Constr Approx (2015) 41:269–296
DOI 10.1007/s00365-014-9272-0

CONSTRUCTIVE
APPROXIMATION

Greedy Approximation in Convex Optimization

V. N. Temlyakov

Received: 29 August 2013 / Revised: 4 August 2014 / Accepted: 16 September 2014 /
Published online: 21 January 2015

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 a few 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 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.

Keywords Sparse · Optimization · Greedy · Banach space · Convergence rate

Communicated by Joel A. Tropp.

Research was supported by NSF Grant DMS-1160841.

V. N. Temlyakov (✉)
University of South Carolina, Columbia, SC, USA
e-mail: temlyakovv@gmail.com

V. N. Temlyakov
Steklov Institute of Mathematics, Moscow, Russia

Mathematics Subject Classification Primary: 41A46 · Secondary: 65K05 · 41A65 · 46B20

1 Introduction

To sparse approximate solutions to convex optimization problems, let us apply the technique known in nonlinear approximation as *greedy approximation*. Greedy approximation has important applications in signal processing, optimization, statistics, and stochastic PDEs. Greedy algorithms are thoroughly studied in approximation theory, functional analysis, learning theory, signal processing, and optimization. Very often, researchers working in one area are not aware of parallel techniques developed in other areas of research. The main goal of this paper is to demonstrate how typical methods developed in greedy approximation in Banach spaces can be used for studying convex optimization problems.

A typical problem of sparse approximation is the following [15,40]. Let X be a Banach space with norm $\|\cdot\|$ and $\mathcal{D}$ be a set of elements of X. For a given $\mathcal{D}$, consider the set of all m-term linear combinations with respect to $\mathcal{D}$ (m-sparse with respect to $\mathcal{D}$ elements):

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

We are interested in approximation of a given $f \in X$ by elements of $\Sigma_m(\mathcal{D})$. The best we can do is

$$\sigma_m(f, \mathcal{D}) := \inf_{x \in \Sigma_m(\mathcal{D})} \|f - x\|. \tag{1}$$

Greedy algorithms in approximation theory are designed to provide a simple way to build good approximants of f from $\Sigma_m(\mathcal{D})$. Clearly, problem (1) is an optimization problem of $E_f(x) := \|f - x\|$ over the manifold $\Sigma_m(\mathcal{D})$.

A typical problem of convex optimization is to find an approximate solution to the problem

$$\inf_x E(x) \tag{2}$$

under the assumption that E is a convex function. In the case that we are optimizing over the whole space X, it is called an *unconstrained optimization problem*. In many cases, we are interested either in optimizing over x of special structure (for instance, $x \in \Sigma_m(\mathcal{D})$, as above) or in optimizing over x from a given domain D (*constrained optimization problem*). Greedy algorithms are used for finding an approximate solution of special structure for problem (2).

Usually in convex optimization, the function E is defined on a finite dimensional space $\mathbb{R}^n$ [9,31]. 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 motivating argument to study this problem in the infinite dimensional setting is that in many contemporary numerical applications, ambient space $\mathbb{R}^n$ involves a large dimension n 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 [46] that in many engineering applications, researchers are interested in an approximate solution of problem (2) as a linear combination of a few 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, [1,3–5,11,12,17,19,22–24,34,45,46]). We refer the reader to the papers [1] and [23] for concise surveys of recent results on greedy algorithms from the point of view of convex optimization and signal processing.

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 sparse approximation both with regard to bases and with regard to redundant systems. In fact one fundamental principle allows us to build good algorithms both for arbitrary redundant systems and for very simple, well structured bases such as the Haar basis: 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. Varying that functional and the ways of constructing (choosing coefficients of the linear combination) the m-term approximant from the already found m elements of the dictionary yields different types of greedy algorithms. For instance, if the corresponding linear combination is a convex combination, then it is the Relaxed Greedy Algorithm (Frank-Wolfe-type algorithm) studied in Sect. 2. In Sect. 3, we study the Chebyshev-type greedy algorithm, which is known in signal processing under the name Fully Corrective Forward Greedy Selection. We use the name Chebyshev in this algorithm because at the approximation step of the algorithm, we use a best approximation operator which bears the name of the *Chebyshev projection* or the *Chebyshev operator*. In the case of Hilbert space, the Chebyshev projection is the orthogonal projection, and it is reflected in the name of the algorithm. In this paper, we discuss the *weak* version of the greedy algorithms. The term *weak* in the definition of these algorithms means that at the greedy step of selection of a new element of the dictionary, we do not shoot for the optimal element of the dictionary which realizes the corresponding supremum, but are satisfied with a weaker property than being optimal. The obvious reason for this is that we do not know in general that the optimal element exists. Another, practical reason is that the weaker the assumption, the easier it is to satisfy it and, therefore, the easier it is to realize in practice. We note that results of this paper provide the same upper bounds for the rate of convergence for the weak versions of the algorithms (in the case $t_k = t$) as for the strong versions ($t = 1$) of the algorithms.

At 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 (loss 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 to find a sparse, with respect to $\mathcal{D}$, solution of the problem (2). We consider three greedy algorithms here: the Weak Chebyshev Greedy Algorithm (WCGA), the Weak Relaxed Greedy Algorithm (WRGA), and the Weak Greedy Algorithm with Free Relaxation (WGAFR). The names of these algorithms used above are from approximation theory. The WCGA is a generalization to a Banach space setting [36] of the Weak Orthogonal Greedy Algorithm (WOGA), which is very important in signal processing and compressed sensing. The WOGA is known in signal processing under the name Weak Orthogonal Matching Pursuit (WOMP). It is used for exact recovery of sparse signals and for approximation of signals by sparse ones. An analog of the WCGA in convex optimization was introduced in [34] under the name Fully Corrective Forward Greedy Selection. The WRGA is the approximation theory analog of the classical Frank-Wolfe algorithm, introduced in [21] and studied in many papers (see, for instance, [12,14,17,18,22,23]). This algorithm was rediscovered in statistics and approximation theory in [2] and [25] (see [40] for further discussion).

Two different ideas have been used in the WRGA and WCGA. The first idea was that of relaxation (we use here terminology from approximation theory). The corresponding algorithms, (including WRGA) were designed for approximation of functions from the convex hull $\operatorname{conv}(\mathcal{D})$ of the given system (dictionary) $\mathcal{D}$. The second idea is used in the WCGA to build the best approximant from the $\operatorname{span}(\varphi_1, \dots, \varphi_m)$ of already chosen elements $\varphi_j \in \mathcal{D}$, instead of the use of only one element φ_m for an update of the approximant, as is done in WRGA.

The realization of both ideas resulted in the construction of algorithms (WRGA and WCGA) that are good for approximation of functions from $\operatorname{conv}(\mathcal{D})$. The advantage of WCGA over WRGA is that WCGA (under some assumptions on the weakness sequence τ) converges for each $f \in X$ in any uniformly smooth Banach space. The WRGA is simpler than the WCGA in the sense of computational complexity. However, the WRGA has limited applicability. It converges only for elements of the closure of the convex hull of a dictionary.

The WGAFR [38] combines good features of both the WRGA and the WCGA algorithms. The WGAFR performs in the same way as the WCGA from the point of view of convergence and rate of convergence, and outperforms the WCGA in terms of computational complexity. In the WGAFR, we are optimizing over two parameters at each step of the algorithm. In other words, we are looking for the best approximation from a 2-dimensional linear subspace at each step. As far as we know, an analog of the WGAFR has not been studied in optimization.

A number of greedy algorithms were successfully used in compressed sensing. The early theoretical results [27] on the widths did not consider the question of practical recovery methods. The celebrated contribution of the work by Candes-Tao and Donoho was to show that the recovery can be done by the ℓ_1 minimization. While the ℓ_1 minimization technique plays an important role in designing computationally tractable recovery methods, its complexity is still impractical for many applications. An attractive alternative to the ℓ_1 minimization is a family of greedy algorithms. They include the Orthogonal Greedy Algorithm [called the Orthogonal Matching Pursuit (OMP) in signal processing], the Regularized Orthogonal Matching Pursuit [30], Compressive Sampling Matching Pursuit (CoSaMP) [29], and the Subspace Pursuit (SP) [13]. The OMP is simpler than CoSaMP and SP, however, at the time of invention of

CoSaMP and SP these algorithms provided exact recovery of sparse signals and the Lebesgue-type inequalities for dictionaries satisfying the Restricted Isometry Property (RIP) [13,29]. The corresponding results for the OMP were not known at that time. Later, Zhang [48] proved exact recovery of sparse signals and the Lebesgue-type inequalities for the OMP under RIP condition on a dictionary. The reader can find an extension of Zhang's results for the Chebyshev Greedy Algorithm in [28] and [43].

The idea of extending greedy-type algorithms designed for approximation to solving optimization problems with sparsity constraints was recently used in a number of papers. Blumensath [4,5] extends the Iterative Hard Thresholding (IHT) algorithm [8] to the optimization setting. Bahmani et al. [1] extend the CoSaMP to the optimization setting. Variants of the coordinate descent algorithms were considered in [3] and [19]. This confirms that the step from greedy approximation to greedy optimization to which this paper is devoted is a timely and important step. We only provide a theoretical analysis of the algorithms here. The earlier version of this paper is published in [41].

Before we proceed to a detailed discussion, we formulate some novel contributions of the paper.

Novelty We provide a unified program to study convergence and rate of convergence of greedy algorithms. This way works in both the approximation theory setting and in the convex optimization setting. It works for an arbitrary dictionary $\mathcal{D}$. We show that this technique works both for the classical Frank-Wolfe algorithm and for the new algorithms in convex optimization (WCGA(co) and WGAFR(co)). We introduce and study a new algorithm, the Weak Greedy Algorithm with Free Relaxation designed for convex optimization. It combines good features of known algorithms—the simplicity of the Frank-Wolfe-type algorithms and the power of fully corrective algorithms. All the theorems in Sects. 2–4 are new results. In the setting with an arbitrary dictionary $\mathcal{D}$, these results are sharp. This follows from sharpness of the approximation theory analogs of these results (see [40], Ch. 6). Better rate of convergence results can be obtained under stronger assumptions on E for dictionaries satisfying certain conditions [32,34,44].

We note that in the study of greedy-type algorithms in approximation theory [40], emphasis is put on the theory of approximation with respect to an arbitrary dictionary $\mathcal{D}$. The reader can find a detailed discussion of applications of greedy-type algorithms in approximation, classification, and boosting in [46]. The reader can find examples of specific dictionaries of interest in [17,22,24,40,45], and [23]. We present some results on sparse solutions for convex optimization problems in the setting with an arbitrary dictionary $\mathcal{D}$. In this paper, we analyze algorithms with exact evaluations. There are known results on algorithms with approximate evaluation: for the Frank-Wolfe algorithm, see [18]; for other convex optimization algorithms, see [16] and [46]; for approximation algorithms, see [37]. Clearly, the corresponding dictionary element φ_m that we choose at the greedy step may not be unique. Our results apply for any realization (any choice of φ_m) of the algorithms.

Greedy algorithms for convex optimization 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 span $\mathcal{D}$ is X. For notational convenience, in this paper symmetric dictionaries are considered. Results of the paper also hold for non-symmetric dictionaries with straight forward modifications. For instance, no modifications are needed in the case of WRGA(co); in the case of WCGA(co), we need to work at the greedy step (C1) with absolute values of the quantities $\langle -E'(G_{m-1}), \varphi_m\rangle$ and $\langle -E'(G_{m-1}), g\rangle$. We denote the closure (in X) of the convex hull of $\mathcal{D}$ by $A_1(\mathcal{D})$. In other words $A_1(\mathcal{D})$ is the closure of conv($\mathcal{D}$). We use this notation because it has become a standard notation in relevant greedy approximation literature.

It is pointed out in [20] that there has been considerable interest in solving the convex unconstrained optimization problem

$$\min_x \frac{1}{2}\|y-\Phi x\|_2^2 + \lambda\|x\|_1, \tag{3}$$

where $x \in \mathbb{R}^n$, $y \in \mathbb{R}^k$, Φ is an $k \times n$ matrix, λ is a nonnegative parameter, $\|v\|_2$ denotes the Euclidian norm of v, and $\|v\|_1$ is the ℓ_1 norm of v. Problems of the form (3) have become familiar over the past three decades, particularly in statistical and signal processing contexts. Problem (3) is closely related to the following convex constrained optimization problem:

$$\min_x \frac{1}{2}\|y-\Phi x\|_2^2 \quad \text{subject to} \quad \|x\|_1 \le A. \tag{4}$$

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 Φ. The condition $y \in A_1(\mathcal{D})$ is equivalent to the existence of $x \in \mathbb{R}^m$ such that $y = \Phi x$ and

$$\|x\|_1 := |x_1| + \cdots + |x_m| \le 1. \tag{5}$$

As a direct corollary of Theorems 6.8 and 6.23 from [40], 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:

$$\|y_k\|_2 \le Ck^{-1/2}, \tag{6}$$

where y_k is the residual after k iterations. The bound (6) holds for any $\mathcal{D}$ (any Φ).

For further discussion of an optimization problem more general than (3) and its application in machine learning, we refer the reader to [47].

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:

$$\rho(E,u) := \frac{1}{2} \sup_{x\in D, \|y\|=1} |E(x+uy) + E(x-uy) - 2E(x)|. \tag{7}$$

A typical assumption in convex optimization is of the form ($\|y\| = 1$):

$$|E(x + uy) - E(x) - \langle E'(x), uy\rangle| \le Cu^2,$$

which corresponds to the case $\rho(E, u)$ of order u^2. We assume that E is Fréchet differentiable. Then convexity of E implies that for any x, y,

$$E(y) \ge E(x) + \langle E'(x), y - x\rangle, \tag{8}$$

or, in other words,

$$E(x) - E(y) \le \langle E'(x), x - y\rangle = \langle -E'(x), y - x\rangle. \tag{9}$$

The above assumptions that E is a Fréchet differentiable convex function and D is a bounded domain guarantee that $\inf_x E(x) > -\infty$.

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.

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 (4) and below) an example which is of interest in applications in compressed sensing. We mention another example that attracted a lot of attention in the recent literature (see, for instance, [17,23,24,45]). 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 Frank-Wolfe-Type Algorithm

In this section, we discuss an algorithm for finding a sparse approximate solution for a constrained optimization problem

$$\inf_{x\in A_1(\mathcal{D})} E(x).$$

The Frank-Wolfe algorithm was introduced in [21] for solving a constrained optimization problem. Later similar algorithms were used in statistics and approximation theory. A characteristic feature of the Frank-Wolfe-type algorithm is that it is a greedy-type algorithm that builds at each iteration a new approximant as a convex combination of the previous approximant and a new element chosen from the dictionary. There are several versions of this algorithm. The reader can find a corresponding discussion in [23] and [40]. In this section, we study a generalization for optimization problems of relaxed greedy algorithms in Banach spaces considered in [36].

Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given weakness sequence of numbers $t_k \in [0, 1]$, $k = 1, \dots$.

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:

(R1) $\varphi_m := \varphi_m^{r,\tau} \in \mathcal{D}$ is any element satisfying

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

(R2) 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 2.1 It follows from the definition of the WRGA that the sequence $\{E(G_m)\}$ is a nonincreasing sequence.

We call the WRGA(co) *relaxed* because at the mth step of the algorithm we use a linear combination (convex combination) of the previous approximant G_{m-1} and a new element φ_m. The relaxation parameter λ_m in the WRGA(co) is chosen at the mth step depending on E. We use in algorithms for convex optimization the notation G_m (G comes from Greedy) used in greedy approximation algorithms to stress that an mth approximant G_m is obtained by a greedy algorithm. Standard optimization theory notation for it is x^m. In the case $E(x) = E(f, x, q) := \|f - x\|^q$, $f \in X$, $q \geq 1$, the WRGA(co) coincides with the Weak Relaxed Greedy Algorithm from approximation theory (see [40], S. 6.3).

We proceed to a theorem on convergence of the WRGA(co). 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 2.2 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

$$\rho(E, u) = \theta t_m u. \tag{10}$$

Remark 2.3 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 (10) 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 2.4 *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 that 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

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

The following theorem gives the upper bound on the rate of convergence. In the case of the strong version of the algorithm ($t_k = 1$), the corresponding rates of convergence were obtained in [14].

Theorem 2.5 *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 $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \leq 1$, $k = 1, 2, \ldots$, we have for any $f \in A_1(\mathcal{D})$ that*

$$E(G_m) - E(f) \leq \left(C_1(q,\gamma) + C_2(q,\gamma) \sum_{k=1}^{m} t_k^p \right)^{1-q}, \quad p := \frac{q}{q-1},$$

with positive constants $C_1(q,\gamma)$, $C_2(q,\gamma)$ which may depend only on q and γ.

3 The Weak Chebyshev Greedy Algorithm

In this section, we study the WCGA designed for convex optimization. Here we are interested in finding a solution to the unconstrained convex optimization problem (2) that is sparse with respect to a given dictionary $\mathcal{D}$. This algorithm provides a sequence $\{G_m\}$ of sparse approximants such that under mild conditions on E, the sequence $\{E(G_m)\}$ converges to the minimal value of E. Moreover, if we know that the point of minimum of E satisfies some conditions, then we guarantee the rate of convergence of $\{E(G_m)\}$. We prove the results for general infinite dimensional Banach space X. In case X is finite dimensional, say, of dimension N, the complexity of the mth iteration will increase to the complexity of the original optimization problem when $m = N$. This means that it makes sense to use the WCGA(co) with the number of iterations much smaller than N.

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:

(C1) $\varphi_m := \varphi_m^{c,\tau} \in \mathcal{D}$ is any element satisfying

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

(C2) 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 Φ_m at which E attains the minimum:

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

In the case $E(x) = E(f, x, q) := \|f - x\|^q$, $f \in X$, $q \geq 1$, the WCGA(co) coincides with the Weak Chebyshev Greedy Algorithm from approximation theory (see [40], S. 6.2).

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 that for any $\theta \in (0, \theta_0]$, we have*

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

Then

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

Here are two simple corollaries of Theorem 3.1.

Corollary 3.2 *Let a convex function E have modulus of smoothness $\rho(E, u)$ of power type $1 < q \leq 2$, that is, $\rho(E, u) \leq \gamma u^q$. Assume that*

$$\sum_{m=1}^\infty t_m^p = \infty, \quad p = \frac{q}{q-1}. \tag{11}$$

Then

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

Corollary 3.3 *Let a convex function E be uniformly smooth. Assume that $t_k = t$, $k = 1, 2, \dots$. Then*

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

We now proceed to the rate of convergence results.

Theorem 3.4 *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^ϵ from D such that*

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

with some number $B \geq 1$. Then we have for the WCGA(co) ($p := q/(q-1)$),

$$E(G_m) - \inf_{x\in D} E(x) \leq \max\left(2\epsilon, C(q,\gamma)B^q\left(C(E,q,\gamma) + \sum_{k=1}^{m} t_k^p\right)^{1-q}\right). \quad (12)$$

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 mth element φ_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 Sect. 2, the relaxation parameter λ_m in the WRGA(co) is chosen at the mth step depending on E. The following modification of the above idea of relaxation in greedy approximation will be studied in this section [38].

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:

(F1) $\varphi_m \in \mathcal{D}$ is any element satisfying

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

(F2) Find w_m and λ_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 nonincreasing sequence.

In the case $E(x) = E(f, x, q) := \|f - x\|^q$, $f \in X$, $q \geq 1$, the WGAFR(co) coincides with the Weak Greedy Algorithm with Free Relaxation from approximation theory (see [40], S. 6.4).

We now formulate 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 4.2 *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*

$$\sum_{m=1}^{\infty} t_m \xi_m(\rho, \tau, \theta) = \infty. \tag{13}$$

Then, for the WGAFR(co), we have

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

The following theorem gives the rate of convergence.

Theorem 4.3 *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^ϵ from D such that*

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

with some number $B \geq 1$. Then we have ($p := q/(q-1)$) for the WGAFR(co)

$$E(G_m) - \inf_{x\in D} E(x) \leq \max\left(2\epsilon, C_1(E, q, \gamma)B^q \left(C_2(E, q, \gamma) + \sum_{k=1}^{m} t_k^p\right)^{1-q}\right). \tag{14}$$

5 Discussion

The technique used in this paper is a modification of the corresponding technique developed in approximation theory (see [36,39] and the book [40]). We now discuss this in more detail. In nonlinear approximation, we use greedy algorithms, for instance

WCGA and WGAFR, for solving a sparse approximation 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 $F_{f_{m-1}}$. For a nonzero element $f \in X$, we let F_f denote a norming (peak) functional for f that is a functional with the following properties:

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

The existence of such a functional is guaranteed by the Hahn-Banach theorem. The norming functional F_f is a linear functional (in other words is an element of the dual to X space X^*) which can be explicitly written in some cases. In a Hilbert space, F_f can be identified with $f\|f\|^{-1}$. In the real L_p, $1 < p < \infty$, it can be identified with $f|f|^{p-2}\|f\|_p^{1-p}$. The following proposition is well known (see, [40], p. 336).

Proposition 5.1 *Let X be a uniformly smooth Banach space. Then, for any $x \neq 0$ and y, we have*

$$F_x(y) = \left(\frac{d}{du}\|x + uy\|\right)(0) = \lim_{u \to 0}(\|x + uy\| - \|x\|)/u. \tag{15}$$

Proposition 5.1 says that 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 [10] 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. We note that from the definition of modulus of smoothness, we get the following inequality:

$$0 \leq \|x + uy\| - \|x\| - uF_x(y) \leq 2\|x\|\rho(u\|y\|/\|x\|). \tag{16}$$

In the proofs of approximation theory results, we use inequality (16) and the trivial inequality

$$\|x + uy\| \geq F_x(x + uy) = \|x\| + uF_x(y). \tag{17}$$

In the proofs of optimization theory results, we use Lemma 6.3 instead of inequality (16) and the convexity inequality (8) instead of (17). The rest of the proofs use the same technique of solving the corresponding recurrent inequalities.

We stress that an important contribution of this paper is the fact that it provides convergence and rate of convergence results for an arbitrary dictionary. The authors of the paper [17] make the following comment on the importance of a step from the standard coordinate system basis to a special redundant dictionary: "In this paper, we overcome this conceptual obstacle by considering *all possible* (normalized) rank-one matrices as coordinates. This set of matrices forms an overcomplete and uncountable infinite basis of the space of matrices. We show that a simple strategy of performing a coordinate descent on this lifted space actually converges to the right solution."

Our smoothness assumption on E was used in the proofs of all theorems from Sects. 2–4 in the form of Lemma 6.3. 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

$$E(x+uy)-E(x)-u\langle E'(x),y\rangle \le 2\rho(E,u\|y\|),\quad x\in D. \tag{18}$$

Moreover, in Sect. 2, where we consider the WRGA(co), the approximants G_m are forced to stay in the convex hull $A_1(\mathcal{D})$. Therefore, in Theorems 2.4 and 2.5, we can use the following inequality instead of (18):

$$E(x+u(y-x))-E(x)-u\langle E'(x),y-x\rangle \le 2\rho(E,u\|y-x\|), \tag{19}$$

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

We note that smoothness assumptions in the form of (19) with $\rho(E,u\|y-x\|)$ replaced by $C\|y-x\|^q$ were used in many papers [12,14,18,22,45]. For instance, the authors of [45] studied the version of WRGA(co) with weakness sequence $t_k=1$, $k=1,2,\dots$. They proved Theorem 2.5 in this case. Their proof, like our proof in Sect. 2, is very close to the corresponding proof from greedy approximation (see [36,39] Sect. 3.3 or [40] Sect. 6.3).

We now make some general remarks on the results of this paper. A typical problem of convex optimization is to find an approximate solution to the problem

$$w:=\inf_x E(x). \tag{20}$$

In this paper, we study sparse (with respect to a given dictionary $\mathcal{D}$) solutions of (20). This means that we are solving the following problem instead of (20). For a given dictionary $\mathcal{D}$, consider the set of all m-term linear combinations with respect to $\mathcal{D}$:

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

We solve the following *sparse optimization problem*:

$$w_m:=\inf_{x\in\Sigma_m(\mathcal{D})} E(x). \tag{21}$$

In this paper, we have used greedy-type algorithms to approximately solve problem (21). Results of the paper show that it turns out that greedy-type algorithms with respect to $\mathcal{D}$ solve problem (20) 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 6.2). However, at the greedy step of our algorithms, we choose

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

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

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 even be 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 a Banach space X. If instead of a 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 the Banach space X. Then our assumption that the set $D := \{x : E(x) \leq E(0)\}$ is bounded is satisfied. However, this set is not necessarily bounded in the norm generated by $A_1(\mathcal{D})$.

All three greedy-type algorithms studied in this paper use the derivative E' of the objective function E. The derivative E' is a linear functional on X, or, in other words, is an element of the dual space X^*. By analogy with approximation theory, we call these algorithms *dual greedy algorithms*. An important feature of the dual greedy algorithms is that they can be modified into a weak form. We obtained our results for any weakness sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $t_k \in [0, 1]$, $k = 1, 2, \ldots$. In particular, some of the t_k could be equal to zero. In the case $t_k = 0$, we can choose any element $\varphi_k \in \mathcal{D}$ at the greedy step. This allows us to shape our approximant at the kth iteration by other criteria than greedy selection. However, our results guarantee convergence and rate of convergence for τ satisfying the corresponding conditions.

On the example of three greedy-type algorithms, we demonstrated how the technique developed in greedy approximation theory can be modified for finding sparse solutions of convex optimization problems. We discussed in this paper only three greedy-type algorithms. There are many other greedy-type algorithms studied in approximation theory. Some of them—the ones providing expansions—are generalized for the convex optimization problem in [42]. There is an important class of greedy-type algorithms, namely the thresholding-type algorithms, that was not generalized for the convex optimization problem.

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 the 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

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

Algorithms that use the first derivative of the objective function E are called *first-order* optimization algorithms.

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

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

Algorithms that only use the values of the objective function E are called *zero-order* optimization algorithms.

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 [42] and plan to study them in our future work.

After finding $\varphi_m \in \mathcal{D}$, we can proceed in different ways. We now list some typical steps that are motivated by the corresponding steps in greedy approximation theory [40]. These steps or their variants are used in different optimization algorithms (see, for instance, [6,7,21,26,31,33]).

(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) Shortened 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 *shortened.*

(C) Chebyshev-type (fully corrective) 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

$$G_m := G_{m-1} + c_m\varphi_m. \tag{22}$$

In the case of greedy step **II**, we define G_m by formula (22) with the greedy step **II** modified as follows: $\varphi_m \in \mathcal{D}$ is an element satisfying

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

All algorithms studied in this paper fall into the category **I** of algorithms with the gradient-type greedy step. The reader can find a study of algorithms with the E-greedy-type step in [16] and [42]. The step (C) corresponds to the Weak Chebyshev Greedy Algorithm from Sect. 3, and step (F) corresponds to the Weak Greedy Algorithm with Free Relaxation from Sect. 4. The reader can find a detailed study of algorithms with step (G) in [42].

6 Proofs

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

Lemma 6.1 *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 $\phi \in L$.

Lemma 6.2 *For any bounded linear functional F and any dictionary $\mathcal{D}$, we have*

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

See [40], p. 343 for the proof.

We will often use the following simple lemma.

Lemma 6.3 *Let E be Fréchet differentiable convex function. Then the following inequality holds for $x \in D$:*

$$0 \le E(x+uy) - E(x) - u\langle E'(x), y\rangle \le 2\rho(E, u\|y\|). \tag{23}$$

Proof The left inequality follows directly from (8). Next, from the definition of modulus of smoothness, it follows that

$$E(x+uy) + E(x-uy) \le 2(E(x) + \rho(E, u\|y\|)). \tag{24}$$

Inequality (8) gives

$$E(x-uy) \ge E(x) + \langle E'(x), -uy\rangle = E(x) - u\langle E'(x), y\rangle. \tag{25}$$

Combining (24) and (25), we obtain

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

This proves the second inequality. □

We begin with the proofs of Theorems 3.1 and 3.4 for the WCGA(co) because these proofs are the simplest ones. Then we present proofs for results on the WRGA(co) and the WGAFR(co). Some parts of these proofs are similar to the corresponding parts of proofs of Theorems 3.1 and 3.4. We will not duplicate these parts and refer the reader to the proofs of Theorems 3.1 and 3.4.

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

Lemma 6.4 *Let E be a uniformly smooth convex function with modulus of smoothness $\rho(E, u)$. Take a number $\epsilon \ge 0$ and an element f^ϵ from D such that*

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

with some number $B \ge 1$. Then we have for the WCGA(co),

$$\begin{aligned}E(G_m) - E(f^\epsilon) &\le E(G_{m-1}) - E(f^\epsilon)\\ &\quad + \inf_{\lambda\ge 0}(-\lambda t_m B^{-1}(E(G_{m-1}) - E(f^\epsilon)) + 2\rho(E,\lambda))\end{aligned}$$

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

Proof The main idea of the proof is the same as in the proof of the corresponding one-step improvement inequality for the WCGA (see, for instance, [40], p. 343–344). 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 6.4 is trivial. Assume $E(G_{m-1}) - E(f^\epsilon) > 0$. By Lemma 6.3, we have for any λ,

$$E(G_{m-1} + \lambda\varphi_m) \leq E(G_{m-1}) - \lambda\langle -E'(G_{m-1}), \varphi_m\rangle + 2\rho(E, \lambda), \tag{26}$$

and by (C1) from the definition of the WCGA(co) and Lemma 6.2, we get

$$\begin{aligned}\langle -E'(G_{m-1}), \varphi_m\rangle &\geq t_m \sup_{g\in\mathcal{D}}\langle -E'(G_{m-1}), g\rangle \\ &= t_m \sup_{\phi\in A_1(\mathcal{D})}\langle -E'(G_{m-1}), \phi\rangle \geq t_m B^{-1}\langle -E'(G_{m-1}), f^\epsilon\rangle.\end{aligned}$$

By Lemma 6.1 with $x_L = G_{m-1}$ and by convexity (9), 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,

$$\begin{aligned}E(G_m) &\leq \inf_{\lambda\geq 0} E(G_{m-1} + \lambda\varphi_m) \\ &\leq E(G_{m-1}) + \inf_{\lambda\geq 0}(-\lambda t_m B^{-1}(E(G_{m-1}) - E(f^\epsilon)) + 2\rho(E, \lambda),\end{aligned} \tag{27}$$

which proves the lemma. □

Proof of Theorem 3.1 The definition of the WCGA(co) implies that $\{E(G_m)\}$ is a nonincreasing sequence. Therefore we have

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

Define

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

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^ϵ such that

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

with some $B \geq 1$. Then, by Lemma 6.4, we get

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

Let us specify $\theta := \min\left(\theta_0, \frac{\alpha}{8B}\right)$ and take $\lambda = \xi_m(\rho, \tau, \theta)$. 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. □

Proof of Theorem 3.4 Define

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

The sequence $\{a_n\}$ is nonincreasing. If $a_n \leq 0$ for some $n \leq m$, then $E(G_m) - E(f^\epsilon) \leq 0$ and $E(G_m) - \inf_{x \in D} E(x) \leq \epsilon$, which implies (12). Thus we assume that $a_n > 0$ for $n \leq m$.

By Lemma 6.4, we have

$$a_m \leq a_{m-1} + \inf_{\lambda \geq 0}\left(-\frac{\lambda t_m a_{m-1}}{B} + 2\gamma\lambda^q\right). \tag{28}$$

Choose λ from the equation

$$\frac{\lambda t_m a_{m-1}}{B} = 4\gamma\lambda^q,$$

which implies that

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

Let

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

Using the notation $p := \frac{q}{q-1}$, we get from (28),

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

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

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

□

We now need a simple known lemma [35].

Lemma 6.5 *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 6.5 with $y_k := a_k^{\frac{1}{q-1}}$, $n = 0$, $w_k = t_m^p/(A_q B^p)$, we get

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

which implies

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

Theorem 3.4 is now proved.

Proof of Theorems 2.4 and 2.5 This proof is similar to the Proof of Theorems 3.1 and 3.4. Instead of Lemma 6.4, we use the following lemma. □

Lemma 6.6 *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 for the WRGA(co),*

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

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 (26), we have for any λ,

$$\begin{aligned} &E(G_{m-1} + \lambda(\varphi_m - G_{m-1})) \\ &\quad \le E(G_{m-1}) - \lambda\langle -E'(G_{m-1}), \varphi_m - G_{m-1}\rangle + 2\rho(E, 2\lambda), \end{aligned} \tag{29}$$

and by (R1) from the definition of the WRGA(co) and Lemma 6.2, we get

$$\begin{aligned} \langle -E'(G_{m-1}), \varphi_m - G_{m-1}\rangle &\ge t_m \sup_{g\in\mathcal{D}} \langle -E'(G_{m-1}), g - G_{m-1}\rangle \\ = t_m \sup_{\phi\in A_1(\mathcal{D})} \langle -E'(G_{m-1}), \phi - G_{m-1}\rangle &\ge t_m\langle -E'(G_{m-1}), f - G_{m-1}\rangle. \end{aligned}$$

By (9), we obtain

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

Thus,

$$\begin{aligned} E(G_m) &\le \inf_{0\le\lambda\le 1} E(G_{m-1} + \lambda(\varphi_m - G_{m-1})) \\ &\le E(G_{m-1}) + \inf_{0\le\lambda\le 1} (-\lambda t_m(E(G_{m-1}) - E(f)) + 2\rho(E, 2\lambda), \end{aligned} \tag{30}$$

which proves the lemma. □

The remaining part of the proof uses the inequality (30) in the same way relation (27) was used in the Proof of Theorems 3.1 and 3.4. The only additional difficulty here is that we are optimizing over $0 \le \lambda \le 1$. In the Proof of Theorem 2.4, we choose $\theta = \alpha/8$, assuming that α 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 2.5. Define

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

The sequence $\{a_n\}$ is nonincreasing. If $a_n \le 0$ for some $n \le m$, then $E(G_m) - E(f) \le 0$, which implies Theorem 2.5. Thus we assume that $a_n > 0$ for $n \le m$. We obtain from Lemma 6.6,

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

We choose λ from the equation

$$\lambda t_m a_{m-1} = 4\gamma(2\lambda)^q \tag{31}$$

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 the first n steps and then choose λ from (31). Then we get for $k \le n$,

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

and

$$a_n \le a_0 \prod_{k=1}^{n}(1 - t_k/2). \tag{32}$$

For $k > n$, we have

$$a_k \le a_{k-1}(1 - \lambda t_k/2), \quad \lambda = \left(\frac{t_m a_{m-1}}{2^{2+q}\gamma}\right)^{\frac{1}{q-1}}. \tag{33}$$

As in the Proof of Theorem 3.4, we obtain, using Lemma 6.5,

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

By (32), 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.

Proof of Theorems 4.2 and 4.3 We begin with an analog of Lemma 6.4. □

Lemma 6.7 *Let E be a uniformly smooth convex function with modulus of smoothness $\rho(E,u)$. Take a number $\epsilon \ge 0$ and an element f^ϵ from D such that*

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

with some number $B \ge 1$. Then we have for the WGAFR(co),

$$\begin{aligned} &E(G_m) - E(f^\epsilon) \le E(G_{m-1}) - E(f^\epsilon) \\ &\quad + \inf_{\lambda\ge 0}(-\lambda t_m B^{-1}(E(G_{m-1}) - E(f^\epsilon)) + 2\rho(E, C_0\lambda)) \end{aligned}$$

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

Proof By the definition of G_m,

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

As in the arguments in the Proof of Lemma 6.4, we use Lemma 6.3

$$\begin{aligned} &E(G_{m-1} + \lambda\varphi_m - wG_{m-1}) \le E(G_{m-1}) \\ &\quad -\lambda\langle -E'(G_{m-1}), \varphi_m\rangle - w\langle E'(G_{m-1}), G_{m-1}\rangle + 2\rho(E, \|\lambda\varphi_m - wG_{m-1}\|) \end{aligned} \tag{34}$$

and estimate

$$\begin{aligned} \langle -E'(G_{m-1}), \varphi_m\rangle &\ge t_m \sup_{g\in\mathcal{D}} \langle -E'(G_{m-1}), g\rangle \\ &= t_m \sup_{\phi\in A_1(\mathcal{D})} \langle -E'(G_{m-1}), \phi\rangle \ge t_m B^{-1}\langle -E'(G_{m-1}), f^\epsilon\rangle. \end{aligned}$$

We set $w^* := \lambda t_m B^{-1}$ and obtain

$$\begin{aligned} &E(G_{m-1} - w^* G_{m-1} + \lambda\varphi_m) \\ &\quad \le E(G_{m-1}) - \lambda t_m B^{-1}\langle -E'(G_{m-1}), f^\epsilon - G_{m-1}\rangle. \end{aligned} \tag{35}$$

By (9), we obtain

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

Thus,

$$\begin{aligned} E(G_m) &\le E(G_{m-1}) + \inf_{\lambda\ge 0}(-\lambda t_m B^{-1}(E(G_{m-1}) - E(f^\epsilon)) \\ &\quad + 2\rho(E, \|\lambda\varphi_m - w^* G_{m-1}\|). \end{aligned} \tag{36}$$

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 $\|G_{m-1}\| \leq C_1 := \operatorname{diam}(D)$. Thus, under assumption $B \geq 1$, 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. □

By Remark 4.1, $\{E(G_m)\}$ is a nonincreasing sequence. Therefore we have

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

Define

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

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^ϵ such that

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

with some $B \geq 1$. Then, by Lemma 6.7, we get

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

Let us specify $\theta := \min\left(\theta_0, \frac{\alpha}{8B}\right)$ 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 Theorem 4.2.

We proceed to the Proof of Theorem 4.3. Define

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

By Lemma 6.7, we have

$$a_m \leq a_{m-1} + \inf_{\lambda \geq 0} \left(-\frac{\lambda t_m a_{m-1}}{B} + 2\gamma (C_0 \lambda)^q \right). \tag{37}$$

Choose λ from the equation

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

The rest of the proof repeats the argument from the Proof of Theorem 3.4.

Acknowledgments 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. The author is also very grateful to the referees and the editor whose detailed comments helped to substantially improve the presentation of the results.

References

1. Bahmani, S., Raj, B., Boufounos, P.: Greedy sparsity-constrained optimization. arXiv:1203.5483v3 [stat.ML]. 6 Jan 2013
2. Barron, A.R.: Universal approximation bounds for superposition of n sigmoidal functions. IEEE Trans. Inf. Theory **39**, 930–945 (1993)
3. Beck, A., Eldar, Y.C.: Sparsity constrained nonlinear optimization: optimality conditions and algorithms. arXiv:1203.4580v1 [cs.IT]. 20 March 2012
4. Blumensath, T.: Compressed sensing with nonlinear observations. Preprint 1–9 (2010)
5. Blumensath, T.: Compressed sensing with nonlinear observations and related nonlinear optimization problems. arXiv:1205.1650v1 [cs.IT]. 8 May 2012
6. Blumensath, T., Davies, M.E.: Gradient pursuits. IEEE Trans. Signal Process. **56**, 2370–2382 (2008)
7. Blumensath, T., Davies, M.E.: Stagewise weak gradient pursuits. IEEE Trans. Signal Process. **57**, 4333–4346 (2009)
8. Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. **27**, 265–274 (2009)
9. Borwein, J.M., Lewis, A.S.: Convex Analysis and Nonlinear Optimization: Theory and Examples. Canadian Mathematical Society, Springer, Berlin (2006)
10. Borwein, J., Guirao, A.J., Hajek, P., Vanderwerff, J.: Uniformly convex functions an Banach spaces. Proc. Am. Math. Soc. **137**(3), 1081–1091 (2009)
11. Chandrasekaran, V., Recht, B., Parrilo, P.A., Willsky, A.S.: The convex geometry of linear inverse problems. In: Proceedings of the 48th Annual Allerton Conference on Communication, Control and Computing, pp. 699–703 (2010)

12. Clarkson, K.L.: Coresets, sparse greedy approximation, and the Frank-Wolfe algorithm. ACM Trans. Algorithms **6**, 63 (2010)
13. Dai, W., Milenkovic, O.: Subspace pursuit for compressive sensing signal reconstruction. IEEE Trans. Inf. Theory **55**, 2230–2249 (2009)
14. Demyanov, V.F., Rubinov, A.M.: Approximate Methods in Optimization Problems. Elsevier, Amsterdam (1970)
15. DeVore, R.A.: Nonlinear approximation. Acta Numer. **7**, 51–150 (1998)
16. DeVore, R.A., Temlyakov, V.N.: Convex Optimization on Banach Spaces. arXiv:1401.0334v1 [stat.ML]. 1 Jan 2014
17. Dudik, M., Harchaoui, Z., Malick, J.: Lifted coordinate descent for learning with trace-norm regularization. In: AISTATS (2012)
18. Dunn, J.C., Harshbarger, S.: Conditional gradient algorithms with open loop step size rules. J. Math. Anal. Appl. **62**, 432–444 (1978)
19. Elad, M., Matalon, B., Zibulevsky, M.: Coordinate and subspace optimization methods for linear least squares with non-quadratic regularization. Appl. Comput. Harmonic Anal. **23**, 346–367 (2007)
20. Figueiredo, M.A.T., Nowak, R.D., Wright, S.J.: Gradient projection for sparse reconstruction: application to compressed sensing and other inverse problems. IEEE Sel. Top. Signal Process. **1**, 586–597 (2007)
21. Frank, M., Wolfe, P.: An algorithm for quadratic programming. Nav. Res. Logist. Q. **3**, 95–110 (1956)
22. Jaggi, M.: Sparse Convex Optimization Methods for Machine Learning, PhD thesis, ETH Zürich, (2011)
23. Jaggi, M.: Revisiting Frank-Wolfe: projection-free sparse convex optimization. In: Proceedings of the 30th International Conference on Machine Learning, Atlanta, Georgia (2013)
24. Jaggi, M., Sulovský, M.: A Simple Algorithm for Nuclear Norm Regularized Problems. ICML (2010)
25. Jones, L.K.: A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training. Ann. Statist. **20**, 608–613 (1992)
26. Karmanov, V.G.: Mathematical Programming. Mir Publishers, Moscow (1989)
27. Kashin, B.S.: Widths of certain finite-dimensional sets and classes of smooth functions. Izv. Acad. Nauk SSSR Ser. Mat **41**, 334–351 (1977)
28. Livshitz, E., Temlyakov, V.: Sparse Approximation and Recovery by Greedy Algorithms. arXiv:1303.3595v1 [math.NA]. 14 March 2013
29. Needell, D., Tropp, J.A.: CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmonic Anal. **26**, 301–321 (2009)
30. Needell, D., Vershynin, R.: Uniform uncertainty principle and signal recovery via orthogonal matching pursuit. Found. Comp. Math. **9**, 317–334 (2009)
31. Nesterov, Yu.: Introductory Lectures on Convex Optimization: A Basic Course. Kluwer Academic Publishers, Boston (2004)
32. Nguyen, H., Petrova, G.: Greedy Strategies for Convex Optimization. arXiv:1401.1754v1 [math.NA]. 8 Jan 2014
33. Pshenichnyi, B.N., Danilin, YuM: Numerical Methods in Extremal Problems [in Russian]. Nauka, Moscow (1975)
34. Shalev-Shwartz, S., Srebro, N., Zhang, T.: Trading accuracy for sparsity in optimization problems with sparsity constrains. SIAM J. Optim. **20**(6), 2807–2832 (2010)
35. Temlyakov, V.N.: Weak greedy algorithms. Adv. Comput. Math. **12**, 213–227 (2000)
36. Temlyakov, V.N.: Greedy algorithms in Banach spaces. Adv. Comput. Math. **14**, 277–292 (2001)
37. Temlyakov, V.N.: Greedy-type approximation in Banach spaces and applications. Constr. Approx. **21**, 257–292 (2005)
38. Temlyakov, V.N.: Relaxation in greedy approximation. Constr. Approx. **28**, 1–25 (2008)
39. Temlyakov, V.N.: Greedy approximation. Acta Numer. **17**, 235–409 (2008)
40. Temlyakov, V.N.: Greedy Approximation. Cambridge University Press, Cambridge (2011)
41. Temlyakov, V.N.: Greedy Approximation in Convex Optimization. arXiv:1206.0392v1 [stat.ML]. 2 June 2012 (see also IMI Preprint, 2012:03, 1–25).
42. Temlyakov, V.N.: Greedy expansions in convex optimization. In: Proceedings of the Steklov Institute of Mathematics, vol. 284, pp. 244–262, (2014) (see also arXiv:1206.0393v1 [stat.ML]. 2 Jun 2012)
43. Temlyakov, V.N.: Sparse Approximation and Recovery by Greedy Algorithms in Banach Spaces. arXiv:1303.6811v1 [stat.ML]. 27 March 2013

44. Temlyakov, V.N.: Chebyshev Greedy Algorithm in Convex Optimization. arXiv:1312.1244v1 [stat.ML]. 4 Dec 2013
45. Tewari, A., Ravikumar, P., Dhillon, I.S.: Greedy Algorithms for Structurally Constrained High Dimensional Problems, prerint 1–10 (2012)
46. Zhang, T.: Sequential greedy approximation for certain convex optimization problems. IEEE Trans. Inf. Theory **49**(3), 682–691 (2003)
47. Zhang, T.: Adaptive forward–backward greedy algorithm for learning sparse representations. IEEE Trans. Inf. Theory **57**, 4689–4708 (2011)
48. Zhang, T.: Sparse recovery with orthogonal matching pursuit under RIP. IEEE Trans. Inf. Theory **57**, 6215–6221 (2011)