Chebushev Greedy Algorithm in convex optimization

V.N. Temlyakov *

December 5, 2013

Abstract

Chebyshev Greedy Algorithm is a generalization of the well known Orthogonal Matching Pursuit defined in a Hilbert space to the case of Banach spaces. We apply this algorithm for constructing sparse approximate solutions (with respect to a given dictionary) to convex optimization problems. Rate of convergence results in a style of the Lebesgue-type inequalities are proved.

1 Introduction

We study sparse approximate solutions to convex optimization problems. We apply the technique developed in nonlinear approximation known under the name of greedy approximation. A typical problem of convex optimization is to find an approximate solution to the problem

$$\inf_x E(x)$$

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 [1], [3]). Recent needs of numerical analysis call for consideration of the above optimization problem on an infinite dimensional space, for instance, a space of

---

*University of South Carolina and Steklov Institute of Mathematics. Research was supported by NSF grant DMS-1160841
continuous functions. Thus, we consider a convex function \( E \) defined on a Banach space \( X \). This paper is a follow up to papers [6], [7], and [4]. We refer the reader to the above mentioned papers for a detailed discussion and justification of importance of greedy methods in optimization problems.

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 (\( \| g \| \leq 1 \)),

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

and the closure of \( \text{span} \mathcal{D} \) is \( X \). For notational convenience in this paper symmetric dictionaries are considered. Results of the paper also hold for non-symmetric dictionaries with straightforward modifications. 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 \( \text{conv}(\mathcal{D}) \). We use this notation because it has become a standard notation in relevant greedy approximation literature.

We assume that \( E \) is Fréchet differentiable and that the set

\[
D := \{ x : E(x) \leq E(0) \}
\]

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

\[
\rho(E, u) := \frac{1}{2} \sup_{x \in \mathcal{D}, \|y\| = 1} |E(x + uy) + E(x - uy) - 2E(x)|. \tag{1.2}
\]

We say that \( E \) is uniformly smooth if \( \rho(E, u) = o(u), \ u \to 0 \).

We defined and studied in [6] the following generalization of the Weak Chebyshev Greedy Algorithm (see [5], Ch. 6) for convex optimization.

**Weak Chebyshev Greedy Algorithm (WCGA(co)).** Let \( \tau := \{ t_k \}_{k=1}^{\infty}, \ t_k \in (0, 1], \ k = 1, 2, \ldots, \) be a weakness sequence. We define \( G_0 := 0 \). Then for each \( m \geq 1 \) we have the following inductive definition.

1. \( \varphi_m := \varphi_m^{\tau} \in \mathcal{D} \) 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

\[
\Phi_m := \Phi_m^{\tau} := \text{span}\{\varphi_j\}_{j=1}^{m},
\]
and define $G_m := G^c_{m,\tau}$ to be the point from $\Phi_m$ at which $E$ attains the minimum:

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

We consider here along with the WCGA(co) the following greedy algorithm.

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

1. $\varphi_m := \varphi^E_{m,\tau} \in \mathcal{D}$ is any element satisfying (assume existence)

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

2. Define $\Phi_m := \Phi_{m,\tau} := \text{span}\{\varphi_j\}_{j=1}^m$,

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

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

The EGCA(co) is in a style of $X$-Greedy algorithms studied in approximation theory (see [5], Ch. 6). In a special case of $X = \mathbb{R}^d$ and $\mathcal{D}$ is a canonical basis of $\mathbb{R}^d$ the EGCA(co) was introduced and studied in [4]. Convergence and rate of convergence of the WCGA(co) were studied in [6]. For instance, the following rate of convergence theorem was proved in [6].

**Theorem 1.1.** 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 $f^c$ from $\mathcal{D}$ such that

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

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

$$E(G_m) - \inf_{x \in \mathcal{D}} E(x) \leq \max \left( 2\epsilon, C(q, \gamma)B^q \left( C(E, q, \gamma) + \sum_{k=1}^m t_k^p \right)^{1-q} \right). \quad (1.3)$$

We will use the following notations. Let $f_0$ be a point of minimum of $E$:

$$E(f_0) = \inf_{x \in \mathcal{D}} E(x).$$
We denote for $m = 1, 2, \ldots$

$$f_m := f_0 - G_m.$$  

In particular, if the point of minimum $f_0$ belongs to $A_1(D)$, then Theorem 1.1 in the case $t_k = t \in (0, 1)$, $k = 1, \ldots$, with $\epsilon = 0$, $B = 1$, gives

$$E(G_m) - E(f_0) \leq C(q, \gamma, t)m^{1-q}.$$  \hspace{1cm} (1.4)

Inequality (1.4) uses only information that $f_0 \in A_1(D)$. Theorem 1.1 is designed in a way that the convergence rate is determined by smoothness of $E$ and complexity of $f_0$. Our way of measuring complexity of the element $f_0$ in Theorem 1.1 is based on $A_1(D)$. Given a dictionary $D$ we say that $f_0$ is simple with respect to $D$ if $f_0 \in A_1(D)$. Next, let for every $\epsilon > 0$ an element $f^\epsilon$ be such that

$$E(f^\epsilon) \leq E(f_0) + \epsilon, \quad f^\epsilon / A(\epsilon) \in A_1(D)$$

with some number $A(\epsilon)$ (the smaller the $A(\epsilon)$ the better). Then we say that complexity of $f_0$ is bounded (bounded from above) by the function $A(\epsilon)$.

We apply algorithms which at the $m$th iteration provide an $m$-term polynomial $G_m$ with respect to $D$. The approximant belongs to the domain $D$ of our interest. Then on one hand we always have the lower bound

$$E(G_m) - \inf_{x \in D} E(x) \geq \inf_{x \in D \cap \Sigma_m(D)} E(x) - \inf_{x \in D} E(x)$$

where $\Sigma_m(D)$ is a collection of all $m$-term polynomials with respect to $D$. On the other hand if we know $f_0$ then the best we can do with our algorithms is to get

$$\|f_0 - G_m\| = \sigma_m(f_0, D)$$

where $\sigma_m(f_0, D)$ is the best $m$-term approximation of $f_0$ with respect to $D$. Then we can aim at building algorithms that provide an error $E(G_m) - E(f_0)$ comparable to $\rho(E, \sigma_m(f_0, D))$. It would be in a style of the Lebesgue-type inequalities. However, it is known from greedy approximation theory that there is no Lebesgue-type inequalities which hold for an arbitrary dictionary even in the case of Hilbert spaces. There are the Lebesgue-type inequalities for special dictionaries. We refer the reader to [5], [2], [8], [9] for results on the Lebesgue-type inequalities. In this paper we obtain rate of convergence results for the WCGA(co) in a style of the Lebesgue-type inequalities.
We will use the following assumptions on properties of $E$.

**E1. Smoothness.** We assume that $E$ is a convex function with

$$
\rho(E, u) \leq \gamma u^2.
$$

**E2. Restricted strong convexity.** We assume that for any $S$-sparse element $f$ we have

$$
E(f) - E(f_0) \geq \beta \|f - f_0\|^2.
$$

Here is one assumption on the dictionary $D$ that we will use (see [8]). For notational simplicity we formulate it for a countable dictionary $D = \{g_i\}_{i=1}^{\infty}$.

**A.** We say that $f = \sum_{i \in T} x_i g_i$ has $\ell_1$ incoherence property with parameters $S$, $V$, and $r$ if for any $A \subset T$ and any $\Lambda$ such that $A \cap \Lambda = \emptyset$, $|A| + |\Lambda| \leq S$ we have for any $\{c_i\}

$$
\sum_{i \in A} |x_i| \leq V |A|^r \|f_A - \sum_{i \in \Lambda} c_i g_i\|, \quad f_A := \sum_{i \in A} x_i g_i.
$$

(1.6)

A dictionary $D$ has $\ell_1$ incoherence property with parameters $K$, $S$, $V$, and $r$ if for any $A \subset B$, $|A| \leq K$, $|B| \leq S$ we have for any $\{c_i\}_{i \in B}$

$$
\sum_{i \in A} |c_i| \leq V |A|^r \|\sum_{i \in B} c_i g_i\|.
$$

The following theorem is the main result of the paper.

**Theorem 1.2.** Let $E$ satisfy assumptions **E1** and **E2**. Suppose for a point of minimum $f_0$ we have $\|f_0 - f^*\| \leq \epsilon$ with $K$-sparse $f := f^*$ satisfying property **A**. Then for the WCGA(co) with weakness parameter $t$ we have for $K + m \leq S$

$$
E(G_m) - E(f_0) \leq \max \left( \left( E(0) - E(f_0) \right) \exp \left( -c_1 m \frac{K}{K^2} \right), 8(\gamma^2 / \beta) \epsilon^2 \right) + 2 \gamma \epsilon^2,
$$

where $c_1 := \frac{\beta^2}{64 \gamma^2}.$

Let us apply Theorem 1.2 in a particular case $r = 1/2$. If we assume that $\sigma_K(f_0, D) \leq C_1 K^{-s}$ then for $m$ of order $K \ln K$ Theorem 1.2 with $\epsilon = C_1 K^{-s}$ provides the bound

$$
E(G_m) - E(f_0) \leq C_2 K^{-2s}.
$$

Note that $K^{-2s}$ is of order $\rho(E, K^{-s})$ in our case.

In the case of direct application of the Weak Chebyshev Greedy Algorithm to the element $f_0$ the corresponding results in a style of the Lebesgue-type inequalities are known (see [2] and [8]).
2 Proofs

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

$$E(y) \geq E(x) + \langle E'(x), y - x \rangle$$

(2.1)

or, in other words,

$$E(x) - E(y) \leq \langle E'(x), x - y \rangle = \langle -E'(x), y - x \rangle.$$  

(2.2)

We will often use the following simple lemma (see [6]).

**Lemma 2.1.** Let $E$ be Fréchet differentiable convex function. Then the following inequality holds for $x \in D$

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

(2.3)

The following two simple lemmas are well-known (see [5], Chapter 6 and [6], Section 2).

**Lemma 2.2.** 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 2.3.** 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 of Theorem 1.2. Let

$$f := f^\epsilon = \sum_{i \in T} x_i g_i, \quad g_i \in \mathcal{D}, \quad |T| = K.$$

We examine $n$ iterations of the algorithm for $n = 1, \ldots, m$. Denote by $T^n$ the set of indices of $g_i$ picked by the WCGA(co) after $n$ iterations, $\Gamma^n :=$
$T \setminus T^n$. Denote as above by $A_1(D)$ the closure in $X$ of the convex hull of the symmetric dictionary $D$. We will bound from above $a_n := E(G_n) - E(f^\epsilon)$. Assume $\|f_{n-1}\|^2 \geq 4(\gamma/\beta)\epsilon^2$ for all $n = 1, \ldots, m$. Denote $A_n := \Gamma^{n-1}$ and

$$ f_{A_n} := f^\epsilon_{A_n} := \sum_{i \in A_n} x_i g_i, \quad \|f_{A_n}\|_1 := \sum_{i \in A_n} |x_i|. $$

The following lemma is used in our proof.

**Lemma 2.4.** Let $E$ be a uniformly smooth convex function with modulus of smoothness $\rho(E, u)$. Take a number $\epsilon \geq 0$ and a $K$-sparse element $f^\epsilon = \sum_{i \in T} x_i g_i$ from $D$ such that

$$ \|f_0 - f^\epsilon\| \leq \epsilon. $$

Then we have for the WCGA(co)

$$ E(G_n) - E(f^\epsilon) \leq E(G_{n-1}) - E(f^\epsilon) + \inf_{\lambda \geq 0} \left(-\lambda t \|f_{A_n}\|_1^{-1}(E(G_{n-1}) - E(f^\epsilon)) + 2\rho(E, \lambda)\right), $$

for $n = 1, 2, \ldots$.

**Proof.** It follows from the definition of WCGA(co) that $E(0) \geq E(G_1) \geq E(G_2) \ldots$. Therefore, if $E(G_{n-1}) - E(f^\epsilon) \leq 0$ then the claim of Lemma 2.4 is trivial. Assume $E(G_{n-1}) - E(f^\epsilon) > 0$. By Lemma 2.1 we have for any $\lambda$

$$ E(G_{n-1} + \lambda \varphi_n) \leq E(G_{n-1}) - \lambda(-E'(G_{n-1}), \varphi_n) + 2\rho(E, \lambda) \quad (2.4) $$

and by (1) from the definition of the WCGA(co) and Lemma 2.3 we get

$$ \langle -E'(G_{n-1}), \varphi_n \rangle \geq t \sup_{g \in D} \langle -E'(G_{n-1}), g \rangle = t \sup_{\phi \in A_1(D)} \langle -E'(G_{n-1}), \phi \rangle \geq t \|f_{A_n}\|_1^{-1} \langle -E'(G_{n-1}), f_{A_n} \rangle. $$

By Lemma 2.2 and (2.2) we obtain

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

Thus,

$$ E(G_n) \leq \inf_{\lambda \geq 0} E(G_{n-1} + \lambda \varphi_n) \leq E(G_{n-1}) + \inf_{\lambda \geq 0} \left(-\lambda t \|f_{A_n}\|_1^{-1}(E(G_{n-1}) - E(f^\epsilon)) + 2\rho(E, \lambda)\right), \quad (2.5) $$

which proves the lemma.
Denote
\[ a_n := E(G_n) - E(f^\epsilon). \]

From (2.5) we obtain
\[ a_n \leq a_{n-1} + \inf_{\lambda \geq 0} \left( -\lambda t \frac{a_{n-1}}{\|f_{A_n}\|_1} + 2\rho(E, \lambda) \right). \tag{2.6} \]

By assumption E1 we have \( \rho(E, u) \leq \gamma u^2 \). We get from (2.6)
\[ a_n \leq a_{n-1} + \inf_{\lambda \geq 0} \left( -\lambda t a_{n-1} \frac{1}{\|f_{A_n}\|_1} + 2\gamma \lambda^2 \right). \]

Let \( \lambda_1 \) be a solution of
\[ \frac{\lambda t a_{n-1}}{2\|f_{A_n}\|_1} = 2\gamma \lambda^2, \quad \lambda_1 = \frac{t a_{n-1}}{4\gamma \|f_{A_n}\|_1}. \]

Our assumption A (see (1.6)) gives
\[
\|f_{A_n}\|_1 = \|(f^\epsilon - G_{n-1})_{A_n}\|_1 \leq VK^r \|f^\epsilon - G_{n-1}\|
\leq VK^r (\|f_0 - G_{n-1}\| + \|f_0 - f^\epsilon\|) \leq VK^r (\|f_{n-1}\| + \epsilon). \tag{2.7}
\]

We bound from below \( a_{n-1} = E(G_{n-1}) - E(f^\epsilon) \). By our smoothness assumption and Lemma 2.1
\[ E(f^\epsilon) - E(f_0) \leq 2\gamma \|f^\epsilon - f_0\|^2 \leq 2\gamma \epsilon^2. \]

Therefore,
\[ a_{n-1} = E(G_{n-1}) - E(f^\epsilon) = E(G_{n-1}) - E(f_0) + E(f_0) - E(f^\epsilon)
\geq E(G_{n-1}) - E(f_0) - 2\gamma \epsilon^2. \]

By restricted strong convexity assumption E2
\[ E(G_{n-1}) - E(f_0) \geq \beta \|G_{n-1} - f_0\|^2 = \beta \|f_{n-1}\|^2. \]

Thus
\[ a_{n-1} \geq \beta \|f_{n-1}\|^2 - 2\gamma \epsilon^2. \tag{2.8} \]

Specify
\[ \lambda = \frac{t \beta \|f_{A_n}\|_1}{32\gamma (VK^r)^2}. \]
Then, using (2.7) and (2.8) we get
\[ \frac{\lambda}{\lambda_1} = \frac{\beta\|f_{A_n}\|_1^2}{8(VK^r)^2a_{n-1}} \leq \frac{\beta(\|f_{n-1}\| + \epsilon)^2}{8\beta\|f_{n-1}\|^2 - 2\gamma\epsilon^2}. \] (2.9)

By our assumption \( \|f_{n-1}\|^2 \geq 4(\gamma/\beta)\epsilon^2 \) and a trivial inequality \( \beta \leq 2\gamma \) we obtain from (2.9) that \( \lambda \leq \lambda_1 \) and therefore
\[ a_n \leq a_{n-1} \left(1 - \frac{\beta t^2}{64\gamma(VK^r)^2}\right), \quad n = 1, \ldots, m. \]

Denote \( c_1 := \frac{\beta t^2}{64\gamma V^2r} \). Then
\[ a_m \leq a_0 \exp\left(-\frac{c_1 m}{K^2r}\right). \] (2.10)

We obtained (2.10) under assumption \( \|f_{n-1}\|^2 \geq 4(\gamma/\beta)\epsilon^2, \quad n = 1, \ldots, m. \) If \( \|f_{n-1}\|^2 < 4(\gamma/\beta)\epsilon^2 \) for some \( n \in [1, m] \) then \( a_{m-1} \leq a_{n-1} \leq 2\gamma\|f_{n-1}\|^2 \leq 8(\gamma^2/\beta)\epsilon^2 \). Therefore,
\[ a_m \leq \max\left(a_0 \exp\left(-\frac{c_1 m}{K^2r}\right), 8(\gamma^2/\beta)\epsilon^2\right). \]

Next, we have
\[ E(G_m) - E(f_0) = a_m + E(f^\epsilon) - E(f_0) \leq a_m + 2\gamma\epsilon^2. \]

This completes the proof of Theorem 1.2.

The above technique of studying the WCGA(co) works for the EGCA(co) as well. Instead of Lemma 2.4 we have the following one.

**Lemma 2.5.** Let \( E \) be a uniformly smooth convex function with modulus of smoothness \( \rho(E, u) \). Take a number \( \epsilon \geq 0 \) and a \( K \)-sparse element \( f^\epsilon \) from \( D \) such that
\[ \|f_0 - f^\epsilon\| \leq \epsilon. \]

Then we have for the EGCA(co)
\[ E(G_n) - E(f^\epsilon) \leq E(G_{n-1}) - E(f^\epsilon) \]
\[ + \inf_{\lambda \geq 0} (-\lambda\|f_{A_n}\|^{-1}_1(E(G_{n-1}) - E(f^\epsilon)) + 2\rho(E, \lambda)), \]
for \( n = 1, 2, \ldots \).
Proof. In the proof of Lemma 2.4 we did not use a specific form of the $G_{n-1}$ as the one generated by the $(n-1)$th iteration of the WCGA(co), we only used that $G_{n-1} \in D$. Let $G_{n-1}$ be from the $(n-1)$th iteration of the EGCA(co) and let $\varphi_m^t$, $t \in (0, 1)$, be such that

$$\langle -E'(G_{n-1}), \varphi_m^t \rangle \geq t \sup_{g \in D} \langle -E'(G_{n-1}), g \rangle.$$ 

Then the above proof of Lemma 2.4 gives

$$\inf_{\lambda \geq 0} E(G_{n-1} + \lambda \varphi_m^t) \leq \inf_{\lambda \geq 0} (-\lambda t \| f_A^n \|_1^{-1} (E(G_{n-1}) - E(f^e)) + 2\rho(E, \lambda)). \quad (2.11)$$

Definition of the EGCA(co) implies

$$E(G_m) \leq \inf_{c} E(G_{n-1} + c \varphi_m) \leq \inf_{\lambda \geq 0} E(G_{n-1} + \lambda \varphi_m^t). \quad (2.12)$$

Combining (2.11) and (2.12) and taking into account that $E(G_m)$ does not depend on $t$, we complete the proof of Lemma 2.5.

The following theorem is derived from Lemma 2.5 in the same way as Theorem 1.2 was derived from Lemma 2.4.

**Theorem 2.1.** Let $E$ satisfy assumptions $E1$ and $E2$. Suppose for a point of minimum $f_0$ we have $\| f_0 - f^e \| \leq \epsilon$ with $K$-sparse $f := f^e$ satisfying property $A$. Then for the EGCA(co) we have for $K + m \leq S$

$$E(G_m) - E(f_0) \leq \max \left( (E(0) - E(f_0)) \exp \left( -\frac{c_1 m}{K^2 \epsilon^2} \right), 8\left(\frac{\gamma^2}{\beta}\right) \epsilon^2 \right) + 2\epsilon^2,$$

where $c_1 := \frac{\beta}{64\gamma V^2}$. 

10
References

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

[2] E. Livshitz and V. Temlyakov, Sparse approximation and recovery by greedy algorithms, Preprint, 2013.

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

[4] 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.

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

[6] V.N. Temlyakov, Greedy approximation in convex optimization, arXiv: 1206.0392v1 [stat.ML] 2 Jun 2012 (see also IMI Preprint, 2012:03, 1–25).

[7] V.N. Temlyakov, Greedy expansions in convex optimization, arXiv: 1206.0393v1 [stat.ML] 2 Jun 2012 (see also IMI Preprint, 2012:03, 1–27).

[8] V.N. Temlyakov, Sparse approximation and recovery by greedy algorithms in Banach spaces, arXiv: 1303.6811v1 [stat.ML] 27 Mar 2013.

[9] T. Zhang, Sparse Recovery with Orthogonal Matching Pursuit under RIP, IEEE Transactions on Information Theory, 57 (2011), 6215–6221.