2017

Convergence and Rate of Convergence of Approximate Greedy-Type Algorithms

Anton Dereventsov
University of South Carolina

Convergence and Rate of Convergence of Approximate Greedy-Type Algorithms

by

Anton Dereventsov

Specialist Degree
Lomonosov Moscow State University 2012

Submitted in Partial Fulfillment of the Requirements

for the Degree of Doctor of Philosophy in

Mathematics

College of Arts and Sciences

University of South Carolina

2017

Accepted by:

Vladimir Temlyakov, Major Professor

Stephen Dilworth, Committee Member

Peter Binev, Committee Member

Ognian Trifonov, Committee Member

Vladimir Gudkov, Committee Member

Cheryl L. Addy, Vice Provost and Dean of the Graduate School

Acknowledgments

I would like to thank Dr. Vladimir Temlyakov for his continual advice, invaluable discussions of new concepts, and guidance regarding the research process and writing of this manuscript.

I am also thankful to Dr. Stephen Dilworth for his vast knowledge of Banach space theory, eagerness to discuss ideas, and readiness to answer any and all questions.

Thanks to Rachel Harrison, without whose support and editing finesse I never would have learned the importance of the Oxford comma.

And special thanks to SukiBoo Nugget, who doesn't do much, but somehow is always enough.

Abstract

In this dissertation we study the questions of convergence and rate of convergence of greedy-type algorithms under imprecise step evaluations. Such algorithms are in demand as the issue of calculation errors appears naturally in applications.

We address the question of strong convergence of the Chebyshev Greedy Algorithm (CGA), which is a generalization of the Orthogonal Greedy Algorithm (also known as the Orthogonal Matching Pursuit), and show that the class of Banach spaces for which the CGA converges for all dictionaries and objective elements is strictly between smooth and uniformly smooth Banach spaces.

We analyze an application-oriented modification of the CGA, the generalized Approximate Chebyshev Greedy Algorithm (gAWCGA), in which we are allowed to perform every operation of the algorithm with a controlled inaccuracy in the form of both relative and absolute errors. Such permission is essential for numerical applications and simplifies realization of the algorithm. We obtain necessary and sufficient conditions for the convergence of the gAWCGA in all uniformly smooth Banach spaces, all dictionaries and all elements.

Greedy algorithms in convex optimization have been of particular interest recently. We discuss algorithms that do not use the derivative of the objective function, and thus offer an alternative to traditional methods of convex minimization. We recall two known algorithms: the Relaxed E-Greedy Algorithm (REGA(co)) and the E-Greedy Algorithm with Free Relaxation (EGAFR(co)), and introduce the Rescaled Relaxed E-Greedy Algorithm for convex optimization (RREGA(co)), which is computationally simpler than the EGAFR(co) and does not suffer the limitations of the REGA(co).

Table of Contents

Chapter 1

Introduction

Applications like signal and image processing often require that a signal/picture is decomposed with respect to a fixed collection of elements. Furthermore, it is desirable that the decomposition is sparse with respect to the selected collection, as such representation will require less memory to store. This problem can be formalized in the following way: find an m-term approximation of an element f of a Hilbert (or, more generally, Banach) space X by a linear combination of elements of a fixed set $\mathcal{D}$ (called dictionary). This statement is the general problem of sparse approximation.

Greedy algorithms are designed specifically to obtain such approximations. For an element $f \in X$ and a dictionary $\mathcal{D}$, a general greedy algorithm iteratively produces sequences of approximations $\{G_m\}_{m=1}^{\infty}$ and remainders $\{f_m\}_{m=1}^{\infty}$ in the following way: on each iteration m it chooses an atom $\phi_m \in \mathcal{D}$ that is close in some sense to the previous remainder f_{m-1}, and then builds the next approximation G_m using the chosen atom ϕ_m.

This sequential nature of greedy algorithms is favorable in applications as it guarantees that an approximation G_m is supported on at most m elements of a dictionary $\mathcal{D}$, and thus allows us to obtain sparse approximations of f with respect to $\mathcal{D}$. Additionally, there is an immediate regulation between the sparsity and the accuracy of the approximation which allows us to acquire the optimal approximation for each particular problem.

Essentially, a greedy algorithm is determined by two things: how it chooses the next atom $\phi_m \in \mathcal{D}$ and how it constructs the next approximation G_m. In the case of a Hilbert space it is natural to choose ϕ_m as an element that maximizes the inner product $|\langle \cdot, f_{m-1} \rangle|$. There are two classical approaches to building G_m: to use all atoms that were chosen until the current iteration $(\phi_1, \ldots, \phi_m)$, or to use only the last one (ϕ_m). Usually algorithms that favor the first approach are more computationally complicated but tend to provide an approximation iteration-wise faster. On the other hand, algorithms that use the second approach are generally simpler computationally and might be advantageous in some cases, as they change only one coefficient in the decomposition on each iteration and, therefore, provide an expansion. However, they usually require more iterations to achieve the required accuracy.

One well-known greedy algorithm for the Hilbert space setting is the Orthogonal Greedy Algorithm (OGA), also known as the Orthogonal Matching Pursuit (see e.g. Pati, Rezaiifar, and Krishnaprasad 1993 or DeVore and Temlyakov 1996). In order to construct an approximation G_m, the OGA takes the orthogonal projection of f on the subspace generated by all the chosen atoms $\phi_1, \ldots, \phi_m$.

Definition (OGA)**.** Set $f_0 = f \in H$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ (we assume existence) such that

$$|\langle \phi_n, f_{n-1} \rangle| = \sup_{g \in \mathcal{D}} |\langle g, f_{n-1} \rangle|,$$

2. denote $\Phi_n = \operatorname{span}\{\phi_j\}_{j=1}^n$ and take

$$G_n = \operatorname{Proj}_{\Phi_n}(f),$$

3. set $f_n = f - G_n$.

Another famous algorithm that uses the simpler approach in constructing approximations is the Pure Greedy Algorithm (PGA), also known as the Matching Pursuit (see e.g. Mallat and Z. Zhang 1993 or DeVore and Temlyakov 1996). Instead of projecting on the whole subspace, the PGA only projects the previous remainder f_{m-1} on the newly chosen atom ϕ_m.

Definition (PGA)**.** Set $f_0 = f \in H$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ (we assume existence) such that

$$|\langle \phi_n, f_{n-1} \rangle| = \sup_{g \in \mathcal{D}} |\langle g, f_{n-1} \rangle|,$$

2. denote $\lambda_n = \langle \phi_n, f_{n-1} \rangle$ and take

$$G_n = G_{n-1} + \lambda_n \phi_n,$$

3. set $f_n = f - G_n$.

Both algorithms converge for all Hilbert spaces X, all dictionaries $\mathcal{D}$ and elements f, and are widely used in applications. While the OGA generally provides faster convergence rates, the PGA can be substantially computationally simpler, especially on higher iterations. For a more detailed analysis on the OGA and the PGA, we refer the reader to the book Temlyakov 2011 and the short paper Dereventsov 2012.

It is important to note that the stated algorithms are generally not realizable since the supremum of the inner product might not be attainable on the dictionary. To overcome this problem, usually the "weak" version of an algorithm is used. In a weak form, the original condition on ϕ_m is replaced by the following one

$$|\langle \phi_m, f_{m-1} \rangle| \geq t_m \sup_{g \in \mathcal{D}} |\langle g, f_{m-1} \rangle|$$

with some $0 \leq t_m < 1$.

The weak versions of the OGA and the PGA (called the Weak Orthogonal Greedy Algorithm (WOGA) and the Weak Greedy Algorithm (WGA), respectively) were introduced in Temlyakov 2000. The convergence of these weak algorithms for all dictionaries $\mathcal{D}$ and elements f was proven in case

$$\sum_{n=1}^{\infty} t_n^2 = \infty \quad \text{for the WOGA}$$

and

$$\sum_{n=1}^{\infty} \frac{t_n}{n} = \infty \quad \text{for the WGA.}$$

A weak greedy algorithm is always realizable as long as all $t_n < 1$. However, it still might be hard to run an algorithm due to difficulties in evaluating the inner product (and/or the projection in the OGA). Hence, it is natural for numerical applications to assume that the steps of an algorithm are performed with some errors. This idea was considered in Gribonval and Nielsen 2001, which resulted in the Approximate Weak Greedy Algorithm (AWGA) — a modification of the WGA, which allows relative errors in calculating the coefficients of the decomposition.

Definition (AWGA)**.** Set $f_0 = f \in H$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ such that

$$|\langle \phi_n, f_{n-1} \rangle| \geq t_m \sup_{g \in \mathcal{D}} |\langle g, f_{n-1} \rangle|,$$

2. denote $\lambda_n = (1 + \varepsilon_n)\langle \phi_n, f_{n-1} \rangle$ and take

$$G_n = G_{n-1} + \lambda_n \phi_n,$$

3. set $f_n = f - G_n$.

It was shown that the AWCGA converges for all dictionaries and elements if

$$\sum_{n=1}^{\infty} \frac{t_n(1 - \varepsilon_n^2)}{n} = \infty.$$

In the AWCGA, sequences $\{t_n\}_{n=1}^{\infty}$ and $\{\varepsilon_n\}_{n=1}^{\infty}$ represent the allowable inaccuracies in performing the calculations and can be changed to fit the algorithm for the current problem.

However some problems are not modeled well by relative errors: for example, a scale gives the same margin of error on each weighting regardless of the weight of the object. Thus, it seems logical to consider greedy algorithms which additionally allow absolute errors in step evaluations. This idea was implemented in Galatenko and Livshitz 2005, where the authors proposed the generalized Approximate Weak Greedy Algorithm (gAWCGA) — a further modification of the WGA with both relative and absolute errors.

Definition (gAWGA)**.** Set $f_0 = f \in H$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ such that

$$|\langle \phi_n, f_{n-1} \rangle| \geq t_n \sup_{g \in \mathcal{D}} |\langle g, f_{n-1} \rangle| - q_n,$$

2. denote $\lambda_n = (1 + \varepsilon_n)\langle \phi_n, f_{n-1} \rangle + \xi_n$ and take

$$G_n = G_{n-1} + \lambda_n \phi_n,$$

3. set $f_n = f - G_n$.

In the gAWCGA there are four inaccuracy sequences: $\{t_n\}_{n=1}^{\infty}$, $\{q_n\}_{n=1}^{\infty}$, $\{\varepsilon_n\}_{n=1}^{\infty}$ and $\{\xi_n\}_{n=1}^{\infty}$, which make this algorithm more flexible for various applications. The gAWCGA converges for all dictionaries and elements if the following conditions hold

$$\sum_{n=1}^{\infty} \frac{t_n \sqrt{1 - \varepsilon_n^2}}{n} = \infty,$$
$$\sum_{n=1}^{\infty} \frac{\xi_n^2}{1 - \varepsilon_n^2} < \infty,$$
$$\sum_{n=1}^{\infty} q_n^2 (1 - \varepsilon_n^2) < \infty.$$

While greedy algorithms in Hilbert spaces are well studied and widespread, some applications require approximation in non-Hilbert norms (see e.g. Donahue et al. 1997), which can be achieved by generalizing greedy algorithms to the Banach space setting. The immediate question that arises is how to choose the next atom $\phi_m \in \mathcal{D}$ in a space X without an inner product. There are two proposed ways to make this choice (see Temlyakov 2011, chapter 6):

1. calculate the norm directly, i.e. $\{\phi_m, \lambda_m\} = \operatorname*{argmin}_{\phi\in\mathcal{D},\, \lambda\in\mathbb{R}} \|f_{m-1} - \lambda\phi\|$;
2. utilize norming functionals, i.e. $\phi_m = \operatorname*{argmax}_{\phi\in\mathcal{D}} |F_{f_{m-1}}(\phi)|$.

Algorithms of the first type are called X-greedy algorithms. One such algorithm is the X-Greedy Algorithm (XGA) — a direct generalization of the PGA, which was introduced in Temlyakov 2003.

Definition (XGA)**.** Set $f_0 = f \in X$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ (we assume existence) and $\lambda_n \in \mathbb{R}$ such that

$$\{\phi_n, \lambda_n\} = \operatorname*{argmin}_{\phi\in\mathcal{D},\, \lambda\in\mathbb{R}} \|f - (G_{n-1} + \lambda\phi)\|\,,$$

2. take

$$G_n = G_{n-1} + \lambda_n\phi_n,$$

3. set $f_n = f - G_n$.

Some results on the convergence of the XGA were presented in Dubinin 1997, Livshitz 2003, Dilworth et al. 2008 and Livshitz 2010. However, to the best of this author's knowledge, there are no known results on the convergence of the XGA for general Banach spaces, dictionaries and elements. Nevertheless, there are some modifications of the XGA that perform well (see e.g. Livshitz 2003 or Section 6.8 in the book Temlyakov 2011).

Greedy algorithms in Banach spaces that use the second approach are called dual greedy algorithms. One distinguished dual greedy algorithm is the Chebyshev Greedy Algorithm (CGA) — a generalization of the OGA, which was introduced and studied in Temlyakov 2001.

Definition (CGA)**.** Set $f_0 = f \in X$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ (we assume existence) such that

$$|F_{f_{n-1}}(\phi_n)| = \sup_{g\in\mathcal{D}} |F_{f_{n-1}}(g)|,$$

2. denote $\Phi_n = \operatorname{span}\{\phi_j\}_{j=1}^n$ and find any $G_n \in \Phi_n$ satisfying

$$\|f - G_n\| = \inf_{G\in\Phi_n} \|f - G\|,$$

3. set $f_n = f - G_n$.

It is known that the CGA performs well in a wide class of Banach spaces (see e.g. Temlyakov 2001 or Dilworth, Kutzarova, and Temlyakov 2002). In Chapter 3 we further discuss the question of strong convergence. Namely, we establish that the class of Banach spaces for which the CGA converges for all dictionaries and elements is strictly between smooth and uniformly smooth Banach spaces.

Similarly to the OGA, in order to solve the question of realizability, the weak version of the CGA (called WCGA) was introduced in Temlyakov 2001.

Definition (WCGA)**.** Set $f_0 = f \in X$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ such that

$$|F_{f_{n-1}}(\phi_n)| \geq t_n \sup_{g\in\mathcal{D}} |F_{f_{n-1}}(g)|,$$

2. denote $\Phi_n = \operatorname{span}\{\phi_j\}_{j=1}^n$ and find any $G_n \in \Phi_n$ satisfying

$$\|f - G_n\| = \inf_{G\in\Phi_n} \|f - G\|,$$

3. set $f_n = f - G_n$.

It was shown that the WCGA converges in all uniformly smooth Banach spaces X with the modulus of smoothness of non-trivial power type $1 < q \leq 2$ for all dictionaries $\mathcal{D}$ and elements $f \in X$ as long as

$$\sum_{n=1}^{\infty} t_n^p = \infty,$$

where $p = q/(q-1)$. Moreover, it is known that this condition is sharp.

One fundamental drawback of the WCGA is that calculating the projection of an element on a subspace in a Banach space might be computationally unfeasible, especially on high iterations. Additionally, it might be hard to evaluate norming functionals and/or to find the next atom ϕ_m due to a large dictionary size. To simplify the realization of the WCGA, the simplified version was proposed in Temlyakov 2005 — the Approximate Weak Chebyshev Greedy Algorithm (AWCGA), in which we are allowed to evaluate the norming functional F_{f_m}, to choose ϕ_m, and to find G_m with some relative errors.

Definition (AWCGA)**.** Set $f_0 = f$ and for each $n \geq 1$

1. take any functional F_{n-1} satisfying

$$\|F_{n-1}\| \leq 1 \quad \text{and} \quad F_{n-1}(f_{n-1}) \geq (1-\delta_{n-1}) \|f_{n-1}\|,$$

2. find any $\phi_n \in \mathcal{D}$ such that

$$|F_{n-1}(\phi_n)| \geq t_n \sup_{g \in \mathcal{D}} |F_{n-1}(g)|,$$

3. denote $\Phi_n = \operatorname{span}\{\phi_j\}_{j=1}^n$ and find any $G_n \in \Phi_n$ satisfying

$$\|f - G_n\| \leq (1+\eta_n) \inf_{G \in \Phi_n} \|f - G\|,$$

4. set $f_n = f - G_n$.

It was proven that the AWCGA converges in a Banach space X with the modulus of smoothness of power type $1 < q \leq 2$ for all dictionaries $\mathcal{D}$ and elements $f \in X$ if the following conditions hold:

$$\sum_{n=1}^{\infty} t_n^p = \infty,$$

$$\delta_n = o(t_{n+1}^p),$$

$$\eta_n = o(t_{n+1}^p),$$

where $p = q/(q-1)$. Similarly to the WCGA, the first condition is sharp.

We note that while the AWCGA uses relative errors, greedy algorithms with absolute errors in Banach spaces were considered in Donahue et al. 1997.

It is clear that as we increase error sequences, the algorithm becomes easier to run, but at the same time it may stop converging in some cases. That is why it is important to establish the necessary and sufficient conditions on the error sequences that guarantee convergence for all dictionaries and elements.

In Chapter 4 we further discuss the issue of simplified step evaluation for the CGA. Namely, we introduce the generalized Approximate Weak Chebyshev Greedy Algorithm (gAWCGA) — a modification of the CGA with both relative and absolute inaccuracies.

Definition (gAWCGA)**.** Set $f_0 = f$ and for each $n \geq 1$

1. take any functional F_{n-1} satisfying

$$\|F_{n-1}\| \leq 1 \quad \text{and} \quad F_{n-1}(f_{n-1}) \geq (1-\delta_{n-1}) \|f_{n-1}\| - \delta'_{n-1},$$

2. find any $\phi_n \in \mathcal{D}$ such that

$$|F_{n-1}(\phi_n)| \geq t_n \sup_{g \in \mathcal{D}} |F_{n-1}(g)| - t'_n,$$

3. denote $\Phi_n = \operatorname{span}\{\phi_j\}_{j=1}^n$ and find any $G_n \in \Phi_n$ satisfying

$$\|f - G_n\| \leq (1+\eta_n) \inf_{G \in \Phi_n} \|f - G\| + \eta'_n,$$

4. set $f_n = f - G_n$.

We investigate how the convergence of the gAWCGA depends on these errors and establish conditions on the sequences that guarantee convergence of the algorithm in all uniformly smooth Banach spaces. The novelty of our approach is that we only require that the error sequences contain infinitely many sufficiently small values rather than requiring that the whole sequence being sufficiently small; i.e. we do not demand that every iteration of the algorithm is adequately precise and allow some "bad" steps. Concretely, we show that the gAWCGA converges in a Banach space X with the modulus of smoothness of power type $1 < q \leq 2$ for all dictionaries $\mathcal{D}$ and elements $f \in X$ if the following conditions hold for a subsequence $\{n_k\}_{k=1}^{\infty}$:

$$\sum_{k=1}^{\infty} t_{n_k+1}^p = \infty, \qquad t'_{n_k+1} = o(t_{n_k+1}),$$

$$\delta_{n_k} = o(t_{n_k+1}^p), \qquad \delta'_{n_k} = o(t_{n_k+1}^p),$$

$$\eta_{n_k} = o(t_{n_k+1}^p), \qquad \eta'_{n_k} = o(t_{n_k+1}^p),$$

where $p = q/(q-1)$.

These conditions are weaker than the known conditions for the AWCGA, and, more importantly, we prove that they are sharp. Moreover, we investigate how these inaccuracies affect the rate of convergence of the gAWCGA, and estimate the inaccuracy parameters which provide the convergence rate of the same order as that of the CGA.

Recently, greedy algorithms found their application in the field of convex optimization (see e.g. Shalev-Shwartz, Srebro, and T. Zhang 2010, Clarkson 2010, Tewari, Ravikumar, and Dhillon 2011, DeVore and Temlyakov 2014, Temlyakov 2015, and Nguyen and Petrova 2016). The general problem of convex optimization is to minimize a convex real-valued function E defined on a real Banach space $(X, \|\cdot\|)$. The problem of greedy approximation, while seemingly different, can be viewed as a special case of the convex optimization problem with $E(x) = \|f - x\|$. It turns out that greedy algorithms can be adapted for solving this problem for a general convex function E.

One advantage of a greedy algorithm is that it naturally produces a sparse minimizer, which is often a desirable property (for example, in statistical classification some form of regularization or sparsification is often used to prevent model overfitting). Moreover, since greedy algorithms are iterative, we control the trade-off between accuracy and sparsity, and can obtain the optimal solution for the current minimization problem.

Another benefit of the greedy approach is that often the usual methods of convex optimization depend on the dimensionality of the space (see e.g. Nemirovski 1995), which makes them not preferable for general use, while greedy algorithms are designed to work in infinite-dimensional spaces, thus naturally eliminating the problem of dimensionality.

An adaptation of X-greedy algorithms for convex minimization is especially interesting since such algorithms would not require the derivative of the objective function E, unlike traditional methods such as gradient descent, the Frank–Wolfe algorithm, and their modifications.

The first X-greedy algorithms for convex minimization — the Relaxed E-Greedy Algorithm (REGA(co)) and E-Greedy Algorithm with Free Relaxation (EGAFR(co)) — were introduced in DeVore and Temlyakov 2014.

Definition (REGA(co))**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n \in [0; 1]$ that

$$\{\phi_n, \lambda_n\} = \operatorname*{argmin}_{\substack{\phi \in \mathcal{D} \\ 0 \leq \lambda \leq 1}} E((1-\lambda)x_{n-1} + \lambda\phi),$$

2. set $x_n = (1-\lambda_n)x_{n-1} + \lambda_n\phi_n$.

Definition (EGAFR(co))**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n, \mu_n \in \mathbb{R}$ that

$$\{\phi_n, \lambda_n, \mu_n\} = \operatorname*{argmin}_{\substack{\phi \in \mathcal{D} \\ \lambda, \mu \in \mathbb{R}}} E(\mu x_{n-1} + \lambda\phi),$$

2. set $x_n = \mu_n x_{n-1} + \lambda_n\phi_n$.

From the definitions of these algorithms it is easy to see that the REGA(co) is naturally limited to the convex hull of the dictionary $\mathcal{D}$. The EGAFR(co) does not suffer this limitation but is more computationally challenging since the minimization is performed by two variables on each iteration. It is therefore desirable to obtain an algorithm that combines the computational simplicity of the REGA(co) with the unrestricted nature of the EGAFR(co).

In Chapter 5 we propose an algorithm that possesses these properties. Specifically, we introduce the Rescaled Relaxed E-Greedy Algorithm (RREGA(co)) — a new E-greedy algorithm which performs an additional rescaling step on each iteration.

Definition (RREGA(co))**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n \in \mathbb{R}$ that

$$\{\phi_n, \lambda_n\} = \operatorname*{argmin}_{\substack{\phi \in \mathcal{D} \\ \lambda \in \mathbb{R}}} E(x_{n-1} + \lambda \phi),$$

2. choose such $\mu_n \geq 0$ that

$$\mu_n = \operatorname*{argmin}_{\mu \geq 0} E(\mu(x_{n-1} + \lambda_n \phi_n)),$$

3. set $x_n = \mu_n \left(x_{n-1} + \lambda_n \phi_n\right)$.

As in greedy approximation, the algorithms in convex optimization might be computationally challenging or even unfeasible due to possible difficulties in evaluating the objective function E and/or choosing the next atom ϕ_m. Hence, it is natural to consider simplified versions of the stated algorithms which allow inexact step evaluations. Since the setting of convex optimization is more general than that of greedy approximation, it is possible that the objective function E takes negative values, and therefore it is preferable to consider the absolute errors rather than the relative ones. Such approximate versions of the REGA(co) and EGAFR(co) (REGA$\{\delta_n\}$ and EGAFR$\{\delta_n\}$ respectively) were considered in Temlyakov 2016.

Definition (REGA$\{\delta_n\}$)**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n \in [0; 1]$ that

$$E((1-\lambda_n)x_{n-1} + \lambda_n \phi_n) \leq \min_{\substack{\phi \in \mathcal{D} \\ 0 \leq \lambda \leq 1}} E((1-\lambda)x_{n-1} + \lambda \phi) + \delta_n,$$

2. set $x_n = (1-\lambda_n)x_{n-1} + \lambda_n \phi_n$.

Definition (EGAFR$\{\delta_n\}$)**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n, \mu_n \in \mathbb{R}$ that

$$E(\mu_n x_{n-1} + \lambda_n \phi_n) \leq \min_{\substack{\phi \in \mathcal{D} \\ \lambda, \mu \in \mathbb{R}}} E(\mu x_{n-1} + \lambda \phi) + \delta_n,$$

2. set $x_n = \mu_n x_{n-1} + \lambda_n \phi_n$.

We propose a simplified version of the RREGA — the Approximate Rescaled Relaxed E-Greedy Algorithm (ARREGA(co)), in which we are allowed to perform the choice of the next atom ϕ_m and the rescaling parameter μ_m with an absolute inaccuracy. For simplicity we consider the same inaccuracy δ_m for the two steps of the ARREGA; however the similar results follow for any version of the ARREGA with minor changes in proofs.

Definition (ARREGA(co))**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n \in \mathbb{R}$ that

$$E\left(x_{n-1} + \lambda_n \phi_n\right) \leq \inf_{\phi \in \mathcal{D}, \lambda \in \mathbb{R}} E\left(x_{n-1} + \lambda \phi\right) + \delta_n,$$

2. find such $\mu_n \geq 0$ that

$$E(\mu_n(x_{n-1} + \lambda_n \phi_n)) \leq \min_{\mu \geq 0} E(\mu(x_{n-1} + \lambda_n \phi_n)) + \delta_n,$$

3. set $x_n = \mu_n(x_{n-1} + \lambda_n \phi_n)$.

We show that the stated algorithms converge if $\delta_n \to 0$ as $n \to \infty$. Moreover, we establish exactly how these inaccuracies affect the rate of convergence of the ARREGA(co).

Additionally, we demonstrate the behavior of the REGA(co), the EGAFR(co), and the RREGA(co) on a few practical examples.

Chapter 2

Preliminaries

In this chapter we introduce the relevant definitions and results that will be used throughout the dissertation.

Let $(X, \|\cdot\|)$ be a real Banach space. By S_X and B_X we denote the unit sphere and the closed unit ball of X respectively, i.e.

$$S_X = \{x \in X : \|x\| = 1\} \quad \text{and} \quad B_X = \{x \in X : \|x\| \leq 1\}.$$

A dictionary $\mathcal{D}$ is a set of elements of X such that $\overline{\operatorname{span}}\, \mathcal{D} = X$ and elements of $\mathcal{D}$ are normalized, i.e. $\|g\| = 1$ for any $g \in \mathcal{D}$. For convenience we assume that all dictionaries are symmetric, i.e. if $g \in \mathcal{D}$ then $-g \in \mathcal{D}$. Conventionally, the elements of a dictionary are called atoms. By $A_1(\mathcal{D})$ we denote the closure of the convex hull of a dictionary $\mathcal{D}$, and by $A_0(\mathcal{D})$ we denote all the linear combinations of the elements of a dictionary $\mathcal{D}$.

For any non-zero element $x \in X$, let F_x denote a norming functional of x, i.e. such a functional that $\|F_x\|_{X^*} = 1$ and $F_x(x) = \|x\|$. The existence of such a functional is guaranteed by the Hahn-Banach theorem. In particular, it is easy to see that in a Hilbert space $(H, \langle \cdot, \cdot \rangle)$ for any $x \in H \setminus \{0\}$ and $y \in H$

$$F_x(y) = \frac{\langle x, y \rangle}{\|x\|},$$

in $(\ell_p, \|\cdot\|_p)$ for $1 < p < \infty$

$$F_x(y) = \frac{\sum \operatorname{sgn}(x_n) |x_n|^{p-1}\, y_n}{\|x\|_p^{p-1}},$$

and in a general Banach space $(X, \|\cdot\|)$

$$F_x(y) = \lim_{t \to 0} \frac{\|x + ty\| - \|x\|}{t}.$$

A function $E : X \to \mathbb{R}$ is convex if for any $x, y \in X$ and $t \in [0; 1]$

$$E(tx + (1-t)y) \leq tE(x) + (1-t)E(y).$$

We say that H_x is a support functional for E at $x \in X$ if for any $y \in X$

$$H_x(y) \leq E(x + y) - E(x).$$

If E is a convex function, a support functional exists at any point $x \in X$.

We say that a function $E : X \to \mathbb{R}$ is Gâteaux-differentiable at $x \in X$ if there is a bounded linear function $E'_x : X \to \mathbb{R}$ such that for any $y \in S_X$

$$E'_x(y) = \frac{d}{dt} E(x + ty)\Big|_{t=0} = \lim_{t \to 0} \frac{E(x + ty) - E(x)}{t}, \qquad (2.1)$$

where $E'_x(y)$ is called the Gâteaux derivative of E at x in direction y. In that case, the support functional H_x is unique and

$$H_x = E'_x.$$

A function E is Gâteaux-differentiable on $X_0 \subset X$ if it is Gâteaux-differentiable at every point $x \in X_0$.

An element $x \in X$ is a point of (Gâteaux) smoothness of X if the norm $\|\cdot\|$ is Gâteaux differentiable at x. In that case

$$\frac{d}{dt} \|x + ty\| \Big|_{t=0} = \lim_{t \to 0} \frac{\|x + ty\| - \|x\|}{t} = F_x(y), \qquad (2.2)$$

i.e. there is a unique norming functional F_x (see e.g. Beauzamy 1982). We say that a Banach space X is (Gâteaux) smooth if every element $x \in X \setminus \{0\}$ is a point of (Gâteaux) smoothness, i.e. for any non-zero x the norming functional F_x is unique.

In particular, the space L_p is smooth for any $1 < p < \infty$, while L_1 and L_∞ are not smooth.

Additionally, we say that a function/norm is Fréchet differentiable if the limit in (2.1)/(2.2) is uniform for every $y \in S_X$.

It is known that the performance of greedy algorithms is tightly connected to the smoothness of a space/function. In particular, the smoothness of a space/function is essential for the convergence of greedy algorithms, but not sufficient. Thus, we introduce a stronger characterization of smoothness.

For a Banach space X, the modulus of smoothness $\rho_X(u)$ is defined by

$$\rho_X(u) = \rho(u, \|\cdot\|, X) = \sup_{\|x\|=\|y\|=1} \frac{\|x+uy\| + \|x-uy\|}{2} - 1. \tag{2.3}$$

Note that the modulus of smoothness is an even and convex function and, therefore, $\rho_X(u)$ is non-decreasing on $(0, \infty)$. A Banach space is uniformly smooth if $\rho_X(u) = o(u)$ as $u \to 0$. We say that the modulus of smoothness $\rho_X(u)$ is of power type $1 \le q \le 2$ if $\rho(u) \le \gamma u^q$ for some $\gamma > 0$. It follows from the definition that every Banach space has a modulus of smoothness of power type 1 and that every Hilbert space has a modulus of smoothness of power type 2.

Denote by $\mathcal{P}_q$ the class of all Banach spaces with the modulus of smoothness of nontrivial power type $1 < q \le 2$. In particular, it is known (see Lemma B.1 from Donahue et al. 1997) that the modulus of smoothness $\rho_p(u)$ of L_p space satisfies

$$\rho_p(u) \le \begin{cases} \frac{1}{p} u^p & 1 < p \le 2 \\ \frac{p-1}{2} u^2 & 2 \le p < \infty \end{cases},$$

hence $L_p \in \mathcal{P}_q$, where $q = \min\{p; 2\}$.

For functions on Banach spaces, the notion of uniform smoothness is slightly different. For convenience, we will restrict ourselves to convex functions.

The modulus of smoothness $\rho(u, E, S)$ of a convex function $E : X \to \mathbb{R}$ on a convex set $S \subset X$ is defined as follows:

$$\rho(u, E, S) = \sup_{\substack{x \in S, \\ y \in S_X}} \frac{E(x+uy) + E(x-uy)}{2} - E(x).$$

The function E is uniformly smooth on S if $\rho(u, E, S) = o(u)$ as $u \to 0$. We say that the modulus of smoothness $\rho(u, E, S)$ is of power type $1 \le q \le 2$ if $\rho(u, E, S) \le \gamma u^q$ for some $\gamma > 0$.

We note that, in comparison with the modulus of smoothness of a norm, the modulus of smoothness of a function additionally depends on the chosen set S. That is because a norm is a positive homogeneous function, thus its smoothness on the whole space is defined by its smoothness on the unit sphere, which is not the case for a general convex function.

Denote by $\mathcal{P}_q(S, X)$ the class of all uniformly smooth on $S \subset X$ convex functions with the modulus of smoothness of power type $1 < q \le 2$. Note that $\mathcal{P}_q(S, X)$ is completely different from the class $\mathcal{P}_q$ of uniformly smooth Banach spaces with the norms of nontrivial power type since any uniformly smooth norm $\|\cdot\|_X$ is not uniformly smooth as a function on any convex subset $S \subset X$ containing 0. However, it is shown in Borwein et al. 2009 that if $\|\cdot\|_X \in \mathcal{P}_q$ then $E(\cdot) = \|\cdot\|_X^q \in \mathcal{P}_q(S, X)$ for any convex $S \subset X$.

Chapter 3

Chebyshev Greedy Algorithm

In this chapter we introduce the Chebyshev Greedy Algorithm and show that the class of Banach spaces for which the algorithm converges for all dictionaries and objective elements is strictly between smooth and uniformly smooth Banach spaces.

We begin with the definition of the Orthogonal Greedy Algorithm (OGA, see DeVore and Temlyakov 1996), also known as the Orthogonal Matching Pursuit (see Pati, Rezaiifar, and Krishnaprasad 1993). Let $(H, \langle \cdot, \cdot \rangle)$ be a real Hilbert space, $\mathcal{D}$ be a dictionary, and $f \in H$ be an objective element. Then the OGA of f with respect to $\mathcal{D}$ is defined as follows.

Definition (OGA)**.** Set $f_0 = f \in H$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ such that

$$\langle f_{n-1}, \phi_n \rangle = \sup_{g \in \mathcal{D}} \langle f_{n-1}, g \rangle,$$

2. denote $\Phi_n = \operatorname{span}\{\phi_j\}_{j=1}^n$ and take

$$G_n = \operatorname{Proj}_{\Phi_n}(f),$$

3. set $f_n = f - G_n$.

The Chebyshev Greedy Algorithm (see Temlyakov 2001) is a generalization of the OGA to the Banach space setting that utilizes norming functionals to measure how close two elements of a Banach space are. Let $(X, \|\cdot\|)$ be a real Banach space, $\mathcal{D}$ be a dictionary, and $f \in X$ be an objective element. Then the CGA of f with respect to $\mathcal{D}$ is defined as follows.

Definition (CGA)**.** Set $f_0 = f \in X$ and for each $n \geq 1$

1. find any $\phi_n \in \mathcal{D}$ (we assume existence) such that

$$F_{f_{n-1}}(\phi_n) = \sup_{g \in \mathcal{D}} F_{f_{n-1}}(g),$$

2. denote $\Phi_n = \operatorname{span}\{\phi_j\}_{j=1}^n$ and find any $G_n \in \Phi_n$ satisfying

$$\|f - G_n\| = \inf_{G \in \Phi_n} \|f - G\|,$$

3. set $f_n = f - G_n$.

Note that if X is a Hilbert space then the CGA coincides with the OGA. We say that the CGA of f converges if every realization of the algorithm provides a sequence of approximations $\{G_n\}_{n=1}^\infty$ that converges to f. Conversely, we say that the CGA diverges if there exists such a realization that $G_n \not\to f$ as $n \to \infty$.

We also note that these algorithms are largely theoretical since an element selected on the first step might not exist in a general case. Moreover, we cannot expect that in practice operations like finding norming functionals and approximants are exact on all steps. For that reason we consider a more application-oriented version of the CGA — the gAWCGA, which we discuss in details in chapter 4.

We first recall known results on convergence of the CGA. It is shown in Temlyakov 2001 that the CGA converges in all uniformly smooth Banach spaces for all dictionaries and all objective elements of the space. However, the uniform smoothness of the space is not necessary: it is shown in Dilworth, Kutzarova, and Temlyakov 2002 that every separable reflexive Banach space X admits an equivalent norm for which the CGA converges for any dictionary $\mathcal{D}$ and any element $f \in X$. Furthermore, one can find a separable reflexive Banach space that does not admit an equivalent uniformly smooth norm (e.g. see Beauzamy 1982). Thus, the condition of uniform smoothness of a space can be weakened. In particular, it is shown in Dilworth, Kutzarova, and

Temlyakov 2002 that if a reflexive Banach space X has the Kadec-Klee property and a Fréchet differentiable norm, then the CGA converges for any dictionary $\mathcal{D}$ and any element $f \in X$. Thus, uniform smoothness of the space is sufficient but not necessary for the convergence of the CGA.

On the other hand, it is shown in Dubinin 1997 that the smoothness of the space is equivalent to a decrease of norms of the remainders of the CGA for any dictionary $\mathcal{D}$ and any element $f \in X$; thus it may seem that the smoothness might be the necessary and sufficient condition. We disprove that hypothesis by constructing an example of a smooth Banach space, a dictionary, and an element, for which the CGA diverges. For completeness we will show the necessity of smoothness of the space as well.

3.1 The Necessity of Smoothness

In this section we justify the necessity of smoothness of a space for the convergence of the CGA. The following proposition shows that if X is not smooth, then for some dictionary $\mathcal{D}$ and some function f, the CGA of f does not converge even if f is a finite linear combination of the elements of the dictionary.

Proposition 3.1.1. *In any non-smooth Banach space X there exists a dictionary $\mathcal{D}$ and an element $f \in A_0(\mathcal{D})$ such that the CGA of f does not converge to f.*

Proof. Since X is not smooth, there exists an element $f \in S_X$ with two norming functionals F and F' such that $F \not\equiv F'$. Then there exists an element $g \in X$ such that $F(g) \neq F'(g)$. Without loss of generality assume that $F(g) > F'(g)$. Denote

$$g_0 = \alpha_0 \left(g - \frac{F(g) + F'(g)}{2} f \right) \quad \text{and} \quad g_1 = \alpha_1 \left(g - F(g) f \right), \tag{3.1}$$

where $\alpha_0 = \left\| g - \frac{F(g)+F'(g)}{2} f \right\|^{-1}$ and $\alpha_1 = \left\| g - F(g) f \right\|^{-1}$. Note that $F(g_0) > 0$ and

$F'(g_0) < 0$. Let $\{e_j\}_{j\in\Lambda}$ be a dictionary in X. Consider the set of indices

$$\Lambda' = \left\{ j \in \Lambda : e_j - \frac{F(e_j)}{F(g_0)} g_0 \neq 0 \right\}.$$

Define for any $j \in \Lambda'$

$$e'_j = \beta_j \left(e_j - \frac{F(e_j)}{F(g_0)} g_0 \right), \quad \text{where} \quad \beta_j = \left\| e_j - \frac{F(e_j)}{F(g_0)} g_0 \right\|^{-1}. \tag{3.2}$$

We claim that $\mathcal{D} = \{\pm g_0, \pm g_1\} \cup \{\pm e'_j\}_{j\in\Lambda'}$ is a dictionary as well. Indeed, take any $h \in X$ and pick any $\epsilon > 0$. Then, since $\{e_j\}_{j\in\Lambda}$ is a dictionary, there exist coefficients $\{a_j\}_{j\in\Lambda}$ such that $\left\| h - \sum_{j\in\Lambda} a_j e_j \right\| < \epsilon$. Since

$$\begin{aligned}
\sum_{j\in\Lambda} a_j e_j &= \sum_{j\in\Lambda'} a_j \left(\beta_j^{-1} e'_j + \frac{F(e_j)}{F(g_0)} g_0 \right) + \sum_{j\in\Lambda\setminus\Lambda'} a_j \frac{F(e_j)}{F(g_0)} g_0 \\
&= \left(\sum_{j\in\Lambda} a_j \frac{F(e_j)}{F(g_0)} \right) g_0 + \sum_{j\in\Lambda'} \frac{a_j}{\beta_j} e'_j \\
&= \frac{F(h)}{F(g_0)} g_0 + \sum_{j\in\Lambda'} \frac{a_j}{\beta_j} e'_j,
\end{aligned}$$

then $h \in \overline{\operatorname{span} \mathcal{D}}$, and $\mathcal{D}$ is a dictionary. Note that $f \in \operatorname{span}\{g_0, g_1\}$, and thus $f \in A_0(\mathcal{D})$. However, we claim that element g_0 does not approximate f, i.e.

$$\operatorname*{argmin}_{\mu\in\mathbb{R}} \|f - \mu g_0\| = 0.$$

Indeed, for any $\mu > 0$

$$\|f + \mu g_0\| \geq F(f + \mu g_0) = 1 + \mu F(g_0) > \|f\|,$$

$$\|f - \mu g_0\| \geq F'(f - \mu g_0) = 1 - \mu F'(g_0) > \|f\|.$$

Additionally, the choice of the elements (3.1) and (3.2) of the dictionary $\mathcal{D}$ provides

$$\begin{aligned}
&F(g_0) > 0, \\
&F(g_1) = 0, \\
&F(e'_j) = 0, \quad \text{for any } j \in \Lambda'.
\end{aligned}$$

Then consider the following realization of CGA of f: for any $n \geq 1$ choose $\phi_n = g_0$ and $f_n = f$. Hence $\|f_n\| \not\to 0$ and CGA does not converge. $\square$

3.2 The Insufficiency of Smoothness

In this section we prove the insufficiency of smoothness of a space for the convergence of the CGA. Concretely, we demonstrate a smooth Banach space, a dictionary, and an element, for which the CGA diverges.

To construct the desired Banach space, we adopt the technique that was used in Donahue et al. 1997 for proving the necessity of smoothness of a space for the convergence of the incremental approximation. Namely, we re-norm ℓ_1 space by introducing the sequence of recursively defined semi-norms $\{\vartheta_n\}_{n=1}^{\infty}$, each of which is the ℓ_{p_n}-norm of the previously calculated semi-norm ϑ_{n-1} and the n-th coordinate of the element, where the sequence $\{p_n\}_{n=1}^{\infty}$ decreases to 1 sufficiently fast. The reason for such a complicated approach is that the constructed space must be smooth but not uniformly smooth, which is already a non-trivial task. We note that an analogous space was used in Livshitz 2003 to prove the insufficiency of smoothness of a space for the convergence of the X-Greedy Algorithm.

Let $\{p_n\}_{n=1}^{\infty}$ be a such non-increasing sequence that $p_n > 1$ for any $n \geq 1$, and

$$\sum_{n=1}^{\infty} \left(1 - \frac{1}{p_n}\right) < \infty. \tag{3.3}$$

Let $\{e_n\}_{n=1}^{\infty}$ be the canonical basis in ℓ_1. Consider a sequence of non-linear functionals $\{\vartheta_n\}_{n=0}^{\infty}$ defined as follows: for any $x = \sum_{n=1}^{\infty} x_n e_n \in \ell_1$

$$\vartheta_0(x) = 0, \text{ and for any } n \geq 1$$
$$\vartheta_n(x) = \left(\vartheta_{n-1}^{p_n}(x) + |x_n|^{p_n}\right)^{1/p_n}.$$

In particular,

$$\vartheta_1(x) = |x_1|,$$
$$\vartheta_2(x) = \left(|x_1|^{p_2} + |x_2|^{p_2}\right)^{1/p_2},$$
$$\vartheta_3(x) = \left(\left(|x_1|^{p_2} + |x_2|^{p_2}\right)^{p_3/p_2} + |x_3|^{p_3}\right)^{1/p_3}.$$

We claim that ϑ_n is a norm on ℓ_1^n. Indeed, for any $x \in \ell_1^n$

$$\vartheta_n(x) = 0 \text{ if and only if } x = 0,$$

$$\vartheta_n(\lambda x) = |\lambda|\vartheta_n(x) \text{ for any } \lambda \in \mathbb{R}.$$

We prove triangle inequality for $\vartheta_n(\cdot)$ using induction by n. The base case $n = 1$ is obvious. Then, using Minkowski's inequality, we obtain for any $n > 1$ and any $x, y \in \ell_1^n$

$$\begin{aligned}
\vartheta_n(x+y) &= \left(\vartheta_{n-1}^{p_n}(x+y) + |x_n + y_n|^{p_n}\right)^{1/p_n} \\
&\le \left((\vartheta_{n-1}(x) + \vartheta_{n-1}(y))^{p_n} + (|x_n| + |y_n|)^{p_n}\right)^{1/p_n} \\
&\le \left(\vartheta_{n-1}^{p_n}(x) + |x_n|^{p_n}\right)^{1/p_n} + \left(\vartheta_{n-1}^{p_n}(y) + |y_n|^{p_n}\right)^{1/p_n} \\
&= \vartheta_n(x) + \vartheta_n(y).
\end{aligned}$$

Define the space X as

$$X = \{x \in \ell_1 : \lim_{n\to\infty} \vartheta_n(x) < \infty\},$$

and the norm $\|\cdot\|_X$ on X as

$$\|x\|_X = \lim_{n\to\infty} \vartheta_n(x).$$

Note that for any $x \in \ell_1$ the sequence $\{\vartheta_n(x)\}_{n=0}^{\infty}$ is non-decreasing, and, therefore, the limit always exists. Moreover, for any $n \ge 1$

$$\vartheta_n(x) \le \vartheta_{n-1}(x) + |x_n| \le \sum_{k=1}^{n} |x_k|,$$

and, by Hölder's inequality,

$$\begin{aligned}
\sum_{k=1}^{n} |x_k| &\le 2^{1-\frac{1}{p_2}} \vartheta_2(x) + \sum_{k=3}^{n} |x_k| \le 2^{1-\frac{1}{p_2}} \left(\vartheta_2(x) + \sum_{k=3}^{n} |x_k|\right) \\
&\le 2^{1-\frac{1}{p_2}} \left(2^{1-\frac{1}{p_3}} \vartheta_3(x) + \sum_{k=4}^{n} |x_k|\right) \le 2^{\sum_{k=2}^{3}\left(1-\frac{1}{p_k}\right)} \left(\vartheta_3(x) + \sum_{k=4}^{n} |x_k|\right) \\
&\dots \\
&\le 2^{\sum_{k=2}^{n-1}\left(1-\frac{1}{p_k}\right)} \left(\vartheta_{n-1}(x) + |x_n|\right) \le 2^{\sum_{k=2}^{n}\left(1-\frac{1}{p_k}\right)} \vartheta_n(x).
\end{aligned}$$

Therefore, by taking the limit by $n \to \infty$, we obtain for any $x \in X$

$$\rho \left\| x \right\|_1 \leq \left\| x \right\|_X \leq \left\| x \right\|_1 , \tag{3.4}$$

where $\rho = 2^{-\sum_{k=1}^{\infty}\left(1-\frac{1}{p_k}\right)} > 0$ by the choice of $\{p_n\}_{n=1}^{\infty}$ (3.3). Hence, the $\|\cdot\|_X$-norm is equivalent to $\|\cdot\|_1$-norm, and $X = (\ell_1, \|\cdot\|_X)$ is a Banach space. We note that while we impose condition (3.3) to obtain the norms equivalence, the weaker restrictions on the rate of decay of $\{p_n\}_{n=1}^{\infty}$ might be used (see Proposition 1 from Dowling et al. 1997).

Next, we show that the constructed space X is smooth. Namely, we prove that for any element $x \in X$ there is a unique norming functional F_x.

First, we establish the dual of X. Let $\{e_n^*\}_{n=1}^{\infty}$ be the canonical basis in ℓ_∞. Consider the sequence of numbers $\{q_n\}_{n=1}^{\infty}$ given by

$$q_n = \frac{p_n}{p_n - 1}.$$

Similarly, we define the sequence of functionals $\{\nu_n\}_{n=0}^{\infty}$ as follows: for any sequence $a = \sum_{n=1}^{\infty} a_n e_n^* \in \ell_\infty$

$$\nu_0(a) = 0, \text{ and for any } n \geq 1$$
$$\nu_n(a) = \left(\nu_{n-1}^{q_n}(a) + |a_n|^{q_n}\right)^{1/q_n} .$$

Consider the space

$$X^* = \{a \in \ell_\infty : \lim_{n\to\infty} \nu_n(a) < \infty\},$$

equipped with the norm

$$\|a\|_{X^*} = \lim_{n\to\infty} \nu_n(a).$$

In the same way as above we show that $\|\cdot\|_{X^*}$-norm and $\|\cdot\|_\infty$-norm are equivalent. For any $n \geq 1$

$$\nu_n(a) \geq \sup_{k\leq n} |a_k|,$$

and

$$\begin{aligned}
\nu_n(a) &= (\nu_{n-1}^{q_n}(a) + |a_n|^{q_n})^{1/q_n} \le 2^{\frac{1}{q_n}} \max\{\nu_{n-1}(a), |a_n|\} \\
&\le 2^{\frac{1}{q_{n-1}}+\frac{1}{q_n}} \max\{\nu_{n-2}(a), |a_{n-1}|, |a_n|\} \\
&\dots \\
&\le 2^{\sum_{k=3}^n \frac{1}{q_k}} \max\{\nu_2(a), |a_3|, \dots, |a_n|\} \\
&\le 2^{\sum_{k=2}^n \frac{1}{q_k}} \max\{|a_1|, |a_2|, \dots, |a_n|\}.
\end{aligned}$$

Therefore, by taking the limit by $n \to \infty$, we obtain for any $a \in X^*$

$$\|a\|_\infty \le \|a\|_{X^*} \le \rho^{-1} \|a\|_\infty ,$$

i.e. the $\|\cdot\|_{X^*}$-norm is equivalent to $\|\cdot\|_\infty$-norm, and $X^* = (\ell_\infty, \|\cdot\|_{X^*})$ is a Banach space.

We claim that X^* is the dual of X. Indeed, for any $x \in X$ and any $a \in X^*$ the Hölder's inequality provides for any $N \in \mathbb{N}$

$$\begin{aligned}
\sum_{n=1}^N |a_n||x_n| &\le \nu_2(a)\vartheta_2(x) + \sum_{n=3}^N |a_n||x_n| \\
&\dots \\
&\le \nu_{N-1}(a)\vartheta_{N-1}(x) + |a_N||x_N| \le \nu_N(a)\vartheta_N(x),
\end{aligned}$$

and therefore

$$|a(x)| = \lim_{N\to\infty} \left|\sum_{n=1}^N a_n x_n\right| \le \lim_{N\to\infty} \sum_{n=1}^N |a_n||x_n| \le \|a\|_{X^*} \|x\|_X .$$

Similarly, using induction we obtain for any functional $a(x) = \sum_{n=1}^\infty a_j x_j$ on X

$$\sup_{x\in S_X} a(x) = \|a\|_{X^*} ,$$

which completes the proof of the claim.

Consider the spaces $X^n = (\ell_1^n, \vartheta_n(\cdot))$ and $X^{*n} = (\ell_\infty^n, \nu_n(\cdot))$ — the initial segments of X and X^* respectively. We use induction to show that for any $n \ge 1$ the space

X^{*n} is strictly convex. Indeed, $X^{*1} = (\mathbb{R}, |\cdot|)$ is strictly convex, and for any $n > 1$

$$X^{*n} = X^{*n-1} \oplus_{q_n} \mathbb{R}$$

is strictly convex as a q_n-sum of strictly convex spaces with $1 < q_n < \infty$ (see, e.g., Beauzamy 1982). Therefore X^n is smooth as a predual of a strictly convex space X^{*n} (e.g. Beauzamy 1982).

Lastly, we need the following technical lemma.

Lemma 3.2.1. *Let $x = \sum_{n=1}^{\infty} x_n e_n$ be an element in X and $F_x = \sum_{n=1}^{\infty} a_n e_n^*$ be a norming functional for x. Then for any $m \in \mathbb{N}$*

$$F_x^m = \frac{\sum_{n=1}^{m} a_n e_n^*}{\nu_m(a)}$$

is a norming functional for $x^m = \sum_{n=1}^{m} x_n e_n \in X^m$.

Proof. Assume that $F_x^m(x^m) < \|x^m\|_{X^m} = \vartheta_m(x)$, i.e. F_x^m is not a norming functional for x^m. Then

$$\sum_{n=1}^{m} a_n x_n < \nu_m(a)\vartheta_m(x)$$

and for any $N > m$ by Hölder's inequality

$$\begin{aligned} F_x(x) = \sum_{n=1}^{\infty} a_n x_n &\leq \sum_{n=1}^{\infty} |a_n||x_n| \\ &< \nu_m(a)\vartheta_m(x) + \sum_{n=m+1}^{\infty} |a_n||x_n| \\ &\leq \nu_N(a)\vartheta_N(x) + \sum_{n=N+1}^{\infty} |a_n||x_n|. \end{aligned}$$

Taking the limit as $N \to \infty$ we get

$$F_x(x) < \|a\|_{X^*} \|x\|_X = \|x\|_X \,,$$

which contradicts $F_x(x) = \|x\|$. □

Finally, we prove the smoothness of X in the following elegant way.

Lemma 3.2.2 (S.J. Dilworth)**.** *The space $X = (\ell_1, \|\cdot\|_X)$ is smooth.*

Proof. Assume that there is an element $x \in X$ with two distinct norming functionals: $F_x = \sum_{n=1}^{\infty} a_n e_n^*$ and $G_x = \sum_{n=1}^{\infty} b_n e_n^*$. Then Lemma 3.2.1 and the smoothness of initial segments provide for any $N \in \mathbb{N}$

$$\frac{\sum_{n=1}^{N} a_n e_n^*}{\nu_N(a)} = F_x^N = G_x^N = \frac{\sum_{n=1}^{N} b_n e_n^*}{\nu_N(b)}.$$

Find such $m \in \mathbb{N}$ that $a_m \neq b_m$. Then, taking the limit as $N \to \infty$ and taking into account that $\|a\|_{X^*} = \|b\|_{X^*} = 1$, we get

$$a_m = \lim_{N\to\infty} \frac{a_m}{\nu_N(a)} = \lim_{N\to\infty} F_x^N(e_m) = \lim_{N\to\infty} G_x^N(e_m) = \lim_{N\to\infty} \frac{b_m}{\nu_N(b)} = b_m,$$

which contradicts $a_m \neq b_m$ and thus X is smooth. □

We now need to establish the norming functionals in X. Take any element $x = \sum_{n=1}^{\infty} x_n e_n$ in X and consider a sequence of functionals $\{\mathcal{F}_x^n\}_{n=0}^{\infty}$ defined as follows: for any $y = \sum_{n=1}^{\infty} y_n e_n \in X$

$$\begin{aligned}
&\mathcal{F}_x^0(y) = 0, \text{ and for any } n \geq 1 \\
&\mathcal{F}_x^n(y) = \frac{\vartheta_{n-1}^{p_n-1}(x)\mathcal{F}_x^{n-1}(y) + \operatorname{sgn} x_n |x_n|^{p_n-1} y_n}{\vartheta_n^{p_n-1}(x)} \\
&\qquad\quad = \vartheta_n^{1-p_{n+1}}(x) \sum_{k=1}^{n} \left(\operatorname{sgn} x_k |x_k|^{p_k-1} y_k \prod_{j=k}^{n} \vartheta_j^{p_{j+1}-p_j}(x) \right).
\end{aligned}$$

Lemma 3.2.3. *Let $x = \sum_{n=1}^{m} x_n e_n$ be an element in X. Then $\mathcal{F}_x^m$ is the norming functional for x.*

Proof. We will use induction by m. For $m = 1$

$$\mathcal{F}_x^1(y) = \operatorname{sgn} x_1 \, y_1,$$

and $\mathcal{F}_x^1(x) = \vartheta_1(x) = \|x\|_X$, $|\mathcal{F}_x^1(y)| = \vartheta_1(y) = \|y\|_X$. For $m > 1$

$$\mathcal{F}_x^m(y) = \frac{\vartheta_{m-1}^{p_m-1}(x)\mathcal{F}_x^{m-1}(y) + \operatorname{sgn} x_m |x_m|^{p_m-1} y_m}{\vartheta_m^{p_m-1}(x)}.$$

Then

$$\mathcal{F}_x^m(x) = \frac{\vartheta_{m-1}^{p_m}(x) + |x_m|^{p_m}}{\vartheta_m^{p_m-1}(x)} = \vartheta_m(x) = \|x\|_X \, ,$$

and induction hypothesis and Hölder's inequality provide

$$\begin{aligned} |\mathcal{F}_x^m(y)| &\le \frac{\vartheta_{m-1}^{p_m-1}(x)|\mathcal{F}_x^{m-1}(y)| + |x_m|^{p_m-1}|y_m|}{\vartheta_m^{p_m-1}(x)} \\ &\le \frac{\vartheta_{m-1}^{p_m-1}(x)\vartheta_{m-1}(y) + |x_m|^{p_m-1}|y_m|}{\vartheta_m^{p_m-1}(x)} \\ &\le (\vartheta_{m-1}^{p_m}(y) + |y_m|^{p_m})^{1/p_m} = \vartheta_n(y) = \|y\|_X \, . \end{aligned}$$

□

Thus, we have established the norming functionals $\mathcal{F}_n$ in the initial segments X^n. In particular, for any $x, y \in X$

$$\begin{aligned} \mathcal{F}_x^1(y) &= \operatorname{sgn} x_1 \, y_1, \\ \mathcal{F}_x^2(y) &= \frac{\operatorname{sgn} x_1 |x_1|^{p_2-1} y_1 + \operatorname{sgn} x_2 |x_2|^{p_2-1} y_2}{\vartheta_2^{p_2-1}(x)}, \\ \mathcal{F}_x^3(y) &= \frac{(\operatorname{sgn} x_1 |x_1|^{p_2-1} y_1 + \operatorname{sgn} x_2 |x_2|^{p_2-1} y_2) \, \vartheta_2^{p_3-p_2}(x) + \operatorname{sgn} x_3 |x_3|^{p_3-1} y_3}{\vartheta_3^{p_3-1}(x)}. \end{aligned}$$

We now choose a dictionary $\mathcal{D}$ in X and an element $f \in X$ such that CGA of f diverges. Without loss of generality assume $t_n = 1$ for each $n \ge 1$, i.e. an element of the dictionary that maximizes $F_{f_{n-1}}$ is chosen on each step. Let

$$\begin{aligned} g_0 &= e_1 + e_2 + e_3, \\ g_k &= e_k + e_{k+1} \text{ for each } k \ge 1, \end{aligned}$$

and take $\mathcal{D} = \{\pm g_n / \|g_n\|_X\}_{n=0}^{\infty}$. Note that for any $k \ge 1$

$$\|g_k\|_X = 2^{1/p_{k+1}} \le 2^{1/p_2} < \left(1 + 2^{p_3/p_2}\right)^{1/p_3} = \|g_0\|_X \, . \tag{3.5}$$

Take $f = e_1 \in X$, then $f = g_0 - g_2 \in A_0(\mathcal{D})$. We will show that the CGA diverges even for such a simple element. We claim that for any $m \ge 1$

$$\phi_m = \pm g_m / \|g_m\|_X \, , \tag{3.6}$$

where by $\pm$ we understand some sign — plus or minus. We will prove this claim using induction by m.

Consider the case $m = 1$. Lemma 3.2.3 provides $F_f = \mathcal{F}_f^1$, thus

$$|\mathcal{F}_f^1(g_0)| = 1,$$
$$|\mathcal{F}_f^1(g_1)| = 1,$$
$$|\mathcal{F}_f^1(g_k)| = 0 \text{ for any } k > 1.$$

Then estimate (3.5) guarantees that $\phi_1 = \pm g_1 / \|g_1\|_X$.

Consider the case $m > 1$. By induction hypothesis the elements

$$\pm g_1 / \|g_1\|_X, \pm g_2 / \|g_2\|_X, \dots, \pm g_{m-1} / \|g_{m-1}\|_X$$

were chosen on previous steps. Then $f_{m-1} = \sum_{n=1}^m c_n e_n$ for some coefficients $\{c_n\}_{n=1}^m$, and therefore $F_{f_{m-1}} = \mathcal{F}_{f_{m-1}}^m$ by Lemma 3.2.3. Note that $f_{m-1} \in X^m$, which is a uniformly smooth space since it is smooth and finitely-dimensional. Hence, applying Lemma G we obtain that $F_{f_{m-1}}(g_k) = 0$ for any $k = 1, \dots, m-1$, i.e.

$$\mathcal{F}_{f_{m-1}}^m(g_1) = \frac{\operatorname{sgn} c_1 |c_1|^{p_2-1} + \operatorname{sgn} c_2 |c_2|^{p_2-1}}{\vartheta_2^{p_2-1}(f_{m-1}) \dots \vartheta_m^{p_m-1}(f_{m-1})} = 0,$$
$$\mathcal{F}_{f_{m-1}}^m(g_2) = \frac{\operatorname{sgn} c_2 |c_2|^{p_2-1} \vartheta_2^{p_3-p_2}(f_{m-1}) + \operatorname{sgn} c_3 |c_3|^{p_3-1}}{\vartheta_3^{p_3-1}(f_{m-1}) \dots \vartheta_m^{p_m-1}(f_{m-1})} = 0,$$
$$\dots$$
$$\mathcal{F}_{f_{m-1}}^m(g_{m-1}) = \frac{\operatorname{sgn} c_{m-1} |c_{m-1}|^{p_{m-1}-1} \vartheta_{m-1}^{p_m-p_{m-1}}(f_{m-1}) + \operatorname{sgn} c_m |c_m|^{p_m-1}}{\vartheta_m^{p_m-1}(f_{m-1})} = 0.$$

From these equalities we derive

$$|c_2|^{p_2-1} = |c_1|^{p_2-1},$$
$$|c_3|^{p_3-1} = |c_2|^{p_2-1} \vartheta_2^{p_3-p_2}(f_{m-1}),$$
$$\dots$$
$$|c_m|^{p_m-1} = |c_{m-1}|^{p_{m-1}-1} \vartheta_{m-1}^{p_m-p_{m-1}}(f_{m-1}),$$

which imply that for any $k = 3, \dots, m$

$$|c_k|^{p_k - 1} = |c_1|^{p_2 - 1} \prod_{n=2}^{k-1} \vartheta_n^{p_{n+1} - p_n}(f_{m-1}). \tag{3.7}$$

Therefore

$$|\mathcal{F}_{f_{m-1}}^m(g_0)| = |\mathcal{F}_{f_{m-1}}^m(g_0 - g_1)| = \vartheta_m^{1 - p_{m+1}}(f_{m-1}) \left(|c_3|^{p_3 - 1} \prod_{j=3}^{m} \vartheta_j^{p_{j+1} - p_j}(f_{m-1}) \right),$$

$$|\mathcal{F}_{f_{m-1}}^m(g_m)| = \frac{|c_m|^{p_m - 1}}{\vartheta_m^{p_m - 1}(f_{m-1})},$$

$$|\mathcal{F}_{f_{m-1}}^m(g_k)| = 0 \text{ for any } k \in \mathbb{N} \setminus \{m\}.$$

Thus, by (3.7)

$$|\mathcal{F}_{f_{m-1}}^m(g_0)| = \vartheta_m^{1 - p_{m+1}}(f_{m-1}) \left(|c_1|^{p_2 - 1} \prod_{j=2}^{m} \vartheta_j^{p_{j+1} - p_j}(f_{m-1}) \right) = |\mathcal{F}_{f_{m-1}}^m(g_m)|,$$

and estimate (3.5) guarantees that $\phi_m = \pm g_m / \|g_m\|_X$, which completes the proof of assumption (3.6).

Hence, the element $\pm g_0 / \|g_0\|_X$ will not be chosen and $\Phi_n = \operatorname{span}\{g_1, \dots, g_n\}$ for any $n \geq 1$. Then the equivalence of the norms (3.4) provides

$$\|f_n\|_X = \inf_{G \in \Phi_n} \|f - G\|_X \geq \rho \inf_{G \in \Phi_n} \|f - G\|_1 = \rho \not\to 0 \text{ as } n \to \infty,$$

i.e. the CGA of f diverges.

Chapter 4

Generalized Approximate Weak Chebyshev Greedy Algorithm

In Chapter 3 we introduced the Chebyshev Greedy Algorithm and discussed the class of Banach spaces for which the algorithm converges. Specifically, we established that this class is strictly between smooth and uniformly smooth Banach spaces.

In this chapter we introduce the generalized Approximate Weak Chebyshev Greedy Algorithm — an application-oriented modification of the CGA — and analyze its convergence in uniformly smooth Banach spaces. In the gAWCGA it is allowed on every step of the algorithm to pick a sub-optimal element of the dictionary as well as to perform all calculations with some controlled inaccuracies (in term of both absolute and relative errors), thus making the realization of the algorithm always possible, as well as making it computationally easier.

We define the following sequences, which represent the inaccuracies in calculating the steps of the gAWCGA. A weakness sequence $\{(t_n, t'_n)\}_{n=1}^{\infty}$ (represents inaccuracies in choosing atoms $\{\phi_n\}_{n=1}^{\infty}$) is such that $0 \leq t_n \leq 1$ and $t'_n \geq 0$ for all $n \geq 1$. A perturbation sequence $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$ (represents inaccuracies in computing norming functionals $\{F_n\}_{n=0}^{\infty}$) is such that $\delta_n \geq 0$ and $\delta'_n \geq 0$ for all $n \geq 0$. An error sequence $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ (represents inaccuracies in computing approximations $\{G_n\}_{n=1}^{\infty}$) is such that $\eta_n \geq 0$ and $\eta'_n \geq 0$ for all $n \geq 1$. By η_∞ and η'_∞ we denote the least upper bounds of the sequences $\{\eta_n\}_{n=1}^{\infty}$ and $\{\eta'_n\}_{n=1}^{\infty}$, respectively.

For a Banach space X, a dictionary $\mathcal{D}$, and an element $f \in X$, the generalized Approximate Weak Chebyshev Greedy Algorithm with a weakness sequence $\{(t_n, t'_n)\}_{n=1}^{\infty}$, a perturbation sequence $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$, and an error sequence $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ is defined as follows.

Definition (gAWCGA)**.** Set $f_0 = f$ and for each $n \geq 1$

1. take any functional F_{n-1} satisfying

$$\|F_{n-1}\| \leq 1 \quad \text{and} \quad F_{n-1}(f_{n-1}) \geq (1 - \delta_{n-1}) \|f_{n-1}\| - \delta'_{n-1}, \tag{4.1}$$

2. find any $\phi_n \in \mathcal{D}$ such that

$$F_{n-1}(\phi_n) \geq t_n \sup_{g \in \mathcal{D}} F_{n-1}(g) - t'_n, \tag{4.2}$$

3. denote $\Phi_n = \operatorname{span}\{\phi_j\}_{j=1}^{n}$ and find any $G_n \in \Phi_n$ satisfying

$$\|f - G_n\| \leq (1 + \eta_n) \inf_{G \in \Phi_n} \|f - G\| + \eta'_n, \tag{4.3}$$

4. set $f_n = f - G_n$.

Note that if for every $n \geq 1$ either $t_n < 1$ or $t'_n > 0$ then there exists a possible realization of the algorithm for any Banach space X, any dictionary $\mathcal{D}$, and any element $f \in X$. We say that the gAWCGA of f converges if every realization of the algorithm provides a sequence of approximations $\{G_n\}_{n=1}^{\infty}$ that converges to f. Conversely, we say that the gAWCGA diverges if there exists such a realization that $G_n \not\to f$ as $n \to \infty$.

If there are no inaccuracies, i.e. $t_n = 1$ and $t'_n = \delta_{n-1} = \delta'_{n-1} = \eta_n = \eta'_n = 0$ for all $n \geq 1$ then the gAWCGA coincides with the CGA. Note also that if $t'_n = \delta_{n-1} = \delta'_{n-1} = \eta_n = \eta'_n = 0$ for all $n \geq 1$ then the gAWCGA coincides with the WCGA which was studied in Temlyakov 2001 and Dilworth, Kutzarova, and Temlyakov 2002. In the case $t'_n = \delta'_{n-1} = \eta'_n = 0$ the gAWCGA coincides with the AWCGA which was studied in Temlyakov 2005 and Dereventsov 2016.

4.1 Convergence of the gAWCGA

In this section we investigate the behavior of the gAWCGA in a uniformly smooth Banach space X and obtain the necessary and sufficient conditions on the weakness, perturbation, and error sequences that guarantee the convergence of the gAWCGA for all dictionaries $\mathcal{D} \subset X$ and all elements $f \in X$. We understand the necessity of conditions in the following way: if at least one of the stated conditions does not hold, one can find a uniformly smooth Banach space X, a dictionary $\mathcal{D}$, and an element $f \in X$ such that the gAWCGA of f with the given weakness, perturbation, and error sequences, diverges. We note that in our case such a Banach space and dictionary need not be complicated. In fact, we demonstrate that an example of the divergent gAWCGA can be found even in ℓ_p space with the canonical basis as a dictionary. We also note that while we are interested in the question of strong convergence of the CGA and its modifications, the more general setting was considered in Dilworth, Kutzarova, and Temlyakov 2002.

We begin this section by recalling the known results concerning the convergence of the CGA and its modifications in uniformly smooth Banach spaces.

For a weakness sequence $\{t_n\}_{n=1}^{\infty}$ and a number $0 < \theta \leq 1/2$ we define a sequence of positive numbers $\{\xi_n\}_{n=1}^{\infty}$ which satisfy the equality $\rho(\xi_n) = \theta t_n \xi_n$ for each $n \geq 1$. It is shown in Temlyakov 2001 that if a Banach space is uniformly smooth then for any $0 < \theta \leq 1/2$ the sequence $\{\xi_n\}_{n=1}^{\infty}$ exists and is uniquely determined by the sequence $\{t_n\}_{n=1}^{\infty}$.

The first result states the sufficient conditions for the convergence of the WCGA.

Theorem A (Temlyakov 2001, Theorem 2.1)**.** *The WCGA with a weakness sequence $\{t_n\}_{n=1}^{\infty}$ converges for any uniformly smooth Banach space X, any dictionary $\mathcal{D}$, and any element $f \in X$ if for any $0 < \theta \leq 1/2$*

$$\sum_{n=1}^{\infty} t_n \xi_n = \infty.$$

The next theorem gives the sufficient conditions for the convergence of the AWCGA.

Theorem B (Temlyakov 2005, Theorem 2.2)**.** *The AWCGA with a weakness sequence* $\{t_n\}_{n=1}^{\infty}$*, a perturbation sequence* $\{\delta_n\}_{n=0}^{\infty}$*, and an error sequence* $\{\eta_n\}_{n=1}^{\infty}$ *converges for any uniformly smooth Banach space* X*, any dictionary* $\mathcal{D}$*, and any element* $f \in X$ *if*

$$\eta_\infty = \sup_{n \geq 1} \eta_n < \infty,$$

and if for any $0 < \theta \leq 1/2$ *the following conditions hold:*

$$\sum_{n=1}^{\infty} t_n \xi_n = \infty,$$

$$\delta_n = o(t_n \xi_n),$$

$$\eta_n = o(t_n \xi_n).$$

We will prove the following theorem that states that a similar result holds for the convergence of the gAWCGA with somewhat weaker restrictions on the approximation parameters. Specifically, we require the parameters to be sufficiently small only along some increasing sequence of natural numbers $\{n_k\}_{k=1}^{\infty}$.

Theorem 4.1.1 (Dereventsov 2017, Theorem 1)**.** *The gAWCGA with a weakness sequence* $\{(t_n, t'_n)\}_{n=1}^{\infty}$*, a perturbation sequence* $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$*, and an error sequence* $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ *converges for any uniformly smooth Banach space* X*, any dictionary* $\mathcal{D}$*, and any element* $f \in X$ *if*

$$\eta_\infty = \sup_{n \geq 1} \eta_n < \infty, \quad \lim_{n \to \infty} \eta'_n = 0, \tag{4.4}$$

and if there exists a subsequence $\{n_k\}_{k=1}^{\infty}$ *such that for any* $0 < \theta \leq 1/2$ *the following* *conditions hold:*

$$\sum_{k=1}^{\infty} t_{n_k+1}\xi_{n_k+1} = \infty, \tag{4.5}$$

$$t'_{n_k+1} = o(t_{n_k+1}), \tag{4.6}$$

$$\delta_{n_k} = o(t_{n_k+1}\xi_{n_k+1}), \tag{4.7}$$

$$\delta'_{n_k} = o(t_{n_k+1}\xi_{n_k+1}), \tag{4.8}$$

$$\eta_{n_k} = o(t_{n_k+1}\xi_{n_k+1}), \tag{4.9}$$

$$\eta'_{n_k} = o(t_{n_k+1}\xi_{n_k+1}). \tag{4.10}$$

If the modulus of smoothness of a space is of a nontrivial power type, the previous theorems can be rewritten in a form that states the necessary and sufficient conditions for the convergence.

Theorem C (Temlyakov 2001, Corollary 2.1)**.** *The WCGA with a weakness sequence* $\{t_n\}_{n=1}^{\infty}$ *converges for any uniformly smooth Banach space* $X \in \mathcal{P}_q$*, and any dictionary* $\mathcal{D}$*, any element* $f \in X$ *if and only if*

$$\sum_{n=1}^{\infty} t_n^p = \infty,$$

where $p = q/(q-1)$*.*

The next theorem gives the necessary and sufficient conditions for the convergence of the AWCGA.

Theorem D (Dereventsov 2016, Theorem 1)**.** *The AWCGA with a weakness sequence* $\{t_n\}_{n=1}^{\infty}$*, a perturbation sequence* $\{\delta_n\}_{n=0}^{\infty}$*, and an error sequence* $\{\eta_n\}_{n=1}^{\infty}$ *converges for any uniformly smooth Banach space* $X \in \mathcal{P}_q$*, any dictionary* $\mathcal{D}$*, and any element* $f \in X$ *if and only if*

$$\eta_\infty = \sup_{n\geq 1} \eta_n < \infty,$$

and if there exists a subsequence $\{n_k\}_{k=1}^{\infty}$ *such that the following conditions hold:*

$$\sum_{k=1}^{\infty} t_{n_k+1}^p = \infty,$$

$$\delta_{n_k} = o(t_{n_k+1}^p),$$

$$\eta_{n_k} = o(t_{n_k+1}^p),$$

where $p = q/(q-1)$.

We will prove the following result that states the necessary and sufficient conditions for the convergence of the gAWCGA.

Theorem 4.1.2 (Dereventsov 2017, Theorem 2)**.** *The gAWCGA with a weakness sequence* $\{(t_n, t'_n)\}_{n=1}^{\infty}$, *a perturbation sequence* $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$, *and an error sequence* $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ *converges for any uniformly smooth Banach space* $X \in \mathcal{P}_q$, *any dictionary* $\mathcal{D}$, *and any element* $f \in X$ *if and only if*

$$\eta_\infty = \sup_{n \geq 1} \eta_n < \infty, \quad \lim_{n \to \infty} \eta'_n = 0, \tag{4.11}$$

and if there exists a subsequence $\{n_k\}_{k=1}^{\infty}$ *such that the following conditions hold:*

$$\sum_{k=1}^{\infty} t_{n_k+1}^p = \infty, \tag{4.12}$$

$$t'_{n_k+1} = o(t_{n_k+1}), \tag{4.13}$$

$$\delta_{n_k} = o(t_{n_k+1}^p), \tag{4.14}$$

$$\delta'_{n_k} = o(t_{n_k+1}^p), \tag{4.15}$$

$$\eta_{n_k} = o(t_{n_k+1}^p), \tag{4.16}$$

$$\eta'_{n_k} = o(t_{n_k+1}^p), \tag{4.17}$$

where $p = q/(q-1)$.

The following corollary states that if the weakness parameter $\{t_n\}_{n=1}^{\infty}$ is separated from zero (e.g. $t_n = t > 0$ for all n) then the gAWCGA converges as long as η'_n goes to zero and other inaccuracy parameters go to zero along the same subsequence.

Corollary 4.1.3. *Let* $\liminf_{n\to\infty} t_n > 0$*. Then the gAWCGA with a weakness sequence* $\{(t_n, t'_n)\}_{n=1}^{\infty}$*, a perturbation sequence* $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$*, and a bounded error sequence* $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ *with* $\lim_{n\to\infty} \eta'_n = 0$*, converges for any uniformly smooth Banach space* $X \in \mathcal{P}_q$*, any dictionary* $\mathcal{D}$*, and any element* $f \in X$ *if and only if*

$$\liminf_{n\to\infty} \left(t'_{n+1} + \delta_n + \delta'_n + \eta_n \right) = 0.$$

The last two corollaries state that the conditions for the convergence of the gAWCGA are the same as for the WCGA if inaccuracy sequences are from the ℓ_1 space.

Corollary 4.1.4. *Let* $\{t'_n\}_{n=1}^{\infty} \in \ell_1$*,* $\{\delta_n\}_{n=0}^{\infty} \in \ell_1$*,* $\{\delta'_n\}_{n=0}^{\infty} \in \ell_1$*,* $\{\eta_n\}_{n=1}^{\infty} \in \ell_1$*, and* $\{\eta'_n\}_{n=1}^{\infty} \in \ell_1$*. Then the gAWCGA with a weakness sequence* $\{(t_n, t'_n)\}_{n=1}^{\infty}$*, a perturbation sequence* $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$*, and a bounded error sequence* $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ *converges for any uniformly smooth Banach space* X*, any dictionary* $\mathcal{D}$*, and any element* $f \in X$ *if for any* $0 < \theta \le 1/2$

$$\sum_{n=1}^{\infty} t_n \xi_n = \infty.$$

Corollary 4.1.5. *Let* $\{t'_n\}_{n=1}^{\infty} \in \ell_1$*,* $\{\delta_n\}_{n=0}^{\infty} \in \ell_1$*,* $\{\delta'_n\}_{n=0}^{\infty} \in \ell_1$*,* $\{\eta_n\}_{n=1}^{\infty} \in \ell_1$*, and* $\{\eta'_n\}_{n=1}^{\infty} \in \ell_1$*. Then the gAWCGA with a weakness sequence* $\{(t_n, t'_n)\}_{n=1}^{\infty}$*, a perturbation sequence* $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$*, and a bounded error sequence* $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ *converges for any uniformly smooth Banach space* $X \in \mathcal{P}_q$*, any dictionary* $\mathcal{D}$*, and any element* $f \in X$ *if and only if*

$$\sum_{n=1}^{\infty} t_n^p = \infty.$$

4.2 Rate of Convergence of the gAWCGA

In this section we analyze the relation between the error sequences and the rate of convergence of the gAWCGA in uniformly smooth Banach spaces. In particular, we give such estimates on the inaccuracy parameters that the convergence rate of the gAWCGA is of the same order as the convergence rate of the WCGA. Moreover, we show the trade-off between inaccuracy parameters and the convergence rate.

It is known that in order to get a nontrivial rate of approximation, an additional requirement has to be imposed on an objective element. Traditionally for this area, we restrict to the elements from the class $A_1(\mathcal{D})$ — the closure of the convex hull of $\mathcal{D}$. We start with the known result for the rate of convergence of the WCGA.

Theorem E (Temlyakov 2001, Theorem 2.2)**.** *Let $X \in \mathcal{P}_q$ be a Banach space and $\mathcal{D}$ be any dictionary. Let $\{t_n\}_{n=1}^{\infty}$ be a weakness sequence. Then for any $f \in A_1(\mathcal{D})$ the WCGA of f satisfies the estimate*

$$\|f_n\| \leq 2(2\gamma)^{1/q} \left(1 + \sum_{k=1}^{n} t_k^p\right)^{-1/p},$$

where $p = q/(q-1)$.

In particular, Theorem E implies that the CGA satisfies the estimate

$$\|f_n\| \leq 2(2\gamma)^{1/q} n^{-1/p}. \tag{4.18}$$

The next result states the rate of convergence of an adaptive AWCGA, where adaptive means that the perturbation and error sequences are determined by the AWCGA applied to a given element $f \in A_1(\mathcal{D})$. This theorem gives such an estimate on inaccuracy parameters that the convergence rate of the AWCGA is the same as of the WCGA.

Theorem F (Temlyakov 2005, Theorem 2.4)**.** *Let $X \in \mathcal{P}_q$ be a Banach space and $\mathcal{D}$ be any dictionary. Let $\{t_n\}_{n=1}^{\infty}$ be a weakness sequence, $\{\delta_n\}_{n=0}^{\infty}$ be a perturbation sequence, and $\{\eta_n\}_{n=1}^{\infty}$ be an error sequence satisfying*

$$\delta_n = t_{n+1}^p \, \|f_n\|^p \, 3^{-p} \left(64 \, (8\gamma)^{p/q}\right)^{-1}, \quad n \geq 0; \tag{4.19}$$

$$\eta_n = t_{n+1}^p E_n^p 3^{-p} \left(64 \, (8\gamma)^{p/q}\right)^{-1}, \quad n \geq 1, \tag{4.20}$$

where $p = q/(q-1)$. *Then for any* $f \in A_1(\mathcal{D})$ *the AWCGA of* f *satisfies the estimate*

$$\|f_n\| \leq 8\gamma^{1/q}\left(1 + \sum_{k=1}^{n} t_k^p\right)^{-1/p}.$$

We prove the following theorem, which is a generalization of Theorem F for the gAWCGA but with conditions (4.19) and (4.20) imposed only on some subsequence $\{n_k\}_{k=1}^{\infty}$. For convenience denote $\eta_0 = \eta_0' = n_0 = 0$.

Theorem 4.2.1. *Let* $X \in \mathcal{P}_q$ *be a Banach space and* $\mathcal{D}$ *be any dictionary. Let* $\{(t_n, t_n')\}_{n=1}^{\infty}$ *be a weakness sequence,* $\{(\delta_n, \delta_n')\}_{n=0}^{\infty}$ *be a perturbation sequence, and* $\{(\eta_n, \eta_n')\}_{n=1}^{\infty}$ *be an error sequence satisfying for some subsequence* $\{n_k\}_{k=0}^{\infty}$

$$t'_{n_k+1} = T_k t_{n_k+1} E_{n_k}, \tag{4.21}$$

$$\delta_{n_k} = D_k t_{n_k+1}^p \min\{\|f_{n_k}\|^p, 1\}, \tag{4.22}$$

$$\delta'_{n_k} = D'_k t_{n_k+1}^p \min\{\|f_{n_k}\|^{p+1}, 1\}, \tag{4.23}$$

$$\eta_{n_k} = H_k t_{n_k+1}^p E_{n_k}^p, \tag{4.24}$$

$$\eta'_{n_k} = H'_k t_{n_k+1}^p E_{n_k}^{p+1}, \tag{4.25}$$

where $p = q/(q-1)$ *and the non-negative sequences* $\{T_k\}_{k=0}^{\infty}$, $\{D_k\}_{k=0}^{\infty}$, $\{D'_k\}_{k=0}^{\infty}$, $\{H_k\}_{k=0}^{\infty}$, *and* $\{H'_k\}_{k=0}^{\infty}$ *are such that for any* $k \geq 0$

$$\alpha_k := (2q\gamma)^{-1/q} p^{-1/p}(1 - T_k - D_k - D'_k) - (3 + H_k + H'_k)(D_k + D'_k + H_k + H'_k)^{1/p} > 0. \tag{4.26}$$

Then for any $f \in A_1(\mathcal{D})$ *the gAWCGA of* f *satisfies for any* $m \geq 0$

$$\|f_{n_m}\| \leq (1 + 3^{-p})\left(1 + \sum_{k=0}^{m-1} \alpha_k^p t_{n_k+1}^p\right)^{-1/p}.$$

Theorem 4.2.1 describes how the error sequences affect the rate of convergence of the gAWCGA and shows the trade-offs between different inaccuracy parameters and the convergence rate. We note that while an estimate on the rate of convergence of the gAWCGA is provided only on the steps $\{n_k\}_{k=0}^{\infty}$, the general estimate for $\|f_n\|$ can be obtained using condition (4.2):

$$\|f_n\| \leq (1+\eta_n)\left(1+\sum_{k=0}^{N-1}\alpha_k^p t_{n_k+1}^p\right)^{-1/p} + \eta_n',$$

where $N = \max\{N \in \mathbb{Z}^+ : n_N \leq n\}$.

Note also that once a subsequence $\{n_k\}_{k=0}^{\infty}$ for which conditions (4.21)–(4.25) hold is found, only the choice of elements ϕ_{n_k} is essential for the rate of convergence, so arbitrary elements ϕ_j can be chosen on other steps.

We state several corollaries that give concrete bounds for the inaccuracy sequences. The following results give an example of such error sequences that the convergence rate of the gAWCGA is of the same order as the one of the CGA (4.18).

Corollary 4.2.2. *Let $X \in \mathcal{P}_q$ be a Banach space and $\mathcal{D}$ be a dictionary. Let $f \in A_1(\mathcal{D})$, $0 < \tau \leq 1$, and the sequences $\{t_n'\}_{n=1}^{\infty}$, $\{(\delta_n, \delta_n')\}_{n=0}^{\infty}$, and $\{(\eta_n, \eta_n')\}_{n=1}^{\infty}$ be such that for any $n \geq 1$*

$$\begin{aligned}
t_{n+1}' &= 12^{-p}\gamma^{1-p}\tau E_n,\\
\delta_n &= .25\, 12^{-p}\gamma^{1-p}\tau^p \min\{\|f_n\|^p, 1\},\\
\delta_n' &= .25\, 12^{-p}\gamma^{1-p}\tau^p \min\{\|f_n\|^{p+1}, 1\},\\
\eta_n &= .25\, 12^{-p}\gamma^{1-p}\tau^p E_n^p,\\
\eta_n' &= .25\, 12^{-p}\gamma^{1-p}\tau^p E_n^{p+1},
\end{aligned}$$

where $p = q/(q-1)$. Then the gAWCGA of f with the weakness sequence $\{(\tau, t_n')\}_{n=1}^{\infty}$, the perturbation sequence $\{(\delta_n, \delta_n')\}_{n=0}^{\infty}$, and the error sequence $\{(\eta_n, \eta_n')\}_{n=1}^{\infty}$ satisfies for any $n \geq 1$ the estimate

$$\|f_n\| \leq \frac{16\gamma^{1/q}}{\tau} n^{-1/p}.$$

The following corollary is similar to the previous one but imposes restrictions on the sequences t_n', δ_n, and δ_n' only on some subsequence $\{n_k\}_{k=1}^{\infty}$ with bounded increments. Thus, the gAWCGA will converge with the same rate as the CGA as long as adequately precise computations are made sufficiently often.

Corollary 4.2.3. *Let $X \in \mathcal{P}_q$ be a Banach space and $\mathcal{D}$ be a dictionary. Let $f \in A_1(\mathcal{D})$, $0 < \tau \le 1$, and the sequences $\{t'_n\}_{n=1}^{\infty}$, $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$, and $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ be such that for some subsequence $\{n_k\}_{k=1}^{\infty}$ with $M = \sup_{k \in \mathbb{N}} |n_{k+1} - n_k| < \infty$*

$$
\begin{aligned}
t'_{n_k+1} &= 12^{-p}\gamma^{1-p}\tau E_{n_k}, \\
\delta_{n_k} &= .25\, 12^{-p}\gamma^{1-p}\tau^p \min\{\|f_{n_k}\|^p, 1\}, \\
\delta'_{n_k} &= .25\, 12^{-p}\gamma^{1-p}\tau^p \min\{\|f_{n_k}\|^{p+1}, 1\}, \\
\eta_k &= .25\, 12^{-p}\gamma^{1-p}\tau^p E_k^p, \\
\eta'_k &= .25\, 12^{-p}\gamma^{1-p}\tau^p E_k^{p+1},
\end{aligned}
$$

where $p = q/(q-1)$. Then the gAWCGA of f with the weakness sequence $\{(\tau, t'_n)\}_{n=1}^{\infty}$, the perturbation sequence $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$, and the error sequence $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ satisfies for any $n \ge M$ the estimate

$$\|f_n\| \le \frac{16\gamma^{1/q} M^{1/p}}{\tau} n^{-1/p}.$$

If the modulus of smoothness of a space is known, we might get weaker restrictions on the error sequences and better estimates on the convergence rate. In the following two results, we give the rate of convergence of the adaptive gAWCGA for ℓ_p spaces.

Corollary 4.2.4. *Let $X = \ell_q$, $1 < q \le 2$ and $\mathcal{D}$ be a dictionary. Let $f \in A_1(\mathcal{D})$, $0 < \tau \le 1$, and the sequences $\{t'_n\}_{n=1}^{\infty}$, $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$, and $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ be such that* *for any* $n \geq 1$

$$t'_{n+1} = \frac{8}{12^p}\tau E_n,$$
$$\delta_n = \frac{1}{12^p 2p}\tau^p \min\{\|f_n\|^p, 1\},$$
$$\delta'_n = \frac{1}{12^p 2p}\tau^p \min\{\|f_n\|^{p+1}, 1\},$$
$$\eta_n = \frac{1}{12^p 2p}\tau^p E_n^p,$$
$$\eta'_n = \frac{1}{12^p 2p}\tau^p E_n^{p+1}.$$

where $p = q/(q-1)$. *Then the gAWCGA of* f *with the weakness sequence* $\{(\tau, t'_n)\}_{n=1}^{\infty}$, *the perturbation sequence* $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$, *and the error sequence* $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ *satisfies for any* $n \geq 1$ *the estimate*

$$\|f_n\| \leq \frac{5(p/2)^{1/p}}{\tau} n^{-1/p}.$$

Corollary 4.2.5. *Let* $X = \ell_q$, $2 \leq q < \infty$ *and* $\mathcal{D}$ *be a dictionary. Let* $f \in A_1(\mathcal{D})$, $0 < \tau \leq 1$, *and the sequences* $\{t'_n\}_{n=1}^{\infty}$, $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$, *and* $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ *be such that for any* $n \geq 1$

$$t'_{n+1} = \frac{1}{18}\tau E_n,$$
$$\delta_n = \frac{1}{576(q-1)}\tau^2 \min\{\|f_n\|^2, 1\},$$
$$\delta'_n = \frac{1}{576(q-1)}\tau^2 \min\{\|f_n\|^3, 1\},$$
$$\eta_n = \frac{1}{576(q-1)}\tau^2 E_n^2,$$
$$\eta'_n = \frac{1}{576(q-1)}\tau^2 E_n^3.$$

Then the gAWCGA of f *with the weakness sequence* $\{(\tau, t'_n)\}_{n=1}^{\infty}$, *the perturbation sequence* $\{(\delta_n, \delta'_n)\}_{n=0}^{\infty}$, *and the error sequence* $\{(\eta_n, \eta'_n)\}_{n=1}^{\infty}$ *satisfies for any* $n \geq 1$ *the estimate*

$$\|f_n\| \leq \frac{5\sqrt{q-1}}{\tau} n^{-1/2}.$$

4.3 Proofs for Section 4.1

In this section we give proofs of Theorems 4.1.1 and 4.1.2. First, we recall some known results.

Lemma G (Temlyakov 2001, Lemma 2.1). *Let X be a uniformly smooth Banach space and L be a finite-dimensional subspace of X. For any $f \in X \setminus L$ denote by f_L the best approximant of f from L. Then for any $\phi \in L$*

$$F_{f-f_L}(\phi) = 0.$$

Lemma H (Temlyakov 2001, Lemma 2.2). *For any bounded linear functional F and any dictionary $\mathcal{D}$*

$$\sup_{g \in \mathcal{D}} F(g) = \sup_{g \in A_1(\mathcal{D})} F(g).$$

We will also use several technical results from Temlyakov 2005 rewritten for the gAWCGA.

Lemma 4.3.1. *Let X be a Banach space with the modulus of smoothness $\rho(u)$. Then for any $\phi \in \Phi_n$*

$$|F_n(\phi)| \leq \beta_n(\phi) := \inf_{\lambda > 0} \frac{1}{\lambda} \left(\delta_n + \eta_n + \frac{\delta'_n + \eta'_n}{\|f_n\|} + 2\rho(\lambda \|\phi\|) \right).$$

Proof. Take any ϕ from Φ_n. By the definition of the modulus of smoothness (2.3) for any $\lambda > 0$

$$\|f_n - \lambda\phi\| + \|f_n + \lambda\phi\| \leq 2 \|f_n\| \left(1 + \rho \left(\frac{\lambda \|\phi\|}{\|f_n\|} \right) \right).$$

Assume that $F_n(\phi) \geq 0$ (case $F_n(\phi) < 0$ is handled similarly). Then, using (4.1), we obtain

$$\|f_n + \lambda\phi\| \geq F_n (f_n + \lambda\phi) \geq (1 - \delta_n) \|f_n\| - \delta'_n + \lambda F_n(\phi),$$

thus

$$\|f_n - \lambda\phi\| \leq \|f_n\| \left(1 + \delta_n + 2\rho \left(\frac{\lambda \|\phi\|}{\|f_n\|} \right) \right) + \delta'_n - \lambda F_n(\phi).$$

On the other hand, by (4.3)

$$\|f_n - \lambda\phi\| \geq E_n \geq (1+\eta_n)^{-1}\left(\|f_n\| - \eta_n'\right) \geq (1-\eta_n)\|f_n\| - \eta_n'.$$

Therefore

$$\lambda F_n(\phi) \leq \|f_n\| \left(\delta_n + \eta_n + 2\rho\left(\frac{\lambda\|\phi\|}{\|f_n\|}\right)\right) + \delta_n' + \eta_n'$$

and, since the inequality holds for any $\lambda > 0$,

$$F_n(\phi) \leq \inf_{\lambda>0} \frac{1}{\lambda}\left(\delta_n + \eta_n + \frac{\delta_n' + \eta_n'}{\|f_n\|} + 2\rho(\lambda\|\phi\|)\right) = \beta_n(\phi).$$

□

Lemma 4.3.2. *Let X be a Banach space with the modulus of smoothness $\rho(u)$. Take a number $\epsilon \geq 0$ and two elements f and h from X such that $\|f-h\| \leq \epsilon$ and $h/A \in A_1(\mathcal{D})$ with some number $A = A(\epsilon) > 0$. Then*

$$|F_n(\phi_{n+1})| \geq t_{n+1}A^{-1}\left((1-\delta_n)\|f_n\| - \delta_n' - \beta_n(G_n) - \epsilon\right) - t_{n+1}'.$$

Proof. Condition (4.2) and Lemma H provide

$$|F_n(\phi_{n+1})| \geq t_{n+1}\sup_{g\in\mathcal{D}}|F_n(g)| - t_{n+1}' = t_{n+1}\sup_{g\in A_1(\mathcal{D})}|F_n(g)| - t_{n+1}'.$$

Taking $g = h/A \in A_1(\mathcal{D})$ we obtain

$$\sup_{g\in A_1(\mathcal{D})}|F_n(g)| \geq A^{-1}|F_n(h)| \geq A^{-1}\left(|F_n(f)| - \epsilon\right)$$
$$\geq A^{-1}\left(|F_n(f_n)| - |F_n(G_n)| - \epsilon\right).$$

Hence condition (4.1) and Lemma 4.3.1 provide

$$|F_n(\phi_{n+1})| \geq t_{n+1}A^{-1}\left((1-\delta_n)\|f_n\| - \delta_n' - \beta_n(G_n) - \epsilon\right) - t_{n+1}'.$$

□

Lemma 4.3.3. *Let X be a Banach space with the modulus of smoothness $\rho(u)$. Take a number $\epsilon \geq 0$ and two elements f and h from X such that $\|f-h\| \leq \epsilon$ and $h/A \in A_1(\mathcal{D})$ with some number $A = A(\epsilon) > 0$. Then for any $m > n$*

$$E_m \leq \inf_{\mu\geq 0} \|f_n\| \left[1 + \delta_n + \frac{\delta_n'}{\|f_n\|} + 2\rho\left(\frac{\mu}{\|f_n\|}\right) \right.$$
$$\left. - \frac{\mu t_{n+1}}{A\,\|f_n\|}\Big((1-\delta_n)\,\|f_n\| - \delta_n' - \beta_n(G_n) - \epsilon\Big)\right] + \mu t_{n+1}'.$$

Proof. By the definition of the modulus of smoothness (2.3) for any $\mu \geq 0$

$$\|f_n - \mu\phi_{n+1}\| + \|f_n + \mu\phi_{n+1}\| \leq 2\,\|f_n\| \left(1 + \rho\left(\frac{\mu}{\|f_n\|}\right)\right).$$

Assume that $F_n(\phi_{n+1}) \geq 0$ (case $F_n(\phi_{n+1}) < 0$ is handled similarly). Then, using (4.1) and Lemma 4.3.2, we get

$$\begin{aligned}\|f_n + \mu\phi_{n+1}\| &\geq F_n\left(f_n + \mu\phi_{n+1}\right) \geq (1-\delta_n)\,\|f_n\| - \delta_n' + \mu|F_n(\phi_{n+1})| \\ &\geq (1-\delta_n)\,\|f_n\| - \delta_n' \\ &\quad + \mu t_{n+1} A^{-1}\left((1-\delta_n)\,\|f_n\| - \delta_n' - \beta_n(G_n) - \epsilon\right) - \mu t_{n+1}'.\end{aligned}$$

Thus

$$\|f_n - \mu\phi_{n+1}\| \leq \|f_n\| \left(1 + \delta_n + \frac{\delta_n'}{\|f_n\|} + 2\rho\left(\frac{\mu}{\|f_n\|}\right)\right)$$
$$- \mu t_{n+1} A^{-1}\Big((1-\delta_n)\,\|f_n\| - \delta_n' - \beta_n(G_n) - \epsilon\Big) + \mu t_{n+1}'.$$

On the other hand, since $E_m \leq E_{n+1} \leq \|f_n - \mu\phi_{n+1}\|$ for any $\mu \geq 0$,

$$E_m \leq \|f_n\| \left[1 + \delta_n + \frac{\delta_n'}{\|f_n\|} + 2\rho\left(\frac{\mu}{\|f_n\|}\right) \right.$$
$$\left. - \frac{\mu t_{n+1}}{A\,\|f_n\|}\Big((1-\delta_n)\,\|f_n\| - \delta_n' - \beta_n(G_n) - \epsilon\Big)\right] + \mu t_{n+1}'.$$

Taking an infimum over all $\mu \geq 0$ completes the proof. □

We are now ready to prove Theorem 4.1.1.

Proof of Theorem 4.1.1. Assume that for some element $f \in X$ the gAWCGA does not converge. Note that then the monotone sequence $\{E_n\}_{n=1}^{\infty}$ does not converge to 0 since otherwise conditions (4.4) would imply

$$\lim_{n\to\infty} \|f_n\| \leq \lim_{n\to\infty} ((1+\eta_\infty)E_n + \eta_n') = 0.$$

Thus there exists a number $\alpha > 0$ such that for any $n \geq 1$

$$\|f_n\| \geq E_n \geq \alpha. \tag{4.27}$$

Denote $C_f = (2+\eta_\infty)\|f\| + \eta_\infty' < \infty$, where $\eta_\infty = \sup_{n\geq 1} \eta_n$ and $\eta_\infty' = \sup_{n\geq 1} \eta_n'$. Then inequality (4.3) gives for any $n \geq 1$

$$\begin{aligned} \|f_n\| &\leq (1+\eta_\infty)\|f\| + \eta_\infty' \leq C_f, \\ \|G_n\| &\leq \|f_n\| + \|f\| \leq C_f. \end{aligned} \tag{4.28}$$

Let $\{n_k\}_{k=1}^{\infty}$ be a subsequence for which conditions of the theorem hold. Then

$$\begin{aligned} \beta_{n_k}(G_{n_k}) &= \inf_{\lambda>0} \frac{1}{\lambda}\left(\delta_{n_k} + \eta_{n_k} + \frac{\delta_{n_k}' + \eta_{n_k}'}{\|f_{n_k}\|} + 2\rho(\lambda\|G_{n_k}\|)\right) \\ &\leq \inf_{\lambda>0} \frac{1}{\lambda}\left(\delta_{n_k} + \eta_{n_k} + \frac{\delta_{n_k}' + \eta_{n_k}'}{\alpha} + 2\rho(\lambda C_f)\right) \end{aligned}$$

and, due to conditions (4.6)–(4.10) and the inequality $0 \leq \theta t_n \xi_n \leq 1$, there exists a number $K \geq 1$ such that for any $k \geq K$ the following estimates hold with $\theta = \frac{\alpha^2}{24AC_f}$:

$$\left(\frac{1}{2} - \delta_{n_k}\right)\alpha - \delta_{n_k}' - \beta_{n_k}(G_{n_k}) \geq \frac{\alpha}{4}, \tag{4.29}$$

$$\delta_{n_k} + \frac{\delta_{n_k}'}{\alpha} \leq \theta\xi_{n_k+1}t_{n_k+1}, \tag{4.30}$$

$$(1+\eta_{n_k})(1-3\theta\xi_{n_k+1}t_{n_k+1}) \leq 1 - 2\theta\xi_{n_k+1}t_{n_k+1}, \tag{4.31}$$

$$\eta_{n_k}' + \alpha\xi_{n_k+1}t_{n_k+1}' \leq \alpha\theta\xi_{n_k+1}t_{n_k+1}. \tag{4.32}$$

Take $\epsilon = \alpha/2$ and find an element $h \in X$ such that $\|f-h\| \leq \epsilon$ and $h/A \in A_1(\mathcal{D})$ for some $A > 0$. Then Lemma 4.3.3, assumption (4.27), and estimates (4.28) and (4.29) provide for any $k \geq K$

$$E_{n_{k+1}} \leq \inf_{\mu \geq 0} \|f_{n_k}\| \left[1 + \delta_{n_k} + \frac{\delta'_{n_k}}{\alpha} + 2\rho\left(\frac{\mu}{\alpha}\right) - \frac{\mu t_{n_k+1}}{AC_f}\left(\left(\frac{1}{2} - \delta_{n_k}\right)\alpha - \delta'_{n_k} - \beta_{n_k}(G_{n_k})\right)\right] + \mu t'_{n_k+1}$$
$$\leq \inf_{\mu \geq 0} \|f_{n_k}\| \left[1 + \delta_{n_k} + \frac{\delta'_{n_k}}{\alpha} + 2\rho\left(\frac{\mu}{\alpha}\right) - \frac{\alpha\mu t_{n_k+1}}{4AC_f}\right] + \mu t'_{n_k+1}.$$

By taking $\mu = \alpha\xi_{n_k+1}$, and using estimates (4.30)–(4.32), and condition (4.3) we obtain

$$\begin{aligned} E_{n_{k+1}} &\leq \|f_{n_k}\| \left(1 + \delta_{n_k} + \frac{\delta'_{n_k}}{\alpha} - 4\theta\xi_{n_k+1}t_{n_k+1}\right) + \alpha\xi_{n_k+1}t'_{n_k+1} \\ &\leq \|f_{n_k}\| \left(1 - 3\theta\xi_{n_k+1}t_{n_k+1}\right) + \alpha\xi_{n_k+1}t'_{n_k+1} \\ &\leq E_{n_k}\left(1 - 2\theta\xi_{n_k+1}t_{n_k+1}\right) + \eta'_{n_k} + \alpha\xi_{n_k+1}t'_{n_k+1} \\ &\leq E_{n_k}\left(1 - \theta\xi_{n_k+1}t_{n_k+1}\right). \end{aligned} \tag{4.33}$$

Note that condition (4.5) implies that the infinite product $\prod_{k=1}^{\infty}\left(1 - \theta\xi_{n_k+1}t_{n_k+1}\right)$ diverges to 0. Then, recursively applying estimate (4.33), we obtain for sufficiently big $N \geq K$

$$\begin{aligned} E_{n_{N+1}} &\leq E_{n_K} \prod_{k=K}^{N}\left(1 - \theta\xi_{n_k+1}t_{n_k+1}\right) \\ &\leq \|f\| \prod_{k=K}^{N}\left(1 - \theta\xi_{n_k+1}t_{n_k+1}\right) \\ &< \alpha, \end{aligned}$$

which contradicts assumption (4.27). Therefore

$$\lim_{n\to\infty} E_n = 0,$$

i.e. the gAWCGA of f converges to f. $\square$

We will use the following simple lemma to prove Theorem 4.1.2.

Lemma 4.3.4. *Let $q > 1$, $a \geq 0$ and $b \geq 1$. Then*

$$(a + b^q)^{1/q} \leq a + b.$$

Proof. Due to the convexity of $(1 + x)^q$ we have for any $x \geq 0$

$$(1 + x)^q \geq 1 + qx.$$

Then by taking $x = a/b$ we get

$$(a + b)^q = b^q(1 + x)^q \geq b^q(1 + qx) = b^q + aqb^{q-1} \geq a + b^q.$$

□

Proof of Theorem 4.1.2. We start with the proof of the sufficiency. Assume that conditions (4.12)–(4.17) hold for some subsequence $\{n_k\}_{k=1}^{\infty}$. Choose any number $0 < \theta \leq 1/2$ and find the corresponding sequence $\{\xi_n\}_{n=1}^{\infty}$. Then using the definition $\rho(\xi_n) = \theta t_n \xi_n$ and the estimate $\rho(u) \leq \gamma u^q$, we derive

$$\xi_n \geq \left(\frac{\theta}{\gamma} t_n\right)^{p-1}.$$

Thus for any $n \geq 1$

$$t_n^p \leq \left(\frac{\gamma}{\theta}\right)^{p-1} t_n \xi_n,$$

and conditions (4.12)–(4.17) imply that conditions (4.5)–(4.10) hold for the subsequence $\{n_k\}_{k=1}^{\infty}$ and any $0 < \theta \leq 1/2$. Therefore Theorem 4.1.1 guarantees convergence of the gAWCGA for any dictionary $\mathcal{D}$ and any element $f \in X$.

We prove the necessity of the stated conditions by giving a counterexample. Namely, we assume that at least one of conditions (4.11)–(4.17) fails, and give an example of such a Banach space $X \in \mathcal{P}_q$, a dictionary $\mathcal{D}$, and an element $f \in \mathcal{D}$ that the gAWCGA of f diverges.

Let $X = \ell_q \in \mathcal{P}_q$ and $\mathcal{D} = \{\pm e_n\}_{n=0}^{\infty}$, where $\{e_n\}_{n=0}^{\infty}$ is the canonical basis in ℓ_q.

Assume that condition (4.11) fails, i.e. that there exist a subsequence $\{n_k\}_{k=1}^{\infty}$ and a number $\alpha > 0$ such that for any $k \geq 1$

$$\eta_{n_k} \geq \alpha k \quad \text{or} \quad \eta'_{n_k} \geq \alpha.$$

Then take a positive non-increasing sequence $\{a_j\}_{j=1}^{\infty} \in \ell_q$ such that

$$a_1 \geq \alpha \quad \text{and} \quad \left(\sum_{j=n_k+1}^{\infty} a_j^q \right)^{1/q} \geq k^{-1}$$

for any $k \geq 1$. Denote $f = \sum_{j=1}^{\infty} a_j e_j \in \ell_q$ and consider the following realization of the gAWCGA of f:

For $n \notin \{n_k\}_{k=1}^{\infty}$ choose F_{n-1} to be the norming functional for f_{n-1}, $\phi_n = e_n$, and $G_n = \sum_{j=1}^{n} a_j e_j$.

For $n \in \{n_k\}_{k=1}^{\infty}$ choose F_{n-1} to be the norming functional for f_{n-1}, $\phi_n = e_n$ and $G_n = \alpha e_1 + \sum_{j=1}^{n} a_j e_j$, which is possible since

$$\|f_{n_k}\|_q = \left(\alpha^q + \sum_{j=n_k+1}^{\infty} a_j^q \right)^{1/q} \leq \alpha + E_{n_k},$$

and either

$$\|f_{n_k}\|_q \leq E_{n_k} + \eta'_{n_k} \quad \text{or} \quad \|f_{n_k}\|_q \leq (1 + \alpha k) E_{n_k} \leq (1 + \eta_{n_k}) E_{n_k}.$$

Then for any $k \geq 1$ norm of the remainder $\|f_{n_k}\|_q \geq \alpha$, hence $\|f_n\|_q \not\to 0$ and the gAWCGA of f diverges.

Assume now that conditions (4.12)–(4.17) do not hold, i.e. for any subsequence $\{n_k\}_{k=1}^{\infty}$ at least one of the following statements fails:

$$\sum_{k=1}^{\infty} t_{n_k+1}^p = \infty,$$

$$t'_{n_k+1} = o(t_{n_k+1})$$

$$\delta_{n_k} = o(t_{n_k+1}^p),$$

$$\delta'_{n_k} = o(t_{n_k+1}^p),$$

$$\eta_{n_k} = o(t_{n_k+1}^p),$$

$$\eta'_{n_k} = o(t_{n_k+1}^p).$$

For a number $\alpha > 0$ define sets

$$\Lambda_1 = \{n > 1 : \delta_{n-1} + \delta'_{n-1} \geq \alpha t_n^p \text{ or } \eta_{n-1} + \eta'_{n-1} \geq \alpha t_n^p \text{ or } t'_n \geq \alpha^{1/p} t_n\}$$

and $\Lambda_2 = \mathbb{N} \setminus \Lambda_1$. We claim that there exists an $\alpha > 0$ such that

$$\sum_{j \in \Lambda_2} t_j^p < \infty. \tag{4.34}$$

Indeed, if $\sum_{j\in\Lambda_2} t_j^p = \infty$ for any $\alpha > 0$ then for every $k \geq 1$ consider $\alpha(k) = 1/k$, and choose a sequence of disjoint finite sets $\{\Gamma_k\}_{k=1}^\infty$, where $\Gamma_k \subset \Lambda_2(k)$ is such that $\sum_{j\in\Gamma_k} t_j^p \geq 1$. Hence by considering the union $\cup_{k=1}^\infty (\Gamma_k + \{-1\})$ (where $+$ denotes the Minkowski addition), we receive the subsequence for which conditions (4.12)–(4.17) hold, which contradicts the aforementioned assumption. Fix an $\alpha > 0$ for which claim (4.34) holds, and find corresponding sets Λ_1 and Λ_2.

If $|\Lambda_1| < \infty$ then $\sum_{j=1}^\infty t_j^p < \infty$. Take $f = e_0 + \sum_{j=1}^\infty t_j^{p/q} e_j$ and consider the following realization of the gAWCGA of f:

For each $n \geq 1$ choose F_{n-1} to be the norming functional for f_{n-1}, $\phi_n = e_n$, and $G_n = \sum_{j=1}^n t_j^{p/q} e_j$. Then for any $n \geq 1$ norm of the remainder $\|f_n\|_q \geq 1$, hence the gAWCGA of f diverges.

Consider the case $|\Lambda_1| = \infty$. Take any such non-negative sequence $\{a_j\}_{j\in\Lambda_1}$ that $a_j \leq 1$ for any $j \geq 1$, $\sum_{j\in\Lambda_1} a_j^q \geq 1/\alpha$ and $\sum_{j\in\Lambda_1} a_j^p < \infty$. Denote

$$f = \alpha^{1/q} \beta \left(\sum_{j\in\Lambda_1} a_j e_j + \sum_{j\in\Lambda_2} t_j^{p/q} e_j \right),$$

where

$$\beta = \left(\eta_\infty + \eta'_\infty + \alpha \left(\sum_{j\in\Lambda_1} a_j^q + \sum_{j\in\Lambda_2} t_j^p \right) \right)^{-1/q} \leq 1.$$

We claim that for some realization of the gAWCGA of f the indices from Λ_1 will not be chosen. Namely, we show that there exists such a realization that for any $n \geq 1$ the set Γ_n of indices of e_j chosen on the first n steps of the algorithm and the n-th remainder f_n satisfy the following relations:

$$\begin{aligned}
&\Gamma_n \cap \Lambda_1 = \varnothing, \\
&f_n = \beta(\eta_n + \eta_n')^{1/q} e_1 + \alpha^{1/q}\beta \left(\sum_{j\in\Lambda_1} a_j e_j + \sum_{j\in\Lambda_2^{(n)}} t_j^{p/q} e_j \right),
\end{aligned} \tag{4.35}$$

where $\Lambda_2^{(n)} = \Lambda_2 \setminus \Gamma_n$. Consider the following realization of the gAWCGA of f:

For $n = 1$ choose

$$F_0(x) = F_f(x) = \frac{\sum_{j\in\Lambda_1} a_j^{q/p} x_j + \sum_{j\in\Lambda_2} t_j x_j}{(\alpha\beta^q)^{-1/p} \|f\|_q^{q/p}}.$$

Then, since $a_j \le 1$, we get

$$\begin{aligned}
&F_0(e_0) = 0, \\
&F_0(e_j) \le (\alpha\beta^q)^{1/p} \|f\|_q^{-q/p}, \text{ for any } j \in \Lambda_1, \\
&F_0(e_j) = t_j(\alpha\beta^q)^{1/p} \|f\|_q^{-q/p} \text{ for any } j \in \Lambda_2,
\end{aligned}$$

and choosing $\phi_1 = e_1$ satisfies (4.2) since $1 \in \Lambda_2$. Thus $\Gamma_1 = \{1\}$, and taking

$$f_1 = \beta(\eta_1 + \eta_1')^{1/q} e_1 + \alpha^{1/q}\beta \left(\sum_{j\in\Lambda_1} a_j e_j + \sum_{j\in\Lambda_2^{(1)}} t_j^{p/q} e_j \right)$$

satisfies (4.3) since the estimate

$$\beta \le E_1 = \alpha^{1/q}\beta \left(\sum_{j\in\Lambda_1} a_j^q + \sum_{j\in\Lambda_2^{(1)}} t_j^p \right)^{1/q} \le 1$$

and Lemma 4.3.4 provide

$$\begin{aligned}
\|f_1\|_q &= \beta \left(\eta_1 + \eta_1' + \alpha \left(\sum_{j\in\Lambda_1} a_j^q + \sum_{j\in\Lambda_2^{(1)}} t_j^p \right) \right)^{1/q} \\
&\le \beta(\eta_1 + \eta_1') + E_1 \\
&\le (1 + \eta_1)E_1 + \eta_1'.
\end{aligned}$$

Hence for $n = 1$ claim (4.35) holds.

For $n \geq 1$, provided

$$f_n = \beta(\eta_n + \eta_n')^{1/q} e_1 + \alpha^{1/q}\beta \left(\sum_{j\in\Lambda_1} a_j e_j + \sum_{j\in\Lambda_2^{(n)}} t_j^{p/q} e_j \right),$$

taking

$$F_n(x) = \frac{(\delta_n + \delta_n')^{1/p} x_0 + (\eta_n + \eta_n')^{1/p} x_1 + \alpha^{1/p} \left(\sum_{j\in\Lambda_1} a_j^{q/p} x_j + \sum_{j\in\Lambda_2^{(n)}} t_j x_j \right)}{\left(\beta^{-q}(1+\delta_n+\delta_n') \left\|f_n\right\|_q^q\right)^{1/p}}$$

satisfies (4.1) since the estimate

$$\beta \leq \left\|f_n\right\|_q = \beta \left(\eta_n + \eta_n' + \alpha \left(\sum_{j\in\Lambda_1} a_j^q + \sum_{j\in\Lambda_2^{(n)}} t_j^p \right) \right)^{1/q} \leq 1$$

and Hölder's inequality provide

$$|F_n(x)| \leq \frac{\left(\delta_n + \delta_n' + \beta^{-q} \left\|f_n\right\|_q^q\right)^{1/p} \left(\sum_{j=0}^{\infty} x_j^q \right)^{1/q}}{\left(\beta^{-q}(1+\delta_n+\delta_n') \left\|f_n\right\|_q^q\right)^{1/p}} \leq \left\|x\right\|_q$$

and

$$F_n(f_n) = \frac{\left\|f_n\right\|_q^q}{(1+\delta_n+\delta_n')^{1/p} \left\|f_n\right\|_q^{q/p}} \geq (1-\delta_n) \left\|f_n\right\|_q - \delta_n',$$

where the last inequality holds since $\left\|f_n\right\|_q \leq 1$ and

$$\begin{aligned}
&(1+\delta_n+\delta_n')^{1/p}((1-\delta_n)\left\|f_n\right\|_q - \delta_n') \\
&\quad \leq (1+\delta_n+\delta_n')^{1/p}(1-\delta_n-\delta_n')\left\|f_n\right\|_q \\
&\quad\quad = (1-(\delta_n+\delta_n')^2)^{1/p}(1-\delta_n-\delta_n')^{1/q}\left\|f_n\right\|_q \\
&\quad\quad\quad \leq \left\|f_n\right\|_q .
\end{aligned}$$

Hence such choice of a functional is admissible. Let $A_n = \left(\beta^{-q}(1+\delta_n+\delta_n')\left\|f_n\right\|_q^q\right)^{-1/p}$.

Then, since $a_j \leq 1$, we get

$$
\begin{aligned}
&F_n(e_0) = (\delta_n + \delta_n')^{1/p} A_n, \\
&F_n(e_1) = (\eta_n + \eta_n')^{1/p} A_n, \\
&F_n(e_j) \leq \alpha^{1/p} A_n \text{ for any } j \in \Lambda_1, \\
&F_n(e_j) = t_j \alpha^{1/p} A_n \text{ for any } j \in \Lambda_2^{(n)}, \\
&F_n(e_j) = 0 \text{ for any } j \in \Gamma_n \setminus \{0, 1\}.
\end{aligned}
$$

If $n+1 \in \Lambda_2$ we choose $\phi_{n+1} = e_{n+1}$. Otherwise $n+1 \in \Lambda_1$, and by definition of the set at least one of the following inequalities holds:

$$
\begin{aligned}
&F_n(e_0) \geq t_{n+1} \alpha^{1/p} A_n \geq t_{n+1} \alpha^{1/p} A_n - t_{n+1}', \\
&F_n(e_1) \geq t_{n+1} \alpha^{1/p} A_n \geq t_{n+1} \alpha^{1/p} A_n - t_{n+1}', \\
&t_{n+1} \sup_{g \in \mathcal{D}} F_n(g) - t_{n+1}' \leq t_{n+1} \alpha^{1/p} A_n - \alpha^{1/p} t_{n+1} \leq 0.
\end{aligned}
$$

Then we choose $\phi_{n+1} = e_0$ or $\phi_{n+1} = e_1$. In either case $\Gamma_{n+1} \cap \Lambda_1 = \varnothing$ and taking

$$
f_{n+1} = \beta(\eta_{n+1} + \eta_{n+1}')^{1/q} e_1 + \alpha^{1/q} \beta \left(\sum_{j \in \Lambda_1} a_j e_j + \sum_{j \in \Lambda_2^{(n+1)}} t_j^{p/q} e_j \right)
$$

satisfies (4.3) since the estimate

$$
\beta \leq E_{n+1} = \alpha^{1/q} \beta \left(\sum_{j \in \Lambda_1} a_j^q + \sum_{j \in \Lambda_2^{(n+1)}} t_j^p \right)^{1/q} \leq 1
$$

and Lemma 4.3.4 provide

$$
\begin{aligned}
\|f_{n+1}\|_q &= \beta \left(\eta_{n+1} + \eta_{n+1}' + \alpha \left(\sum_{j \in \Lambda_1} a_j^q + \sum_{j \in \Lambda_2^{(n+1)}} t_j^p \right) \right)^{1/q} \\
&\leq \beta(\eta_{n+1} + \eta_{n+1}') + E_{n+1} \\
&\leq (1 + \eta_{n+1}) E_{n+1} + \eta_{n+1}'.
\end{aligned}
$$

Hence claim (4.35) holds for any $n \geq 1$. Thus $\|f_n\| \geq \beta \not\to 0$ and the gAWCGA of f diverges. □

Corollaries 4.1.4 and 4.1.5 are obtained using Theorems 4.1.1 and 4.1.2, and the following simple fact.

Lemma I (Lemma 2 from Dereventsov 2016)**.** *Let* $\{a_n\}_{n=1}^{\infty}$ *and* $\{b_n\}_{n=1}^{\infty}$ *be any such non-negative sequences that*

$$\sum_{n=1}^{\infty} a_n < \infty \quad \textit{and} \quad \sum_{n=1}^{\infty} b_n = \infty.$$

Then there exists a subsequence $\{n_k\}_{k=1}^{\infty}$ *such that*

$$\sum_{k=1}^{\infty} b_{n_k} = \infty \quad \textit{and} \quad a_{n_k} = o(b_{n_k}).$$

We also note that in corollary 4.1.5 the sequence $\{t'_n\}_{n=1}^{\infty}$ can be taken from ℓ_p rather than ℓ_1, we consider it to be from ℓ_1 only for the simplicity of formulation.

4.4 Proofs for Section 4.2

In this section we give proofs of Theorem 4.2.1 and Corollaries 4.2.2–4.2.5.

Proof of Theorem 4.2.1. Take any $f \in A_1(\mathcal{D})$. Then Lemma 4.3.3 applied with the subsequence $\{n_k\}_{k=1}^{\infty}$ provides with $\epsilon = 0$ and $A = 1$

$$\begin{aligned} E_{n_{k+1}} \leq \inf_{\mu \geq 0} \|f_{n_k}\| \Bigg[1 + \delta_{n_k} + \frac{\delta'_{n_k}}{\|f_{n_k}\|} + 2\rho\left(\frac{\mu}{\|f_{n_k}\|}\right) \\ - \frac{\mu}{\|f_{n_k}\|}\Big(t_{n_k+1}\Big((1-\delta_{n_k})\|f_{n_k}\| - \delta'_{n_k} - \beta_{n_k}(G_{n_k})\Big) - t'_{n_k+1}\Big)\Bigg], \end{aligned} \tag{4.36}$$

where

$$\begin{aligned} \beta_{n_k}(G_{n_k}) &= \inf_{\lambda > 0} \frac{1}{\lambda}\left(\delta_{n_k} + \eta_{n_k} + \frac{\delta'_{n_k} + \eta'_{n_k}}{\|f_{n_k}\|} + 2\rho(\lambda\|G_{n_k}\|)\right) \\ &\leq \inf_{\lambda > 0} \frac{1}{\lambda}\Big((D_k + D'_k + H_k + H'_k)t_{n_k+1}^p \|f_{n_k}\|^p + 2\gamma\lambda^q \|G_{n_k}\|^q\Big) \\ &= (2q\gamma)^{1/q} p^{1/p} \|G_{n_k}\| (D_k + D'_k + H_k + H'_k)^{1/p} t_{n_k+1} \|f_{n_k}\|. \end{aligned}$$

Denote $\tilde{D}_k = D_k + D'_k$ and $\tilde{H}_k = H_k + H'_k$. Then using condition (4.3), the estimate $\|E_n\| \le \|f\| \le 1$ for any $n \ge 0$, and the triangle inequality we get

$$\|f_{n_k}\| \le (1+\eta_{n_k})E_{n_k} + \eta'_{n_k} \le \left(1 + \tilde{H}_k t^p_{n_k+1} E^p_{n_k}\right) E_{n_k}, \tag{4.37}$$
$$\|G_{n_k}\| \le \|f\| + \|f_{n_k}\| \le 1 + \left(1 + \tilde{H}_k t^p_{n_k+1} E^p_{n_k}\right) E_{n_k} \le 2 + \tilde{H}_k,$$

and thus

$$\beta_{n_k}(G_{n_k}) \le \gamma_0 \Xi_k \|f_{n_k}\|,$$

where $\gamma_0 = (2q\gamma)^{1/q} p^{1/p}$ and $\Xi_k = (2+\tilde{H}_k)(\tilde{D}_k + \tilde{H}_k)^{1/p}$. Therefore

$$\begin{aligned}
&\frac{\mu}{\|f_{n_k}\|}\left(t_{n_k+1}\left((1-\delta_{n_k})\|f_{n_k}\| - \delta'_{n_k} - \beta_{n_k}(G_{n_k})\right) - t'_{n_k+1}\right)\\
&\quad \ge \frac{\mu}{\|f_{n_k}\|}\left(t_{n_k+1}((1-D_k)\|f_{n_k}\| - D'_k\|f_{n_k}\| - \gamma_0\Xi_k\|f_{n_k}\|) - T_k t_{n_k+1}\|f_{n_k}\|\right)\\
&\quad = \mu t_{n_k+1}\left(1 - T_k - \tilde{D}_k - \gamma_0\Xi_k\right),
\end{aligned}$$

and

$$\begin{aligned}
&\inf_{\mu\ge 0}\left[2\rho\left(\frac{\mu}{\|f_{n_k}\|}\right) - \frac{\mu}{\|f_{n_k}\|}\left(t_{n_k+1}\left((1-\delta_{n_k})\|f_{n_k}\| - \delta'_{n_k} - \beta_{n_k}(G_{n_k})\right) - t'_{n_k+1}\right)\right]\\
&\qquad \le \inf_{\mu\ge 0}\left[\mu^q 2\gamma\|f_{n_k}\|^{-q} - \mu t_{n_k+1}\left(1 - T_k - \tilde{D}_k - \gamma_0\Xi_k\right)\right]\\
&\qquad = -\gamma_0^{-p}\left(1 - T_k - \tilde{D}_k - \gamma_0\Xi_k\right)^p t^p_{n_k+1}\|f_{n_k}\|^p\\
&\qquad = -\left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - \Xi_k\right)^p t^p_{n_k+1}\|f_{n_k}\|^p.
\end{aligned}$$

Hence by substituting this estimate in (4.36) we obtain

$$\begin{aligned}
E_{n_k+1} &\le \|f_{n_k}\|\left[1 + \delta_{n_k} + \frac{\delta'_{n_k}}{\|f_{n_k}\|} - \left(\gamma_0^{-1}(1-T_k-\tilde{D}_k) - \Xi_k\right)^p t^p_{n_k+1}\|f_{n_k}\|^p\right]\\
&\le \|f_{n_k}\|\left[1 - \left(\left(\gamma_0^{-1}(1-T_k-\tilde{D}_k) - \Xi_k\right)^p - \tilde{D}_k\right) t^p_{n_k+1}\|f_{n_k}\|^p\right]\\
&\le E_{n_k}\left[1 + \tilde{H}_k t^p_{n_k+1}E^p_{n_k}\right]\left[1 - \left(\left(\gamma_0^{-1}(1-T_k-\tilde{D}_k) - \Xi_k\right)^p - \tilde{D}_k\right) t^p_{n_k+1}E^p_{n_k}\right]\\
&\le E_{n_k}\left[1 - \left(\left(\gamma_0^{-1}(1-T_k-\tilde{D}_k) - \Xi_k\right)^p - \tilde{D}_k - \tilde{H}_k\right) t^p_{n_k+1}E^p_{n_k}\right],
\end{aligned}$$

where the last two inequalities hold since $\Xi_k = (2 + \tilde{H}_k)(\tilde{D}_k + \tilde{H}_k)^{1/p}$ and

$$
\begin{aligned}
\left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - \Xi_k\right)^p - \tilde{D}_k &\geq \left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - (\Xi_k + \tilde{D}_k^{1/p})\right)^p \\
&\geq \left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - (3 + \tilde{H}_k)(\tilde{D}_k + \tilde{H}_k)^{1/p}\right)^p \\
&= \alpha_k^p > 0
\end{aligned}
$$

by condition (4.26). By the same argument

$$
\begin{aligned}
0 < \alpha_k^p &= \left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - (3 + \tilde{H}_k)(\tilde{D}_k + \tilde{H}_k)^{1/p}\right)^p \\
&= \left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - \Xi_k - (\tilde{D}_k + \tilde{H}_k)^{1/p}\right)^p \\
&\leq \left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - \Xi_k\right)^p - \tilde{D}_k - \tilde{H}_k
\end{aligned}
$$

and therefore

$$
\begin{aligned}
E_{n_{k+1}} &\leq E_{n_k}\left[1 - \left(\left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - \Xi_k\right)^p - \tilde{D}_k - \tilde{H}_k\right) t_{n_k+1}^p E_{n_k}^p\right] \\
&\leq E_{n_k}\left(1 - \alpha_k^p t_{n_k+1}^p E_{n_k}^p\right). \qquad (4.38)
\end{aligned}
$$

We claim that $\alpha_k < 1$ for any $k \geq 0$. Indeed, note that for any $u > 0$ and any $x, y \in S_X$

$$
\gamma u^q \geq \rho(u) \geq \frac{1}{2}\left(\|x + uy\| + \|x - uy\| - 2\right).
$$

In particular, taking $u = 2$ and $x = y$ we get $\gamma \geq 2^{-q}$. Then

$$
\begin{aligned}
\alpha_k^p &= \left(\gamma_0^{-1}(1 - T_k - \tilde{D}_k) - (3 + \tilde{H}_k)(\tilde{D}_k + \tilde{H}_k)^{1/p}\right)^p \\
&\leq \gamma_0^{-p} = \frac{q-1}{q^p (2\gamma)^{p-1}} \leq \frac{2(q-1)}{q^p} < \frac{2(q-1)}{q} = \frac{2}{p} \leq 1
\end{aligned}
$$

and estimate (4.38) implies

$$
E_{n_{k+1}}^p \leq E_{n_k}^p \left(1 - \alpha_k^p t_{n_k+1}^p E_{n_k}^p\right)^p \leq E_{n_k}^p \left(1 - \alpha_k^p t_{n_k+1}^p E_{n_k}^p\right). \qquad (4.39)
$$

We use induction by m to show that for any $m \geq 0$

$$
E_{n_m} \leq \left(1 + \sum_{k=0}^{m-1} \alpha_k^p t_{n_k+1}^p\right)^{-1/p}. \qquad (4.40)
$$

For $m = 0$ we have

$$E_{n_0} \leq E_0 = \|f\| \leq 1,$$

hence the estimate holds. Assume that it holds for m. Then inequality (4.39) provides

$$\begin{aligned} E_{n_{m+1}}^{-p} &\geq E_{n_m}^{-p} \left(1 - \alpha_m^p t_{n_m+1}^p E_{n_m}^p\right)^{-1} \geq E_{n_m}^{-p} \left(1 + \alpha_m^p t_{n_m+1}^p E_{n_m}^p\right) \\ &= E_{n_m}^{-p} + \alpha_m^p t_{n_m+1}^p \geq 1 + \sum_{k=0}^{m} \alpha_k^p t_{n_k+1}^p, \end{aligned}$$

since by the assumption estimate (4.40) is correct for m. Thus the induction holds. Then for any $m \geq 0$

$$\begin{aligned} \|f_{n_m}\| &\leq (1 + \eta_{n_m}) E_{n_m} + \eta'_{n_m} \leq (1 + \tilde{H}_m) E_{n_m} \\ &< (1 + 3^{-p}) \left(1 + \sum_{k=0}^{m-1} \alpha_k^p t_{n_k+1}^p\right)^{-1/p} \end{aligned}$$

since condition (4.26) provides

$$\begin{aligned} 0 < \alpha_m &= \gamma_0^{-1}(1 - T_m - \tilde{D}_m) - (3 + \tilde{H}_m)(\tilde{D}_m + \tilde{H}_m)^{1/p} \\ &\leq 1 - 3\tilde{H}_m^{1/p}. \end{aligned}$$

□

Proof of Corollary 4.2.2. We will show that the error sequences satisfy condition (4.26). For the given sequences we have $T_k = 12^{-p}\gamma^{1-p}$ and $D_k = D'_k = H_k = H'_k = .25\,12^{-p}\gamma^{1-p}$ for all $k \geq 0$. First, note that $(2q\gamma)^{1/q}p^{1/p} \leq 3\gamma^{1/q}$. Then, using the estimate $\gamma \geq 2^{-q}$, we obtain

$$\begin{aligned} (2q\gamma)^{1/q}p^{1/p}\alpha_k &\geq 1 - T_k - D_k - D'_k - 3\gamma^{1/q}(3 + H_k + H'_k)(D_k + D'_k + H_k + H'_k)^{1/p} \\ &\geq 1 - 1.5\,12^{-p}\gamma^{1-p} - 3\gamma^{1/q}(3 + .5\,12^{-p}\gamma^{1-p})\left(12^{-p}\gamma^{1-p}\right)^{1/p} \\ &\geq 1 - 1.5\,6^{-p} - .25(3 + .5\,6^{-p}) = .25 - 1.625\,6^{-p} \geq .2 \end{aligned}$$

and

$$\alpha_k \geq .2\,(2q\gamma)^{-1/q}p^{-1/p} \geq \frac{1}{15}\gamma^{-1/q} > 0. \tag{4.41}$$

Thus, condition (4.26) is satisfied and estimate (4.40) provides

$$
\begin{aligned}
\|f_n\| &\le (1+\eta_n)E_n + \eta_n' \\
&\le (1+.5\,6^{-p})\left(1+\sum_{k=1}^{n}\alpha_{k-1}^p\tau^p\right)^{-1/p} \\
&\le \frac{73}{72}\left(\left(\frac{1}{15}\gamma^{-1/q}\tau\right)^p n\right)^{-1/p} \\
&< \frac{16\gamma^{1/q}}{\tau}n^{-1+1/q}.
\end{aligned}
$$

□

Proof of Corollary 4.2.3. Fix a subsequence $\{n_k\}_{k=1}^\infty$ for which the conditions of the corollary hold. Let $N = \max\{N \in \mathbb{N} : n_N \le n\}$. Since $\sup_{k\in\mathbb{N}} |n_{k+1} - n_k| = M < \infty$, we estimate $N = \lfloor n/M \rfloor$ and $N+1 \ge n/M$. Then, using estimates (4.41) and (4.40), we get

$$
\begin{aligned}
\|f_n\| &\le (1+\eta_n)E_n + \eta_n' \le (1+.5\,6^{-p})E_{n_N} \\
&\le \frac{289}{288}\left(1+\sum_{k=0}^{N-1}\alpha_{k-1}^p\tau^p\right)^{-1/p} \le \frac{289}{288}\left(1+\left(\frac{1}{15}\gamma^{-1/q}\tau\right)^p N\right)^{-1/p} \\
&< \frac{16\gamma^{1/q}}{\tau}(N+1)^{-1/p} \le \frac{16\gamma^{1/q}M^{1/p}}{\tau}n^{-1+1/q}.
\end{aligned}
$$

□

Proof of Corollary 4.2.4. It is known that $\ell_q \in \mathcal{P}_q$ for $1 < q \le 2$ and that the modulus of smoothness $\rho_q(u)$ of ℓ_q space satisfies

$$
\rho_q(u) \le \frac{1}{q}u^q
$$

(see Lemma B.1 from Donahue et al. 1997). thus $\gamma = 1/q$. First, we estimate α_k.

$$
\begin{aligned}
\alpha_k &= \frac{1}{2(p/2)^{1/p}}\left(1-\frac{8}{12^p}-\frac{1}{12^p p}\right) - \left(3+\frac{1}{12^p p}\right)\left(\frac{2}{12^p p}\right)^{1/p} \\
&= \frac{1}{2(p/2)^{1/p}}\left(\frac{1}{2}-\frac{48p+7}{12^p 6p}\right) \ge \frac{1}{2(p/2)^{1/p}}\left(\frac{1}{2}-\frac{103}{1728}\right) \\
&= \frac{1}{(p/2)^{1/p}}\frac{761}{3456} > 0.
\end{aligned}
$$

Thus, condition (4.26) is satisfied and estimate (4.40) provides

$$\begin{aligned}
\|f_n\| &\le (1+\eta_n)E_n + \eta_n' \\
&\le \left(1+\frac{1}{12^p p}\right)\left(1+\sum_{k=1}^{n}\alpha_{k-1}^p\tau^p\right)^{-1/p} \\
&\le \frac{289}{288}\frac{3456(p/2)^{1/p}}{761\tau}n^{-1/p} \\
&< \frac{5(p/2)^{1/p}}{\tau}n^{-1/p}.
\end{aligned}$$

□

Proof of Corollary 4.2.5. It is known that $\ell_q \in \mathcal{P}_2$ for $2 \le q < \infty$ and that the modulus of smoothness $\rho_q(u)$ of ℓ_q space satisfies

$$\rho_q(u) \le \frac{q-1}{2}\,u^2$$

(see Lemma B.1 from Donahue et al. 1997). Thus $\gamma = (q-1)/2$. First, we estimate α_k.

$$\begin{aligned}
\alpha_k &= \frac{1}{2\sqrt{q-1}}\left(1-\frac{1}{18}-\frac{1}{288(q-1)}\right) - \left(3+\frac{1}{288(q-1)}\right)\frac{1}{12\sqrt{q-1}} \\
&\ge \frac{1}{2\sqrt{q-1}}\frac{761}{1728} = \frac{761}{3456\sqrt{q-1}} > 0.
\end{aligned}$$

Thus, condition (4.26) is satisfied and estimate (4.40) provides

$$\begin{aligned}
\|f_n\| &\le (1+\eta_n)E_n + \eta_n' \\
&\le \left(1+\frac{1}{288(q-1)}\right)\left(1+\sum_{k=1}^{n}\alpha_{k-1}^2\tau^2\right)^{-1/2} \\
&\le \frac{289}{288}\frac{3456\sqrt{q-1}}{761\tau}n^{-1/2} \\
&< \frac{5\sqrt{q-1}}{\tau}n^{-1/2}.
\end{aligned}$$

□

Chapter 5

Greedy Algorithms for Convex Optimization

In this chapter we discuss the application of greedy algorithms in the field of convex optimization.

A general problem of convex optimization is to minimize a given convex function E on a Banach space X. It turns out that techniques from greedy approximation can be applied to efficiently solve such problems. The reasoning behind this is as follows: the process of greedy expansion can be viewed as solving a particular convex minimization problem, since for a given element $f \in X$, after m iterations a greedy algorithm returns an approximation G_m which is built to minimize the norm of $f - G_m$; i.e. a greedy algorithm minimizes the convex function $E(x) = \|f - x\|$.

Additionally, the restrictions imposed on the objective function E in convex optimization are often the same as the restrictions imposed on Banach space X and the element f in greedy approximation.

For example, one typical condition in convex optimization is smoothness of the objective function, which corresponds to the condition of smoothness of the space in greedy approximation. In fact, it is shown in Borwein et al. 2009 that the moduli of smoothness in both fields are connected in a straightforward fashion.

Another common restriction of convex optimization is the structure of the point of minimum $z_0 = \operatorname{argmin}_{x \in X} E(x)$. Often it is assumed that the atomic norm of z_0 with respect to some set $\mathcal{A}$ is bounded, i.e. $\|z_0\|_{\mathcal{A}} \leq M < \infty$, which is essentially the same condition as $f \in A_1(\mathcal{D})$ in greedy approximation.

Besides the similar statements and imposed conditions, the vital argument for using greedy algorithms is that they are designed to produce a sparse representation with respect to the selected dictionary. Thus, a greedy algorithm would naturally provide a sparse minimizer, which is often desirable in convex optimization. Moreover, since greedy algorithms are iterative, we control the trade-off between the accuracy and the sparsity, and obtain the optimal solution for the current minimization problem.

Recall that in greedy approximation, algorithms can be divided into two categories: dual greedy algorithms and X-greedy algorithms.

Dual greedy algorithms are the ones that use norming functionals, and, since a norming functional is the derivative of the norm, an adaptation of a dual algorithm for convex optimization would utilize the derivative of the objective function E. Such algorithms were considered, for instance, in Clarkson 2010, Tewari, Ravikumar, and Dhillon 2011, Gao and Petrova 2015, and Nguyen and Petrova 2016.

Such derivative-based algorithms are common for convex optimization. However, they require the exact value of derivative E' of the objective function E, which might be unknown or hard to approximate if the values of E are not precise, or if it is computationally hard to evaluate $E(x)$. In such cases, X-greedy algorithms can offer an alternative approach to the problem.

X-greedy algorithms use direct norm evaluations, and therefore an adaptation of an X-greedy algorithm for convex optimization (called an E-greedy algorithm) would only rely on function evaluations and would not require the derivative. This approach was first employed in T. Zhang 2003 and then generalized in DeVore and Temlyakov 2014. We will recall the related results and propose a new E-greedy algorithm for convex optimization.

In section 5.1 we compare some of the E-greedy algorithms for convex optimization, and in section 5.2 we discuss the convergence results for the stated algorithms. In section 5.3 we discuss the approximate versions of these algorithms. In section 5.5

we prove the stated results.

5.1 E-Greedy Algorithms for Convex Optimization

In this section we discuss some known E-greedy type algorithms, propose a new one, and give the convergence results for the stated algorithms. Let $(X, \|\cdot\|)$ be a real Banach space, $\mathcal{D}$ be a dictionary, and $E : X \to \mathbb{R}$ be a convex function.

We begin with the Relaxed E-Greedy Algorithm (REGA(co)), which was introduced and studied for special objective functions in T. Zhang 2003 under the name of Greedy Sequential Approximation, and then in DeVore and Temlyakov 2014 for a wider class of functions.

Definition (REGA(co))**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n \in [0;1]$ that

$$\{\phi_n, \lambda_n\} = \underset{\substack{\phi \in \mathcal{D} \\ 0 \leq \lambda \leq 1}}{\operatorname{argmin}} E((1-\lambda)x_{n-1} + \lambda\phi),$$

2. set $x_n = (1-\lambda_n)x_{n-1} + \lambda_n\phi_n$.

Note that the REGA(co) is naturally limited: since each approximation x_m is a convex combination of the previous approximation x_{m-1} and the next atom ϕ_m, the REGA(co) is restricted only to the class $A_1(\mathcal{D})$ and will not obtain elements from $X \setminus A_1(\mathcal{D})$. The benefit of this limitation is that when searching for the optimal step size λ_m, one only has to search on the interval $[0;1]$.
This way of updating the approximation is in the style of the Frank–Wolfe algorithm (see Frank and Wolfe 1956) and has been used in many algorithms recently (see e.g. Figueiredo, Nowak, and Wright 2007, Jaggi 2013, Shalev-Shwartz, Srebro, and T. Zhang 2010, Tewari, Ravikumar, and Dhillon 2011).

The next algorithm is the E-Greedy Algorithm with Free Relaxation (EGAFR(co)), which was introduced in DeVore and Temlyakov 2014.

Definition (EGAFR(co))**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n, \mu_n \in \mathbb{R}$ that

$$\{\phi_n, \lambda_n, \mu_n\} = \operatorname*{argmin}_{\substack{\phi\in\mathcal{D} \\ \lambda,\mu\in\mathbb{R}}} E(\mu x_{n-1} + \lambda\phi),$$

2. set $x_n = \mu_n x_{n-1} + \lambda_n \phi_n$.

The EGAFR(co) is not limited to the $A_1(\mathcal{D})$ class since an approximation x_m is generally not a convex but a linear combination of the previous approximation x_{m-1} and the next atom ϕ_m. However, the search for the optimal parameters λ_m, μ_m is more complicated in the EGAFR(co) since one has to perform a two-dimensional line search as opposed to a one-dimensional one in the REGA(co).

We introduce here the Rescaled Relaxed E-Greedy Algorithm (RREGA(co)), which attempts to combine the computational simplicity of the REGA(co) and the unrestrained nature of the EGAFR(co). Specifically, the RREGA(co) constructs the approximation x_m in two steps: it chooses the best direction ϕ_m, and then rescales $x_{m-1} + \lambda_m \phi_m$. A similar approach was used for the Rescaled Pure Greedy Algorithm in Petrova 2015, which was adapted for convex optimization in Gao and Petrova 2015.

Definition (RREGA(co))**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n \geq 0$ that

$$\{\phi_n, \lambda_n\} = \operatorname*{argmin}_{\substack{\phi\in\mathcal{D} \\ \lambda\geq 0}} E(x_{n-1} + \lambda\phi),$$

2. choose such $\mu_n \geq 0$ that

$$\mu_n = \operatorname*{argmin}_{\mu\geq 0} E(\mu(x_{n-1} + \lambda_n \phi_n)),$$

3. set $x_n = \mu_n \left(x_{n-1} + \lambda_n \phi_n\right)$.

Similarly to the EGAFR(co), the RREGA(co) is not limited to the $A_1(\mathcal{D})$ and is computationally simpler since one iteration of the RREGA(co) requires two one-dimensional line searches as opposed to a two-dimensional line search in the EGAFR(co).

5.2 Convergence of the E-Greedy Algorithms for Convex Optimization

In this section we discuss the convergence and the rate of convergence of the greedy algorithms from section 5.1. Since the convex optimization setting is much more general than that of approximation theory, we need to impose additional conditions on the objective function E. For a Banach space X and a function $E : X \to \mathbb{R}$ denote

$$\Omega := \{x \in X : E(x) \le E(0)\} \subset X.$$

We assume that the minimum exists and is attained, i.e.

$$z_0 := \operatorname*{argmin}_{x \in X} E(x) \in \Omega$$

and that support functionals are bounded on Ω, i.e. there exists a constant M_Ω such that for any support functional H_x at $x \in \Omega$

$$\|H_x\|_{X^*} \le M_\Omega.$$

We now state the known convergence results for the REGA(co) and EGAFR(co) from DeVore and Temlyakov 2014 (in case $q = 2$, the following theorem was proven in T. Zhang 2003).

Theorem J (DeVore and Temlyakov 2014, Theorem 1.1)**.** *Let $\mathcal{D}$ be a dictionary, $E : X \to \mathbb{R}$ be a convex uniformly smooth on $A_1(\mathcal{D}) \cap \Omega$ function, and $z_0 \in A_1(\mathcal{D})$. Then the REGA(co) of E converges.*

Additionally, if $E \in \mathcal{P}_q(\Omega, X)$, then the REGA(co) provides for any $m \ge 1$

$$E(x_m) - E(z_0) \le C(q, \gamma)\, m^{1-q}.$$

Theorem K (DeVore and Temlyakov 2014, Theorem 1.2)**.** *Let $E : X \to \mathbb{R}$ be a convex uniformly smooth on Ω function. Then the EGAFR(co) of E converges for any dictionary $\mathcal{D}$.*
Additionally, if $E \in \mathcal{P}_q(\Omega, X)$ and $z_0 \in A_1(\mathcal{D})$ then the EGAFR(co) provides for any $m \geq 1$

$$E(x_m) - E(z_0) \leq C(E, q, \gamma)\, m^{1-q}.$$

We will prove the following convergence result for the RREGA(co).

Theorem 5.2.1. *Let $E : X \to \mathbb{R}$ be a convex uniformly smooth on Ω function. Then the RREGA(co) of E converges for any dictionary $\mathcal{D}$.*
Additionally, if $E \in \mathcal{P}_q(\Omega, X)$ and $z_0 \in A_1(\mathcal{D})$ then the RREGA(co) provides for any $m \geq 1$

$$E(x_m) - E(z_0) \leq 8\gamma\, m^{1-q}. \tag{5.1}$$

Note that in DeVore and Temlyakov 2014, Theorem K is stated in a more general way: in the case $E \in \mathcal{P}_q(\Omega, X)$, it provides the convergence rate of the EGAFR(co) regardless of whether or not z_0 is in $A_1(\mathcal{D})$. Concretely, it states that for $E \in \mathcal{P}_q(\Omega, X)$ the EGAFR(co) provides

$$E(x_m) - E(z_0) \leq C(q, \gamma, E)\epsilon_m,$$

where

$$\epsilon_m = \inf\{\epsilon > 0 : A(\epsilon)^q m^{1-q} \leq \epsilon\}$$

and

$$A(\epsilon) = \inf\{M > 0 : \exists y \in X \text{ such that } y/M \in A_1(\mathcal{D}) \text{ and } E(y) - E(z_0) < \epsilon\}.$$

Theorem 5.2.1 also can be formulated in such a way with appropriate changes in the proof. However, for the simplicity of presentation, we state and prove a more direct form.

Besides the computational simplicity, another advantage of the RREGA(co) over the EGAFR(co) is that the intervals of all possible values of the parameters λ_m and μ_m can be determined if some additional information is known. We will assume that C_Ω — a bound for the diameter of the set $\Omega = \{x \in X : E(x) \le E(0)\}$, is given, i.e.

$$\operatorname{diam} \Omega \le C_\Omega.$$

Moreover, in some cases we can significantly reduce the search interval for λ_m if a lower bound for the objective function E is known, i.e. there is a constant C_E such that

$$E(x) \ge C_E \text{ for any } x \in \Omega.$$

This bound is given naturally in problems like regression modeling and statistical classification, as a loss function is bounded from below by 0.

Theorem 5.2.2. *Let $E : X \to \mathbb{R}$ be a convex uniformly smooth on Ω function. Then the parameters λ_m and μ_m of the RREGA(co) of E with respect to any dictionary $\mathcal{D}$ satisfy the following estimates for any $m \ge 1$*

$$\lambda_m \in [0; C_\Omega] \quad \textit{and} \quad \mu_m \in \left[0; \frac{C_\Omega}{\|x_{m-1} + \lambda_m \phi_m\|}\right].$$

Moreover, if $E \in \mathcal{P}_q(X, \Omega)$, $z_0 \in A_1(\mathcal{D})$, and $E(x) \ge C_E$ on Ω, then for any $m \ge 1$ one can take

$$\begin{aligned} \lambda_m &\in \left[0; \min\left\{C_\Omega,\ \left(\frac{E(x_m) - C_E}{2\gamma q}\right)^{p-1}\right\}\right], \\ \mu_m &\in \left[0; \frac{\min\{1, C_\Omega\}}{\|x_{m-1} + \lambda_m \phi_m\|}\right]. \end{aligned} \tag{5.2}$$

Note that while estimates (5.2) generally do not guarantee the optimal values of λ_m and μ_m, they still provide the convergence rate (5.1).

5.3 Approximate E-Greedy Algorithms for Convex Optimization

In this section we discuss approximate versions of the algorithms from section 5.1 as well as their convergence.

Let $\{\delta_n\}_{n=1}^{\infty}$ be a non-negative sequence (called a weakness sequence) that represents the absolute values of the computational inaccuracies. The following approximate versions of the REGA and the EGAFR (REGA$\{\delta_n\}$ and EGAFR$\{\delta_n\}$ respectively) were analyzed in Temlyakov 2016 (for a special case $\delta_n = \delta$ for all $n \geq 1$ these algorithms were considered in DeVore and Temlyakov 2014).

Definition (REGA$\{\delta_n\}$)**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n \in [0;1]$ that

$$E((1-\lambda_n)x_{n-1} + \lambda_n\phi_n) \leq \min_{\substack{\phi\in\mathcal{D}\\ 0\leq\lambda\leq 1}} E((1-\lambda)x_{n-1} + \lambda\phi) + \delta_n,$$

2. set $x_n = (1-\lambda_n)x_{n-1} + \lambda_n\phi_n$.

Definition (EGAFR$\{\delta_n\}$)**.** Set $x_0 = 0$ and for each $n \geq 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n, \mu_n \in \mathbb{R}$ that

$$E(\mu_n x_{n-1} + \lambda_n\phi_n) \leq \min_{\substack{\phi\in\mathcal{D}\\ \lambda,\mu\in\mathbb{R}}} E(\mu x_{n-1} + \lambda\phi) + \delta_n,$$

2. set $x_n = \mu_n x_{n-1} + \lambda_n\phi_n$.

Theorem L (Temlyakov 2016, Theorems 3.3 and 3.4)**.** *Let $E : X \to \mathbb{R}$ be a convex uniformly smooth on $A_1(\mathcal{D})$ function, $z_0 \in A_1(\mathcal{D})$ and $\{\delta_n\}_{n=1}^{\infty}$ be a weakness sequence with*

$$\lim_{n\to\infty} \delta_n = 0.$$

Then the REGA(co) of E converges for any dictionary $\mathcal{D}$.
Additionally, if $E \in \mathcal{P}_q(\Omega, X)$, $z_0 \in A_1(\mathcal{D})$ and $\delta_n \leq \delta n^{-q}$ for some $\delta > 0$ and all $n \geq 1$, then

$$E(x_m) - E(z_0) \leq C(q,\gamma,E,\delta)\, m^{1-q}.$$

Theorem M (Temlyakov 2016, Proposition 3.1)**.** *Let $E : X \to \mathbb{R}$ be a convex uniformly smooth on Ω function and $\{\delta_n\}_{n=1}^{\infty}$ be a weakness sequence with*

$$\lim_{n\to\infty} \delta_n = 0.$$

Then the EGAFR(co) of E converges for any dictionary $\mathcal{D}$.
Additionally, if $E \in \mathcal{P}_q(\Omega, X)$, $z_0 \in A_1(\mathcal{D})$ and $\delta_n \le \delta n^{-q}$ for some $\delta > 0$ and all $n \ge 1$, then

$$E(x_m) - E(z_0) \le C(q, \gamma, E, \delta)\, m^{1-q}.$$

We introduce the Approximate Rescaled Relaxed E-Greedy Algorithm for convex optimization (ARREGA(co)) — an application-oriented version of the RREGA in which steps of the algorithm might be performed not exactly, but with some inaccuracies. We describe these inaccuracies in form of the absolute errors on both steps of the algorithm.

Definition (ARREGA(co))**.** Set $x_0 = 0$ and for each $n \ge 1$

1. find any such $\phi_n \in \mathcal{D}$ and $\lambda_n \ge 0$ that

$$E\left(x_{n-1} + \lambda_n \phi_n\right) \le \inf_{\phi\in\mathcal{D},\lambda\ge 0} E\left(x_{n-1} + \lambda\phi\right) + \delta_n,$$

2. find such $\mu_n \ge 0$ that

$$E(\mu_n(x_{n-1} + \lambda_n\phi_n)) \le \min_{\mu\ge 0} E(\mu(x_{n-1} + \lambda_n\phi_n)) + \delta_n,$$

3. set $x_n = \mu_n(x_{n-1} + \lambda_n\phi_n)$.

Note that while we impose the same inaccuracy $\delta_m$on both steps of the algorithm, any other approximate version of the RREGA can be considered, and the corresponding results will hold with appropriate changes in formulation and minor changes in proofs.

Note also that while parameter bounds for the ARREGA are generally weaker than the ones for the RREGA, one can still use the bounds from Theorem 5.2.2, which will result in the monotone decrease of the sequence $\{E(x_m)\}_{m=0}^{\infty}$, regardless of values of the weakness sequence $\{\delta_n\}_{n=1}^{\infty}$.

We will prove the following convergence result for the ARREGA(co).

Theorem 5.3.1. *Let $E : X \to \mathbb{R}$ be a convex uniformly smooth on Ω function and $\{\delta_n\}_{n=1}^{\infty}$ be a weakness sequence with*

$$\lim_{n\to\infty} \delta_n = 0.$$

Then the ARREGA(co) of E converges for any dictionary $\mathcal{D}$.
Additionally, if $E \in \mathcal{P}_q(\Omega, X)$ and $z_0 \in A_1(\mathcal{D})$ then

$$E(x_m) - E(z_0) \leq \left(8\gamma + 2\sum_{n=1}^{m} \delta_n n^{q-1}\right) m^{1-q}.$$

The following corollary offers concrete conditions on the weakness sequence, under which the convergence rate of the ARREGA(co) is of the same order as the convergence rate of the RREGA(co) (see (5.1)).

Corollary 5.3.2. *Let $\{\delta_n\}_{n=1}^{\infty}$ be a weakness sequence such that for any $n \geq 1$*

$$\delta_n \leq \delta n^{-r},$$

with some $\delta > 0$ and $r > q$, where $q \in (1; 2]$. Then for any dictionary $\mathcal{D}$ and any convex function $E \in \mathcal{P}_q(\Omega, X)$ with $z_0 \in A_1(\mathcal{D})$ the ARREGA(co) provides

$$E(x_m) - E(z_0) \leq \left(8\gamma + 2\delta\left(1 + (r-q)^{-1}\right)\right) m^{1-q}.$$

We note that while the previous corollary does not provide an estimate if $r = q$, one can obtain a similar result in this case by using a more sophisticated technique. It is additionally required, however, that a lower bound C_E on the objective function is given.

Proposition 5.3.3. *Let* $\{\delta_n\}_{n=1}^{\infty}$ *be a weakness sequence such that for any* $n \geq 1$

$$\delta_n \leq \delta n^{-q},$$

with some $\delta > 0$ *and* $q \in (1;2]$. *Then for any dictionary* $\mathcal{D}$ *and any convex function* $E \in \mathcal{P}_q(\Omega, X)$ *such that* $z_0 \in A_1(\mathcal{D})$ *and* $E(x) \geq C_E$ *for all* $x \in \Omega$, *the ARREGA(co) provides*

$$E(x_m) - E(z_0) \leq \max\left\{32\gamma,\, 4p\gamma\delta^{1/p},\, E(0) - C_E\right\} m^{1-q},$$

where $p = q/(q-1)$.

5.4 Implementation of E-Greedy Algorithms for Convex Optimization

In this section we demonstrate a few examples of the practical implementation of the greedy algorithms discussed in section 5.1.

In the examples below, we minimize the function $E : \ell_p^{(100)} \to \mathbb{R}$ with some $p \in (1, \infty)$ with respect to some dictionary $\mathcal{D}$ which is randomly generated for each example as described further. Let $\{e_k\}_{k=1}^{100}$ be the canonical basis in $\ell_p^{(100)}$ and $\{c_k^n\}_{k=1,n=1}^{100,200}$ be a random sequence of real numbers uniformly distributed on the interval $[-100, 100]$, i.e. for any $1 \leq k \leq 100$ and $1 \leq n \leq 200$

$$c_k^n \sim U(-100; 100).$$

Take the dictionary $\mathcal{D}$ as follows:

$$\mathcal{D} = \{\pm g_n\}_{n=1}^{200}, \quad \text{where } g_n = \frac{\sum_{k=1}^{100} c_k^n e_k}{\left(\sum_{k=1}^{100} |c_k^n|^p\right)^{1/p}}.$$

We will also use elements $f, h \in A_1(\mathcal{D})$ which are obtained in the following way:

$$f = \sum_{n=1}^{50} a_n^f g_{\sigma(n)}, \text{ where } a_n^f \sim \mathcal{N}(0, 100) \text{ and } \sum_{n=1}^{50} |a_n^f| = 1,$$

$$h = \sum_{n=1}^{50} a_n^h g_{\sigma(n)}, \text{ where } a_n^h \sim \mathcal{N}(0, 100) \text{ and } \sum_{n=1}^{50} |a_n^h| = 1,$$

where σ is a random permutation of the set $\{1, 2, 3, \ldots, 200\}$.

Example 1. $E(x) = \dfrac{\|f - x\|_4}{\|f\|_4} \; : \; \ell_4^{(100)} \to \mathbb{R}$

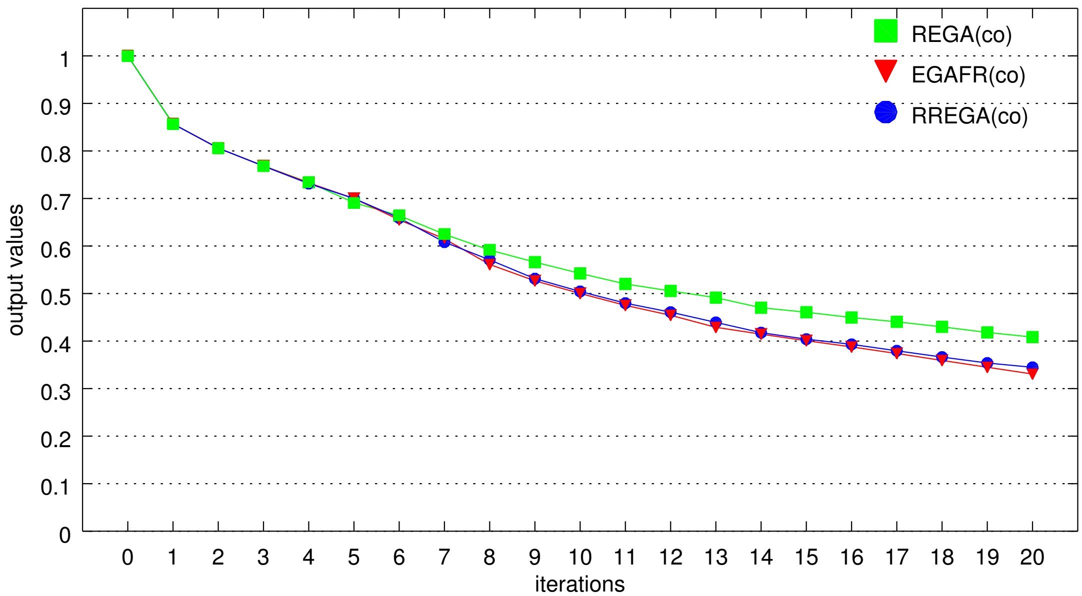

Figure 5.1 First 20 iterations of the REGA, the EGAFR, and the RREGA

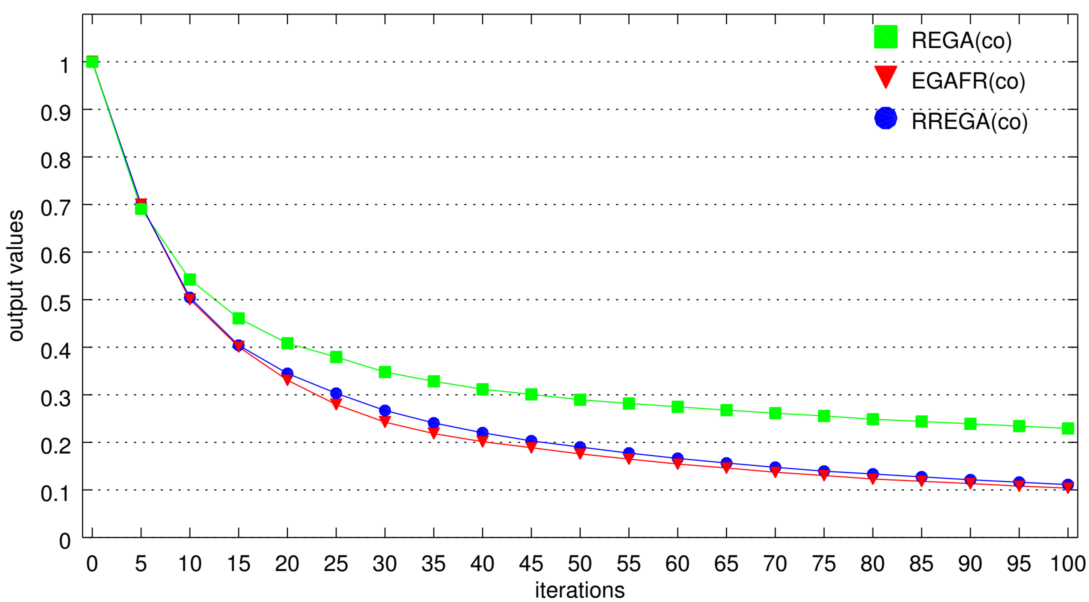

Figure 5.2 First 100 iterations of the REGA, the EGAFR, and the RREGA

Example 2. $E(x) = \dfrac{\|f - x\|_3^3 + .01\,\|x\|_{1.2}^{1.2}}{\|f\|_3^3} \;:\; \ell_6^{(100)} \to \mathbb{R}$

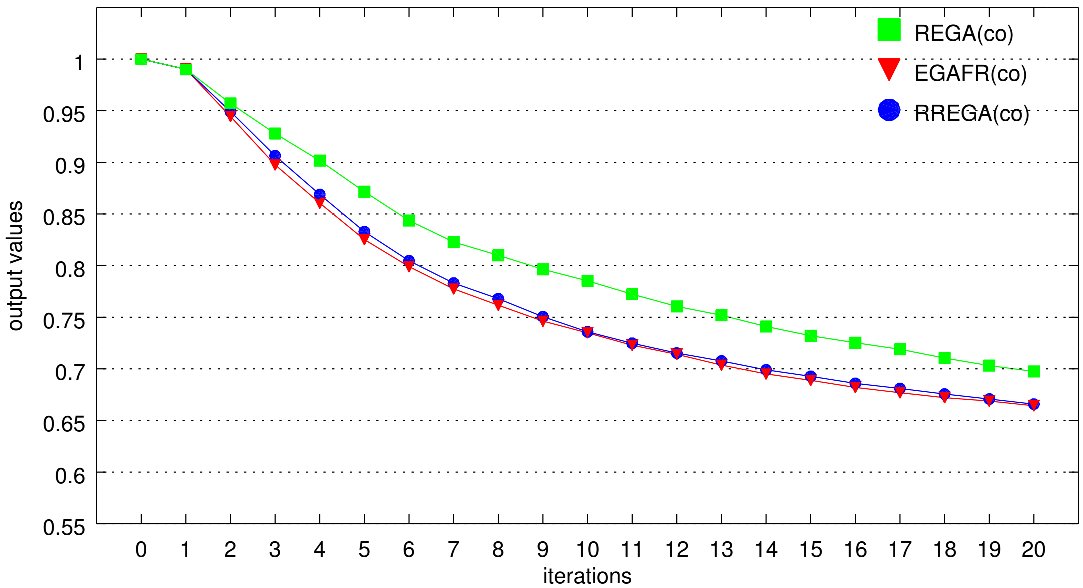

Figure 5.3 First 20 iterations of the REGA, the EGAFR, and the RREGA

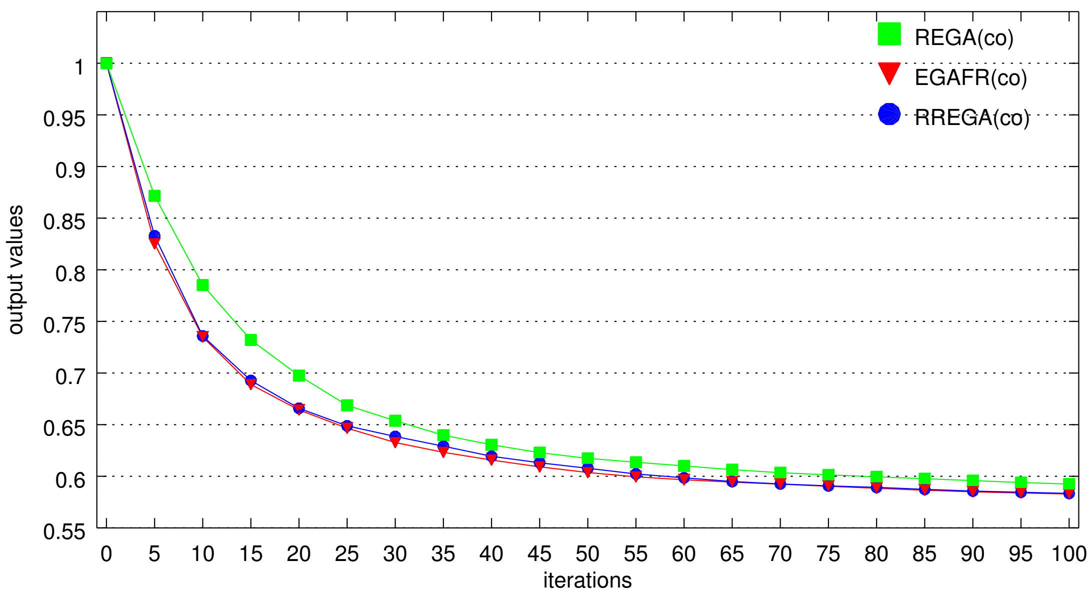

Figure 5.4 First 100 iterations of the REGA, the EGAFR, and the RREGA

Example 3. $E(x) = \dfrac{2\,\|f_1 - x\|_7^7 + \|f_2 - x\|_5^5 + .005\,\|x\|_{1.3}^{1.3}}{2\,\|f_1\|_7^7 + \|f_2\|_5^5} \;:\; \ell_9^{(100)} \to \mathbb{R}$

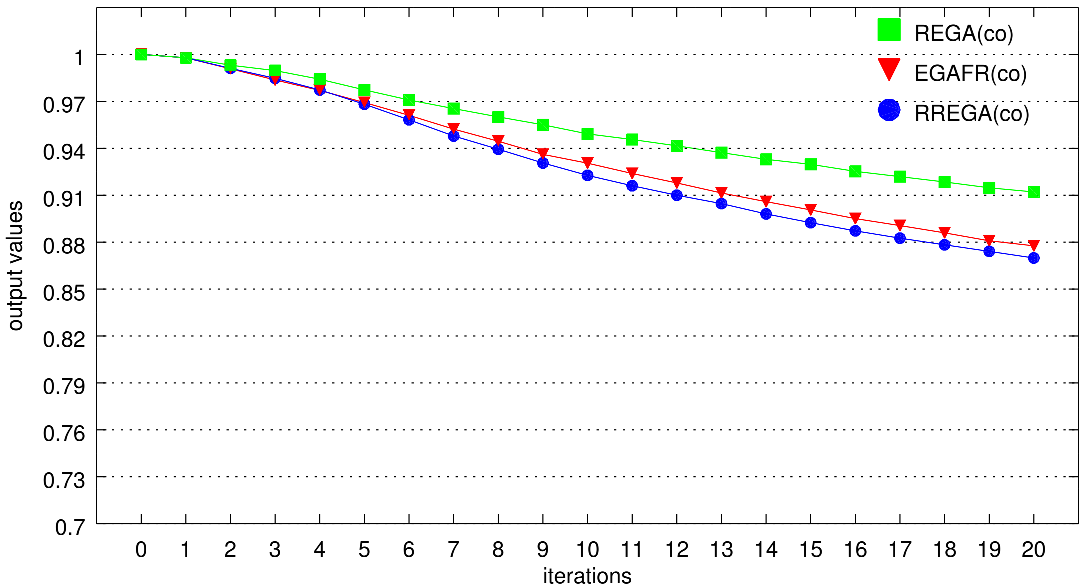

Figure 5.5 First 20 iterations of the REGA, the EGAFR, and the RREGA

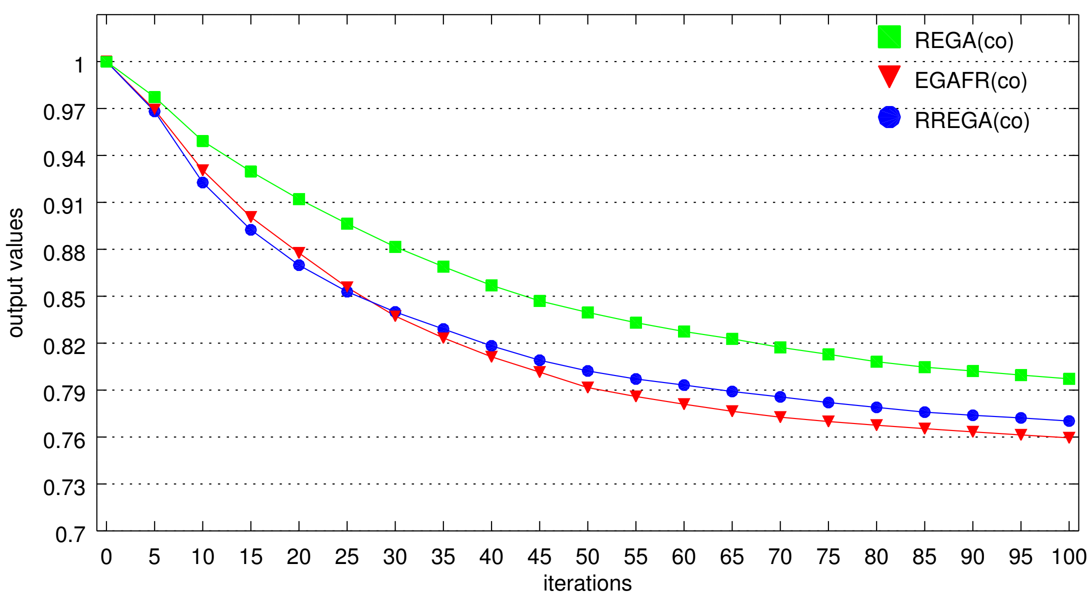

Figure 5.6 First 100 iterations of the REGA, the EGAFR, and the RREGA

5.5 Proofs for Sections 5.2 and 5.3

In this section we prove Theorems 5.2.1, 5.2.2 and 5.3.1. We recall several technical results that were originally stated for a Fréchet-differentiable functions but can be generalized for arbitrary convex functions with minor changes in the proofs.

Lemma N (Temlyakov 2015, Lemma 1.1). *Let $E : X \to \mathbb{R}$ be a convex function and Ω be a convex subset of X. Then for any $x \in \Omega$, $y \in X$, any support functional H_x and $u \geq 0$*

$$0 \leq E(x + uy) - E(x) - uH_x(y) \leq 2\rho(u \left\| y \right\|, \Omega).$$

Lemma O (Temlyakov 2015, Lemma 2.1). *Let $E : X \to \mathbb{R}$ be a convex uniformly smooth on $\Omega \subset X$ function. Let L be a finite-dimensional subspace of X and $z_0 \in \Omega \cap L$ be a point at which E attains its minimum on L, i.e.*

$$z_0 = \operatorname*{argmin}_{x \in L} E(x).$$

Then there exists support functional H_{z_0} such that for any $y \in L$

$$H_{z_0}(y) = 0.$$

Lemma 5.5.1. *Let E be a convex uniformly smooth on Ω function and $\mathcal{D}$ be a dictionary. Let $\epsilon > 0$ and $y \in X$ be any such element that $\|z_0 - y\| < \epsilon$ and $Ay \in A_1(\mathcal{D})$ for some $A = A(y, \mathcal{D}) > 0$. Then the ARREGA(co) of E provides for any $m \geq 0$*

$$E(x_{m+1}) \leq \inf_{\lambda \geq 0} \Big(E(x_m) - \lambda A(E(x_m) - E(z_0) - \epsilon M_\Omega) + 2\rho(\lambda, \Omega) \Big) + 2\delta_{m+1}.$$

Proof of Lemma 5.5.1. The definition of the ARREGA(co) provides for any $m \geq 0$

$$\begin{aligned} E(x_{m+1}) &= E\left(\mu_{m+1}(x_m + \lambda_{m+1}\phi_{m+1})\right) \leq E\left(x_m + \lambda_{m+1}\phi_{m+1}\right) + \delta_m \\ &\leq \inf_{\phi \in \mathcal{D}, \lambda \geq 0} E\left(x_m + \lambda\phi\right) + 2\delta_{m+1}, \end{aligned}$$

and Lemmas N and H provide for any support functional H_{x_m}

$$\begin{aligned}\inf_{\phi\in\mathcal{D},\lambda\geq 0} E\left(x_m+\lambda\phi\right) &\leq \inf_{\phi\in\mathcal{D},\lambda\geq 0} E(x_m)+\lambda H_{x_m}(\phi)+2\rho(\lambda\left\|\phi\right\|,\Omega)\\ &\leq \inf_{\lambda\geq 0} E(x_m)+\lambda H_{x_m}(Ay)+2\rho(\lambda,\Omega)\\ &\leq \inf_{\lambda\geq 0} E(x_m)+\lambda A(H_{x_m}(z_0)+\epsilon M_\Omega)+2\rho(\lambda,\Omega).\end{aligned}$$

Let $H^*_{x_m}$ be the support functional from Lemma O, then

$$H^*_{x_m}(z_0)=H^*_{x_m}(z_0-x_m)\leq E(z_0)-E(x_m),$$

and therefore

$$E(x_{m+1})\leq \inf_{\lambda\geq 0}\Big(E(x_m)-\lambda A(E(x_m)-E(z_0)-\epsilon M_\Omega)+2\rho(\lambda,\Omega)\Big)+2\delta_{m+1}.$$

□

We will also need the following technical result, which is an analogue of Lemma 4.2 from Nguyen and Petrova 2016, rewritten for the ARREGA.

Lemma 5.5.2. *Let $\{a_m\}_{m=0}^\infty$ and $\{\delta_m\}_{m=1}^\infty$ be such sequences of non-negative numbers that for any $m\geq 1$*

$$a_m\leq a_{m-1}\left(1-\beta a_{m-1}^p\right)+\delta_m$$

for some $\beta>0$ and $p\geq 1$. Then for any $m\geq 1$

$$a_m\leq\left(\beta^{-1/p}+\sum_{n=1}^m\delta_n n^{1/p}\right)m^{-1/p}.$$

Proof of Lemma 5.5.2. Denote for each $m\geq 1$

$$b_m=a_{m-1}\left(1-\beta a_{m-1}^p\right)\geq 0.$$

We will use induction by m to prove the desired estimate. The base of induction holds since $(1-x)^{-p}\geq(1-x)^{-1}\geq 1+x$ for any $0\leq x<1$, and thus

$$b_1^{-p}\geq a_0^{-p}(1-\beta a_0^p)^{-p}\geq a_0^{-p}(1+\beta a_0^p)=(a_0^{-p}+\beta)\geq\beta.$$

Hence

$$a_1 \leq b_1 + \delta_1 \leq \beta^{-1/p} + \delta_1.$$

Assume that the estimate holds for some m. Then

$$\begin{aligned} b_{m+1}^{-p} &\geq a_m^{-p}(1-\beta a_m^p)^{-p} \geq a_m^{-p}(1+\beta a_m^p) = a_m^{-p} + \beta \\ &\geq \left(\beta^{-1/p} + \sum_{n=1}^{m} \delta_n n^{1/p}\right)^{-p} m + \beta \\ &\geq \left(\beta^{-1/p} + \sum_{n=1}^{m} \delta_n n^{1/p}\right)^{-p} (m+1). \end{aligned}$$

Thus

$$b_{m+1} \leq \left(\beta^{-1/p} + \sum_{n=1}^{m} \delta_n n^{1/p}\right)(m+1)^{-1/p}$$

and

$$a_{m+1} \leq b_{m+1} + \delta_{m+1} = \left(\beta^{-1/p} + \sum_{n=1}^{m+1} \delta_n n^{1/p}\right)(m+1)^{-1/p},$$

which proves the induction assumption. □

We now prove Theorem 5.3.1. Theorem 5.2.1 follows by taking $\delta_n = 0$ for all $n \geq 1$.

Proof of Theorem 5.3.1. Let $\{\delta_n\}_{n=1}^{\infty}$ be a weakness sequence with $\lim_{n\to\infty} \delta_n = 0$. Assume that for some uniformly smooth convex function $E(x)$ and a dictionary $\mathcal{D}$ the ARREGA(co) does not converge, i.e. $E(x_m) \not\to E(z_0)$ as $m \to \infty$. Then there exists $\alpha \in (0; E(0)]$ and a subsequence $\{n_k\}_{k=1}^{\infty}$ such that for any $k \geq 1$

$$E(x_{n_k}) - E(z_0) > \alpha. \tag{5.3}$$

Let $\epsilon = \alpha/4M_\Omega$ and take any such $y \in X$ that $Ay \in A_1(D)$ and $\|y - z_0\| < \epsilon$ for some $A = A(y, \mathcal{D}) \in (0; 1]$. Since E is uniformly smooth on Ω, there exists $\lambda_0 \in (0; 1]$ such that

$$\rho(\lambda_0, \Omega) \leq \frac{\lambda_0 A\alpha}{16}. \tag{5.4}$$

For a fixed value of λ_0 find such $N \in \mathbb{N}$ that for any $n \geq N$

$$\delta_n \leq \min\left\{\frac{\alpha}{4}, \frac{\lambda_0 A\alpha}{32}\right\}. \tag{5.5}$$

We will show that for sufficiently big k

$$E(x_{n_k}) - E(z_0) \leq \alpha.$$

Indeed, take any $m \geq N$. If $E(x_m) - E(z_0) < \alpha/2$, then by the definition of the ARREGA(co) and estimate (5.5) we have

$$E(x_{m+1}) \leq E(x_m) + 2\delta_{m+1} \leq E(x_m) + \alpha/2. \tag{5.6}$$

If $E(x_m) - E(z_0) \geq \alpha/2$, then Lemma 5.5.1 and estimates (5.4) and (5.5) provide

$$\begin{aligned} E(x_{m+1}) &\leq \inf_{\lambda \geq 0} \Big(E(x_m) - \lambda A(E(x_m) - E(z_0) - \alpha/4) + 2\rho(\lambda, \Omega) \Big) + 2\delta_{m+1} \\ &\leq \inf_{\lambda \geq 0} \left(E(x_m) - \lambda A\alpha/4 + 2\rho(\lambda, \Omega) \right) + 2\delta_{m+1} \\ &\leq E(x_m) - \lambda_0 A\alpha/8 + 2\delta_{m+1} \leq E(x_m) - \lambda_0 A\alpha/16. \end{aligned} \tag{5.7}$$

Note that for any $m \geq N$ we get from the definition of the ARREGA(co)

$$E(x_m) \leq \inf_{\mu \geq 0} E\left(\mu(x_{m-1} + \lambda_m \phi_m)\right) + \delta_m \leq E(0) + \delta_m \leq E(0) + \alpha/4.$$

Let $K = \lceil 16(E(0) - E(z_0))/\lambda_0 A\alpha \rceil$, then estimates (5.6) and (5.7) guarantee

$$E(x_{n_{N+K}}) - E(z_0) \leq \alpha,$$

which contradicts assumption (5.3). Therefore the ARREGA(co) of f with respect to $\mathcal{D}$ converges.

Now assume that $z_0 \in A_1(\mathcal{D})$ and $E \in \mathcal{P}_q(\Omega, X)$, i.e. $\rho(u) \leq \gamma u^q$ for some $\gamma > 0$. Denote $p = q/(q-1) \in [2, +\infty)$. Then Lemma 5.5.1 provides with $y = z_0$, $A = 1$, and $\epsilon = 0$

$$\begin{aligned} E(x_{m+1}) &\leq \inf_{\lambda \geq 0} E(x_m) - \lambda(E(x_m) - E(z_0)) + 2\rho(\lambda, \Omega) + 2\delta_{m+1} \\ &\leq \inf_{\lambda \geq 0} E(x_m) - \lambda(E(x_m) - E(z_0)) + 2\gamma\lambda^q + 2\delta_{m+1} \\ &= E(x_m) - (q-1)(2\gamma)^{1-p} q^{-p} (E(x_m) - E(z_0))^p + 2\delta_{m+1}, \end{aligned}$$

where the infimum is attained at

$$\lambda = \left(\frac{E(x_m) - E(z_0)}{2\gamma q}\right)^{p-1} \geq 0. \tag{5.8}$$

Denote $a_m = E(x_m) - E(z_0)$, then

$$a_{m+1} \leq a_m - (q-1)(2\gamma)^{1-p} q^{-p} a_m^p + 2\delta_{m+1} \tag{5.9}$$

and Lemma 5.5.2 guarantees for any $m \geq 1$

$$\begin{aligned} a_m &\leq \left(((q-1)(2\gamma)^{1-p}q^{-p})^{1-q} + 2\sum_{n=1}^{m} \delta_n n^{q-1}\right) m^{1-q} \\ &\leq \left(8\gamma + 2\sum_{n=1}^{m} \delta_n n^{q-1}\right) m^{1-q}. \end{aligned}$$

□

Proof of Proposition 5.3.3. The proposition follows from estimate (5.9) and the following lemma.

Lemma P (Temlyakov 2016, Lemma 3.4)**.** *Let $q \in (1, 2]$ and $p = q/(q-1)$. Assume that a sequence $\{\delta_n\}_{n=1}^{\infty}$ is such that*

$$\delta_n \leq \delta n^{-q}.$$

Suppose a nonnegative sequence $\{a_n\}_{n=1}^{\infty}$ satisfies the inequalities

$$a_{n+1} \leq a_n - \beta a_n^p + \delta_{n+1}$$

with some $\beta \in (0; 1]$. Then

$$a_n \leq C(q, \delta, \beta, a_0)\, n^{1-q}.$$

It follows from the proof of Lemma P (see Lemma 2.1 in Temlyakov 1999 and Remark 3.1 in Temlyakov 2016) that one can take

$$C(q, \delta, \beta, a_0) = 2^{(1-q^2)/2} A,$$

where $A > 0$ is such that

$$\beta A^p - 2A - \delta > 0 \quad \text{and} \quad A > a_0.$$

In particular, taking

$$A = \max\left\{\left(\frac{4}{\beta}\right)^{q-1}, \left(\frac{2\delta}{\beta}\right)^{1/p}, a_0\right\}$$

satisfies both conditions. Substituting values from estimate (5.9) and performing straightforward estimates completes the proof. □

Proof of Theorem 5.2.2. Denote $\widetilde{x}_m = x_{m-1} + \lambda_m \phi_m$, i.e. the m-th approximation before the rescaling. Then since $\|x\| \le C_\Omega$ for any $x \in \Omega$, we get

$$\lambda_m = \|\lambda_m \phi_m\| = \|\widetilde{x}_m - x_{m-1}\| \le C_\Omega + \|x_{m-1}\|$$

and

$$\mu_m = \frac{\|x_m\|}{\|\widetilde{x}_m\|} \le \frac{C_\Omega}{\|\widetilde{x}_m\|} = \frac{C_\Omega}{\|x_{m-1} + \lambda_m \phi_m\|}.$$

Assume now that $E \in \mathcal{P}_q(X, \Omega)$ and $z_0 \in A_1(\mathcal{D})$. Then it follows from the proof of Theorem 5.2.1 (see (5.8)) that for any $m \ge 1$ we can take

$$\begin{aligned} \lambda_m &= \left(\frac{E(x_{m-1}) - E(z_0)}{2\gamma q}\right)^{p-1} \\ &\le \left(\frac{H_{x_{m-1}}(x_{m-1} - z_0)}{2\gamma q}\right)^{p-1} \le \left(\frac{M_\Omega(1 + \|x_{m-1}\|)}{2\gamma q}\right)^{p-1}. \end{aligned}$$

Additionally, if $E(x) \ge C_E$ for any $x \in \Omega$, we obtain

$$\lambda_m \le \left(\frac{E(x_{m-1}) - C_E}{2\gamma q}\right)^{p-1}.$$

□

Bibliography

Beauzamy, B. (1982). *Introduction to Banach spaces and their geometry*. Vol. 86. Notas de matematica (Amsterdam, Netherlands). North-Holland Publishing Company.

Borwein, J. et al. (2009). "Uniformly convex functions on Banach spaces". In: *Proceedings of the American Mathematical Society* 137.3, pp. 1081–1091.

Clarkson, K.L. (2010). "Coresets, sparse greedy approximation, and the Frank–Wolfe algorithm". In: *ACM Transactions on Algorithms (TALG)* 6.4, p. 63.

Dereventsov, A.V. (2012). "Comparison of the convergence rate of pure greedy and orthogonal greedy algorithms". In: *Mathematical Notes* 92.3, pp. 485–489.

— (2016). "On the Approximate Weak Chebyshev Greedy Algorithm in uniformly smooth Banach spaces". In: *Journal of Mathematical Analysis and Applications* 436.1, pp. 288–304.

— (2017). "On the generalized Approximate Weak Chebyshev Greedy Algorithm". In: *Studia Mathematica* 237.2, pp. 153–175.

DeVore, R.A. and V.N. Temlyakov (1996). "Some remarks on greedy algorithms". In: *Advances in computational Mathematics* 5.1, pp. 173–187.

— (2014). "Convex optimization on Banach spaces". In: *arXiv preprint:1401.0334*.

Dilworth, S.J., D. Kutzarova, and V.N. Temlyakov (2002). "Convergence of some greedy algorithms in Banach spaces". In: *Journal of Fourier Analysis and Applications* 8.5, pp. 489–506.

Dilworth, S.J. et al. (2008). "Weak convergence of greedy algorithms in Banach spaces". In: *Journal of Fourier Analysis and Applications* 14.5, pp. 609–628.

Donahue, M.J. et al. (1997). "Rates of convex approximation in non-Hilbert spaces". In: *Constructive Approximation* 13.2, pp. 187–220.

Dowling, P.N. et al. (1997). “The optimality of James’s distortion theorems”. In: *Proceedings of the American Mathematical Society* 125.1, pp. 167–174.

Dubinin, V.V. (1997). *Greedy algorithms and applications*. Doctoral Dissertation. University of South Carolina.

Figueiredo, M.A.T., R.D. Nowak, and S.J. Wright (2007). “Gradient projection for sparse reconstruction: Application to compressed sensing and other inverse problems”. In: *IEEE Journal of selected topics in signal processing* 1.4, pp. 586–597.

Frank, M. and P. Wolfe (1956). “An algorithm for quadratic programming”. In: *Naval research logistics quarterly* 3.1-2, pp. 95–110.

Galatenko, V.V. and E.D. Livshitz (2005). “Generalized approximate weak greedy algorithms”. In: *Mathematical Notes* 78.1, pp. 170–184.

Gao, Z. and G. Petrova (2015). “Rescaled Pure Greedy Algorithm for Convex Optimization”. In: *arXiv preprint arXiv: 1505.03606*.

Gribonval, R. and M. Nielsen (2001). “Approximate weak greedy algorithms”. In: *Advances in Computational Mathematics* 14.4, pp. 361–378.

Jaggi, M. (2013). “Revisiting Frank–Wolfe: Projection-Free Sparse Convex Optimization.” In: *ICML (1)*, pp. 427–435.

Livshitz, E.D. (2003). “Convergence of greedy algorithms in Banach spaces”. In: *Mathematical Notes* 73.3, pp. 342–358.

— (2010). “The convergence of the greedy algorithm with respect to the Haar system in the space $L_p(0,1)$”. In: *Sbornik: Mathematics* 201.2, p. 253.

Mallat, S.G. and Z. Zhang (1993). “Matching pursuits with time-frequency dictionaries”. In: *IEEE Transactions on signal processing* 41.12, pp. 3397–3415.

Nemirovski, A.S. (1995). “Information-based complexity of convex programming”. In: *Lecture Notes*.

Nguyen, H. and G. Petrova (2016). “Greedy strategies for convex optimization”. In: *Calcolo*, pp. 1–18.

Pati, Y.C., R. Rezaiifar, and P.S. Krishnaprasad (1993). “Orthogonal matching pursuit: Recursive function approximation with applications to wavelet decomposition”. In: *Signals, Systems and Computers, 1993. 1993 Conference Record of The Twenty-Seventh Asilomar Conference on*. IEEE, pp. 40–44.

Petrova, G. (2015). “Rescaled pure greedy algorithm for Hilbert and Banach spaces”. In: *Applied and Computational Harmonic Analysis* 41.3, pp. 852–866.

Shalev-Shwartz, S., N. Srebro, and T. Zhang (2010). “Trading accuracy for sparsity in optimization problems with sparsity constraints”. In: *SIAM Journal on Optimization* 20.6, pp. 2807–2832.

Temlyakov, V.N. (1999). “Greedy algorithms and m-term approximation with regard to redundant dictionaries”. In: *Journal of Approximation Theory* 98.1, pp. 117–145.

— (2000). “Weak greedy algorithms”. In: *Advances in Computational Mathematics* 12.2, pp. 213–227.

— (2001). “Greedy algorithms in Banach spaces”. In: *Advances in Computational Mathematics* 14.3, pp. 277–292.

— (2003). “Nonlinear methods of approximation”. In: *Foundations of Computational Mathematics* 3.1, pp. 33–107.

— (2005). “Greedy-type approximation in Banach spaces and applications”. In: *Constructive Approximation* 21.2, pp. 257–292.

— (2011). *Greedy approximation*. Cambridge Monographs on Applied and Computational Mathematics. Cambridge University Press.

— (2015). “Greedy approximation in convex optimization”. In: *Constructive Approximation* 41.2, pp. 269–296.

— (2016). “Convergence and rate of convergence of some greedy algorithms in convex optimization”. In: *Proceedings of the Steklov Institute of Mathematics* 293.1, pp. 325–337.

Tewari, A., P.K. Ravikumar, and I.S. Dhillon (2011). “Greedy algorithms for structurally constrained high dimensional problems”. In: *Advances in Neural Information Processing Systems*, pp. 882–890.

Zhang, T. (2003). “Sequential Greedy Approximation for certain Convex Optimization problems”. In: *IEEE Transactions on Information Theory* 49.3, pp. 682–691.