RESCALED PURE GREEDY ALGORITHM FOR CONVEX OPTIMIZATION

ZHEMING GAO, GUERGANA PETROVA

Abstract. We suggest a new greedy strategy for convex optimization in Banach spaces and prove its convergent rates under a suitable behavior of the modulus of uniform smoothness of the objective function. We show that this algorithm is a generalization of the recently discovered Rescaled Pure Greedy Algorithm for approximation in Hilbert spaces.

AMS subject classification: 65K05, 90C25, 41A46.
Key Words: Greedy Algorithms, Convex Optimization, Rates of Convergence.

1. Introduction

The main goal in convex optimization is the development and analysis of algorithms for solving the problem

$$\inf_{x\in\Omega} E(x), \tag{1.1}$$

where E is a given convex function and Ω is a bounded convex subset of a Banach space X. E is called the *objective* function and satisfies the convexity condition

$$E(\gamma x + \delta y) \le \gamma E(x) + \delta E(y), \quad x, y \in \Omega, \quad \gamma, \delta \ge 0, \quad \gamma + \delta = 1.$$

While the classical convex optimization deals with objective functions E defined on subsets Ω in $\mathbb{R}^n$ for moderate values of n, see [2], some of the new applications require that the dimension n is quite large or even infinite. The design of algorithms for such cases is quite challenging since typical convergent results involve n, and therefore deteriorate severely with the growth of n. This is the so-called curse of dimensionality. Recently, there has been an increased interest in developing greedy based strategies for solving (1.1) with provable convergence rate depending only on the properties of E and not on the dimension n of the underlying space. These algorithms provide approximations $\{E(x_m)\}$, $m = 1, 2, \ldots$, to the solution of (1.1), with x_m being a linear combination of m elements from a given dictionary $\mathcal{D} \subset X$. A dictionary is any set $\mathcal{D}$ of norm one elements from X whose span is dense in X. An example of a dictionary is any Shauder basis for X, or a union

This research was supported by the NSF Grant DMS 1521067, and by the DARPA Grant HR0011619523 through Oak Ridge National Laboratory.

of several Shauder bases. The current greedy based algorithms for convex optimization, see [12, 8, 9, 4], are concerned with finding approximations x_m, $m = 0, 1, 2, \ldots$, to the global minimizer $\bar{x}$ of the objective function E. Common features in these algorithms are choosing an initial approximation $x_0 = 0$, selecting the set Ω as

$$\Omega := \{x \in X \ : \ E(x) \leq E(0) = E(x_0)\},$$

(note that $\bar{x} \in \Omega$), and generating the sequence $E_m := E(x_m) \approx E(\bar{x})$, $m = 1, 2, \ldots$, recursively, using the dictionary $\mathcal{D}$. Some methods, such as the Weak Chebychev Greedy Algorithm, see [8], provide at the m-th step an approximant $x_m \approx \bar{x}$, determined as

$$x_m := \text{argmin}_{x \in span\{\varphi_{j_1}, \ldots, \varphi_{j_m}\}} E(x),$$

where $\varphi_{j_1}, \ldots, \varphi_{j_m}$ are suitably chosen elements from $\mathcal{D}$. Others, choose x_m as a solution to

$$x_m := \text{argmin}_{\omega, \lambda \in I\!R} E(\omega x_{m-1} + \lambda \varphi_m),$$

or

$$x_m := \text{argmin}_{\lambda \in [0,1]} E((1-\lambda) x_{m-1} + \lambda \varphi_m),$$

for a suitably chosen element $\varphi_m \in \mathcal{D}$, with $x_{m-1} \approx \bar{x}$ being the approximation to $\bar{x}$, generated at the previous step. Convergence rates for the above mentioned algorithms are proved to be of order $\mathcal{O}(m^{1-q})$, where q is a parameter related to the smoothness of the objective function E. From computational point of view, the latter two approaches are more efficient, since they require solving two or one dimensional optimization problems at each step. Note that some of these approaches can be applied only to objective functions E whose minimum is attained in the convex hull of the dictionary $\mathcal{D}$, since at step m the approximant x_m is derived as a convex combination of x_{m-1} and φ_m.

In this paper, we introduce a new greedy algorithm based on only one dimensional optimization at each step, which does not require the solution of (1.1) to belong to the convex hull of $\mathcal{D}$, and has the optimal convergence rate $\mathcal{O}(m^{1-q})$. The proposed algorithm is similar to the *EGA($\mathcal{C}$)* algorithm from [8] . However, the convergence rate of *EGA($\mathcal{C}$)* is shown to be $\mathcal{O}(m^{-r})$, for any $r \in (0, \frac{q-1}{q+1})$, while we achieve the optimal convergence rate of $\mathcal{O}(m^{1-q})$. Our new algorithm can be viewed as a generalization of the Rescaled Pure Greedy Algorithm (*RPGA*) for approximating functions in Hilbert and Banach spaces, recently introduced in [7], and thus we call it *RPGA(co)*.

This paper is organized as follows. In §2, we list several definitions and known results about convex functions. The *RPGA(co)* is presented in §3, where we prove its convergence rate. In §4, we provide the weak version of the algorithm. We discuss in §5 the fact that the *RPGA(co)*, applied to the objective function E, defined on a Hilbert space H as $E(x) := \|x - \bar{x}\|^2$ is the recently introduced *RPGA* for approximating $\bar{x} \in H$, see [7].

2. Preliminaries

We now introduce some notation and state several well known facts about Banach spaces and convex functions.

2.1. **The Banach space X.** A set of functions $\mathcal{D} := \{\varphi\} \subset X$ is called a dictionary for the Banach space X with norm $\|\cdot\|$, if $\|\varphi\| = 1$ for every element $\varphi \in \mathcal{D}$, and the closure of $span(\mathcal{D})$ is X. Any Shauder basis for X can be viewed as a dictionary for X, but the main idea behind dictionaries is to cover redundant families such as frames. A common example of dictionaries is the union of several Shauder bases.

For a general dictionary $\mathcal{D} \subset X$ and any $M > 0$, we define the set $\mathcal{A}_1^o(\mathcal{D}, M)$ of elements in X as

$$\mathcal{A}_1^o(\mathcal{D}, M) := \{x = \sum_{k\in\Lambda} c_k(x)\varphi_k : \varphi_k \in \mathcal{D},\ |\Lambda| < \infty,\ \sum_{k\in\Lambda} |c_k(x)| \le M\},$$

where $|\Lambda|$ is the cardinality of the index set Λ. If we denote by $\mathcal{A}_1(\mathcal{D}, M)$ the closure of $\mathcal{A}_1^o(\mathcal{D}, M)$ in X, we consider the class $\mathcal{A}_1(\mathcal{D})$, defined as the union of $\mathcal{A}_1(\mathcal{D}, M)$ over all possible $M > 0$. For any $x \in \mathcal{A}_1(\mathcal{D})$, we denote

$$\|x\|_1 := \inf\{M :\ x \in \mathcal{A}_1(\mathcal{D}, M)\}.$$

In what follows, we assume that the minimizer $\bar{x}$ of the objective function E is such that $\bar{x} \in \mathcal{A}_1(\mathcal{D})$. Our algorithm will generate approximants $x_m \in X$ to $\bar{x}$, where each x_m is a sum of at most m terms from the dictionary $\mathcal{D}$.

2.2. **The objective function E.** Let us first recall that a function F is Frechet differentiable at $x \in \Omega$ if there exists a bounded linear functional, denoted by $F'(x) \in X^*$, such that

$$\lim_{h\to 0} \frac{|F(x+h) - F(x) - \langle F'(x), h\rangle|}{\|h\|} = 0.$$

Here, we use the notation $\langle L, h\rangle := L(h)$ to denote the action of the functional $L \in X^*$ on the element $h \in X$.

The following lemmas are well known facts and we simply state them.

Lemma 2.1. *Let $E : X \to \mathbb{R}$ be a Frechet differentiable function at each point in Ω and convex on X. Then, for all $x \in \Omega$ and $x' \in X$,*

$$\langle E'(x), x - x'\rangle \ge E(x) - E(x').$$

Lemma 2.2. *Let E be a Frechet differentiable convex function, defined on a convex domain Ω. Then E has a global minimum at $\bar{x} \in \Omega$ if and only if $E'(\bar{x}) = 0$.*

Lemma 2.3. *Let F be a Frechet differentiable function and φ is a fixed element in X. If $x^* \in X$ is such that*

$$F(x^*) = \min_{t\in\mathbb{R}} F(t\varphi),$$

then, $\langle F'(x^*), x^* \rangle = 0$.

Next, we consider convex functions $E : X \to I\!R$ that attain a global minimum at $\bar{x} \in X$ and satisfy the following two assumptions.

- **Condition 0:** E has Frechet derivative $E'(x) \in X^*$ at each point in $\Omega := \{x \in X \ : \ E(x) \le E(0)\}$, the set Ω is bounded, and
$$\|E'(x)\| \le M_0, \quad x \in \Omega.$$
- **Uniform Smoothness (US):** There are constants $\alpha > 0$, $M > 0$, and $1 < q \le 2$, such that for all x, x' with $\|x - x'\| \le M$, $x \in \Omega$,
$$E(x') - E(x) - \langle E'(x), x' - x \rangle \le \alpha \|x' - x\|^q. \tag{2.2}$$

The **US** condition on E is closely related to a condition on the modulus of smoothness ρ of E (or to a condition on the modulus of uniform smoothness ρ_1 of E). We refer the reader to [6], where these relations are discussed.

We point out that when looking for the global minimizer $\bar{x}$ of E, we can restrict ourselves to the set
$$\Omega := \{x : \ E(x) \le E(0)\},$$
since $\bar{x} \in \Omega$. In what follows, we will consider the minimization problem (1.1) over this set. Note that this is a convex set as a level set of a convex function.

We also will use the following lemma, proved in [6]. Other versions of this lemma have been discussed in [10].

Lemma 2.4. *Let* $\ell > 0$, $r > 0$, $B > 0$, *and* $\{a_m\}_{m=0}^{\infty}$ *and* $\{r_m\}_{m=1}^{\infty}$ *be sequences of non-negative numbers satisfying the inequalities*
$$a_J \le B, \quad a_{m+1} \le a_m(1 - \frac{r_{m+1}}{r} a_m^{\ell}), \quad m = J+1, J+2, \dots .$$
Then, we have for $m = J+1, J+2, \dots,$
$$a_m \le \max\{1, \ell^{-1/\ell}\} r^{1/\ell} (rB^{-\ell} + \Sigma_{k=J+1}^{m} r_k)^{-1/\ell}. \tag{2.3}$$

3. The Rescaled Pure Greedy Algorithm for Convex Optimization

In this section, we first describe our new convex optimization algorithm *RPGA(co)*$(\mu, \mathcal{D})$ with parameter μ and dictionary $\mathcal{D}$. Then, we prove the rate of convergence of $E(x_m) \to E(\bar{x})$, $m \to \infty$, where the sequence $\{x_m\}_{m=0}^{\infty} \subset X$ is the output of the algorithm.

Let E be a convex function, defined on the Banach space X, that satisfies the **US** condition (2.2) with $\alpha > 0$ and $1 < q \le 2$. Then, the *RPGA(co)*$(\mu, \mathcal{D})$ with parameter μ and dictionary $\mathcal{D}$ is as follows:

Step 0**:**

- Define $x_0 := 0$. If $E'(x_0) = 0$, stop the algorithm and define $x_k := x_0$, $k \ge 1$.

Step m**:**

- Assume x_{m-1} has been defined and $E'(x_{m-1}) \neq 0$. Do the following:
  - choose a direction $\varphi_{j_m} \in \mathcal{D}$ such that
$$|\langle E'(x_{m-1}), \varphi_{j_m}\rangle| = \sup_{\varphi\in\mathcal{D}} |\langle E'(x_{m-1}), \varphi\rangle|.$$
  - compute λ_m, given by the formula
$$\lambda_m := \operatorname{sgn}\{\langle E'(x_{m-1}), \varphi_{j_m}\rangle\}\,(\alpha\mu)^{-\frac{1}{q-1}}\,|\langle E'(x_{m-1}), \varphi_{j_m}\rangle|^{\frac{1}{q-1}}.$$
  - compute
$$\hat{x}_m := x_{m-1} - \lambda_m\varphi_{j_m}, \quad t_m := \operatorname{argmin}_{t\in\mathbb{R}} E(t\hat{x}_m).$$
  - define the next approximant as
$$x_m := t_m\hat{x}_m.$$
- If $E'(x_m) = 0$, stop the algorithm and define $x_k = x_m$, for $k > m$.
- If $E'(x_m) \neq 0$, proceed to Step $m + 1$.

Remark 3.1. *Let us observe that:*

- *all elements $x_m \in \Omega$, since x_m is obtained as a minimization of E over a set containing the* 0.
- *the algorithm will either stop when $E'(x_m) = 0$ (and its output will be the sequence $x_1, x_2, \ldots, x_m, x_m, x_m, \ldots$, where $x_m = \bar{x}$, because of Lemma* 2.2*), or will continue generating the infinite sequence $x_1, x_2, \ldots, x_m, x_{m+1}, \ldots$.*

At each step, the *RPGA(co)*$(\mu, \mathcal{D})$ performs a minimization of the objective function E along only a one dimensional space. This univariate optimization problem is called line search and is well studied in optimization theory, see [5]. If at Step m we were to use $\hat{x}_m$ as next approximant to $\bar{x}$ and not the minimizer x_m of E along the line generated by $\hat{x}_m$, then the algorithm would be very similar to the *EGA(*$\mathcal{C}$*)* from [8] . However, in contrast to *EGA(*$\mathcal{C}$*)*, the suggested algorithm achieves the optimal convergence rate of $\mathcal{O}(m^{1-q})$. In addition, it provides convergent results for all cases of $\bar{x} \in \mathcal{A}_1(\mathcal{D})$, and not only for $\bar{x}$ in the convex hull of the dictionary $\mathcal{D}$.

The following theorem is the main result of our paper, giving the optimal convergence rate for the *RPGA(co)*$(\mu, \mathcal{D})$, applied to E.

Theorem 3.2. *Let the convex function E satisfy* **Condition 0**, *the* **US** *condition* (2.2)*, and its minimizer $\bar{x} \in \mathcal{A}_1(\mathcal{D})$. Then, the RPGA(co)$(\mu, \mathcal{D})$ with parameter μ, where $\mu > \max\{1, \alpha^{-1}M_0M^{1-q}\}$, applied to E and a dictionary $\mathcal{D} = \{\varphi\}$ outputs a sequence $\{x_m\}_{m=0}^{\infty}$ of approximants to $\bar{x}$, where the error $e_m := E(x_m) - E(\bar{x})$ satisfies the inequality*

$$e_m \leq C_1 m^{1-q}, \quad k \geq 1, \quad \textit{with} \quad C_1 = C_1(q, \alpha, E, \mu).$$

Proof: If at Step k_0 the *RPGA(co)*$(\mu, \mathcal{D})$ had stopped, we had recovered the minimizer $\bar{x}$, namely we have that $\bar{x} = x_{k_0}$. Since the algorithm sets

$x_m = x_{k_0} = \bar{x}$ for $m > k_0$, then the error $e_m = 0$ for $m \geq k_0$. For values of $m < k_0$, or when the algorithm had not stopped, we have at Step m that

$$\|(x_{m-1} - \lambda_m \varphi_{j_m}) - x_{m-1}\| = \left(\frac{|\langle E'(x_{m-1}), \varphi_{j_m}\rangle|}{\alpha\mu}\right)^{\frac{1}{q-1}} \leq M,$$

because of the restriction $\mu > \max\{1, \alpha^{-1} M_0 M^{1-q}\}$ on the parameter μ and the fact that the elements of the dictionary $\mathcal{D}$ are normalized with norm 1. Therefore, we can apply the **US** condition (2.2) to $(x_{m-1} - \lambda_m \varphi_{j_m})$ and x_{m-1} and obtain that

$$\begin{aligned} E(\hat{x}_m) &= E(x_{m-1} - \lambda_m \varphi_{j_m}) \leq E(x_{m-1}) - \lambda_m \langle E'(x_{m-1}), \varphi_{j_m}\rangle + \alpha|\lambda_m|^q \\ &= E(x_{m-1}) - \frac{\mu - 1}{\mu} (\alpha\mu)^{-\frac{1}{q-1}} |\langle E'(x_{m-1}), \varphi_{j_m}\rangle|^{q/(q-1)}, \end{aligned}$$

where we use the value of λ_m and the fact that $\|\varphi_{j_m}\| = 1$. Since x_m is the minimizer of the function E along the line generated by $\hat{x}_m$, we have that $E(x_m) \leq E(\hat{x}_m)$, and thus, using the above inequality, we derive

$$E(x_m) \leq E(x_{m-1}) - \frac{\mu - 1}{\mu} (\alpha\mu)^{-\frac{1}{q-1}} |\langle E'(x_{m-1}), \varphi_{j_m}\rangle|^{\frac{q}{q-1}}. \tag{3.4}$$

Next, we provide a lower bound for $|\langle E'(x_{m-1}), \varphi_{j_m}\rangle|$. Let us fix $\varepsilon > 0$ and choose a representation for $\bar{x}$ via the elements from the dictionary $\mathcal{D}$, $\bar{x} = \sum_{\varphi \in \mathcal{D}} c_\varphi^\varepsilon \varphi$, such that

$$\sum_{\varphi \in \mathcal{D}} |c_\varphi^\varepsilon| < \|\bar{x}\|_1 + \varepsilon.$$

It follows from Lemma 2.3 and the definition of x_{m-1} that $\langle E'(x_{m-1}), x_{m-1}\rangle = 0$, and therefore

$$\begin{aligned} \langle E'(x_{m-1}), x_{m-1} - \bar{x}\rangle &= -\langle E'(x_{m-1}), \bar{x}\rangle = -\sum_\varphi c_\varphi^\varepsilon \langle E'(x_{m-1}), \varphi\rangle \\ &\leq |\langle E'(x_{m-1}), \varphi_{j_m}\rangle| \sum_\varphi |c_\varphi^\varepsilon| \\ &< |\langle E'(x_{m-1}), \varphi_{j_m}\rangle| (\|\bar{x}\|_1 + \varepsilon), \end{aligned}$$

where we have used the choice of φ_{j_m}. We let $\varepsilon \to 0$ and obtain that

$$\langle E'(x_{m-1}), x_{m-1} - \bar{x}\rangle \leq |\langle E'(x_{m-1}), \varphi_{j_m}\rangle| \|\bar{x}\|_1.$$

The latter inequality and Lemma 2.1, applied for $x = x_{m-1}$ and $x' = \bar{x}$ give

$$\|\bar{x}\|_1^{-1} e_{m-1} \leq |\langle E'(x_{m-1}), \varphi_{j_m}\rangle|, \tag{3.5}$$

which is the desired estimate from below for $|\langle E'(x_{m-1}), \varphi_{j_m}\rangle|$. In particular,

$$e_0 \leq |\langle E'(x_0), \varphi_{j_1}\rangle| \|\bar{x}\|_1 \leq \|E'(0)\| \|\bar{x}\|_1. \tag{3.6}$$

We substitute (3.3) in (3.4) and derive

$$E(x_m) \leq E(x_{m-1}) - \frac{\mu - 1}{\mu} (\alpha\mu)^{-\frac{1}{q-1}} \|\bar{x}\|_1^{-\frac{q}{q-1}} e_{m-1}^{\frac{q}{q-1}}.$$

Subtracting $E(\bar{x})$ from both sides gives

$$e_m \le e_{m-1}\left(1 - \frac{\mu-1}{(\alpha\mu^q\|\bar{x}\|_1^q)^{\frac{1}{q-1}}} e_{m-1}^{\frac{1}{q-1}}\right).$$

Now we apply Lemma 2.4 for the sequences $\{e_m\}_{m=0}^{\infty}$, $\{r_m\}_{m=1}^{\infty}$, with $r_m = 1$, and

$$J = 0, \quad B = \|E'(0)\|\|\bar{x}\|_1, \quad \ell = \frac{1}{q-1}, \quad r = \frac{(\alpha\mu^q\|\bar{x}\|_1^q)^{\frac{1}{q-1}}}{\mu-1},$$

and derive that

$$e_m \le \frac{\alpha\mu^q\|\bar{x}\|_1^q}{(\mu-1)^{q-1}}\left(\frac{\|\bar{x}\|_1}{\mu-1}\left(\frac{\alpha\mu^q}{\|E'(0)\|}\right)^{\frac{1}{q-1}} + m\right)^{1-q}, \quad m = 1, 2, \dots.$$

The proof is completed. □

Remark 3.3. *One can further optimize the constant $C_1 = C_1(q, \alpha, E, \mu)$ with respect to μ and select a specific value for $\mu > \max\{1, \alpha^{-1}M_0M^{1-q}\}$ that will guarantee the best convergence rate in terms of best constants.*

Careful analysis of the above proof shows that a similar theorem holds in the following case.

Theorem 3.4. *Let the convex function $E : X \to \mathbb{R}$ be Frechet differentiable with minimizer $\bar{x} \in \mathcal{A}_1(\mathcal{D})$. Let there be constants $\alpha > 0$ and $1 < q \le 2$, such that for all x, $x' \in X$,*

$$E(x') - E(x) - \langle E'(x), x' - x\rangle \le \alpha\|x' - x\|^q.$$

Then, the application of the RPGA(co)$(\mu, \mathcal{D})$ with parameter $\mu > 1$ and a dictionary $\mathcal{D} = \{\varphi\}$ outputs a sequence $\{x_m\}_{m=0}^{\infty}$, such that the error $e_m := E(x_m) - E(\bar{x})$ satisfies the inequality

$$e_m \le C_1 m^{1-q}, \quad m = 1, 2, \dots, \quad \text{where} \quad C_1 = C_1(q, \alpha, E, \mu).$$

4. The Weak Rescaled Pure Greedy Algorithm for Convex Optimization

In this section, we describe the weak version of our algorithm, which we denote by *WRPGA(co)*$(\{\ell_m\}, \{\mu_m\}, \mathcal{D})$, with weakness sequence $\{\ell_m\}$, $\ell_m \in (0, 1]$ $m = 1, 2, \dots$, and parameter sequence $\{\mu_m\}$, where

$$\mu_m > \max\{1, \alpha^{-1}M_0M^{1-q}\}, \quad m = 1, 2, \dots.$$

In the case when $\ell_m = 1$ and $\mu_m = \mu$, $m = 1, 2, \dots$, the weak version of the algorithm coincides with the *RPGA(co)*$(\mu, \mathcal{D})$. The weakness sequence $\{\ell_m\}$ allows us to have some freedom in the selection of the next direction φ_{j_m}, while the parameter sequence $\{\mu_m\}$ gives more choices in how much to advance along the selected direction φ_{j_m}.

WRPGA(co)$(\{\ell_m\}, \{\mu_m\}, \mathcal{D})$:

Step 0:

- Define $x_0^{\rm w} := 0$. If $E'(x_0^{\rm w}) = 0$, stop the algorithm and define $x_k^{\rm w} := x_0^{\rm w}$, $k \geq 1$.

Step m:

- Assume $x_{m-1}^{\rm w}$ has been defined and $E'(x_{m-1}^{\rm w}) \neq 0$. Do the following:
  - choose a direction $\varphi_{j_m} \in \mathcal{D}$ such that

$$|\langle E'(x_{m-1}^{\rm w}), \varphi_{j_m}\rangle| \geq \ell_m \sup_{\varphi\in\mathcal{D}} |\langle E'(x_{m-1}^{\rm w}), \varphi\rangle|.$$

  - compute λ_m given by the formua

$$\lambda_m := \operatorname{sgn}\{\langle E'(x_{m-1}^{\rm w}), \varphi_{j_m}\rangle\} \, (\alpha\mu_m)^{-\frac{1}{q-1}} \, |\langle E'(x_{m-1}^{\rm w}), \varphi_{j_m}\rangle|^{\frac{1}{q-1}}.$$

  - compute

$$\hat{x}_m^{\rm w} := x_{m-1}^{\rm w} - \lambda_m \varphi_{j_m}, \quad t_m := \operatorname{argmin}_{t\in\mathbb{R}} E(t\hat{x}_m^{\rm w}).$$

  - define the next approximant as

$$x_m^{\rm w} = t_m \hat{x}_m^{\rm w}.$$

- If $E'(x_m^{\rm w}) = 0$, stop the algorithm and define $x_k^{\rm w} := x_m^{\rm w}$, for $k > m$.
- If $E'(x_m^{\rm w}) \neq 0$, proceed to Step $m+1$.

As noted before, all elements of the generated by the algorithm sequence $\{x_k^{\rm w}\}$ belong to the set Ω. The next theorem is the main result about the convergence rate of the *WRPGA(co)*$(\{\ell_m\}, \{\mu_m\}, \mathcal{D})$.

Theorem 4.1. *Let the convex function E satisfy* **Condition 0** *and the* **US** *condition and its minimizer $\bar{x} \in \mathcal{A}_1(\mathcal{D})$. Then, the application of the WRPGA(co)$(\{\ell_m\}, \{\mu_m\}, \mathcal{D})$ with a weakness sequence $\{\ell_m\}$, parameter sequence $\{\mu_m\}$, where $\mu_m > \max\{1, \alpha^{-1} M_0 M^{1-q}\}$, and a dictionary $\mathcal{D} = \{\varphi\}$ outputs a sequence $\{x_m^{\rm w}\}_{m=0}^{\infty}$, such that the following inequality holds*

$$e_m^{\rm w} := E(x_m^{\rm w}) - E(\bar{x}) \leq \alpha \|x\|_1^q \left(C_1 + \sum_{j=1}^{m} (\mu_j - 1) \left(\frac{\ell_j}{\mu_j}\right)^{\frac{q}{q-1}} \right)^{1-q}, \quad m \geq 1,$$

with $C_1 = C_1(q, \alpha, E)$.

Proof: Similarly to the proof of Theorem 3.2, we have that for $m \geq 1$,

$$E(x_m^{\rm w}) \leq E(x_{m-1}^{\rm w}) - \frac{\mu_m - 1}{\mu_m} \, (\alpha\mu_m)^{-\frac{1}{q-1}} \, |\langle E'(x_{m-1}^{\rm w}), \varphi_{j_m}\rangle|^{\frac{q}{q-1}}. \tag{4.7}$$

The same way one can easily derive the lower estimate

$$\|\bar{x}\|_1^{-1} \ell_m e_{m-1} \leq |\langle E'(x_{m-1}^{\rm w}), \varphi_{j_m}\rangle|.$$

We use (4.7) and the lower estimate for $|\langle E'(x_{m-1}^{\rm w}), \varphi_{j_m}\rangle|$ to obtain

$$e_m^{\rm w} \leq e_{m-1}^{\rm w} \left(1 - \frac{\mu_m - 1}{\mu_m} \, (\alpha\mu_m)^{-\frac{1}{q-1}} \, \ell_m^{\frac{q}{q-1}} \, \|\bar{x}\|_1^{-\frac{q}{q-1}} \, [e_{m-1}^{\rm w}]^{\frac{1}{q-1}} \right).$$

We also have that, by definition $e_0 = E(0) - E(\bar{x})$. We now apply Lemma 2.4 for the sequences $\{e_m^{\rm w}\}$ and $\{r_m\}$ with $J = 0$, $B = E(0) - E(\bar{x})$,

$$r_m = (\mu_m - 1)\left(\frac{\ell_m}{\mu_m}\right)^{\frac{q}{q-1}}, \quad \ell = \frac{1}{q-1} > 0, \quad r = (\alpha\|\bar{x}\|_1^q)^{\frac{1}{q-1}},$$

and derive that

$$e_m^{\rm w} \le \alpha\|\bar{x}\|_1^q \left(\left(\frac{\alpha\|\bar{x}\|_1^q}{E(0) - E(\bar{x})}\right)^{\frac{1}{q-1}} + \sum_{j=1}^{m}(\mu_j - 1)\left(\frac{\ell_j}{\mu_j}\right)^{\frac{q}{q-1}}\right)^{1-q}.$$

The proof is completed. □

As in the previous case, careful analysis can provide the best choice for parameters $\{\mu_m\}$ in terms of optimal constants. Also, a similar theorem holds in the following case.

Theorem 4.2. *Let the convex function $E : X \to \mathbb{R}$ be Frechet differentiable with minimizer $\bar{x} \in \mathcal{A}_1(\mathcal{D})$. Let there be constants $\alpha > 0$ and $1 < q \le 2$, such that for all x, $x' \in X$,*

$$E(x') - E(x) - \langle E'(x), x' - x\rangle \le \alpha\|x' - x\|^q.$$

Then, the application of the WRPGA(co)($\{\ell_m\}, \{\mu_m\}, \mathcal{D}$) with a weakness sequence $\{\ell_m\}$, parameter sequence $\{\mu_m\}$, $\mu_m > 1$, and a dictionary $\mathcal{D} = \{\varphi\}$ outputs a sequence $\{x_m^{\rm w}\}_{m=0}^{\infty}$, such that the following inequality holds

$$e_m^{\rm w} := E(x_m^{\rm w}) - E(\bar{x}) \le \alpha\|x\|_1^q \left(C_1 + \sum_{j=1}^{m}(\mu_j - 1)\left(\frac{\ell_j}{\mu_j}\right)^{\frac{q}{q-1}}\right)^{1-q}, \quad k \ge 1,$$

with $C_1 = C_1(q, \alpha, E)$.

5. Approximation in Hilbert spaces

Let us consider the case when X is a Hilbert space H with a norm, induced by the scalar product, namely $\|\cdot\| = (\cdot,\cdot)^{1/2}$, and the objective function $E : H \to \mathbb{R}$ is

$$E(x) := \|x - \bar{x}\|^2, \tag{5.8}$$

where $\bar{x}$ is a fixed element in the Hilbert space H. Note that E is Frechet differentiable on H and its Frechet derivative at x is the linear functional $E'(x)$ that acts on $h \in H$ as

$$\langle E'(x), h\rangle = 2(x - \bar{x}, h).$$

Moreover, we have that

$$\begin{aligned} E(x') - E(x) - \langle E'(x), x' - x\rangle &= \|x' - \bar{x}\|^2 - \|x - \bar{x}\|^2 - 2(x - \bar{x}, x' - x) \\ &= \|x - x'\|^2, \end{aligned}$$

and therefore the function E, defined in (5.8), satisfies the conditions of Theorem 3.4 with $\alpha = 1$ and $q = 2$. Clearly, our algorithm can be applied for finding the minimizer $\bar{x}$ of E, and in this particular case it is the following.

RPGA(co)$(\mu, \mathcal{D})$

Step 0**:**

- Define $x_0 := 0$. If $\bar{x} = 0$, stop the algorithm and define $x_k := x_0 = \bar{x}$, $k \geq 1$.

Step m**:**

- Assume x_{m-1} has been defined and $x_{m-1} \neq \bar{x}$. Do the following:
  - choose a direction $\varphi_{j_m} \in \mathcal{D}$ such that
$$|(x_{m-1} - \bar{x}, \varphi_{j_m})| = \sup_{\varphi \in \mathcal{D}} |(x_{m-1} - \bar{x}, \varphi)|.$$
  - compute λ_m given by the formula
$$\lambda_m := \frac{2}{\mu}(x_{m-1} - \bar{x}, \varphi_{j_m}).$$
  - compute
$$\hat{x}_m := x_{m-1} - \lambda_m \varphi_{j_m}, \quad t_m := \frac{(\bar{x}, \hat{x}_m)}{\|\hat{x}_m\|^2}.$$
  - define the next approximant as
$$x_m = t_m \hat{x}_m.$$
- If $x_m = \bar{x}$, stop the algorithm and define $x_k := x_m = \bar{x}$, for $k > m$.
- If $x_m \neq \bar{x}$, proceed to Step $m + 1$.

Each term x_m of the output sequence $\{x_m\}$ of this algorithm can be viewed as a m-term approximant to $\bar{x}$ from the dictionary $\mathcal{D}$, and therefore, the algorithm can be viewed as a greedy algorithm for approximating the element $\bar{x} \in H$. The convergence theorem in this case can be either directly proved, or one can just apply Theorem 3.4 for the function under investigation $E(x) = \|x - \bar{x}\|^2$. However, a direct proof gives a better understanding of the constants, appearing in the inequalities, as demonstrated below.

Theorem 5.1. *Let $\bar{x} \in \mathcal{A}_1(\mathcal{D}) \subset H$. The RPGA$(\mu, \mathcal{D})$, with parameter $\mu > 1$ and a dictionary $\mathcal{D}$ outputs a sequence $\{x_m\}$ of approximations to $\bar{x}$ satisfying the following error estimate*

$$\|\bar{x} - x_m\| \leq \frac{\mu}{2\sqrt{\mu - 1}} \|\bar{x}\|_1 (m+1)^{-1/2}, \quad m = 1, 2 \dots . \tag{5.9}$$

Proof: The proof follows the arguments from Theorem 3.2 and is omitted here. □

Remark 5.2. *The RPGA(co)$(2, \mathcal{D})$, with $\mu = 2$, applied to the convex function $E(x) := \|x - \bar{x}\|^2$, is the recently discovered Rescaled Pure Greedy Algorithm (RPGA) for approximating elements in H via linear combinations of at most m dictionary elements (see* [7] *for detailed discussion of this algorithm). The latter was suggested as a modification to the Pure Greedy Algorithm, known also as Matching Pursuit, and an alternative to the Relaxed Greedy Algorithm or the Orthogonal Greedy Algorithm for Hilbert spaces. Theorem 5.1 with $\mu = 2$ is Theorem 3.1 from* [7]. *Note that the optimal*

value of the constant $C = \frac{\mu}{2\sqrt{\mu-1}}$ *in* (5.9) *is achieved when* $\mu = 2$ *and this optimal value is* $C = 1$.

References

[1] J. Borwein, A. Guiro, P. Hajek, and J. Vanderwerff, *Uniformly convex functions on Banach Spaces*, Proc. Amer. Math. Soc., **137**, 1081–1091, 2009.

[2] S. Boyd, L. Vandenberghe, Convex optimization, Cambridge University Press, 2009.

[3] R. DeVore, V. Temlyakov, *Some remarks on greedy algorithms*, Advances in Computational Math., **5**, 173–187, 1996.

[4] R. DeVore, V. Temlyakov, *Convex optimization on Banach spaces*, Foundations of Computational Mathematics, **16**(2), 369–394, 2016.

[5] A. Nemirovski, *Optimization II: Numerical methods for nonlinear continuous optimization*, Lecture Notes, Israel Institute of Technology, 1999.

[6] H. Nguyen, G. Petrova, *Greedy strategies for convex optimization*, submitted, *arXiv:1401.1754*.

[7] G. Petrova, *Rescaled Pure Greedy Algorithm for Hilbert and Banach Spaces*, submitted, *arXiv:1505.03604*.

[8] V. Temlyakov, *Greedy expansions in convex optimization*, Proceedings of the Steklov Institute of Mathematics, **284**(1), 244–262, 2014.

[9] V. Temlyakov, *Greedy approximation in convex optimization*, Constr. Approx., **41**(2), 269–296, 2015.

[10] V. Temlyakov, Greedy approximation, Cambridge monographs on Applied and Computational Mathematics, Cambridge University Press, 2011.

[11] C. Zalinescu, Convex Analysis in General Vector Spaces, World Scientific Publishing Co. Inc., River Edge, NJ, 2002.

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

P. Komarek, *Logistic Regression for Data Mining and High-Dimensional Classification*, Chapter 4, 22-28, PhD Dissertation, Dept. of Math Sciences, Carnegie Mellon University, 2004.

Guergana Petrova
Department of Mathematics, Texas A&M University, College Station, TX 77843, USA
gpetrova@math.tamu.edu
Zheming Gao
Department of Operations Research, NCSU, Raleigh, NC 27695, USA
zgao5@ncsu.edu