsilon$ from $D$ such that*

$$E(f^\epsilon) \leq \inf_{x \in D} E(x) + \epsilon, \quad f^\epsilon / A(\epsilon) \in A_1(\mathcal{D}),$$

*with some number $A(\epsilon) \geq 1$. Then we have ($p := q/(q-1)$)*

$$E(G_m) - \inf_{x \in D} E(x) \leq \max\left(2\epsilon, C(q,\gamma)A(\epsilon)^q \left(C(E,q,\gamma) + \sum_{k=1}^{m} t_k^p\right)^{1-q}\right). \tag{4.5}$$

*Proof.* Denote

$$a_n := E(G_n) - E(f^\epsilon).$$

By Lemma 4.1 we have

$$a_m \leq a_{m-1} + \inf_{\lambda \geq 0}\left(-\frac{\lambda t_m a_{m-1}}{A(\epsilon)} + 2\gamma(C_0\lambda)^q\right). \tag{4.6}$$

Choose $\lambda$ from the equation

$$\frac{\lambda t_m a_{m-1}}{A(\epsilon)} = 4\gamma(C_0\lambda)^q.$$

The rest of the proof repeats the argument from the proof of Theorem 2.2. $\square$

# 5 Comments

We already mentioned in the Introduction that the technique used in this paper is a slight modification of the corresponding technique developed in approximation theory (see [14], [16] and the book [17]). We now discuss this in more detail. We pointed out in the Introduction that at a first glance the settings of approximation and optimization problems are very different. In the approximation problem an element $f \in X$ is given and our task is to find a sparse approximation of it. In optimization theory an energy function $E(x)$ is given and we should find an approximate sparse solution to the minimization problem. It turns out that the same technique can be used for solving both problems. In nonlinear approximation we use greedy algorithms, for instance WCGA and WGAFR, for solving this problem. The greedy step is the one where we look for $\varphi_m \in \mathcal{D}$ satisfying

$$F_{f_{m-1}}(\varphi_m) \geq t_m \sup_{g\in\mathcal{D}} F_{f_{m-1}}(g).$$

This step is based on the norming functional $F_{f_{m-1}}$. As we pointed out in the Introduction the norming functional $F_{f_{m-1}}$ is the derivative of the norm function $E(x) := \|x\|$. Clearly, we can reformulate our problem of approximation of $f$ as an optimization problem with $E(x) := \|f-x\|$. It is a convex function, however, it is not a uniformly smooth function in the sense of smoothness of convex functions. A way out of this problem is to consider $E(f,x,q) := \|f-x\|^q$ with appropriate $q$. For instance, it is known (see [4]) that if $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$, then $E(f,x,q)$ is a uniformly smooth convex function with modulus of smoothness of order $u^q$. Next,

$$E'(f,x,q) = -q\|f-x\|^{q-1}F_{f-x}.$$

Therefore, the algorithms WCGA(co), WRGA(co) and WGAFR(co) coincide in this case with the corresponding algorithms WCGA, WRGA and WGAFR

from approximation theory. In the proofs of approximation theory results we use inequality (1.2) and the trivial inequality

$$\|x+uy\| \ge F_x(x+uy) = \|x\| + uF_x(y). \tag{5.1}$$

In the proofs of optimization theory results we use Lemma 1.1 instead of inequality (1.2) and the convexity inequality (1.9) instead of (5.1). The rest of the proofs uses the same technique of solving the corresponding recurrent inequalities.

Our smoothness assumption on $E$ was used in the proofs of all theorems from Sections 2–4 in the form of Lemma 1.1. This means that in all those theorems the assumption that $E$ has modulus of smoothness $\rho(E,u)$ can be replaced by the assumption that $E$ satisfies the inequality

$$E(x+uy) - E(x) - u\langle E'(x), y\rangle \le 2\rho(E, u\|y\|), \quad x \in D. \tag{5.2}$$

Moreover, in Section 3, where we consider the WRGA(co), the approximants $G_m$ are forced to stay in the $A_1(\mathcal{D})$. Therefore, in Theorems 3.1 and 3.2 we can use the following inequality instead of (5.2)

$$E(x+u(y-x)) - E(x) - u\langle E'(x), y-x\rangle \le 2\rho(E, u\|y-x\|), \tag{5.3}$$

for $x, y \in A_1(\mathcal{D})$ and $u \in [0,1]$.

We note that smoothness assumptions in the form of (5.3) with $\rho(E, u\|y-x\|)$ replaced by $C\|y-x\|^q$ were used in [20]. The authors studied the version of WRGA(co) with weakness sequence $t_k = 1$, $k = 1, 2, \dots$. They proved Theorem 3.2 in this case. Their proof alike our proof in Section 3 is very close to the corresponding proof from greedy approximation (see [14], [16] Section 3.3 or [17] Section 6.3).

We now make some general remarks on the results of this paper. As we already pointed out in Introduction a typical problem of convex optimization is to find an approximate solution to the problem

$$w := \inf_x E(x). \tag{5.4}$$

In this paper we are interested in sparse (with respect to a given dictionary $\mathcal{D}$) solutions of (5.4). This means that we are solving the following problem instead of (5.4). For a given dictionary $\mathcal{D}$ consider the set of all $m$-term polynomials with respect to $\mathcal{D}$:

$$\Sigma_m(\mathcal{D}) := \{x \in X : x = \sum_{i=1}^m c_i g_i, \quad g_i \in \mathcal{D}\}.$$

We solve the following *sparse optimization problem*

$$w_m := \inf_{x \in \Sigma_m(\mathcal{D})} E(x). \tag{5.5}$$

In this paper we have used greedy-type algorithms to solve (approximately) problem (5.5). Results of the paper show that it turns out that greedy-type algorithms with respect to $\mathcal{D}$ solve problem (5.4) too.

We are interested in a solution from $\Sigma_m(\mathcal{D})$. Clearly, when we optimize a linear form $\langle F, g\rangle$ over the dictionary $\mathcal{D}$ we obtain the same value as optimization over the convex hull $A_1(\mathcal{D})$. We often use this property (see Lemma 2.2). However, at the greedy step of our algorithms we choose

(1) $\varphi_m := \varphi_m^{c,\tau} \in \mathcal{D}$ is **any** element satisfying

$$\langle -E'(G_{m-1}), \varphi_m\rangle \ge t_m \sup_{g\in\mathcal{D}} \langle -E'(G_{m-1}), g\rangle.$$

Thus if we replace the dictionary $\mathcal{D}$ by its convex hull $A_1(\mathcal{D})$ we may take an element satisfying the above greedy condition which is not from $\mathcal{D}$ and could be even an infinite combination of the dictionary elements.

Next, we begin with a Banach space $X$ and a convex function $E(x)$ defined on this space. Properties of this function $E$ are formulated in terms of Banach space $X$. If instead of Banach space $X$ we consider another Banach space, for instance, the one generated by $A_1(\mathcal{D})$ as a unit ball then the properties of $E$ will change. For instance, a typical example of $E$ could be $E(x) := \|f - x\|^q$ with $\|\cdot\|$ being the norm of Banach space $X$. Then our assumption that the set $D := \{x : E(x) \le E(0)\}$ is bounded is satisfied. However, this set is not necessarily bounded in the norm generated by $A_1(\mathcal{D})$.

**Acknowledgements.** This paper was motivated by the IMA Annual Program Workshop "Machine Learning: Theory and Computation" (March 26–30, 2012), in particular, by talks of Steve Wright and Pradeep Ravikumar. The author is very thankful to Arkadi Nemirovski for an interesting discussion of the results and for his remarks.

# References


[1] T. Blumensath and M.E. Davies, Gradient Pursuits, IEEE Transactions in Signal Processing, **56** (2008), 2370–2382.

[2] T. Blumensath and M.E. Davies, Stagewise Weak Gradient Pursuits, IEEE Transactions in Signal Processing, **57** (2009), 4333–4346.

[3] J.M. Borwein and A.S. Lewis, Convex Analysis and Nonlinear Optimization. Theory and Examples, Canadian Mathematical Society, Springer, 2006.

[4] J. Borwein, A.J. Guirao, P. Hajek, and J. Vanderwerff, Uniformly convex functions an Banach spaces, Proceedings of the American Mathematical Society, **137(3)** (2009), 1081–1091.

[5] V. Chandrasekaran, B. Recht, P.A. Parrilo, and A.S. Willsky, The convex geometry of linear inverse problems, Proceedings of the 48th Annual Allerton Conference on Communication, Control and Computing, 2010, 699–703.

[6] R.A. DeVore, Nonlinear approximation, Acta Numerica, **7** (1998), 51–150.

[7] M.A.T. Figueiredo, R.D. Nowak, and S.J. Wright, Gradient projection for Sparse Reconstruction: Application to Compressed Sensing and Other Inverse Problems, IEEE, Selected Topics in Signal Processing, **1** (2007), 586–597.

[8] M. Frank and P. Wolfe, An algorithm for quadratic programming, Naval Research Logistics Quarterly, **3** (1956), 95–110.

[9] V.G. Karmanov, Mathematical Programming, Mir Publishers, Moscow, 1989.

[10] Yu. Nesterov, Introductory Lectures on Convex Optimization: A Basic Course, Kluwer Academic Publishers, Boston, 2004.

[11] B.N. Pshenichnyi and Yu.M. Danilin, Numerical Methods in Extremal Problems [in Russian], Nauka,. Moscow, 1975.

[12] S. Shalev-Shwartz, N. Srebro, and T. Zhang, Trading accuracy for sparsity in optimization problems with sparsity constrains, SIAM Journal on Optimization, **20(6)** (2010), 2807–2832.

[13] V.N. Temlyakov, Weak Greedy Algorithms, Adv. Comput. Math., **12** (2000), 213–227.

[14] V.N. Temlyakov, Greedy algorithms in Banach spaces, Adv. Comput. Math., **14** (2001), 277–292.

[15] V.N. Temlyakov, Relaxation in greedy approximation, Constructive Approximation, **28** (2008), 1–25.

[16] V.N. Temlyakov, Greedy approximation, Acta Numerica, **17** (2008), 235–409.

[17] V.N. Temlyakov, Greedy approximation, Cambridge University Press, 2011.

[18] V.N. Temlyakov, Greedy approximation in convex optimization, IMI Preprint, 2012:03, 1–25;

[19] V.N. Temlyakov, Greedy expansions in convex optimization, IMI Preprint, 2012:03, 1–27;

[20] A. Tewari, P. Ravikumar, and I.S. Dhillon, Greedy Algorithms for Structurally Constrained High Dimensional Problems, prerint, (2012), 1–10.

[21] T. Zhang, Sequential greedy approximation for certain convex optimization problems, IEEE Transactions on Information Theory, **49(3)** (2003), 682–691.