Acta Numerica (2008), pp. 235–409
doi: 10.1017/S0962492906380014

© Cambridge University Press, 2008
Printed in the United Kingdom

Greedy approximation

V. N. Temlyakov
University of South Carolina,
Columbia, 29208, USA
E-mail: temlyak@math.sc.edu

In this survey we discuss properties of specific methods of approximation that belong to a family of greedy approximation methods (greedy algorithms). It is now well understood that we need to study nonlinear sparse representations in order to significantly increase our ability to process (compress, denoise, *etc.*) large data sets. Sparse representations of a function are not only a powerful analytic tool but they are utilized in many application areas such as image/signal processing and numerical computation. The key to finding sparse representations is the concept of m-term approximation of the target function by the elements of a given system of functions (dictionary). The fundamental question is how to construct good methods (algorithms) of approximation. Recent results have established that greedy-type algorithms are suitable methods of nonlinear approximation in both m-term approximation with regard to bases, and m-term approximation with regard to redundant systems. It turns out that there is one fundamental principle that allows us to build good algorithms, both for arbitrary redundant systems and for very simple well-structured bases, such as the Haar basis. This principle is the use of a greedy step in searching for a new element to be added to a given m-term approximant.

CONTENTS

CHAPTER 1: Greedy approximation with respect to bases

CHAPTER 2: GA with respect to dictionaries: Hilbert spaces

CHAPTER 3: GA with respect to dictionaries: Banach spaces

Preface

This section provides a general introduction. Each chapter has its own more specific introduction. A generic problem of mathematical and numerical analysis is to approximate a function f from a Banach space X in the norm $\|\cdot\|$ of this space. There are two major approaches to this problem that lead to two different branches of mathematical analysis. In approach (I) we begin with an assumption of how an approximant should look. In other words, in approach (I) we specify the form of an approximant. In approach (II) we begin with an assumption on the information available about f. Here are typical settings that fall into approach (I).

(Ia) An approximant comes from a given linear subspace L_n of dimension n (algebraic polynomials of degree $n-1$, trigonometric polynomials of appropriate order, splines with $n-1$ fixed knots).

(Ib) An approximant comes from a nonlinear set (rational functions; m-term approximant with respect to a given system, splines with fixed number of free knots).

The following are typical settings for approach (II).

(IIa) Information on f is given by a vector $(f(x_1),\dots,f(x_n))$, for some given set $(x_1,\dots,x_n)$ of points, or some set that we can choose depending on the problem.

(IIb) The above setting (IIa) has a more general formulation with the functionals $f(x_j)$ replaced by arbitrary linear functionals $\lambda_j(f)$.

In this survey we mostly concentrate on approach (I). We will only touch upon approach (II) in Section 2.9 in a discussion of compressed sensing. Approach (II) is the main issue for information-based complexity (see Traub, Wasilkowski and Wozniakowski (1988)). A very important question in approach (I) is how to choose an appropriate form of approximants. This question has been intensely studied in approximation theory, and resulted in the invention of the concept of *width.* In 1936 A. N. Kolmogorov introduced the following quantity (known as Kolmogorov's width) for a compact $F \subset X$:

$$d_n(F,X) := \inf_{L_n} \sup_{f\in F} \inf_{a\in L_n} \|f-a\|,$$

where L_n is an n-dimensional linear subspace of X. The Kolmogorov width $d_n(F,X)$ of a compact F is an important characteristic of F that states that the best we can achieve in approximating functions from F by elements of linear subspaces of dimension n is $d_n(F,X)$. Therefore, if one can find an n-dimensional subspace L_n^* and an approximation method $A_n : F \to L_n^*$ such that, for any $f \in F$,

$$\|f - A_n(f)\| \le (1+\epsilon) d_n(F,X),$$

then A_n is an almost ideal approximation method for F with respect to the Kolmogorov width. Thus, the concept of width provides a very nice theoretical way to compare optimal approximation methods. The major drawback of this approach from a practical point of view is that, in order to initialize a procedure of selection of L_n^* and A_n, we need to know the function class F. In many contemporary practical problems we have no idea which class to choose in place of F.

There are two ways to overcome the above problem. The first one is to return (in spirit) to the classical setting that goes back to Chebyshev and Weierstrass. In this setting, we fix *a priori* the form of the approximant (say, approximation by algebraic polynomials of degree n, as in the case of Chebyshev and Weierstrass) and look for an approximation method that is optimal, or near-optimal, for each individual function from X. For example, the approximation method that picks the algebraic polynomial of degree n of best approximation to f in X is an optimal method of approximation by algebraic polynomials of degree n. However, such an obvious optimal method of approximation may not be good from the perspective of practical implementation. This leads to the following natural setting. We specify not only the form of the approximant, but also choose a specific method of approximation (for instance, one known to be suitable for practical implementation). Now, we have a precise mathematical problem of studying the efficiency of our specific method of approximation. We discuss this problem in detail here. It turns out that a convenient and flexible way of measuring the efficiency of a specific approximation method is to derive the corresponding Lebesgue-type inequalities. Remember that we would like this method to work for all functions; therefore, it should at least converge for each $f \in X$, and hence convergence is a fundamental theoretical problem. In this survey we thoroughly discuss the problem of convergence for greedy algorithms.

The second way to overcome the above-mentioned drawback of a method based on the concept of width consists in weakening the *a priori* assumption that f is an element of F. Instead of looking for an approximation method that is optimal (or near-optimal) for a given single class F, we look for an approximation method that is near-optimal for each class from a given collection $\mathcal{F}$ of classes. Such a method is called *universal* for $\mathcal{F}$. Universal algorithms have been studied in approximation theory (see Temlyakov (1988, 2003*a*)) and in learning theory (see, for instance, Temlyakov (2005*c*)). In this survey we do not further discuss universal algorithms, but refer the reader to the survey of Temlyakov (2003*a*).

In this survey we discuss properties of specific methods of approximation that belong to a family of greedy approximation methods (greedy algorithms). Realizing approach (I) mentioned above, we need to specify the form of the approximant. We use a concept of *sparsity* in dealing with this problem. It is now well understood that we need to study nonlinear sparse representations in order to significantly increase our ability to process (compress, denoise, *etc.*) large data sets. Sparse representations are not only a powerful analytic tool but are used in many application areas, such as image/signal processing and numerical computation. The key to finding sparse representations is the concept of the m-term approximation of the target function by the elements of a given system of functions: a dictionary. Since the elements of the dictionary used in the m-term approximation are allowed to depend on the function being approximated, this type of approximation is very efficient when the approximants can be found. Thus, we specify the form of our approximant as an m-term approximant with regard to a given system of functions. It is clear that this method of approximation is a particular case of nonlinear methods of approximation.

The past decade has seen great success in studying nonlinear approximation, motivated by numerous applications: see the surveys by DeVore (1998) and Temlyakov (2003*a*). Nonlinear approximation is important in applications because of its concise representations and increased computational efficiency. Two types of nonlinear approximation are frequently employed in applications. Adaptive methods are used in PDE solvers, while m-term approximation, considered here, is used in image/signal/data processing, as well as in the design of neural networks. The fundamental question of nonlinear approximation is how to devise good constructive methods, or algorithms, of nonlinear approximation. This problem has two levels of nonlinearity. The first level of nonlinearity is m-term approximation with regard to bases. In this problem one can use the unique function expansion with regard to a given basis to build an approximant. Nonlinearity enters by looking for m-term approximants with terms (*i.e.*, basis elements in the approximant) allowed to depend on the given function. We discuss m-term approximation with regard to bases in detail in Chapter 1. On the second level of nonlinearity, we replace a basis by a more general system which is not necessarily minimal, for example, a redundant system, or dictionary. This setting is much more complicated than the bases case; however, there is a solid justification of the importance of redundant systems in both theoretical questions and in practical applications: see, for instance, Schmidt (1906), Huber (1985) and Donoho (2001). In Chapters 2 and 3 we discuss approximation by linear combinations of elements that are taken from a redundant (overcomplete) system of elements. We give a brief discussion of the question: Why do we need redundant systems? Answering this question, we first mention three classical redundant systems that are used in different areas of mathematics. Perhaps the first example of m-term approximation with regard to a redundant dictionary was considered by Schmidt (1906), who considered the approximation of functions $f(x,y)$ of two variables by bilinear forms $\sum_{i=1}^{m} u_i(x)v_i(y)$ in $L_2([0,1]^2)$. This problem is closely connected to properties of the integral operator $J_f(g) := \int_0^1 f(x,y)g(y)\,\mathrm{d}y$ with kernel $f(x,y)$.

Another example which is well known in statistics is the projection pursuit regression problem. We formulate the related setting in the language of function theory. Given a bounded domain $\Omega \subset \mathbb{R}^d$, the problem is to approximate a given function $f \in L_2(\Omega)$ by a sum of ridge functions, *i.e.*, by $\sum_{j=1}^m r_j(\omega_j \cdot x)$, for $x, \omega_j \in \mathbb{R}^d$, $j = 1, \dots, m$, where r_j, $j = 1, \dots, m$, are univariate functions.

The third example is from signal processing. In signal processing the most popular methods of approximation are wavelets and the system of Gabor functions $\{g_{a,b}(x-c) : g_{a,b}(x) := \mathrm{e}^{\mathrm{i}ax}\mathrm{e}^{-bx^2},\ a, c \in \mathbb{R},\ b \in \mathbb{R}_+\}$. The Gabor system gives more flexibility in constructing an approximant but it is a redundant, not minimal, system. It also seems natural to use redundant systems in modelling analysing elements for the visual system; see the discussion in Donoho (2001).

Thus, in order to address the contemporary needs of approximation theory and computational mathematics, a very general model of approximation with regard to a redundant system, or dictionary, has been considered in many recent papers. As such a model, we choose a Banach space X whose elements are our target functions, and an approximating system which can be any subset $\mathcal{D}$ of elements of this space such that the closure of $\operatorname{span} \mathcal{D}$ coincides with X. We would like to have an algorithm to construct m-term approximants that, at each step, adds only one new element from $\mathcal{D}$ and keeps elements of $\mathcal{D}$ previously obtained. This requirement is an analogue of *on-line* computation that is very desirable in practical algorithms. Clearly, we are looking for good algorithms which, at least, converge for each target function. It is not obvious that such an algorithm exists in a setting at the above level of generality (X, $\mathcal{D}$ are arbitrary).

The fundamental question is how to construct good methods, or algorithms, of approximation. Recent results have established that greedy-type algorithms are suitable methods of nonlinear approximation, in both m-term approximation with regard to bases, and m-term approximation with regard to redundant systems. It turns out that there is one fundamental principle that allows us to build good algorithms both for arbitrary redundant systems and for very simple well-structured bases such as the Haar basis. This principle is the use of a greedy step in searching for a new element $g_m(f) \in \mathcal{D}$ to be added to a given m-term approximant, by which we mean that $g_m(f) \in \mathcal{D}$ should maximize a certain functional determined by information from the previous steps of the algorithm. We obtain different types of greedy algorithms by varying the above-mentioned functional and also by using different ways of constructing the m-term approximant (*i.e.*, choosing coefficients of the linear combination) from the previously found m elements of the dictionary. In Chapters 2 and 3 we present different greedy-type algorithms, beginning with a very simple and very natural Pure Greedy Algorithm in a Hilbert space, and ending with its rather complicated modifications in a Banach space. The general goal of different modifications is to prepare the corresponding greedy algorithms for practical implementation. We discuss this issue in detail in Chapters 2 and 3.

It is known that in many numerical problems, users are satisfied with a Hilbert space setting and do not consider a more general setting in a Banach space. We now give one remark that justifies our interest in Banach spaces. The first argument is an *a priori* argument that the spaces L_p are very natural, and should be studied along with the L_2-space. The second argument is an *a posteriori* argument. The study of greedy approximation in Banach spaces discovered that one very important characteristic of a Banach space X that governs the behaviour of greedy approximations is the *modulus of smoothness* $\rho(u)$ of X (see Section 3.1 for details). It is known that the spaces L_p, $2 \leq p < \infty$ have moduli of smoothness of the same order u^2. Thus, many results that are known for the Hilbert space L_2, and were proved using the special structure of a Hilbert space, can be generalized to the Banach spaces L_p, $2 \leq p < \infty$. The new proofs use only the geometry of the unit sphere of the space expressed in the form $\rho(u) \leq \gamma u^2$.

The theory of greedy approximation is developing rapidly and results are spread over hundreds of papers by different authors. There are several surveys that discuss greedy approximation: see DeVore (1998), Temlyakov (2003*a*), Konyagin and Temlyakov (2002), Wojtaszczyk (2002*a*) and Temlyakov (2006*b*). There are no books on greedy approximation at present. We decided to include in this survey proofs of the most important and typical results. In the majority of cases these proofs are not technically involved and allow the reader to understand a phenomenon much better than merely stating results. We have tried to make the presentation of ideas and techniques of greedy approximation sufficiently systematic to be used in a graduate course on greedy approximation.

We will use C, $C(p,d)$, $C_{p,d}$, *etc.*, to denote various positive constants, the indexes indicating dependence on other parameters. We will use the following symbols for brevity. For two non-negative sequences $a = \{a_n\}_{n=1}^{\infty}$ and $b = \{b_n\}_{n=1}^{\infty}$, the relation, or order inequality, $a_n \ll b_n$ means that there is a number $C(a,b)$ such that, for all n, we have $a_n \leq C(a,b)\, b_n$; and the relation $a_n \asymp b_n$ means that $a_n \ll b_n$ and $b_n \ll a_n$. Other notation is defined in the text itself.

CHAPTER ONE
Greedy approximation with respect to bases

1.1. Introduction

It is well known that in many problems it is very convenient to represent a function by a series with respect to a given system of functions. For example, in 1807 Fourier suggested representing a 2π-periodic function by its series (now known as the Fourier series) with respect to the trigonometric system. A very important feature of the trigonometric system that made it attractive for the representation of periodic functions is orthogonality. For an orthonormal system $\mathcal{B} := \{b_n\}_{n=1}^{\infty}$ of a Hilbert space H with an inner product $\langle \cdot, \cdot \rangle$, one can construct a Fourier series of an element f in the following way:

$$f \sim \sum_{n=1}^{\infty} \langle f, b_n \rangle b_n. \tag{1.1.1}$$

If the system $\mathcal{B}$ is a basis for H, then the series in (1.1.1) converges to f in H and (1.1.1) provides the unique representation

$$f = \sum_{n=1}^{\infty} \langle f, b_n \rangle b_n \tag{1.1.2}$$

of f with respect to $\mathcal{B}$. This representation has nice approximative properties. By Parseval's identity,

$$\|f\|^2 = \sum_{n=1}^{\infty} |\langle f, b_n \rangle|^2, \tag{1.1.3}$$

we obtain a convenient way to calculate, or estimate, the norm $\|f\|$.

It is known that the partial sums

$$S_m(f, \mathcal{B}) := \sum_{n=1}^{m} \langle f, b_n \rangle b_n \tag{1.1.4}$$

provide the best approximation, that is, defining

$$E_m(f, \mathcal{B}) := \inf_{\{c_n\}} \left\| f - \sum_{n=1}^{m} c_n b_n \right\| \tag{1.1.5}$$

to be the distance of f from the span$\{b_1, \dots, b_m\}$, we have

$$\|f - S_m(f, \mathcal{B})\| = E_m(f, \mathcal{B}). \tag{1.1.6}$$

Identities (1.1.3) and (1.1.6) are fundamental properties of Hilbert spaces and their orthonormal bases. These properties make the theory of approximation in H from the span$\{b_1, \dots, b_m\}$, or linear approximation theory, simple and convenient.

The situation becomes more complicated when we replace a Hilbert space H by a Banach space X. In a Banach space X we consider a Schauder basis Ψ instead of an orthonormal basis $\mathcal{B}$ in H. In Section 1.2 we discuss Schauder bases in detail. If $\Psi := \{\psi_n\}_{n=1}^{\infty}$ is a Schauder basis for X, then for any $f \in X$ there exists a unique representation

$$f = \sum_{n=1}^{\infty} a_n(f)\psi_n$$

that converges in X.

Theorem 1.2.3 states that the partial sum operators S_m, defined by

$$S_m(f, \Psi) := \sum_{n=1}^{m} a_n(f)\psi_n,$$

are uniformly bounded operators from X to X. In other words, there exists a constant B such that, for any $f \in X$ and any m, we have

$$\|S_m(f, \Psi)\| \le B\|f\|.$$

This inequality implies the following analogue of (1.1.6): for any $f \in X$,

$$\|f - S_m(f, \Psi)\| \le (B + 1)E_m(f, \Psi), \tag{1.1.7}$$

where

$$E_m(f, \Psi) := \inf_{\{c_n\}} \left\| f - \sum_{n=1}^{m} c_n\psi_n \right\|.$$

Inequality (1.1.7) shows that the $S_m(f, \Psi)$ provides near-best approximation from $\operatorname{span}\{\psi_1, \dots, \psi_m\}$. Thus, if we are satisfied with near-best approximation instead of best approximation, then the linear approximation theory with respect to Schauder bases becomes simple and convenient. The partial sums $S_m(\cdot, \Psi)$ provide near-best approximation for any individual element of X.

Motivated by computational issues, researchers became interested in nonlinear approximation with regard to a given system instead of linear approximation. For example, in the case of representation (1.1.2) in a Hilbert space, one can take an approximant of the form

$$S_\Lambda(f, \mathcal{B}) := \sum_{n\in\Lambda} \langle f, b_n\rangle b_n, \quad |\Lambda| = m,$$

instead of an approximant $S_m(f, \mathcal{B})$ from an m-dimensional linear subspace. Then the two approximants $S_m(f, \mathcal{B})$ and $S_\Lambda(f, \mathcal{B})$ have the same sparsity: both are linear combinations of m basis elements. However, we can achieve a better approximation error with $S_\Lambda(f, \mathcal{B})$ than with $S_m(f, \mathcal{B})$ if we choose Λ in the right way. In the case of a Hilbert space and an orthonormal basis $\mathcal{B}$, an optimal choice Λ_m of Λ is obvious: Λ_m is a set of m indices with the biggest (in absolute value) coefficients $\langle f, b_n\rangle$. Then, by Parseval's identity (1.1.3), we obtain

$$\|f - S_{\Lambda_m}(f, \mathcal{B})\| \le \|f - S_m(f, \mathcal{B})\|.$$

Also, it is clear that the $S_{\Lambda_m}(f, \mathcal{B})$ realizes the best m-term approximation of f with regard to $\mathcal{B}$,

$$\|f - S_{\Lambda_m}(f, \mathcal{B})\| = \sigma_m(f, \mathcal{B}) := \inf_{\Lambda:|\Lambda|=m} \inf_{\{c_n\}} \left\| f - \sum_{n\in\Lambda} c_n b_n \right\|. \tag{1.1.8}$$

The approximant $S_{\Lambda_m}(f, \mathcal{B})$ can be obtained as a realization of m iterations of the following *greedy approximation step*. For a given $f \in H$ we choose at a *greedy step* an index n_1 with the biggest $|\langle f, b_{n_1}\rangle|$. At a greedy approximation step we build a new element $f_1 := f - \langle f, b_{n_1}\rangle b_{n_1}$.

The identity (1.1.8) shows that the greedy approximation works perfectly in nonlinear approximation in a Hilbert space with regard to an orthonormal basis $\mathcal{B}$.

This chapter is devoted to a systematic study of greedy approximation in Banach spaces. In Section 1.2 we discuss the following natural question. Equation (1.1.8) proves the existence of the best m-term approximant in a Hilbert space with respect to an orthonormal basis. Further, we discuss existence of the best m-term approximant in a Banach space with respect to a Schauder basis. That discussion illustrates that the situation with existence theorems is much more complex in Banach spaces than in Hilbert spaces. We also give some sufficient conditions on a Schauder basis that guarantee existence of the best m-term approximant. However, the problem is far from being completely solved.

The central issue of this chapter is the following question. Which bases are suitable for greedy approximation? Greedy approximation with regard to a Schauder basis is defined in a similar way to the greedy approximation with regard to an orthonormal basis (see above). The greedy algorithm picks the terms with the biggest (in absolute value) coefficients from the expansion

$$f = \sum_{n=1}^{\infty} a_n(f)\psi_n, \tag{1.1.9}$$

and gives a *greedy approximant*

$$G_m(f, \Psi) := S_{\Lambda_m}(f, \Psi) := \sum_{n\in\Lambda_m} a_n(f)\psi_n.$$

Here, Λ_m is such that $|\Lambda_m| = m$ and

$$\min_{n\in\Lambda_m} |a_n(f)| \ge \max_{n\notin\Lambda_m} |a_n(f)|.$$

We note that we need some restrictions on the basis Ψ (see Sections 1.3 and 1.4 for a detailed discussion) in order to be able to run the greedy algorithm for each $f \in X$. It is sufficient to assume that Ψ is normalized. We make this assumption for our further discussion in the Introduction.

An application of the greedy algorithm can also be seen as a rearrangement of the series from (1.1.9) in a special way: according to the size of coefficients. Let

$$|a_{n_1}| \geq |a_{n_2}| \geq \cdots .$$

Then

$$G_m(f, \Psi) = \sum_{j=1}^{m} a_{n_j}(f)\psi_{n_j}.$$

Thus, the greedy approximant $G_m(f, \Psi)$ is a partial sum of the rearranged series

$$\sum_{j=1}^{\infty} a_{n_j}(f)\psi_{n_j}. \tag{1.1.10}$$

An immediate question with (1.1.10) is: When does this series converge? The theory of convergence of rearranged series is a classical topic in analysis. A series converges *unconditionally* if every rearrangement of this series converges. A basis Ψ of a Banach space X is said to be an *unconditional basis* if, for every $f \in X$, its expansion (1.1.9) converges unconditionally. For a set of indices Λ define

$$S_\Lambda(f, \Psi) := \sum_{n\in\Lambda} a_n(f)\psi_n.$$

It is well known that if Ψ is unconditional then there exists a constant K such that, for any Λ,

$$\|S_\Lambda(f, \Psi)\| \leq K\|f\|. \tag{1.1.11}$$

This inequality is similar to $\|S_m(f, \Psi)\| \leq B\|f\|$ and implies an analogue of (1.1.7):

$$\|f - S_\Lambda(f, \Psi)\| \leq (K + 1)E_\Lambda(f, \Psi), \tag{1.1.12}$$

where

$$E_\Lambda(f, \Psi) := \inf_{\{c_n\}} \left\| f - \sum_{n\in\Lambda} c_n\psi_n \right\|.$$

Inequality (1.1.12) indicates that in the case of an unconditional basis Ψ it is sufficient for finding near-best m-term approximant to optimize only over the sets of indices Λ. The greedy algorithm $G_m(\cdot, \Psi)$ gives a simple recipe for building Λ_m: pick the indices with biggest coefficients. In Section 1.3 we discuss in detail when the above simple recipe provides a near-best m-term approximant. It turns out that the assumption that Ψ is merely unconditional does not guarantee that $G_m(\cdot, \Psi)$ provides a near-best m-term approximation. We also discuss a new class of bases (*greedy bases*) that has the property that $G_m(f, \Psi)$ provides a near-best m-term approximation for each $f \in X$. We show that the class of greedy bases is a proper subclass of the class of unconditional bases.

It follows from the definition of an unconditional basis that any rearrangement of the series in (1.1.9) converges. It is known that it converges to f. The rearrangement (1.1.10) is a specific rearrangement of (1.1.9). Clearly, for an unconditional basis Ψ, (1.1.10) converges to f. It turns out that unconditionality of Ψ is not a necessary condition for convergence of (1.1.10) for each $f \in X$. Bases that have the property of convergence of (1.1.10) for each $f \in X$ are exactly the *quasi-greedy bases* (see Section 1.4).

Let us summarize our discussion of bases in Banach spaces. Schauder bases are natural for convergence of $S_m(f, \Psi)$ and convenient for linear approximation theory. Other classical bases, namely, unconditional bases, are natural for convergence of all rearrangements of expansions. The needs of nonlinear approximation, or, more specifically, the needs of greedy approximation lead us to new concepts of bases: greedy bases and quasi-greedy bases. The relations between these bases are the following:

$$\begin{aligned}&\{\text{greedy bases}\} \subset \{\text{unconditional bases}\} \subset \\ &\{\text{quasi-greedy bases}\} \subset \{\text{Schauder bases}\}.\end{aligned}$$

All the inclusions $\subset$ are proper inclusions. In this chapter we provide a justification of the importance of the new classes of bases. With a belief in the importance of greedy bases and quasi-greedy bases, we discuss here the following natural questions. Could we weaken a rule of building $G_m(f, \Psi)$ and still have good approximation and convergence properties? We answer this question in Sections 1.5 and 1.6. What can be said about classical systems, say, the Haar system and the trigonometric system, in this regard? We discuss this question in Sections 1.3 and 1.7. How to build the approximation theory (mostly, direct and inverse theorems) for m-term approximation with regard to greedy-type bases? Section 1.8 is devoted to this question.

1.2. Schauder bases in Banach spaces

Schauder bases in Banach spaces are used to associate a sequence of numbers with an element $f \in X$: these are the coefficients of f with respect to a basis. This helps in studying properties of a Banach space X. We begin with some classical results on Schauder bases: see, for instance, Lindenstrauss and Tzafriri (1977).

Definition 1.2.1. A sequence $\Psi := \{\psi_n\}_{n=1}^{\infty}$ in a Banach space X is called a Schauder basis of X (basis of X) if, for any $f \in X$, there exists a unique sequence $\{a_n\}_{n=1}^{\infty} := \{a_n(f)\}_{n=1}^{\infty}$ such that

$$f = \sum_{n=1}^{\infty} a_n \psi_n.$$

Let

$$S_0(f) := 0, \quad S_m(f) := S_m(f, \Psi) := \sum_{n=1}^{m} a_n(f)\psi_n.$$

For a fixed basis Ψ, consider the following quantity:

$$|||f||| := \sup_m \|S_m(f, \Psi)\|.$$

It is clear that, for any $f \in X$ we have

$$\|f\| \leq |||f||| < \infty. \tag{1.2.1}$$

It is easy to see that $|||\cdot|||$ provides a norm on the linear space X. Denote this new normed linear space by X^s.

Proposition 1.2.2. The space X^s is a Banach space.

Theorem 1.2.3. Let X be a Banach space with a Schauder basis Ψ. Then the operators $S_m : X \to X$ are bounded linear operators and

$$\sup_m \|S_m\| < \infty.$$

The proof of this theorem is based on the following fundamental theorem of Banach.

Theorem 1.2.4. Let U, V be Banach spaces and T be a bounded linear one-to-one operator from V to U. Then the inverse operator T^{-1} is a bounded linear operator from U to V.

We specify $U = X$, $V = X^s$, and let T be the identity map. It follows from (1.2.1) that T is a bounded operator from V to U. Thus, by Theorem 1.2.4, T^{-1} is also bounded. This means that there exists a constant C such that, for any $f \in X$, we have $|||f||| \leq C\|f\|$. This completes the proof of Theorem 1.2.3.

The operators $\{S_m\}_{m=1}^{\infty}$ are called the natural projections associated with a basis Ψ. The number $\sup_m \|S_m\|$ is called the basis constant of the basis Ψ. A basis whose basis constant is one is called a *monotone basis.* It is clear that an orthonormal basis in a Hilbert space is a monotone basis. Every Schauder basis Ψ is monotone with respect to the norm $|||f||| := \sup_m \|S_m(f, \Psi)\|$, which was already used above. Indeed, we have

$$|||S_m(f)||| = \sup_n \|S_n(S_m(f))\| = \sup_{1 \leq n \leq m} \|S_n(f)\| \leq |||f|||.$$

The above remark means that for any Schauder basis Ψ of X we can renorm X (take X^s) to make the basis Ψ monotone for a new norm.

Theorem 1.2.5. Let $\{x_n\}_{n=1}^\infty$ be a sequence of elements in a Banach space X. Then $\{x_n\}_{n=1}^\infty$ is a Schauder basis of X if and only if the following three conditions hold.

(a) $x_n \neq 0$ for all n.

(b) There is a constant K such that, for every choice of scalars $\{a_i\}_{i=1}^\infty$ and integers $n < m$, we have

$$\left\|\sum_{i=1}^n a_i x_i\right\| \leq K \left\|\sum_{i=1}^m a_i x_i\right\|.$$

(c) The closed linear span of $\{x_n\}_{n=1}^\infty$ coincides with X.

We note that for a basis Ψ with the basis constant K, we have for any $f \in X$

$$\|f - S_m(f,\Psi)\| \leq (K+1) \inf_{\{c_k\}} \left\| f - \sum_{k=1}^m c_k \psi_k \right\|.$$

Thus, the partial sums $S_m(f,\Psi)$ provide near-best approximation from $\operatorname{span}\{\psi_1, \dots, \psi_m\}$.

Let a Banach space X, with a basis $\Psi = \{\psi_k\}_{k=1}^\infty$, be given. In order to understand the efficiency of an algorithm providing an m-term approximation we compare its accuracy with the best-possible accuracy when an approximant is a linear combination of m terms from Ψ. We define the best m-term approximation with regard to Ψ as follows:

$$\sigma_m(f) := \sigma_m(f,\Psi)_X := \inf_{c_k,\Lambda} \left\| f - \sum_{k\in\Lambda} c_k \psi_k \right\|_X,$$

where the infimum is taken over coefficients c_k and sets of indices Λ with cardinality $|\Lambda| = m$. We note that in the above definition of $\sigma_m(f,\Psi)_X$ the system Ψ may be any system of elements from X, not necessarily a basis of X.

An immediate natural question is when the best m-term approximant exists. This question is a more difficult problem than the corresponding problem in the case of linear approximation. The problem of existence of best m-term approximant with regard to a basis has not been studied thoroughly. We present here some results in this direction.

Let us proceed to the approximation problem setting. Let a subset $A \subset X$ be given. For any $f \in X$, let

$$d(f,A) := d(f,A)_X := \inf_{a\in A} \|f - a\|$$

denote the distance from f to A, or in other words the best approximation error of f by elements from A in the norm of X. To illustrate some relevant techniques in this direction, let us prove existence theorems in the following two settings.

S1 Let $X = L_p(0, 2\pi)$, $1 \le p < \infty$, or $X = L_\infty(0, 2\pi) := \mathcal{C}(0, 2\pi)$ be the set of 2π-periodic functions. Consider A to be the set Σ_m of all complex trigonometric polynomials or $\Sigma_m(R)$ of all real trigonometric polynomials which have at most m non-zero coefficients:

$$\Sigma_m := \left\{ t : t = \sum_{k\in\Lambda} c_k \mathrm{e}^{\mathrm{i}kx}, \quad \#\Lambda \le m \right\},$$

$$\Sigma_m(R) := \left\{ t : t = \sum_{k\in\Lambda_1} a_k \cos kx + \sum_{k\in\Lambda_2} b_k \sin kx, \quad \#\Lambda_1 + \#\Lambda_2 \le m \right\}.$$

We will also use the following notation in this case:

$$\sigma_m(f, \mathcal{T})_X := d(f, \Sigma_m)_X.$$

S2 Let $X = L_p(0, 1)$, $1 \le p < \infty$ and let A be the set Σ_m^S of piecewise constant functions with at most $m - 1$ break-points at $(0, 1)$.

In the setting S2 we prove here the following existence theorem (see DeVore and Lorenz (1993), p. 363).

Theorem 1.2.6. For any $f \in L_p(0, 1)$, $1 \le p < \infty$, there exists $g \in \Sigma_m^S$ such that

$$d(f, \Sigma_m^S)_p = \|f - g\|_p.$$

Proof. Fix the break-points $0 = y_0 \le y_1 \le \cdots \le y_{m-1} \le y_m = 1$, let $y := (y_0, \dots, y_m)$, and let $S_0(y)$ be the set of piecewise constant functions with break-points $y_1, \dots, y_{m-1}$. Further, let

$$\mathrm{e}_m^y(f)_p := \inf_{a\in S_0(y)} \|f - a\|_p.$$

From the definition of $d(f, \Sigma_m^S)_p$, there exists a sequence y^i such that

$$\mathrm{e}_m^{y^i}(f)_p \to d(f, \Sigma_m^S)_p$$

when $i \to \infty$. Considering a subsequence of $\{y^i\}$, if necessary we can assume that $y^i \to y^*$ for some $y^* \in \mathbb{R}^{m+1}$. Now we consider only those indices j for which $y^*_{j-1} \ne y^*_j$. Let Λ denote the corresponding set of indices. Take a positive number ϵ satisfying

$$\epsilon < \min_{j\in\Lambda} (y^*_j - y^*_{j-1})/3,$$

and consider i such that

$$\|y^* - y^i\|_\infty < \epsilon, \quad \text{where} \quad \|y\|_\infty := \max_k |y_k|. \tag{1.2.2}$$

By the existence theorem in the case of approximation by elements of a subspace of finite dimension, for each y^i there exists

$$g(f, y^i, c^i) := \sum_{j=1}^{m} c^i_j \chi_{[y^i_{j-1}, y^i_j]},$$

where χ_E denotes the characteristic function of a set E, with the property

$$\|f - g(f, y^i, c^i)\|_p = \mathrm{e}^{y^i}_m(f)_p.$$

For i satisfying (1.2.2) and $j \in \Lambda$ we have $|c^i_j| \le C(f, \epsilon)$, which allows us to assume (passing to a subsequence if necessary) the convergence

$$\lim_{i\to\infty} c^i_j = c_j, \quad j \in \Lambda.$$

Consider

$$g(f, c) := \sum_{j\in\Lambda} c_j \chi_{[y^*_{j-1}, y^*_j]}.$$

Let $U_\epsilon(y) := \cup_j (y_j - \epsilon, y_j + \epsilon)$ and introduce $G := [0, 1] \setminus U_\epsilon(y^*)$. Then we have

$$\int_G |f - g(f, c)|^p = \lim_{i\to\infty} \int_G |f - g(f, y^i, c^i)|^p \le d(f, \Sigma^S_m)^p_p.$$

Letting $\epsilon \to 0$, we complete the proof. □

We proceed now to the trigonometric case S1. We will give the proof in the general d-variable case for $\mathcal{T}^d := \mathcal{T} \times \cdots \times \mathcal{T}$ (d times) because this generality does not introduce any complication. The following theorem was essentially proved in Baishanski (1983). The presented proof is taken from Temlyakov (1998*c*).

Theorem 1.2.7. Let $1 \le p \le \infty$. For any $f \in L_p(\mathbb{T}^d)$ and any $m \in \mathbb{N}$, there exists a trigonometric polynomial t_m of the form

$$t_m(x) = \sum_{n=1}^{m} c_n \mathrm{e}^{\mathrm{i}(k^n, x)}, \tag{1.2.3}$$

such that

$$\sigma_m(f, \mathcal{T}^d)_p = \|f - t_m\|_p. \tag{1.2.4}$$

Proof. We prove this theorem by induction. Let us use the abbreviated notation $\sigma_m(f)_p := \sigma_m(f, \mathcal{T}^d)_p$.

First step. Let $m = 1$. We assume $\sigma_1(f)_p < \|f\|_p$, because in the case $\sigma_1(f)_p = \|f\|_p$ the proof is trivial: we take $t_1 = 0$. We now prove that polynomials of the form $c\,\mathrm{e}^{\mathrm{i}(k,x)}$ with big $|k|$ cannot provide approximation with error close to $\sigma_1(f)_p$. This will allow us to restrict the search for an optimal approximant $c_1 \mathrm{e}^{\mathrm{i}(k^1,x)}$ to a finite number of k^1, which in turn will imply the existence.

We introduce a parameter $N \in \mathbb{N}$, which will be specified later on, and consider the following polynomials:

$$\mathcal{K}_N(u) := \sum_{|k|<N} \left(1 - \frac{|k|}{N}\right) \mathrm{e}^{\mathrm{i}ku}, \quad u \in \mathbb{T}, \tag{1.2.5}$$

and

$$\mathcal{K}_N(x) := \prod_{j=1}^{d} \mathcal{K}_N(x_j), \quad x = (x_1, \dots, x_d) \in \mathbb{T}^d.$$

The functions $\mathcal{K}_N$ are the Fejér kernels. These polynomials have the following property (for (1.2.6) see Zygmund (1959, Chapter 3, Section 3)):

$$\|\mathcal{K}_N\|_1 = 1, \quad N = 1, 2, \dots. \tag{1.2.6}$$

Consider the operator

$$(K_N(g))(x) = (2\pi)^{-d} \int_{\mathbb{T}^d} \mathcal{K}_N(x - y) g(y) \,\mathrm{d}y. \tag{1.2.7}$$

Let

$$e_N(g) := \|g - K_N(g)\|_p. \tag{1.2.8}$$

It is known that for any $f \in L_p(\mathbb{T}^d)$ we have $e_N \to 0$ as $N \to \infty$. For fixed N take any $k \in \mathbb{Z}^d$ such that $\|k\|_\infty \geq N$. Consider $g(x) = f(x) - c\,\mathrm{e}^{\mathrm{i}(k,x)}$ with some c. Using (1.2.5) and (1.2.6), we get on the one hand

$$\|K_N(f)\|_p = \|K_N(g)\|_p \leq \|g\|_p. \tag{1.2.9}$$

On the other hand, we have

$$\|K_N(f)\|_p \geq \|f\|_p - \|f - K_N(f)\|_p \geq \|f\|_p - e_N(f). \tag{1.2.10}$$

Therefore, combining (1.2.9) and (1.2.10) we obtain, for all k, $\|k\|_\infty \geq N$, and any c,

$$\|f(x) - c\,\mathrm{e}^{\mathrm{i}(k,x)}\|_p \geq \|f\|_p - e_N(f). \tag{1.2.11}$$

Making N big enough, we get

$$\|f\|_p - e_N(f) \geq (\|f\|_p + \sigma_1(f)_p)/2. \tag{1.2.12}$$

Relations (1.2.11) and (1.2.12) imply

$$\sigma_1(f)_p = \inf_{c, \|k\|_\infty < N} \|f(x) - c\,\mathrm{e}^{\mathrm{i}(k,x)}\|_p,$$

which completes the proof for $m = 1$, by the existence theorem in the case of approximation by elements of a subspace of finite dimension.

General step. Assume that Theorem 1.2.7 has already been proved for $m-1$. We prove it for m. If $\sigma_m(f)_p = \sigma_{m-1}(f)_p$, we are done, by the induction assumption. Let $\sigma_m(f)_p < \sigma_{m-1}(f)_p$. The idea of the proof in the general step is similar to that in the first step.

Take any $k^1, \dots, k^m$. Assume $\|k^j\|_\infty \le \|k^m\|_\infty$, $j = 1, \dots, m-1$, and $\|k^m\|_\infty > N$. We prove that a polynomial with frequencies $k^1, \dots, k^m$ does not provide good approximation. Take any numbers $c_1, \dots, c_m$, and consider

$$f_{m-1}(x) := f(x) - \sum_{j=1}^{m-1} c_j \mathrm{e}^{\mathrm{i}(k^j,x)},$$

$$g(x) := f_{m-1}(x) - c_m \mathrm{e}^{\mathrm{i}(k^m,x)}.$$

Then, replacing f by f_{m-1}, we get in the same way as above the estimate

$$\left\| f(x) - \sum_{j=1}^{m} c_j \mathrm{e}^{\mathrm{i}(k^j,x)} \right\|_p \ge \sigma_{m-1}(f)_p - e_N(f). \tag{1.2.13}$$

We remark here that the analogue to (1.2.10) looks as follows:

$$\begin{aligned} \|K_N(f_{m-1})\|_p &\ge \sigma_{m-1}(K_N(f))_p \\ &\ge \sigma_{m-1}(f)_p - \|f - K_N(f)\|_p \\ &\ge \sigma_{m-1}(f)_p - e_N(f). \end{aligned}$$

Making N big enough, we derive from (1.2.13) that

$$\sigma_m(f)_p = \inf\left(\inf_{c_j, j=1,\dots,m} \left\| f(x) - \sum_{j=1}^{m} c_j \mathrm{e}^{\mathrm{i}(k^j,x)} \right\|_p \right),$$

where the infimum is taken over k^j satisfying the restriction $\|k^j\|_\infty \le N$ for all $j = 1, \dots, m$. In order to complete the proof of Theorem 1.2.7, it remains to remark that, by the existence theorem in the case of approximation by elements of a subspace of finite dimension, the inside infimum can always be replaced by minimum, and the outside infimum is taken over a finite set. This completes the proof. □

Concerning the problem of uniqueness of the best approximant, we will only make a remark that shows that in the m-term nonlinear approximation we can hardly expect the unicity. Let us consider problem S1 on best m-term trigonometric approximation in a particular case $X = L_2(0, 2\pi)$. Take

$$f(x) = \sum_{k=1}^{n} \mathrm{e}^{\mathrm{i}kx}.$$

Clearly, $\sigma_1(f)_2 = (n-1)^{1/2}$ and each $\mathrm{e}^{\mathrm{i}kx}$, $k = 1, \dots, n$ may serve as a best approximant.

We can prove the following existence theorem (see Temlyakov (2001*a*)) in a similar way to the proof of Theorem 1.2.7.

Theorem 1.2.8. Let Ψ be a monotone basis of X. Then, for any $x \in X$ and any $m \in \mathbb{N}$, there exist Λ_m, $|\Lambda_m| \le m$, and $\{c_i^* : i \in \Lambda_m\}$ such that

$$\left\| f - \sum_{i\in\Lambda_m} c_i^* \psi_i \right\| = \sigma_m(f, \Psi).$$

Here is one more existence theorem from Temlyakov (2001*a*).

Theorem 1.2.9. Let Ψ be a normalized ($\|\psi_k\| = 1$, $k = 1, \dots$) Schauder basis of X with the additional property that ψ_k converges weakly to 0. Then, for any $f \in X$, and any $m \in \mathbb{N}$, there exist Λ_m, $|\Lambda_m| \le m$, and $\{c_i^* : i \in \Lambda_m\}$ such that

$$\left\| f - \sum_{i\in\Lambda_m} c_i^* \psi_i \right\| = \sigma_m(f, \Psi).$$

Proof. The proof is a development of ideas from Baishanski (1983). In order to sketch the idea of the proof, let us consider first the case $m = 1$. Let

$$\|f - c_{k_n}\psi_{k_n}\| \to \sigma_1(f, \Psi), \quad n \to \infty. \tag{1.2.14}$$

If

$$\liminf_{n\to\infty} k_n < \infty,$$

then there exists k and a sequence $\{a_n\}$ such that

$$\|f - a_n\psi_k\| \to \sigma_1(f, \Psi), \quad n \to \infty. \tag{1.2.15}$$

Using the fact that Ψ is a Schauder basis, we infer from (1.2.15) that the sequence $\{a_n\}$ is bounded. Choosing a convergent subsequence of $\{a_n\}$, we construct an a such that

$$\|f - a\psi_k\| = \sigma_1(f, \Psi),$$

which proves existence in this case. Assume now that

$$\lim_{n\to\infty} k_n = \infty.$$

Let F_f be a norming (peak) functional for f: $F_f(f) = \|f\|$, $\|F_f\| = 1$. Then

$$\|f - c_{k_n}\psi_{k_n}\| \ge F_f(f - c_{k_n}\psi_{k_n}) = \|f\| - c_{k_n}F_f(\psi_{k_n}). \tag{1.2.16}$$

Relation (1.2.14) implies boundedness of $\{c_{k_n}\}$, and therefore, by weak convergence to 0 of $\{\psi_k\}$, we get from (1.2.16) and (1.2.14) that

$$\sigma_1(f, \Psi) = \|f\|.$$

Thus we can take 0 as a best approximant. Let us now consider the general case of m-term approximation. Let

$$f^n := \sum_{j=1}^{m} c^n_{k^n_j} \psi_{k^n_j}, \quad k^n_1 < k^n_2 < \cdots < k^n_m,$$

be such that

$$\|f - f^n\| \to \sigma_m(f, \Psi).$$

Then we have

$$|c^n_{k^n_j}| \le M \tag{1.2.17}$$

for all n, j with some constant M. Assume that we have

$$\begin{aligned} \liminf_{n\to\infty} k^n_j &< \infty, \quad \text{for some (possibly none) } j = 1, \dots, l \le m, \\ \lim_{n\to\infty} k^n_j &= \infty, \quad \text{for some (possibly none) } j = l+1, \dots, m. \end{aligned}$$

Then, as in the case of $m = 1$ we find Λ, $|\Lambda| \le l$, and a subsequence $\{n_s\}_{s=1}^{\infty}$ such that

$$\sum_{k\in\Lambda} c^{n_s}_k \psi_k \to \sum_{k\in\Lambda} c_k \psi_k =: y. \tag{1.2.18}$$

Consider the norming functional F_{f-y}. We have from (1.2.17), (1.2.18) and weak convergence of $\{\psi_k\}$ to 0 that

$$F_{f-y}(f^{n_s} - y) \to 0, \quad \text{as } s \to \infty.$$

Thus

$$\begin{aligned} \|f - y\| = F_{f-y}(f - y) &= F_{f-y}(f - f^{n_s} + f^{n_s} - y) \\ &\le \|f - f^{n_s}\| + |F_{f-y}(f^{n_s} - y)| \to \sigma_m(f, \Psi), \end{aligned}$$

as $s \to \infty$. This implies that

$$\|f - y\| = \sigma_m(f, \Psi),$$

which completes the proof of Theorem 1.2.9. □

The following observation is from Wojtaszczyk (2002*b*).

Remark 1.2.10. It is clear from the proof of Theorem 1.2.9 that the condition of weak convergence of ψ_k to 0 can be replaced by the condition $y(\psi_k) \to 0$ for every $y \in Y$. Here, $Y \subset X^*$ is such that, for all $f \in X$,

$$\|f\| = \sup_{y\in Y, \|y\|\le 1} |y(f)|.$$

Also, Wojtaszczyk (2002*b*) contains an example of an unconditional basis Ψ and an element f such that the best m-term approximation of f with regard to Ψ does not exist.

1.3. Greedy bases

Let a Banach space X, with a basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$, be given. We assume that $\|\psi_k\| \geq C > 0$, $k = 1, 2, \ldots$, and consider the following theoretical greedy algorithm. For a given element $f \in X$ we consider the expansion

$$f = \sum_{k=1}^{\infty} c_k(f, \Psi)\psi_k. \tag{1.3.1}$$

For an element $f \in X$ we say that a permutation ρ of the positive integers is decreasing if

$$|c_{k_1}(f, \Psi)| \geq |c_{k_2}(f, \Psi)| \geq \cdots, \tag{1.3.2}$$

where $\rho(j) = k_j$, for $j = 1, 2, \ldots$, and write $\rho \in D(f)$. If the inequalities are strict in (1.3.2), then $D(f)$ consists of only one permutation. We define the mth greedy approximant of f, with respect to the basis Ψ corresponding to a permutation $\rho \in D(f)$, by the formula

$$G_m(f) := G_m(f, \Psi) := G_m(f, \Psi, \rho) := \sum_{j=1}^{m} c_{k_j}(f, \Psi)\psi_{k_j}.$$

We note that there is another natural greedy-type algorithm based on ordering $\|c_k(f, \Psi)\psi_k\|$ instead of ordering absolute values of coefficients. In this case we do not need the restriction $\|\psi_k\| \geq C > 0$, $k = 1, 2, \ldots$. Let $\Lambda_m(f)$ be a set of indices such that

$$\min_{k \in \Lambda_m(f)} \|c_k(f, \Psi)\psi_k\| \geq \max_{k \notin \Lambda_m(f)} \|c_k(f, \Psi)\psi_k\|.$$

We define $G_m^X(f, \Psi)$ by the formula

$$G_m^X(f, \Psi) := S_{\Lambda_m(f)}(f, \Psi), \quad \text{where } S_E(f) := S_E(f, \Psi) := \sum_{k \in E} c_k(f, \Psi)\psi_k.$$

It is clear that for a normalized basis ($\|\psi_k\| = 1$, $k = 1, 2, \ldots$) the above two greedy algorithms coincide. It is also clear that the above greedy algorithm $G_m^X(\cdot, \Psi)$ can be considered as a greedy algorithm $G_m(\cdot, \Psi')$, with $\Psi' := \{\psi_k/\|\psi_k\|\}_{k=1}^{\infty}$ being a normalized version of the Ψ. Thus, we will concentrate on studying the algorithm $G_m(\cdot, \Psi)$. In the above definition of $G_m(\cdot, \Psi)$ we impose an extra condition on a basis Ψ: $\inf_k \|\psi_k\| > 0$. This restriction allows us to define $G_m(f, \Psi)$ for all $f \in X$. For the sake of completeness we will also discuss the case

$$\inf_k \|\psi_k\| = 0. \tag{1.3.3}$$

In this case we define the $G_m(f, \Psi)$ in the same way as above, but only for f of a special form:

$$f = \sum_{k \in Y} c_k(f, \Psi)\psi_k, \quad |Y| < \infty. \tag{1.3.4}$$

The above algorithm $G_m(\cdot, \Psi)$ is a simple algorithm which describes the theoretical scheme for m-term approximation of an element f. We call this algorithm the Greedy Algorithm (GA). In order to understand the efficiency of this algorithm we compare its accuracy with the best-possible accuracy when an approximant is a linear combination of m terms from Ψ. We define the best m-term approximation error with respect to Ψ as follows:

$$\sigma_m(f) := \sigma_m(f, \Psi)_X := \inf_{c_k, \Lambda} \left\| f - \sum_{k \in \Lambda} c_k \psi_k \right\|_X,$$

where the infimum is taken over coefficients c_k and sets of indices Λ with cardinality $|\Lambda| = m$. The best we can achieve with the algorithm G_m is

$$\|f - G_m(f, \Psi, \rho)\|_X = \sigma_m(f, \Psi)_X,$$

or the slightly weaker

$$\|f - G_m(f, \Psi, \rho)\|_X \le G\sigma_m(f, \Psi)_X, \tag{1.3.5}$$

for all elements $f \in X$, and with a constant $G = C(X, \Psi)$ independent of f and m. It was mentioned in the Introduction (see (1.1.8)) that, when $X = H$ is a Hilbert space and $\mathcal{B}$ is an orthonormal basis, we have

$$\|f - G_m(f, \mathcal{B}, \rho)\|_H = \sigma_m(f, \mathcal{B})_H.$$

Let us begin our discussion with an important class of bases: wavelet-type bases. For $X = L_p$, we will write p instead of L_p. Let $\mathcal{H} := \{H_k\}_{k=1}^\infty$ denote the Haar basis on $[0, 1)$ normalized in $L_2(0, 1)$. We denote by $\mathcal{H}_p := \{H_{k,p}\}_{k=1}^\infty$ the Haar basis $\mathcal{H}$ renormalized in $L_p(0, 1)$, which is defined as follows: $H_{1,p} = 1$ on $[0, 1)$ and, for $k = 2^n + l$, $l = 1, 2, \dots, 2^n$, $n = 0, 1, \dots,$

$$H_{k,p} = \begin{cases} 2^{n/p}, & x \in [(2l-2)2^{-n-1}, (2l-1)2^{-n-1}), \\ -2^{n/p}, & x \in [(2l-1)2^{-n-1}, 2l2^{-n-1}), \\ 0, & \text{otherwise.} \end{cases}$$

We will use the following definition of the L_p-equivalence of bases. We say that $\Psi = \{\psi_k\}_{k=1}^\infty$ is L_p-equivalent to $\Phi = \{\phi_k\}_{k=1}^\infty$ if for any finite set Λ and any coefficients c_k, $k \in \Lambda$, we have

$$C_1(p, \Psi, \Phi) \left\| \sum_{k \in \Lambda} c_k \phi_k \right\|_p \le \left\| \sum_{k \in \Lambda} c_k \psi_k \right\|_p \le C_2(p, \Psi, \Phi) \left\| \sum_{k \in \Lambda} c_k \phi_k \right\|_p$$

with two positive constants $C_1(p,\Psi,\Phi), C_2(p,\Psi,\Phi)$ which may depend on p, Ψ, and Φ. For sufficient conditions on Ψ to be L_p-equivalent to $\mathcal{H}$, see Frazier and Jawerth (1990) and DeVore, Konyagin and Temlyakov (1998). In particular, it is known that all reasonable univariate wavelet-type bases are L_p-equivalent to $\mathcal{H}$ for $1 < p < \infty$. We proved the following theorem in Temlyakov (1998*a*).

Theorem 1.3.1. Let $1 < p < \infty$ and let a basis Ψ be L_p-equivalent to the Haar basis $\mathcal{H}$. Then, for any $f \in L_p(0,1)$, we have

$$\|f - G_m^p(f,\Psi)\|_p \leq C(p,\Psi)\sigma_m(f,\Psi)_p$$

with a constant $C(p,\Psi)$ independent of f and m.

By a simple renormalization argument we obtain the following version of Theorem 1.3.1.

Theorem 1.3.1A. Let $1 < p < \infty$ and let a basis Ψ be L_p-equivalent to the Haar basis $\mathcal{H}_p$. Then, for any $f \in L_p(0,1)$ and any $\rho \in D(f)$, we have

$$\|f - G_m(f,\Psi,\rho)\|_p \leq C(p,\Psi)\sigma_m(f,\Psi)_p$$

with a constant $C(p,\Psi)$ independent of f, ρ, and m.

We note that Temlyakov (1998*a*) also contains a generalization of Theorem 1.3.1 to the multivariate Haar basis obtained by the multi-resolution analysis procedure. These theorems motivated us to consider the general setting of greedy approximation in Banach spaces. We concentrated on studying bases which satisfy (1.3.5) for all individual functions. Definitions 1.3.2–1.3.4, below, are from Konyagin and Temlyakov (1999*a*).

Definition 1.3.2. We call a basis Ψ a greedy basis if, for every $f \in X$ (in the case $\inf_k \|\psi_k\| > 0$) and for f of the form (1.3.4) (in the case $\inf_k \|\psi_k\| = 0$), there exists a permutation $\rho \in D(f)$ such that the inequality

$$\|f - G_m(f,\Psi,\rho)\|_X \leq G\sigma_m(f,\Psi)_X \tag{1.3.6}$$

holds with a constant independent of f, m.

Theorem 1.3.1A shows that each basis Ψ which is L_p-equivalent to the univariate Haar basis $\mathcal{H}_p$ is a greedy basis for $L_p(0,1)$, $1 < p < \infty$. We note that for a Hilbert space each orthonormal basis is a greedy basis with a constant $G = 1$ (see (1.3.6)).

We now give the definitions of unconditional and democratic bases.

Definition 1.3.3. A basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$ of a Banach space X is said to be unconditional if, for every choice of signs $\theta = \{\theta_k\}_{k=1}^{\infty}$, $\theta_k = 1$ or -1,

$k = 1, 2, \dots$, the linear operator M_θ defined by

$$M_\theta\left(\sum_{k=1}^{\infty} a_k\psi_k\right) = \sum_{k=1}^{\infty} a_k\theta_k\psi_k$$

is a bounded operator from X into X.

Definition 1.3.4. We say that a basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$ is a democratic basis for X if there exists a constant $D := D(X, \Psi)$ such that, for any two finite sets of indices P and Q with the same cardinality $|P| = |Q|$, we have

$$\left\|\sum_{k\in P}\psi_k\right\| \le D\left\|\sum_{k\in Q}\psi_k\right\|.$$

We proved in Konyagin and Temlyakov (1999*a*) the following theorem.

Theorem 1.3.5. A basis is greedy if and only if it is unconditional and democratic.

This theorem gives a characterization of greedy bases. Further investigations (Temlyakov 1998*b*, Cohen, DeVore and Hochmuth 2000, Kerkyacharian and Picard 2004, Gribonval and Nielsen 2001*b*, Kamont and Temlyakov 2004) showed that the concept of greedy bases is very useful in direct and inverse theorems of nonlinear approximation and also in applications in statistics.

Let us make a remark on bases Ψ that satisfy condition (1.3.3). In this case the greedy algorithm $G_m(\cdot, \Psi)$ is defined only for f of the form (1.3.4). However, if Ψ is a greedy basis, then by Theorem 1.3.5 it is democratic, and therefore satisfies the condition $\inf_k \|\psi_k\| > 0$. Thus, there are no greedy bases satisfying (1.3.3).

An interesting generalization of m-term approximation was considered in Cohen *et al.* (2000). Let $\Psi = \{\psi_I\}_I$ be a basis indexed by dyadic intervals. Take an α and assign to each index set Λ the following measure:

$$\Phi_\alpha(\Lambda) := \sum_{I\in\Lambda} |I|^\alpha.$$

In the case $\alpha = 0$ we get $\Phi_0(\Lambda) = |\Lambda|$. An analogue of best m-term approximation is as follows:

$$\inf_{\Lambda:\Phi_\alpha(\Lambda)\le m}\ \inf_{c_I, I\in\Lambda}\left\|f - \sum_{I\in\Lambda} c_I\psi_I\right\|_p.$$

A detailed study of this type of approximation (restricted approximation) can be found in Cohen *et al.* (2000).

We now elaborate on the idea of assigning to each basis element ψ_k a non-negative weight w_k. We discuss weight-greedy bases and prove a criterion for weight-greedy bases similar to that for greedy bases.

Let Ψ be a basis for X. As above, if $\inf_n \|\psi_n\| > 0$ then $c_n(f) \to 0$ as $n \to \infty$, where

$$f = \sum_{n=1}^{\infty} c_n(f)\psi_n.$$

Then we can rearrange the coefficients $\{c_n(f)\}$ in the decreasing order

$$|c_{n_1}(f)| \geq |c_{n_2}(f)| \geq \cdots,$$

and define the mth greedy approximant as

$$G_m(f, \Psi) := \sum_{k=1}^{m} c_{n_k}(f)\psi_{n_k}. \tag{1.3.7}$$

In the case $\inf_n \|\psi_n\| = 0$ we define $G_m(f, \Psi)$ by (1.3.7) for f of the form

$$f = \sum_{n \in Y} c_n(f)\psi_n, \quad |Y| < \infty. \tag{1.3.8}$$

Let a weight sequence $w = \{w_n\}_{n=1}^{\infty}$, $w_n > 0$, be given. For $\Lambda \subset \mathbb{N}$, denote $w(\Lambda) := \sum_{n \in \Lambda} w_n$. For a positive real number $v > 0$ define

$$\sigma_v^w(f, \Psi) := \inf_{\{b_n\}, \Lambda : w(\Lambda) \leq v} \left\| f - \sum_{n \in \Lambda} b_n \psi_n \right\|,$$

where Λ are finite.

We present results from Kerkyacharian, Picard and Temlyakov (2006).

Definition 1.3.6. We call a basis Ψ a weight-greedy basis (w-greedy basis) if for any $f \in X$ in the case $\inf_n \|\psi_n\| > 0$ or for any $f \in X$ of the form (1.3.8) in the case $\inf_n \|\psi_n\| = 0$, we have

$$\|f - G_m(f, \Psi)\| \leq C_G \sigma^w_{w(\Lambda_m)}(f, \Psi),$$

where Λ_m is obtained from the representation

$$G_m(f, \Psi) = \sum_{n \in \Lambda_m} c_n(f)\psi_n, \quad |\Lambda_m| = m.$$

Definition 1.3.7. We call a basis Ψ a weight-democratic basis (w-democratic basis) if, for any finite $A, B \subset \mathbb{N}$ such that $w(A) \leq w(B)$, we have

$$\left\| \sum_{n \in A} \psi_n \right\| \leq C_D \left\| \sum_{n \in B} \psi_n \right\|.$$

Theorem 1.3.8. A basis Ψ is a w-greedy basis if and only if it is unconditional and w-democratic.

Proof. **I** We first prove the implication

$$\text{unconditional} + w\text{-democratic} \quad \Rightarrow \quad w\text{-greedy}.$$

Let f be any function, or a function of the form (1.3.8) if $\inf_n \|\psi_n\| = 0$. Consider

$$G_m(f,\Psi) = \sum_{n\in Q} c_n(f)\psi_n =: S_Q(f).$$

We take any finite set $P \subset \mathbb{N}$ satisfying $w(P) \le w(Q)$. Then our assumption $w_n > 0$, $n \in \mathbb{N}$ implies that either $P = Q$ or $Q \setminus P$ is non-empty. As in the Introduction, let

$$E_P(f,\Psi) := \inf_{\{b_n\}} \left\| f - \sum_{n\in P} b_n\psi_n \right\|.$$

Then, by unconditionality of Ψ, we have (see (1.1.12))

$$\|f - S_P(f)\| \le (K+1)E_P(f,\Psi). \tag{1.3.9}$$

This (with $P = Q$) completes the proof in the case $\sigma^w_{w(Q)}(f,\Psi) = E_Q(f,\Psi)$. Suppose that $\sigma^w_{w(Q)}(f,\Psi) < E_Q(f,\Psi)$. Clearly, we may now consider only those P that satisfy the following two conditions:

$$w(P) \le w(Q) \quad \text{and} \quad E_P(f,\Psi) < E_Q(f,\Psi).$$

For P satisfying the above conditions we have $Q \setminus P \ne \emptyset$. We estimate

$$\|f - S_Q(f)\| \le \|f - S_P(f)\| + \|S_P(f) - S_Q(f)\|. \tag{1.3.10}$$

We have

$$S_P(f) - S_Q(f) = S_{P\setminus Q}(f) - S_{Q\setminus P}(f). \tag{1.3.11}$$

As for (1.3.9) we get

$$\|S_{Q\setminus P}(f)\| \le KE_P(f,\Psi). \tag{1.3.12}$$

It remains to estimate $\|S_{P\setminus Q}(f)\|$. By unconditionality and w-democracy in the case of a real Banach space X, we have

$$\|S_{P\setminus Q}(f)\| \le 2K \max_{n\in P\setminus Q} |c_n(f)| \left\| \sum_{n\in P\setminus Q} \psi_n \right\| \tag{1.3.13}$$

$$\le 2KC_D \min_{n\in Q\setminus P} |c_n(f)| \left\| \sum_{n\in Q\setminus P} \psi_n \right\| \le C(K)C_D \|S_{Q\setminus P}(f)\|.$$

In the case of a complex Banach space X the above inequalities hold with $2K$ replaced by $4K$. Combining (1.3.9)–(1.3.13), we complete the proof of part I. □

II We now prove the implication

$$w\text{-greedy} \quad \Rightarrow \quad \text{unconditional} + w\text{-democratic}.$$

IIa We begin with the following one:

$$w\text{-greedy} \quad \Rightarrow \quad \text{unconditional}.$$

We will prove a slightly stronger statement.

Lemma 1.3.9. Let Ψ be a basis such that, for any f of the form (1.3.8), we have

$$\|f - G_m(f,\Psi)\| \le CE_\Lambda(f,\Psi),$$

where

$$G_m(f,\Psi) = \sum_{n\in\Lambda} c_n(f)\psi_n.$$

Then Ψ is unconditional.

Proof. It is clear that it is sufficient to prove that there exists a constant C_0 such that, for any finite Λ and any f of the form (1.3.8), we have

$$\|S_\Lambda(f)\| \le C_0\|f\|.$$

Let f and Λ be given and $\Lambda \subset [1, M]$. Consider

$$f_M := S_{[1,M]}(f).$$

Then $\|f_M\| \le C_B\|f\|$. We take a $b > \max_{1\le n\le M} |c_n(f)|$ and define a new function

$$g := f_M - S_\Lambda(f_M) + b\sum_{n\in\Lambda}\psi_n.$$

Then

$$G_m(g,\Psi) = b\sum_{n\in\Lambda}\psi_n, \quad m := |\Lambda|,$$

and

$$E_\Lambda(g,\Psi) \le \|f_M\|.$$

Thus,

$$\|f_M - S_\Lambda(f_M)\| = \|g - G_m(g,\Psi)\| \le CE_\Lambda(g,\Psi) \le C\|f_M\|.$$

Therefore,

$$\|S_\Lambda(f)\| = \|S_\Lambda(f_M)\| \le C_0\|f\|. \qquad \square$$

IIb It remains to prove the implication

$$w\text{-greedy} \quad \Rightarrow \quad w\text{-democratic}.$$

First, let $A, B \subset \mathbb{N}$, $w(A) \le w(B)$, be such that $A \cap B = \emptyset$. Consider

$$f := \sum_{n\in A}\psi_n + (1+\epsilon)\sum_{n\in B}\psi_n, \quad \epsilon > 0.$$

Then

$$G_m(f,\Psi) = (1+\epsilon)\sum_{n\in B}\psi_n, \quad m := |B|,$$

and

$$E_A(f,\Psi) \le \Big\|\sum_{n\in B}\psi_n\Big\|(1+\epsilon).$$

Therefore, by the w-greedy assumption we get

$$\Big\|\sum_{n\in A}\psi_n\Big\| \le C(1+\epsilon)\Big\|\sum_{n\in B}\psi_n\Big\|.$$

Now let A, B be any finite subsets of $\mathbb{N}$ for which $w(A) \le w(B)$. Then, using the unconditionality of Ψ proved in IIa and the above part of IIb, we obtain

$$\begin{aligned}\Big\|\sum_{n\in A}\psi_n\Big\| &\le \Big\|\sum_{n\in A\setminus B}\psi_n\Big\| + \Big\|\sum_{n\in A\cap B}\psi_n\Big\| \\ &\le C\Big\|\sum_{n\in B\setminus A}\psi_n\Big\| + K\Big\|\sum_{n\in B}\psi_n\Big\| \le C_1\Big\|\sum_{n\in B}\psi_n\Big\|.\end{aligned}$$

This completes the proof of Theorem 1.3.8. □

Theorems 1.3.5 and 1.3.8 show that *greedy* = *unconditional* + *democratic.* We now show that unconditionality does not imply democracy, and *vice versa.*

Unconditionality does not imply democracy. This follows from properties of the multivariate Haar system $\mathcal{H}^2 = \mathcal{H}\times\mathcal{H}$ defined as the tensor product of the univariate Haar systems $\mathcal{H}$ (see (1.3.14) below).

Democracy does not imply unconditionality. Let X be the set of all real sequences $x = (x_1, x_2, \dots)$ such that

$$\|x\|_X = \sup_{N\in\mathbb{N}}\Big|\sum_{n=1}^{N} x_n\Big|$$

is finite. Clearly, X equipped with the norm $\|\cdot\|_X$ is a Banach space. Let $\psi_k \in X$, $k = 1, 2, \dots$, be defined as $(\psi_k)_n = 1$ if $n = k$ and $(\psi_k)_n = 0$ otherwise. Let X_0 denote the subspace of X generated by the elements ψ_k. It is easy to see that $\{\psi_k\}$ is a democratic basis in X_0. However, it is not an unconditional basis, since

$$\Big\|\sum_{k=1}^{m}\psi_k\Big\|_X = m,$$

but

$$\left\|\sum_{k=1}^{m}(-1)^k\psi_k\right\|_X = 1.$$

We let $\mathcal{H}_p := \{H_{k,p}\}_{k=1}^{\infty}$ be the Haar basis $\mathcal{H}$ renormalized in $L_p([0,1))$. We define the multivariate Haar basis $\mathcal{H}_p^d$ to be the tensor product of the univariate Haar bases: $\mathcal{H}_p^d := \mathcal{H}_p \times \cdots \times \mathcal{H}_p$;

$$H_{\mathbf{n},p}(x) := H_{n_1,p}(x_1)\cdots H_{n_d,p}(x_d), \quad x = (x_1,\dots,x_d), \quad \mathbf{n} = (n_1,\dots,n_d).$$

Supports of functions $H_{\mathbf{n},p}$ are arbitrary dyadic parallelepipeds (intervals). It is known (see Temlyakov (2002*a*)) that the tensor product structure of the multivariate wavelet bases makes them universal for approximation of anisotropic smoothness classes with different anisotropy. It is also known that the study of such bases is more difficult than the study of the univariate bases. In many cases we need to develop new techniques and in some cases we encounter new phenomena. For instance, it turns out that the democratic property does not hold for the multivariate Haar basis $\mathcal{H}_p^d$ for $p \neq 2$. The following relation is known for $1 < p < \infty$:

$$\sup_{f\in L_p} \|f - G_m(f,\mathcal{H}_p^d)\|_p/\sigma_m(f,\mathcal{H}_p^d) \asymp (\log m)^{(d-1)|1/2-1/p|}. \tag{1.3.14}$$

The lower bound in (1.3.14) was proved by R. Hochmuth; the upper bound in (1.3.14) was proved in the case $d = 2$, $4/3 \le p \le 4$, and was conjectured for all d, $1 < p < \infty$, in Temlyakov (1998*b*). The conjecture was proved in Wojtaszczyk (2000).

Let us return to the problem of finding a near-best m-term approximant of $f \in X$ with regard to a basis Ψ. This problem consists of two subproblems. First, we need to identify a set Λ_m of m indices that can be used in achieving near-best m-term approximation of f. Second, we need to find the coefficients $\{c_k\}$, $k \in \Lambda_m$, such that the approximant $\sum_{k\in\Lambda} c_k\psi_k$ provides near-best approximation of f. It is clear from the properties of an unconditional basis Ψ that, for any $f \in X$ and any Λ, we have (see (1.1.12))

$$\left\|f - \sum_{k\in\Lambda} c_k(f,\Psi)\psi_k\right\| \le C \inf_{\{c_k\}}\left\|f - \sum_{k\in\Lambda} c_k\psi_k\right\|.$$

Therefore, in the case of an unconditional basis Ψ the second subproblem is easy: we can always choose the expansion coefficients $c_k(f,\Psi)$, $k \in \Lambda$. Theorem 1.3.5 shows that if a basis Ψ is simultaneously unconditional and democratic then the first subproblem is also easy: it follows from the definition of greedy basis that the algorithm of choosing the m biggest in absolute-value coefficients gives the set Λ_m.

It would be very interesting to understand how we can find Λ_m in the case when we only know that Ψ is unconditional. The following special case of the above problem is of great interest: $X = L_p([0,1]^d)$, $d \geq 2$, Ψ is the multivariate Haar basis $\mathcal{H}_p^d$, $1 < p < \infty$. It is known from Temlyakov (1998*b*), Wojtaszczyk (2000) and Kamont and Temlyakov (2004) that the function

$$\mu(m, \mathcal{H}_p^d) := \sup_{k \leq m} \left(\sup_{\Lambda:|\Lambda|=k} \left\| \sum_{\mathbf{n}\in\Lambda} H_{\mathbf{n},p} \right\|_p \Big/ \inf_{\Lambda:|\Lambda|=k} \left\| \sum_{\mathbf{n}\in\Lambda} H_{\mathbf{n},p} \right\|_p \right)$$

plays a very important role in estimates of the m-term greedy approximation in terms of the best m-term approximation. For instance (see Temlyakov (1998*b*)),

$$\|f - G_m^{L_p}(f, \mathcal{H}_p^d)\|_p \leq C(p,d)\mu(m, \mathcal{H}_p^d)\sigma_m(f, \mathcal{H}_p^d)_p, \quad 1 < p < \infty. \quad (1.3.15)$$

The following theorem gives, in particular, upper bounds for $\mu(m, \mathcal{H}_p^d)$.

Theorem 1.3.10. Let $1 < p < \infty$. Then, for any Λ, $|\Lambda| = m$, we have for $2 \leq p < \infty$

$$C_{p,d}^1 m^{1/p} \min_{\mathbf{n}\in\Lambda} |c_\mathbf{n}| \leq \left\| \sum_{\mathbf{n}\in\Lambda} c_\mathbf{n} H_{\mathbf{n},p} \right\|_p \leq C_{p,d}^2 m^{1/p} (\log m)^{h(p,d)} \max_{\mathbf{n}\in\Lambda} |c_\mathbf{n}|,$$

and for $1 < p \leq 2$

$$C_{p,d}^3 m^{1/p} (\log m)^{-h(p,d)} \min_{\mathbf{n}\in\Lambda} |c_\mathbf{n}| \leq \left\| \sum_{\mathbf{n}\in\Lambda} c_\mathbf{n} H_{\mathbf{n},p} \right\|_p \leq C_{p,d}^4 m^{1/p} \max_{\mathbf{n}\in\Lambda} |c_\mathbf{n}|,$$

where $h(p,d) := (d-1)|1/2 - 1/p|$.

Theorem 1.3.10 for $d = 1$, $1 < p < \infty$ was proved in Temlyakov (1998*a*), and for $d = 2$, $4/3 \leq p \leq 4$ it was proved in Temlyakov (1998*b*). Theorem 1.3.10 in the general case was proved in Wojtaszczyk (2000). It is known (Temlyakov 2002*c*) that the extra log factors in Theorem 1.3.10 are sharp.

Let Ψ be a normalized basis for $L_p([0,1))$. For the space $L_p([0,1)^d)$ we define $\Psi^d := \Psi \times \cdots \times \Psi$ (d times), and

$$\psi_\mathbf{n}(x) := \psi_{n_1}(x_1)\cdots\psi_{n_d}(x_d), \quad \text{for } x = (x_1,\dots,x_d), \quad \mathbf{n} = (n_1,\dots,n_d).$$

In Kerkyacharian *et al.* (2006) we proved the following theorem using a proof whose structure is similar to that from Wojtaszczyk (2000).

Theorem 1.3.11. Let $1 < p < \infty$ and let Ψ be a greedy basis for $L_p([0,1))$. Then, for any Λ, $|\Lambda| = m$, we have for $2 \leq p < \infty$

$$C_{p,d}^5 m^{1/p} \min_{\mathbf{n}\in\Lambda} |c_\mathbf{n}| \leq \left\| \sum_{\mathbf{n}\in\Lambda} c_\mathbf{n} \psi_\mathbf{n} \right\|_p \leq C_{p,d}^6 m^{1/p} (\log m)^{h(p,d)} \max_{\mathbf{n}\in\Lambda} |c_\mathbf{n}|,$$

and for $1 < p \le 2$

$$C^7_{p,d} m^{1/p} (\log m)^{-h(p,d)} \min_{\mathbf{n}\in\Lambda} |c_{\mathbf{n}}| \le \Big\| \sum_{\mathbf{n}\in\Lambda} c_{\mathbf{n}} \psi_{\mathbf{n}} \Big\|_p \le C^8_{p,d} m^{1/p} \max_{\mathbf{n}\in\Lambda} |c_{\mathbf{n}}|,$$

where $h(p,d) := (d-1)|1/2 - 1/p|$.

Inequality (1.3.15) was extended in Wojtaszczyk (2000) to a normalized unconditional basis Ψ for X instead of $\mathcal{H}^d_p$ for $L_p([0,1)^d)$. Therefore, as a corollary of Theorem 1.3.11 we obtain the following inequality for a greedy basis Ψ (for $L_p([0,1))$)

$$\|f - G^{L_p}_m(f, \Psi^d)\|_p \le C(\Psi, d, p)(\log m)^{h(p,d)} \sigma_m(f, \Psi^d)_p, \quad 1 < p < \infty. \tag{1.3.16}$$

1.4. Quasi-greedy and almost greedy bases

In Section 1.3 we imposed the condition

$$\inf_k \|\psi_k\| > 0 \tag{1.4.1}$$

on a basis Ψ, to define $G_m(f, \Psi)$ for all $f \in X$. We noticed that in the case of a greedy basis this condition is always satisfied. In this section we assume that (1.4.1) is satisfied.

Let us discuss the question of weakening the requirement that a basis be a greedy basis. We begin with a concept of quasi-greedy basis that was introduced in Konyagin and Temlyakov (1999*a*).

Definition 1.4.1. We call a basis Ψ a quasi-greedy basis if, for every $f \in X$ and every permutation $\rho \in D(f)$, we have

$$\|G_m(f, \Psi, \rho)\|_X \le C\|f\|_X \tag{1.4.2}$$

with a constant C independent of f, m, and ρ.

It is clear that (1.4.2) is weaker then (1.3.6). Wojtaszczyk (2000) proved the following theorem.

Theorem 1.4.2. *A basis Ψ is quasi-greedy if and only if, for any $f \in X$ and any $\rho \in D(f)$, we have*

$$\|f - G_m(f, \Psi, \rho)\| \to 0 \quad \text{as } m \to \infty. \tag{1.4.3}$$

Theorem 1.4.2 allows us to use (1.4.3) as an equivalent definition of a quasi-greedy basis. We give one more equivalent definition of a quasi-greedy basis.

Definition 1.4.3. We say that a basis Ψ is quasi-greedy if there exists a constant C_Q such that, for any $f \in X$ and any finite set of indices Λ having the property

$$\min_{k\in\Lambda}|c_k(f)| \ge \max_{k\notin\Lambda}|c_k(f)|, \tag{1.4.4}$$

we have

$$\|S_\Lambda(f,\Psi)\| \le C_Q\|f\|. \tag{1.4.5}$$

It is clear that for elements f with the unique decreasing rearrangement of coefficients ($\#D(f)=1$), inequalities (1.4.2) and (1.4.5) are equivalent. By slightly modifying the coefficients and using the continuity argument we deduce that (1.4.2) and (1.4.5) are equivalent for general f.

We now continue a discussion from Section 1.3 of relations between the following concepts: greedy basis, unconditional basis, democratic basis, and quasi-greedy basis. Theorem 1.3.5 states that *greedy = unconditional + democratic*. It is clear from the definition of quasi-greedy basis that an unconditional basis is always a quasi-greedy basis. We now give an example from Konyagin and Temlyakov (1999*a*) of a basis that is quasi-greedy and democratic (even superdemocratic) and is not an unconditional basis.

It is clear that an unconditional and democratic basis Ψ satisfies the following inequality:

$$\Bigl\|\sum_{k\in P}\theta_k\psi_k\Bigr\| \le D_S\Bigl\|\sum_{k\in Q}\epsilon_k\psi_k\Bigr\|, \tag{1.4.6}$$

for any two finite sets P and Q, $|P|=|Q|$, and any choices of signs $\theta_k=\pm1$, $k\in P$, and $\epsilon_k=\pm1$, $k\in Q$.

Definition 1.4.4. We say that a basis Ψ is a superdemocratic basis if it satisfies (1.4.6).

Theorem 1.3.5 implies that a greedy basis is a superdemocratic one. Now we will construct an example of a superdemocratic quasi-greedy basis which is not an unconditional basis, and therefore, by Theorem 1.3.5, is not a greedy basis.

Let X be the set of all real sequences $x=(x_1,x_2,\dots)\in l_2$ such that

$$\|x\|_1 = \sup_{N\in\mathbb{N}}\Bigl|\sum_{n=1}^{N} x_n/\sqrt{n}\Bigr|$$

is finite. Clearly, X equipped with the norm

$$\|\cdot\| = \max(\|\cdot\|_{l_2},\|\cdot\|_1)$$

is a Banach space. Let $\psi_k\in X$, $k=1,2,\dots$, be defined as $(\psi_k)_n=1$ if $n=k$ and $(\psi_k)_n=0$ otherwise. Let X_0 denote the subspace of X generated by the elements ψ_k. It is easy to see that $\Psi=\{\psi_k\}$ is a democratic basis in X_0. Moreover, it is superdemocratic: for any $k_1,\dots,k_m$ and for any choice of signs,

$$\sqrt{m} \leq \left\| \sum_{j=1}^{m} \pm \psi_{k_j} \right\| < 2\sqrt{m}. \tag{1.4.7}$$

Indeed, we have

$$\left\| \sum_{j=1}^{m} \pm \psi_{k_j} \right\|_{l_2} = \sqrt{m},$$

$$\left\| \sum_{j=1}^{m} \pm \psi_{k_j} \right\|_{1} \leq \sum_{j=1}^{m} 1/\sqrt{j} < 2\sqrt{m},$$

and (1.4.7) follows. However, Ψ is not an unconditional basis since, for $m \geq 2$,

$$\left\| \sum_{k=1}^{m} \psi_k/\sqrt{k} \right\| \geq \sum_{k=1}^{m} 1/k \asymp \log m,$$

but

$$\left\| \sum_{k=1}^{m} (-1)^k \psi_k/\sqrt{k} \right\| \asymp \sqrt{\log m}.$$

We now prove that the basis Ψ constructed above is a quasi-greedy basis. Assume $\|f\| = 1$. Then, by definition of $\|\cdot\|$ we have

$$\sum_{k=1}^{\infty} |c_k(f)|^2 \leq 1, \tag{1.4.8}$$

and for any M

$$\left| \sum_{k=1}^{M} c_k(f) k^{-1/2} \right| \leq 1. \tag{1.4.9}$$

It is clear that for any Λ we have

$$\|S_\Lambda(f, \Psi)\|_{l_2} \leq \|f\|_{l_2} \leq 1. \tag{1.4.10}$$

We now estimate $\|S_\Lambda(f, \Psi)\|_1$. Let Λ be any finite set of indices satisfying (1.4.4), and let

$$\alpha := \min_{k \in \Lambda} |c_k(f)|.$$

If $\alpha = 0$, then $S_\Lambda(f, \Psi) = f$ and (1.4.5) holds. Therefore consider $\alpha > 0$, and, for any N, let

$$\Lambda^+(N) := \{k \in \Lambda : k > N\}, \qquad \Lambda^-(N) := \{k \in \Lambda : k \leq N\}.$$

By Hölder's inequality we have, for any N,

$$\sum_{k\in\Lambda^+(N)} |c_k(f)|k^{-1/2} \le \Big(\sum_{k\in\Lambda^+(N)} |c_k(f)|^{3/2}\Big)^{2/3}\Big(\sum_{k>N} k^{-3/2}\Big)^{1/3} \tag{1.4.11}$$

$$\ll N^{-1/6}\Big(\sum_{k\in\Lambda^+(N)} |c_k(f)|^{3/2}(|c_k(f)|/\alpha)^{1/2}\Big)^{2/3} \ll (\alpha^2 N)^{-1/6}.$$

Choose $N_\alpha := [\alpha^{-2}]+1$. Then, for any $M \le N_\alpha$ we have by (1.4.9) that

$$\Big|\sum_{k\in\Lambda^-(M)} c_k(f)k^{-1/2}\Big| \le \Big|\sum_{k=1}^{M} c_k(f)k^{-1/2}\Big| + \Big|\sum_{k\notin\Lambda^-(M),k\le M} c_k(f)k^{-1/2}\Big|$$

$$\le 1 + \alpha\sum_{k=1}^{M} k^{-1/2} \le 1 + 2\alpha M^{1/2} \ll 1. \tag{1.4.12}$$

For $M > N_\alpha$, we get using (1.4.11) and (1.4.12)

$$\Big|\sum_{k\in\Lambda^-(M)} c_k(f)k^{-1/2}\Big| \le \Big|\sum_{k\in\Lambda^-(N_\alpha)} c_k(f)k^{-1/2}\Big| + \sum_{k\in\Lambda^+(N_\alpha)} |c_k(f)|k^{-1/2} \ll 1.$$

Thus

$$\|S_\Lambda(f,\Psi)\|_1 \le C,$$

which completes the proof.

The above example and Theorem 1.3.5 show that a quasi-greedy basis is not necessarily a greedy basis. Further results on quasi-greedy bases can be found in Wojtaszczyk (2000) and Dilworth, Kalton, Kutzarova and Temlyakov (2003).

The above discussion shows that a quasi-greedy basis is not necessarily an unconditional basis. However, quasi-greedy bases have some properties that are close to those of unconditional bases. We formulate two of them (see, for instance, Konyagin and Temlyakov (2002)).

Lemma 1.4.5. Let Ψ be a quasi-greedy basis. Then, for any two finite sets of indices $A \subseteq B$ and coefficients $0 < t \le |a_j| \le 1$, $j \in B$, we have

$$\Big\|\sum_{j\in A} a_j\psi_j\Big\| \le C(X,\Psi,t)\Big\|\sum_{j\in B} a_j\psi_j\Big\|.$$

It will be convenient to define the quasi-greedy constant K to be the least constant such that

$$\|G_m(f)\| \le K\|f\| \quad \text{and} \quad \|f - G_m(f)\| \le K\|f\|, \quad f \in X.$$

Lemma 1.4.6. Suppose Ψ is a quasi-greedy basis with a quasi-greedy constant K. Then, for any real numbers a_j and any finite set of indices P, we have

$$(4K^2)^{-1}\min_{j\in P}|a_j|\left\|\sum_{j\in P}\psi_j\right\| \le \left\|\sum_{j\in P}a_j\psi_j\right\| \le 2K\max_{j\in P}|a_j|\left\|\sum_{j\in P}\psi_j\right\|.$$

We note that the mth greedy approximant $G_m(x,\Psi)$ changes if we renormalize the basis Ψ (replace it by a basis $\{\lambda_n\psi_n\}$). This gives us more flexibility in adjusting a given basis Ψ for greedy approximation. Let us make one observation from Konyagin and Temlyakov (2003*a*) along these lines.

Proposition 1.4.7. Let $\Psi = \{\psi_n\}_{n=1}^{\infty}$ be a normalized basis for a Banach space X. Then the basis $\{e_n\}_{n=1}^{\infty}$, $e_n := 2^n\psi_n$, $n = 1,2,\dots$ is a quasi-greedy basis in X.

We proceed to an intermediate concept of *almost greedy basis.* This concept was introduced and studied in Dilworth *et al.* (2003). Let

$$f = \sum_{k=1}^{\infty} c_k(f)\psi_k.$$

We define the following expansional best m-term approximation of f:

$$\tilde{\sigma}_m(f) := \tilde{\sigma}_m(f,\Psi) := \inf_{\Lambda,|\Lambda|=m}\left\|f - \sum_{k\in\Lambda}c_k(f)\psi_k\right\|.$$

It is clear that

$$\sigma_m(f,\Psi) \le \tilde{\sigma}_m(f,\Psi).$$

It is also clear that for an unconditional basis Ψ we have

$$\tilde{\sigma}_m(f,\Psi) \le C\sigma_m(f,\Psi).$$

Definition 1.4.8. We call a basis Ψ an almost greedy basis if, for every $f \in X$, there exists a permutation $\rho \in D(f)$ such that we have the inequality

$$\|f - G_m(f,\Psi,\rho)\|_X \le C\tilde{\sigma}_m(f,\Psi)_X, \tag{1.4.13}$$

with a constant independent of f and m.

The following proposition follows from the proof of Theorem 3.3 of Dilworth *et al.* (2003) (see Theorem 1.4.10 below).

Proposition 1.4.9. If Ψ is an almost greedy basis then (1.4.13) holds for any permutation $\rho \in D(f)$.

The following characterization of almost greedy bases was obtained in Dilworth *et al.* (2003).

Theorem 1.4.10. Suppose Ψ is a basis of a Banach space. The following are equivalent.

A Ψ is almost greedy.

B Ψ is quasi-greedy and democratic.

C For any (respectively, every) $\lambda > 1$ there is a constant $C = C_\lambda$ such that

$$\|f - G_{[\lambda m]}(f, \Psi)\| \leq C_\lambda \sigma_m(f, \Psi).$$

In order to give the reader an idea of relations between $\tilde{\sigma}$ and σ we present an estimate for $\tilde{\sigma}_n(f, \Psi)$ in terms of $\sigma_m(f, \Psi)$ for a quasi-greedy basis Ψ. For a basis Ψ we define the fundamental function

$$\varphi(m) := \sup_{|A| \leq m} \left\| \sum_{k \in A} \psi_k \right\|.$$

We also need the following function:

$$\phi(m) := \inf_{|A| = m} \left\| \sum_{k \in A} \psi_k \right\|.$$

The following inequality was obtained in Dilworth *et al.* (2003).

Theorem 1.4.11. Let Ψ be a quasi-greedy basis. Then, for any m and r there exists a set E, $|E| \leq m + r$ such that

$$\|f - S_E(f, \Psi)\| \leq C\left(1 + \frac{\varphi(m)}{\phi(r+1)}\right)\sigma_m(f, \Psi).$$

In Section 1.3, in addition to bases Ψ satisfying (1.4.1), we discussed a more general case that included bases satisfying (1.3.3). In the latter case we defined the greedy algorithm $G_m(f, \Psi)$ for functions f of the form (1.3.4). We gave a definition of a greedy basis in the general case, which included those bases satisfying (1.3.3). However, the characterization of greedy bases given by Theorem 1.3.5 excluded bases satisfying (1.3.3). We note that a similar attempt to include bases Ψ satisfying (1.3.3) into the consideration of quasi-greedy bases does not work. Indeed, let Ψ be a normalized unconditional basis and consider a renormalized basis $\Psi' := \{\psi'_k := k^{-3}\psi_k\}$. Clearly, Ψ' is also an unconditional basis, and therefore inequality (1.4.2) is satisfied for any f of the form (1.3.4). However, for the function

$$f := \sum_{k=1}^{\infty} k^{-2}\psi_k = \sum_{k=1}^{\infty} k\psi'_k,$$

we cannot apply the algorithm $G_m(\cdot, \Psi')$ because the expansion coefficients are not bounded.

1.5. Weak Greedy Algorithms with respect to bases

The greedy approximant $G_m(f,\Psi)$ considered in Sections 1.3 and 1.4 was defined to be the sum

$$\sum_{j=1}^{m} c_{k_j}(f,\Psi)\psi_{k_j}$$

of the expansion terms with the m biggest coefficients in absolute value (see (1.3.2)). In this section we discuss a more flexible way to construct a greedy approximant. The rule for choosing the expansion terms for approximation will be weaker than in the greedy algorithm $G_m(\cdot,\Psi)$. Instead of taking m terms with the biggest coefficients we now take m terms with near-biggest coefficients. We proceed to a formal definition of the Weak Greedy Algorithm with regard to a basis Ψ. We assume here that Ψ satisfies (1.4.1).

Let $t \in (0,1]$ be a fixed parameter. For a given basis Ψ and a given $f \in X$, let $\Lambda_m(t)$ be any set of m indices such that

$$\min_{k\in\Lambda_m(t)} |c_k(f,\Psi)| \ge t \max_{k\notin\Lambda_m(t)} |c_k(f,\Psi)|, \tag{1.5.1}$$

and define

$$G_m^t(f) := G_m^t(f,\Psi) := \sum_{k\in\Lambda_m(t)} c_k(f,\Psi)\psi_k.$$

We call it the Weak Greedy Algorithm (WGA) with the weakness sequence $\{t\}$ (the weakness parameter t). We note that the WGA with regard to a basis was introduced in the very first paper (see Temlyakov (1998*a*)) on greedy bases. It is clear that $G_m^1(f,\Psi) = G_m(f,\Psi)$. It is also clear that, in the case $t < 1$, we have more flexibility in building a weak greedy approximant $G_m^t(f,\Psi)$ than in building $G_m(f,\Psi)$: it is one advantage of a weak greedy approximant $G_m^t(f,\Psi)$. The question is: How much does this flexibility affect efficiency of the algorithm? Surprisingly, it turns out that the effect is minimal: it is only reflected in a multiplicative constant (see below).

We begin our discussion with the case when Ψ is a greedy basis. It was proved in Temlyakov (1998*a*) that, when $X = L_p$, $1 < p < \infty$, and Ψ is the Haar system $\mathcal{H}_p$ normalized in L_p, we have

$$\|f - G_m^t(f,\mathcal{H}_p)\|_{L_p} \le C(p,t)\sigma_m(f,\mathcal{H}_p)_{L_p}, \tag{1.5.2}$$

for any $f \in L_p$. It was noted in Konyagin and Temlyakov (2002) that the proof of (1.5.2) from Temlyakov (1998*a*) works for any greedy basis, not merely the Haar system $\mathcal{H}_p$. Thus, we have the following result.

Theorem 1.5.1. For any greedy basis Ψ of a Banach space X, and any $t \in (0,1]$, we have

$$\|f - G_m^t(f,\Psi)\|_X \le C(\Psi,t)\sigma_m(f,\Psi)_X, \tag{1.5.3}$$

for each $f \in X$.

We now consider the Weak Greedy Algorithm with regard to a quasi-greedy basis Ψ. It was proved in Konyagin and Temlyakov (2002) that the weak greedy approximant has properties similar to the greedy approximant.

Theorem 1.5.2. Let Ψ be a quasi-greedy basis. Then, for a fixed $t \in (0,1]$ and any m, we have for any $f \in X$

$$\|G^t_m(f,\Psi)\| \le C(t)\|f\|. \tag{1.5.4}$$

The following theorem from Konyagin and Temlyakov (2002) is essentially due to Wojtaszczyk (2000).

Theorem 1.5.3. Let Ψ be a quasi-greedy basis for a Banach space X. Then, for any fixed $t \in (0,1]$, we have for each $f \in X$ that

$$G^t_m(f,\Psi) \to f \quad \text{as } m \to \infty.$$

Let us now proceed to an almost greedy basis Ψ. The following result was established in Konyagin and Temlyakov (2002).

Theorem 1.5.4. Let Ψ be an almost greedy basis. Then, for $t \in (0,1]$ we have for any m

$$\|f - G^t_m(f,\Psi)\| \le C(t)\tilde{\sigma}_m(f,\Psi). \tag{1.5.5}$$

Proof. We drop Ψ from the notation for the sake of brevity. Take any $\epsilon > 0$ and find P, $|P| = m$ such that

$$\|f - S_P(f)\| \le \tilde{\sigma}_m(f) + \epsilon.$$

Let $Q := \Lambda_m(t)$ with $\Lambda_m(t)$ from the definition of $G^t_m(f)$. Then

$$\|f - G^t_m(f)\| \le \|f - S_P(f)\| + \|S_P(f) - S_Q(f)\|. \tag{1.5.6}$$

We have

$$S_P(f) - S_Q(f) = S_{P\setminus Q}(f) - S_{Q\setminus P}(f). \tag{1.5.7}$$

Let us first estimate $\|S_{Q\setminus P}(f)\|$. Denote $f_1 := f - S_P(f)$. Then

$$S_{Q\setminus P}(f) = S_{Q\setminus P}(f_1).$$

Next,

$$\begin{aligned}\min_{k\in Q\setminus P}|c_k(f_1)| = \min_{k\in Q\setminus P}|c_k(f)| \ge \min_{k\in Q}|c_k(f)| \\ \ge t\max_{k\notin Q}|c_k(f)| \ge t\max_{k\notin Q}|c_k(f_1)| = t\max_{k\notin Q\setminus P}|c_k(f_1)|.\end{aligned}$$

Thus $Q \setminus P = \Lambda_n(t)$ for f_1 with $n := |Q \setminus P|$. By Theorem 1.5.2 we have

$$\|S_{Q\setminus P}(f)\| \le C_1(t)\|f_1\|. \tag{1.5.8}$$

We now estimate $\|S_{P\setminus Q}(f)\|$. From the definition of Q we easily derive

$$at \le b, \quad \text{where} \quad a := \max_{k\in P\setminus Q} |c_k(f)|, \quad b := \min_{k\in Q\setminus P} |c_k(f)|. \tag{1.5.9}$$

By Lemma 1.4.6 (see Lemma 2.1 from Dilworth *et al.* (2003)),

$$\|S_{P\setminus Q}(f)\| \le 2Ka \Big\| \sum_{k\in P\setminus Q} \psi_k \Big\| \tag{1.5.10}$$

and (see Lemma 2.2 from Dilworth *et al.* (2003))

$$\|S_{Q\setminus P}(f)\| \ge (4K^2)^{-1} b \Big\| \sum_{k\in Q\setminus P} \psi_k \Big\|. \tag{1.5.11}$$

By Theorem 1.4.10 an almost greedy basis is a democratic basis. Thus we obtain

$$\Big\| \sum_{k\in P\setminus Q} \psi_k \Big\| \le D \Big\| \sum_{k\in Q\setminus P} \psi_k \Big\|. \tag{1.5.12}$$

Combining (1.5.6)–(1.5.12) we obtain (1.5.5). Theorem 1.5.4 is proved. □

We now discuss the stability of the greedy-type property of a basis. Let $0 < a \le \lambda_k \le b < \infty$, $k = 1, 2, \dots$ and for a basis $\Psi = \{\psi_k\}$ consider $\Psi^\lambda := \{\lambda_k \psi_k\}$. The following theorem is from Konyagin and Temlyakov (2002). We note that the case for quasi-greedy bases was proved in Wojtaszczyk (2000).

Theorem 1.5.5. Let a basis Ψ have one of the following properties:

(1) greedy,
(2) almost greedy,
(3) quasi-greedy.

Then the basis Ψ^λ has the same property.

Proof. Let $f \in X$ and

$$f = \sum_k c_k(f)\psi_k = \sum_k c_k(f)\lambda_k^{-1}\lambda_k\psi_k.$$

Consider

$$G_m(f, \Psi^\lambda) = \sum_{k\in\Lambda_m} (c_k(f)\lambda_k^{-1})\lambda_k\psi_k.$$

Then, using $\lambda_k \in [a, b]$ and the definition of the $G_m(f, \Psi^\lambda)$, we obtain

$$\min_{k\in\Lambda_m} |c_k(f)| \ge a \min_{k\in\Lambda_m} |c_k(f)|\lambda_k^{-1} \ge a \max_{k\notin\Lambda_m} |c_k(f)|\lambda_k^{-1} \ge \frac{a}{b} \max_{k\notin\Lambda_m} |c_k(f)|.$$

Therefore, the set Λ_m can be interpreted as a $\Lambda_m(t)$ with $t = a/b$ with regard to the basis Ψ. It remains to apply the corresponding results for $G_m^t(f, \Psi)$: (1.5.3) in case (1), (1.5.4) in case (3), and (1.5.5) in case (2). This completes the proof of Theorem 1.5.5. □

Kamont and Temlyakov (2004) studied the following modification of the above weak-type greedy algorithm as a way to further weaken restriction (1.5.1). We call this modification the Weak Greedy Algorithm (WGA) with a weakness sequence $\tau = \{t_k\}$. Let a weakness sequence $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \in [0, 1]$, $k = 1, \dots$ be given. We define the WGA by induction. We take an element $f \in X$, and at the first step we let

$$\Lambda_1(\tau) := \{n_1\}, \qquad G_1^\tau(f, \Psi) := c_{n_1}\psi_{n_1},$$

with any n_1 satisfying

$$|c_{n_1}| \geq t_1 \max_n |c_n|,$$

where we write $c_n := c_n(f, \Psi)$ for brevity. Assume we have already defined

$$G_{m-1}^\tau(f, \Psi) := G_{m-1}^{X,\tau}(f, \Psi) := \sum_{n \in \Lambda_{m-1}(\tau)} c_n \psi_n.$$

Then, at the mth step we define

$$\Lambda_m(\tau) := \Lambda_{m-1}(\tau) \cup \{n_m\}, \qquad G_m^\tau(f, \Psi) := G_m^{X,\tau}(f, \Psi) := \sum_{n \in \Lambda_m(\tau)} c_n \psi_n,$$

with any $n_m \notin \Lambda_{m-1}(\tau)$ satisfying

$$|c_{n_m}| \geq t_m \max_{n \notin \Lambda_{m-1}(\tau)} |c_n|.$$

Thus, for $f \in X$ the WGA builds a rearrangement of a subsequence of the expansion (1.3.1). If Ψ is an unconditional basis then we also have the limit $G_m^\tau(f, \Psi) \to f^*$. It is clear that in this case $f^* = f$ if and only if the sequence $\{n_k\}_{k=1}^{\infty}$ contains indices of all non-zero $c_n(f, \Psi)$. We say that the WGA corresponding to Ψ and τ is convergent if, for any realization $G_m^\tau(f, \Psi)$, we have

$$\|f - G_m^\tau(f, \Psi)\| \to 0 \quad \text{as } m \to \infty,$$

for all $f \in X$.

We formulate here only one theorem from Kamont and Temlyakov (2004).

Theorem 1.5.6. Let $2 \leq p < \infty$, $d \geq 1$ and let Ψ be a normalized unconditional basis in $L_p([0, 1]^d)$. Let $\tau = \{t_n : n \geq 1\}$ be a weakness sequence. Then the WGA corresponding to Ψ and τ converges if and only if $\tau \notin l_p$.

1.6. Thresholding and minimal systems

In this section we briefly discuss some further generalizations. Here, we assume that X is a quasi-Banach space and replace a basis by a complete minimal system. In addition, we consider the Weak Thresholding Algorithm and prove that its convergence is equivalent to convergence of the Weak Greedy Algorithm (see Proposition 1.6.3). Thresholding algorithms are very useful in statistics (see, for instance, Donoho and Johnstone (1994)).

Let X be a quasi-Banach space (real or complex) with the quasi-norm $\|\cdot\|$ such that for all $x, y \in X$ we have $\|x+y\| \leq \alpha(\|x\|+\|y\|)$ and $\|tx\| = |t|\|x\|$. It is well known (see Kalton, Beck and Roberts (1984, Lemma 1.1)) that there is a p, $0 < p \leq 1$, such that

$$\left\|\sum_n x_n\right\| \leq 4^{1/p}\left(\sum_n \|x_n\|^p\right)^{1/p}. \tag{1.6.1}$$

Let $\{e_n\} \subset X$ be a complete minimal system in X with the conjugate (dual) system $\{e_n^*\} \subset X^*$ ($e_n^*(e_n) = 1$, $e_n^*(e_k) = 0$, $k \neq n$). We assume that $\sup_n \|e_n^*\| < \infty$. This implies that for each $x \in X$ we have

$$\lim_{n\to\infty} e_n^*(x) = 0. \tag{1.6.2}$$

Any element $x \in X$ has a formal expansion

$$x \sim \sum_n e_n^*(x) e_n, \tag{1.6.3}$$

and various types of convergence of the series (1.6.3) can be studied. In this section we deal with greedy-type approximations with regard to the system $\{e_n\}$. We note that in this section we use the notation x and $\{e_n\}$ for an element and for a system, respectively, differing from the notation f and Ψ used in previous sections, to emphasize that we are now in a more general setting. It will be convenient for us to define a unique 'greedy ordering' in this section. For any $x \in X$ we define the greedy ordering for x as the map $\rho : \mathbb{N} \to \mathbb{N}$ for which $\{j : e_j^*(x) \neq 0\} \subset \rho(\mathbb{N})$, and such that, if $j < k$, then either $|e_{\rho(j)}^*(x)| > |e_{\rho(k)}^*(x)|$ or $|e_{\rho(j)}^*(x)| = |e_{\rho(k)}^*(x)|$ and $\rho(j) < \rho(k)$. The mth greedy approximation is given by

$$G_m(x) := G_m(x, \{e_n\}) := \sum_{j=1}^m e_{\rho(j)}^*(x) e_{\rho(j)}.$$

The system $\{e_n\}$ is a quasi-greedy system (Konyagin and Temlyakov 1999*a*) if there exists a constant C such that $\|G_m(x)\| \leq C\|x\|$ for all $x \in X$ and $m \in \mathbb{N}$. Wojtaszczyk (2000) proved that these are precisely the systems for which $\lim_{m\to\infty} G_m(x) = x$ for all x. If, as in Section 1.4, a quasi-greedy system $\{e_n\}$ is a basis, then we say that $\{e_n\}$ is a quasi-greedy basis. As we mentioned above, it is clear that any unconditional basis is a quasi-greedy basis. We note that there are conditional quasi-greedy bases $\{e_n\}$ in some Banach spaces. Hence, for such a basis $\{e_n\}$ there exists a permutation of $\{e_n\}$ which forms a quasi-greedy system but not a basis. This remark justifies the study of the class of quasi-greedy systems rather than the class of quasi-greedy bases.

Greedy approximations are close to thresholding approximations (sometimes they are called *thresholding greedy approximations*). Thresholding approximations are defined by

$$T_\epsilon(x) := \sum_{|e_j^*(x)|\geq\epsilon} e_j^*(x)e_j, \quad \epsilon > 0.$$

Clearly, for any $\epsilon > 0$ there exists an m such that $T_\epsilon(x) = G_m(x)$. Therefore, if $\{e_n\}$ is a quasi-greedy system then

$$\forall x \in X \quad \lim_{\epsilon\to 0} T_\epsilon(x) = x. \tag{1.6.4}$$

Conversely, following the Remark from Wojtaszczyk (2000, pp. 296–297), it is easy to show that condition (1.6.4) implies that $\{e_n\}$ is a quasi-greedy system.

As in Section 1.5, one can define the Weak Thresholding Approximation. Fix $t \in (0,1)$. For $\epsilon > 0$ let

$$D_{t,\epsilon}(x) := \{j : t\epsilon \leq |e_j^*(x)| < \epsilon\}.$$

The Weak Thresholding Approximations are defined as all possible sums

$$T_{\epsilon,D}(x) = \sum_{|e_j^*(x)|\geq\epsilon} e_j^*(x)e_j + \sum_{j\in D} e_j^*(x)e_j,$$

where $D \subseteq D_{t,\epsilon}(x)$. We say that the Weak Thresholding Algorithm converges for $x \in X$, and write $x \in \mathrm{WT}\{e_n\}(t)$ if, for any $D(\epsilon) \subseteq D_{t,\epsilon}$,

$$\lim_{\epsilon\to 0} T_{\epsilon,D(\epsilon)}(x) = x.$$

It is clear that the above relation is equivalent to

$$\lim_{\epsilon\to 0} \sup_{D\subseteq D_{t,\epsilon}(x)} \|x - T_{\epsilon,D}(x)\| = 0.$$

We proved in Konyagin and Temlyakov (2003*a*) (see Theorem 1.6.1 below) that the set $\mathrm{WT}\{e_n\}(t)$ does not depend on $t \in (0,1)$. Therefore, we can drop t from the notation: $\mathrm{WT}\{e_n\} = \mathrm{WT}\{e_n\}(t)$.

It turns out that the Weak Thresholding Algorithm has more regularity than the Thresholding Algorithm: we will see that the set $\mathrm{WT}\{e_n\}$ is linear. On the other hand, by ‘weakening’ the Thresholding Algorithm (making convergence stronger), we do not narrow the convergence set too much. It is known that for many natural classes of sets $Y \subseteq X$ the convergence of $T_\epsilon(x)$ to x for all $x \in Y$ is equivalent to the condition $Y \subseteq \mathrm{WT}\{e_n\}$. In particular, it can be derived from Wojtaszczyk (2000, Proposition 3) that the above two conditions are equivalent for $Y = X$.

We suppose that X and $\{e_n\}$ satisfy the conditions stated in the beginning of this section. The following two theorems were proved in Konyagin and Temlyakov (2003*a*).

Theorem 1.6.1. Let $t, t' \in (0,1)$, $x \in X$. Then the following conditions are equivalent.

(1) $\lim_{\epsilon\to 0} \sup_{D\subseteq D_{t,\epsilon}(x)} \|T_{\epsilon,D}(x) - x\| = 0$.

(2) $\lim_{\epsilon\to 0} T_\epsilon(x) = x$ and

$$\lim_{\epsilon\to 0} \sup_{D\subseteq D_{t,\epsilon}(x)} \left\| \sum_{j\in D} e_j^*(x)e_j \right\| = 0. \tag{1.6.5}$$

(3) $\lim_{\epsilon\to 0} T_\epsilon(x) = x$ and

$$\lim_{\epsilon\to 0} \sup_{|a_j|\le 1 (j\in D_{t,\epsilon}(x))} \left\| \sum_{j\in D_{t,\epsilon}(x)} a_j e_j^*(x)e_j \right\| = 0. \tag{1.6.6}$$

(4) $\lim_{\epsilon\to 0} T_\epsilon(x) = x$ and

$$\lim_{\epsilon\to 0} \sup_{|b_j|<\epsilon (j:|e_j^*(x)|\ge\epsilon)} \left\| \sum_{j:|e_j^*(x)|\ge\epsilon} b_j e_j \right\| = 0. \tag{1.6.7}$$

(5) $\lim_{\epsilon\to 0} \sup_{D\subseteq D_{t',\epsilon}(x)} \|T_{\epsilon,D}(x) - x\| = 0$.

So, the set $\mathrm{WT}\{e_n\}(t)$ defined above is indeed independent of $t \in (0,1)$.

Theorem 1.6.2. The set $\mathrm{WT}\{e_n\}$ is linear.

Let us discuss relations between the Weak Thresholding Algorithm $T_{\epsilon,D}(x)$ and the Weak Greedy Algorithm $G_m^t(x)$. We define $G_m^t(x)$ with regard to a minimal system $\{e_n\}$ in the same way as it was defined for a basis Ψ. For a given system $\{e_n\}$ and $t \in (0,1]$, we denote for $x \in X$ and $m \in \mathbb{N}$ by $W_m(t)$ any set of m indices such that

$$\min_{j\in W_m(t)} |e_j^*(x)| \ge t \max_{j\notin W_m(t)} |e_j^*(x)|, \tag{1.6.8}$$

and define

$$G_m^t(x) := G_m^t(x,\{e_n\}) := S_{W_m(t)}(x) := \sum_{j\in W_m(t)} e_j^*(x)e_j.$$

It is clear that for any $t \in (0,1]$ and any $D \subseteq D_{t,\epsilon}(x)$ there exist m and $W_m(t)$ satisfying (1.6.8) such that

$$T_{\epsilon,D}(x) = S_{W_m(t)}(x). \tag{1.6.9}$$

Thus the convergence $G_m^t(x) \to x$ as $m \to \infty$ implies the convergence $T_{\epsilon,D}(x) \to x$ as $\epsilon \to 0$ for any $t \in (0,1]$. We will now prove (see Konyagin and Temlyakov (2003*a*, Proposition 2.2)) that for $t \in (0,1)$ the inverse is also true.

Proposition 1.6.3. Let $t \in (0,1)$ and $x \in X$. Then the following two conditions are equivalent:

$$\lim_{\epsilon \to 0} \sup_{D \subseteq D_{t,\epsilon}(x)} \|T_{\epsilon,D}(x) - x\| = 0, \tag{1.6.10}$$

$$\lim_{m \to \infty} \|G_m^t(x) - x\| = 0, \tag{1.6.11}$$

for any realization $G_m^t(x)$.

Proof. The implication (1.6.11) ⇒ (1.6.10) is simple and follows from the remark following (1.6.9). We prove that (1.6.10) ⇒ (1.6.11). Let

$$\epsilon_m := \max_{j \notin W_m(t)} |e_j^*(x)|.$$

Clearly $\epsilon_m \to 0$ as $m \to \infty$. We have

$$G_m^t(x) = T_{2\epsilon_m}(x) + \sum_{j \in D_m} e_j^*(x) e_j, \tag{1.6.12}$$

with D_m having the following property: for any $j \in D_m$,

$$t\epsilon_m \le |e_j^*(x)| < 2\epsilon_m.$$

Thus, by condition (5) from Theorem 1.6.1, for $t' = t/2$ we obtain (1.6.11).

Proposition 1.6.3 is now proved. □

Proposition 1.6.3 and Theorem 1.6.1 imply that the convergence set of the Weak Greedy Algorithm $G_m^t(\cdot)$ does not depend on $t \in (0,1)$ and coincides with $\mathrm{WT}\{e_n\}$. By Theorem 1.6.2 this set is a linear set.

Let us make a comment on the case $t = 1$ that is not covered by Proposition 1.6.3. It is clear that $T_\epsilon(x) = G_m(x)$ with some m, and therefore $G_m(x) \to x$ as $m \to \infty$ implies $T_\epsilon(x) \to x$ as $\epsilon \to 0$. It is also not difficult to understand that, in general, $T_\epsilon(x) \to x$ as $\epsilon \to 0$ does not imply $G_m(x) \to x$ as $m \to \infty$. This can be done, for instance, by considering the trigonometric system in the space L_p, $p \neq 2$, and using the Rudin–Shapiro polynomials (see Temlyakov (1998*c*)). However, if, for the trigonometric system, we put the Fourier coefficients with equal absolute values in a natural order (say, lexicographic), then, in the case $1 < p < \infty$, by Riesz's theorem we obtain convergence of $G_m(f)$ from convergence of $T_\epsilon(f)$. Results from Konyagin and Skopina (2001) show that the situation is different for $p = 1$. In this case the natural order does not help to derive convergence of $G_m(f)$ from convergence of $T_\epsilon(f)$.

1.7. Greedy approximation with respect to the trigonometric system

The first results (see Theorem 1.3.1) on greedy approximation with regard to bases showed that the Haar basis and other bases similar to it are very well designed for greedy approximation. In this section we discuss another classical system, namely, the trigonometric system from the point of view of greedy approximation. It is well known that the trigonometric system is not an unconditional basis for L_p, $p \neq 2$. Therefore, by Theorem 1.3.5 it is not a greedy basis for L_p, $p \neq 2$. In this section we mostly discuss convergence properties of the Weak Greedy Algorithm with regard to the trigonometric system. It is a non-trivial problem. We will demonstrate how it relates to some deep results in harmonic and functional analysis.

Consider a periodic function $f \in L_p(\mathbb{T}^d)$, $1 \leq p \leq \infty$, $(L_\infty(\mathbb{T}^d) = \mathcal{C}(\mathbb{T}^d))$, defined on the d-dimensional torus $\mathbb{T}^d$. Let a number $m \in \mathbb{N}$ and a number $t \in (0,1]$ be given, and let Λ_m be a set of $k \in \mathbb{Z}^d$ with the properties

$$\min_{k \in \Lambda_m} |\hat{f}(k)| \geq t \max_{k \notin \Lambda_m} |\hat{f}(k)|, \quad |\Lambda_m| = m, \tag{1.7.1}$$

where

$$\hat{f}(k) := (2\pi)^{-d} \int_{\mathbb{T}^d} f(x) \mathrm{e}^{-\mathrm{i}(k,x)}\, \mathrm{d}x$$

is a Fourier coefficient of f. We define

$$G_m^t(f) := G_m^t(f, \mathcal{T}^d) := S_{\Lambda_m}(f) := \sum_{k \in \Lambda_m} \hat{f}(k) \mathrm{e}^{\mathrm{i}(k,x)},$$

and call it an mth weak greedy approximant of f with regard to the trigonometric system $\mathcal{T}^d := \{\mathrm{e}^{\mathrm{i}(k,x)}\}_{k \in \mathbb{Z}^d}$, $\mathcal{T} := \mathcal{T}^1$. We write $G_m(f) = G_m^1(f)$ and call it an mth greedy approximant. Clearly, an mth weak greedy approximant and even an mth greedy approximant may not be unique. In this section we do not impose any extra restrictions on Λ_m in addition to (1.7.1). Thus, theorems formulated below hold for any choice of Λ_m satisfying (1.7.1), or, in other words, for any realization $G_m^t(f)$ of the weak greedy approximation.

We will discuss in detail only results concerning convergence of the WGA with regard to the trigonometric system. T. W. Körner (1996), answering a question raised by Carleson and Coifman, constructed a function from $L_2(\mathbb{T})$ and then, in Körner (1999), a continuous function such that $\{G_m(f, \mathcal{T})\}$ diverges almost everywhere. It was proved in Temlyakov (1998*c*) for $p \neq 2$, and in Cordoba and Fernandez (1998) for $p < 2$, that there exists an $f \in L_p(\mathbb{T})$ such that $\{G_m(f, \mathcal{T})\}$ does not converge in L_p. It was remarked in Temlyakov (2003*a*) that the method from Temlyakov (1998*c*) gives a little more.

(1) There exists a continuous function f such that $\{G_m(f,\mathcal{T})\}$ does not converge in $L_p(\mathbb{T})$ for any $p > 2$.

(2) There exists a function f that belongs to any $L_p(\mathbb{T})$, $p < 2$, such that $\{G_m(f,\mathcal{T})\}$ does not converge in measure.

Thus the above negative results show that the condition $f \in L_p(\mathbb{T}^d)$, $p \neq 2$, does not guarantee convergence of $\{G_m(f,\mathcal{T})\}$ in the L_p-norm. The main goal of this section is to complement the survey of Temlyakov (2003*a*) by recent results in the following setting: find an additional (to $f \in L_p$) condition on f to guarantee that $\|f - G_m(f,\mathcal{T})\|_p \to 0$ as $m \to \infty$. In Konyagin and Temlyakov (2003*b*) we proved the following theorem.

Theorem 1.7.1. Let $f \in L_p(\mathbb{T}^d)$, $2 < p \le \infty$, and let $q > p' := p/(p-1)$. Assume that f satisfies the condition

$$\sum_{|k|>n} |\hat{f}(k)|^q = o(n^{d(1-q/p')})$$

where $|k| := \max_{1\le j\le d} |k_j|$. Then we have

$$\lim_{m\to\infty} \|f - G_m^t(f,\mathcal{T}^d)\|_p = 0.$$

It was proved in Konyagin and Temlyakov (2003*b*) that Theorem 1.7.1 is sharp.

Proposition 1.7.2. For each $2 < p \le \infty$ there exists $f \in L_p(\mathbb{T}^d)$ such that

$$|\hat{f}(k)| = O(|k|^{-d(1-1/p)}),$$

and the sequence $\{G_m(f)\}$ diverges in L_p.

Let us make some comments. For a given set Λ denote

$$E_\Lambda(f)_p := \inf_{c_k, k\in\Lambda} \Big\| f - \sum_{k\in\Lambda} c_k \mathrm{e}^{\mathrm{i}(k,x)} \Big\|_p, \quad S_\Lambda(f) := \sum_{k\in\Lambda} \hat{f}(k)\mathrm{e}^{\mathrm{i}(k,x)}.$$

Define a special domain

$$Q(n) := \{k : |k| \le n^{1/d}\}.$$

Remark 1.7.3. Theorem 1.7.1 implies that if $f \in L_p$, $2 < p \le \infty$, and

$$E_{Q(n)}(f)_2 = o(n^{1/p-1/2}),$$

then $G_m^t(f) \to f$ in L_p.

Remark 1.7.4. The proof of Proposition 1.7.2 (see Konyagin and Temlyakov (2003*b*)) implies that there is an $f \in L_p(\mathbb{T}^d)$ such that

$$E_{Q(n)}(f)_\infty = O(n^{1/p-1/2})$$

and $\{G_m(f)\}$ diverges in L_p, $2 < p \le \infty$.

We note that Remark 1.7.3 can also be obtained from some general inequalities for $\|f - G^t_m(f)\|_p$. As in the above general definition of best m-term approximation, we define the best m-term approximation with regard to $\mathcal{T}^d$:

$$\sigma_m(f)_p := \sigma_m(f, \mathcal{T}^d)_p := \inf_{k^j \in \mathbb{Z}^d, c_j} \left\| f - \sum_{j=1}^{m} c_j \mathrm{e}^{\mathrm{i}(k^j, x)} \right\|_p .$$

The following inequality was proved in Temlyakov (1998*c*) for $t = 1$ and in Konyagin and Temlyakov (2003*b*) for general t.

Theorem 1.7.5. For each $f \in L_p(\mathbb{T}^d)$ and any $0 < t \le 1$ we have

$$\|f - G^t_m(f)\|_p \le (1 + (2 + 1/t)m^{h(p)})\sigma_m(f)_p, \quad 1 \le p \le \infty, \tag{1.7.2}$$

where $h(p) := |1/2 - 1/p|$.

It was proved in Temlyakov (1998*c*) that the inequality (1.7.2) is sharp: there is a positive absolute constant C such that, for each m and $1 \le p \le \infty$, there exists a function $f \ne 0$ with the property

$$\|G_m(f)\|_p \ge C m^{h(p)} \|f\|_p. \tag{1.7.3}$$

The above inequality (1.7.3) shows that the trigonometric system is not a quasi-greedy basis for L_p, $p \ne 2$. We formulate one more inequality from Konyagin and Temlyakov (2003*b*).

Theorem 1.7.6. Let $2 \le p \le \infty$. Then, for any $f \in L_p(\mathbb{T}^d)$ and any Q, $|Q| \le m$, we have

$$\|f - G^t_m(f)\|_p \le \|f - S_Q(f)\|_p + (3 + 1/t)(2m)^{h(p)} E_Q(f)_2.$$

We present some results from Konyagin and Temlyakov (2003*b*) that are formulated in terms of the Fourier coefficients. For $f \in L_1(\mathbb{T}^d)$ let $\{\hat{f}(k(l))\}_{l=1}^{\infty}$ denote the decreasing rearrangement of $\{\hat{f}(k)\}_{k \in \mathbb{Z}^d}$, *i.e.*,

$$|\hat{f}(k(1))| \ge |\hat{f}(k(2))| \ge \cdots .$$

Let $a_n(f) := |\hat{f}(k(n))|$.

Theorem 1.7.7. Let $2 < p < \infty$ and let a decreasing sequence $\{A_n\}_{n=1}^{\infty}$ satisfy the condition

$$A_n = o(n^{1/p-1}) \quad \text{as } n \to \infty.$$

Then, for any $f \in L_p(\mathbb{T}^d)$ with the property $a_n(f) \le A_n$, $n = 1, 2, \ldots$, we have

$$\lim_{m \to \infty} \|f - G^t_m(f)\|_p = 0.$$

We also proved in Konyagin and Temlyakov (2003*b*) that, for any decreasing sequence $\{A_n\}$ satisfying

$$\limsup_{n\to\infty} A_n n^{1-1/p} > 0,$$

there exists a function $f \in L_p$ such that $a_n(f) \le A_n$, $n = 1, \dots,$ whose sequence $\{G_m(f)\}$ of greedy approximants is divergent in L_p.

In Konyagin and Temlyakov (2003*b*) we proved a necessary and sufficient condition on the majorant $\{A_n\}$ to guarantee, under the assumption that f is a continuous function, the uniform convergence of greedy approximants to a function f.

Theorem 1.7.8. Let a decreasing sequence $\{A_n\}_{n=1}^{\infty}$ satisfy the condition $(\mathcal{A}_\infty)$:

$$\sum_{M<n\le e^M} A_n = o(1) \quad \text{as } M \to \infty.$$

Then, for any $f \in \mathcal{C}(\mathbb{T})$ with the property $a_n(f) \le A_n$, $n = 1, 2, \dots,$ we have

$$\lim_{m\to\infty} \|f - G_m^t(f, \mathcal{T})\|_\infty = 0.$$

The condition $(\mathcal{A}_\infty)$ is very close to the convergence of the series $\sum_n A_n$; if the condition $(\mathcal{A}_\infty)$ holds then we have

$$\sum_{n=1}^{N} A_n = o(\log_*(N)), \quad \text{as } N \to \infty,$$

where a function $\log_*(u)$ is defined to be bounded for $u \le 0$ and to satisfy $\log_*(u) = \log_*(\log u) + 1$ for $u > 0$. The function $\log_*(u)$ grows more slowly than any iterated logarithmic function.

The condition $(\mathcal{A}_\infty)$ in Theorem 1.7.8 is sharp.

Theorem 1.7.9. Assume that a decreasing sequence $\{A_n\}_{n=1}^{\infty}$ does not satisfy the condition $(\mathcal{A}_\infty)$. Then there exists a function $f \in \mathcal{C}(\mathbb{T})$ with the property $a_n(f) \le A_n$, $n = 1, 2, \dots,$ and such that we have

$$\limsup_{m\to\infty} \|f - G_m(f, \mathcal{T})\|_\infty > 0$$

for some realization $G_m(f, \mathcal{T})$.

In Konyagin and Temlyakov (2005) we concentrated on imposing extra conditions in the following form. We assume that for some sequence $\{M(m)\}$, $M(m) > m$, we have

$$\|G_{M(m)}(f) - G_m(f)\|_p \to 0 \quad \text{as } m \to \infty.$$

When p is an even number, or $p = \infty$, we found, in Konyagin and Temlyakov

(2005), necessary and sufficient conditions on the growth of the sequence $\{M(m)\}$ to provide convergence $\|f - G_m(f)\|_p \to 0$ as $m \to \infty$. We proved the following theorem in Konyagin and Temlyakov (2005).

Theorem 1.7.10. Let $p = 2q$, $q \in \mathbb{N}$, be an even integer, $\delta > 0$. Assume that $f \in L_p(\mathbb{T})$ and there exists a sequence of positive integers $M(m) > m^{1+\delta}$ such that

$$\|G_m(f) - G_{M(m)}(f)\|_p \to 0 \quad \text{as } m \to \infty.$$

Then we have

$$\|G_m(f) - f\|_p \to 0 \quad \text{as } m \to \infty.$$

In Konyagin and Temlyakov (2005) we proved that the condition $M(m) > m^{1+\delta}$ cannot be replaced by the condition $M(m) > m^{1+o(1)}$.

Theorem 1.7.11. For any $p \in (2, \infty)$ there exists a function $f \in L_p(\mathbb{T})$ with an $L_p(\mathbb{T})$-divergent sequence $\{G_m(f)\}$ of greedy approximations with the following property. For any sequence $\{M(m)\}$ such that $m \le M(m) \le m^{1+o(1)}$, we have

$$\|G_{M(m)}(f) - G_m(f)\|_p \to 0, \quad \text{as } m \to \infty.$$

In Konyagin and Temlyakov (2005) we also considered the case $p = \infty$, and proved necessary and sufficient conditions for convergence of greedy approximations in the uniform norm. For a mapping $\alpha : W \to W$ we let α_k denote its k-fold iteration: $\alpha_k := \alpha \circ \alpha_{k-1}$.

Theorem 1.7.12. Let $\alpha : \mathbb{N} \to \mathbb{N}$ be strictly increasing. Then the following conditions are equivalent.

(a) For some $k \in \mathbb{N}$ and for any sufficiently large $m \in \mathbb{N}$, we have the inequality $\alpha_k(m) > e^m$.

(b) If $f \in C(\mathbb{T})$ and

$$\left\|G_{\alpha(m)}(f) - G_m(f)\right\|_\infty \to 0, \quad \text{as } m \to \infty,$$

then

$$\left\|f - G_m(f)\right\|_\infty \to 0 \quad \text{as } m \to \infty.$$

In order to illustrate the techniques used in the proofs of the above results we discuss some inequalities that were used in proving Theorems 1.7.10 and 1.7.12. The reader will also see from the further discussion a connection to some deep results in harmonic analysis. The general style of these inequalities is as follows. A function that has a sparse representation with regard to the trigonometric system cannot be approximated in L_p by functions with small Fourier coefficients. We begin our discussion with some concepts introduced in Konyagin and Temlyakov (2005) that are useful in proving such inequalities. The following new characteristic of a Banach space L_p plays an important role in such inequalities. We introduce some more notation. Let Λ be a finite subset of $\mathbb{Z}^d$. We let $|\Lambda|$ denote its cardinality and let $\mathcal{T}(\Lambda)$ be the span of $\{\mathrm{e}^{\mathrm{i}(k,x)}\}_{k\in\Lambda}$. Denote

$$\Sigma_m(\mathcal{T}) = \cup_{\Lambda:|\Lambda|\le m}\mathcal{T}(\Lambda).$$

For $f \in L_p$, $F \in L_{p'}$, $1 \le p \le \infty$, $p' = p/(p-1)$, we write

$$\langle F, f\rangle := \int_{\mathbb{T}^d} F\bar{f}\,\mathrm{d}\mu, \qquad \mathrm{d}\mu := (2\pi)^{-d}\,\mathrm{d}x.$$

Definition 1.7.13. Let Λ be a finite subset of $\mathbb{Z}^d$ and $1 \le p \le \infty$. We call a set $\Lambda' := \Lambda'(p,\gamma)$, $\gamma \in (0,1]$, a (p,γ)-dual to Λ if, for any $f \in \mathcal{T}(\Lambda)$, there exists $F \in \mathcal{T}(\Lambda')$ such that $\|F\|_{p'} = 1$ and $\langle F, f\rangle \ge \gamma\|f\|_p$.

Let $D(\Lambda, p, \gamma)$ denote the set of all (p,γ)-dual sets Λ'. The following function is important for us:

$$v(m,p,\gamma) := \sup_{\Lambda:|\Lambda|=m}\ \inf_{\Lambda'\in D(\Lambda,p,\gamma)} |\Lambda'|.$$

We note that in a particular case $p = 2q$, $q \in \mathbb{N}$ we have

$$v(m,p,1) \le m^{p-1}. \tag{1.7.4}$$

This follows immediately from the form of the norming functional F for $f \in L_p$:

$$F = f^{q-1}(\bar{f})^q\|f\|_p^{1-p}. \tag{1.7.5}$$

In Konyagin and Temlyakov (2005) we used the quantity $v(m,p,\gamma)$ in greedy approximation. We first prove a lemma.

Lemma 1.7.14. Let $2 \le p \le \infty$. For any $h \in \Sigma_m(\mathcal{T})$ and any $g \in L_p$, we have

$$\|h+g\|_p \ge \gamma\|h\|_p - v(m,p,\gamma)^{1-1/p}\|\{\hat{g}(k)\}\|_{\ell_\infty}.$$

Proof. Let $h \in \mathcal{T}(\Lambda)$ with $|\Lambda| = m$ and let $\Lambda' \in D(\Lambda,p,\gamma)$. Then, using the Definition 1.7.13 we find $F(h,\gamma) \in \mathcal{T}(\Lambda')$ such that

$$\|F(h,\gamma)\|_{p'} = 1 \quad \text{and} \quad \langle F(h,\gamma), h\rangle \ge \gamma\|h\|_p.$$

We have

$$\langle F(h,\gamma), h\rangle = \langle F(h,\gamma), h+g\rangle - \langle F(h,\gamma), g\rangle \le \|h+g\|_p + |\langle F(h,\gamma), g\rangle|.$$

Next,

$$|\langle F(h,\gamma), g\rangle| \le \|\{\hat{F}(h,\gamma)(k)\}\|_{\ell_1}\|\{\hat{g}(k)\}\|_{\ell_\infty}.$$

Using $F(h,\gamma) \in \mathcal{T}(\Lambda')$ and the Hausdorff–Young theorem (Zygmund 1959, Chapter 12, Section 1.2), we obtain

$$\begin{aligned}\|\{\hat{F}(h,\gamma)(k)\}\|_{\ell_1} &\le |\Lambda'|^{1-1/p}\|\{\hat{F}(h,\gamma)(k)\}\|_{\ell_p} \\ &\le |\Lambda'|^{1-1/p}\|F(h,\gamma)\|_{p'} = |\Lambda'|^{1-1/p}.\end{aligned}$$

We now combine the above inequalities and use the definition of $v(m,p,\gamma)$. □

Definition 1.7.15. Let X be a finite-dimensional subspace of L_p, $1 \le p \le \infty$. We call a subspace $Y \subset L_{p'}$ a (p,γ)-dual to X, $\gamma \in (0,1]$, if for any $f \in X$ there exists $F \in Y$ such that $\|F\|_{p'} = 1$ and $\langle F, f\rangle \ge \gamma\|f\|_p$.

As above, let $D(X,p,\gamma)$ denote the set of all (p,γ)-dual subspaces Y. Consider the following function:

$$w(m,p,\gamma) := \sup_{X:\dim X=m} \inf_{Y\in D(X,p,\gamma)} \dim Y.$$

We begin our discussion with a particular case: $p = 2q$, $q \in \mathbb{N}$. Let X be given and let $e_1,\dots,e_m$ form a basis of X. Using the Hölder inequality for n functions $f_1,\dots,f_n \in L_n$, we have

$$\int |f_1\cdots f_n|\,\mathrm{d}\mu \le \|f_1\|_n \cdots \|f_n\|_n.$$

Setting $f_i = |e_j|^{p'}$, $n = p-1$, we deduce that any function of the form

$$\prod_{i=1}^m |e_i|^{k_i}, \quad k_i \in \mathbb{N}, \quad \sum_{i=1}^m k_i = p-1,$$

belongs to $L_{p'}$. It now follows from (1.7.5) that

$$w(m,p,1) \le m^{p-1}, \quad p = 2q, \quad q \in \mathbb{N}. \tag{1.7.6}$$

There is a general theory of the uniform approximation property (UAP) which provides some estimates for $w(m,p,\gamma)$ and $v(m,p,\gamma)$. We give some definitions from this theory. For a given subspace X of L_p, $\dim X = m$, and a constant $K > 1$, let $k_p(X,K)$ be the smallest k such that there is an operator $I_X : L_p \to L_p$, with $I_X(f) = f$ for $f \in X$, $\|I_X\|_{L_p\to L_p} \le K$, and $\operatorname{rank} I_X \le k$. Define

$$k_p(m,K) := \sup_{X:\dim X=m} k_p(X,K),$$

and let us discuss how $k_p(m,K)$ can be used in estimating $w(m,p,\gamma)$. Consider the dual operator I_X^* to I_X. Then $\|I_X^*\|_{L_{p'}\to L_{p'}} \le K$ and $\operatorname{rank} I_X^* \le k_p(m,K)$. Let $f \in X$, $\dim X = m$, and let F_f be the norming functional for f. Define

$$F := I_X^*(F_f)/\|I_X^*(F_f)\|_{p'}.$$

Then, for any $f \in X$,

$$\langle f, I_X^*(F_f)\rangle = \langle I_X(f), F_f\rangle = \langle f, F_f\rangle = \|f\|_p$$

and

$$\|I_X^*(F_f)\|_{p'} \le K$$

imply

$$\langle f, F\rangle \ge K^{-1}\|f\|_p.$$

Therefore

$$w(m,p,K^{-1}) \le k_p(m,K). \tag{1.7.7}$$

We note that the behaviour of functions $w(m,p,\gamma)$ and $k_p(m,K)$ may be very different. Bourgain (1992) proved that, for any $p \in (1,\infty)$, $p \ne 2$, the function $k_p(m,K)$ grows faster than any polynomial in m. The estimate (1.7.6) shows that, in the particular case $p = 2q$, $q \in \mathbb{N}$, the growth of $w(m,p,\gamma)$ is at most polynomial. This means that we cannot expect to obtain accurate estimates for $w(m,p,K^{-1})$ using inequality (1.7.7). We give one more application of the UAP in the style of Lemma 1.7.14.

Lemma 1.7.16. Let $2 \le p \le \infty$. For any $h \in \Sigma_m(\mathcal{T})$ and any $g \in L_p$ we have

$$\|h+g\|_p \ge K^{-1}\|h\|_p - k_p(m,K)^{1/2}\|g\|_2, \tag{1.7.8}$$

$$\|h+g\|_p \ge K^{-2}\|h\|_p - k_p(m,K)\|\{\hat g(k)\}\|_{\ell_\infty}. \tag{1.7.9}$$

Proof. Let $h \in \mathcal{T}(\Lambda)$, $|\Lambda| = m$. Take $X = \mathcal{T}(\Lambda)$ and consider the operator I_X provided by the UAP. Let $\psi_1,\dots,\psi_M$ form an orthonormal basis for the range Y of the operator I_X. Then $M \le k_p(m,K)$. Let

$$I_X(\mathrm{e}^{\mathrm{i}(k,x)}) = \sum_{j=1}^M c_j^k \psi_j.$$

Then the property $\|I_X\|_{L_p\to L_p} \le K$ implies

$$\left(\sum_{j=1}^M |c_j^k|^2\right)^{1/2} = \|I_X(\mathrm{e}^{\mathrm{i}(k,x)})\|_2 \le \|I_X(\mathrm{e}^{\mathrm{i}(k,x)})\|_p \le K.$$

Consider along with the operator I_X, the new operator,

$$A := (2\pi)^{-d}\int_{\mathbb{T}^d} T_t I_X T_{-t}\,\mathrm{d}t,$$

where T_t is the shift operator: $T_t(f) = f(\cdot + t)$. Then

$$A(\mathrm{e}^{\mathrm{i}(k,x)}) = \sum_{j=1}^M c_j^k (2\pi)^{-d}\int_{\mathbb{T}^d} \mathrm{e}^{-\mathrm{i}(k,t)}\psi_j(x+t)\,\mathrm{d}t = \left(\sum_{j=1}^M c_j^k \hat\psi_j(k)\right)\mathrm{e}^{\mathrm{i}(k,x)}.$$

Let

$$\lambda_k := \sum_{j=1}^{M} c_j^k \hat{\psi}_j(k).$$

We have

$$\sum_k |\lambda_k|^2 \le \sum_k \left(\sum_{j=1}^{M} |c_j^k|^2\right)\left(\sum_{j=1}^{M} |\hat{\psi}(k)|^2\right) \le K^2 M.$$

Also, $\lambda_k = 1$ for $k \in \Lambda$. For the operator A we have

$$\|A\|_{L_p \to L_p} \le K \quad \text{and} \quad \|A\|_{L_2 \to L_\infty} \le K M^{1/2}.$$

Therefore

$$\|A(h+g)\|_p \le K\|h+g\|_p$$

and

$$\|A(h+g)\|_p \ge \|h\|_p - K M^{1/2}\|g\|_2.$$

This proves the first inequality.

Consider the operator $B := A^2$. Then

$$B(h) = h, \quad h \in \mathcal{T}(\Lambda), \quad \|B\|_{L_p \to L_p} \le K^2,$$

and

$$\|B(f)\|_\infty \le K^2 M \|\{\hat{f}(k)\}\|_{\ell_\infty}.$$

Now, on the one hand

$$\|B(h+g)\|_p \le K^2\|h+g\|_p,$$

and on the other hand

$$\|B(h+g)\|_p = \|h + B(g)\|_p \ge \|h\|_p - K^2 M \|\{\hat{g}(k)\}\|_{\ell_\infty}.$$

This proves inequality (1.7.9). □

Theorem 1.7.17. For any $h \in \Sigma_m(\mathcal{T})$ and any $g \in L_\infty$ we have

$$\|h+g\|_\infty \ge K^{-1}\|h\|_\infty - \mathrm{e}^{C(K)m/2}\|g\|_2,$$
$$\|h+g\|_\infty \ge K^{-2}\|h\|_\infty - \mathrm{e}^{C(K)m}\|\{\hat{g}(k)\}\|_{\ell_\infty}.$$

Proof. This theorem is a direct corollary of Lemma 1.7.16 and the known estimate

$$k_\infty(m, K) \le \mathrm{e}^{C(K)m}$$

(see Figiel, Johnson and Schechtman (1988)). □

As we have already mentioned, $k_p(m, K)$ increases faster than any polynomial. In Konyagin and Temlyakov (2005) we improved inequality (1.7.8) by using other arguments.

Lemma 1.7.18. Let $2 \le p \le \infty$. For any $h \in \Sigma_m(\mathcal{T})$ and any $g \in L_p$, we have

$$\|h+g\|_p^p \ge 2^{-p-1}\|h\|_p^p - 2m^{p/2}\|h\|_p^{p-2}\|g\|_2^2. \tag{1.7.10}$$

We mention two inequalities from Konyagin and Temlyakov (2003*b*) in the style of the inequalities in Lemmas 1.7.14–1.7.18.

Lemma 1.7.19. Let $2 \le p < \infty$ and $h \in L_p$, $\|h\|_p \ne 0$. Then, for any $g \in L_p$ we have

$$\|h\|_p \le \|h+g\|_p + (\|h\|_{2p-2}/\|h\|_p)^{p-1}\|g\|_2.$$

Lemma 1.7.20. Let $h \in \Sigma_m(\mathcal{T})$, $\|h\|_\infty = 1$. Then, for any function g such that $\|g\|_2 \le \frac{1}{4}(4\pi m)^{-m/2}$, we have

$$\|h+g\|_\infty \ge 1/4.$$

We proceed to estimate $v(m,p,\gamma)$ and $w(m,p,\gamma)$ for $p \in [2,\infty)$. In the special case of even p, we have by (1.7.4) and (1.7.6) that

$$v(m,p,1) \le m^{p-1}, \quad w(m,p,1) \le m^{p-1}.$$

The following bound was proved in Konyagin and Temlyakov (2005).

Lemma 1.7.21. Let $2 \le p < \infty$, and let $\alpha := p/2 - [p/2]$. Then we have

$$v(m,p,\gamma) \le m^{c(\alpha,\gamma)m^{1/2}+p-1}.$$

1.8. Greedy-type bases; direct and inverse theorems

Theorem 1.3.1 points out the importance of bases L_p-equivalent to the Haar basis. We will now discuss necessary and sufficient conditions for f to have a prescribed decay of $\{\sigma_m(f,\Psi)\}$ under the assumption that Ψ is L_p-equivalent to the Haar basis $\mathcal{H}$, $1 < p < \infty$. We will express these conditions in terms of coefficients $\{c_n(f)\}$ of the expansion

$$f = \sum_{n=1}^{\infty} c_n(f)\psi_n.$$

The direct theorems of approximation theory provide bounds of approximation error (in our case $\sigma_m(f,\Psi)$) in terms of smoothness properties of a function f. These theorems are also known under the name of Jackson-type inequalities. The inverse theorems of approximation theory (also known as Bernstein-type inequalities) provide some smoothness properties of a function f from the sequence of approximation errors (in our case $\{\sigma_m(f,\Psi)\}$). It is well understood in approximation theory (see Petrushev (1988), DeVore and Lorenz (1993) and DeVore (1998)) how the Jackson-type and Bernstein-type inequalities can be used in order to characterize the corresponding approximation spaces. In the case of our interest, when we study best m-term approximation with regard to bases that are L_p-equivalent to the Haar basis, the theory of Jackson and Bernstein inequalities has been developed in Cohen *et al.* (2000). It was used in Cohen *et al.* (2000) for a description of approximation spaces defined in terms of $\{\sigma_m(f, \Psi)\}$. We want to point out that in the special case of bases that are L_p-equivalent to the Haar basis (and also for some more general bases) there exists a simple direct way to describe the approximation spaces defined in terms of $\{\sigma_m(f, \Psi)\}$ (Temlyakov 1998*b*, Kamont and Temlyakov 2004, Kerkyacharian and Picard 2004). We present results from Temlyakov (1998*b*) here. The following lemma from Temlyakov (1998*a*) (see Lemmas 3.1 and 3.2) plays the key role in this consideration.

Lemma 1.8.1. Let a basis Ψ be L_p-equivalent to $\mathcal{H}_p$, $1 < p < \infty$. Then, for any finite Λ and $a \le |c_n| \le b$, $n \in \Lambda$, we have

$$C_1(p, \Psi)a(|\Lambda|)^{1/p} \le \left\| \sum_{n\in\Lambda} c_n\psi_n \right\|_p \le C_2(p, \Psi)b(|\Lambda|)^{1/p}. \tag{1.8.1}$$

We note that the results that follow use only the assumption that Ψ is a greedy basis satisfying (1.8.1). We formulate a general statement and then consider several important particular examples of the rate of decrease of $\{\sigma_m(f, \Psi)_p\}$. We begin by introducing some notation. For a sequence $\mathcal{E} = \{\epsilon_k\}_{k=0}^{\infty}$ of positive numbers monotonically decreasing to zero (we write $\mathcal{E} \in$ MDP), we define inductively a sequence $\{N_s\}_{s=0}^{\infty}$ of non-negative integers:

$$N_0 = 0, \quad \text{and } N_s \text{ is the smallest integer satisfying} \tag{1.8.2}$$
$$\epsilon_{N_s} < 2^{-s}, \qquad d_s := \max(N_{s+1} - N_s, 1).$$

We are going to consider the following examples of sequences.

Example 1.8.1. Take $\epsilon_0 = 1$ and $\epsilon_k = k^{-r}$, $r > 0$, $k = 1, 2, \ldots$. Then

$$N_s \asymp 2^{s/r} \quad \text{and} \quad d_s \asymp 2^{s/r}.$$

Example 1.8.2. Fix $0 < b < 1$ and take $\epsilon_k = 2^{-k^b}$, $k = 0, 1, 2, \ldots$. Then

$$N_s = s^{1/b} + O(1) \quad \text{and} \quad d_s \asymp s^{1/b-1}.$$

Let $f \in L_p$. Rearrange the sequence $\|c_n(f)\psi_n\|_p$ in decreasing order,

$$\|c_{n_1}(f)\psi_{n_1}\|_p \ge \|c_{n_2}(f)\psi_{n_2}\|_p \ge \cdots,$$

and define

$$a_k(f, p) := \|c_{n_k}(f)\psi_{n_k}\|_p.$$

We now give some inequalities for $a_k(f, p)$ and $\sigma_m(f, \Psi)_p$. We will use the abbreviations $\sigma_m(f)_p := \sigma_m(f, \Psi)_p$ and $\sigma_0(f)_p := \|f\|_p$.

Lemma 1.8.2. For any two positive integers $N < M$ we have

$$a_M(f,p) \le C(p,\Psi)\sigma_N(f)_p(M-N)^{-1/p}.$$

Proof. By Theorem 1.3.1 we have, for all m,

$$\|f - G^p_m(f,\Psi)\|_p \le C(p,\Psi)\sigma_m(f)_p.$$

Hence, and by definition of G^p_m, we get

$$J := \Big\|\sum_{k=N+1}^{M} c_{n_k}(f)\psi_{n_k}\Big\|_p \le C(p,\Psi)(\sigma_N(f)_p + \sigma_M(f)_p). \tag{1.8.3}$$

Next, we have for $k \in (N, M]$,

$$\|c_{n_k}(f)\psi_{n_k}\|_p \ge \|c_{n_M}(f)\psi_{n_M}\|_p = a_M(f,p),$$

and by Lemma 1.8.1 we obtain

$$a_M(f,p)(M-N)^{1/p} \le C(p,\Psi)J. \tag{1.8.4}$$

Relations (1.8.3) and (1.8.4) imply the conclusion of Lemma 1.8.2. □

Lemma 1.8.3. For any sequence $m_0 < m_1 < m_2 < \cdots$ of non-negative integers we have

$$\sigma_{m_s}(f)_p \le C(p,\Psi)\sum_{l=s}^{\infty} a_{m_l}(f,p)(m_{l+1}-m_l)^{1/p}.$$

Proof. We have

$$\sigma_{m_s}(f)_p \le \Big\|\sum_{k>m_s} c_{n_k}(f)\psi_{n_k}\Big\|_p \le \sum_{l=s}^{\infty}\Big\|\sum_{k\in(m_l,m_{l+1}]} c_{n_k}(f)\psi_{n_k}\Big\|_p.$$

Hence, using Lemma 1.8.1,

$$\sigma_{m_s}(f)_p \le C(p,\Psi)\sum_{l=s}^{\infty} a_{m_l}(f,p)(m_{l+1}-m_l)^{1/p},$$

as required. □

Theorem 1.8.4. Assume a given sequence $\mathcal{E} \in$ MDP satisfies the conditions

$$\epsilon_{N_s} \ge C_1 2^{-s}, \qquad d_{s+1} \le C_2 d_s, \qquad s = 0,1,2,\dots.$$

Then we have the equivalence

$$\sigma_n(f)_p \ll \epsilon_n \iff a_{N_s}(f,p) \ll 2^{-s}d_s^{-1/p}.$$

Proof. We first prove ⇒. If $N_{s+1} > N_s$, then we use Lemma 1.8.2 with $M = N_{s+1}$ and $N = N_s$,

$$a_{N_{s+1}}(f,p) \le C(p,\Psi)\sigma_{N_s}(f)_p d_s^{-1/p} \le C(p,\Psi)2^{-s-1}(d_{s+1}/C_2)^{-1/p},$$

which implies the statement of Theorem 1.8.4 in this case. Let $N_{s+1} = N_s = \cdots = N_{s-j} > N_{s-j-1}$. The assumption $\epsilon_{N_s} \geq C_1 2^{-s}$ combined with the definition of N_s: $\epsilon_{N_s} < 2^{-s}$ imply that $j \leq C_3$. Then, from the above case we get

$$a_{N_{s-j}}(f,p) \ll 2^{-s+j}(d_{s-j})^{-1/p},$$

and therefore

$$a_{N_{s+1}}(f,p) \ll 2^{-s-1}(d_{s+1})^{-1/p}.$$

The implication ⇒ has been proved.

We now prove the inverse statement ⇐. Using Lemma 1.8.3, we get

$$\sigma_{N_s}(f)_p \ll \sum_{l=s}^{\infty} a_{N_l}(f,p)(N_{l+1}-N_l)^{1/p} \ll \sum_{l=s}^{\infty} 2^{-l} \ll 2^{-s} \ll \epsilon_{N_s},$$

and for $n \in [N_s, N_{s+1})$

$$\sigma_n(f)_p \leq \sigma_{N_s}(f)_p \ll \epsilon_{N_s}(f)_p \ll 2^{-s} \ll \epsilon_{N_{s+1}}(f)_p \leq \epsilon_n(f)_p. \qquad \square$$

Corollary 1.8.5. Theorem 1.8.4 applied to Examples 1.8.1 and 1.8.2 gives the following relations:

$$\sigma_m(f)_p \ll (m+1)^{-r} \iff a_n(f,p) \ll n^{-r-1/p}, \tag{1.8.5}$$

$$\sigma_m(f)_p \ll 2^{-m^b} \iff a_n(f,p) \ll 2^{-n^b} n^{(1-1/b)/p}. \tag{1.8.6}$$

Remark 1.8.6. Making use of Lemmas 1.8.2 and 1.8.3 we can prove a version of Corollary 1.8.5 with the sign $\ll$ replaced by $\asymp$.

Theorem 1.8.4 and Corollary 1.8.5 are in the spirit of the classical Jackson–Bernstein direct and inverse theorems in linear approximation theory, where conditions of the form

$$E_n(f)_p \ll \epsilon_n, \quad \text{or} \quad \|E_n(f)_p/\epsilon_n\|_{l_\infty} < \infty \tag{1.8.7}$$

are imposed on the corresponding sequences of approximating characteristics. It is well known (see DeVore (1998)) that, in studying many questions of approximation theory, it is convenient to consider, along with the restriction (1.8.7), its following generalization:

$$\|E_n(f)_p/\epsilon_n\|_{l_q} < \infty. \tag{1.8.8}$$

Lemmas 1.8.2 and 1.8.3 are also useful in handling this more general case. For instance, in the particular case of Example 1.8.1 we get the following statement.

Theorem 1.8.7. Let $1 < p < \infty$ and $0 < q < \infty$. Then, for any positive r we have the equivalence relation

$$\sum_m \sigma_m(f)_p^q m^{rq-1} < \infty \iff \sum_n a_n(f,p)^q n^{rq-1+q/p} < \infty.$$

Remark 1.8.8. The condition

$$\sum_n a_n(f,p)^q n^{rq-1+q/p} < \infty$$

with $q = \beta := (r+1/p)^{-1}$ takes a very simple form:

$$\sum_n a_n(f,p)^\beta = \sum_n \|c_n(f)\psi_n\|_p^\beta < \infty. \tag{1.8.9}$$

In the case $\Psi = \mathcal{H}_p$, condition (1.8.9) is equivalent to f being in Besov space $B^r_\beta(L_\beta)$.

Corollary 1.8.9. Theorem 1.8.7 implies the following relation:

$$\sum_m \sigma_m(f,\mathcal{H})_p^\beta m^{r\beta-1} < \infty \iff f \in B^r_\beta(L_\beta),$$

where $\beta := (r+1/p)^{-1}$.

The statement similar to Corollary 1.8.9 for free-knot spline approximation was proved in Petrushev (1988). Corollary 1.8.9 and further results in this direction can be found in DeVore and Popov (1988) and DeVore, Jawerth and Popov (1992). We want to remark here that conditions in terms of $a_n(f,p)$ are convenient in applications. For instance, relation (1.8.5) can be rewritten using the idea of thresholding. For a given $f \in L_p$ denote

$$T(\epsilon) := \#\{a_k(f,p) : a_k(f,p) \geq \epsilon\}.$$

Then (1.8.5) is equivalent to

$$\sigma_m(f)_p \ll (m+1)^{-r} \iff T(\epsilon) \ll \epsilon^{-(r+1/p)^{-1}}.$$

For further results in this direction see DeVore (1998), Cohen *et al.* (2000) and Oswald (2001).

The above direct and inverse Theorem 1.8.7 that holds for greedy bases satisfying (1.8.1) was extended in Kerkyacharian and Picard (2004) to the case of quasi-greedy bases satisfying (1.8.1). Kerkyacharian and Picard (2004) say that a basis Ψ of a Banach space X has the p-Temlyakov property if there exists $0 < C < \infty$ such that, for any finite set of indices Λ, we have

$$C^{-1}\Big(\min_{n\in\Lambda}|c_n|\Big)|\Lambda|^{1/p} \leq \Big\|\sum_{n\in\Lambda} c_n\psi_n\Big\|_X \leq C\Big(\max_{n\in\Lambda}|c_n|\Big)|\Lambda|^{1/p}. \tag{1.8.10}$$

Now let

$$f = \sum_{k=1}^\infty c_k(f)\psi_k$$

and

$$|c_{k_1}| \geq |c_{k_2}| \geq \cdots$$

be a decreasing reordering of the coefficients. The following result is from Kerkyacharian and Picard (2004).

Theorem 1.8.10. Let Ψ be a quasi-greedy basis.

(1) If Ψ has the p-Temlyakov property (1.8.10), then for any $0 < r < \infty$, $0 < q < \infty$ we have

$$\sum_m \sigma_m(f,\Psi)_X^q m^{rq-1} < \infty \iff \sum_n |c_{k_n}(f)|^q n^{rq-1+q/p} < \infty. \tag{1.8.11}$$

(2) If (1.8.11) holds with some $r > 0$, then Ψ has the p-Temlyakov property (1.8.10).

We note that property (1.8.10) implies that Ψ is democratic. Therefore, by Theorem 1.4.10 a quasi-greedy basis satisfying (1.8.10) is an almost greedy basis. The basis $\mathcal{H}_p^d$ is not a democratic basis for L_p, $p \neq 2$, $d > 1$. So, we cannot apply the above results in this case. Some direct and inverse theorems for $\mathcal{H}_p^d$ are obtained in Kamont and Temlyakov (2004).

1.9. Some further results

We begin our discussion with the case of $X = L_p$, $p = 1$ or $p = \infty$ and $\Psi = \mathcal{H}_p^d$. It turns out that efficiency of greedy algorithms $G_m(\cdot, \mathcal{H}_p^d)$, $p = 1, \infty$, drops down dramatically compared with the case $1 < p < \infty$. We formulate a result from Temlyakov (1998*b*).

Theorem 1.9.1. Let $p = 1$ or $p = \infty$. Then we have for each $f \in L_p$

$$\|f - G_m(f, \mathcal{H}_p^d)\|_p \leq (3m+1)\sigma_m(f, \mathcal{H}^d)_p.$$

The extra factor $(3m + 1)$ cannot be replaced by a factor $c(m)$ such that $c(m)/m \to 0$ as $m \to \infty$.

This particular result indicates that there are problems with greedy approximation in L_1 and in $\mathcal{C}$ with regard to the Haar basis. We note that, as is proved in Oswald (2001), the extra factor $3m+1$ is the best-possible extra factor in Theorem 1.9.1. The greedy-type bases have nice properties and they are important in nonlinear m-term approximation. Therefore, one of the new directions of research in functional analysis and in approximation theory is to understand which Banach spaces may have such bases. Another direction is to understand in which Banach spaces some classical bases are of greedy type. Some results in this direction can be derived immediately from known results on Banach spaces that have unconditional bases, and from the characterization Theorem 1.3.5. For instance, it is well known that the spaces L_1 and $\mathcal{C}$ do not have unconditional bases. Therefore, Theorem 1.3.5 implies that there is no greedy basis in L_1 and in $\mathcal{C}$.

It was proved in Dilworth, Kutzarova and Wojtaszczyk (2002) that the Haar basis $\mathcal{H}_1$ is not a quasi-greedy basis for L_1. We saw in Section 1.6 that the use of the Weak Greedy Algorithm has some advantages over the Greedy Algorithm. Theorem 1.6.2 states that the convergence set $\mathrm{WT}\{e_n\}$ of the WGA is linear for any $t \in (0,1)$, while the convergence set may not be linear for the Greedy Algorithm. Recently, Gogyan (2006) proved that, for any $t \in (0,1)$ and for any $f \in L_1(0,1)$, there exists a realization of the WGA with respect to the Haar basis that converges to f in L_1.

It was proved in Dilworth *et al.* (2002) that there exists an increasing sequence of integers $\{n_j\}$ such that the lacunary Haar system $\{H^1_{2^{n_j}+l}; l = 1,\dots,2^{n_j}, j = 1,2,\dots\}$ is a quasi-greedy basis for its linear span in L_1. Gogyan (2005) proved that the above property holds if either $\{n_j\}$ is a sequence of all even numbers or $\{n_j\}$ is a sequence of all odd numbers. We also note that the space $L_1(0,1)$ has a quasi-greedy basis (Dilworth, Kalton and Kutzarova 2003). The reader can find further results on existence (and non-existence) of quasi-greedy and almost greedy bases in Dilworth *et al.* (2003). In particular, it is proved in Dilworth *et al.* (2003) that $\mathcal{C}[0,1]$ does not have quasi-greedy bases.

We pointed out in Section 1.7 that the trigonometric system is not a quasi-greedy basis for L_p, $p \neq 2$. The question of when (and for which weights w) the trigonometric system forms a quasi-greedy basis for a weighted space $L_p(w)$ was studied in Nielsen (2006). The author proved that this can happen only for $p = 2$ and, whenever the system forms a quasi-greedy basis, the basis must be a Riesz basis.

Theorem 1.3.1A shows that, in the case when a basis Ψ is L_p-equivalent to the Haar basis $\mathcal{H}_p$, $1 < p < \infty$, the Greedy Algorithm $G_m(f,\Psi)$ provides near-best approximation for each individual function $f \in L_p$. For a function class $F \subset X$, let

$$\sigma_m(F,\Psi)_X := \sup_{f\in F} \sigma_m(F,\Psi)_X,$$

$$G_m(F,\Psi)_X := \sup_{f\in F} \|f - G_m(f,\Psi)\|_X.$$

Obviously, if $G_m(\cdot,\Psi)$ provides near-best approximation for each individual function, then it provides near-best approximation for each function class F:

$$G_m(F,\Psi)_X \le C\sigma_m(F,\Psi)_X.$$

In Section 1.7 we pointed out that the trigonometric system is not a quasi-greedy basis for L_p, $p \neq 2$ (see (1.7.3)). Thus, the trigonometric system is not a greedy basis for L_p, $p \neq 2$, and for some functions $f \in L_p$, $p \neq 2$, $G_m(f,\mathcal{T})$, does not provide near-best approximation. However, it was proved in Temlyakov (1998*c*) that in many cases the algorithm $G_m(\cdot,\mathcal{T})$ is optimal for a given class of functions. The reader can find further results on $\sigma_m(F,\mathcal{T}^d)_p$ and $G_m(F,\mathcal{T}^d)_p$ for different classes F in DeVore and Temlyakov (1995) and Temlyakov (1998*c*, 2000*a*, 2002*a*).

Consideration of approximation in a function class leads to a concept of the *optimal* (*best*) *basis* for a given class. The first results for best basis approximation were given by Kashin (1985), who showed that, for any orthonormal basis Ψ and any $0<\alpha\le 1$, we have

$$\sigma_m(\operatorname{Lip}\alpha,\Psi)_{L_2}\ge cm^{-\alpha}, \tag{1.9.1}$$

where the constant c depends only on α. It follows from this that any of the standard wavelet or Fourier bases are best for the Lipschitz classes, when the approximation is carried out in L_2 and the competition is held over all orthonormal bases. The estimate (1.9.1) rests on some fundamental estimates for the best basis approximation of finite-dimensional hypercubes using orthonormal bases.

The problem of best basis selection was studied in Coifman and Wickerhauser (1992). Donoho (1993, 1997) also studied the problem of best bases for a function class F. He calls a basis Ψ from a collection $\mathbb{B}$ best for F if

$$\sigma_m(F,\Psi)_X=O(m^{-\alpha}),\quad m\to\infty,$$

and no other basis Ψ' from $\mathbb{B}$ satisfies

$$\sigma_n(F,\Psi')_X=O(n^{-\beta}),\quad n\to\infty,$$

for a value of $\beta>\alpha$. Donoho has shown that in some cases it is possible to determine a best basis (in the above sense) for the class F by intrinsic properties of how the class gets represented with respect to the basis. In Donoho's analysis (as was the case for Kashin as well) the space X is L_2 (or equivalently any Hilbert space), and the competition for a best basis takes place over all complete orthonormal systems (*i.e.*, $\mathbb{B}$ consists of all complete orthonormal bases for L_2).

In DeVore, Petrova and Temlyakov (2003) we continued to study the problem of optimal basis selection with regard to natural collections of bases. We worked on the following problem in this direction. We say that a function class F is aligned to the basis Ψ if, whenever $f=\sum a_k\psi_k$ is in F, then

$$\sum a_k'\psi_k\in F\quad\text{for any}\quad |a_k'|\le c|a_k|,\quad k=1,2,\dots,$$

where $c>0$ is a fixed constant. We pointed out in DeVore *et al.* (2003) that the results from Kashin (1985) and Donoho (1993) imply the following result.

Theorem 1.9.2. Let Φ be an orthonormal basis for a Hilbert space H and let F be a function class aligned with Φ such that, for some $\alpha>0$, $\beta\in\mathbb{R}$, we have

$$\limsup_{m\to\infty} m^{\alpha}(\log m)^{\beta}\sigma_m(F,\Phi)>0.$$

Then, for any orthonormal basis B we have

$$\limsup_{m\to\infty} m^\alpha(\log m)^\beta \sigma_m(F,B) > 0.$$

We have obtained in DeVore *et al.* (2003) a generalization of this important result in the following direction. We replaced the Hilbert space with the Banach space and also widened the search for optimal basis selection from the collection of orthonormal bases to the collection of unconditional bases. Here is the corresponding theorem from DeVore *et al.* (2003).

Theorem 1.9.3. Let Ψ be a normalized unconditional basis for X with the property

$$\Big\|\sum_{j\in A}\psi_j\Big\|_X \asymp (\#A)^\mu,$$

for some $\mu > 0$. Assume that the function class F is aligned with Ψ, and for some $\alpha > 0$, $\beta \in \mathbb{R}$ we have

$$\limsup_{m\to\infty} m^\alpha(\log m)^\beta \sigma_m(F,\Psi) > 0.$$

Then, for any unconditional basis B we have

$$\limsup_{m\to\infty} m^\alpha(\log m)^{\alpha+\beta} \sigma_m(F,B) > 0. \tag{1.9.2}$$

Theorem 1.9.3 is weaker than Theorem 1.9.2 in the sense that we have an extra factor $(\log m)^\alpha$ in (1.9.2). Recently, Bednorz (2006) proved Theorem 1.9.3 with (1.9.2) replaced by (1.9.3):

$$\limsup_{m\to\infty} m^\alpha(\log m)^\beta \sigma_m(F,B) > 0. \tag{1.9.3}$$

The following nonlinear analogues of the Kolmogorov widths and the ortho-widths (see, for instance, Temlyakov (1989*a*)) were considered in Temlyakov (2000*a*, 2002*a*, 2003*a*). Let a function class F and a Banach space X be given. Assume that, on the basis of some additional information, we know that our basis for m-term approximation should satisfy some structural properties, for instance, it has to be orthogonal. Let $\mathbb{B}$ be a collection of bases satisfying a given property.

I Define an analogue of the Kolmogorov width

$$\sigma_m(F,\mathbb{B})_X := \inf_{\Psi\in\mathbb{B}}\sup_{f\in F}\sigma_m(f,\Psi)_X.$$

II Define an analogue of the orthowidth

$$\gamma_m(F,\mathbb{B})_X := \inf_{\Psi\in\mathbb{B}}\sup_{f\in F}\|f - G_m(f,\Psi)\|_X.$$

In the papers cited above some results were obtained when $\mathbb{B} = \mathbb{O}$, the set of orthonormal bases, and F is either a multivariate smoothness class of an anisotropic Sobolev–Nikol'skii kind, or a class of functions with bounded mixed derivatives.

We conclude this section with a very recent result from Wojtaszczyk (2006). Theorem 1.3.1 says that the univariate Haar basis $\mathcal{H}$ is a greedy basis for $L_p := L_p([0,1])$, $1 < p < \infty$. The spaces L_p are examples of rearrangement-invariant spaces. Let us recall that a rearrangement-invariant space of functions defined on $[0,1]$ is a Banach space X with norm $\|\cdot\|$ whose elements are measurable (in the sense of Lebesgue) functions defined on $[0,1]$ satisfying the following conditions.

(1) If $f \in X$ and g is a measurable function such that $|g(x)| \le |f(x)|$ almost everywhere, then $g \in X$ and $\|g\| \le \|f\|$.

(2) If $f \in X$ and g has the same distribution as f, *i.e.*, for all λ,

$$\text{measure}(\{x \in [0,1] : f(x) \le \lambda\}) = \text{measure}(\{x \in [0,1] : g(x) \le \lambda\}),$$

then $g \in X$ and $\|g\| = \|f\|$.

The following result was proved in Wojtaszczyk (2006).

Theorem 1.9.4. Let X be a rearrangement-invariant space on $[0,1]$. If the Haar system normalized in X is a greedy basis for X, then $X = L_p([0,1])$ with some $1 < p < \infty$.

It is a very interesting result that singles out the L_p-spaces with $1 < p < \infty$ from the collection of rearrangement-invariant spaces. Theorem 1.9.4 emphasizes the importance of the L_p-spaces in the theory of greedy approximation.

1.10. Systems L_p-equivalent to $\mathcal{H}$

In the previous sections of this chapter we have presented elements of a general theory of greedy-type bases. In this section we concentrate on construction of greedy bases and related bases that are useful in approximation of functions in the L_p-norm. Theorem 1.3.1 indicates importance of bases that are L_p-equivalent to the Haar basis $\mathcal{H}$. It says that such bases are greedy bases for L_p, $1 < p < \infty$. Theorem 1.3.1 addresses the case of $L_p([0,1])$. The same proof works for the $L_p(\mathbb{R})$. In this section we will give some sufficient conditions on a system of functions in order to be L_p-equivalent to the Haar basis. It is more convenient to give these conditions in the case of $L_p(\mathbb{R})$. These results are a part of general Littlewood–Paley theory. We begin in this section by introducing various forms of Littlewood–Paley theory for systems of functions. From the univariate wavelet ψ, we can construct efficient bases for $L_2(\mathbb{R})$ and other function spaces by dilation and shifts (see, for instance, DeVore (1998)). For example, the functions

$$\psi_{j,k} := 2^{k/2}\psi(2^k \cdot -j), \quad j,k \in \mathbb{Z},$$

form a stable basis (orthogonal basis in the case of an orthogonal wavelet ψ) for $L_2(\mathbb{R})$.

It is convenient to use a different indexing for the functions $\psi_{j,k}$. Let $D := D(\mathbb{R})$ denote the set of dyadic intervals. Each such interval I is of the form $I = [j2^{-k}, (j+1)2^{-k}]$. We define

$$\psi_I := \psi_{j,k}, \quad I = [j2^{-k}, (j+1)2^{-k}]. \tag{1.10.1}$$

Thus the basis $\{\psi_{j,k}\}_{j,k\in\mathbb{Z}}$ is the same as $\{\psi_I\}_{I\in D(\mathbb{R})}$.

We consider in this section systems of functions $\{\eta(I,\cdot)\}_{I\in D}$ defined on $\mathbb{R}$. If $1 < p < \infty$, we say that a family of real-valued functions $\eta(I,\cdot)$, $I \in D$, satisfies the *strong Littlewood–Paley property* for p if, for any finite sequence (c_I) of real numbers, we have

$$\Big\|\sum_{I\in D} c_I\eta(I,\cdot)\Big\|_p \asymp \Big\|\Big(\sum_{I\in D}[c_I\eta(I,\cdot)]^2\Big)^{1/2}\Big\|_p \tag{1.10.2}$$

with constants of equivalency depending at most on p. Here and later we use the notation $A \asymp B$ to mean that there are two constants $C_1, C_2 > 0$ such that

$$C_1A \le B \le C_2A.$$

We shall indicate what the constants depend on (in the case of (1.10.2) they may depend on p).

Here is a useful remark concerning (1.10.2). From the validity of (1.10.2) for finite sequences, we can deduce its validity for infinite sequences by a limiting argument. For example, if $(c_I)_{I\in D}$ is an infinite sequence for which the sum on the left-hand side of (1.10.2) converges in $L_p(\mathbb{R})$ with respect to some ordering of the $I \in D$, then the right-hand side of (1.10.2) will converge with respect to the same ordering and the right-hand side of (1.10.2) will be less than a multiple of the left. Likewise, we can reverse the roles of the left- and right-hand sides. Similar remarks hold for other statements such as (1.10.2).

We use *strong Littlewood–Paley inequality* to differentiate (1.10.2) from other possible forms of Littlewood–Paley inequalities. For example, the Littlewood–Paley inequalities for the complex exponentials take a different form (see Zygmund (1959, Chapter XV)). Another point of interest in our considerations is the following:

$$\Big\|\sum_{I\in D} c_I\eta(I,\cdot)\Big\|_p \asymp \Big\|\Big(\sum_{I\in D}[c_I\chi_I]^2\Big)^{1/2}\Big\|_p. \tag{1.10.3}$$

We use the notation χ for the characteristic function of $[0,1]$ and χ_I for its $L_2(\mathbb{R})$-normalized, shifted dilates given by (1.10.1) (with $\psi = \chi$).

The two forms (1.10.2) and (1.10.3) are equivalent under very mild conditions on the functions $\eta(I,\cdot)$. To see this, we shall use the Hardy–Littlewood maximal operator, which is defined for a locally integrable function g on $\mathbb{R}$ by

$$Mg(x) := \sup_{J \ni x} \frac{1}{|J|} \int_J |g(y)| \,\mathrm{d}y$$

with the supremum taken over all intervals J that contain x. It is well known that M is a bounded operator on $L_p(\mathbb{R})$ for all $1 < p \leq \infty$. The Fefferman–Stein inequality (Fefferman and Stein 1972) bounds the mapping M on sequences of functions. We shall only need the following special case of this inequality, which says that for any functions $\eta(I,\cdot)$ and constants c_I, $I \in D$, we have for $1 < p \leq \infty$,

$$\left\| \left(\sum_{I \in D} (c_I M\eta(I,\cdot))^2 \right)^{1/2} \right\|_p \leq A \left\| \left(\sum_{I \in D} (c_I \eta(I,\cdot))^2 \right)^{1/2} \right\|_p, \tag{1.10.4}$$

with an absolute constant A.

Consider now as an example the equivalence of (1.10.2). If the functions $\eta(I,\cdot)$, $I \in D$, satisfy

$$|\eta(I,x)| \leq CM\chi_I(x), \quad \chi_I(x) \leq CM\eta(I,x), \quad \text{for almost all } x \in \mathbb{R}, \tag{1.10.5}$$

then, using (1.10.4), we see that (1.10.2) holds if and only if (1.10.3) holds. The first inequality in (1.10.5) is a decay condition on $\eta(I,\cdot)$. For example, if $\eta(I,\cdot)$ is given by the normalized, shifted dilates of the function ψ, $\eta(I,\cdot) = \psi_I$, then the first inequality in (1.10.5) holds whenever

$$|\psi(x)| \leq C[\max(1,|x|)]^{-\lambda}, \quad \text{for almost all } x \in \mathbb{R},$$

with $\lambda \geq 1$. The second condition in (1.10.5) is extremely mild. For example, it is always satisfied when the family $\eta(I,\cdot)$ is generated by the shifted dilates of a non-zero function ψ.

Suppose that we have in hand two families $\eta(I,\cdot), \mu(I,\cdot)$, $I \in D(\mathbb{R})$. We shall use the notation $\{\eta(I,\cdot)\}_{I\in D} \prec \{\mu(I,\cdot)\}_{I\in D}$ if there is a constant $C > 0$ such that

$$\left\| \sum_{I \in D} c_I \eta(I,\cdot) \right\|_p \leq C \left\| \sum_{I \in D} c_I \mu(I,\cdot) \right\|_p \tag{1.10.6}$$

holds for all finite sequences $(c_I)_{I\in D}$ with C independent of the sequence. If $\{\eta(I,\cdot)\}_{I\in D} \prec \{\mu(I,\cdot)\}_{I\in D}$ and $\{\mu(I,\cdot)\}_{I\in D} \prec \{\eta(I,\cdot)\}_{I\in D}$, then we write $\{\eta(I,\cdot)\}_{I\in D} \approx \{\mu(I,\cdot)\}_{I\in D}$ and say that these systems are L_p-equivalent.

Given two families $\eta(I,\cdot),\mu(I,\cdot)$, $I\in D(\mathbb{R})$, we define the operator T which maps $\mu(I,\cdot)$ into $\eta(I,\cdot)$ for all $I\in D$, and we extend T to finite linear combinations of the $\mu(I,\cdot)$ by linearity. Then (1.10.6) holds if and only if T is a bounded operator with respect to the L_p-norm, and $\{\mu(I,\cdot)\}_{I\in D}\prec\{\eta(I,\cdot)\}_{I\in D}$ holds if and only if T has a bounded inverse with respect to the L_p-norm.

The strong Littlewood–Paley inequalities (1.10.3) are the same as the L_p-equivalence $\{\eta(I,\cdot)\}\approx\{H_I\}$. We begin with a presentation of sufficient conditions in order that $\{\eta(I,\cdot)\}\prec\{H_I\}$. Let ξ_I, $I\in D$, denote the centre of the dyadic interval I. We shall assume in this section that $\eta(I,\cdot)$, $I\in D$, is a family of univariate functions that satisfy the following assumptions.

A1 There is an $\epsilon>0$, and a constant C_1 such that, for all $t\in\mathbb{R}$ and all $J\in D$, we have

$$|\eta(J,\xi_J+t|J|)|\le C_1|J|^{-1/2}(1+|t|)^{-1-\epsilon}.$$

A2 There is an $\epsilon>0$ and a constant C_2 and a partition of $[-1/2,1/2]$ into intervals $J_1,\dots,J_m$ that are dyadic with respect to $[-1/2,1/2]$, such that, for any $J\in D$, any $j\in\mathbb{Z}$, and any t_1,t_2 in the interior of the same interval J_k, $k=1,\dots,m$, we have

$$\begin{aligned}|\eta(J,\xi_J+j|J|+t_1|J|)-\eta(J,\xi_J+j|J|+t_2|J|)|\\ \le C_2|J|^{-1/2}(1+|j|)^{-1-\epsilon}|t_2-t_1|^{\epsilon},\end{aligned}$$

A3 For any $J\in D$, we have

$$\int_{\mathbb{R}}\eta(J,x)\,\mathrm{d}x=0.$$

When $\eta(J,\cdot)=\psi_J$ for a function ψ, it is enough to check these assumptions for $J=[0,1]$, *i.e.*, for the function ψ alone. They follow for all other dyadic intervals J by dilation and translation.

Condition A1 is a standard decay assumption and A3 is the zero moment condition. Condition A2 requires that the functions $\eta(I,\cdot)$ be piecewise in Lip ϵ.

Let T be the linear operator which satisfies

$$T\Big(\sum_{I\in D}c_IH_I\Big)=\sum_{I\in D}c_I\eta(I,\cdot) \tag{1.10.7}$$

for each finite linear combination $\sum_{I\in D}c_IH_I$ of the H_I. We wish to show that

$$\Big\|T\Big(\sum_{I\in D}c_IH_I\Big)\Big\|_p\le C\Big\|\sum_{I\in D}c_IH_I\Big\|_p$$

for each such sum. From this it would follow that T extends (by continuity) to a bounded operator on all of $L_p(\mathbb{R})$ and therefore $\{\eta(I,\cdot)\}\prec\{H_I\}$.

We can expand $\eta(J,\cdot)$ into its Haar decomposition. Let

$$\lambda(I,J) := \int_{\mathbb{R}} \eta(J,x)H_I(x)\,\mathrm{d}x, \tag{1.10.8}$$

so that

$$\eta(J,\cdot) = \sum_{I\in D} \lambda(I,J)H_I.$$

It follows that

$$T\biggl(\sum_{J\in D} c_J H_J\biggr) = \sum_{I\in D}\sum_{J\in D} \lambda(I,J)c_J H_I. \tag{1.10.9}$$

Thus the mapping T is tied to the bi-infinite matrix $\Lambda := (\lambda(I,J))_{I,J\in D}$, which maps the sequence $c := (c_J)$ to the sequence

$$(c'_I) := \Lambda c.$$

One approach to proving Littlewood–Paley inequalities is to show that the matrix Λ decays sufficiently fast away from the diagonal (see Frazier and Jawerth (1990, Section 3)). Following Frazier and Jawerth (1990), we say that a matrix $A = (a(I,J))_{I,J\in D}$ is *almost diagonal* if, for some $\epsilon > 0$, we have

$$|a(I,J)| \le C\omega(I,J), \tag{1.10.10}$$

with

$$\omega(I,J) := \biggl(1 + \frac{|\xi_I - \xi_J|}{\max(|I|,|J|)}\biggr)^{-1-\epsilon} \biggl(\min\biggl(\frac{|I|}{|J|},\frac{|J|}{|I|}\biggr)\biggr)^{(1+\epsilon)/2}. \tag{1.10.11}$$

In DeVore *et al.* (1998) we used the following special case of a theorem of Frazier and Jawerth (1990, Theorem 3.3), concerning almost diagonal operators.

Theorem 1.10.1. If $(a(I,J))_{I,J\in\mathcal{D}}$ is an almost diagonal matrix, then the operator A defined by

$$A\biggl(\sum_{J\in\mathcal{D}} c_J H_J\biggr) := \sum_{I\in\mathcal{D}}\sum_{J\in\mathcal{D}} a(I,J)c_J H_I \tag{1.10.12}$$

is bounded on $L_p(\mathbb{R})$ for each $1 < p < \infty$.

In DeVore *et al.* (1998) we proved the following theorems.

Theorem 1.10.2. If $\eta(I,\cdot)$, $I \in D$, satisfy assumptions A1–A3, then the operator T defined by (1.10.7) is bounded from $L_p(\mathbb{R})$ into itself for each $1 < p < \infty$.

Corollary 1.10.3. If $\eta(I,\cdot)$, $I \in \mathcal{D}$, satisfy assumptions A1–A3, then $\{\eta(I,\cdot\}_{I\in D} \prec \{H_I\}_{I\in D}$.

We can use a duality argument to give sufficient conditions that the operator T of (1.10.7) is boundedly invertible. For this, we assume that $\eta(I,\cdot)$, $I \in D$, is a family of functions for which there is a dual family $\eta^*(I,\cdot)$, $I \in D$, that satisfies

$$\langle \eta(I,\cdot), \eta^*(J,\cdot)\rangle = \delta(I,J), \quad I,J \in D.$$

Theorem 1.10.4. If the functions $\eta^*(I,\cdot)$, $I \in D$, satisfy assumptions A1–A3, then $\{H_I\}_{I\in D} \prec \{\eta(I,\cdot\}_{I\in D}$.

Theorem 1.10.5. If the systems of functions $\{\eta(I,\cdot)\}_{I\in D}$, $\{\eta^*(I,\cdot)\}_{I\in D}$, satisfy assumptions A1–A3, then the system $\{\eta(I,\cdot)\}_{I\in D}$ is L_p-equivalent to the Haar system $\{H_I\}_{I\in D}$ for $1 < p < \infty$.

It is known from different results (see DeVore *et al.* (1992), DeVore (1998), Temlyakov (2003*a*)) that wavelets are well designed for nonlinear approximation. We present here one general result in this direction. We fix $p \in (1,\infty)$ and consider in $L_p([0,1]^d)$ a basis $\Psi := \{\psi_I\}_{I\in D}$ indexed by dyadic intervals I of $[0,1]^d$, $I = I_1 \times \cdots \times I_d$, I_j is a dyadic interval of $[0,1]$, $j = 1,\dots,d$, which satisfies certain properties. Set $L_p := L_p(\Omega)$ with a normalized Lebesgue measure on Ω, $|\Omega| = 1$. First of all we assume that, for all $1 < q,p < \infty$ and $I \in D$, where $D := D([0,1]^d)$ is the set of all dyadic intervals of $[0,1]^d$, we have

$$\|\psi_I\|_p \asymp \|\psi_I\|_q |I|^{1/p-1/q}, \tag{1.10.13}$$

with constants independent of I. This property can be easily checked for a given basis.

Next, assume that for any $s = (s_1,\dots,s_d) \in \mathbb{Z}^d$, $s_j \ge 0$, $j = 1,\dots,d$, and any $\{c_I\}$, we have for $1 < p < \infty$

$$\Big\|\sum_{I\in D_s} c_I\psi_I\Big\|_p^p \asymp \sum_{I\in D_s} \|c_I\psi_I\|_p^p, \tag{1.10.14}$$

where

$$D_s := \{I = I_1 \times \cdots \times I_d \in D : |I_j| = 2^{-s_j}, \quad j = 1,\dots,d\}.$$

This assumption allows us to estimate the L_p-norm of a dyadic block in terms of Fourier coefficients.

The third assumption is that Ψ is a basis satisfying the following version (weak form) of the Littlewood–Paley inequality, as follows. Let $1 < p < \infty$ and let $f \in L_p$ have the expansion

$$f = \sum_I f_I \psi_I.$$

We assume that

$$\lim_{\min_j \mu_j \to \infty} \Big\| f - \sum_{s_j \le \mu_j, j=1,\dots,d} \sum_{I \in D_s} f_I \psi_I \Big\|_p = 0, \tag{1.10.15}$$

and

$$\|f\|_p \asymp \Big\| \Big(\sum_s \Big| \sum_{I \in D_s} f_I \psi_I \Big|^2 \Big)^{1/2} \Big\|_p. \tag{1.10.16}$$

Let $\mu \in \mathbb{Z}^d$, $\mu_j \ge 0$, $j = 1, \dots, d$. Denote by $\Psi(\mu)$ the subspace of polynomials of the form

$$\psi = \sum_{s_j \le \mu_j, j=1,\dots,d} \sum_{I \in D_s} c_I \psi_I.$$

We now define a function class. Let $R = (R_1, \dots, R_d)$, $R_j > 0$, $j = 1, \dots, d$, and

$$g(R) := \Big(\sum_{j=1}^d R_j^{-1} \Big)^{-1}.$$

For any natural number l, define

$$\Psi(R, l) := \Psi(\mu), \qquad \mu_j = [g(R)l/R_j], \quad j = 1, \dots, d.$$

We define the class $H_q^R(\Psi)$ as the set of functions $f \in L_q$ representable in the form

$$f = \sum_{l=1}^\infty t_l, \quad t_l \in \Psi(R, l), \quad \|t_l\|_q \le 2^{-g(R)l}.$$

We proved in Temlyakov (2002*a*) the following theorem.

Theorem 1.10.6. Let $1 < q, p < \infty$ and $g(R) > (1/q - 1/p)_+$. Then, for Ψ satisfying (1.10.13)–(1.10.16), we have

$$\sup_{f \in H_q^R(\Psi)} \|f - G_m^{L_p}(f, \Psi)\|_p \ll m^{-g(R)}.$$

In the periodic case the basis $U^d := U \times \cdots \times U$ can be used in place of Ψ in Theorem 1.10.6. We define the system $U := \{U_I\}$ in the univariate case. Denote

$$U_n^+(x) := \sum_{k=0}^{2^n-1} \mathrm{e}^{\mathrm{i}kx} = \frac{\mathrm{e}^{\mathrm{i}2^n x} - 1}{\mathrm{e}^{\mathrm{i}x} - 1}, \quad n = 0, 1, 2, \dots,$$

$$U_{n,k}^+(x) := \mathrm{e}^{\mathrm{i}2^n x} U_n^+(x - 2\pi k 2^{-n}), \quad k = 0, 1, \dots, 2^n - 1,$$

$$U_{n,k}^-(x) := \mathrm{e}^{-\mathrm{i}2^n x} U_n^+(-x + 2\pi k 2^{-n}), \quad k = 0, 1, \dots, 2^n - 1.$$

We normalize the system of functions $\{U^+_{n,k}, U^-_{n,k}\}$ in L_2 and enumerate it by dyadic intervals. We write

$$U_I(x) := 2^{-n/2} U^+_{n,k}(x) \quad \text{with} \quad I = [(k+1/2)2^{-n}, (k+1)2^{-n}),$$
$$U_I(x) := 2^{-n/2} U^-_{n,k}(x) \quad \text{with} \quad I = [k2^{-n}, (k+1/2)2^{-n}),$$

and

$$U_{[0,1)}(x) := 1.$$

Wojtaszczyk (1997) proved that U is an unconditional basis of L_p, $1 < p < \infty$. It is well known that $H^R_q(U^d)$ is equivalent to the standard anisotropic multivariate periodic Hölder–Nikol'skii classes NH^R_p. We define these classes in the following way (see Nikol'skii (1975)). The class NH^R_p, $R = (R_1, \dots, R_d)$ and $1 \le p \le \infty$, is the set of periodic functions $f \in L_p([0, 2\pi]^d)$ such that, for each $l_j = [R_j] + 1$, $j = 1, \dots, d$, the following relations hold:

$$\|f\|_p \le 1, \qquad \|\Delta^{l_j,j}_t f\|_p \le |t|^{R_j}, \quad j = 1, \dots, d, \tag{1.10.17}$$

where $\Delta^{l,j}_t$ is the lth difference with step t in the variable x_j. For $d = 1$, NH^R_p coincides with the standard Hölder class H^R_p. Theorem 1.10.6 gives the following result.

Theorem 1.10.7. Let $1 < q, p < \infty$; then for R such that $g(R) > (1/q - 1/p)_+$, we have

$$\sup_{f \in NH^R_q} \|f - G^{L_p}_m(f, U^d)\|_p \ll m^{-g(R)}.$$

We also proved in Temlyakov (2002*a*) that the basis U^d is an optimal orthonormal basis for approximation of classes NH^R_q in L_p:

$$\sigma_m(NH^R_q, \mathbb{O})_p \asymp \sigma_m(NH^R_q, U^d)_p \asymp m^{-g(R)} \tag{1.10.18}$$

for $1 < q < \infty$, $2 \le p < \infty$, $g(R) > (1/q - 1/p)_+$. Here $\mathbb{O}$ is a collection of orthonormal bases. It is important to note that Theorem 1.10.7 guarantees that the estimate in (1.10.18) can be realized by greedy algorithm $G^{L_p}_m(\cdot, U^d)$ with regard to U^d. Another important feature of (1.10.18) is that the basis U^d is optimal (in the sense of order) for each class NH^R_q independently of $R = (R_1, \dots, R_d)$ and q. This property is known as universality for a collection of classes (in the above case, the collection $\{NH^R_q\}$). Further discussion of this important issue can be found in Temlyakov (2002*a*, 2003*a*).

CHAPTER TWO
Greedy approximation with respect to dictionaries: Hilbert spaces

2.1. Introduction

We discuss greedy approximation with regard to redundant systems in this chapter. Greedy approximation is a special form of nonlinear approximation. The basic idea behind nonlinear approximation is that the elements used in the approximation do not come from a fixed linear space but are allowed to depend on the function being approximated. The standard problem in this regard is the problem of m-term approximation, where one fixes a basis and aims to approximate a target function f by a linear combination of m terms of the basis. We discussed this problem in detail in Chapter 1. When the basis is a wavelet basis or a basis of other waveforms, then this type of approximation is the starting point for compression algorithms. An important feature of approximation using a basis

$$\Psi := \{\psi_k\}_{k=1}^{\infty}$$

of a Banach space X is that each function $f \in X$ has a unique representation

$$f = \sum_{k=1}^{\infty} c_k(f)\psi_k, \tag{2.1.1}$$

and we can identify f with the set of its coefficients $\{c_k(f)\}_{k=1}^{\infty}$. The problem of m-term approximation with regard to a basis has been studied thoroughly and rather complete results have been established (see Chapter 1). In particular, it was established that the greedy-type algorithm which forms a sum of m terms with the largest $\|c_k(f)\psi_k\|_X$ out of expansion (2.1.1) realizes in many cases near-best m-term approximation for function classes (DeVore *et al.* 1992) and even for individual functions (see Chapter 1).

Recently, there has emerged another more complicated form of nonlinear approximation, which we call highly nonlinear approximation. It takes many forms but has the basic ingredient that a basis is replaced by a larger system of functions that is usually redundant. We call such systems dictionaries. On the one hand, redundancy offers much promise for greater efficiency in terms of the approximation rate, but on the other hand gives rise to highly non-trivial theoretical and practical problems. The problem of characterizing approximation rate for a given function or function class is now much more substantial and results are quite fragmentary. However, such results are very important for understanding what this new type of approximation offers. Perhaps the first example of this type was considered by Schmidt (1906), who studied the approximation of functions $f(x, y)$ of two variables by bilinear forms,

$$\sum_{i=1}^{m} u_i(x)v_i(y),$$

in $L_2([0,1]^2)$. This problem is closely connected with properties of the integral operator

$$J_f(g) := \int_0^1 f(x,y)g(y)\,\mathrm{d}y$$

with kernel $f(x,y)$. Schmidt (1906) gave an expansion (known as the Schmidt expansion)

$$f(x,y) = \sum_{j=1}^{\infty} s_j(J_f)\phi_j(x)\psi_j(y),$$

where $\{s_j(J_f)\}$ is a non-increasing sequence of singular numbers of J_f, *i.e.*, $s_j(J_f) := \lambda_j(J_f^* J_f)^{1/2}$, where $\{\lambda_j(A)\}$ is the sequence of eigenvalues of an operator A, and J_f^* is the adjoint operator to J_f. The two sequences $\{\phi_j(x)\}$ and $\{\psi_j(y)\}$ form orthonormal sequences of eigenfunctions of the operators $J_f J_f^*$ and $J_f^* J_f$, respectively. He also proved that

$$\left\| f(x,y) - \sum_{j=1}^{m} s_j(J_f)\phi_j(x)\psi_j(y) \right\|_{L_2} = \inf_{u_j, v_j \in L_2,\quad j=1,\dots,m} \left\| f(x,y) - \sum_{j=1}^{m} u_j(x)v_j(y) \right\|_{L_2}.$$

It was understood later that the above best bilinear approximation can be realized by the following greedy algorithm. Assume c_j, $u_j(x)$, $v_j(y)$, $\|u_j\|_{L_2} = \|v_j\|_{L_2} = 1$, $j = 1, \dots, m-1$, have been constructed after $m-1$ steps of the algorithm. At the mth step we choose c_m, $u_m(x)$, $v_m(y)$, $\|u_m\|_{L_2} = \|v_m\|_{L_2} = 1$, to minimize

$$\left\| f(x,y) - \sum_{j=1}^{m} c_j u_j(x)v_j(y) \right\|_{L_2}.$$

We call this type of algorithm the Pure Greedy Algorithm (PGA) (see the general definition below).

Another problem of this type which is well known in statistics is the projection pursuit regression problem, mentioned in the Preface. The problem is to approximate in L_2 a given function $f \in L_2$ by a sum of ridge functions, *i.e.*, by

$$\sum_{j=1}^{m} r_j(\omega_j \cdot x), \quad x, \omega_j \in \mathbb{R}^d, \quad j = 1, \dots, m,$$

where r_j, $j = 1, \dots, m$, are univariate functions. The following greedy-type algorithm (projection pursuit) was proposed in Friedman and Stuetzle (1981) to solve this problem. Assume functions $r_1, \dots, r_{m-1}$ and vectors $\omega_1, \dots, \omega_{m-1}$ have been determined after $m-1$ steps of algorithm. Choose at the mth step a unit vector ω_m and a function r_m to minimize the error

$$\left\| f(x) - \sum_{j=1}^{m} r_j(\omega_j \cdot x) \right\|_{L_2}.$$

This is one more example of a Pure Greedy Algorithm. The Pure Greedy Algorithm and some other versions of greedy-type algorithms have recently been intensively studied: see Barron (1993), Donahue, Gurvits, Darken and Sontag (1997), Davis, Mallat and Avellaneda (1997), DeVore and Temlyakov (1996, 1997), Dubinin (1997), Huber (1985), Jones (1987, 1992), Konyagin and Temlyakov (1999*b*), Livshitz (2006, 2007*a*, 2007*b*), Livshitz and Temlyakov (2001, 2003) and Temlyakov (1999, 2000*b*, 2002*b*, 2003*b*). There are several survey papers that discuss greedy approximation with regard to redundant systems: see DeVore (1998) and Temlyakov (2003*a*, 2006*a*). In this chapter we discuss along with the PGA some of its modifications which are more suitable for implementation. This new type of greedy algorithms will be termed Weak Greedy Algorithms.

In order to orient the reader we recall some notation and definitions from the theory of greedy algorithms. Let H be a real Hilbert space with an inner product $\langle \cdot, \cdot \rangle$ and the norm $\|x\| := \langle x, x \rangle^{1/2}$. We say a set $\mathcal{D}$ of functions (elements) from H is a dictionary if each $g \in \mathcal{D}$ has norm one ($\|g\| = 1$) and the closure of $\operatorname{span} \mathcal{D}$ is equal to H. Sometimes it will be convenient for us also to consider the symmetrized dictionary $\mathcal{D}^{\pm} := \{\pm g : g \in \mathcal{D}\}$. In DeVore and Temlyakov (1996) we studied the following two greedy algorithms. If $f \in H$, we let $g = g(f) \in \mathcal{D}$ be the element from $\mathcal{D}$ which maximizes $|\langle f, g \rangle|$ (we make an additional assumption that a maximizer exists) and define

$$G(f) := G(f, \mathcal{D}) := \langle f, g \rangle g \tag{2.1.2}$$

and

$$R(f) := R(f, \mathcal{D}) := f - G(f).$$

Pure Greedy Algorithm (PGA). We define $f_0 := R_0(f) := R_0(f, \mathcal{D}) := f$ and $G_0(f) := G_0(f, \mathcal{D}) := 0$. Then, for each $m \geq 1$, we inductively define

$$G_m(f) := G_m(f, \mathcal{D}) := G_{m-1}(f) + G(R_{m-1}(f)),$$
$$f_m := R_m(f) := R_m(f, \mathcal{D}) := f - G_m(f) = R(R_{m-1}(f)).$$

We note that the Pure Greedy Algorithm is known under the name Matching Pursuit in signal processing (see, for instance, Mallat and Zhang (1993)).

If H_0 is a finite-dimensional subspace of H, we let P_{H_0} be the orthogonal projector from H onto H_0. That is, $P_{H_0}(f)$ is the best approximation to f from H_0.

Orthogonal Greedy Algorithm (OGA). We define $f_0^o := R_0^o(f) := R_0^o(f,\mathcal{D}) := f$ and $G_0^o(f) := G_0^o(f,\mathcal{D}) := 0$. Then, for each $m \geq 1$, we inductively define

$$\begin{aligned} H_m := H_m(f) &:= \operatorname{span}\{g(R_0^o(f)),\dots,g(R_{m-1}^o(f))\},\\ G_m^o(f) := G_m^o(f,\mathcal{D}) &:= P_{H_m}(f),\\ f_m^o := R_m^o(f) := R_m^o(f,\mathcal{D}) &:= f - G_m^o(f). \end{aligned}$$

We remark that for each f we have

$$\|f_m^o\| \leq \|f_{m-1}^o - G_1(f_{m-1}^o,\mathcal{D})\|. \tag{2.1.3}$$

In Section 1.5 we realized that the Weak Greedy Algorithms with regard to bases work as well as the corresponding Greedy Algorithms. In this chapter we study similar modifications of the Pure Greedy Algorithm (PGA) and the Orthogonal Greedy Algorithm (OGA), which we call, respectively, the Weak Greedy Algorithm (WGA) and the Weak Orthogonal Greedy Algorithm (WOGA). We now give the corresponding definitions from Temlyakov (2000*b*). Let a sequence $\tau = \{t_k\}_{k=1}^\infty$, $0 \leq t_k \leq 1$, be given.

Weak Greedy Algorithm (WGA). We define $f_0^\tau := f$. Then, for each $m \geq 1$ we have the following inductive definition.

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

$$|\langle f_{m-1}^\tau, \varphi_m^\tau\rangle| \geq t_m \sup_{g\in\mathcal{D}} |\langle f_{m-1}^\tau, g\rangle|.$$

(2)
$$f_m^\tau := f_{m-1}^\tau - \langle f_{m-1}^\tau, \varphi_m^\tau\rangle \varphi_m^\tau.$$

(3)
$$G_m^\tau(f,\mathcal{D}) := \sum_{j=1}^m \langle f_{j-1}^\tau, \varphi_j^\tau\rangle \varphi_j^\tau.$$

We note that, for a particular case $t_k = t$, $k = 1,2,\dots$, this algorithm was considered in Jones (1987). Thus, the WGA is a generalization of the PGA making it easier to construct an element φ_m^τ at the mth greedy step. We point out that the WGA contains, in addition to the first (greedy) step, the second step (see (2) and (3) in the above definition) where we update an approximant by adding an orthogonal projection of the residual f_{m-1}^τ onto φ_m^τ. Therefore, the WGA provides for each $f \in H$ an expansion into a series (greedy expansion)

$$f \sim \sum_{j=1}^\infty c_j(f)\varphi_j^\tau, \quad c_j(f) := \langle f_{j-1}^\tau, \varphi_j^\tau\rangle.$$

In general it is not an orthogonal expansion but it has some similar properties. The coefficients $c_j(f)$ of an expansion are obtained by the Fourier formulas with f replaced by the residuals f^τ_{j-1}. It is easy to see that

$$\|f^\tau_m\|^2 = \|f^\tau_{m-1}\|^2 - |c_m(f)|^2.$$

We prove convergence of greedy expansion (see, for instance, Theorem 2.2.4 below), and therefore, from the above equality, we get for this expansion an analogue of the Parseval formula for orthogonal expansions:

$$\|f\|^2 = \sum_{j=1}^{\infty} |c_j(f)|^2.$$

Weak Orthogonal Greedy Algorithm (WOGA). We define $f^{o,\tau}_0 := f$, $f^{o,\tau}_1 := f^\tau_1$, and $\varphi^{o,\tau}_1 := \varphi^\tau_1$, where f^τ_1, φ^τ_1 are are given in the above definition of the WGA. Then, for each $m \geq 2$ we have the following inductive definition.

(1) $\varphi^{o,\tau}_m \in \mathcal{D}$ is any element satisfying

$$|\langle f^{o,\tau}_{m-1}, \varphi^{o,\tau}_m\rangle| \geq t_m \sup_{g\in\mathcal{D}} |\langle f^{o,\tau}_{m-1}, g\rangle|.$$

(2) $$G^{o,\tau}_m(f,\mathcal{D}) := P_{H^\tau_m}(f), \quad \text{where} \quad H^\tau_m := \operatorname{span}(\varphi^{o,\tau}_1, \dots, \varphi^{o,\tau}_m).$$

(3) $$f^{o,\tau}_m := f - G^{o,\tau}_m(f,\mathcal{D}).$$

It is clear that G^τ_m and $G^{o,\tau}_m$ in the case $t_k = 1$, $k = 1, 2, \dots$, coincide with the PGA G_m and the OGA G^o_m, respectively. It is also clear that the WGA and the WOGA are more ready for implementation than the PGA and the OGA. The WOGA has the same greedy step as the WGA and differs in the construction of a linear combination of $\varphi_1, \dots, \varphi_m$. In the WOGA we do our best to construct an approximant out of $H_m := \operatorname{span}(\varphi_1, \dots, \varphi_m)$: we take an orthogonal projection onto H_m. Clearly, in this way we lose a property of the WGA to build an expansion into a series in the case of the WOGA. However, this modification pays off in the sense of improving the convergence rate of approximation. To see this, compare Theorems 2.3.5 and 2.3.6.

There is one more greedy-type algorithm that works well for functions from the convex hull of $\mathcal{D}^\pm$, where $\mathcal{D}^\pm := \{\pm g : g \in \mathcal{D}\}$.

For a general dictionary $\mathcal{D}$ we define the class of functions

$$\mathcal{A}^o_1(\mathcal{D}, M) := \left\{ f \in H : f = \sum_{k\in\Lambda} c_k w_k, \quad w_k \in \mathcal{D}, \ \#\Lambda < \infty, \quad \sum_{k\in\Lambda} |c_k| \leq M \right\}$$

and we define $\mathcal{A}_1(\mathcal{D}, M)$ to be the closure (in H) of $\mathcal{A}^o_1(\mathcal{D}, M)$. Furthermore, we define $\mathcal{A}_1(\mathcal{D})$ to be the union of the classes $\mathcal{A}_1(\mathcal{D}, M)$ over all $M > 0$.

For $f \in \mathcal{A}_1(\mathcal{D})$, we define the norm

$$|f|_{\mathcal{A}_1(\mathcal{D})}$$

to be the smallest M such that $f \in \mathcal{A}_1(\mathcal{D}, M)$.

For $M = 1$ we denote $A_1(\mathcal{D}) := \mathcal{A}_1(\mathcal{D}, 1)$. We proceed to discuss the relaxed type of greedy algorithms. We begin with the simplest one.

Relaxed Greedy Algorithm (RGA). Let $f_o^r := R_0^r(f) := R_0^r(f, \mathcal{D}) := f$ and $G_0^r(f) := G_0^r(f, \mathcal{D}) := 0$. For $m = 1$, we define $G_1^r(f) := G_1^r(f, \mathcal{D}) := G_1(f)$ and $f_1^r := R_1^r(f) := R_1^r(f, \mathcal{D}) := R_1(f)$. For a function $h \in H$, let $g = g(h)$ denote the function from $\mathcal{D}^{\pm}$ which maximizes $\langle h, g\rangle$ (we assume the existence of such an element). Then, for each $m \geq 2$ we inductively define

$$G_m^r(f) := G_m^r(f, \mathcal{D}) := \left(1 - \frac{1}{m}\right) G_{m-1}^r(f) + \frac{1}{m} g(R_{m-1}^r(f)),$$
$$f_m^r := R_m^r(f) := R_m^r(f, \mathcal{D}) := f - G_m^r(f).$$

There are several modifications of the Relaxed Greedy Algorithm (see, for instance, Barron (1993) and DeVore and Temlyakov (1996)). Before giving the definition of the Weak Relaxed Greedy Algorithm (WRGA), we make one remark which helps to motivate the corresponding definition. Assume $G_{m-1} \in A_1(\mathcal{D})$ is an approximant to $f \in A_1(\mathcal{D})$ obtained at the $(m-1)$th step. The major idea of relaxation in greedy algorithms is to look for an approximant at the mth step of the form $G_m := (1-a)G_{m-1} + ag$, $g \in \mathcal{D}^{\pm}$, $0 \leq a \leq 1$. This form guarantees that $G_m \in A_1(\mathcal{D})$. Thus we are looking for co-convex approximants. The best we can do at the mth step is to achieve

$$\delta_m := \inf_{g \in \mathcal{D}^{\pm}, 0 \leq a \leq 1} \|f - ((1-a)G_{m-1} + ag)\|.$$

Let $f_n := f - G_n$, $n = 1, \dots, m$. It is clear that for a given $g \in \mathcal{D}^{\pm}$ we have

$$\inf_a \|f_{m-1} - a(g - G_{m-1})\|^2 = \|f_{m-1}\|^2 - \langle f_{m-1}, g - G_{m-1}\rangle^2 \|g - G_{m-1}\|^{-2},$$

and this infimum is attained for

$$a(g) = \langle f_{m-1}, g - G_{m-1}\rangle \|g - G_{m-1}\|^{-2}.$$

Next, it is not difficult to derive from the definition of $A_1(\mathcal{D})$ and from our assumption on existence of a maximizer that, for any $h \in H$ and $u \in A_1(\mathcal{D})$, there exists $g \in \mathcal{D}^{\pm}$ such that

$$\langle h, g\rangle \geq \langle h, u\rangle. \tag{2.1.4}$$

Taking $h = f_{m-1}$ and $u = f$, we get from (2.1.4) that there exists $g_m \in \mathcal{D}^{\pm}$ such that

$$\langle f_{m-1}, g_m - G_{m-1}\rangle \geq \langle f_{m-1}, f - G_{m-1}\rangle = \|f_{m-1}\|^2. \tag{2.1.5}$$

This implies in particular that we get for g_m

$$\|g_m - G_{m-1}\| \geq \|f_{m-1}\| \tag{2.1.6}$$

and $0 \leq a(g_m) \leq 1$. Thus,

$$\delta_m^2 \leq \|f_{m-1}\|^2 - \frac{1}{4} \sup_{g \in \mathcal{D}^{\pm}} \langle f_{m-1}, g - G_{m-1} \rangle^2.$$

We now give the definition of the WRGA for $f \in A_1(\mathcal{D})$.

Weak Relaxed Greedy Algorithm (WRGA). We define $f_0 := f$ and $G_0 := 0$. Then, for each $m \geq 1$ we have the following inductive definition.

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

$$\langle f_{m-1}, \varphi_m - G_{m-1} \rangle \geq t_m \|f_{m-1}\|^2. \tag{2.1.7}$$

(2)

$$\begin{aligned} G_m &:= G_m(f, \mathcal{D}) := (1-\beta_m) G_{m-1} + \beta_m \varphi_m, \\ \beta_m &:= t_m \left(1 + \sum_{k=1}^{m} t_k^2\right)^{-1} \quad \text{for} \quad m \geq 1. \end{aligned}$$

(3)

$$f_m := f - G_m.$$

2.2. Convergence

We begin this section with convergence of the Weak Orthogonal Greedy Algorithm (WOGA). The following theorem was proved in Temlyakov (2000*b*).

Theorem 2.2.1. Assume

$$\sum_{k=1}^{\infty} t_k^2 = \infty. \tag{2.2.1}$$

Then, for any dictionary $\mathcal{D}$ and any $f \in H$, we have for the WOGA

$$\lim_{m \to \infty} \|f_m^{o,\tau}\| = 0. \tag{2.2.2}$$

Remark 2.2.2. It is easy to see that if $\mathcal{D} = \mathcal{B}$, an orthonormal basis, the assumption (2.2.1) is also necessary for convergence (2.2.2) for all f.

Proof of Theorem 2.2.1. Let $f \in H$ and let $\varphi_1^{o,\tau}, \varphi_2^{o,\tau}, \ldots$ be as given in the definition of the WOGA. Let

$$H_n := H_n^{\tau} = \operatorname{span}(\varphi_1^{o,\tau}, \ldots, \varphi_n^{o,\tau}).$$

It is clear that $H_n \subseteq H_{n+1}$, and therefore $\{P_{H_n}(f)\}$ converges to some function v. The following Lemma 2.2.3 says that $v = f$ and completes the proof of Theorem 2.2.1. □

Lemma 2.2.3. Assume that (2.2.1) is satisfied. Then, if $\{f_m^\tau\}_{m=1}^\infty$ or $\{f_m^{o,\tau}\}_{m=1}^\infty$ converges, it converges to zero.

Proof of Lemma 2.2.3. We prove this lemma by contradiction. Let us consider first the case of $\{f_m^\tau\}_{m=1}^\infty$. Assume $f_m^\tau \to u \neq 0$ as $m \to \infty$. It is clear that

$$\sup_{g \in \mathcal{D}} |\langle u, g\rangle| \geq 2\delta$$

with some $\delta > 0$. Therefore, there exists N such that, for all $m \geq N$, we have

$$\sup_{g \in \mathcal{D}} |\langle f_m^\tau, g\rangle| \geq \delta.$$

From the definition of the WGA we get for all $m > N$

$$\|f_m^\tau\|^2 = \|f_{m-1}^\tau\|^2 - |\langle f_{m-1}^\tau, \varphi_m^\tau\rangle|^2 \leq \|f_N^\tau\|^2 - \delta^2 \sum_{k=N+1}^m t_k^2,$$

which contradicts (2.2.1).

We now proceed to the case $\{f_m^{o,\tau}\}_{m=1}^\infty$. Assume $f_m^{o,\tau} \to u \neq 0$ as $m \to \infty$. Then, as in the above proof, there exist $\delta > 0$ and N such that, for all $m \geq N$, we have

$$\sup_{g \in \mathcal{D}} |\langle f_m^{o,\tau}, g\rangle| \geq \delta.$$

Next, as in (2.1.3) we have

$$\|f_m^{o,\tau}\|^2 \leq \|f_{m-1}^{o,\tau}\|^2 - t_m^2 \Big(\sup_{g \in \mathcal{D}} |\langle f_{m-1}^{o,\tau}, g\rangle|\Big)^2 \leq \|f_N^{o,\tau}\|^2 - \delta^2 \sum_{k=N+1}^m t_k^2,$$

which contradicts the divergence of $\sum_k t_k^2$. □

Theorem 2.2.1 and Remark 2.2.2 show that (2.2.1) is a necessary and sufficient condition on weakness sequence $\tau = \{t_k\}$ in order that the WOGA converges for each f and all $\mathcal{D}$. Condition (2.2.1) can be rewritten as $\tau \notin \ell_2$. It turns out that the convergence of the PGA is more delicate. We now proceed to the corresponding results. The following theorem gives a criterion of convergence in a special case of monotone weakness sequences $\{t_k\}$. Sufficiency was proved in Temlyakov (2000*b*) and necessity in Livshitz and Temlyakov (2001).

Theorem 2.2.4. In the class of monotone sequences $\tau = \{t_k\}_{k=1}^\infty$, $1 \geq t_1 \geq t_2 \geq \cdots \geq 0$, the condition

$$\sum_{k=1}^\infty \frac{t_k}{k} = \infty \tag{2.2.3}$$

is necessary and sufficient for convergence of the Weak Greedy Algorithm for each f and all Hilbert spaces H and dictionaries $\mathcal{D}$.

Remark 2.2.5. We note that the sufficiency part of Theorem 2.2.4 (see Temlyakov (2000*b*)) does not need the monotonicity of τ.

Proof of sufficiency condition in Theorem 2.2.4. This proof (see Temlyakov (2000*b*)) is a refinement of the original proof of Jones (1987). The following lemma, Lemma 2.2.6, combined with Lemma 2.2.3, implies sufficiency in Theorem 2.2.4. □

Lemma 2.2.6. Assume (2.2.3) is satisfied. Then $\{f_m^\tau\}_{m=1}^\infty$ converges.

Proof of Lemma 2.2.6. It is easy to derive from the definition of the WGA the following two relations:

$$f_m^\tau = f - \sum_{j=1}^{m} \langle f_{j-1}^\tau, \varphi_j^\tau \rangle \varphi_j^\tau, \tag{2.2.4}$$

$$\|f_m^\tau\|^2 = \|f\|^2 - \sum_{j=1}^{m} |\langle f_{j-1}^\tau, \varphi_j^\tau \rangle|^2. \tag{2.2.5}$$

Let $a_j := |\langle f_{j-1}^\tau, \varphi_j^\tau \rangle|$. We get from (2.2.5) that

$$\sum_{j=1}^{\infty} a_j^2 \le \|f\|^2. \tag{2.2.6}$$

We take any two indices $n < m$ and consider

$$\|f_n^\tau - f_m^\tau\|^2 = \|f_n^\tau\|^2 - \|f_m^\tau\|^2 - 2\langle f_n^\tau - f_m^\tau, f_m^\tau \rangle.$$

Let

$$\theta_{n,m}^\tau := |\langle f_n^\tau - f_m^\tau, f_m^\tau \rangle|.$$

Using (2.2.4) and the definition of the WGA, we obtain, for all $n < m$ and all m such that $t_{m+1} \neq 0$,

$$\theta_{n,m}^\tau \le \sum_{j=n+1}^{m} |\langle f_{j-1}^\tau, \varphi_j^\tau \rangle| |\langle f_m^\tau, \varphi_j^\tau \rangle| \le \frac{a_{m+1}}{t_{m+1}} \sum_{j=1}^{m+1} a_j. \tag{2.2.7}$$

□

We now need a property of ℓ_2-sequences.

Lemma 2.2.7. Assume $y_j \ge 0$, $j = 1, 2, \dots$, and

$$\sum_{k=1}^{\infty} \frac{t_k}{k} = \infty, \qquad \sum_{j=1}^{\infty} y_j^2 < \infty.$$

Then

$$\varliminf_{n\to\infty} \frac{y_n}{t_n} \sum_{j=1}^{n} y_j = 0.$$

Proof. Let $P(\tau) := \{n \in \mathbb{N} : t_n \neq 0\}$. Consider a series

$$\sum_{n\in P(\tau)} \frac{t_n}{n} \frac{y_n}{t_n} \sum_{j=1}^{n} y_j. \tag{2.2.8}$$

We shall prove that this series converges. It is clear that convergence of this series together with the assumption $\sum_{k=1}^{\infty} t_k/k = \infty$ imply the statement of Lemma 2.2.7.

We use the following known fact. If $\{y_j\}_{j=1}^{\infty} \in \ell_2$ then $\{n^{-1}\sum_{j=1}^{n} y_j\}_{n=1}^{\infty} \in \ell_2$ (see Zygmund (1959, Chapter 1, Section 9)). By the Cauchy inequality, we have

$$\sum_{n\in P(\tau)} \frac{t_n}{n} \frac{y_n}{t_n} \sum_{j=1}^{n} y_j \leq \left(\sum_{n=1}^{\infty} y_n^2\right)^{1/2} \left(\sum_{n=1}^{\infty}\left(n^{-1}\sum_{j=1}^{n} y_j\right)^2\right)^{1/2} < \infty.$$

This completes the proof of Lemma 2.2.7. □

Relation (2.2.7) and Lemma 2.2.7 imply that

$$\varliminf_{m\to\infty} \max_{n<m} \theta^{\tau}_{n,m} = 0.$$

It remains to use the following simple lemma.

Lemma 2.2.8. In a Banach space X, let a sequence $\{x_n\}_{n=1}^{\infty}$ be given, such that, for any k, l, we have

$$\|x_k - x_l\|^2 = y_k - y_l + \vartheta_{k,l},$$

where $\{y_n\}_{n=1}^{\infty}$ is a convergent sequence of real numbers and the real sequence $\vartheta_{k,l}$ satisfies the property

$$\varliminf_{l\to\infty} \max_{k<l} |\vartheta_{k,l}| = 0.$$

Then $\{x_n\}_{n=1}^{\infty}$ converges.

The necessary condition in Theorem 2.2.4 was proved in Livshitz and Temlyakov (2001). We do not present it here.

Theorem 2.2.4 solves the problem of convergence of the WGA in the case of monotone weakness sequences. We now consider the case of general weakness sequences. In Theorem 2.2.4 we reduced the proof of convergence of the WGA with weakness sequence τ to some properties of ℓ_2-sequences with regard to τ. The sufficiency part of Theorem 2.2.4 was derived from the following two statements.

Proposition 2.2.9. Let τ be such that, for any $\{a_j\}_{j=1}^\infty \in \ell_2$, $a_j \geq 0$, $j = 1, 2, \dots$, we have

$$\liminf_{n\to\infty} a_n \sum_{j=1}^n a_j/t_n = 0.$$

Then, for any H, $\mathcal{D}$, and $f \in H$ we have

$$\lim_{m\to\infty} \|f_m^\tau\| = 0.$$

Proposition 2.2.10. If τ satisfies condition (2.2.3) then τ satisfies the assumption of Proposition 2.2.9.

We proved in Temlyakov (2002*b*) a criterion on τ for convergence of the WGA. Let us introduce some notation. We define by $\mathcal{V}$ the class of sequences $x = \{x_k\}_{k=1}^\infty$, $x_k \geq 0$, $k = 1, 2, \dots$, with the following property: there exists a sequence $0 = q_0 < q_1 < \cdots$ that may depend on x such that

$$\sum_{s=1}^\infty \frac{2^s}{\Delta q_s} < \infty \tag{2.2.9}$$

and

$$\sum_{s=1}^\infty 2^{-s} \sum_{k=1}^{q_s} x_k^2 < \infty, \tag{2.2.10}$$

where $\Delta q_s := q_s - q_{s-1}$.

Remark 2.2.11. It is clear from this definition that, if $x \in \mathcal{V}$ and for some N and c, we have $0 \leq y_k \leq c x_k$, $k \geq N$, then $y \in \mathcal{V}$.

Theorem 2.2.12. The condition $\tau \notin \mathcal{V}$ is necessary and sufficient for convergence of all realizations of the Weak Greedy Algorithm with weakness sequence τ for each f and all Hilbert spaces H and dictionaries $\mathcal{D}$.

The proof of the sufficiency part of Theorem 2.2.12 is a refinement of the corresponding proof of Theorem 2.2.4. The study of the behaviour of sequences $a_n \sum_{j=1}^n a_j$ for $\{a_j\}_{j=1}^\infty \in \ell_2$, $a_j \geq 0$, $j = 1, 2, \dots$, plays an important role in both proofs. It turns out that the class $\mathcal{V}$ appears naturally in the study of the above-mentioned sequences. We proved the following theorem in Temlyakov (2002*b*).

Theorem 2.2.13. The following two conditions are equivalent:

$$\tau \notin \mathcal{V}, \tag{2.2.11}$$

$$\forall \{a_j\}_{j=1}^\infty \in \ell_2, \quad a_j \geq 0, \quad \liminf_{n\to\infty} a_n \sum_{j=1}^n a_j/t_n = 0. \tag{2.2.12}$$

Theorem 2.2.12 solves the problem of convergence of the WGA in a very general situation. The sufficiency part of Theorem 2.2.12 guarantees that, whenever $\tau \notin \mathcal{V}$, the WGA converges for each f and all $\mathcal{D}$. The necessity part of Theorem 2.2.12 states that, if $\tau \in \mathcal{V}$, then there exist an element f and a dictionary $\mathcal{D}$ such that some realization $G_m^\tau(f, \mathcal{D})$ of the WGA does not converge to f. However, Theorem 2.2.12 leaves open the following interesting and important problem. Let a dictionary $\mathcal{D} \subset H$ be given. Find necessary and sufficient conditions on a weakness sequence τ in order that $G_m^\tau(f, \mathcal{D}) \to f$ for each $f \in H$. The corresponding open problems for special dictionaries are formulated in Temlyakov (2003*a*, pp. 78, 81). They concern the following two classical dictionaries:

$$\Pi_2 := \{u(x)v(y) : u, v \in L_2([0,1]), \quad \|u\|_2 = \|v\|_2 = 1\},$$

and

$$\mathcal{R}_2 := \{g(x) = r(\omega \cdot x) : \|g\|_2 = 1\},$$

where r is a univariate function and $\omega \cdot x$ is the scalar product of x, $\|x\|_{\ell_2} \le 1$, and a unit vector $\omega \in \mathbb{R}^2$.

2.3. Rate of convergence

2.3.1. Upper bounds for approximation by general dictionaries

We shall discuss here approximation from a general dictionary $\mathcal{D}$. We begin with a discussion of the approximation properties of the Relaxed Greedy Algorithm. The result we give below in Theorem 2.3.2 is from DeVore and Temlyakov (1996), and can be found in the paper of Jones (1992) in a different form. We begin with the following elementary lemma about numerical sequences.

Lemma 2.3.1. If $A > 0$ and $\{a_n\}_{n=1}^\infty$ is a sequence of non-negative numbers satisfying $a_1 \le A$ and

$$a_m \le a_{m-1} - \frac{2}{m} a_{m-1} + \frac{A}{m^2}, \quad m = 2, 3, \ldots, \tag{2.3.1}$$

then

$$a_m \le \frac{A}{m}. \tag{2.3.2}$$

Proof. The proof is by induction. Suppose we have

$$a_{m-1} \le \frac{A}{m-1}$$

for some $m \ge 2$. Then, from our assumption (2.3.1) we have

$$a_m \le \frac{A}{m-1}\left(1 - \frac{2}{m}\right) + \frac{A}{m^2} = A\left(\frac{1}{m} - \frac{1}{(m-1)m} + \frac{1}{m^2}\right) \le \frac{A}{m}. \qquad \square$$

If $f \in \mathcal{A}_1^o(\mathcal{D})$, then $f = \sum_j c_j g_j$, for some $g_j \in \mathcal{D}$ and with $\sum_j |c_j| \le 1$. Since the functions g_j all have norm one, it follows that

$$\|f\| \le \sum_j |c_j| \|g_j\| \le 1.$$

Since the functions $g \in \mathcal{D}$ have norm one, it follows that $G_1^r(f) = G_1(f)$ also has norm at most one. By induction, we find that $\|G_m^r(f)\| \le 1$, $m \ge 1$.

Theorem 2.3.2. For the Relaxed Greedy Algorithm we have, for each $f \in A_1(\mathcal{D})$, the estimate

$$\|f - G_m^r(f)\| \le \frac{2}{\sqrt{m}}, \quad m \ge 1. \tag{2.3.3}$$

Proof. We use the abbreviation $r_m := G_m^r(f)$ and $g_m := g(R_{m-1}^r(f))$. From the definition of r_m, we have

$$\|f - r_m\|^2 = \|f - r_{m-1}\|^2 + \frac{2}{m}\langle f - r_{m-1}, r_{m-1} - g_m\rangle + \frac{1}{m^2}\|r_{m-1} - g_m\|^2. \tag{2.3.4}$$

The last term on the right-hand side of (2.3.4) does not exceed $4/m^2$. For the middle term, we have

$$\begin{aligned}\langle f - r_{m-1}, r_{m-1} - g_m\rangle &= \inf_{g \in \mathcal{D}^\pm} \langle f - r_{m-1}, r_{m-1} - g\rangle \\ &= \inf_{\phi \in A_1(\mathcal{D})} \langle f - r_{m-1}, r_{m-1} - \phi\rangle \\ &\le \langle f - r_{m-1}, r_{m-1} - f\rangle = -\|f - r_{m-1}\|^2.\end{aligned}$$

We substitute this in (2.3.4) to obtain

$$\|f - r_m\|^2 \le \left(1 - \frac{2}{m}\right)\|f - r_{m-1}\|^2 + \frac{4}{m^2}. \tag{2.3.5}$$

Thus the theorem follows from Lemma 2.3.1 with $A=4$ and $a_m := \|f - r_m\|^2$. □

We now turn our discussion to the approximation properties of the Pure Greedy Algorithm and the Orthogonal Greedy Algorithm.

We shall need the following simple known lemma (see, for example, DeVore and Temlyakov (1996)).

Lemma 2.3.3. Let $\{a_m\}_{m=1}^\infty$ be a sequence of non-negative numbers satisfying the inequalities

$$a_1 \le A, \quad a_{m+1} \le a_m(1 - a_m/A), \quad m = 1, 2, \ldots.$$

Then we have for each m

$$a_m \le A/m.$$

Proof. The proof is by induction on m. For $m = 1$ the statement is true by assumption. We assume $a_m \le A/m$ and prove that $a_{m+1} \le A/(m+1)$. If $a_{m+1} = 0$ this statement is obvious. Assume therefore that $a_{m+1} > 0$. Then we have

$$a_{m+1}^{-1} \ge a_m^{-1}(1 - a_m/A)^{-1} \ge a_m^{-1}(1 + a_m/A) = a_m^{-1} + A^{-1} \ge (m+1)A^{-1},$$

which implies $a_{m+1} \le A/(m+1)$. □

We now want to estimate the decrease in error provided by one step of the Pure Greedy Algorithm. Let $\mathcal{D}$ be an arbitrary dictionary. If $f \in H$ and

$$\rho(f) := \langle f, g(f)\rangle / \|f\|, \tag{2.3.6}$$

where as before $g(f) \in \mathcal{D}^{\pm}$ satisfies

$$\langle f, g(f)\rangle = \sup_{g \in \mathcal{D}^{\pm}} \langle f, g\rangle,$$

then

$$R(f)^2 = \|f - G(f)\|^2 = \|f\|^2(1 - \rho(f)^2). \tag{2.3.7}$$

The larger $\rho(f)$, the better the decrease of the error in the step of the Pure Greedy Algorithm. The following lemma estimates $\rho(f)$ from below.

Lemma 2.3.4. If $f \in \mathcal{A}_1(\mathcal{D}, M)$, then

$$\rho(f) \ge \|f\|/M. \tag{2.3.8}$$

Proof. It is sufficient to prove (2.3.8) for $f \in \mathcal{A}_1^o(\mathcal{D}, M)$ since the general result follows from this by taking limits. We can write $f = \sum c_k g_k$, where this sum has a finite number of terms and $g_k \in \mathcal{D}$ and $\sum |c_k| \le M$. Hence,

$$\|f\|^2 = \langle f, f\rangle = \left\langle f, \sum c_k g_k \right\rangle = \sum c_k \langle f, g_k\rangle \le M\rho(f)\|f\|,$$

and (2.3.8) follows. □

The following theorem was proved in DeVore and Temlyakov (1996).

Theorem 2.3.5. Let $\mathcal{D}$ be an arbitrary dictionary in H. Then, for each $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have

$$\|f - G_m(f, \mathcal{D})\| \le M m^{-1/6}.$$

Proof. It is enough to prove the theorem for $f \in \mathcal{A}_1(\mathcal{D}, 1)$; the general result then follows by rescaling. We shall use the abbreviated notation $f_m := R_m(f)$ for the residual. Let

$$a_m := \|f_m\|^2 = \|f - G_m(f, \mathcal{D})\|^2, \quad m = 0, 1, \dots, \quad f_0 := f,$$

and define the sequence $\{b_m\}_{m=0}^{\infty}$ by

$$b_0 := 1, \quad b_{m+1} := b_m + \rho(f_m)\|f_m\|, \quad m = 0, 1, \dots .$$

Since $f_{m+1} := f_m + \rho(f_m)\|f_m\|g(f_m)$, we obtain by induction that

$$f_m \in \mathcal{A}_1(\mathcal{D}, b_m), \quad m = 0, 1, \ldots,$$

and consequently we have the following relations for $m = 0, 1, \ldots$:

$$a_{m+1} = a_m(1 - \rho(f_m)^2), \tag{2.3.9}$$

$$b_{m+1} = b_m + \rho(f_m)a_m^{1/2}, \tag{2.3.10}$$

$$\rho(f_m) \geq a_m^{1/2}b_m^{-1}. \tag{2.3.11}$$

The last two relations give

$$b_{m+1} = b_m(1 + \rho(f_m)a_m^{1/2}b_m^{-1}) \leq b_m(1 + \rho(f_m)^2). \tag{2.3.12}$$

Combining this inequality with (2.3.9) we find

$$a_{m+1}b_{m+1} \leq a_m b_m(1 - \rho(f_m)^4),$$

which in turn implies for all m

$$a_m b_m \leq a_0 b_0 = \|f\|^2 \leq 1. \tag{2.3.13}$$

Further, using (2.3.9) and (2.3.11) we get

$$a_{m+1} = a_m(1 - \rho(f_m)^2) \leq a_m(1 - a_m/b_m^2).$$

Since $b_n \leq b_{n+1}$, this gives

$$a_{n+1}b_{n+1}^{-2} \leq a_n b_n^{-2}(1 - a_n b_n^{-2}).$$

Applying Lemma 2.3.3 to the sequence $(a_m b_m^{-2})$ we obtain

$$a_m b_m^{-2} \leq m^{-1}. \tag{2.3.14}$$

Relations (2.3.13) and (2.3.14) imply

$$a_m^3 = (a_m b_m)^2 a_m b_m^{-2} \leq m^{-1}.$$

In other words,

$$\|f_m\| = a_m^{1/2} \leq m^{-1/6},$$

which proves the theorem. □

The next theorem (DeVore and Temlyakov 1996) estimates the error in approximation by the Orthogonal Greedy Algorithm.

Theorem 2.3.6. Let $\mathcal{D}$ be an arbitrary dictionary in H. Then, for each $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have

$$\|f - G_m^o(f, \mathcal{D})\| \leq Mm^{-1/2}.$$

Proof. The proof of this theorem is similar to the proof of Theorem 2.3.5 but technically even simpler. We can again assume that $M = 1$. We let

$f_m^o := R_m^o(f)$ be the residual in the Orthogonal Greedy Algorithm. Then, from the definition of Orthogonal Greedy Algorithm, we have

$$\|f_{m+1}^o\| \le \|f_m^o - G_1(f_m^o, \mathcal{D})\|. \tag{2.3.15}$$

From (2.3.7), we obtain

$$\|f_{m+1}^o\|^2 \le \|f_m^o\|^2(1 - \rho(f_m^o)^2). \tag{2.3.16}$$

By the definition of the Orthogonal Greedy Algorithm, $G_m^o(f) = P_{H_m} f$ and hence $f_m^o = f - G_m^o(f)$ is orthogonal to $G_m^o(f)$. Using this as in the proof of Lemma 2.3.4, we obtain

$$\|f_m^o\|^2 = \langle f_m^o, f\rangle \le \rho(f_m^o)\|f_m^o\|.$$

Hence,

$$\rho(f_m^o) \ge \|f_m^o\|.$$

Using this inequality in (2.3.16), we find

$$\|f_{m+1}^o\|^2 \le \|f_m^o\|^2(1 - \|f_m^o\|^2).$$

In order to complete the proof it remains to apply Lemma 2.3.3 with $A = 1$ and $a_m = \|f_m^o\|^2$. □

2.3.2. Upper estimates for weak-type greedy algorithms

We begin this subsection with an error estimate for the Weak Orthogonal Greedy Algorithm. The following theorem from Temlyakov (2000*b*) is a generalization of Theorem 2.3.6.

Theorem 2.3.7. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Then, for each $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have*

$$\|f - G_m^{o,\tau}(f, \mathcal{D})\| \le M\left(1 + \sum_{k=1}^m t_k^2\right)^{-1/2}.$$

We now turn to the Weak Relaxed Greedy Algorithms. The following theorem from Temlyakov (2000*b*) shows that the WRGA performs on the $\mathcal{A}_1(\mathcal{D}, M)$ similar to the WOGA. We note that the approximation step of building the $G_m(f, \mathcal{D})$ in the WRGA is simpler than the corresponding step of building the $G_m^{o,\tau}(f, \mathcal{D})$ in the WOGA.

Theorem 2.3.8. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Then, for each $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have for the Weak Relaxed Greedy Algorithm*

$$\|f - G_m(f, \mathcal{D})\| \le 2M\left(1 + \sum_{k=1}^m t_k^2\right)^{-1/2}.$$

We now proceed to the Weak Greedy Algorithm. The construction of an approximant $G^\tau_m(f,\mathcal{D})$ in the WGA is the simplest out of the three types of algorithms (WGA, WOGA, WRGA) discussed here. We pointed out above that the WGA provides for each $f \in H$ an expansion into a series that satisfies an analogue of the Parseval formula. The following theorem from Temlyakov (2000*b*) gives the upper bounds for the residual $\|f^\tau_m\|$ of the WGA that are not as good as in Theorems 2.3.7 and 2.3.8 for the WOGA and WRGA, respectively. The next theorem, Theorem 2.3.10, shows that the bound (2.3.17) is sharp in a certain sense.

Theorem 2.3.9. Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume $\tau := \{t_k\}_{k=1}^\infty$ is a non-increasing sequence. Then, for $f \in \mathcal{A}_1(\mathcal{D}, M)$ we have

$$\|f - G^\tau_m(f,\mathcal{D})\| \le M\left(1 + \sum_{k=1}^m t_k^2\right)^{-t_m/2(2+t_m)}. \tag{2.3.17}$$

In a particular case $\tau = \{t\}$, $(t_k = t,\ k = 1, 2, \ldots)$, (2.3.17) gives

$$\|f - G^t_m(f,\mathcal{D})\| \le M(1 + mt^2)^{-t/(4+2t)}, \quad 0 < t \le 1. \tag{2.3.18}$$

This estimate implies the inequality

$$\|f - G^t_m(f,\mathcal{D})\| \le C_1(t)m^{-at}|f|_{\mathcal{A}_1(\mathcal{D})}, \tag{2.3.19}$$

with the exponent at approaching 0 linearly in t. We proved in Livshitz and Temlyakov (2003) that this exponent cannot decrease to 0 at a slower rate than linear.

Theorem 2.3.10. There exists an absolute constant $b > 0$ such that, for any $t > 0$, we can find a dictionary $\mathcal{D}_t$ and a function $f_t \in \mathcal{A}_1(\mathcal{D}_t)$ such that, for some realization $G^t_m(f_t, \mathcal{D}_t)$ of the Weak Greedy Algorithm, we have

$$\liminf_{m\to\infty} \|f_t - G^t_m(f_t,\mathcal{D}_t)\| m^{bt}/|f_t|_{\mathcal{A}_1(\mathcal{D}_t)} > 0. \tag{2.3.20}$$

Remark 2.3.11. The estimate (2.3.18) implies that for small t the parameter a in (2.3.19) can be taken close to 1/4. The proof from Livshitz and Temlyakov (2003) implies that the parameter b in (2.3.20) can be taken close to $(\ln 2)^{-1}$.

We now discuss some further results on the rate of convergence of the PGA and related results on greedy expansions. Theorem 2.3.5 states that for a general dictionary $\mathcal{D}$ the Pure Greedy Algorithm provides the estimate

$$\|f - G_m(f,\mathcal{D})\| \le |f|_{\mathcal{A}_1(\mathcal{D})} m^{-1/6}.$$

The above estimate was improved a little in Konyagin and Temlyakov (1999*b*) to

$$\|f - G_m(f,\mathcal{D})\| \le 4|f|_{\mathcal{A}_1(\mathcal{D})} m^{-11/62}.$$

We now discuss recent progress on the following open problem (Temlyakov 2003*a*, p. 65, Open Problem 3.1). This problem is a central theoretical problem in greedy approximation in Hilbert spaces.

Open problem. Find the order of decay of the sequence

$$\gamma(m) := \sup_{f,\mathcal{D},\{G_m\}} \big(\|f - G_m(f,\mathcal{D})\| |f|_{\mathcal{A}_1(\mathcal{D})}^{-1}\big),$$

where the supremum is taken over all dictionaries $\mathcal{D}$, all elements $f \in \mathcal{A}_1(\mathcal{D}) \setminus \{0\}$ and all possible choices of $\{G_m\}$.

Recently, the known upper bounds in approximation by the Pure Greedy Algorithm were improved in Sil'nichenko (2004). Sil'nichenko proved the estimate

$$\gamma(m) \le Cm^{-\frac{s}{2(2+s)}},$$

where s is a solution from $[1, 1.5]$ of the equation

$$(1+x)^{\frac{1}{2+x}}\left(\frac{2+x}{1+x}\right) - \frac{1+x}{x} = 0.$$

Numerical calculations of s (see Sil'nichenko (2004)) give

$$\frac{s}{2(2+s)} = 0.182\cdots > 11/62.$$

The technique used in Sil'nichenko (2004) is a further development of a method from Konyagin and Temlyakov (1999*b*).

There is also some progress in the lower estimates. The estimate

$$\gamma(m) \ge Cm^{-0.27},$$

with a positive constant C, was proved in Livshitz and Temlyakov (2003). For previous lower estimates see Temlyakov (2003*a*, p. 59). Very recently Livshitz (2007*b*), using the technique from Livshitz and Temlyakov (2003), proved the following lower estimate:

$$\gamma(m) \ge Cm^{-0.1898}. \tag{2.3.21}$$

We mentioned above that the PGA and its generalization the Weak Greedy Algorithm (WGA) give, for every element $f \in H$, a convergent expansion in a series with respect to a dictionary $\mathcal{D}$. We discuss a further generalization of the WGA that also provides a convergent expansion. We consider here a generalization of the WGA obtained by introducing to it a tuning parameter $b \in (0, 1]$ (see Temlyakov (2007*a*)). Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, and a parameter $b \in (0, 1]$ be given. We define the Weak Greedy Algorithm with parameter b as follows.

Weak Greedy Algorithm with parameter b (WGA(b)). We define $f_0^{\tau,b} := f$. Then, for each $m \ge 1$ we have the following inductive definition.

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

$$|\langle f_{m-1}^{\tau,b}, \varphi_m^{\tau,b}\rangle| \geq t_m \sup_{g\in\mathcal{D}} |\langle f_{m-1}^{\tau,b}, g\rangle|.$$

(2) $$f_m^{\tau,b} := f_{m-1}^{\tau,b} - b\langle f_{m-1}^{\tau,b}, \varphi_m^{\tau,b}\rangle \varphi_m^{\tau,b}.$$

(3) $$G_m^{\tau,b}(f,\mathcal{D}) := b\sum_{j=1}^{m} \langle f_{j-1}^{\tau,b}, \varphi_j^{\tau,b}\rangle \varphi_j^{\tau,b}.$$

We note that the WGA(b) can be seen as a realization of the Approximate Greedy Algorithm studied in Gribonval and Nielsen (2001*a*) and Galatenko and Livshitz (2003, 2005).

We point out that the WGA(b), like the WGA, contains, in addition to the first (greedy) step, the second step (see (2) and (3) in the above definition) where we update an approximant by adding an orthogonal projection of the residual $f_{m-1}^{\tau,b}$ onto $\varphi_m^{\tau,b}$ multiplied by b. The WGA(b), therefore, provides, for each $f \in H$, an expansion into a series (greedy expansion):

$$f \sim \sum_{j=1}^{\infty} c_j(f)\varphi_j^{\tau,b}, \quad c_j(f) := b\langle f_{j-1}^{\tau,b}, \varphi_j^{\tau,b}\rangle.$$

We begin with a convergence result from Temlyakov (2007*a*).

Theorem 2.3.12. *Let $\tau \notin \mathcal{V}$. Then the WGA(b) with $b \in (0,1]$ converges for each f and all Hilbert spaces H and dictionaries $\mathcal{D}$.*

Theorem 2.3.12 is an extension of the corresponding result for the WGA (see Theorem 2.2.12).

We proved in Temlyakov (2007*a*) the following convergence rate of the WGA(b).

Theorem 2.3.13. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume $\tau := \{t_k\}_{k=1}^\infty$ is a non-increasing sequence and $b \in (0,1]$. Then, for $f \in A_1(\mathcal{D})$ we have*

$$\|f - G_m^{\tau,b}(f,\mathcal{D})\| \leq \left(1 + b(2-b)\sum_{k=1}^{m} t_k^2\right)^{-(2-b)t_m/2(2+(2-b)t_m)}. \tag{2.3.22}$$

This theorem is an extension of the corresponding result for the WGA (see Theorem 2.3.9). In the particular case $t_k = 1$, $k = 1, 2, \dots$, we get the following rate of convergence:

$$\|f - G_m^{1,b}(f,\mathcal{D})\| \leq Cm^{-r(b)}, \quad r(b) := \frac{2-b}{2(4-b)}.$$

We note that $r(1) = 1/6$ and $r(b) \to 1/4$ as $b \to 0$. Thus we can offer the following observation. At each step of the Pure Greedy Algorithm we can choose a fixed fraction of the optimal coefficient for that step instead of the optimal coefficient itself. Surprisingly, this leads to better upper estimates than those known for the Pure Greedy Algorithm.

2.4. Greedy algorithms for systems that are not dictionaries

In this section we discuss greedy algorithms with regard to a system $\mathcal{G}$ that is not a dictionary. Here, we will discuss a variant of the RGA that is a generalization of the version of the RGA suggested by Barron (1993). Let H be a real Hilbert space and let $\mathcal{G} := \{g\}$ be a system of elements $g \in H$ such that $\|g\| \le C_0$. Usually, in the theory of greedy algorithms we consider approximation with regard to a dictionary $\mathcal{D}$. One of the properties of a dictionary $\mathcal{D}$ is that the closure of $\operatorname{span} \mathcal{D}$ is equal to H. In this section we do not assume that the system $\mathcal{G}$ is a dictionary. In particular, we do not assume that the closure of $\operatorname{span} \mathcal{G}$ is H. This setting is motivated by applications in learning theory (see Section 2.8). We present here results from Temlyakov (2005*d*). Let $\mathcal{G}^{\pm} := \{\pm g : g \in \mathcal{G}\}$ denote the symmetrized system $\mathcal{G}$, and let $\theta > 0$.

RGA($\boldsymbol{\theta}$) with respect to $\boldsymbol{\mathcal{G}}$. For $f \in H$ we define $f_0 := f$, $G_0 := G_0(f) := 0$. Then, for each $n \ge 1$ we have the following inductive definition.

(1) $\varphi_n \in \mathcal{G}^{\pm}$ is an element satisfying (we assume existence)

$$\langle f_{n-1}, \varphi_n \rangle = \max_{g \in \mathcal{G}^{\pm}} \langle f_{n-1}, g \rangle.$$

(2) $$G_n := G_n(f) := \left(1 - \frac{\theta}{n+\theta}\right) G_{n-1} + \frac{\theta}{n+\theta}\varphi_n, \quad f_n := f - G_n.$$

Let $A_1(\mathcal{G})$ denote the closure in H of the convex hull of $\mathcal{G}^{\pm}$. Then, for $f \in H$ there exists a unique element $f' \in A_1(\mathcal{G})$ such that

$$d(f, A_1(\mathcal{G}))_H = \|f - f'\| \le \|f - \phi\|, \quad \phi \in A_1(\mathcal{G}). \tag{2.4.1}$$

In analysis of the RGA(θ) we will use the following simple lemma (see DeVore and Temlyakov (1996) and Lemma 2.3.1 above for a variant of this lemma). Our analysis is similar to that of DeVore and Temlyakov (1996) and Lee, Bartlett and Williamson (1996).

Lemma 2.4.1. Let a sequence $\{a_n\}_{n=0}^{\infty}$ of non-negative numbers satisfy the relations (with $\beta > 1$, $B > 0$)

$$a_n \le \frac{n}{n+\beta} a_{n-1} + \frac{B}{(n+\beta)^2}, \quad n = 1, 2, \dots, \qquad a_0 \le \frac{B}{(\beta-1)\beta}.$$

Then, for all n,

$$a_n \le \frac{B}{(\beta-1)(n+\beta)}.$$

Proof. Setting $A := B/(\beta - 1)$, we obtain by induction

$$a_n \leq \frac{A}{n-1+\beta}\frac{n}{n+\beta}+\frac{B}{(n+\beta)^2} = \frac{A}{n+\beta}-\frac{A(\beta-1)}{(n+\beta)(n-1+\beta)}+\frac{B}{(n+\beta)^2}.$$

Taking into account the inequality

$$\frac{A(\beta-1)}{(n+\beta)(n-1+\beta)} \geq \frac{A(\beta-1)}{(n+\beta)^2} = \frac{B}{(n+\beta)^2},$$

we complete the proof. □

Theorem 2.4.2. For $\theta > 1$ there exists a constant $C(\theta)$ such that, for any $f \in H$, we have

$$\|f_n\|^2 \leq d(f, A_1(\mathcal{G}))_H^2 + C(\theta)(\|f\| + C_0)^2 n^{-1}.$$

Proof. From the definition of G_n and f_n we get, setting $\alpha := \frac{\theta}{n+\theta}$,

$$f_n = f - G_n = (1-\alpha)f_{n-1} + \alpha(f - \varphi_n)$$

and

$$\|f_n\|^2 = (1-\alpha)^2\|f_{n-1}\|^2 + 2\alpha(1-\alpha)\langle f_{n-1}, f-\varphi_n\rangle + \alpha^2\|f-\varphi_n\|^2. \quad (2.4.2)$$

It is known and easy to check that for any $h \in H$ we have

$$\sup_{g\in\mathcal{G}^{\pm}} \langle h, g\rangle = \sup_{\phi\in A_1(\mathcal{G})} \langle h, \phi\rangle. \quad (2.4.3)$$

Denote f' as above and set $f^* := f - f'$. Using (2.4.3) and the definition of φ_n, we obtain from (2.4.2)

$$\begin{aligned}\|f_n\|^2 &\leq (1-\alpha)^2\|f_{n-1}\|^2 + 2\alpha(1-\alpha)\langle f_{n-1}, f-f'\rangle + \alpha^2\|f-\varphi_n\|^2\\ &= (1-\alpha)(\|f_{n-1}\|^2 - \alpha\|f_{n-1}\|^2 + 2\alpha\langle f_{n-1}, f^*\rangle - \alpha\|f^*\|^2)\\ &\quad + \alpha(1-\alpha)\|f^*\|^2 + \alpha^2\|f-\varphi_n\|^2\\ &\leq (1-\alpha)\|f_{n-1}\|^2 + \alpha\|f^*\|^2 + \alpha^2\|f-\varphi_n\|^2.\end{aligned}$$

This implies

$$\|f_n\|^2 - \|f^*\|^2 \leq (1-\alpha)(\|f_{n-1}\|^2 - \|f^*\|^2) + \alpha^2(\|f\| + C_0)^2.$$

Setting $a_n := \|f_n\|^2 - \|f^*\|^2$, $\beta := \theta$, and applying Lemma 2.4.1, we complete the proof. □

Theorem 2.4.3. For $\theta > 1/2$ there exists a constant $C := C(\theta, C_0)$ such that, for any $f \in H$, we have

$$\|f' - G_n(f)\|^2 \leq C/n.$$

Proof. If $f \in A_1(\mathcal{G})$ then the statement of Theorem 2.4.3 follows from known results (see Barron (1993) and Theorem 2.3.2). If $d(f, A_1(\mathcal{G})) > 0$, then property (2.4.1) implies that, for any $\phi \in A_1(\mathcal{G})$, we have

$$\langle f^*, \phi - f' \rangle \le 0. \tag{2.4.4}$$

It follows from the definition of f_n that

$$f_n = \left(1 - \frac{\theta}{n+\theta}\right) f_{n-1} + \frac{\theta}{n+\theta}(f - \varphi_n).$$

We set $f'_n := f_n - f^*$. Then, we get from the above representation

$$f'_n = \left(1 - \frac{\theta}{n+\theta}\right) f'_{n-1} + \frac{\theta}{n+\theta}(f' - \varphi_n).$$

We note that $f'_n = f' - G_n(f)$. Let us estimate

$$\begin{aligned}\|f'_n\|^2 = \|f'_{n-1}\|^2 &\left(1 - \frac{\theta}{n+\theta}\right)^2 \\ &+ \frac{2\theta}{n+\theta}\left(1 - \frac{\theta}{n+\theta}\right)\langle f'_{n-1}, f' - \varphi_n\rangle + \frac{\theta^2}{(n+\theta)^2}\|f' - \varphi_n\|^2.\end{aligned} \tag{2.4.5}$$

Next,

$$\begin{aligned}\langle f'_{n-1}, f' - \varphi_n \rangle &= \langle f'_{n-1} + f^*, f' - \varphi_n \rangle - \langle f^*, f' - \varphi_n \rangle \\ &= \langle f_{n-1}, f' - \varphi_n \rangle + \langle f^*, \varphi_n - f' \rangle.\end{aligned} \tag{2.4.6}$$

First, we prove that

$$\langle f_{n-1}, f' - \varphi_n \rangle \le 0. \tag{2.4.7}$$

It easily follows from $f' \in A_1(\mathcal{G})$ that

$$\langle f_{n-1}, f' \rangle \le \max_{g \in \mathcal{G}^{\pm}} \langle f_{n-1}, g \rangle. \tag{2.4.8}$$

By the definition of φ_n we get

$$\max_{g \in \mathcal{G}^{\pm}} \langle f_{n-1}, g \rangle = \langle f_{n-1}, \varphi_n \rangle. \tag{2.4.9}$$

Thus, (2.4.7) follows from (2.4.8) and (2.4.9).

Secondly, we note that (2.4.4) implies

$$\langle f^*, \varphi_n - f' \rangle \le 0. \tag{2.4.10}$$

Therefore, by (2.4.6), (2.4.7), and (2.4.10) we obtain

$$\langle f'_{n-1}, f' - \varphi_n \rangle \le 0. \tag{2.4.11}$$

Substitution of (2.4.11) in (2.4.5) gives

$$\|f'_n\|^2 \le \|f'_{n-1}\|^2\left(1 - \frac{2\theta}{n+\theta}\right) + \frac{\theta^2}{(n+\theta)^2}(\|f'_{n-1}\|^2 + \|f' - \varphi_n\|^2). \tag{2.4.12}$$

Using bounds $\|f'_{n-1}\| \leq C_0$ and $\|f' - \varphi_n\| \leq 2C_0$, we find

$$\|f'_n\|^2 \leq \|f'_{n-1}\|^2\left(1 - \frac{2\theta}{n+\theta}\right) + 5C_0^2\theta^2/(n+\theta)^2.$$

We note that

$$1 - \frac{2\theta}{n+\theta} < 1 - \frac{2\theta}{n+2\theta}.$$

We now apply Lemma 2.4.1 with $a_n = \|f'_n\|^2$, $\beta = 2\theta$, and get

$$\|f'_n\|^2 \leq C(\theta, C_0)/n. \tag{2.4.13}$$

This completes the proof. □

2.5. Saturation property of greedy-type algorithms

In this section we shall give an example from DeVore and Temlyakov (1996) which shows that replacing a dictionary $\mathcal{B}$ given by an orthogonal basis by a non-orthogonal redundant dictionary $\mathcal{D}$ may damage the efficiency of the Pure Greedy Algorithm. The dictionary $\mathcal{D}$ in our example differs from the dictionary $\mathcal{B}$ by the one addition of the element g for a certain suitably chosen function g.

Let $\mathcal{B} := \{h_k\}_{k=1}^{\infty}$ be an orthonormal basis in a Hilbert space H. Consider the following element:

$$g := Ah_1 + Ah_2 + aA\sum_{k\geq 3}(k(k+1))^{-1/2}h_k, \tag{2.5.1}$$

with

$$A := (33/89)^{1/2} \quad \text{and} \quad a := (23/11)^{1/2}.$$

Then $\|g\| = 1$. We define the dictionary $\mathcal{D} := \mathcal{B} \cup \{g\}$.

Theorem 2.5.1. For the function

$$f := h_1 + h_2,$$

we have

$$\|f - G_m(f)\| \geq m^{-1/2}, \quad m \geq 4.$$

Proof. We shall examine the steps of the Pure Greedy Algorithm applied to the function $f = h_1 + h_2$. We shall use the abbreviated notation $f_m := R_m(f) := f - G_m(f)$ for the residual at step m.

First step. We have

$$\langle f, g\rangle = 2A > 1, \quad |\langle f, h_k\rangle| \leq 1, \quad k = 1, 2, \dots.$$

This implies

$$G_1(f, D) = \langle f, g\rangle g,$$

and

$$f_1 = f - \langle f, g\rangle g = (1-2A^2)(h_1+h_2) - 2aA^2\sum_{k\ge 3}(k(k+1))^{-1/2}h_k.$$

Second step. We have

$$\langle f_1, g\rangle = 0, \quad \langle f_1, h_k\rangle = (1-2A^2), \quad k=1,2, \quad \langle f_1, h_3\rangle = -aA^2 3^{-1/2}.$$

Comparing $\langle f_1, h_1\rangle$ and $|\langle f_1, h_3\rangle|$ we get

$$|\langle f_1, h_3\rangle| = (23/89)(33/23)^{1/2} > 23/89 = 1-2A^2 = \langle f_1, h_1\rangle.$$

This implies that the second approximation $G_1(f_1, D)$ is $\langle f_1, h_3\rangle h_3$ and

$$f_2 = f_1 - \langle f_1, h_3\rangle h_3 = (1-2A^2)(h_1+h_2) - 2aA^2\sum_{k\ge 4}(k(k+1))^{-1/2}h_k.$$

Third step. We have

$$\begin{aligned}\langle f_2, g\rangle &= -\langle f_1, h_3\rangle\langle h_3, g\rangle = (A/2)(23/89),\\ \langle f_2, h_1\rangle &= \langle f_2, h_2\rangle = 1-2A^2 = 23/89,\\ \langle f_2, h_4\rangle &= -aA^2 5^{-1/2} = -(23/89)(99/115)^{1/2}.\end{aligned}$$

Therefore, the third approximation should be $\langle f_2, h_1\rangle h_1$ or $\langle f_2, h_2\rangle h_2$. Let us take the first of these so that

$$f_3 = f_2 - \langle f_2, h_1\rangle h_1.$$

Fourth step. It is clear that for all $k \ne 1$ we have

$$\langle f_3, h_k\rangle = \langle f_2, h_k\rangle.$$

This equality and the calculations from the third step show that it is sufficient to compare $\langle f_3, h_2\rangle$ and $\langle f_3, g\rangle$. We have

$$\langle f_3, g\rangle = \langle f_2, g\rangle - \langle f_2, h_1\rangle\langle h_1, g\rangle = -(23/89)(A/2).$$

This means that

$$f_4 = f_3 - \langle f_3, h_2\rangle h_2 = -2aA^2\sum_{k\ge 4}(k(k+1))^{-1/2}h_k. \tag{2.5.2}$$

***m*th step (*m* > 4).** We prove by induction that for all $m \ge 4$ we have

$$f_m = -2aA^2\sum_{k\ge m}(k(k+1))^{-1/2}h_k. \tag{2.5.3}$$

For $m=4$ this relation follows from (2.5.2). We assume we have proved (2.5.3) for some m and derive that (2.5.3) also holds true for $m+1$. To find f_{m+1}, we only have to compare the two inner products: $\langle f_m, h_m\rangle$ and $\langle f_m, g\rangle$. We have

$$|\langle f_m, h_m\rangle| = 2aA^2(m(m+1))^{-1/2},$$

and

$$|\langle f_m, g\rangle| = 2a^2A^3 \sum_{k\geq m} (k(k+1))^{-1} = 2a^2A^3m^{-1}.$$

Since

$$(|\langle f_m, g\rangle|/|\langle f_m, h_m\rangle|)^2 = (aA)^2(1+1/m) \leq 345/356 < 1,$$

we have that

$$|\langle f_m, g\rangle| < |\langle f_m, h_m\rangle|, \quad m \geq 4.$$

This proves (2.5.3) with m replaced by $m+1$.

From (2.5.3), we obtain

$$\|f - G_m(f, D)\| = \|f_m\| = 2aA^2m^{-1/2} > m^{-1/2}, \quad m \geq 4. \qquad \square$$

2.6. Lebesgue-type inequalities for greedy approximation

Lebesgue proved the following inequality: for any 2π-periodic continuous function f we have

$$\|f - S_n(f)\|_\infty \leq \left(4 + \frac{4}{\pi^2}\ln n\right) E_n(f)_\infty, \tag{2.6.1}$$

where $S_n(f)$ is the nth partial sum of the Fourier series of f and $E_n(f)_\infty$ is the error of the best approximation of f by the trigonometric polynomials of order n in the uniform norm $\|\cdot\|_\infty$. The inequality (2.6.1) relates the error of a particular method (S_n) of approximation by the trigonometric polynomials of order n to the best-possible error $E_n(f)_\infty$ of approximation by the trigonometric polynomials of order n. By a Lebesgue-type inequality we mean an inequality that provides an upper estimate for the error of a particular method of approximation of f by elements of a special form, say, form $\mathcal{A}$, by the best-possible approximation of f by elements of the form $\mathcal{A}$. In the case of approximation with regard to bases (or minimal systems), the Lebesgue-type inequalities are known both in linear and in nonlinear settings (see Chapter 1 and surveys by Konyagin and Temlyakov (2002) and Temlyakov (2003*a*)). It would be very interesting to prove the Lebesgue-type inequalities for redundant systems (dictionaries). However, there are substantial difficulties.

We begin our discussion with the Pure Greedy Algorithm (PGA). It is natural to compare performance of the PGA with the best m-term approximation with regard to a dictionary $\mathcal{D}$. We let $\Sigma_m(\mathcal{D})$ denote the collection of all functions (elements) in H which can be expressed as a linear combination of at most m elements of $\mathcal{D}$. Thus, each function $s \in \Sigma_m(\mathcal{D})$ can be written in the form

$$s = \sum_{g\in\Lambda} c_g g, \quad \Lambda \subset \mathcal{D}, \quad \#\Lambda \leq m,$$

where the c_g are real or complex numbers. In some cases, it may be possible to write an element from $\Sigma_m(\mathcal{D})$ in this form in more than one way. The space $\Sigma_m(\mathcal{D})$ is not linear: the sum of two functions from $\Sigma_m(\mathcal{D})$ is generally not in $\Sigma_m(\mathcal{D})$.

For a function $f \in H$ we define its best m-term approximation error:

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

It seems there is no hope of proving a non-trivial Lebesgue-type inequality for the PGA in the case of an arbitrary dictionary $\mathcal{D}$. This pessimism is based on the following result from DeVore and Temlyakov (1996) (see Section 2.5).

Let $\mathcal{B} := \{h_k\}_{k=1}^{\infty}$ be an orthonormal basis in a Hilbert space H. Consider the following element:

$$g := Ah_1 + Ah_2 + aA \sum_{k \geq 3} (k(k+1))^{-1/2} h_k,$$

with

$$A := (33/89)^{1/2} \quad \text{and} \quad a := (23/11)^{1/2}.$$

Then $\|g\| = 1$. We define the dictionary $\mathcal{D} = \mathcal{B} \cup \{g\}$. It was proved in DeVore and Temlyakov (1996) (see Section 2.5 above) that, for the function

$$f = h_1 + h_2,$$

we have

$$\|f - G_m(f, \mathcal{D})\| \geq m^{-1/2}, \quad m \geq 4.$$

It is clear that $\sigma_2(f, \mathcal{D}) = 0$.

Therefore, we look for conditions on a dictionary $\mathcal{D}$ that allow us to prove Lebesgue-type inequalities. The condition $\mathcal{D} = \mathcal{B}$, an orthonormal basis for H, guarantees that

$$\|R_m(f, \mathcal{B})\| = \sigma_m(f, \mathcal{B}).$$

This is an ideal situation. The results that we will discuss here concern the case when we replace an orthonormal basis $\mathcal{B}$ by a dictionary that is, in a certain sense, not far from an orthonormal basis.

Let us begin with results from Donoho, Elad and Temlyakov (2007) that are close to results from Temlyakov (1999). We give a definition of a λ-quasi-orthogonal dictionary with depth D. When $D = \infty$ this definition coincides with the definition of a λ-quasi-orthogonal dictionary from Temlyakov (1999).

Definition 2.6.1. We say $\mathcal{D}$ is a λ-quasi-orthogonal dictionary with depth D if, for any $n \in [1, D]$ and any $g_i \in \mathcal{D}$, $i = 1, \ldots, n$, there exists a collection

$\varphi_j \in \mathcal{D}$, $j = 1, \dots, J$, $J \le N := \lambda n$, with the properties

$$g_i \in X_J := \operatorname{span}(\varphi_1, \dots, \varphi_J), \quad i = 1, \dots, n,$$

and for any $f \in X_J$ we have

$$\max_{1 \le j \le J} |\langle f, \varphi_j \rangle| \ge N^{-1/2} \|f\|.$$

Remark 2.6.2. It is clear that an orthonormal dictionary is a 1-quasi-orthogonal dictionary.

The following theorem for $D = \infty$ was established in Temlyakov (1999). It is pointed out in Donoho *et al.* (2007) that the proof from Temlyakov (1999) also works in the case $D < \infty$, and gives the following result.

Theorem 2.6.3. Let a given dictionary $\mathcal{D}$ be λ-quasi-orthogonal with depth D, and let $0 < r < (2\lambda)^{-1}$ be a real number. Then, for any f such that

$$\sigma_m(f, \mathcal{D}) \le m^{-r}, \quad m = 1, 2, \dots, D,$$

we have

$$\|f_m\| = \|f - G_m(f, \mathcal{D})\| \le C(r, \lambda) m^{-r}, \quad m \in [1, D/2].$$

In this section we consider dictionaries that have become popular in signal processing. Denote

$$M(\mathcal{D}) := \sup_{g \ne h; g, h \in \mathcal{D}} |\langle g, h \rangle|,$$

the coherence parameter of a dictionary $\mathcal{D}$. For an orthonormal basis $\mathcal{B}$ we have $M(\mathcal{B}) = 0$. It is clear that the smaller $M(\mathcal{D})$, the more $\mathcal{D}$ resembles an orthonormal basis. However, we should note that in the case $M(\mathcal{D}) > 0$ the $\mathcal{D}$ can be a redundant dictionary. We showed in Donoho *et al.* (2007) (see Proposition 2.1) that a dictionary with coherence $M := M(\mathcal{D})$ is a $(1 + 4\delta)$-quasi-orthogonal dictionary with depth δ/M, for any $\delta \in (0, 1/7]$. Therefore, Theorem 2.6.3 applies to M-coherent dictionaries. We proved in Donoho *et al.* (2007) a general Lebesgue-type inequality for the PGA with regard to an M-coherent dictionary.

Theorem 2.6.4. Let a dictionary $\mathcal{D}$ have the mutual coherence $M = M(\mathcal{D})$. Then, for any $S \le 1/(2M)$ we have the following inequality:

$$\|f_S\|^2 \le 2\|f\|(\sigma_S(f, \mathcal{D}) + 5MS\|f\|). \tag{2.6.2}$$

As a direct corollary of this theorem we obtain the following inequality for functions f that allow an S-sparse representation in $\mathcal{D}$ ($\sigma_S(f) = 0$):

$$\|f_S\| \le (10MS)^{1/2} \|f\|.$$

Inequality (2.6.2) is the first Lebesgue-type inequality for the PGA in the case of incoherent dictionary $\mathcal{D}$.

We now proceed to a discussion of the Orthogonal Greedy Algorithm (OGA). It is clear from the definition of the OGA that at each step we have (see (2.1.3))

$$\|f_m^o\|^2 \le \|f_{m-1}^o\|^2 - |\langle f_{m-1}^o, g(f_{m-1}^o)\rangle|^2.$$

We noted in Donoho *et al.* (2007) that the use of this inequality instead of the equality

$$\|f_m\|^2 = \|f_{m-1}\|^2 - |\langle f_{m-1}, g(f_{m-1})\rangle|^2,$$

which holds for the PGA, allows us to prove an analogue of Theorem 2.6.3 for the OGA. The proof repeats the corresponding proof from Temlyakov (1999). We formulate this as a remark.

Remark 2.6.5. Theorem 2.6.3 holds for the OGA instead of the PGA (for $\|f_m^o\|$ instead of $\|f_m\|$).

The first general Lebesgue-type inequality for the OGA for the M-coherent dictionary was obtained in Gilbert, Muthukrishnan and Strauss (2003). They proved that

$$\|f_m^o\| \le 8m^{1/2}\sigma_m(f) \quad \text{for } m < 1/(32M).$$

The constants in this inequality were improved in Tropp (2004) (see also Donoho, Elad and Temlyakov (2006)):

$$\|f_m^o\| \le (1+6m)^{1/2}\sigma_m(f) \quad \text{for } m < 1/(3M). \tag{2.6.3}$$

We proved in Donoho *et al.* (2007) an analogue of (2.6.2) for the OGA.

Theorem 2.6.6. Let a dictionary $\mathcal{D}$ have the mutual coherence $M = M(\mathcal{D})$. Then, for any $S \le 1/(2M)$ we have the following inequalities:

$$\|f_S^o\|^2 \le 2\|f_k^o\|(\sigma_{S-k}(f_k^o) + 3MS\|f_k^o\|), \quad 0 \le k \le S. \tag{2.6.4}$$

Inequality (2.6.4) can be used to improve (2.6.3) for small m. We proved in Donoho *et al.* (2007) the following inequality.

Theorem 2.6.7. Let a dictionary $\mathcal{D}$ have the mutual coherence $M = M(\mathcal{D})$. Assume $m \le 0.05M^{-2/3}$. Then, for $l \ge 1$ satisfying $2^l \le \log m$ we have

$$\|f_{m(2^l-1)}^o\| \le 6m^{2^{-l}}\sigma_m(f).$$

Corollary 2.6.8. Let a dictionary $\mathcal{D}$ have the mutual coherence $M = M(\mathcal{D})$. Assume $m \le 0.05M^{-2/3}$. Then we have

$$\|f_{[m\log m]}^o\| \le 24\sigma_m(f). \tag{2.6.5}$$

Inequality (2.6.5) is an almost perfect Lebesgue-type inequality. It has the following two deficiencies. First, clearly we would like to replace $[m\log m]$ by m or Cm in the number of iterations of the OGA. However, this is a minor drawback of (2.6.5). Second, as is stated in Corollary 2.6.8, inequality (2.6.5) holds for only a small number of iterations: $m \le 0.05M^{-2/3}$. It would be interesting to know if we can push the limit from $M^{-2/3}$ to a natural limit of M^{-1}.

The above results show that the smaller the coherence parameter $M(\mathcal{D})$, the better the performance of the OGA. In particular, (2.6.3) implies that if f is S-sparse with respect to $\mathcal{D}$ ($\sigma_S(f,\mathcal{D}) = 0$), then $f_S^o = 0$, provided that $S < 1/(3M)$. This means that the OGA exactly recovers S-sparse elements with respect to the M-coherent dictionary $\mathcal{D}$ if $S < 1/(3M)$. This is a very nice property of M-coherent dictionaries, which is important in applications (see Donoho *et al.* (2006) for a discussion). Therefore, it is very desirable to build dictionaries with a small coherent parameter $M(\mathcal{D})$. A rigorous setting in this regard is the following. Let $H = \mathbb{R}^d$ and let the cardinality of $\mathcal{D}$ be equal to N ($|\mathcal{D}| = N$). Find

$$c(N,d) := \inf_{\mathcal{D},|\mathcal{D}|=N} M(\mathcal{D})$$

and describe the *Grassmannian dictionaries* for which $M(\mathcal{D}) = c(N,d)$, $|\mathcal{D}| = N$.

In a special case when $\mathcal{D}$ is assumed to be a frame, the dictionaries (frames) described above are known as Grassmannian frames. The theory of Grassmannian frames is a beautiful mathematical theory that has connections to areas such as spherical codes, algebraic geometry, graph theory, and sphere packings (see Stromberg and Heath (2003)). Some fundamental problems of this theory are still open. For instance, it is known that in the case of frames we have

$$c^{\text{frame}}(N,d) \ge \left(\frac{N-d}{d(N-1)}\right)^{1/2}. \tag{2.6.6}$$

However, it is not known for which pairs (N,d) we have equality in (2.6.6).

2.7. Some further remarks

In the Preface we mentioned two classical examples of redundant dictionaries:

$$\Pi_2 := \bigl\{u(x_1)v(x_2) : (x_1,x_2) \in [0,1]^2,\ \|u\|_{L_2([0,1])} = \|v\|_{L_2([0,1])} = 1\bigr\},$$

$$\mathcal{R}_2 := \bigl\{r(\omega\cdot x) : x,\omega \in \mathbb{R}^d,\ \|x\|_{\ell_2} \le 1,\ \|\omega\|_{\ell_2} = 1,\ \|r(\omega\cdot x)\|_{L_2(B_2^d)} = 1\bigr\}.$$

The reader can find detailed discussion of the m-term approximation with regard to these dictionaries in the survey of Temlyakov (2003*a*). The dictionary Π_2 is a very interesting example from the point of view of greedy approximation. As mentioned in the Preface, we have for each $f \in L_2([0,1]^2)$

$$\|f - G_m(f,\Pi_2)\|_{L_2([0,1]^2)} = \sigma_m(f,\Pi_2)_{L_2([0,1]^2)}, \tag{2.7.1}$$

for the PGA. This means that the Π_2 is ideally designed for greedy approximation. Clearly, all the general results of this chapter apply in the case $\mathcal{D} = \Pi_2$. Surprisingly, we do not quite understand how to use specific properties of Π_2 in greedy approximation. For instance (see the end of Section 2.2), there has been no progress on the following open problem (see Temlyakov (2003*a*), p.78). Find the necessary and sufficient conditions on a weakness sequence τ to guarantee convergence of the WGA with regard to Π_2 for each $f \in L_2([0,1]^2)$.

We note that the Schmidt expansion formula for $f \in L_2([0,1]^2)$,

$$f(x_1, x_2) = \sum_{j=1}^{\infty} s_j(J_f)\phi_j(x_1)\psi_j(x_2),$$

points out the importance of the sequence $\{s_j(J_f)\}$ of singular numbers of the integral operator J_f associated with f. There is an extensive literature devoted to estimating $s_j(J_f)$ and $\sigma_m(f, \Pi_2)$ (in different norms) in terms of smoothness of f. We mention some of the papers: Fredholm (1903), Weyl (1911), Hille and Tamarkin (1931), Smithies (1937), Birman and Solomyak (1977), Cochran (1977) and Temlyakov (1989*b*, 1990, 1992*a*, 1992*b*, 1993*b*). For a further discussion see the survey of Temlyakov (2003*a*).

The dictionary $\mathcal{R}_2$ is not as good as Π_2 for greedy approximation. There are some weaker analogues of (2.7.1) for greedy approximation with regard to $\mathcal{R}_2$ (see Maiorov, Oskolkov and Temlyakov (2002)). However, as in the case of Π_2, we do not know how to use specific features of $\mathcal{R}_2$ in greedy approximation. There has also been no progress on an open problem (see Temlyakov (2003*a*, p. 81)), similar to the one mentioned above, on convergence of the WGA with regard to $\mathcal{R}_2$.

We now proceed to a discussion of some recent results on simultaneous greedy approximation. A new ingredient of the papers of Lutoborski and Temlyakov (2003), Leviatan and Temlyakov (2005, 2006) and Temlyakov (2004) is the move from approximating a single element f, to simultaneous approximation of a set of elements $f^1, \dots, f^N$. We will give a description of some results from Lutoborski and Temlyakov (2003), Leviatan and Temlyakov (2006) and Temlyakov (2004). The main purpose of the above papers is to construct greedy-type expansions,

$$f^i \sim \sum_{j=1}^{\infty} c_j^i(f)\varphi_j, \quad c_j^i(f) := \langle f_{j-1}^i, \varphi_j \rangle, \tag{2.7.2}$$

for a given finite set of elements $f^1, \dots, f^N$, simultaneously with the same sequence $\{\varphi_j\}$ for all f^i, $i = 1, \dots, N$. The first result in this direction was obtained in Lutoborski and Temlyakov (2003). The Vector Greedy Algorithms that are designed for the purpose of constructing mth greedy approximants, simultaneously for a given finite number of elements, were introduced and studied in Lutoborski and Temlyakov (2003).

Vector Weak Greedy Algorithm (VWGA). Let a vector of elements $f^i \in H$, $i = 1, \dots, N$ be given. We write $f_0^{i,v,\tau} := f^i$. Then, for each $m \geq 1$ we have the following inductive definition.

(1) Let $\varphi_m^{v,\tau} \in \mathcal{D}$ be any element satisfying

$$\max_i |\langle f_{m-1}^{i,v,\tau}, \varphi_m^{v,\tau} \rangle| \geq t_m \max_i \sup_{g \in \mathcal{D}} |\langle f_{m-1}^{i,v,\tau}, g \rangle|. \tag{2.7.3}$$

(2)
$$f_m^{i,v,\tau} := f_{m-1}^{i,v,\tau} - \langle f_{m-1}^{i,v,\tau}, \varphi_m^{v,\tau} \rangle \varphi_m^{v,\tau}, \quad i = 1, \dots, N.$$

(3)
$$G_m^{v,\tau}(f^i, \mathcal{D}) := \sum_{j=1}^{m} \langle f_{j-1}^{i,v,\tau}, \varphi_j^{v,\tau} \rangle \varphi_j^{v,\tau}, \quad i = 1, \dots, N.$$

It was proved in Lutoborski and Temlyakov (2003) that the VWGA converges for $\tau \notin \mathcal{V}$. Therefore the VWGA with $\tau \notin \mathcal{V}$ provides the convergent expansions

$$f^i = \sum_{j=1}^{\infty} b_j^i g_j, \qquad g_j \in \mathcal{D},$$

with the property

$$\|f^i\|^2 = \sum_{j=1}^{\infty} |b_j^i|^2, \qquad i = 1, \dots, N.$$

The following estimate of the rate of convergence of the VWGA was obtained in Lutoborski and Temlyakov (2003).

Theorem 2.7.1. Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k = t$, $k = 1, \dots,$ $0 < t < 1$. Then, for any vector of elements $f^1, \dots, f^N$, $f^i \in A_1(\mathcal{D})$, $i = 1, \dots, N$, we have

$$\sum_{i=1}^{N} \|f_m^{i,v,\tau}\|^2 \leq \left(1 + \frac{mt^2}{N}\right)^{-t/(2N+t)} N^{\frac{2N+2t}{2N+t}}.$$

Comparing Theorem 2.3.9 with $\tau = \{t\}$ with Theorem 2.7.1, we see that the exponent $t/(2N+t)$ of decay is seriously affected by the number N of simultaneously approximated elements. Also, simultaneous approximation brings an extra factor, $N^{\frac{2N+2t}{2N+t}} \asymp N$. In Leviatan and Temlyakov (2006) we improve the exponent of decay, replacing $t/(2N+t)$ by $t/(2N^{1/2}+t)$, and we get the worse constant N^2 instead of N. Here is the corresponding theorem from Leviatan and Temlyakov (2006).

Theorem 2.7.2. Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume $\tau := \{t_k\}_{k=1}^{\infty}$ is a non-increasing sequence. Then, for any vector of elements $f^1, \dots, f^N$, $f^i \in A_1(\mathcal{D})$, $i = 1, \dots, N$, we have

$$\sum_{i=1}^{N} \|f_m^{i,v,\tau}\|^2 \le N^2 \left(1 + \frac{1}{N} \sum_{k=1}^{m} t_k^2\right)^{\frac{-t_m}{2N^{1/2}+t_m}}.$$

In addition to the VWGA the following two modifications of the VWGA were considered in Leviatan and Temlyakov (2006). The modifications differ from the VWGA only in the first step. In the first step of the Simultaneous Weak Greedy Algorithm 1 (SWGA1), we have the following.

(1)[SWGA1] We look for any $\varphi_m^{s1,\tau} \in \mathcal{D}$ satisfying

$$\left(\sum_{i=1}^{N} |\langle f_{m-1}^i, \varphi_m^{s1,\tau}\rangle|^2\right)^{1/2} \ge t_m \max_i \sup_{g \in \mathcal{D}} |\langle f_{m-1}^i, g\rangle|, \quad f_{m-1}^i := f_{m-1}^{i,s1,\tau}. \tag{2.7.4}$$

The first step of the Simultaneous Weak Greedy Algorithm 2 (SWGA2) is then as follows.

(1)[SWGA2] We look for any $\varphi_m^{s2,\tau} \in \mathcal{D}$ satisfying

$$\left(\sum_{i=1}^{N} |\langle f_{m-1}^i, \varphi_m^{s2,\tau}\rangle|^2\right)^{1/2} \ge t_m \sup_{g \in \mathcal{D}} \left(\sum_{i=1}^{N} |\langle f_{m-1}^i, g\rangle|^2\right)^{1/2}, \quad f_{m-1}^i := f_{m-1}^{i,s2,\tau}. \tag{2.7.5}$$

Clearly, any φ_m satisfying (2.7.3) or (2.7.5) also satisfies (2.7.4). Thus, any upper estimate for the SWGA1 yields an upper estimate for both the VWGA and the SWGA2. It was proved in Leviatan and Temlyakov (2006) that Theorem 2.7.2 holds for both variants of the Simultaneous Weak Greedy Algorithm.

We proved in Temlyakov (2004) the following estimate that improves the estimates in Theorems 2.7.1 and 2.7.2. It combines good features of estimates from Theorems 2.7.1 and 2.7.2. We proved in Temlyakov (2004) an estimate with the exponent $t/(2N^{1/2} + t)$ from Theorem 2.7.2 and with the constant N as in Theorem 2.7.1. Let s stand for either v or $s1$ or $s2$.

Theorem 2.7.3. Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k = t \in (0, 1]$, $k = 1, 2, \dots$. Then, for any vector of elements $f^1, \dots, f^N$, $f^i \in A_1(\mathcal{D})$, $i = 1, \dots, N$, we have

$$\sum_{i=1}^{N} \|f_m^{i,s,\tau}\|^2 \le N \left(1 + \frac{1}{N} m t^2\right)^{\frac{-t}{2N^{1/2}+t}}.$$

Theorem 2.7.4. Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume $\tau := \{t_k\}_{k=1}^{\infty}$ is a non-increasing sequence. Then, for any vector of elements

$f^1,\dots,f^N$, $f^i\in A_1(\mathcal{D})$, $i=1,\dots,N$, we have

$$\sum_{i=1}^{N}\|f_m^{i,s,\tau}\|^2\le CN\biggl(N+\sum_{k=1}^{m}t_k^2\biggr)^{\frac{-t_m}{2N^{1/2}+t_m}},$$

with an absolute constant $C=\mathrm{e}^{2/\mathrm{e}}<3$.

Let us make some comments on proofs of Theorems 2.7.1–2.7.4. The proof of Theorem 2.7.1 from Lutoborski and Temlyakov (2003) is a modification of the proof of Theorem 2.3.9 from Temlyakov (2000*b*) to the vector case. This proof does not use Theorem 2.3.9. The proof of Theorem 2.7.2 from Leviatan and Temlyakov (2006) directly uses Theorem 2.3.9. In Leviatan and Temlyakov (2006) we interpret a simultaneous approximation of $f^1,\dots,f^N$ in H with respect to $\mathcal{D}$ as an approximation of $F=(f^1,\dots,f^N)$ in $H_N:=H\times\cdots\times H$ with respect to a special dictionary $\mathcal{D}_N\subset H_N$ built from $\mathcal{D}$. The proof of Theorems 2.7.3 and 2.7.4 is more like the proof of Theorem 2.7.1. It is a modification of the proof of Theorem 2.3.9.

2.8. Application of greedy algorithms in learning theory

We discuss in this section some mathematical aspects of supervised learning theory. Supervised learning, or learning from examples, refers to a process that builds on the base of available data of inputs x_i and outputs y_i, $i=1,\dots,m$, a function that best represents the relation between the inputs $x\in X$ and the corresponding outputs $y\in Y$. The central question is how well this function estimates the outputs for general inputs. This is a big area of research both in non-parametric statistics and in learning theory.

A standard mathematical framework for the setting of the above learning problem is the following (Cucker and Smale 2001, Poggio and Smale 2003, DeVore, Kerkyacharian, Picard and Temlyakov 2004, 2006, Konyagin and Temlyakov 2004, 2007, Temlyakov 2005*c*, 2005*d*, 2006*d*). Let $X\subset\mathbb{R}^d$, $Y\subset\mathbb{R}$ be Borel sets, and let ρ be a Borel probability measure on $Z=X\times Y$. For $f:X\to Y$ define *the error*

$$\mathcal{E}(f):=\int_Z(f(x)-y)^2\,\mathrm{d}\rho.$$

Consider $\rho(y|x)$, the conditional (with respect to x) probability measure on Y, and ρ_X, the marginal probability measure on X (for $S\subset X$, $\rho_X(S)=\rho(S\times Y)$). Here we consider only bounded sets Y and, therefore, there exists a regular conditional probability $\rho(\cdot|x)$. Define $f_\rho(x)$ to be the conditional expectation of y with respect to measure $\rho(\cdot|x)$. The function f_ρ is known in statistics as the *regression function* of ρ. It is clear that if $f_\rho\in L_2(\rho_X)$ then it minimizes the error $\mathcal{E}(f)$ over all $f\in L_2(\rho_X)$: $\mathcal{E}(f_\rho)\le\mathcal{E}(f)$, $f\in L_2(\rho_X)$. Thus, in the sense of error $\mathcal{E}(\cdot)$, the regression function f_ρ is optimal to describe the relation between inputs $x \in X$ and outputs $y \in Y$. Now, our goal is to find an estimator $f_{\mathbf{z}}$, given data $\mathbf{z} = ((x_1, y_1), \dots, (x_m, y_m))$ that approximates f_ρ well with high probability. We assume that (x_i, y_i), $i = 1, \dots, m$ are independent and distributed according to ρ. We note that it is easy to see that, for any $f \in L_2(\rho_X)$,

$$\mathcal{E}(f) - \mathcal{E}(f_\rho) = \|f - f_\rho\|^2_{L_2(\rho_X)}.$$

The fundamental problem of learning theory is how to build a good estimator. It is well known in statistics that the following way of building $f_{\mathbf{z}}$ provides a near-optimal estimator in many cases. First, choose the right hypothesis space $\mathcal{H}$. Second, construct $f_{\mathbf{z},\mathcal{H}} \in \mathcal{H}$ as the empirical optimum (least squares estimator). We explain this in more detail. We define

$$f_{\mathbf{z},\mathcal{H}} = \arg\min_{f \in \mathcal{H}} \mathcal{E}_{\mathbf{z}}(f),$$

where

$$\mathcal{E}_{\mathbf{z}}(f) := \frac{1}{m} \sum_{i=1}^{m} (f(x_i) - y_i)^2$$

is the *empirical error* (*risk*) of f. This $f_{\mathbf{z},\mathcal{H}}$ is called the *empirical optimum* or the *Least Squares Estimator* (LSE). Clearly, a crucial role in this approach is played by a choice of the hypothesis space $\mathcal{H}$. In other words, we need to begin our construction of an estimator with a decision on what should be the form of the estimator. In this section we discuss only the case relevant to the use of nonlinear approximation, in particular, greedy approximation in such a construction. We want to construct a good estimator that will provide high accuracy and that will be practically implementable. We will discuss a realization of this plan in several stages. We begin with results on accuracy. We will give a presentation in a rather general form of nonlinear approximation.

Let $\mathcal{D}(n,q) := \{g_l^n\}_{l=1}^{N_n}$, $n \in \mathbb{N}$, $N_n \le n^q$, $q \ge 1$, be a system of bounded functions defined on X. We will consider a sequence $\{\mathcal{D}(n,q)\}_{n=1}^{\infty}$ of such systems. In building an estimator, based on $\mathcal{D}(n,q)$, we are going to use n-term approximations with regard to $\mathcal{D}(n,q)$:

$$G_\Lambda := \sum_{l \in \Lambda} c_l g_l^n, \quad |\Lambda| = n. \tag{2.8.1}$$

A standard assumption that we make in supervised learning theory is that $|y| \le M$ almost surely. This implies that we always assume that $|f_\rho| \le M$. Denoting $\|f\|_{B(X)} := \sup_{x \in X} |f(x)|$, we rewrite the above assumption in the form $\|f_\rho\|_{B(X)} \le M$. It is clear that with such an assumption it is natural to restrict our search to estimators $f_{\mathbf{z}}$ satisfying the same inequality $\|f_{\mathbf{z}}\|_{B(X)} \le M$. Now, in learning theory there are two standard ways to go. In the first approach, (I), we are looking for an estimator of the form (2.8.1) with an extra condition

$$\|G_\Lambda\|_{B(X)} \le M. \tag{2.8.2}$$

In the second approach, (II), we take an approximant G_Λ of the form (2.8.1) and truncate it, *i.e.*, consider $T_M(G_\Lambda)$, where T_M is a truncation operator: $T_M(u) = u$ if $|u| \le M$ and $T_M(u) = M \operatorname{sign} u$ if $|u| \ge M$. Then automatically $\|T_M(G_\Lambda)\|_{B(X)} \le M$.

Let us look in more detail at the hypothesis spaces generated in the above two cases. In case (I) we use the following compacts in $B(X)$ as a source of estimators:

$$F_n(q) := \left\{ f : \exists \Lambda \subset [1, N_n], |\Lambda| = n, f = \sum_{l\in\Lambda} c_l g_l^n, \|f\|_{B(X)} \le M \right\}.$$

An important feature of $F_n(q)$ is that it is a collection of sparse (at most n terms) estimators. An important drawback is that it may not be easy to check if (2.8.2) is satisfied for a particular G_Λ of the form (2.8.1).

In case (II) we use the following sets in $B(X)$ as a source of estimators:

$$F_n^T(q) := \left\{ f : \exists \Lambda \subset [1, N_n], |\Lambda| = n, f = T_M\left(\sum_{l\in\Lambda} c_l g_l^n \right) \right\}.$$

An obvious good feature of $F_n^T(q)$ is that by definition we have $\|f\|_{B(X)} \le M$ for any f from $F_n^T(q)$. An important drawback is that $F_n^T(q)$ has (in general) a rather complex structure. In particular, applying the truncation operator T_M to G_Λ we lose (in general) the sparseness property of G_Λ.

Now, when we have specified our hypothesis spaces, we can look for an existing theory that provides the corresponding error bounds. The general theory is well developed in case (I). We will use a variant of such a general theory developed in Temlyakov (2005*c*). This theory is based on the following property of compacts $F_n(q)$, formulated in terms of covering numbers:

$$N(F_n(q), \epsilon, B(X)) \le (1 + 2M/\epsilon)^n n^{qn}. \tag{2.8.3}$$

We now formulate the corresponding results from Temlyakov (2005*c*). For a compact Θ in a Banach space B we let $N(\Theta, \epsilon, B)$ denote the covering number, that is, the minimal number of balls of radius ϵ, with centres in Θ, needed to cover Θ. Let a and b be two positive numbers. Consider a collection $\mathcal{K}(a, b)$ of compact subsets K_n in $B(X)$ that are contained in the M-ball of $B(X)$ and satisfy the following covering numbers condition:

$$N(K_n, \epsilon, B(X)) \le (a(1 + 1/\epsilon))^n n^{bn}, \quad n = 1, 2, \ldots. \tag{2.8.4}$$

The following theorem was proved in Temlyakov (2005*c*). We begin with the definition of our estimator. As above, let $\mathcal{K} := \mathcal{K}(a, b)$ be a collection of compacts K_n in $B(X)$ satisfying (2.8.4).

We take a parameter $A \geq 1$ and consider the following Penalized Least Squares Estimator (PLSE):

$$f_{\mathbf{z}}^A := f_{\mathbf{z}}^A(\mathcal{K}) := f_{\mathbf{z},K_{n(\mathbf{z})}},$$

with

$$n(\mathbf{z}) := \arg \min_{1\leq j\leq m} \left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},K_j}) + \frac{Aj \ln m}{m} \right).$$

For a set L of a Banach space B, let

$$d(\Theta, L)_B := \sup_{f\in\Theta} \inf_{g\in L} \|f - g\|_B.$$

Theorem 2.8.1. For $\mathcal{K} := \{K_n\}_{n=1}^\infty$ satisfying (2.8.4) and $M > 0$, there exists $A_0 := A_0(a, b, M)$ such that, for any $A \geq A_0$ and any ρ such that $|y| \leq M$ a.s., we have

$$\|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)}^2 \leq \min_{1\leq j\leq m} \left(3d(f_\rho, K_j)_{L_2(\rho_X)}^2 + \frac{4Aj \ln m}{m} \right),$$

with probability $\geq 1 - m^{-c(M)A}$.

It is clear from (2.8.3) and from the definition of $F_n(q)$ that we can apply Theorem 2.8.1 to the sequence of compacts $\{F_n(q)\}$ and obtain the following error bound with probability $\geq 1 - m^{-c(M)A}$:

$$\|f_{\mathbf{z}}^A - f_\rho\|_{L_2(\rho_X)}^2 \leq \min_{1\leq j\leq m} \left(3d(f_\rho, F_j(q))_{L_2(\rho_X)}^2 + \frac{4Aj \ln m}{m} \right). \tag{2.8.5}$$

We note that inequality (2.8.5) is the Lebesgue-type inequality (see Section 2.6). Indeed, on the left-hand side of (2.8.5) we have an error of a particular estimator $f_{\mathbf{z}}^A$ built as the PLSE and on the right-hand side of (2.8.5) we have $d(f_\rho, F_j(q))_{L_2(\rho_X)}$: the best error that we can get using estimators from $F_j(q)$, $j = 1, 2, \dots$. We recall that by construction $f_{\mathbf{z}}^A \in F_{n(\mathbf{z})}(q)$.

Let us now discuss an application of the theory from Temlyakov (2005c) in case (II). We cannot apply that theory directly to the sequence of sets $\{F_n^T(q)\}$ because we do not know if these sets satisfy the covering number condition (2.8.4). However, we can modify the sets $F_n^T(q)$ to make them satisfy condition (2.8.4). Let $c \geq 0$ and define

$$\begin{aligned} F_n^T(q, c) := \Bigg\{ f : \exists G_\Lambda := \sum_{l\in\Lambda} c_l g_l^n, \Lambda \subset [1, N_n], |\Lambda| = n, \\ \|G_\Lambda\|_{B(X)} \leq C_2 n^c, f = T_M(G_\Lambda) \Bigg\} \end{aligned}$$

with some fixed $C_2 \geq 1$. Then, using the inequality

$$|T_M(f_1(x)) - T_M(f_2(x))| \leq |f_1(x) - f_2(x)|, \quad \text{for } x \in X,$$

it is easy to get that

$$N(F_n^T(q,c),\epsilon,B(X)) \le (2C_2(1+1/\epsilon))^n n^{(q+c)n}.$$

Therefore, (2.8.4) is satisfied with $a = 2C_2$ and $b = q + c$. We note that, from a practical point of view, an extra restriction $\|G_\Lambda\|_{B(X)} \le C_2 n^c$ is not a big constraint.

The above estimators (built as the PLSE) are very good from the theoretical point of view. Their error bounds satisfy Lebesgue-type inequalities. However, they are not good from the point of view of implementation. For example, there is no simple algorithm to find $f_{\mathbf{z},F_n(q)}$ because $F_n(q)$ is a union of $\binom{N_n}{n}$ M-balls of n-dimensional subspaces. Thus, finding an exact LSE $f_{\mathbf{z},F_n(q)}$ is practically impossible. We now use a remark from Temlyakov (2005*c*) that allows us to build an approximate LSE with good approximation error. We proceed to the definition of the Penalized Approximate Least Squares Estimator (PALSE) (see Temlyakov (2005*c*)). Let $\delta := \{\delta_{j,m}\}_{j=1}^m$ be a sequence of non-negative numbers. We define $f_{\mathbf{z},\delta,K_j}$ as an estimator satisfying the relation

$$\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\delta,K_j}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},K_j}) + \delta_{j,m}. \tag{2.8.6}$$

In other words, $f_{\mathbf{z},\delta,K_j}$ is an approximation to the least squares estimator $f_{\mathbf{z},K_j}$.

Next, we take a parameter $A \ge 1$ and define the Penalized Approximate Least Squares Estimator (PALSE)

$$f^A_{\mathbf{z},\delta} := f^A_{\mathbf{z},\delta}(\mathcal{K}) := f_{\mathbf{z},\delta,K_{n(\mathbf{z})}},$$

with

$$n(\mathbf{z}) := \arg\min_{1\le j\le m}\left(\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\delta,K_j}) + \frac{Aj\ln m}{m}\right).$$

The theory developed in Temlyakov (2005*c*) gives the following error estimate.

Theorem 2.8.2. Under the assumptions of Theorem 2.8.1 we have

$$\|f^A_{\mathbf{z},\delta} - f_\rho\|^2_{L_2(\rho_X)} \le \min_{1\le j\le m}\left(3d(f_\rho,K_j)^2_{L_2(\rho_X)} + \frac{4Aj\ln m}{m} + 2\delta_{j,m}\right),$$

with probability $\ge 1 - m^{-c(M)A}$.

We point out here that the approximate least squares estimator $f_{\mathbf{z},\delta,K_j}$ approximates the least squares estimator $f_{\mathbf{z},K_j}$ in the sense that $\mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},\delta,K_j}) - \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},K_j})$ is small, and not in the sense that $\|f_{\mathbf{z},\delta,K_j} - f_{\mathbf{z},K_j}\|$ is small. Theorem 2.8.2 guarantees a good error bound for any penalized estimator built from $\{f_{\mathbf{z},\delta,K_j}\}$ satisfying (2.8.6). We will use greedy algorithms in building an approximate estimator. We now present results from Temlyakov (2005*d*).

We will need more specific compacts $F(n,q)$ and will impose some restrictions on g_l^n. We assume that $\|g_l^n\|_{B(X)} \le C_1$ for all n and l. We consider the following compacts instead of $F_n(q)$:

$$F(n,q) := \left\{ f : \exists \Lambda \subset [1,N_n], |\Lambda| = n, f = \sum_{l\in\Lambda} c_l g_l^n, \sum_{l\in\Lambda} |c_l| \le 1 \right\}.$$

Then we have $\|f\|_{B(X)} \le C_1$ for any $f \in F(n,q)$ and $\|f\|_{B(X)} \le M$ if $M \ge C_1$. Let $\mathbf{z} = (z_1,\dots,z_m)$, $z_i = (x_i,y_i)$, be given. Consider the following system of vectors in $\mathbb{R}^m$:

$$v^{j,l} := (g_l^j(x_1),\dots,g_l^j(x_m)), \quad l \in [1,N_j].$$

We equip the $\mathbb{R}^m$ with the norm $\|v\| := (m^{-1}\sum_{i=1}^m v_i^2)^{1/2}$. Then

$$\|v^{j,l}\| \le \|g_l^j\|_{B(X)} \le C_1.$$

Consider the system $\mathcal{G} := \{v^{j,l}\}_{l=1}^{N_j}$ in $H = \mathbb{R}^m$ with the norm $\|\cdot\|$ defined above. Finding the estimator

$$f_{\mathbf{z},F(j,q)} = \sum_{l\in\Lambda} c_l g_l^j, \quad \sum_{l\in\Lambda} |c_l| \le 1, \quad |\Lambda| = j, \quad \Lambda \subset [1,N_j],$$

is equivalent to finding best j-term approximant of $y \in \mathbb{R}^m$ from the $A_1(\mathcal{G})$ in the space H. We apply the RGA(θ) from Section 2.4 with $\theta = 2$ with respect to $\mathcal{G}$ to y and find, after j steps, an approximant

$$v^j := \sum_{l\in\Lambda'} a_l v^{j,l}, \quad \sum_{l\in\Lambda'} |a_l| \le 1, \quad |\Lambda'| = j, \quad \Lambda' \subset [1,N_j],$$

such that

$$\|y - v^j\|^2 \le d(y, A_1(\mathcal{G}))^2 + Cj^{-1}, \quad C = C(M,C_1).$$

We define an estimator

$$\hat{f}_{\mathbf{z}} := \hat{f}_{\mathbf{z},F(j,q)} := \sum_{l\in\Lambda'} a_l g_l^j.$$

Then $\hat{f}_{\mathbf{z}} \in F(j,q)$ and

$$\mathcal{E}_{\mathbf{z}}(\hat{f}_{\mathbf{z},F(j,q)}) \le \mathcal{E}_{\mathbf{z}}(f_{\mathbf{z},F(j,q)}) + Cj^{-1}.$$

We let $\delta := \{Cj^{-1}\}_{j=1}^m$, and define for $A \ge 1$

$$f_{\mathbf{z},\delta}^A := \hat{f}_{\mathbf{z},F(n(\mathbf{z}),q)},$$

with

$$n(\mathbf{z}) := \arg\min_{1\le j\le m} \left(\mathcal{E}_{\mathbf{z}}(\hat{f}_{\mathbf{z},F(j,q)}) + \frac{Aj\ln m}{m} \right).$$

By Theorem 2.8.2 we have for $A \geq A_0(M)$

$$\|f^A_{\mathbf{z},\delta} - f_\rho\|^2_{L_2(\rho_X)} \leq \min_{1\leq j\leq m}\left(3d(f_\rho, F(j,q))^2 + \frac{4Aj\ln m}{m} + 2Cj^{-1}\right) \quad (2.8.7)$$

with probability $\geq 1 - m^{-c(M)A}$.

In particular, (2.8.7) means that the estimator $f^A_{\mathbf{z},\delta}$ is an estimator that provides the error

$$\|f^A_{\mathbf{z},\delta} - f_\rho\|^2_{L_2(\rho_X)} \ll \left(\frac{\ln m}{m}\right)^{\frac{2r}{1+2r}}$$

for f_ρ such that $d(f_\rho, F(j,q))_{L_2(\rho_X)} \ll j^{-r}$, $r \leq 1/2$. We note that the estimator $f^A_{\mathbf{z},\delta}$ is based on the greedy algorithm and it can easily be implemented.

We now describe an application of greedy algorithm in learning theory from Barron, Cohen, Dahmen and DeVore (2005). In this application one can use the Orthogonal Greedy Algorithm or the following variant of the Relaxed Greedy Algorithm.

Let $\alpha_1 := 0$ and $\alpha_m := 1 - 2/m$, $m \geq 2$. We set $f_0 := f$, $G_0 := 0$ and inductively define two sequences $\{\beta_m\}_{m=1}^\infty$, $\{\varphi_m\}_{m=1}^\infty$ as follows:

$$(\beta_m, \varphi_m) := \arg\min_{\beta\in\mathbb{R}, g\in\mathcal{D}} \|f - (\alpha_m G_{m-1} + \beta g)\|.$$

Then we set

$$f_m := f_{m-1} - \beta_m\varphi_m, \quad G_m := G_{m-1} + \beta_m\varphi_m.$$

For systems $\mathcal{D}(n,q)$ the following estimator is considered in Barron *et al.* (2005). As above, let $\mathbf{z} = (z_1, \dots, z_m)$, $z_i = (x_i, y_i)$, be given. Consider the following system of vectors in $\mathbb{R}^m$:

$$v^{j,l} := (g^j_l(x_1), \dots, g^j_l(x_m)), \quad l \in [1, N_j].$$

We equip the $\mathbb{R}^m$ with the norm $\|v\| := (m^{-1}\sum_{i=1}^m v_i^2)^{1/2}$ and normalize the above system of vectors. Denote the new system of vectors by $\mathcal{G}_j$. Now we apply either the OGA or the version of the RGA defined above to the vector $y \in \mathbb{R}$ with respect to the system $\mathcal{G}_j$. As in the case discussed above of the system $\mathcal{G}$, we obtain an estimator $\hat{f}_j$. Next, we look for the penalized estimator built from the estimators $\{\hat{f}_j\}$ in the following way. Let

$$n(\mathbf{z}) := \arg\min_{1\leq j\leq m}\left(\mathcal{E}_{\mathbf{z}}(T_M(\hat{f}_j)) + \frac{Aj\log m}{m}\right).$$

Define

$$\hat{f} := T_M(\hat{f}_{n(\mathbf{z})}).$$

Assuming that the systems $\mathcal{D}(n,q)$ are normalized in $L_2(\rho_X)$, Barron *et al.* (2005) proved the following error estimate.

Theorem 2.8.3. There exists $A_0(M)$ such that for $A \geq A_0$ we have the following bound for the expectation of the error:

$$E(\|f_\rho - \hat{f}\|^2_{L_2(\rho_X)}) \leq \min_{1\leq j\leq m}(C(A,M,q)j\log m/m \\ + \inf_{h\in \operatorname{span}\mathcal{D}(j,q)}(2\|f_\rho - h\|^2_{L_2(\rho_X)} + 8\|h\|^2_{\mathcal{A}_1(\mathcal{D}(j,q))}/j)). \tag{2.8.8}$$

Let us make a comparison of (2.8.8) with (2.8.7). First of all, (2.8.8) gives an error bound for the expectation and (2.8.7) gives an error bound with high probability. In this sense (2.8.7) is better than (2.8.8). However, the condition $\|g^n_l\|_{B(X)} \leq C_1$ imposed on the systems $\mathcal{D}(n,q)$ in order to obtain (2.8.7) is more restrictive than the corresponding assumption for (2.8.8).

2.9. A remark on compressed sensing

Recently, compressed sensing (compressive sampling) has attracted a lot of attention from both mathematicians and computer scientists. Compressed sensing refers to a problem of *economical* recovery of an unknown vector $u \in \mathbb{R}^m$ from the information provided by linear measurements $\langle u,\varphi_j\rangle$, $\varphi_j \in \mathbb{R}^m$, $j = 1,\dots,n$. The goal is to design an algorithm that finds (approximates) u from the information $y = (\langle u,\varphi_1\rangle,\dots,\langle u,\varphi_n\rangle) \in \mathbb{R}^n$. The crucial step here is to build a *sensing* set of vectors $\varphi_j \in \mathbb{R}^m$, $j = 1,\dots,n$ that is *good* for all vectors $u \in \mathbb{R}^m$. Clearly, the terms *economical* and *good* should be clarified in a mathematical setting of the problem. A natural variant of such a setting, which we discuss here, uses the concept of *sparsity.* We call a vector $u \in \mathbb{R}^m$ k-sparse if it has at most k non-zero coordinates. Now, for a given pair (m,n) we want to understand what is the biggest sparsity $k(m,n)$ such that there exists a set of vectors $\varphi_j \in \mathbb{R}^m$, $j = 1,\dots,n$ and economical algorithm A mapping y into $\mathbb{R}^m$ in such a way that, for any u of sparsity $k(m,n)$, one would have an exact recovery $A(u) = u$. In other words, we want to describe matrices Φ with rows $\varphi_j \in \mathbb{R}^m$, $j = 1,\dots,n$, such that there exists an economical algorithm of solving the following sparse recovery problem.

The sparse recovery problem can be stated as the problem of finding the sparsest vector $u^0 := u^0_\Phi(y) \in \mathbb{R}^m$:

$$\min \|v\|_0 \quad \text{subject to } \Phi v = y, \tag*{(P_0)}$$

where $\|v\|_0 := |\operatorname{supp}(v)|$. D. Donoho and co-authors (see, for instance, Chen, Donoho and Saunders (2001), Donoho *et al.* (2006) and the history therein) have suggested an economical algorithm (Basis Pursuit) and have begun a systematic study of the following question. For which measurement matrices Φ should the highly non-convex combinatorial optimization problem (P_0) be equivalent to its convex relaxation problem

$$\min \|v\|_1 \quad \text{subject to} \quad \Phi v = y, \tag{P_1}$$

where $\|v\|_1$ denotes the ℓ_1-norm of the vector $v \in \mathbb{R}^m$? Denote the solution to (P_1) by $A_\Phi(y)$. It is known that the problem (P_1) can be solved by linear programming techniques. The ℓ_1-minimization algorithm A_Φ from (P_1) is an economical algorithm that we consider in this section. It is known (see, for instance, Donoho *et al.* (2006)) that for M-coherent matrices Φ we have $u^0_\Phi(\Phi u) = A_\Phi(\Phi u) = u$, provided u is k-sparse with $k < (1 + 1/M)/2$. This allows us to build rather simple deterministic matrices Φ with $k(m, n) \asymp n^{1/2}$ and recover A_Φ from (P_1) with the ℓ_1-minimization algorithm.

Recent progress (see surveys by Candès (2006) and DeVore (2006)) in compressed sensing has resulted in proving the existence of matrices Φ with $k(m, n) \asymp n/\log(m/n)$, which is substantially larger than $n^{1/2}$. We proceed to a detailed discussion of these recent results.

We begin with results from Donoho (2006). Donoho formulated the following three properties of matrices Φ with ℓ_2-normalized columns, and proved the existence of matrices satisfying these conditions. Let T be a subset of indices from $[1, m]$. Let Φ_T denote a matrix consisting of columns of Φ with indices from T.

CS1 The minimal singular value of Φ_T is $\geq \eta_1 > 0$ uniformly in T, satisfying $|T| \leq \rho n/\log m$.

CS2 Let W_T denote the range of Φ_T. Assume that for any T satisfying $|T| \leq \rho n/\log m$, we have

$$\|w\|_1 \geq \eta_2 n^{1/2} \|w\|_2, \quad \forall w \in W_T, \quad \eta_2 > 0.$$

CS3 Denote $T^c := \{j\}_{j=1}^m \setminus T$. For any T, $|T| \leq \rho n/\log m$ and for any $w \in W_T$, we have for any v satisfying $\Phi_{T^c} v = w$

$$\|v\|_{\ell_1(T^c)} \geq \eta_3 (\log(m/n))^{-1/2} \|w\|_1, \quad \eta_3 > 0.$$

It is proved in Donoho (2006) that if Φ satisfies CS1–CS3, then there exists $\rho_0 > 0$ such that $u^0_\Phi(\Phi u) = A_\Phi(\Phi u) = u$ provided $|\operatorname{supp} u| \leq \rho_0 n/\log m$. Analysis in Donoho (2006) relates the compressed sensing problem to the problem of estimating the Kolmogorov widths and their dual, the Gel'fand widths.

We give the corresponding definitions. For a compact $F \subset \mathbb{R}^m$, the Kolmogorov width is given by

$$d_n(F, \ell_p) := \inf_{L_n : \dim L_n \leq n} \sup_{f \in F} \inf_{a \in L_n} \|f - a\|_p,$$

where L_n is a linear subspace of $\mathbb{R}^m$ and $\|\cdot\|_p$ denotes the ℓ_p-norm. The

Gel'fand width is defined by

$$d^n(F,\ell_p) := \inf_{V_n} \sup_{f\in F\cap V_n} \|f\|_p,$$

where the infimum is taken over linear subspaces V_n with dimension $\geq m-n$. It is well known that the Kolmogorov and the Gel'fand widths are related by the duality formula. For instance, when $F = B_p^m$ is a unit ℓ_p-ball in $\mathbb{R}^m$ and $1 \leq q,p \leq \infty$, we have

$$d_n(B_p^m,\ell_q) = d^n(B_{q'}^m,\ell_{p'}), \quad p' := p/(p-1). \tag{2.9.1}$$

In the particular case $p = 2$, $q = \infty$ of our interest, (2.9.1) gives

$$d_n(B_2^m,\ell_\infty) = d^n(B_1^m,\ell_2). \tag{2.9.2}$$

It has been established in approximation theory (see Kashin (1977) and Garnaev and Gluskin (1984)) that

$$d_n(B_2^m,\ell_\infty) \leq C((1+\log(m/n))/n)^{1/2}. \tag{2.9.3}$$

In other words, it was proved (see (2.9.3) and (2.9.2)) that for any pair (m,n) there exists a subspace V_n, $\dim V_n \geq m-n$ such that, for any $x \in V_n$, we have

$$\|x\|_2 \leq C((1+\log(m/n))/n)^{1/2}\|x\|_1. \tag{2.9.4}$$

It was understood in Donoho (2006) that properties of the null space $\mathcal{N}(\Phi) := \{x : \Phi x = 0\}$ of a measurement matrix Φ play an important role in the compressed sensing problem. Donoho (2006) introduced the following two characteristics of Φ formulated in terms of $\mathcal{N}(\Phi)$:

$$w(\Phi,F) := \sup_{x\in F\cap\mathcal{N}(\Phi)} \|x\|_2$$

and

$$\nu(\Phi,T) := \sup_{x\in\mathcal{N}(\Phi)} \|x_T\|_1/\|x\|_1,$$

where x_T is a restriction of x onto T: $(x_T)_j = x_j$ for $j \in T$ and $(x_T)_j = 0$ otherwise. He proved that if Φ obeys the following two conditions,

$$\nu(\Phi,T) \leq \eta_1, \quad |T| \leq \rho_1 n/\log m, \tag{A1}$$

$$w(\Phi,B_1^m) \leq \eta_2((\log m)/n)^{1/2}, \tag{A2}$$

then for any $u \in B_1^m$ we have

$$\|u - A_\Phi(\Phi u)\|_2 \leq C((\log m)/n)^{1/2}.$$

We now proceed to the contribution of E. Candès, J. Romberg and T. Tao published in a series of papers (see Candès and Tao (2005)). They introduced the following Restricted Isometry Property (RIP) of a sensing matrix Φ: $\delta_S < 1$ is the S-restricted isometry constant of Φ if it is the smallest quantity such that

$$(1-\delta_S)\|c\|_2^2 \le \|\Phi_T c\|_2^2 \le (1+\delta_S)\|c\|_2^2$$

for all subsets T with $|T| \le S$ and all coefficient sequences $\{c_j\}_{j\in T}$. Candès and Tao (2005) proved that if $\delta_{2S} + \delta_{3S} < 1$, then for S-sparse u we have $A_\Phi(\Phi u) = u$ (recovery by ℓ_1-minimization is exact). They also proved existence of sensing matrices Φ obeying the condition $\delta_{2S} + \delta_{3S} < 1$ for large values of sparsity $S \asymp n/\log(m/n)$. For a positive number a denote

$$\sigma_a(v)_1 := \min_{w\in\mathbb{R}^m:|\operatorname{supp}(w)|\le a} \|v-w\|_1.$$

Candès, Romberg and Tao (2006) proved that if $\delta_{3S} + 3\delta_{4S} < 2$, then

$$\|u - A_\Phi(\Phi u)\|_2 \le CS^{-1/2}\sigma_S(u)_1. \tag{2.9.5}$$

We note that properties of the RIP-type matrices have already been employed in Kashin (1977) (see Kashin and Temlyakov (2007) for further discussion) for the widths estimation. The inequality (2.9.3) with an extra factor $(1+\log m/n)$ was established in Kashin (1977). The proof in Kashin (1977) is based on properties of a random matrix Φ with elements $\pm 1/\sqrt{n}$.

Further investigation of the compressed sensing problem was conducted by Cohen, Dahmen and DeVore (2007). They proved that if Φ satisfies the RIP of order $2k$ with $\delta_{2k} < \delta < 1/3$, then

$$\|u - A_\Phi(\Phi u)\|_1 \le \frac{2+2\delta}{1-3\delta}\sigma_k(u)_1. \tag{2.9.6}$$

The above inequality is the Lebesgue-type inequality (see Section 2.6) for the approximation method $u \to A_\Phi(\Phi u)$. In Cohen *et al.* (2007) the inequality (2.9.6) was called *instance optimality.* In the proof of (2.9.6) the authors used the following property (null space property) of matrices Φ satisfying the RIP of order $3k/2$: for any $x \in \mathcal{N}(\Phi)$ and any T with $|T| \le k$, we have

$$\|x\|_1 \le C\|x_{T^c}\|_1. \tag{2.9.7}$$

The null space property (2.9.7) is closely related to the property (A1) from Donoho (2006). The proof of (2.9.6) from Cohen *et al.* (2007) gives an inequality similar to (2.9.6) under the assumption that Φ has the null space property (2.9.7) with $C < 2$.

We now discuss results of Kashin and Temlyakov (2007). We say that a measurement matrix Φ has a Strong Compressed Sensing Property (SCSP) if, for any $u \in \mathbb{R}^m$, we have

$$\|u - A_\Phi(\Phi u)\|_2 \le Ck^{-1/2}\sigma_k(u)_1, \tag{2.9.8}$$

for $k \asymp n/\log(m/n)$. We define a Weak Compressed Sensing Property (WCSP) by replacing (2.9.8) by the weaker inequality

$$\|u - A_\Phi(\Phi u)\|_2 \le Ck^{-1/2}\|u\|_1. \tag{2.9.9}$$

We say that Φ satisfies the Width Property (WP) if (2.9.4) holds for the null space $\mathcal{N}(\Phi)$. The main result of the paper Kashin and Temlyakov (2007) states that the above three properties of Φ are equivalent. We proceed to a detailed discussion of results from Kashin and Temlyakov (2007).

We mentioned above that it is known that, for any pair (m, n), $n < m$, there exists a subspace $\Gamma \subset \mathbb{R}^m$ with $\dim \Gamma \ge m - n$ such that

$$\|x\|_2 \le Cn^{-1/2}(\ln(em/n))^{1/2}\|x\|_1, \quad \forall x \in \Gamma. \tag{2.9.10}$$

We will discuss some properties of subspaces Γ satisfying (2.9.10) that are useful in compressed sensing. Let

$$S := S(m,n) := C^{-2}n(\ln(em/n))^{-1}.$$

For $x = (x_1, \dots, x_m) \in \mathbb{R}^m$, define $\operatorname{supp}(x) := \{j : x_j \ne 0\}$.

Lemma 2.9.1. Let Γ satisfy (2.9.10) and $x \in \Gamma$. Then either $x = 0$ or $|\operatorname{supp}(x)| \ge S(m,n)$.

Proof. Assume $x \ne 0$. Then $\|x\|_1 > 0$. Denote $\Lambda := \operatorname{supp}(x)$. We have

$$\|x\|_1 = \sum_{j\in\Lambda} |x_j| \le |\Lambda|^{1/2}\Bigl(\sum_{j\in\Lambda}|x_j|^2\Bigr)^{1/2} \le |\Lambda|^{1/2}\|x\|_2. \tag{2.9.11}$$

Using (2.9.10), we get from (2.9.11)

$$\|x\|_1 \le |\Lambda|^{1/2}S(m,n)^{-1/2}\|x\|_1.$$

Thus

$$|\Lambda| \ge S(m,n). \qquad \square$$

Lemma 2.9.2. Let Γ satisfy (2.9.10) and let $x \ne 0$, $x \in \Gamma$. Then, for any Λ such that $|\Lambda| < S(m,n)/4$,

$$\sum_{j\in\Lambda} |x_j| < \|x\|_1/2.$$

Proof. As in (2.9.11),

$$\sum_{j\in\Lambda}|x_j| \le |\Lambda|^{1/2}S(m,n)^{-1/2}\|x\|_1 < \|x\|_1/2. \qquad \square$$

Lemma 2.9.3. Let Γ satisfy (2.9.10). Suppose $u \in \mathbb{R}^m$ is sparse with $|\operatorname{supp}(u)| < S(m,n)/4$. Then, for any $v = u + x$, $x \in \Gamma$, $x \ne 0$,

$$\|v\|_1 > \|u\|_1.$$

Proof. Let $\Lambda := \operatorname{supp}(u)$. Then

$$\begin{aligned}\|v\|_1 = \sum_{j\in[1,m]} |v_j| &= \sum_{j\in\Lambda} |u_j + x_j| + \sum_{j\notin\Lambda} |x_j| \\ &\geq \sum_{j\in\Lambda} |u_j| - \sum_{j\in\Lambda} |x_j| + \sum_{j\notin\Lambda} |x_j| = \|u\|_1 + \|x\|_1 - 2\sum_{j\in\Lambda} |x_j|.\end{aligned}$$

By Lemma 2.9.2,

$$\|x\|_1 - 2\sum_{j\in\Lambda} |x_j| > 0. \qquad \square$$

Lemma 2.9.3 guarantees that the following algorithm, known as the Basis Pursuit (see A_Φ defined above), will find a sparse u exactly, provided $|\operatorname{supp}(u)| < S(m,n)/4$:

$$u_\Gamma := u + \arg\min_{x\in\Gamma} \|u + x\|_1.$$

Theorem 2.9.4. Let Γ satisfy (2.9.10). Then, for any $u \in \mathbb{R}^m$ and u' such that $\|u'\|_1 \leq \|u\|_1$, $u - u' \in \Gamma$,

$$\|u - u'\|_1 \leq 4\sigma_{S/16}(u)_1, \tag{2.9.12}$$

$$\|u - u'\|_2 \leq (S/16)^{-1/2}\sigma_{S/16}(u)_1. \tag{2.9.13}$$

Proof. It is given that $u - u' \in \Gamma$. Thus, (2.9.13) follows from (2.9.12) and (2.9.10). We now prove (2.9.12). Let Λ, $|\Lambda| = [S/16]$, be the set of indices of coordinates of u that are largest in absolute value. Let u_Λ denote the restriction of u onto this set, *i.e.*, $(u_\Lambda)_j = u_j$ for $j \in \Lambda$ and $(u_\Lambda)_j = 0$ for $j \notin \Lambda$, and let $u^\Lambda := u - u_\Lambda$. Then

$$\sigma_{S/16}(u)_1 = \sigma_{|\Lambda|}(u)_1 = \|u - u_\Lambda\|_1 = \|u^\Lambda\|_1. \tag{2.9.14}$$

We have

$$\|u - u'\|_1 \leq \|(u - u')_\Lambda\|_1 + \|(u - u')^\Lambda\|_1.$$

Next,

$$\|(u - u')^\Lambda\|_1 \leq \|u^\Lambda\|_1 + \|(u')^\Lambda\|_1.$$

Using $\|u'\|_1 \leq \|u\|_1$, we obtain

$$\|(u')^\Lambda\|_1 - \|u^\Lambda\|_1 = \|u'\|_1 - \|u\|_1 - \|u'_\Lambda\|_1 + \|u_\Lambda\|_1 \leq \|(u - u')_\Lambda\|_1.$$

Therefore,

$$\|(u')^\Lambda\|_1 \leq \|u^\Lambda\|_1 + \|(u - u')_\Lambda\|_1$$

and

$$\|u-u'\|_1 \le 2\|(u-u')_\Lambda\|_1 + 2\|u^\Lambda\|_1. \tag{2.9.15}$$

Using the fact $u-u' \in \Gamma$, we estimate

$$\begin{aligned}\|(u-u')_\Lambda\|_1 &\le |\Lambda|^{1/2}\|(u-u')_\Lambda\|_2 \le |\Lambda|^{1/2}\|u-u'\|_2\\ &\le |\Lambda|^{1/2}S^{-1/2}\|u-u'\|_1.\end{aligned} \tag{2.9.16}$$

Our assumption on $|\Lambda|$ guarantees that $|\Lambda|^{1/2}S^{-1/2} \le 1/4$. Using this and substituting (2.9.16) into (2.9.15), we obtain

$$\|u-u'\|_1 \le \|u-u'\|_1/2 + 2\|u^\Lambda\|_1,$$

which gives (2.9.12):

$$\|u-u'\|_1 \le 4\|u^\Lambda\|_1. \qquad \square$$

Corollary 2.9.5. Let Γ satisfy (2.9.10). Then, for any $u \in \mathbb{R}^m$,

$$\|u-u_\Gamma\|_1 \le 4\sigma_{S/16}(u)_1, \tag{2.9.17}$$

$$\|u-u_\Gamma\|_2 \le (S/16)^{-1/2}\sigma_{S/16}(u)_1. \tag{2.9.18}$$

Proposition 2.9.6. Let Γ be such that (2.9.9) holds with u_Γ instead of $A_\Phi(\Phi u)$ and $k = n/\ln(em/n)$. Then Γ satisfies (2.9.10).

Proof. Let $u \in \Gamma$. Then $u_\Gamma = 0$, and we get from (2.9.9)

$$\|u\|_2 \le C(n/\ln(em/n))^{-1/2}\|u\|_1. \qquad \square$$

Theorem 2.9.7. The following three properties of Φ are equivalent: the Strong Compressed Sensing Property, the Weak Compressed Sensing Property, and the Width Property.

Proof. It is obvious that SCSP ⇒ WCSP. Corollary 2.9.5 with $\Gamma = \mathcal{N}(\Phi)$ implies that WP ⇒ SCSP. Proposition 2.9.6 with $\Gamma = \mathcal{N}(\Phi)$ implies that WCSP ⇒ WP. Thus the three properties are equivalent. □

The result (2.9.5) of Candès *et al.* (2006) states that the RIP with $S \asymp n/\log(m/n)$ implies the SCSP. Therefore, by Theorem 2.9.7 it implies the WP.

We note that there are very interesting results on greedy approximation in compressed sensing. We do not discuss these results here, and refer the reader to two of them: Tropp and Gilbert (2005) and Needell and Vershynin (2007).

CHAPTER THREE
Greedy approximation with respect to dictionaries: Banach spaces

3.1. Introduction

In this chapter we make a step from Hilbert spaces to more general Banach spaces. Let X be a Banach space with norm $\|\cdot\|$. We say that a set of elements (functions) $\mathcal{D}$ from X is a dictionary, respectively, symmetric dictionary, if each $g \in \mathcal{D}$ has norm bounded by one ($\|g\| \le 1$),

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

and the closure of $\operatorname{span} \mathcal{D}$ is X. We denote the closure (in X) of the convex hull of $\mathcal{D}$ by $A_1(\mathcal{D})$. We introduce a new norm, associated with a dictionary $\mathcal{D}$, in the dual space X^* by the formula

$$\|F\|_{\mathcal{D}} := \sup_{g\in\mathcal{D}} F(g), \quad F \in X^*.$$

In this chapter we will study greedy algorithms with regard to $\mathcal{D}$. For a non-zero element $f \in X$ we let F_f denote a norming (peak) functional for f:

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

The existence of such a functional is guaranteed by Hahn–Banach theorem.

We begin with a generalization of the Pure Greedy Algorithm. The greedy step of the PGA can be interpreted in two ways. First, we look at the mth step for an element $\varphi_m \in \mathcal{D}$ and a number λ_m satisfying

$$\|f_{m-1} - \lambda_m\varphi_m\|_H = \inf_{g\in\mathcal{D},\lambda} \|f_{m-1} - \lambda g\|_H. \tag{3.1.1}$$

Second, we look for an element $\varphi_m \in \mathcal{D}$ such that

$$\langle f_{m-1}, \varphi_m\rangle = \sup_{g\in\mathcal{D}} \langle f_{m-1}, g\rangle. \tag{3.1.2}$$

In a Hilbert space both versions (3.1.1) and (3.1.2) resulted in the same PGA. In a general Banach space the corresponding versions of (3.1.1) and (3.1.2) lead to different greedy algorithms. The Banach space version of (3.1.1) is straightforward: instead of the Hilbert norm $\|\cdot\|_H$ in (3.1.1) we use the Banach norm $\|\cdot\|_X$. This results in the following greedy algorithm (see Temlyakov (2003*a*)).

***X*-Greedy Algorithm (XGA).** We define $f_0 := f$, $G_0 := 0$. Then, for each $m \ge 1$ we have the following inductive definition.

(1) $\varphi_m \in \mathcal{D}$, $\lambda_m \in \mathbb{R}$ are such that (we assume existence)

$$\|f_{m-1} - \lambda_m\varphi_m\|_X = \inf_{g\in\mathcal{D},\lambda} \|f_{m-1} - \lambda g\|_X. \tag{3.1.3}$$

(2) Define

$$f_m := f_{m-1} - \lambda_m \varphi_m, \qquad G_m := G_{m-1} + \lambda_m \varphi_m.$$

The second version of the PGA in a Banach space is based on the concept of a norming (peak) functional. We note that in a Hilbert space a norming functional F_f acts as follows:

$$F_f(g) = \langle f/\|f\|, g\rangle.$$

Thus, (3.1.2) can be rewritten in terms of the norming functional $F_{f_{m-1}}$ as

$$F_{f_{m-1}}(\varphi_m) = \sup_{g\in\mathcal{D}} F_{f_{m-1}}(g). \tag{3.1.4}$$

This observation leads to the class of dual greedy algorithms. We define the Weak Dual Greedy Algorithm with weakness τ (WDGA(τ)) (see Dilworth, Kutzarova and Temlyakov (2002) and Temlyakov (2003*a*)) that is a generalization of the Weak Greedy Algorithm.

Weak Dual Greedy Algorithm (WDGA($\boldsymbol{\tau}$)). Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$. Then, for each $m \geq 1$ we have the following inductive definition.

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

$$F_{f_{m-1}}(\varphi_m) \geq t_m \|F_{f_{m-1}}\|_{\mathcal{D}}. \tag{3.1.5}$$

(2) Define a_m as

$$\|f_{m-1} - a_m\varphi_m\| = \min_{a\in\mathbb{R}} \|f_{m-1} - a\varphi_m\|.$$

(3) Let

$$f_m := f_{m-1} - a_m\varphi_m.$$

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

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

The uniformly smooth Banach space is the one with the property

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

It is easy to see that for any Banach space X its modulus of smoothness $\rho(u)$ is an even convex function satisfying the inequalities

$$\max(0, u-1) \leq \rho(u) \leq u, \quad u \in (0,\infty).$$

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

Lemma 3.1.1. Let $x \neq 0$. Then

$$0 \leq \|x + uy\| - \|x\| - uF_x(y) \leq 2\|x\|\rho(u\|y\|/\|x\|). \tag{3.1.6}$$

Proof. We have

$$\|x + uy\| \geq F_x(x + uy) = \|x\| + uF_x(y).$$

This proves the first inequality. Next, from the definition of modulus of smoothness it follows that

$$\|x + uy\| + \|x - uy\| \leq 2\|x\|(1 + \rho(u\|y\|/\|x\|)). \tag{3.1.7}$$

Also,

$$\|x - uy\| \geq F_x(x - uy) = \|x\| - uF_x(y). \tag{3.1.8}$$

Combining (3.1.7) and (3.1.8), we obtain

$$\|x + uy\| \leq \|x\| + uF_x(y) + 2\|x\|\rho(u\|y\|/\|x\|).$$

This proves the second inequality. □

Proposition 3.1.2. Let X be a uniformly smooth Banach space. Then, for any $x \neq 0$ and y we have

$$F_x(y) = \left(\frac{\mathrm{d}}{\mathrm{d}u}\|x + uy\|\right)(0) = \lim_{u \to 0}(\|x + uy\| - \|x\|)/u. \tag{3.1.9}$$

Proof. The equality (3.1.9) follows from (3.1.6) and the property that, for a uniformly smooth Banach space, $\lim_{u \to 0} \rho(u)/u = 0$. □

Proposition 3.1.2 shows that in the WDGA we are looking for an element $\varphi_m \in \mathcal{D}$ that provides a big derivative of the quantity $\|f_{m-1} + ug\|$. Thus, we have two classes of greedy algorithms in Banach spaces. The first one is based on a greedy step of the form (3.1.3). We call this class the class of X-greedy algorithms. The second one is based on a greedy step of the form (3.1.5). We call this class the class of dual greedy algorithms. A very important feature of the dual greedy algorithms is that they can be modified into a weak form. The term ‘weak’ in the definition of the WDGA means that, at the greedy step (3.1.5), we do not aim for the optimal element of the dictionary which realizes the corresponding supremum but are satisfied with a weaker property than being optimal. The obvious reason for this is that we do not know, in general, that the optimal one exists. Another, practical reason is that the weaker the assumption, the easier it is satisfied and, therefore, it is easier to realize in practice.

The greedy algorithms defined above (XGA, WDGA) are the generalizations of the PGA and the WGA, studied in Chapter 2, to the case of Banach spaces. The results of Chapter 2 show that the PGA is not the most efficient greedy algorithm for approximation of elements of $A_1(\mathcal{D})$. It was mentioned in Chapter 2 (see Livshitz and Temlyakov (2003) for the proof) that there exist a dictionary $\mathcal{D}$, a positive constant C, and an element $f \in A_1(\mathcal{D})$ such that, for the PGA,

$$\|f_m\| \geq Cm^{-0.27}. \tag{3.1.10}$$

We note that even before the lower estimate (3.1.10) was proved, researchers began looking for other greedy algorithms that provide a good rate of approximation of functions from $A_1(\mathcal{D})$. Two different ideas have been used at this step. The first idea was that of relaxation: see Jones (1992), Barron (1993), DeVore and Temlyakov (1996) and Temlyakov (2000*b*). The corresponding algorithms (for example, the WRGA, studied in Chapter 2) were designed for approximation of functions from $A_1(\mathcal{D})$. These algorithms do not provide an expansion into a series but they have other good features. It was established (see Theorem 2.3.8) for the WRGA with $\tau = \{1\}$ in a Hilbert space that, for $f \in A_1(\mathcal{D})$,

$$\|f_m\| \leq Cm^{-1/2}.$$

Also, for the WRGA we always have $G_m \in A_1(\mathcal{D})$. The latter property, clearly, limits the applicability of the WRGA to the $A_1(\mathcal{D})$.

The second idea was the idea of building the best approximant from the span$(\varphi_1, \dots, \varphi_m)$ instead of the use of only one element φ_m for an update of the approximant. This idea was realized in the Weak Orthogonal Greedy Algorithm (see Chapter 2) in the case of a Hilbert space and in the Weak Chebyshev Greedy Algorithm (WCGA) (see Temlyakov (2001*b*)) in the case of a Banach space.

The realization of both ideas resulted in the construction of algorithms (the WRGA and WCGA) that are good for approximation of functions from $A_1(\mathcal{D})$. We present results on the WCGA in Section 3.2 and results on the WRGA in Section 3.3. The WCGA has the following advantage over the WRGA. It will be proved in Section 3.2 that the WCGA (under some assumptions on the weakness sequence τ) converges for each $f \in X$ in any uniformly smooth Banach space. The WRGA is simpler than the WCGA in the sense of computational complexity. However, the WRGA has limited applicability. It converges only for elements of the closure of the convex hull of a dictionary. In Sections 3.4 and 3.5 we study algorithms that combine good features of both algorithms the WRGA and the WCGA. In the construction of such algorithms we use different forms of relaxation.

The Weak Greedy Algorithm with Free Relaxation (WGAFR) (Temlyakov 2006*c*), studied in Section 3.4, is the most powerful of the versions considered here. We prove convergence of the WGAFR in Theorem 3.4.3. This theorem is the same as the corresponding convergence result for the

WCGA (see Theorem 3.2.4). The results on the rate of convergence for the WGAFR and the WCGA are also the same (see Theorem 3.4.4 and Theorem 3.2.12). Thus, the WGAFR performs in the same way as the WCGA from the point of view of convergence and rate of convergence, and outperforms the WCGA in terms of computational complexity.

In the WGAFR we are optimizing over two parameters w and λ at each step of the algorithm. In other words we are looking for the best approximation from a two-dimensional linear subspace at each step. In the other version of the weak relaxed greedy algorithms (see the GAWR), considered in Section 3.5, we approximate from a one-dimensional linear subspace at each step of the algorithm. This makes computational complexity of these algorithms very close to that of the PGA. The analysis of the GAWR version turns out to be more complicated than the analysis of the WGAFR. Also, the results obtained for the GAWR are not as general as in the case of the WGAFR. For instance, we present results on the GAWR only in the case $\tau = \{t\}$, when the weakness parameter t is the same for all steps.

The XGA and WDGA have a good feature that distinguishes them from all relaxed greedy algorithms, and also from the WCGA. For an element $f \in X$ they provide an expansion into a series,

$$f \sim \sum_{j=1}^{\infty} c_j(f) g_j(f), \quad g_j(f) \in \mathcal{D}, \quad c_j(f) > 0, \quad j = 1, 2, \dots, \tag{3.1.11}$$

such that

$$G_m = \sum_{j=1}^{m} c_j(f) g_j(f), \quad f_m = f - G_m.$$

In Section 3.7 we discuss other greedy algorithms that provide the expansion (3.1.11).

All the algorithms studied in Sections 3.2–3.7 belong to the class of dual greedy algorithms. Results obtained in Sections 3.2–3.7 confirm that dual greedy algorithms provide powerful methods of nonlinear approximation. In Section 3.8 we present some results on the X-greedy algorithms. These results are similar to those for the dual greedy algorithms.

The algorithms studied in Sections 3.2–3.8 are very general approximation methods that work well in an arbitrary uniformly smooth Banach space X for any dictionary $\mathcal{D}$. This motivates an attempt, made in Section 3.9, to modify these theoretical approximation methods in a direction of practical applicability. In Section 3.9 we illustrate this idea by modifying the WCGA. We note that Section 3.6 is also devoted to modification of greedy algorithms in order to make them more practically feasible. The main idea of Section 3.6 is to replace the most difficult (expensive) step of an algorithm, namely the greedy step, by a thresholding step.

In Section 3.10 we give an example of how the greedy algorithms can be used in constructing deterministic cubature formulas with error estimates similar to those for the Monte Carlo Method.

As a typical example of a uniformly smooth Banach space we will use a space L_p, $1 < p < \infty$. It is well known (see, for instance, Donahue *et al.* (1997, Lemma B.1)) that in the case $X = L_p$, $1 \leq p < \infty$ we have

$$\rho(u) \leq u^p/p \quad \text{if} \quad 1 \leq p \leq 2 \quad \text{and} \quad \rho(u) \leq (p-1)u^2/2 \quad \text{if} \quad 2 \leq p < \infty. \tag{3.1.12}$$

It is also known (see Lindenstrauss and Tzafriri (1977, p. 63)) that, for any X with $\dim X = \infty$, we have

$$\rho(u) \geq (1+u^2)^{1/2} - 1,$$

and for every X, $\dim X \geq 2$,

$$\rho(u) \geq Cu^2, \quad C > 0.$$

This limits the power-type modulus of smoothness of non-trivial Banach spaces to the case $1 \leq q \leq 2$.

3.2. The Weak Chebyshev Greedy Algorithm

Let $\tau := \{t_k\}_{k=1}^\infty$ be a given weakness sequence of non-negative numbers $t_k \leq 1$, $k = 1, \dots$. We define first the Weak Chebyshev Greedy Algorithm (WCGA) (see Temlyakov (2001*b*)) that is a generalization for Banach spaces of the Weak Orthogonal Greedy Algorithm.

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

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

$$F_{f_{m-1}^c}(\varphi_m^c) \geq t_m \|F_{f_{m-1}^c}\|_{\mathcal{D}}.$$

(2) Define

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

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

(3) Let

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

Remark 3.2.1. It follows from the definition of the WCGA that the sequence $\{\|f_m^c\|\}$ is a non-increasing sequence.

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

Definition 3.2.2. Let $\rho(u)$ be an even convex function on $(-\infty,\infty)$ with the property: $\rho(2)\ge 1$ and

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

For any $\tau=\{t_k\}_{k=1}^{\infty}$, $0<t_k\le 1$, and $0<\theta\le 1/2$ we define $\xi_m:=\xi_m(\rho,\tau,\theta)$ as a number u satisfying the equation

$$\rho(u)=\theta t_m u. \tag{3.2.1}$$

Remark 3.2.3. Assumptions on $\rho(u)$ imply that the function

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

is a continuous increasing function on $[0,\infty)$ with $s(2)\ge 1/2$. Thus (3.2.1) has a unique solution $\xi_m=s^{-1}(\theta t_m)$ such that $0<\xi_m\le 2$.

The following theorem from Temlyakov (2001*b*) gives a sufficient condition for convergence of the WCGA.

Theorem 3.2.4. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Assume that a sequence $\tau:=\{t_k\}_{k=1}^{\infty}$ satisfies the condition: for any $\theta>0$ we have

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

Then, for any $f\in X$ we have

$$\lim_{m\to\infty}\|f_m^{c,\tau}\|=0.$$

Corollary 3.2.5. Let a Banach space X have modulus of smoothness $\rho(u)$ of power type $1<q\le 2$, that is, $\rho(u)\le\gamma u^q$. Assume that

$$\sum_{m=1}^{\infty} t_m^p=\infty,\quad p=\frac{q}{q-1}. \tag{3.2.2}$$

Then the WCGA converges for any $f\in X$.

Proof. Denote $\rho^q(u):=\gamma u^q$. Then

$$\rho(u)/u\le\rho^q(u)/u,$$

and therefore for any $\theta>0$ we have

$$\xi_m(\rho,\tau,\theta)\ge\xi_m(\rho^q,\tau,\theta).$$

For ρ^q we get from the definition of ξ_m that

$$\xi_m(\rho^q,\tau,\theta)=(\theta t_m/\gamma)^{\frac{1}{q-1}}.$$

Thus (3.2.2) implies that

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

It remains to apply Theorem 3.2.4. □

The following theorem from Temlyakov (2001*b*) gives the rate of convergence of the WCGA for f in $A_1(\mathcal{D})$.

Theorem 3.2.6. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Then, for a sequence $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \leq 1$, $k = 1, 2, \dots$, we have for any $f \in A_1(\mathcal{D})$ that

$$\|f_m^{c,\tau}\| \leq C(q,\gamma)\biggl(1 + \sum_{k=1}^{m} t_k^p\biggr)^{-1/p}, \quad p := \frac{q}{q-1},$$

with a constant $C(q,\gamma)$ which may depend only on q and γ.

We will use the following two simple and well-known lemmas in the proof of the above two theorems.

Lemma 3.2.7. Let X be a uniformly smooth Banach space and let L be a finite-dimensional subspace of X. For any $f \in X \setminus L$, let f_L denote the best approximant of f from L. Then we have

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

for any $\phi \in L$.

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

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

For any λ we have from the definition of $\rho(u)$ that

$$\|f - f_L - \lambda\phi\| + \|f - f_L + \lambda\phi\| \leq 2\|f - f_L\|\biggl(1 + \rho\biggl(\frac{\lambda}{\|f - f_L\|}\biggr)\biggr). \tag{3.2.3}$$

Next

$$\|f - f_L + \lambda\phi\| \geq F_{f-f_L}(f - f_L + \lambda\phi) = \|f - f_L\| + \lambda\beta. \tag{3.2.4}$$

Combining (3.2.3) and (3.2.4) we get

$$\|f - f_L - \lambda\phi\| \leq \|f - f_L\|\biggl(1 - \frac{\lambda\beta}{\|f - f_L\|} + 2\rho\biggl(\frac{\lambda}{\|f - f_L\|}\biggr)\biggr). \tag{3.2.5}$$

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

$$\biggl(1 - \frac{\lambda'\beta}{\|f - f_L\|} + 2\rho\biggl(\frac{\lambda'}{\|f - f_L\|}\biggr)\biggr) < 1.$$

Then (3.2.5) gives

$$\|f - f_L - \lambda'\phi\| < \|f - f_L\|,$$

which contradicts the assumption that $f_L \in L$ is the best approximant of f. □

Lemma 3.2.8. For any bounded linear functional F and any dictionary $\mathcal{D}$, we have

$$\|F\|_{\mathcal{D}} := \sup_{g\in\mathcal{D}} F(g) = \sup_{f\in A_1(\mathcal{D})} F(f).$$

Proof. The inequality

$$\sup_{g\in\mathcal{D}} F(g) \le \sup_{f\in A_1(\mathcal{D})} F(f)$$

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

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

Thus

$$F(f) \le \|F\|\epsilon + F\left(\sum_{i=1}^N a_i^\epsilon g_i^\epsilon\right) \le \epsilon\|F\| + \sup_{g\in\mathcal{D}} F(g),$$

which proves Lemma 3.2.8. □

We will also need one more lemma from Temlyakov (2001*b*).

Lemma 3.2.9. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Take a number $\epsilon \ge 0$ and two elements f, f^ϵ from X such that

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

with some number $A(\epsilon) > 0$. Then we have

$$\|f_m^{c,\tau}\| \le \|f_{m-1}^{c,\tau}\| \inf_{\lambda\ge 0}\left(1 - \lambda t_m A(\epsilon)^{-1}\left(1 - \frac{\epsilon}{\|f_{m-1}^{c,\tau}\|}\right) + 2\rho\left(\frac{\lambda}{\|f_{m-1}^{c,\tau}\|}\right)\right),$$

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

Proof. We have for any λ

$$\|f_{m-1}^c - \lambda\varphi_m^c\| + \|f_{m-1}^c + \lambda\varphi_m^c\| \le 2\|f_{m-1}^c\|\left(1 + \rho\left(\frac{\lambda}{\|f_{m-1}^c\|}\right)\right), \quad (3.2.6)$$

and by (1) from the definition of the WCGA and Lemma 3.2.8 we get

$$\begin{aligned}F_{f^c_{m-1}}(\varphi^c_m) &\geq t_m \sup_{g\in\mathcal{D}} F_{f^c_{m-1}}(g)\\ &= t_m \sup_{\phi\in A_1(\mathcal{D})} F_{f^c_{m-1}}(\phi) \geq t_m A(\epsilon)^{-1} F_{f^c_{m-1}}(f^\epsilon).\end{aligned}$$

By Lemma 3.2.7 we obtain

$$\begin{aligned}F_{f^c_{m-1}}(f^\epsilon) &= F_{f^c_{m-1}}(f + f^\epsilon - f) \geq F_{f^c_{m-1}}(f) - \epsilon\\ &= F_{f^c_{m-1}}(f^c_{m-1}) - \epsilon = \|f^c_{m-1}\| - \epsilon.\end{aligned}$$

Thus, as in (3.2.5) we get from (3.2.6)

$$\begin{aligned}\|f^c_m\| &\leq \inf_{\lambda\geq 0} \|f^c_{m-1} - \lambda\varphi^c_m\| \qquad (3.2.7)\\ &\leq \|f^c_{m-1}\| \inf_{\lambda\geq 0}\left(1 - \lambda t_m A(\epsilon)^{-1}\left(1 - \frac{\epsilon}{\|f^c_{m-1}\|}\right) + 2\rho\left(\frac{\lambda}{\|f^c_{m-1}\|}\right)\right),\end{aligned}$$

which proves the lemma. □

Proof of Theorem 3.2.4. The definition of the WCGA implies that $\{\|f^c_m\|\}$ is a non-increasing sequence. Therefore we have

$$\lim_{m\to\infty} \|f^c_m\| = \alpha.$$

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

$$\|f^c_m\| \geq \alpha.$$

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

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

with some $A(\epsilon)$. Then, by Lemma 3.2.9 we get

$$\|f^c_m\| \leq \|f^c_{m-1}\| \inf_\lambda (1 - \lambda t_m A(\epsilon)^{-1}/2 + 2\rho(\lambda/\alpha)).$$

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

$$\|f^c_m\| \leq \|f^c_{m-1}\|(1 - 2\theta t_m \xi_m).$$

The assumption

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

implies that

$$\|f^c_m\| \to 0 \quad \text{as } m \to \infty.$$

We have a contradiction, which proves the theorem. □

Proof of Theorem 3.2.6. By Lemma 3.2.9 with $\epsilon = 0$ and $A(\epsilon) = 1$, we have for $f \in A_1(\mathcal{D})$ that

$$\|f_m^c\| \le \|f_{m-1}^c\| \inf_{\lambda \ge 0} \biggl(1 - \lambda t_m + 2\gamma\biggl(\frac{\lambda}{\|f_{m-1}^c\|}\biggr)^q\biggr). \tag{3.2.8}$$

Choose λ from the equation

$$\frac{1}{2}\lambda t_m = 2\gamma\biggl(\frac{\lambda}{\|f_{m-1}^c\|}\biggr)^q,$$

which implies that

$$\lambda = \|f_{m-1}^c\|^{\frac{q}{q-1}}(4\gamma)^{-\frac{1}{q-1}} t_m^{\frac{1}{q-1}}.$$

Let

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

Using the notation $p := \frac{q}{q-1}$, we get from (3.2.8)

$$\|f_m^c\| \le \|f_{m-1}^c\|\biggl(1 - \frac{1}{2}\lambda t_m\biggr) = \|f_{m-1}^c\|(1 - t_m^p\|f_{m-1}^c\|^p/A_q).$$

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

$$\|f_m^c\|^p \le \|f_{m-1}^c\|^p(1 - t_m^p\|f_{m-1}^c\|^p/A_q).$$

By an analogue of Lemma 2.3.3 (see Temlyakov (2000*b*, Lemma 3.1)), using the estimate $\|f\|^p \le 1 < A_q$ we get

$$\|f_m^c\|^p \le A_q\biggl(1 + \sum_{n=1}^m t_n^p\biggr)^{-1}$$

which implies

$$\|f_m^c\| \le C(q,\gamma)\biggl(1 + \sum_{n=1}^m t_n^p\biggr)^{-1/p}.$$

Theorem 3.2.6 is now proved. □

Remark 3.2.10. Theorem 3.2.6 also holds for a slightly modified version of the WCGA, the WCGA(1), for which at step (1) we require

$$F_{f_{m-1}^{c(1)}}(\varphi_m^{c(1)}) \ge t_m\|f_{m-1}^{c(1)}\|. \tag{3.2.9}$$

This statement follows from the fact that, in the proof of Theorem 3.2.6, the relation

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

was used only to get (3.2.9).

Proposition 3.2.11. Condition (3.2.2) in Corollary 3.2.5 is sharp.

Proof. Let $1 < q \leq 2$. Consider $X = \ell_q$. It is known (Lindenstrauss and Tzafriri 1977, p. 67) that ℓ_q, $1 < q \leq 2$, is a uniformly smooth Banach space with modulus of smoothness $\rho(u)$ of power type q. Denote $p := \frac{q}{q-1}$ and take any $\{t_k\}_{k=1}^{\infty}$, $0 < t_k \leq 1$, such that

$$\sum_{k=1}^{\infty} t_k^p < \infty. \tag{3.2.10}$$

Choose $\mathcal{D}$ as a standard basis $\{e_j\}_{j=1}^{\infty}$, $e_j := (0,\dots,0,1,0,\dots)$, for ℓ_q. Consider the following realization of the WCGA for

$$f := \Big(1, t_1^{\frac{1}{q-1}}, t_2^{\frac{1}{q-1}}, \dots\Big).$$

First of all, (3.2.10) guarantees that $f \in \ell_q$. Next, it is well known that F_f can be identified as

$$F_f = (1, t_1, t_2, \dots)/\Big(1 + \sum_{k=1}^{\infty} t_k^p\Big)^{1/p} \in \ell_p.$$

At the first step of the WCGA we pick $\varphi_1 = e_2$ and get

$$f_1^c = \Big(1, 0, t_2^{\frac{1}{q-1}}, \dots\Big).$$

We continue with f replaced by f_1 and so on. After m steps we get

$$f_m^c = \Big(1, 0, \dots, 0, t_{m+1}^{\frac{1}{q-1}}, \dots\Big).$$

It is clear that for all m we have $\|f_m^c\|_{\ell_q} \geq 1$. □

The following variant of Theorem 3.2.6 (see Temlyakov (2006*c*)) follows from Lemma 3.2.9.

Theorem 3.2.12. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Take a number $\epsilon \geq 0$ and two elements f, f^ϵ from X such that

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

with some number $A(\epsilon) > 0$. Then we have $(p := q/(q-1))$

$$\|f_m^{c,\tau}\| \leq \max\biggl(2\epsilon, C(q,\gamma)(A(\epsilon)+\epsilon)\biggl(1+\sum_{k=1}^{m} t_k^p\biggr)^{-1/p}\biggr). \tag{3.2.11}$$

3.3. Relaxation; co-convex approximation

In this section we study a generalization for Banach spaces of relaxed greedy algorithms considered in Chapter 2. We present here results from Temlyakov (2001*b*). Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a given weakness sequence of numbers $t_k \in [0,1]$, $k = 1, \ldots$.

Weak Relaxed Greedy Algorithm (WRGA). We define $f_0^r := f_0^{r,\tau} := f$ and $G_0^r := G_0^{r,\tau} := 0$. Then, for each $m \geq 1$ we have the following inductive definition.

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

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

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

$$\|f - ((1-\lambda_m)G_{m-1}^r + \lambda_m\varphi_m^r)\| = \inf_{0\leq\lambda\leq 1} \|f - ((1-\lambda)G_{m-1}^r + \lambda\varphi_m^r)\|$$

and define

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

(3) Let

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

Remark 3.3.1. It follows from the definition of the WRGA that the sequence $\{\|f_m^r\|\}$ is a non-increasing sequence.

We call the WRGA *relaxed* because at the mth step of the algorithm we use a linear combination (convex combination) of the previous approximant G_{m-1}^r and a new element φ_m^r. The relaxation parameter λ_m in the WRGA is chosen at the mth step depending on f. We prove here the analogues of Theorems 3.2.4 and 3.2.6 for the Weak Relaxed Greedy Algorithm.

Theorem 3.3.2. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Assume that a sequence $\tau := \{t_k\}_{k=1}^{\infty}$ satisfies the condition: for any $\theta > 0$ we have

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

Then, for any $f \in A_1(\mathcal{D})$ we have

$$\lim_{m\to\infty} \|f_m^{r,\tau}\| = 0.$$

Theorem 3.3.3. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$. Then, for a sequence $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \le 1$, $k = 1, 2, \dots$, we have for any $f \in A_1(\mathcal{D})$ that

$$\|f_m^{r,\tau}\| \le C_1(q,\gamma)\Big(1 + \sum_{k=1}^{m} t_k^p\Big)^{-1/p}, \quad p := \frac{q}{q-1},$$

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

Proof of Theorems 3.3.2 and 3.3.3. This proof is similar to the proof of Theorems 3.2.4 and 3.2.6. Instead of Lemma 3.2.9 we use the following lemma.

Lemma 3.3.4. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Then, for any $f \in A_1(\mathcal{D})$ we have

$$\|f_m^{r,\tau}\| \le \|f_{m-1}^{r,\tau}\| \inf_{0\le\lambda\le1}\Big(1 - \lambda t_m + 2\rho\Big(\frac{2\lambda}{\|f_{m-1}^{r,\tau}\|}\Big)\Big), \quad m = 1, 2, \dots.$$

Proof. We have

$$f_m^r := f - ((1-\lambda_m)G_{m-1}^r + \lambda_m\varphi_m^r) = f_{m-1}^r - \lambda_m(\varphi_m^r - G_{m-1}^r)$$

and

$$\|f_m^r\| = \inf_{0\le\lambda\le1} \|f_{m-1}^r - \lambda(\varphi_m^r - G_{m-1}^r)\|.$$

As for (3.2.6), we have for any λ

$$\begin{aligned}\|f_{m-1}^r - {}&\lambda(\varphi_m^r - G_{m-1}^r)\| + \|f_{m-1}^r + \lambda(\varphi_m^r - G_{m-1}^r)\| \\ &\le 2\|f_{m-1}^r\|\Big(1 + \rho\Big(\frac{\lambda\|\varphi_m^r - G_{m-1}^r\|}{\|f_{m-1}^r\|}\Big)\Big).\end{aligned} \tag{3.3.1}$$

Next we get for $\lambda \ge 0$

$$\begin{aligned}&\|f_{m-1}^r + \lambda(\varphi_m^r - G_{m-1}^r)\| \\ &\quad\ge F_{f_{m-1}^r}(f_{m-1}^r + \lambda(\varphi_m^r - G_{m-1}^r)) \\ &\quad= \|f_{m-1}^r\| + \lambda F_{f_{m-1}^r}(\varphi_m^r - G_{m-1}^r) \ge \|f_{m-1}^r\| + \lambda t_m \sup_{g\in\mathcal{D}} F_{f_{m-1}^r}(g - G_{m-1}^r) \\ &\quad= \|f_{m-1}^r\| + \lambda t_m \sup_{\phi\in A_1(\mathcal{D}} F_{f_{m-1}^r}(\phi - G_{m-1}^r) \ge \|f_{m-1}^r\| + \lambda t_m\|f_{m-1}^r\|,\end{aligned}$$

applying Lemma 3.2.8 for the last inequality. Using the trivial estimate $\|\varphi_m^r - G_{m-1}^r\| \le 2$, we obtain

$$\|f_{m-1}^r - \lambda(\varphi_m^r - G_{m-1}^r)\| \le \|f_{m-1}^r\|\Big(1 - \lambda t_m + 2\rho\Big(\frac{2\lambda}{\|f_{m-1}^r\|}\Big)\Big), \tag{3.3.2}$$

from (3.3.1), which proves Lemma 3.3.4. □

The remaining part of the proof uses inequality (3.3.2) in the same way relation (3.2.7) was used in the proof of Theorems 3.2.4 and 3.2.6. The only additional difficulty here is that we are optimizing over $0 \le \lambda \le 1$. However, it is easy to check that the corresponding λ chosen in a similar way always satisfies the restriction $0 \le \lambda \le 1$. In the proof of Theorem 3.3.2 we choose $\theta = \alpha/8$ and $\lambda = \alpha\xi_m(\rho,\tau,\theta)/2$, and in the proof of Theorem 3.3.3 we choose λ from the equation

$$\frac{1}{2}\lambda t_m = 2\gamma(2\lambda)^q \|f^r_{m-1}\|^{-q}. \qquad \square$$

Remark 3.3.5. Theorems 3.3.2 and 3.3.3 hold for a slightly modified version of the WRGA, the WRGA(1), for which at step (1) we require

$$F_{f^{r(1)}_{m-1}}(\varphi^{r(1)}_m - G^{r(1)}_{m-1}) \ge t_m \|f^{r(1)}_{m-1}\|. \tag{3.3.3}$$

This follows from the observation that in the proof of Lemma 3.3.4 we used the inequality from step (1) of the WRGA only to derive (3.3.3). It is clear from Lemma 3.2.8 that in the case of approximation of $f \in A_1(\mathcal{D})$, the requirement (3.3.3) is weaker and easier to check than step (1) of the WRGA.

3.4. Free relaxation

Both of the above algorithms, the WCGA and the WRGA, use the functional $F_{f_{m-1}}$ in a search for the mth element φ_m from the dictionary to be used in approximation. The construction of the approximant in the WRGA is different from the construction in the WCGA. In the WCGA we build the approximant G^c_m so as to maximally use the approximation power of the elements $\varphi_1, \dots, \varphi_m$. The WRGA by its definition is designed for approximation of functions from $A_1(\mathcal{D})$. In building the approximant in the WRGA we keep the property $G^r_m \in A_1(\mathcal{D})$. As we mentioned in Section 3.3 the relaxation parameter λ_m in the WRGA is chosen at the mth step depending on f. The following modification of the above idea of relaxation in greedy approximation will be studied in this section (see Temlyakov (2006*c*)).

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

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

$$F_{f_{m-1}}(\varphi_m) \ge t_m \|F_{f_{m-1}}\|_{\mathcal{D}}.$$

(2) Find w_m and λ_m such that

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

and define

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

(3) Let

$$f_m := f - G_m.$$

We begin with the following analogue of Lemma 3.2.9.

Lemma 3.4.1. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Take a number $\epsilon \geq 0$ and two elements f, f^ϵ from X such that

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

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

$$\|f_m\| \leq \|f_{m-1}\| \inf_{\lambda \geq 0} \biggl(1 - \lambda t_m A(\epsilon)^{-1}\Bigl(1 - \frac{\epsilon}{\|f_{m-1}\|}\Bigr) + 2\rho\Bigl(\frac{5\lambda}{\|f_{m-1}\|}\Bigr)\biggr),$$

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

Proof. By the definition of f_m,

$$\|f_m\| \leq \inf_{\lambda \geq 0, w} \|f_{m-1} + wG_{m-1} - \lambda \varphi_m\|.$$

As in the arguments in the proof of Lemma 3.2.9, we use the inequality

$$\begin{aligned} \|f_{m-1} + wG_{m-1} - \lambda\varphi_m\| + \|f_{m-1} - wG_{m-1} + \lambda\varphi_m\| \\ \leq 2\|f_{m-1}\|(1 + \rho(\|wG_{m-1} - \lambda\varphi_m\| / \|f_{m-1}\|)), \end{aligned} \tag{3.4.1}$$

and estimate for $\lambda \geq 0$

$$\begin{aligned} \|f_{m-1} - wG_{m-1} + \lambda\varphi_m\| &\geq F_{f_{m-1}}(f_{m-1} - wG_{m-1} + \lambda\varphi_m) \\ &\geq \|f_{m-1}\| - F_{f_{m-1}}(wG_{m-1}) + \lambda t_m \sup_{g \in \mathcal{D}} F_{f_{m-1}}(g). \end{aligned}$$

By Lemma 3.2.8, we continue:

$$\begin{aligned} &= \|f_{m-1}\| - F_{f_{m-1}}(wG_{m-1}) + \lambda t_m \sup_{\phi \in A_1(\mathcal{D})} F_{f_{m-1}}(\phi) \\ &\geq \|f_{m-1}\| - F_{f_{m-1}}(wG_{m-1}) + \lambda t_m A(\epsilon)^{-1} F_{f_{m-1}}(f^\epsilon) \\ &\geq \|f_{m-1}\| - F_{f_{m-1}}(wG_{m-1}) + \lambda t_m A(\epsilon)^{-1}(F_{f_{m-1}}(f) - \epsilon). \end{aligned}$$

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

$$\|f_{m-1} - w^* G_{m-1} + \lambda\varphi_m\| \geq \|f_{m-1}\| + \lambda t_m A(\epsilon)^{-1}(\|f_{m-1}\| - \epsilon). \tag{3.4.2}$$

Combining (3.4.1) and (3.4.2) we get

$$\begin{aligned} \|f_m\| \leq \|f_{m-1}\| \inf_{\lambda \geq 0} (1 - \lambda t_m A(\epsilon)^{-1}(1 - \epsilon / \|f_{m-1}\|) \\ + 2\rho(\|w^* G_{m-1} - \lambda\varphi_m\| / \|f_{m-1}\|)). \end{aligned}$$

We now estimate

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

Next,

$$\|G_{m-1}\| = \|f - f_{m-1}\| \le 2\|f\| \le 2(\|f^\epsilon\| + \epsilon) \le 2(A(\epsilon) + \epsilon).$$

Thus, under assumption $A(\epsilon) \ge \epsilon$ we get

$$w^*\|G_{m-1}\| \le 2\lambda t_m (A(\epsilon) + \epsilon)/A(\epsilon) \le 4\lambda.$$

Finally,

$$\|w^* G_{m-1} - \lambda\varphi_m\| \le 5\lambda.$$

This completes the proof of Lemma 3.4.1. □

Remark 3.4.2. It follows from the definition of the WGAFR that the sequence $\{\|f_m\|\}$ is a non-increasing sequence.

We now prove a convergence theorem for an arbitrary uniformly smooth Banach space. Modulus of smoothness $\rho(u)$ of a uniformly smooth Banach space is an even convex function such that $\rho(0) = 0$ and $\lim_{u\to 0} \rho(u)/u = 0$. The function $s(u) := \rho(u)/u$, $s(0) := 0$, associated with $\rho(u)$, is a continuous increasing function on $[0,\infty)$. Therefore, the inverse function $s^{-1}(\cdot)$ is well defined.

Theorem 3.4.3. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Assume that a sequence $\tau := \{t_k\}_{k=1}^\infty$ satisfies the following condition. For any $\theta > 0$ we have

$$\sum_{m=1}^{\infty} t_m s^{-1}(\theta t_m) = \infty. \tag{3.4.3}$$

Then, for any $f \in X$ we have for the WGAFR

$$\lim_{m\to\infty} \|f_m\| = 0.$$

Proof. By Remark 3.4.2, $\{\|f_m\|\}$ is a non-increasing sequence. Therefore we have

$$\lim_{m\to\infty} \|f_m\| = \beta.$$

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

$$\|f_m\| \ge \beta.$$

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

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

with some $A(\epsilon) \geq \epsilon$. Then, by Lemma 3.4.1 we get

$$\|f_m\| \leq \|f_{m-1}\| \inf_{\lambda \geq 0} (1 - \lambda t_m A(\epsilon)^{-1}/2 + 2\rho(5\lambda/\beta)).$$

Let us specify $\theta := \beta/(40A(\epsilon))$ and take $\lambda = \beta s^{-1}(\theta t_m)/5$. Then we obtain

$$\|f_m\| \leq \|f_{m-1}\|(1 - 2\theta t_m s^{-1}(\theta t_m)).$$

The assumption

$$\sum_{m=1}^{\infty} t_m s^{-1}(\theta t_m) = \infty$$

implies that

$$\|f_m\| \to 0 \quad \text{as } m \to \infty.$$

We have a contradiction, which proves the theorem. □

Theorem 3.4.4. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Take a number $\epsilon \geq 0$ and two elements f, f^ϵ from X such that

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

with some number $A(\epsilon) > 0$. Then we have for the WGAFR

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

Proof. It is clear that it suffices to consider the case $A(\epsilon) \geq \epsilon$. Otherwise, $\|f_m\| \leq \|f\| \leq \|f^\epsilon\| + \epsilon \leq 2\epsilon$. Also, assume $\|f_m\| > 2\epsilon$ (otherwise Theorem 3.4.4 trivially holds). Then, by Remark 3.4.2, we have for all $k = 0, 1, \dots, m$ that $\|f_k\| > 2\epsilon$. By Lemma 3.4.1 we obtain

$$\|f_k\| \leq \|f_{k-1}\| \inf_{\lambda \geq 0} \left(1 - \lambda t_k A(\epsilon)^{-1}/2 + 2\gamma\left(\frac{5\lambda}{\|f_{k-1}\|}\right)^q\right). \tag{3.4.4}$$

Choose λ from the equation

$$\frac{\lambda t_k}{4A(\epsilon)} = 2\gamma\left(\frac{5\lambda}{\|f_{k-1}\|}\right)^q,$$

which implies that

$$\lambda = \|f_{k-1}\|^{\frac{q}{q-1}} 5^{-\frac{q}{q-1}} (8\gamma A(\epsilon))^{-\frac{1}{q-1}} t_k^{\frac{1}{q-1}}.$$

Define

$$A_q := 4(8\gamma)^{\frac{1}{q-1}} 5^{\frac{q}{q-1}}.$$

Using the notation $p := \frac{q}{q-1}$, we get from (3.4.4)

$$\|f_k\| \le \|f_{k-1}\|\biggl(1 - \frac{1}{4}\frac{\lambda t_k}{A(\epsilon)}\biggr) = \|f_{k-1}\|\biggl(1 - \frac{t_k^p\|f_{k-1}\|^p}{A_q A(\epsilon)^p}\biggr).$$

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

$$\|f_k\|^p \le \|f_{k-1}\|^p\biggl(1 - \frac{t_k^p\|f_{k-1}\|^p}{A_q A(\epsilon)^p}\biggr).$$

By an analogue of Lemma 2.3.3 (see Temlyakov (2000*b*, Lemma 3.1)), using the estimates $\|f\| \le A(\epsilon) + \epsilon$ and $A_q > 1$, we get

$$\|f_m\|^p \le A_q(A(\epsilon)+\epsilon)^p\biggl(1 + \sum_{k=1}^m t_k^p\biggr)^{-1},$$

which implies

$$\|f_m\| \le C(q,\gamma)(A(\epsilon)+\epsilon)\biggl(1 + \sum_{k=1}^m t_k^p\biggr)^{-1/p}.$$

Theorem 3.4.4 is proved. □

3.5. Fixed relaxation

In this section we consider a relaxed greedy algorithm with relaxation prescribed in advance. Let a sequence $\mathbf{r} := \{r_k\}_{k=1}^\infty$, $r_k \in [0,1)$, of relaxation parameters be given. Then, at each step of our new algorithm we build the mth approximant of the form $G_m = (1-r_m)G_{m-1} + \lambda\varphi_m$. With an approximant of this form we are not limited to approximation of functions from $A_1(\mathcal{D})$ as in the WRGA. In this section we study the Greedy Algorithm with Weakness parameter t and Relaxation $\mathbf{r}$ (GAWR$(t,\mathbf{r})$). In addition to the acronym GAWR$(t,\mathbf{r})$ we will use the abbreviated acronym GAWR for the name of this algorithm. We give a general definition of the algorithm in the case of a weakness sequence τ. We present in this section results from Temlyakov (2006*c*).

GAWR$(\boldsymbol{\tau},\mathbf{r})$. Let $\tau := \{t_m\}_{m=1}^\infty$, $t_m \in (0,1]$, be a weakness sequence and let $\mathbf{r} := \{r_m\}_{m=1}^\infty$, $r_m \in [0,1)$, be a relaxation sequence. We define $f_0 := f$ and $G_0 := 0$. Then, for each $m \ge 1$ we have the following inductive definition.

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

$$F_{f_{m-1}}(\varphi_m) \ge t_m\|F_{f_{m-1}}\|_{\mathcal{D}}.$$

(2) Find $\lambda_m \geq 0$ such that

$$\|f - ((1-r_m)G_{m-1} + \lambda_m \varphi_m)\| = \inf_{\lambda \geq 0} \|f - ((1-r_m)G_{m-1} + \lambda \varphi_m)\|$$

and define

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

(3) Let

$$f_m := f - G_m.$$

In the case $\tau = \{t\}$ we write t instead of τ in the notation. We note that in the case $r_k = 0$, $k = 1, \dots,$ when there is no relaxation the GAWR$(\tau, \mathbf{0})$ coincides with the Weak Dual Greedy Algorithm. We now proceed to the GAWR. We begin with an analogue of Lemma 3.2.9.

Lemma 3.5.1. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Take a number $\epsilon \geq 0$ and two elements f, f^ϵ from X such that

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

with some number $A(\epsilon) > 0$. Then we have for the GAWR$(t, \mathbf{r})$

$$\begin{aligned}\|f_m\| \leq \|f_{m-1}\|&(1 - r_m(1 - \epsilon/\|f_{m-1}\|) \\ &+ 2\rho((r_m(\|f\| + A(\epsilon)/t))/((1-r_m)\|f_{m-1}\|)), \quad m = 1, 2, \dots .\end{aligned}$$

Theorem 3.5.2. Let a sequence $\mathbf{r}$ satisfy the conditions

$$\sum_{k=1}^{\infty} r_k = \infty, \quad r_k \to 0 \quad \text{as } k \to \infty.$$

Then the GAWR$(t, \mathbf{r})$ converges in any uniformly smooth Banach space for each $f \in X$ and for all dictionaries $\mathcal{D}$.

Proof. We prove this theorem in two steps.

I First, we prove that $\liminf_{m\to\infty} \|f_m\| = 0$. The proof goes by contradiction. We want to prove that $\liminf_{m\to\infty} \|f_m\| = 0$. Assume the contrary. Then there exists K and $\beta > 0$ such that we have for all $k \geq K$ that $\|f_k\| \geq \beta$. By Lemma 3.5.1, for $m > K$

$$\|f_m\| \leq \|f_{m-1}\| \left(1 - r_m\left(1 - \frac{\epsilon}{\beta}\right) + 2\rho\left(\frac{r_m(\|f\| + A(\epsilon)/t)}{(1-r_m)\beta}\right)\right).$$

We choose $\epsilon := \beta/2$. Using the assumption that X is uniformly smooth and the assumption $r_k \to 0$ as $k \to \infty$, we find $N \geq K$ such that for $m \geq N$ we have

$$2\rho\left(\frac{r_m(\|f\| + A(\epsilon)/t)}{(1-r_m)\beta}\right) \leq r_m/4.$$

Then, for $m > N$,

$$\|f_m\| \le \|f_{m-1}\|(1 - r_m/4).$$

The assumption $\sum_{m=1}^{\infty} r_m = \infty$ implies that $\|f_m\| \to 0$ as $m \to \infty$. The obtained contradiction to the assumption $\beta > 0$ completes the proof of part I.

II Secondly, we prove that $\lim_{m\to\infty} \|f_m\| = 0$. Using the assumption $r_k \to 0$ as $k \to \infty$, we find N_1 such that for $k \ge N_1$ we have $r_k \le 1/2$. For such k we obtain from Lemma 3.5.1

$$\|f_k\| - \epsilon \le (1 - r_k)(\|f_{k-1}\| - \epsilon) + 2\|f_{k-1}\|\rho\left(\frac{Br_k}{\|f_{k-1}\|}\right), \tag{3.5.1}$$

with $B := 2(\|f\| + A(\epsilon)/t)$. Denote $a_k := \|f_{k-1}\| - \epsilon$. We note that from the definition of f_k it follows that

$$a_{k+1} \le a_k + r_k\|f\|. \tag{3.5.2}$$

Using the fact that the function $\rho(u)/u$ is monotone increasing on $[0, \infty)$, we obtain from (3.5.1) for $a_k > 0$

$$\begin{aligned} a_{k+1} &\le a_k\left(1 - r_k + 2\frac{\|f_{k-1}\|}{a_k}\rho\left(\frac{Br_k}{\|f_{k-1}\|}\right)\right) \\ &\le a_k\left(1 - r_k + 2\rho\left(\frac{Br_k}{a_k}\right)\right). \end{aligned} \tag{3.5.3}$$

We now introduce an auxiliary sequence $\{b_k\}$ of positive numbers that is defined by the equation

$$2\rho(Br_k/b_k) = r_k.$$

The property $\rho(u)/u \to 0$ as $u \to 0$ implies $b_k \to 0$ as $k \to \infty$. Inequality (3.5.3) guarantees that for $k \ge N_1$ such that $a_k \ge b_k$, we have $a_{k+1} \le a_k$.

Let

$$U := \{k : k \ge N_1, \quad a_k \ge b_k\}.$$

If the set U is finite then we get

$$\limsup_{k\to\infty} a_k \le \lim_{k\to\infty} b_k = 0.$$

This implies

$$\limsup_{m\to\infty} \|f_m\| \le \epsilon.$$

Consider the case when U is infinite. We note that part I of the proof implies that there is a subsequence $\{k_j\}$ such that $a_{k_j} \le 0$, $j = 1, 2, \dots$. This means that

$$U = \cup_{j=1}^{\infty} [l_j, n_j],$$

with the property $n_{j-1} < l_j - 1$. For $k \notin U$, $k \geq N_1$ we have

$$a_k < b_k. \tag{3.5.4}$$

For $k \in [l_j, n_j]$, we have by (3.5.2) and the monotonicity property of a_k, when $k \in [l_j, n_j]$, that

$$a_k \leq a_{l_j} \leq a_{l_j-1} + r_{l_j-1}\|f\| \leq b_{l_j-1} + r_{l_j-1}\|f\|. \tag{3.5.5}$$

By (3.5.4) and (3.5.5) we obtain

$$\limsup_{k\to\infty} a_k \leq 0 \;\Rightarrow\; \limsup_{m\to\infty} \|f_m\| \leq \epsilon.$$

Taking into account that $\epsilon > 0$ is arbitrary, we complete the proof. □

We now proceed to results on the rate of approximation. We will need the following technical lemma (see Temlyakov (1999, 2006*c*)).

Lemma 3.5.3. Let a sequence $\{a_n\}_{n=1}^\infty$ have the following property. For, given positive numbers $\alpha < \gamma \leq 1$, $A > a_1$, we have, for all $n \geq 2$,

$$a_n \leq a_{n-1} + A(n-1)^{-\alpha}. \tag{3.5.6}$$

If for some $\nu \geq 2$ we have

$$a_\nu \geq A\nu^{-\alpha},$$

then

$$a_{\nu+1} \leq a_\nu(1 - \gamma/\nu). \tag{3.5.7}$$

Then there exists a constant $C(\alpha,\gamma)$ such that, for all $n = 1, 2, \dots$, we have

$$a_n \leq C(\alpha,\gamma)An^{-\alpha}.$$

Theorem 3.5.4. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Let $\mathbf{r} := \{2/(k+2)\}_{k=1}^\infty$. Consider the GAWR$(t,\mathbf{r})$. For a pair of functions f, f^ϵ, satisfying

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

we have

$$\|f_m\| \leq \epsilon + C(q,\gamma)(\|f\| + A(\epsilon)/t)m^{-1+1/q}.$$

Proof. By Lemma 3.5.1 we obtain

$$\|f_k\| - \epsilon \leq (1 - r_k)(\|f_{k-1}\| - \epsilon) + C\gamma\|f_{k-1}\|\left(\frac{r_k(\|f\| + A(\epsilon)/t)}{\|f_{k-1}\|}\right)^q. \tag{3.5.8}$$

Consider, as in the proof of Theorem 3.5.2, the sequence $a_n := \|f_{n-1}\| - \epsilon$. We plan to apply Lemma 3.5.3 to the sequence $\{a_n\}$. We set $\alpha := 1 - 1/q \leq 1/2$. The parameters $\gamma \in (\alpha, 1]$ and A will be chosen later. We note that

$$\|f_m\| \leq \|f_{m-1}\| + r_m\|f\|.$$

Therefore, condition (3.5.6) of Lemma 3.5.3 is satisfied with $A \geq 2\|f\|$. Let $a_k \geq Ak^{-\alpha}$. Then, by (3.5.8) we get

$$\begin{aligned} a_{k+1} &\leq a_k(1 - r_k + C\gamma(r_k(\|f\| + A(\epsilon)/t)/a_k)^q \\ &\leq a_k\left(1 - \frac{2}{k+2} + \frac{C\gamma(\|f\| + A(\epsilon)/t)^q 2^q}{A^q}\frac{k^{\alpha q}}{(k+2)^q}\right). \end{aligned}$$

Setting $A := \max(2\|f\|, 2(2C\gamma)^{1/q}(\|f\| + A(\epsilon)/t))$, we obtain

$$a_{k+1} \leq a_k\left(1 - \frac{3}{2(k+2)}\right).$$

Thus condition (3.5.7) of Lemma 3.5.3 is satisfied with $\gamma = 3/4$. Applying Lemma 3.5.3 we obtain

$$\|f_m\| \leq \epsilon + C(q,\gamma)(\|f\| + A(\epsilon)/t)m^{-1+1/q}. \qquad \square$$

We conclude this section by the following remark. The algorithms GAWR and WGAFR are both dual-type greedy algorithms. The first steps are similar for both algorithms: we use the norming functional $F_{f_{m-1}}$ in the search for an element φ_m. The WGAFR provides more freedom than the GAWR does in choosing good coefficients w_m and λ_m. This results in more flexibility in choosing the weakness sequence $\tau = \{t_m\}$. For instance, condition (3.4.3) of Theorem 3.4.3 is satisfied if $\tau = \{t\}$, $t \in (0,1]$ for any uniformly smooth Banach space. In the case $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$, condition (3.4.3) is satisfied if

$$\sum_{m=1}^{\infty} t_m^p = \infty, \quad p := q/(q-1).$$

3.6. Thresholding algorithms

We begin with a remark on computational complexity of greedy algorithms. The main point of Section 3.4 is in proving that relaxation allows us to build greedy algorithms (see the WGAFR) that are computationally simpler than the WCGA and perform as well as the WCGA. We note that the WCGA and the WGAFR differ in the second step of the algorithm. However, the most computationally involved step of all greedy algorithms is the greedy step (the first step of the algorithm). One of the goals of relaxation was to get rid of the assumption $f \in A_1(\mathcal{D})$ (as in the WRGA). All relaxed greedy algorithms from Sections 3.4 and 3.5 are applicable to (and converge for) any $f \in X$. We want to point out that the information $f \in A_1(\mathcal{D})$ allows us to simplify substantially the greedy step of the algorithm. It is remarked in Section 3.2 (see Remark 3.2.10) that we can replace the first step of the WCGA by the following search criterion:

$$F_{f_{m-1}}(\varphi_m) \geq t_m\|f_{m-1}\|. \tag{3.6.1}$$

A similar remark (see Section 3.3, Remark 3.3.5) holds for the WRGA. The requirement (3.6.1) is weaker than the requirement of the greedy step of the WCGA. However, Theorem 3.2.6 holds for this modification of the WCGA. Relation (3.6.1) is a threshold-type inequality and can be checked more easily than the greedy inequality.

We now consider two algorithms defined and studied in Temlyakov (2006*c*) with a different type of thresholding. These algorithms work for any $f \in X$. We begin with the Dual Greedy Algorithm with Relaxation and Thresholding (DGART).

DGART. We define $f_0 := f$ and $G_0 := 0$. Then, for a given parameter $\delta \in (0, 1/2]$ we have the following inductive definition for $m \geq 1$.

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

$$F_{f_{m-1}}(\varphi_m) \geq \delta. \tag{3.6.2}$$

If there is no $\varphi_m \in \mathcal{D}$ satisfying (3.6.2) then we stop.

(2) Find w_m and λ_m such that

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

and define

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

(3) Let

$$f_m := f - G_m.$$

If $\|f_m\| \leq \delta\|f\|$ then we stop, otherwise we proceed to the $(m + 1)$th iteration.

The following algorithm is a thresholding-type modification of the WCGA. This modification can be applied to any $f \in X$.

Chebyshev Greedy Algorithm with Thresholding (CGAT). For a given parameter $\delta \in (0, 1/2]$, we conduct instead of the greedy step of the WCGA the following thresholding step: find $\varphi_m \in \mathcal{D}$ such that $F_{f_{m-1}}(\varphi_m) \geq \delta$. Choosing such a φ_m, if one exists, we apply steps (2) and (3) of the WCGA. If such φ_m does not exist, then we stop. We also stop if $\|f_m\| \leq \delta\|f\|$.

Theorem 3.6.1. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Take a number $\epsilon \geq 0$ and two elements f, f^ϵ from X such that

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

with some number $A(\epsilon) > 0$. Then the DGART (CGAT) will stop after $m \le C(\gamma)\delta^{-p}\ln(1/\delta)$, $p := q/(q-1)$, iterations with

$$\|f_m\| \le \epsilon + \delta A(\epsilon).$$

Proof. We begin with the error bound. For both algorithms, the DGART and the CGAT, our stopping criterion guarantees that either $\|F_{f_m}\|_{\mathcal{D}} \le \delta$ or $\|f_m\| \le \delta\|f\|$. In the latter case the required bound follows from simple inequalities:

$$\|f\| \le \epsilon + \|f^\epsilon\| \le \epsilon + A(\epsilon).$$

Thus, assume that $\|F_{f_m}\|_{\mathcal{D}} \le \delta$ holds. In the case of the CGAT we apply Lemma 3.2.7 with $L = \operatorname{span}(\varphi_1, \dots, \varphi_m)$ and obtain

$$\|f_m\| = F_{f_m}(f_m) = F_{f_m}(f) \le \epsilon + F_{f_m}(f^\epsilon) \le \epsilon + \|F_{f_m}\|_{\mathcal{D}} A(\epsilon) \le \epsilon + \delta A(\epsilon).$$

For the DGART we apply Lemma 3.2.7 with f_{m-1} and $L = \operatorname{span}(G_{m-1}, \varphi_m)$, and get

$$\begin{aligned}\|f_m\| &= F_{f_m}(f_m) = F_{f_m}(f_{m-1}) = F_{f_m}(f)\\ &\le \epsilon + F_{f_m}(f^\epsilon) \le \epsilon + \|F_{f_m}\|_{\mathcal{D}} A(\epsilon) \le \epsilon + \delta A(\epsilon).\end{aligned}$$

This proves the required bound.

We now proceed to the bound of m. We prove the bound for both algorithms simultaneously. We note that for the DGART

$$\|f_k\| = \inf_{\lambda,w} \|f_{k-1} + wG_{k-1} - \lambda\varphi_k\| \le \inf_{\lambda\ge 0} \|f_{k-1} - \lambda\varphi_k\|.$$

We write for all $k \le m$, $\lambda \ge 0$

$$\|f_{k-1} - \lambda\varphi_k\| + \|f_{k-1} + \lambda\varphi_k\| \le 2\|f_{k-1}\|(1 + \rho(\lambda/\|f_{k-1}\|)). \tag{3.6.3}$$

Next,

$$\|f_{k-1} + \lambda\varphi_k\| \ge F_{f_{k-1}}(f_{k-1} + \lambda\varphi_k) \ge \|f_{k-1}\| + \lambda\delta. \tag{3.6.4}$$

Combining (3.6.3) with (3.6.4), we obtain

$$\|f_k\| \le \inf_{\lambda\ge 0} \|f_{k-1} - \lambda\varphi_k\| \le \inf_{\lambda\ge 0}\big(\|f_{k-1}\| - \lambda\delta + 2\|f_{k-1}\|\gamma(\lambda/\|f_{k-1}\|)^q\big). \tag{3.6.5}$$

Solving the equation $\delta x/2 = 2\gamma x^q$ we get $x_1 = (\delta/(4\gamma))^{1/(q-1)}$. Setting $\lambda := x_1\|f_{k-1}\|$ we obtain

$$\|f_k\| \le \|f_{k-1}\|(1 - \delta x_1/2) = \|f_{k-1}\|(1 - c(\gamma)\delta^p).$$

Thus,

$$\|f_k\| \le \|f\|(1 - c(\gamma)\delta^p)^k.$$

By the stopping condition $\|f_m\| \le \delta\|f\|$, we deduce that $m \le n$, where n is the smallest integer for which

$$(1-c(\gamma)\delta^p)^n \le \delta.$$

This implies

$$m \le C(\gamma)\delta^{-p}\ln(1/\delta). \qquad \Box$$

We proceed to one more thresholding-type algorithm (see Temlyakov (2005*a*)). Keeping in mind possible applications of this algorithm, we do not assume that a dictionary $\mathcal{D}$ is symmetric: $g \in \mathcal{D}$ implies $-g \in \mathcal{D}$. To indicate this we use the notation $\mathcal{D}^+$ for such a dictionary. We do not assume that elements of a dictionary $\mathcal{D}^+$ are normalized ($\|g\| = 1$ if $g \in \mathcal{D}^+$) and assume only that $\|g\| \le 1$ if $g \in \mathcal{D}^+$. By $A_1(\mathcal{D}^+)$ we denote the closure of the convex hull of $\mathcal{D}^+$. Let $\epsilon = \{\epsilon_n\}_{n=1}^\infty$, $\epsilon_n > 0$, $n = 1, 2, \dots$.

Incremental Algorithm with schedule $\boldsymbol{\epsilon}$ (IA($\boldsymbol{\epsilon}$)). Let $f \in A_1(\mathcal{D}^+)$. Denote $f_0^{i,\epsilon} := f$ and $G_0^{i,\epsilon} := 0$. Then, for each $m \ge 1$ we have the following inductive definition.

(1) $\varphi_m^{i,\epsilon} \in \mathcal{D}^+$ is any element satisfying

$$F_{f_{m-1}^{i,\epsilon}}(\varphi_m^{i,\epsilon} - f) \ge -\epsilon_m.$$

(2) Define

$$G_m^{i,\epsilon} := (1-1/m)G_{m-1}^{i,\epsilon} + \varphi_m^{i,\epsilon}/m.$$

(3) Let

$$f_m^{i,\epsilon} := f - G_m^{i,\epsilon}.$$

We note that, as in Lemma 3.2.8, we have for any bounded linear functional F and any $\mathcal{D}^+$

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

Therefore, for any F and any $f \in A_1(\mathcal{D}^+)$,

$$\sup_{g\in\mathcal{D}^+} F(g) \ge F(f).$$

This guarantees existence of $\varphi_m^{i,\epsilon}$.

Theorem 3.6.2. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$. Define

$$\epsilon_n := K_1\gamma^{1/q}n^{-1/p}, \quad p = \frac{q}{q-1}, \quad n = 1, 2, \dots.$$

Then, for any $f \in A_1(\mathcal{D}^+)$ we have

$$\|f_m^{i,\epsilon}\| \le C(K_1)\gamma^{1/q}m^{-1/p}, \quad m = 1, 2 \dots.$$

Proof. We will use the abbreviated notation $f_m := f_m^{i,\epsilon}$, $\varphi_m := \varphi_m^{i,\epsilon}$, $G_m := G_m^{i,\epsilon}$. Writing

$$f_m = f_{m-1} - (\varphi_m - G_{m-1})/m,$$

we immediately obtain the trivial estimate

$$\|f_m\| \leq \|f_{m-1}\| + 2/m. \tag{3.6.6}$$

Since

$$\begin{aligned} f_m &= (1-1/m)f_{m-1} - (\varphi_m - f)/m \\ &= (1-1/m)(f_{m-1} - (\varphi_m - f)/(m-1)), \end{aligned} \tag{3.6.7}$$

we obtain

$$\begin{aligned} &\|f_{m-1} - (\varphi_m - f)/(m-1)\| \\ &\quad \leq \|f_{m-1}\|(1 + 2\rho(2((m-1)\|f_{m-1}\|)^{-1})) + \epsilon_m(m-1)^{-1}, \end{aligned} \tag{3.6.8}$$

in a similar way to (3.6.5). Using the definition of ϵ_m and the assumption $\rho(u) \leq \gamma u^q$, we make the following observation. There exists a constant $C(K_1)$ such that, if

$$\|f_{m-1}\| \geq C(K_1)\gamma^{1/q}(m-1)^{-1/p}, \tag{3.6.9}$$

then

$$2\rho(2((m-1)\|f_{m-1}\|)^{-1}) + \epsilon_m((m-1)\|f_{m-1}\|)^{-1} \leq 1/(4m), \tag{3.6.10}$$

and therefore, by (3.6.7) and (3.6.8),

$$\|f_m\| \leq (1 - 3/(4m))\|f_{m-1}\|. \tag{3.6.11}$$

Taking into account (3.6.6), we apply Lemma 3.5.3 to the sequence $a_n = \|f_n\|$, $n = 1, 2, \ldots$ with $\alpha = 1/p$, $\beta = 3/4$, and complete the proof of Theorem 3.6.2. □

3.7. Greedy expansions

3.7.1. Introduction

From the definition of a dictionary it follows that any element $f \in X$ can be approximated arbitrarily well by finite linear combinations of the dictionary elements. The primary goal of this section is to study representations of an element $f \in X$ by a series

$$f \sim \sum_{j=1}^{\infty} c_j(f) g_j(f), \quad g_j(f) \in \mathcal{D}, \quad c_j(f) > 0, \quad j = 1, 2, \ldots. \tag{3.7.1}$$

In building the representation (3.7.1) we should construct two sequences: $\{g_j(f)\}_{j=1}^{\infty}$ and $\{c_j(f)\}_{j=1}^{\infty}$. In this section the construction of $\{g_j(f)\}_{j=1}^{\infty}$

will be based on ideas used in greedy-type nonlinear approximation (greedy-type algorithms). This justifies the use of the term *greedy expansion* for (3.7.1) considered in the section. The construction of $\{g_j(f)\}_{j=1}^{\infty}$ is, clearly, the most important and difficult part in building the representation (3.7.1). On the basis of the contemporary theory of nonlinear approximation with respect to redundant dictionaries, we may conclude that the method of using a norming functional in greedy steps of an algorithm is the most productive in approximation in Banach spaces. This method was utilized in the Weak Chebyshev Greedy Algorithm and in the Weak Dual Greedy Algorithm. We use this same method in new algorithms considered in this section. A new qualitative result of this section establishes that we have a lot of flexibility in constructing a sequence of coefficients $\{c_j(f)\}_{j=1}^{\infty}$.

Denote

$$r_{\mathcal{D}}(f) := \sup_{F_f} \|F_f\|_{\mathcal{D}} := \sup_{F_f} \sup_{g \in \mathcal{D}} F_f(g).$$

We note that, in general, a norming functional F_f is not unique. This is why we take $\sup_{F_f}$ over all norming functionals of f in the definition of $r_{\mathcal{D}}(f)$. It is known that in the case of uniformly smooth Banach spaces (our primary object here) the norming functional F_f is unique. In such a case we do not need $\sup_{F_f}$ in the definition of $r_{\mathcal{D}}(f)$: we have $r_{\mathcal{D}}(f) = \|F_f\|_{\mathcal{D}}$.

We begin with a description of a general scheme that provides an expansion for a given element f. Later, specifying this general scheme, we will obtain different methods of expansion.

Dual-Based Expansion (DBE). Let $t \in (0,1]$ and $f \neq 0$. Denote $f_0 := f$. Assume $\{f_j\}_{j=0}^{m-1} \subset X$, $\{\varphi_j\}_{j=1}^{m-1} \subset \mathcal{D}$ and a set of coefficients $\{c_j\}_{j=1}^{m-1}$ of expansion have already been constructed. If $f_{m-1} = 0$ then we stop (set $c_j = 0$, $j = m, m+1, \dots$ in the expansion) and get $f = \sum_{j=1}^{m-1} c_j\varphi_j$. If $f_{m-1} \neq 0$ then we conduct the following two steps.

(1) Choose $\varphi_m \in \mathcal{D}$ such that

$$\sup_{F_{f_{m-1}}} F_{f_{m-1}}(\varphi_m) \geq t r_{\mathcal{D}}(f_{m-1}).$$

(2) Define

$$f_m := f_{m-1} - c_m\varphi_m,$$

where $c_m > 0$ is a coefficient either prescribed in advance or chosen from a concrete approximation procedure.

We call the series

$$f \sim \sum_{j=1}^{\infty} c_j\varphi_j \tag{3.7.2}$$

the Dual-Based Expansion of f with coefficients $c_j(f) := c_j$, $j = 1, 2, \dots$ with respect to $\mathcal{D}$.

Denote

$$S_m(f, \mathcal{D}) := \sum_{j=1}^{m} c_j \varphi_j.$$

Then it is clear that

$$f_m = f - S_m(f, \mathcal{D}).$$

We prove some convergence results for the DBE in Sections 3.7.2 and 3.7.3. In Section 3.7.3 we consider a variant of the Dual-Based Expansion with coefficients chosen by a certain simple rule. The rule depends on two numerical parameters, $t \in (0, 1]$ (the weakness parameter from the definition of the DBE) and $b \in (0, 1)$ (the tuning parameter of the approximation method). The rule also depends on a majorant μ of the modulus of smoothness of the Banach space X.

Dual Greedy Algorithm with parameters (t, b, μ) (DGA(t, b, μ)). Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$, and let $\mu(u)$ be a continuous majorant of $\rho(u)$: $\rho(u) \leq \mu(u)$, $u \in [0, \infty)$. For parameters $t \in (0, 1]$, $b \in (0, 1]$ we define sequences $\{f_m\}_{m=0}^{\infty}$, $\{\varphi_m\}_{m=1}^{\infty}$, $\{c_m\}_{m=1}^{\infty}$ inductively. Let $f_0 := f$. If for $m \geq 1$ $f_{m-1} = 0$ then we set $f_j = 0$ for $j \geq m$ and stop. If $f_{m-1} \neq 0$ then we conduct the following three steps.

(1) Take any $\varphi_m \in \mathcal{D}$ such that

$$F_{f_{m-1}}(\varphi_m) \geq t r_{\mathcal{D}}(f_{m-1}). \tag{3.7.3}$$

(2) Choose $c_m > 0$ from the equation

$$\|f_{m-1}\| \mu(c_m / \|f_{m-1}\|) = \frac{tb}{2} c_m r_{\mathcal{D}}(f_{m-1}). \tag{3.7.4}$$

(3) Define

$$f_m := f_{m-1} - c_m \varphi_m. \tag{3.7.5}$$

In Section 3.7.3 we prove the following convergence result.

Theorem 3.7.1. Let X be a uniformly smooth Banach space with the modulus of smoothness $\rho(u)$ and let $\mu(u)$ be a continuous majorant of $\rho(u)$ with the property $\mu(u)/u \downarrow 0$ as $u \to +0$. Then, for any $t \in (0, 1]$ and $b \in (0, 1)$, the DGA(t, b, μ) converges for each dictionary $\mathcal{D}$ and all $f \in X$.

The following result from Section 3.7.3 gives the rate of convergence.

Theorem 3.7.2. Assume X has a modulus of smoothness $\rho(u) \le \gamma u^q$, $q \in (1,2]$. Denote $\mu(u) = \gamma u^q$. Then, for any dictionary $\mathcal{D}$ and any $f \in A_1(\mathcal{D})$, the rate of convergence of the DGA(t,b,μ) is given by

$$\|f_m\| \le C(t,b,\gamma,q)m^{-\frac{t(1-b)}{p(1+t(1-b))}}, \quad p := \frac{q}{q-1}.$$

3.7.2. Convergence of the Dual-Based Expansion

We begin with the following lemma.

Lemma 3.7.3. Let $f \in X$. Assume that the coefficients $\{c_j\}_{j=1}^{\infty}$ of the expansion

$$f \sim \sum_{j=1}^{\infty} c_j\varphi_j, \qquad f_m := f - \sum_{j=1}^{m} c_j\varphi_j$$

are non-negative and satisfy the following two conditions:

$$\sum_{j=1}^{\infty} c_j r_{\mathcal{D}}(f_{j-1}) < \infty, \tag{3.7.6}$$

$$\sum_{j=1}^{\infty} c_j = \infty. \tag{3.7.7}$$

Then

$$\liminf_{m\to\infty} \|f_m\| = 0. \tag{3.7.8}$$

Proof. The proof of this lemma is similar to the proof of Lemma 1 from Ganichev and Kalton (2003). Denote $s_n := \sum_{j=1}^{n} c_j$. Then (3.7.7) implies (see Bary (1961, p. 904)) that

$$\sum_{n=1}^{\infty} c_n/s_n = \infty. \tag{3.7.9}$$

Using (3.7.6), we get

$$\sum_{n=1}^{\infty} s_n r_{\mathcal{D}}(f_{n-1})c_n/s_n = \sum_{n=1}^{\infty} c_n r_{\mathcal{D}}(f_{n-1}) < \infty.$$

Thus, by (3.7.9),

$$\liminf_{n\to\infty} s_n r_{\mathcal{D}}(f_{n-1}) = 0,$$

and also $(s_{n-1} \le s_n)$

$$\liminf_{n\to\infty} s_n r_{\mathcal{D}}(f_n) = 0.$$

Let

$$\lim_{k\to\infty} s_{n_k} r_{\mathcal{D}}(f_{n_k}) = 0. \tag{3.7.10}$$

Consider $\{F_{f_{n_k}}\}$. The unit sphere in the dual X^* is weakly* compact (see Habala, Hájek and Zizler (1996, p. 45)). Let $\{F_i\}_{i=1}^{\infty}$, $F_i := F_{f_{n_{k_i}}}$ be a w^*-convergent subsequence. Denote

$$F := w^* - \lim_{i\to\infty} F_i.$$

We will complete the proof of Lemma 3.7.3 by contradiction. We assume that (3.7.8) does not hold, that is, there exist $\alpha > 0$ and $N \in \mathbb{N}$ such that

$$\|f_m\| \ge \alpha, \quad m \ge N, \tag{3.7.11}$$

and will thence derive a contradiction.

We begin by deducing from (3.7.11) that $F \ne 0$. Indeed, we have

$$F(f) = \lim_{i\to\infty} F_i(f), \tag{3.7.12}$$

and

$$F_i(f) = F_i\biggl(f_{n_{k_i}} + \sum_{j=1}^{n_{k_i}} c_j\varphi_j\biggr) = \|f_{n_{k_i}}\| + \sum_{j=1}^{n_{k_i}} c_j F_i(\varphi_j) \ge \alpha - s_{n_{k_i}} r_{\mathcal{D}}(f_{n_{k_i}}), \tag{3.7.13}$$

for big i. Relations (3.7.12), (3.7.13) and (3.7.10) imply that $F(f) \ge \alpha$, and hence $F \ne 0$. This implies that there exist $g \in \mathcal{D}$ for which $F(g) > 0$. However,

$$F(g) = \lim_{i\to\infty} F_i(g) \le \lim_{i\to\infty} r_{\mathcal{D}}(f_{n_{k_i}}) = 0.$$

We have a contradiction, which completes the proof of Lemma 3.7.3. □

In the paper Temlyakov (2007*b*) we pushed to the extreme the flexibility of choice of the coefficients $c_j(f)$ in (3.7.1). We made these coefficients independent of an element $f \in X$. Surprisingly, for properly chosen coefficients we obtained results for the corresponding dual greedy expansion similar to the above Theorems 3.7.1 and 3.7.2. Even more surprisingly, we obtained similar results for the corresponding X-greedy expansions. We proceed to the formulation of these results. Let $\mathcal{C} := \{c_m\}_{m=1}^{\infty}$ be a fixed sequence of positive numbers. We restrict ourselves to positive numbers because of the symmetry of the dictionary $\mathcal{D}$.

***X*-Greedy Algorithm with coefficients $\mathcal{C}$ (XGA($\mathcal{C}$)).** We define $f_0 := f$, $G_0 := 0$. Then, for each $m \ge 1$ we have the following inductive definition.

(1) $\varphi_m \in \mathcal{D}$ is such that (assuming existence)

$$\|f_{m-1} - c_m\varphi_m\|_X = \inf_{g\in\mathcal{D}} \|f_{m-1} - c_m g\|_X.$$

(2) Let

$$f_m := f_{m-1} - c_m\varphi_m, \qquad G_m := G_{m-1} + c_m\varphi_m.$$

Dual Greedy Algorithm, weakness $\boldsymbol{\tau}$, coefficients $\boldsymbol{\mathcal{C}}$ (DGA($\boldsymbol{\tau}, \boldsymbol{\mathcal{C}}$)). Let $\tau := \{t_m\}_{m=1}^{\infty}$, $t_m \in [0,1]$, be a weakness sequence. We define $f_0 := f$, $G_0 := 0$. Then, for each $m \geq 1$ we have the following inductive definition.

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

$$F_{f_{m-1}}(\varphi_m) \geq t_m \|F_{f_{m-1}}\|_{\mathcal{D}}.$$

(2) Define

$$f_m := f_{m-1} - c_m\varphi_m, \qquad G_m := G_{m-1} + c_m\varphi_m.$$

In the case $\tau = \{t\}$, $t \in (0,1]$, we write t instead of τ in the notation. The first result on convergence properties of the DGA($t, \mathcal{C}$) was obtained in Temlyakov (2007a). We prove it here.

Theorem 3.7.4. Let X be a uniformly smooth Banach space with the modulus of smoothness $\rho(u)$. Assume $\mathcal{C} = \{c_j\}_{j=1}^{\infty}$ is such that $c_j \geq 0$, $j = 1, 2, \dots,$

$$\sum_{j=1}^{\infty} c_j = \infty,$$

and for any $y > 0$,

$$\sum_{j=1}^{\infty} \rho(yc_j) < \infty. \tag{3.7.14}$$

Then, for the DGA($t, \mathcal{C}$) we have

$$\liminf_{m\to\infty} \|f_m\| = 0. \tag{3.7.15}$$

Proof. The proof is by contradiction. Assume (3.7.15) does not hold. Then $\exists \alpha > 0$ and $\exists N \in \mathbb{N}$ such that, for all $m \geq N$,

$$\|f_m\| \geq \alpha > 0.$$

From the definition of the modulus of smoothness we have

$$\|f_{n-1} - c_n\varphi_n\| + \|f_{n-1} + c_n\varphi_n\| \leq 2\|f_{n-1}\|(1 + \rho(c_n/\|f_{n-1}\|)). \tag{3.7.16}$$

Using the definition of φ_n,

$$F_{f_{n-1}}(\varphi_n) \geq t r_{\mathcal{D}}(f_{n-1}), \tag{3.7.17}$$

we get

$$\begin{aligned} \|f_{n-1} + c_n\varphi_n\| &\geq F_{f_{n-1}}(f_{n-1} + c_n\varphi_n) \qquad (3.7.18)\\ &= \|f_{n-1}\| + c_n F_{f_{n-1}}(\varphi_n) \geq \|f_{n-1}\| + c_n t r_{\mathcal{D}}(f_{n-1}). \end{aligned}$$

Combining (3.7.16) and (3.7.18), we get

$$\|f_n\| = \|f_{n-1} - c_n\varphi_n\| \leq \|f_{n-1}\|(1+2\rho(c_n/\|f_{n-1}\|)) - c_n t r_{\mathcal{D}}(f_{n-1}). \tag{3.7.19}$$

We note that by Remark 3.2.3

$$\|f_{n-1}\|\rho(c_n/\|f_{n-1}\|) \leq \alpha\rho(c_n/\alpha), \quad n > N.$$

Therefore, by the assumption (3.7.14)

$$\sum_{n=1}^{\infty} \|f_{n-1}\|\rho(c_n/\|f_{n-1}\|) < \infty. \tag{3.7.20}$$

This and (3.7.19) imply

$$\sum_{n=1}^{\infty} c_n r_{\mathcal{D}}(f_{n-1}) \leq t^{-1}\Bigg(\|f\| + 2\sum_{n=1}^{\infty} \|f_{n-1}\|\rho(c_n/\|f_{n-1}\|)\Bigg) < \infty.$$

It remains to apply Lemma 3.7.3 to complete the proof. □

In Temlyakov (2007*b*) we proved an analogue of Theorem 3.7.4 for the XGA($\mathcal{C}$) and improved upon the convergence in Theorem 3.7.4 in the case of uniformly smooth Banach spaces with power-type modulus of smoothness. Under an extra assumption on $\mathcal{C}$ we replaced lim inf by lim. Here is the corresponding result from Temlyakov (2007*b*).

Theorem 3.7.5. Let $\mathcal{C} \in \ell_q \setminus \ell_1$ be a monotone sequence. Then the DGA($t, \mathcal{C}$) and the XGA($\mathcal{C}$) converge for each dictionary and all $f \in X$ in any uniformly smooth Banach space X with modulus of smoothness $\rho(u) \leq \gamma u^q$, $q \in (1, 2]$.

In Temlyakov (2007*b*) we also addressed a question of rate of approximation for $f \in A_1(\mathcal{D})$. We proved the following theorem.

Theorem 3.7.6. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $q \in (1, 2]$. We set $s := (1 + 1/q)/2$ and $\mathcal{C}_s := \{k^{-s}\}_{k=1}^{\infty}$. Then the DGA($t, \mathcal{C}_s$) and XGA($\mathcal{C}_s$) (for this algorithm $t = 1$) converge for $f \in A_1(\mathcal{D})$ with the following rate: for any $r \in (0, t(1 - s))$,

$$\|f_m\| \leq C(r, t, q, \gamma)m^{-r}.$$

In the case $t = 1$, Theorem 3.7.6 provides the rate of convergence m^{-r} for $f \in A_1(\mathcal{D})$ with r arbitrarily close to $(1 - 1/q)/2$. Theorem 3.7.2 provides a similar rate of convergence. It would be interesting to know if the rate $m^{-(1-1/q)/2}$ is the best that can be achieved in greedy expansions (for each $\mathcal{D}$, any $f \in A_1(\mathcal{D})$, and any X with $\rho(u) \leq \gamma u^q$, $q \in (1, 2]$). We note that there are greedy approximation methods that provide an error bound of the order $m^{1/q-1}$ for $f \in A_1(\mathcal{D})$ (see Temlyakov (2003*a*, 2006*c*) for recent results). However, these approximation methods do not provide an expansion.

3.7.3. A modification of the Weak Dual Greedy Algorithm

We begin this subsection with a proof of Theorem 3.7.1. Here we give a definition of the DGA(τ, b, μ), $\tau = \{t_k\}_{k=1}^{\infty}$, $t_k \in (0,1]$ that coincides with the definition of the DGA(t, b, μ) from Section 3.7.1 in the case $\tau = \{t\}$.

Dual Greedy Algorithm with parameters (τ, b, μ) (DGA(τ, b, μ)). Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$, and let $\mu(u)$ be a continuous majorant of $\rho(u)$: $\rho(u) \le \mu(u)$, $u \in [0, \infty)$. For a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $t_k \in (0,1]$ and a parameter $b \in (0,1]$, we define sequences $\{f_m\}_{m=0}^{\infty}$, $\{\varphi_m\}_{m=1}^{\infty}$, $\{c_m\}_{m=1}^{\infty}$ inductively. Let $f_0 := f$. If $f_{m-1} = 0$ for some $m \ge 1$, then we set $f_j = 0$ for $j \ge m$ and stop. If $f_{m-1} \ne 0$ then we conduct the following three steps.

(1) Take any $\varphi_m \in \mathcal{D}$ such that

$$F_{f_{m-1}}(\varphi_m) \ge t_m r_{\mathcal{D}}(f_{m-1}). \tag{3.7.21}$$

(2) Choose $c_m > 0$ from the equation

$$\|f_{m-1}\|\mu(c_m/\|f_{m-1}\|) = \frac{t_m b}{2} c_m r_{\mathcal{D}}(f_{m-1}). \tag{3.7.22}$$

(3) Define

$$f_m := f_{m-1} - c_m \varphi_m. \tag{3.7.23}$$

Proof of Theorem 3.7.1. In this case $\tau = \{t\}$, $t \in (0,1]$. We have by (3.7.19)

$$\|f_m\| = \|f_{m-1} - c_m\varphi_m\| \le \|f_{m-1}\|(1 + 2\rho(c_m/\|f_{m-1}\|)) - c_m t r_{\mathcal{D}}(f_{m-1}). \tag{3.7.24}$$

Using the choice of c_m, we find

$$\|f_m\| \le \|f_{m-1}\| - t(1-b)c_m r_{\mathcal{D}}(f_{m-1}). \tag{3.7.25}$$

In particular, (3.7.25) implies that $\{\|f_m\|\}$ is a monotone decreasing sequence and

$$t(1-b)c_m r_{\mathcal{D}}(f_{m-1}) \le \|f_{m-1}\| - \|f_m\|.$$

Thus

$$\sum_{m=1}^{\infty} c_m r_{\mathcal{D}}(f_{m-1}) < \infty. \tag{3.7.26}$$

We have the following two cases:

$$\text{(I)} \quad \sum_{m=1}^{\infty} c_m = \infty, \qquad \text{(II)} \quad \sum_{m=1}^{\infty} c_m < \infty.$$

In case (I), by Lemma 3.7.3 we obtain

$$\liminf_{m\to\infty} \|f_m\| = 0 \;\Rightarrow\; \lim_{m\to\infty} \|f_m\| = 0.$$

It remains to consider case (II). We prove convergence in this case by contradiction. Assume

$$\lim_{m\to\infty} \|f_m\| = \alpha > 0. \tag{3.7.27}$$

By (II) we have $f_m \to f_\infty \neq 0$ as $m \to \infty$. We note that by uniform smoothness of X we get

$$\lim_{m\to\infty} \|F_{f_m} - F_{f_\infty}\| = 0.$$

We have $F_{f_\infty} \neq 0$, and therefore there is a $g \in \mathcal{D}$ such that $F_{f_\infty}(g) > 0$. However,

$$F_{f_\infty}(g) = \lim_{m\to\infty} F_{f_m}(g) \le \lim_{m\to\infty} r_{\mathcal{D}}(f_m) = 0. \tag{3.7.28}$$

Indeed, by (3.7.22) and (3.7.27) we get

$$r_{\mathcal{D}}(f_{m-1}) \le \alpha c_m^{-1} \mu(c_m/\alpha) \frac{2}{tb} \to 0,$$

as $m \to \infty$.

Theorem 3.7.1 is proved. □

Remark 3.7.7. It is clear from the above proof that Theorem 3.7.1 holds for an algorithm obtained from the DGA(τ, b, μ), by replacing (3.7.22) by

$$\|f_{m-1}\| \mu(c_m/\|f_{m-1}\|) = \frac{b}{2} c_m F_{f_{m-1}}(\varphi_m). \tag{3.7.29}$$

Also, a parameter b in (3.7.22) and (3.7.29) can be replaced by varying parameters $b_m \in (a, b) \subset (0, 1)$.

We proceed to study the rate of convergence of the DGA(τ, b, μ) in the uniformly smooth Banach spaces with the power-type majorant of modulus of smoothness: $\rho(u) \le \mu(u) = \gamma u^q$, $1 < q \le 2$. We now prove a statement more general than Theorem 3.7.2.

Theorem 3.7.8. Let $\tau := \{t_k\}_{k=1}^\infty$ be a non-increasing sequence $1 \ge t_1 \ge t_2 \cdots > 0$ and $b \in (0, 1)$. Assume X has a modulus of smoothness $\rho(u) \le \gamma u^q$, $q \in (1, 2]$. Denote $\mu(u) = \gamma u^q$. Then, for any dictionary $\mathcal{D}$ and any $f \in A_1(\mathcal{D})$, the rate of convergence of the DGA(τ, b, μ) is given by

$$\|f_m\| \le C(b, \gamma, q)\left(1 + \sum_{k=1}^m t_k^p\right)^{-\frac{t_m(1-b)}{p(1+t_m(1-b))}}, \quad p := \frac{q}{q-1}.$$

Proof. As in (3.7.25), we get

$$\|f_m\| \le \|f_{m-1}\| - t_m(1-b) c_m r_{\mathcal{D}}(f_{m-1}). \tag{3.7.30}$$

Thus we need to estimate $c_m r_{\mathcal{D}}(f_{m-1})$ from below. It is clear that

$$\|f_{m-1}\|_{A_1(\mathcal{D})} = \|f - \sum_{j=1}^{m-1} c_j\varphi_j\|_{A_1(\mathcal{D})} \le \|f\|_{A_1(\mathcal{D})} + \sum_{j=1}^{m-1} c_j. \tag{3.7.31}$$

Denote $b_n := 1 + \sum_{j=1}^n c_j$. Then, by (3.7.31) we get

$$\|f_{m-1}\|_{A_1(\mathcal{D})} \le b_{m-1}.$$

Next, by Lemma 3.2.8 we get

$$\begin{aligned} r_{\mathcal{D}}(f_{m-1}) &= \sup_{g\in\mathcal{D}} F_{f_{m-1}}(g) = \sup_{\varphi\in A_1(\mathcal{D})} F_{f_{m-1}}(\varphi) \\ &\ge \|f_{m-1}\|^{-1}_{A_1(\mathcal{D})} F_{f_{m-1}}(f_{m-1}) \ge \|f_{m-1}\|/b_{m-1}. \end{aligned} \tag{3.7.32}$$

Substituting (3.7.32) into (3.7.30), we get

$$\|f_m\| \le \|f_{m-1}\|(1 - t_m(1-b)c_m/b_{m-1}). \tag{3.7.33}$$

From the definition of b_m we find

$$b_m = b_{m-1} + c_m = b_{m-1}(1 + c_m/b_{m-1}).$$

Using the inequality

$$(1+x)^\alpha \le 1 + \alpha x, \quad 0 \le \alpha \le 1, \quad x \ge 0,$$

we obtain

$$b_m^{t_m(1-b)} \le b_{m-1}^{t_m(1-b)}(1 + t_m(1-b)c_m/b_{m-1}). \tag{3.7.34}$$

Multiplying (3.7.33) and (3.7.34), and using that $t_m \le t_{m-1}$, we get

$$\|f_m\| b_m^{t_m(1-b)} \le \|f_{m-1}\| b_{m-1}^{t_{m-1}(1-b)} \le \|f\| \le 1. \tag{3.7.35}$$

The function $\mu(u)/u = \gamma u^{q-1}$ is increasing on $[0,\infty)$. Therefore the c_m from (3.7.22) is greater than or equal to c'_m from (see (3.7.32))

$$\gamma\|f_{m-1}\|(c'_m/\|f_{m-1}\|)^q = \frac{t_m b}{2} c'_m \|f_{m-1}\|/b_{m-1}, \tag{3.7.36}$$

$$c'_m = \left(\frac{t_m b}{2\gamma}\right)^{\frac{1}{q-1}} \frac{\|f_{m-1}\|^{\frac{q}{q-1}}}{b_{m-1}^{\frac{1}{q-1}}}. \tag{3.7.37}$$

Setting

$$p := \frac{q}{q-1}, \qquad A^{-1} := (1-b)\left(\frac{b}{2\gamma}\right)^{\frac{1}{q-1}} \le 1/2,$$

we obtain

$$\|f_m\| \le \|f_{m-1}\|\left(1 - \frac{t_m^p}{A}\frac{\|f_{m-1}\|^p}{b_{m-1}^p}\right) \tag{3.7.38}$$

from (3.7.30), (3.7.32) and (3.7.37). Noting that $b_m \geq b_{m-1}$, we infer from (3.7.38) that

$$(\|f_m\|/b_m)^p \leq (\|f_{m-1}\|/b_{m-1})^p(1 - A^{-1}t_m^p(\|f_{m-1}\|/b_{m-1})^p). \qquad (3.7.39)$$

Taking into account that $\|f\| \leq 1 < A$, we obtain from (3.7.39) by an analogue of Lemma 2.3.3 (see Temlyakov (2000*b*, Lemma 3.1))

$$(\|f_m\|/b_m)^p \leq A\biggl(1 + \sum_{k=1}^{m} t_k^p\biggr)^{-1}. \qquad (3.7.40)$$

Combining (3.7.35) and (3.7.40), we get

$$\|f_m\| \leq C(b,\gamma,q)\biggl(1 + \sum_{k=1}^{m} t_k^p\biggr)^{-\frac{t_m(1-b)}{p(1+t_m(1-b))}}, \quad p := \frac{q}{q-1}.$$

This completes the proof of Theorem 3.7.8. □

In the case $\tau = \{t\}$, $t \in (0,1]$, we get Theorem 3.7.2 from Theorem 3.7.8.

Remark 3.7.9. Theorem 3.7.8 holds for an algorithm obtained from the DGA(τ, b, μ) by replacing (3.7.22) by (3.7.29).

It follows from the proof of Theorem 3.7.8 that it holds for a modification of the DGA(τ, b, μ) when we replace the quantity $r_{\mathcal{D}}(f_{m-1})$ in the definition by its lower estimate (see (3.7.32)) $\|f_{m-1}\|/b_{m-1}$, with $b_{m-1} := 1+\sum_{j=1}^{m-1} c_j$. Clearly, this modification is more suitable for practical implementation than the DGA(τ, b, μ). We formulate the above remark as a separate result.

Modified Dual Greedy Algorithm (τ, b, μ) (MDGA(τ, b, μ)). Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$ and let $\mu(u)$ be a continuous majorant of $\rho(u)$: $\rho(u) \leq \mu(u)$, $u \in [0, \infty)$. For a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $t_k \in (0,1]$ and a parameter $b \in (0,1)$, we define for $f \in A_1(\mathcal{D})$ sequences $\{f_m\}_{m=0}^{\infty}$, $\{\varphi_m\}_{m=1}^{\infty}$, $\{c_m\}_{m=1}^{\infty}$ inductively. Let $f_0 := f$. If for $m \geq 1$ $f_{m-1} = 0$, then we set $f_j = 0$ for $j \geq m$ and stop. If $f_{m-1} \neq 0$ then we conduct the following three steps.

(1) Take any $\varphi_m \in \mathcal{D}$ such that

$$F_{f_{m-1}}(\varphi_m) \geq t_m\|f_{m-1}\|\biggl(1 + \sum_{j=1}^{m-1} c_j\biggr)^{-1}.$$

(2) Choose $c_m > 0$ from the equation

$$\mu(c_m/\|f_{m-1}\|) = \frac{t_m b}{2} c_m\biggl(1 + \sum_{j=1}^{m-1} c_j\biggr)^{-1}.$$

(3) Define

$$f_m := f_{m-1} - c_m\varphi_m.$$

Theorem 3.7.10. Let $\tau := \{t_k\}_{k=1}^{\infty}$ be a non-increasing sequence $1 \geq t_1 \geq t_2 \cdots > 0$ and $b \in (0,1)$. Assume X has a modulus of smoothness $\rho(u) \leq \gamma u^q$, $q \in (1,2]$. Denote $\mu(u) = \gamma u^q$. Then, for any dictionary $\mathcal{D}$ and any $f \in A_1(\mathcal{D})$, the rate of convergence of the MDGA(τ, b, μ) is given by

$$\|f_m\| \leq C(b,\gamma,q)\Big(1 + \sum_{k=1}^{m} t_k^p\Big)^{-\frac{t_m(1-b)}{p(1+t_m(1-b))}}, \quad p := \frac{q}{q-1}.$$

Let us discuss an application of Theorem 3.7.2 in the case of a Hilbert space. It is well known and easy to check that, for a Hilbert space H,

$$\rho(u) \leq (1+u^2)^{1/2} - 1 \leq u^2/2.$$

Therefore, by Theorem 3.7.2 with $\mu(u) = u^2/2$, the DGA(t,b,μ) provides the following error estimate:

$$\|f_m\| \leq C(t,b)m^{-\frac{t(1-b)}{2(1+t(1-b))}} \quad \text{for } f \in A_1(\mathcal{D}). \tag{3.7.41}$$

The estimate (3.7.41) with $t = 1$ gives

$$\|f_m\| \leq C(b)m^{-\frac{1-b}{2(2-b)}} \quad \text{for } f \in A_1(\mathcal{D}). \tag{3.7.42}$$

The exponent $(1-b)/(2(2-b))$ in this estimate tends to $1/4$ when b tends to 0. Comparing (3.7.42) with the upper estimate for the PGA (see Section 2.3), we observe that the DGA$(1,b,u^2/2)$ with small b has a better upper estimate for the rate of convergence than the known estimates for the PGA. We note also that inequality (2.3.21) indicates that the exponent in the power rate of decay of error for the PGA is less than 0.1898.

Let us figure out how the DGA$(1,b,u^2/2)$ works in Hilbert space. Consider its mth step. Let $\varphi_m \in \mathcal{D}$ be from (3.7.3). Then it is clear that φ_m maximizes $\langle f_{m-1}, g\rangle$ over the dictionary $\mathcal{D}$ and

$$\langle f_{m-1}, \varphi_m\rangle = \|f_{m-1}\| r_{\mathcal{D}}(f_{m-1}).$$

The PGA would use φ_m with the coefficient $\langle f_{m-1}, \varphi_m\rangle$ at this step. The DGA$(1,b,u^2/2)$ uses the same φ_m and only a fraction of $\langle f_{m-1}, \varphi_m\rangle$:

$$c_m = b\|f_{m-1}\| r_{\mathcal{D}}(f_{m-1}). \tag{3.7.43}$$

Thus the choice $b = 1$ in (3.7.43) corresponds to the PGA. However, it is clear from the above considerations that our technique, designed for general Banach spaces, does not work in the case $b = 1$. The above discussion brings us the following surprising observation. The use of a small fraction $(c_m = b\langle f_{m-1}, g\rangle)$ of an optimal coefficient results in an improvement of the upper estimate for the rate of convergence.

3.7.4. Convergence of the WDGA

We now study convergence of the Weak Dual Greedy Algorithm (WDGA) defined in the Introduction of this chapter. We present in this subsection results from Ganichev and Kalton (2003). We will prove the convergence result under an extra assumption on a Banach space X.

Definition 3.7.11. (Property Γ) A uniformly smooth Banach space has property Γ if there is a constant $\beta > 0$ such that, for any $x, y \in X$ satisfying $F_x(y) = 0$, we have

$$\|x + y\| \geq \|x\| + \beta F_{x+y}(y).$$

Property Γ in the above form was introduced in Ganichev and Kalton (2003). This condition (formulated somewhat differently) was considered previously in the context of greedy approximation in Livshitz (2003).

Theorem 3.7.12. Let X be a uniformly smooth Banach space with property Γ. Then the WDGA(τ) with $\tau = \{t\}$, $t \in (0, 1]$, converges for each dictionary and all $f \in X$.

Proof. Let $\{f_m\}_{m=0}^{\infty}$ be a sequence generated by the WDGA(t). Then

$$f_{m-1} = f_m + a_m \varphi_m, \quad F_{f_m}(\varphi_m) = 0. \tag{3.7.44}$$

We use property Γ with $x := f_m$ and $y := a_m \varphi_m$ and obtain

$$\|f_{m-1}\| \geq \|f_m\| + \beta a_m F_{f_{m-1}}(\varphi_m). \tag{3.7.45}$$

This inequality, and monotonicity of the sequence $\{\|f_m\|\}$, imply that

$$\sum_{m=1}^{\infty} a_m F_{f_{m-1}}(\varphi_m) < \infty \;\Rightarrow\; \sum_{m=1}^{\infty} a_m r_{\mathcal{D}}(f_{m-1}) < \infty. \tag{3.7.46}$$

As in the proof of Theorem 3.7.1, we consider separately two cases:

$$\text{(I)} \quad \sum_{m=1}^{\infty} a_m = \infty, \qquad \text{(II)} \quad \sum_{m=1}^{\infty} a_m < \infty.$$

In case (I), by (3.7.46) and Lemma 3.7.3 we obtain

$$\liminf_{m\to\infty} \|f_m\| = 0 \;\Rightarrow\; \lim_{m\to\infty} \|f_m\| = 0.$$

In case (II) we argue by contradiction. Assume

$$\lim_{m\to\infty} \|f_m\| = \alpha > 0.$$

Then, by (II) we have $f_m \to f_\infty \neq 0$ as $m \to \infty$. By uniform smoothness of X we get

$$\lim_{m\to\infty} \|F_{f_m} - F_{f_\infty}\| = 0, \quad \lim_{m\to\infty} \|F_{f_m} - F_{f_{m-1}}\| = 0. \tag{3.7.47}$$

In particular, (3.7.44) and (3.7.47) imply that

$$\lim_{m\to\infty} F_{f_{m-1}}(\varphi_m) = 0 \;\Rightarrow\; \lim_{m\to\infty} r_{\mathcal{D}}(f_m) = 0. \tag{3.7.48}$$

We have $F_{f_\infty} \neq 0$, and therefore there is a $g \in \mathcal{D}$ such that $F_{f_\infty}(g) > 0$. However, by (3.7.47) and (3.7.48),

$$F_{f_\infty}(g) = \lim_{m\to\infty} F_{f_m}(g) \le \lim_{m\to\infty} r_{\mathcal{D}}(f_m) = 0.$$

The obtained contradiction completes the proof.

We now give a direct proof in case (I) that does not use Lemma 3.7.3. By property Γ we get

$$\|f_m\| \le \|f_{m-1}\| - \beta a_m F_{f_{m-1}}(\varphi_m) \le \|f_{m-1}\| - t\beta a_m \|F_{f_{m-1}}\|_{\mathcal{D}}. \tag{3.7.49}$$

Let $\epsilon > 0$, $A(\epsilon) > 0$, and f^ϵ be such that

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

Then

$$\begin{aligned}\|f_{m-1}\| &= F_{f_{m-1}}(f_{m-1}) = F_{f_{m-1}}(f - f^\epsilon + f^\epsilon - G_{m-1})\\ &\le \epsilon + \|F_{f_{m-1}}\|_{\mathcal{D}}(A(\epsilon) + b_m),\end{aligned}$$

where $b_m := \sum_{k=1}^{m-1} a_k$. Therefore,

$$\|F_{f_{m-1}}\|_{\mathcal{D}} \ge (\|f_{m-1}\| - \epsilon)/(A(\epsilon) + b_m). \tag{3.7.50}$$

We complete the proof by obtaining a contradiction. If $\lim_{m\to\infty} \|f_m\| = \alpha > 0$, and $\epsilon := \alpha/2$, then (3.7.49) and (3.7.50) imply

$$\|f_m\| \le \|f_{m-1}\| \left(1 - \frac{t\beta a_m}{2(A(\epsilon) + b_m)}\right).$$

Assumption (I) implies

$$\sum_{m=1}^{\infty} \frac{a_m}{A(\epsilon) + b_m} = \infty \;\Rightarrow\; \|f_m\| \to 0. \qquad \square$$

We now turn to the L_p-spaces. The following results, Proposition 3.7.13 and Theorem 3.7.14, are from Ganichev and Kalton (2003).

Proposition 3.7.13. The L_p-space with $1 < p < \infty$ has property Γ.

Proof. Let $p \in (1, \infty)$. Consider the following function:

$$\phi_p(u) := \frac{u|1+u|^{p-2}(1+u) - u}{|1+u|^p - pu - 1}, \quad u \neq 0, \quad \phi_p(0) := 2/p.$$

We note that $|1+u|^p - pu - 1 > 0$ for $u \neq 0$. Indeed, it is sufficient to check the inequality for $u \ge -1/p$. In this case $|1+u|^p = (1+u)^p > 1 + pu$, $u \neq 0$.

It is easy to check that

$$\lim_{u\to 0} \phi_p(u) = 2/p.$$

Thus, $\phi_p(u)$ is continuous on $(-\infty, \infty)$. This and

$$\lim_{u\to -\infty} \phi_p(u) = \lim_{u\to \infty} \phi_p(u) = 1$$

imply that $\phi_p(u) \leq C_p$.

We now proceed to property Γ. For any two real functions $x(s)$, $y(s)$, the inequality $\phi_p(u) \leq C_p$ implies

$$\begin{aligned} &|x(s)+y(s)|^{p-2}(x(s)+y(s))y(s) - |x(s)|^{p-2}x(s)y(s) \\ &\qquad \leq C_p(|x(s)+y(s)|^p - p|x(s)|^{p-2}x(s)y(s) - |x(s)|^p). \end{aligned} \tag{3.7.51}$$

Suppose that $F_x(y) = 0$. This means that

$$\int |x(s)|^{p-2}x(s)y(s)\,\mathrm{d}s = 0. \tag{3.7.52}$$

Integrating inequality (3.7.51) and taking into account (3.7.52), we get

$$\|x+y\|^{p-1}F_{x+y}(y) \leq C_p(\|x+y\|^p - \|x\|^p). \tag{3.7.53}$$

Next,

$$\|x\| = F_x(x) = F_x(x+y) \leq \|x+y\|.$$

Therefore, (3.7.53) implies

$$F_{x+y}(y) \leq pC_p(\|x+y\| - \|x\|). \tag{3.7.54}$$

It remains to note that (3.7.54) is equivalent to property Γ with $\beta = (pC_p)^{-1}$. □

Combining Theorem 3.7.12 with Proposition 3.7.13 we obtain the following result.

Theorem 3.7.14. Let $p \in (1, \infty)$. Then the WDGA(τ) with $\tau = \{t\}$, $t \in (0, 1]$, converges for each dictionary and all $f \in L_p$.

3.8. Relaxation; *X*-greedy algorithms

In Sections 3.2–3.7 we studied dual greedy algorithms. In this section we define some generalizations of the X-Greedy Algorithm using the idea of relaxation. We begin with an analogue of the WGAFR.

***X*-Greedy Algorithm with Free Relaxation (XGAFR).** We define $f_0 := f$ and $G_0 := 0$. Then, for each $m \geq 1$ we have the following inductive definition.

(1) $\varphi_m \in \mathcal{D}$ and $\lambda_m \geq 0$, w_m are such that

$$\|f - ((1-w_m)G_{m-1} + \lambda_m \varphi_m)\| = \inf_{g\in\mathcal{D},\lambda\geq 0,w} \|f - ((1-w)G_{m-1} + \lambda g)\|$$

and

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

(2) Let

$$f_m := f - G_m.$$

Using this definition, we obtain that for any $t \in (0,1]$

$$\|f_m\| \leq \inf_{\lambda\geq 0,w} \|f - ((1-w)G_{m-1} + \lambda\varphi_m^t)\|,$$

where the $\varphi_m^t \in \mathcal{D}$ is an element satisfying

$$F_{f_{m-1}}(\varphi_m^t) \geq t\|F_{f_{m-1}}\|_{\mathcal{D}}.$$

Setting $t = 1$ we obtain a version of Lemma 3.4.1 for the XGAFR.

Lemma 3.8.1. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Take a number $\epsilon \geq 0$ and two elements f, f^ϵ from X such that

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

with some number $A(\epsilon) \geq \epsilon$. Then we have for the XGAFR

$$\|f_m\| \leq \|f_{m-1}\| \inf_{\lambda\geq 0}\left(1 - \lambda A(\epsilon)^{-1}\left(1 - \frac{\epsilon}{\|f_{m-1}\|}\right) + 2\rho\left(\frac{5\lambda}{\|f_{m-1}\|}\right)\right),$$

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

Theorems 3.4.3 and 3.4.4 were derived from Lemma 3.4.1. In the same way we derive from Lemma 3.8.1 the following analogues of Theorems 3.4.3 and 3.4.4 for the XGAFR.

Theorem 3.8.2. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Then, for any $f \in X$ we have for the XGAFR

$$\lim_{m\to\infty} \|f_m\| = 0.$$

Theorem 3.8.3. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Take a number $\epsilon \geq 0$ and two elements f, f^ϵ from X such that

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

with some number $A(\epsilon) > 0$. Then we have for the XGAFR

$$\|f_m\| \leq \max\bigl(2\epsilon, C(q,\gamma)(A(\epsilon)+\epsilon)(1+m)^{-1/p}\bigr), \quad p := q/(q-1).$$

We now proceed to an analogue of the GAWR.

***X*-Greedy Algorithm with Relaxation r (XGAR(r)).** Given a relaxation sequence $\mathbf{r} := \{r_m\}_{m=1}^\infty$, $r_m \in [0,1)$, we define $f_0 := f$ and $G_0 := 0$. Then, for each $m \geq 1$ we have the following inductive definition.

(1) $\varphi_m \in \mathcal{D}$ and $\lambda_m \geq 0$ are such that

$$\|f - ((1-r_m)G_{m-1} + \lambda_m \varphi_m)\| = \inf_{g \in \mathcal{D}, \lambda \geq 0} \|f - ((1-r_m)G_{m-1} + \lambda g)\|$$

and

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

(2) Let

$$f_m := f - G_m.$$

We note that in the case $r_k = 0$, $k = 1, \dots$, when there is no relaxation, the XGAR($\mathbf{0}$) coincides with the X-Greedy Algorithm. Practically nothing is known about convergence and rate of convergence of the X-Greedy Algorithm. However, relaxation helps to prove convergence results for the XGAR($\mathbf{r}$). Here are analogues of the corresponding results for the GAWR.

Lemma 3.8.4. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u)$. Take a number $\epsilon \geq 0$ and two elements f, f^ϵ from X such that

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

with some number $A(\epsilon) > 0$. Then we have for the XGAR($\mathbf{r}$)

$$\|f_m\| \leq \|f_{m-1}\| \left(1 - r_m\left(1 - \frac{\epsilon}{\|f_{m-1}\|}\right) + 2\rho\left(\frac{r_m(\|f\| + A(\epsilon))}{(1-r_m)\|f_{m-1}\|}\right)\right),$$

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

Theorem 3.8.5. Let a sequence $\mathbf{r} := \{r_k\}_{k=1}^\infty$, $r_k \in [0,1)$, satisfy the conditions

$$\sum_{k=1}^\infty r_k = \infty, \quad \text{and} \quad r_k \to 0 \quad \text{as } k \to \infty.$$

Then the XGAR($\mathbf{r}$) converges in any uniformly smooth Banach space for each $f \in X$ and for all dictionaries $\mathcal{D}$.

Theorem 3.8.6. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Let $\mathbf{r} := \{2/(k+2)\}_{k=1}^\infty$. Consider the XGAR($\mathbf{r}$). For a pair of functions f, f^ϵ satisfying

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

we have

$$\|f_m\| \leq \epsilon + C(q,\gamma)(\|f\| + A(\epsilon)) m^{-1+1/q}.$$

3.9. Greedy algorithms with approximate evaluations and restricted search

In this section we study a modification of the WCGA that is motivated by numerical applications. In this modification, we allow steps of the WCGA to be performed approximately with some error control. We show that the modified version of the WCGA performs as well as the WCGA. We develop the theory of the Approximate Weak Chebyshev Greedy Algorithm in a general setting: X is an arbitrary uniformly smooth Banach space and $\mathcal{D}$ is any dictionary. We begin with some remarks on the WCGA. It is clear that in the case of an infinite dictionary $\mathcal{D}$ there is no direct computationally feasible way to evaluate $\sup_{g\in\mathcal{D}} F_{f^c_{m-1}}(g)$. This makes the greedy step, even in a weak version, very difficult to realize in practice. At the second step of the WCGA we are looking for the best approximant of f from Φ_m. We know that such an approximant exists. However, in practice we cannot find it exactly: we can only find it approximately.

The above observations motivated us to consider a variant of the WCGA with an eye towards practically implementable algorithms. We note that Approximate Weak Greedy Algorithms in Hilbert spaces were studied in Gribonval and Nielsen (2001*a*) and Galatenko and Livshitz (2003, 2005).

In Temlyakov (2005*a*) we studied the following modification of the WCGA. Let three sequences $\tau = \{t_k\}_{k=1}^\infty$, $\delta = \{\delta_k\}_{k=0}^\infty$, $\eta = \{\eta_k\}_{k=1}^\infty$ of numbers from $[0,1]$ be given.

Approximate Weak Chebyshev Greedy Algorithm (AWCGA). We define $f_0 := f_0^{\tau,\delta,\eta} := f$. Then, for each $m \geq 1$ we have the following inductive definition.

(1) F_{m-1} is a functional with properties

$$\|F_{m-1}\| \leq 1, \qquad F_{m-1}(f_{m-1}) \geq \|f_{m-1}\|(1-\delta_{m-1});$$

and $\varphi_m := \varphi_m^{\tau,\delta,\eta} \in \mathcal{D}$ is any element satisfying

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

(2) Define

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

and let

$$E_m(f) := \inf_{\varphi\in\Phi_m} \|f-\varphi\|.$$

Let $G_m \in \Phi_m$ be such that

$$\|f - G_m\| \leq E_m(f)(1+\eta_m).$$

(3) Let

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

The term *approximate* in this definition means that we use a functional F_{m-1} that is an approximation to the norming (peak) functional $F_{f_{m-1}}$ and also that we use an approximant $G_m \in \Phi_m$ which satisfies a weaker assumption than being a best approximant to f from Φ_m. Thus, in the *approximate* version of the WCGA, we have addressed the issue of non-exact evaluation of the norming functional and the best approximant. We did not address the issue of finding the $\sup_{g\in\mathcal{D}} F_{f^c_{m-1}}(g)$. In the paper Temlyakov (2005*b*) we addressed this issue. We did it in two steps. First we considered the corresponding modification of the WCGA, and then the modification of the AWCGA. These modifications are done in the style of the concept of *depth search* from Donoho (2001).

We now consider a countable dictionary $\mathcal{D} = \{\pm\psi_j\}_{j=1}^{\infty}$. We denote $\mathcal{D}(N) := \{\pm\psi_j\}_{j=1}^{N}$. Let $\mathcal{N} := \{N_j\}_{j=1}^{\infty}$ be a sequence of natural numbers.

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

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

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

(2) Define

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

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

(3) Let

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

We formulate some results from Temlyakov (2005*a*, 2005*b*) in a particular case of a uniformly smooth Banach space with modulus of smoothness of power type (see Temlyakov (2005*a*, 2005*b*) for the general case). The following theorem was proved in Temlyakov (2005*a*).

Theorem 3.9.1. Let a Banach space X have modulus of smoothness $\rho(u)$ of power type $1 < q \leq 2$, that is, $\rho(u) \leq \gamma u^q$. Assume that

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

and

$$\delta_m = o(t_m^p), \qquad \eta_m = o(t_m^p).$$

Then the AWCGA converges for any $f \in X$.

We now give two theorems from Temlyakov (2005*b*) on greedy algorithms with restricted search.

Theorem 3.9.2. Let a Banach space X have modulus of smoothness $\rho(u)$ of power type $1 < q \leq 2$, that is, $\rho(u) \leq \gamma u^q$. Assume that $\lim_{m\to\infty} N_m = \infty$ and

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

Then the RWCGA converges for any $f \in X$.

For $b > 0$, $K > 0$, we define the class

$$A_1^b(K, \mathcal{D}) := \{f : d(f, A_1(\mathcal{D}(n)) \leq Kn^{-b}, \quad n = 1, 2, \dots\}.$$

Here, $A_1(\mathcal{D}(n))$ is a convex hull of $\{\pm\psi_j\}_{j=1}^n$, and for a compact set F

$$d(f, F) := \inf_{\phi\in F} \|f - \phi\|.$$

Theorem 3.9.3. Let X be a uniformly smooth Banach space with modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Then, for $t \in (0, 1]$ there exist $C_1(t, \gamma, q, K)$, $C_2(t, \gamma, q, K)$ such that, for $\mathcal{N}$ with $N_m \geq C_1(t, \gamma, q, K)m^{r/b}$, $m = 1, 2, \dots$, we have for any $f \in A_1^b(K, \mathcal{D})$

$$\|f_m^{c,\tau,\mathcal{N}}\| \leq C_2(t, \gamma, q, K)m^{-r}, \quad \tau = \{t\}, \quad r := 1 - 1/q.$$

We note that we can choose an algorithm from Theorem 3.9.3 that satisfies the *polynomial depth search* condition $N_m \leq Cm^a$ from Donoho (2001).

We proceed to an algorithm that combines approximate evaluations with restricted search. Let three sequences $\tau = \{t_k\}_{k=1}^\infty$, $\delta = \{\delta_k\}_{k=0}^\infty$, $\eta = \{\eta_k\}_{k=1}^\infty$ of numbers from $[0, 1]$ be given. Let $\mathcal{N} := \{N_j\}_{j=1}^\infty$ be a sequence of natural numbers.

Restricted Approximate Weak Chebyshev Greedy Algorithm (RAWCGA). We define $f_0 := f_0^{\tau,\delta,\eta,\mathcal{N}} := f$. Then, for each $m \geq 1$ we have the following inductive definition.

(1) F_{m-1} is a functional with properties

$$\|F_{m-1}\| \leq 1, \qquad F_{m-1}(f_{m-1}) \geq \|f_{m-1}\|(1 - \delta_{m-1}),$$

and $\varphi_m := \varphi_m^{\tau,\delta,\eta,\mathcal{N}} \in \mathcal{D}(N_m)$ is any element satisfying

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

(2) Define

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

and let

$$E_m(f) := \inf_{\varphi \in \Phi_m} \|f - \varphi\|.$$

Let $G_m \in \Phi_m$ be such that

$$\|f - G_m\| \le E_m(f)(1 + \eta_m).$$

(3) Let

$$f_m := f_m^{\tau,\delta,\eta,\mathcal{N}} := f - G_m.$$

Theorem 3.9.4. Let a Banach space X have modulus of smoothness $\rho(u)$ of power type $1 < q \le 2$, that is, $\rho(u) \le \gamma u^q$. Assume $\lim_{m\to\infty} N_m = \infty$,

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

and

$$\delta_m = o(t_m^p), \qquad \eta_m = o(t_m^p).$$

Then the RAWCGA converges for any $f \in X$.

We now make some general remarks on m-term approximation with the depth search constraint. The depth search constraint means that for a given m we restrict ourselves to systems of elements (subdictionaries) containing at most $N := N(m)$ elements. Let X be a linear metric space and for a set $\mathcal{D} \subset X$, let $\mathcal{L}_m(\mathcal{D})$ denote the collection of all linear subspaces spanned by m elements of $\mathcal{D}$. For a linear subspace $L \subset X$, the ϵ-neighbourhood $U_\epsilon(L)$ of L is the set of all $x \in X$ which are at a distance not exceeding ϵ from L (*i.e.*, those $x \in X$ which can be approximated to an error not exceeding ϵ by the elements of L). For any compact set $F \subset X$ and any integers $N, m \ge 1$, we define the (N, m)-entropy numbers (see Temlyakov (2003*a*, p. 94))

$$\epsilon_{N,m}(F, X) := \inf_{\#\mathcal{D}=N} \inf\{\epsilon : F \subset \cup_{L\in\mathcal{L}_m(\mathcal{D})} U_\epsilon(L)\}.$$

We let $\Sigma_m(\mathcal{D})$ denote the collection of all functions (elements) in X which can be expressed as a linear combination of at most m elements of $\mathcal{D}$. Thus each function $s \in \Sigma_m(\mathcal{D})$ can be written in the form

$$s = \sum_{g\in\Lambda} c_g g, \quad \Lambda \subset \mathcal{D}, \quad \#\Lambda \le m,$$

where the c_g are real or complex numbers. For a function $f \in X$ we define its best m-term approximation error

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

For a function class $F \subset X$ we define

$$\sigma_m(F) := \sigma_m(F, \mathcal{D}) := \sup_{f\in F} \sigma_m(f, \mathcal{D}).$$

We can express $\sigma_m(F, \mathcal{D})$ as

$$\sigma_m(F, \mathcal{D}) = \inf\{\epsilon : F \subset \cup_{L \in \mathcal{L}_m(\mathcal{D})} U_\epsilon(L)\}.$$

It follows therefore that

$$\inf_{\#\mathcal{D}=N} \sigma_m(F, \mathcal{D}) = \epsilon_{N,m}(F, X).$$

In other words, finding best dictionaries consisting of N elements for m-term approximation of F is the same as finding sets $\mathcal{D}$ which attain the (N, m)-entropy numbers $\epsilon_{N,m}(F, X)$. It is easy to see that $\epsilon_{m,m}(F, X) = d_m(F, X)$, where $d_m(F, X)$ is the Kolmogorov width of F in X. This establishes a connection between (N, m)-entropy numbers and the Kolmogorov widths. One can find a further discussion on the nonlinear Kolmogorov (N, m)-widths and the entropy numbers in Temlyakov (2003*a*).

3.10. An application of greedy algorithms for the discrepancy estimates

Let $1 \le p < \infty$. We recall the definition of the L_p-discrepancy of points $\{\xi^1, \dots, \xi^m\} \subset \Omega_d := [0, 1]^d$. Let $\chi_{[a,b]}(\cdot)$ be the characteristic function of the interval $[a, b]$. For $x, y \in \Omega_d$, let

$$B(x, y) := \prod_{j=1}^{d} \chi_{[0,x_j]}(y_j).$$

Then the L_p-discrepancy of $\xi := \{\xi^1, \dots, \xi^m\} \subset \Omega_d$ is defined by

$$D(\xi, m, d)_p := \left\| \int_{\Omega_d} B(x, y)\, \mathrm{d}y - \frac{1}{m} \sum_{\mu=1}^{m} B(x, \xi^\mu) \right\|_{L_p(\Omega_d)}.$$

It will be convenient for us to study a slight modification of $D(\xi, m, d)_p$. For $a, t \in [0, 1]$, let

$$H(a, t) := \chi_{[0,a]}(t) - \chi_{[a,1]}(t),$$

and for $x, y \in \Omega_d$

$$H(x, y) := \prod_{j=1}^{d} H(x_j, y_j).$$

We define the symmetrized L_p-discrepancy by

$$D^s(\xi, m, d)_p := \left\| \int_{\Omega_d} H(x, y)\, \mathrm{d}y - \frac{1}{m} \sum_{\mu=1}^{m} H(x, \xi^\mu) \right\|_{L_p(\Omega_d)}.$$

The L_∞-discrepancies $D(\xi, m, d)_\infty$ and $D^s(\xi, m, d)_\infty$ are defined in the same way, with the L_p-norm replaced by the L_∞-norm.

Using the identity

$$\chi_{[0,x_j]}(y_j) = \frac{1}{2}(H(1,y_j) + H(x_j,y_j)),$$

we get a simple inequality,

$$D(\xi,m,d)_\infty \le D^s(\xi,m,d)_\infty. \tag{3.10.1}$$

We are interested in ξ with small discrepancy. Consider

$$D(m,d)_p := \inf_\xi D(\xi,m,d)_p, \qquad D^s(m,d)_p := \inf_\xi D^s(\xi,m,d)_p.$$

For $1 < p < \infty$ the following relation is known,

$$D(m,d)_p \asymp m^{-1}(\ln m)^{(d-1)/2} \tag{3.10.2}$$

(see Beck and Chen (1987, p. 5)), with constants in $\asymp$ depending on p and d. The correct order of $D(m,d)_p$, $p = 1,\infty$, for $d \ge 3$ is unknown. The following estimate was obtained in Heinrich, Novak, Wasilkowski and Wozniakowski (2001):

$$D(m,d)_\infty \le Cd^{1/2}m^{-1/2}. \tag{3.10.3}$$

It is pointed out in Heinrich *et al.* (2001) that (3.10.3) is only an existence theorem and even the constant C in (3.10.3) is unknown. Their proof is a probabilistic one. There are also some other estimates in Heinrich *et al.* (2001) with explicit constants. We mention one of them,

$$D(m,d)_\infty \le C(d\ln d)^{1/2}((\ln m)/m)^{1/2}, \tag{3.10.4}$$

with an explicit constant C. The proof of (3.10.4) is also probabilistic.

In this section we apply greedy-type algorithms to obtain upper estimates of $D(m,d)_p$, $1 \le p \le \infty$ in the style of (3.10.3) and (3.10.4). The important feature of our proof is that it is deterministic, and moreover it is constructive. Formally, the optimization problem

$$D(m,d)_p = \inf_\xi D(\xi,m,d)_p$$

is deterministic: one needs to minimize over $\{\xi^1,\dots,\xi^m\} \subset \Omega_d$. However, minimization by itself does not provide any upper estimate. It is known (see Davis *et al.* (1997)) that simultaneous optimization over many parameters ($\{\xi^1,\dots,\xi^m\}$ in our case) is a very difficult problem. We note that

$$D(m,d)_p = \sigma^e_m(J,\mathcal{B})_p := \inf_{g_1,\dots,g_m\in\mathcal{B}} \|J(\cdot) - \frac{1}{m}\sum_{\mu=1}^m g_\mu\|_{L_p(\Omega_d)},$$

where

$$J(x) = \int_{\Omega_d} B(x,y)\,\mathrm{d}y$$

and

$$\mathcal{B} = \{B(x, y) : y \in \Omega_d\}.$$

It was proved in Davis *et al.* (1997) that if an algorithm finds the best m-term approximation for each $f \in \mathbb{R}^N$ for every dictionary $\mathcal{D}$, with the number of elements of order N^k, $k \geq 1$, then this algorithm solves an NP-hard problem. Thus, in nonlinear m-term approximation we look for methods (algorithms) which provide approximation close to best m-term approximation, and at each step solve an optimization problem over only one parameter (ξ^μ in our case). In this section we will provide such an algorithm for estimating $\sigma_m^e(J, \mathcal{B})_p$. We call this algorithm 'constructive' because it provides an explicit construction with feasible one-parameter optimization steps.

We proceed to the construction. We will use in our construction the IA(ϵ) which was studied in Section 3.6. We will use the following corollaries of Theorem 3.6.2.

Corollary 3.10.1. We apply Theorem 3.6.2 for $X = L_p(\Omega_d)$, $p \in [2, \infty)$, $\mathcal{D}^+ = \{H(x, y) : y \in \Omega_d\}$, $f = J^s(x)$, where

$$J^s(x) = \int_{\Omega_d} H(x, y)\, \mathrm{d}y \in A_1(\mathcal{D}^+).$$

Using (3.1.12), we get by Theorem 3.6.2 a constructive set $\xi^1, \ldots, \xi^m$, such that

$$D^s(\xi, m, d)_p = \|(J^s)_m^{i,\epsilon}\|_{L_p(\Omega_d)} \leq Cp^{1/2}m^{-1/2},$$

with absolute constant C.

Corollary 3.10.2. We apply Theorem 3.6.2 for $X = L_p(\Omega_d)$, $p \in [2, \infty)$, $\mathcal{D}^+ = \{B(x, y) : y \in \Omega_d\}$, $f = J(x)$, where

$$J(x) = \int_{\Omega_d} B(x, y)\, \mathrm{d}y \in A_1(\mathcal{D}^+).$$

Using (3.1.12), we get by Theorem 3.6.2 a constructive set $\xi^1, \ldots, \xi^m$, such that

$$D(\xi, m, d)_p = \|J_m^{i,\epsilon}\|_{L_p(\Omega_d)} \leq Cp^{1/2}m^{-1/2},$$

with absolute constant C.

Corollary 3.10.3. We apply Theorem 3.6.2 for $X = L_p(\Omega_d)$, $p \in [2, \infty)$, $\mathcal{D}^+ = \{B(x, y)/\|B(\cdot, y)\|_{L_p(\Omega_d)} : y \in \Omega_d\}$, $f = J(x)$. Using (3.1.12), we get by Theorem 3.6.2 a constructive set $\xi^1,\dots,\xi^m$ such that

$$\left\|\int_{\Omega_d} B(x,y)\,\mathrm{d}y - \frac{1}{m}\sum_{\mu=1}^m \left(\frac{p}{p+1}\right)^d \left(\prod_{j=1}^d (1-\xi_j^\mu)^{-1/p}\right) B(x,\xi^\mu)\right\|_{L_p(\Omega_d)} \le C\left(\frac{p}{p+1}\right)^d p^{1/2} m^{-1/2},$$

with absolute constant C.

We note that in the case $X = L_p(\Omega_d)$, $p \in [2,\infty)$, $\mathcal{D}^+ = \{H(x,y) : x \in \Omega_d\}$, $f = J^s(y)$, the implementation of the IA(ϵ) is a sequence of maximization steps when we maximize functions of d variables. An important advantage of the L_p-spaces is a simple and explicit form of the norming functional F_f of a function $f \in L_p(\Omega_d)$. The F_f acts as (for real L_p-spaces)

$$F_f(g) = \int_{\Omega_d} \|f\|_p^{1-p} |f|^{p-2} f g \,\mathrm{d}y.$$

Thus the IA(ϵ) should find at a step m an approximate solution to the following optimization problem (over $y \in \Omega_d$)

$$\int_{\Omega_d} |f_{m-1}^{i,\epsilon}(x)|^{p-2} f_{m-1}^{i,\epsilon}(x) H(x,y)\,\mathrm{d}x \;\to\; \max.$$

Let us discuss one possible application of the WRGA instead of the IA(ϵ). An obvious change is that instead of the cubature formula

$$\frac{1}{m}\sum_{\mu=1}^m H(x,\xi^\mu),$$

in the case of IA(ϵ), we have the cubature formula

$$\sum_{\mu=1}^m w_\mu^m H(x,\xi^\mu), \qquad \sum_{\mu=1}^m |w_\mu^m| \le 1,$$

in the case of the WRGA. It is a disadvantage of the WRGA. An advantage of the WRGA is that we are more flexible in selecting an element φ_m^r satisfying

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

than an element $\varphi_m^{i,\epsilon}$ satisfying

$$F_{f_{m-1}^{i,\epsilon}}(\varphi_m^{i,\epsilon} - f) \ge -\epsilon_m.$$

We will now derive an estimate for $D(m,d)_\infty$ from Corollary 3.10.2.

Proposition 3.10.4. For any m there exists a constructive set

$$\xi = \{\xi^1, \dots, \xi^m\} \subset \Omega_d$$

such that

$$D(\xi, m, d)_\infty \le C d^{3/2} (\max(\ln d, \ln m))^{1/2} m^{-1/2}, \quad d, m \ge 2 \tag{3.10.5}$$

with an effective absolute constant C.

Proof. We use the inequality

$$D(\xi, m, d)_\infty \le c(d, p) d(3d + 4) D(\xi, m, d)_p^{p/(p+d)}, \tag{3.10.6}$$

from Niederreiter, Tichy and Turnwald (1990), and the estimate for $c(d, p)$

$$c(d, p) \le 3^{1/3} d^{-1+2/(1+p/d)}, \tag{3.10.7}$$

from Heinrich *et al.* (2001). Specifying $p = d \max(\ln d, \ln m)$ and using Corollary 3.10.2 we get (3.10.5) from (3.10.6) and (3.10.7). □

REFERENCES

A. R. Barron (1993), 'Universal approximation bounds for superposition of n sigmoidal functions', *IEEE Trans. Inform. Theory* **39**, 930–945.

A. Barron, A. Cohen, W. Dahmen and R. DeVore (2005), Approximation and learning by greedy algorithms. Manuscript.

B. M. Baishanski (1983), 'Approximation by polynomials of given length', *Illinois J. Math.* **27**, 449–458.

N. K. Bary (1961), *Trigonometric Series*, Nauka, Moscow (in Russian). English translation: Pergamon Press, Oxford (1964).

J. Beck and W. Chen (1987), *Irregularities of Distribution*, Cambridge University Press.

W. Bednorz (2006), Greedy bases are best for m-term approximation. Manuscript.

M. S. Birman and M. Z. Solomyak (1977), 'Estimates of singular numbers of integral operators', *Uspekhi Mat. Nauk* **32**, 17–84. English translation in *Russian Math. Surveys* **32** (1977).

J. Bourgain (1992), 'A remark on the behaviour of L^p-multipliers and the range of operators acting on L^p-spaces', *Israel J. Math.* **79**, 193–206.

E. Candès (2006), Compressive sampling. In *Proc. International Congress of Mathematics* (Madrid 2006), Vol. 3, pp. 1433–1452.

E. Candès, J. Romberg and T. Tao (2006), 'Stable signal recovery from incomplete and inaccurate measurements', *Comm. Pure Appl. Math.* **59**, 1207–1223.

E. Candès and T. Tao (2005), 'Decoding by linear programming', *IEEE Trans. Inform. Theory* **51**, 4203–4215.

B. Carl (1981), 'Entropy numbers, s-numbers, and eigenvalue problems', *J. Funct. Anal.* **41**, 290–306.

S. S. Chen, D. L. Donoho and M. A. Saunders (2001), 'Atomic decomposition by basis pursuit', *SIAM Review* **43**, 129–159.

J. A. Cochran (1977), 'Composite integral operators and nuclearity', *Ark. Mat.* **15**, 215–222.

A. Cohen, R. A. DeVore and R. Hochmuth (2000), 'Restricted nonlinear approximation', *Constr. Approx.* **16**, 85–113.

A. Cohen, W. Dahmen and R. DeVore (2007), Compressed sensing and k-term approximation. Manuscript.

R. R. Coifman and M. V. Wickerhauser (1992), 'Entropy-based algorithms for best-basis selection', *IEEE Trans. Inform. Theory* **38**, 713–718.

A. Cordoba and P. Fernandez (1998), 'Convergence and divergence of decreasing rearranged Fourier series', *SIAM J. Math. Anal.* **29**, 1129–1139.

F. Cucker and S. Smale (2001), 'On the mathematical foundations of learning', *Bull. Amer. Math. Soc.* **39**, 1–49.

G. Davis, S. Mallat and M. Avellaneda (1997), 'Adaptive greedy approximations', *Constr. Approx.* **13**, 57–98.

R. A. DeVore (1998), Nonlinear approximation. In *Acta Numerica*, Vol. 7, Cambridge University Press, pp. 51–150.

R. A. DeVore (2006), Optimal computation. In *Proc. International Congress of Mathematics* (Madrid 2006), Vol. 1, pp. 187–215.

R. DeVore, B. Jawerth and V. Popov (1992), 'Compression of wavelet decompositions', *Amer. J. Math.* **114**, 737–785.

R. DeVore, G. Kerkyacharian, D. Picard and V. Temlyakov (2004), On mathematical methods of learning. IMI Preprint 10, Department of Mathematics, University of South Carolina.

R. DeVore, G. Kerkyacharian, D. Picard and V. Temlyakov (2006), 'Mathematical methods for supervised learning', *Found. Comput. Math.* **6**, 3–58.

R. A. DeVore, S. V. Konyagin and V. N. Temlyakov (1998), 'Hyperbolic wavelet approximation', *Constr. Approx.* **14**, 1–26.

R. A. DeVore and G. G. Lorenz (1993), *Constructive Approximation*, Springer, Berlin.

R. DeVore, G. Petrova and V. N. Temlyakov (2003), 'Best basis selection for approximation in L_p', *Found. Comput. Math.* **3**, 161–185.

R. A. DeVore and V. A. Popov (1988), Interpolation spaces and non-linear approximation. In *Function Spaces and Approximation*, Vol. 1302 of *Lecture Notes in Mathematics*, Springer, pp. 191–205.

R. A. DeVore and V. N. Temlyakov (1995), 'Nonlinear approximation by trigonometric sums', *J. Fourier Anal. Appl.* **2**, 29–48.

R. A. DeVore and V. N. Temlyakov (1996), 'Some remarks on greedy algorithms', *Adv. Comput. Math.* **5**, 173–187.

R. A. DeVore and V. N. Temlyakov (1997), 'Nonlinear approximation in finite-dimensional spaces', *J. Complexity* **13**, 489–508.

S. J. Dilworth, N. J. Kalton and D. Kutzarova (2003), 'On the existence of almost greedy bases in Banach spaces', *Studia Math.* **158**, 67–101.

S. J. Dilworth, N. J. Kalton, D. Kutzarova and V. N. Temlyakov (2003), 'The thresholding greedy algorithm, greedy bases, and duality', *Constr. Approx.* **19**, 575–597.

S. Dilworth, D. Kutzarova and V. Temlyakov (2002), 'Convergence of some greedy algorithms in Banach spaces', *J. Fourier Anal. Appl.* **8**, 489–505.

S. J. Dilworth, D. Kutzarova and P. Wojtaszczyk (2002), 'On approximate ℓ_1 systems in Banach spaces', *J. Approx. Theory* **114**, 214–241.

M. Donahue, L. Gurvits, C. Darken and E. Sontag (1997), 'Rate of convex approximation in non-Hilbert spaces', *Constr. Approx.* **13**, 187–220.

D. L. Donoho (1993), 'Unconditional bases are optimal bases for data compression and for statistical estimation', *Appl. Comput. Harmon. Anal.* **1**, 100–115.

D. L. Donoho (1997), 'CART and best-ortho-basis: A connection', *Ann. Statist.* **25**, 1870–1911.

D. L. Donoho (2001), 'Sparse components of images and optimal atomic decompositions', *Constr. Approx.* **17**, 353–382.

D. Donoho (2006), 'Compressed sensing', *IEEE Trans. Inform. Theory* **52**, 1289–1306.

D. Donoho, M. Elad and V. N. Temlyakov (2006), 'Stable recovery of sparse overcomplete representations in the presence of noise', *IEEE Trans. Inform. Theory* **52**, 6–18.

D. Donoho, M. Elad and V. N. Temlyakov (2007), 'On the Lebesgue type inequalities for greedy approximation', *J. Approx. Theory* **147**, 185–195.

D. Donoho and I. Johnstone (1994), 'Ideal spatial adaptation via wavelet shrinkage', *Biometrica* **81**, 425–455.

V. V. Dubinin (1997), Greedy algorithms and applications. PhD Thesis, University of South Carolina.

C. Fefferman and E. Stein (1972), 'H^p spaces of several variables', *Acta Math.* **129**, 137–193.

T. Figiel, W. B. Johnson and G. Schechtman (1988), 'Factorization of natural embeddings of ℓ_p^n into L_r I', *Studia Math.* **89**, 79–103.

M. Frazier and B. Jawerth (1990), 'A discrete transform and decomposition of distribution spaces', *J. Funct. Anal.* **93**, 34–170.

I. Fredholm (1903), 'Sur une classe d'équations fonctionelles', *Acta Math.* **27**, 365–390.

J. H. Friedman and W. Stuetzle (1981), 'Projection pursuit regression', *J. Amer. Statist. Assoc.* **76**, 817–823.

V. V. Galatenko and E. D. Livshitz (2003), 'On convergence of approximate weak greedy algorithms', *East J. Approx.* **9**, 43–49.

V. V. Galatenko and E. D. Livshitz (2005), 'Generalized approximate weak greedy algorithms', *Math. Notes* **78**, 170–184.

M. Ganichev and N. J. Kalton (2003), 'Convergence of the weak dual greedy algorithm in L_p-spaces', *J. Approx. Theory* **124**, 89–95.

A. Garnaev and E. Gluskin (1984), 'The widths of a Euclidean ball', *Dokl. Akad. Nauk USSR* **277**, 1048–1052. English translation in *Soviet Math. Dokl.* **30**, 200–204.

A. C. Gilbert, S. Muthukrishnan and M. J. Strauss (2003), Approximation of functions over redundant dictionaries using coherence. In *Proc. 14th Annual ACM–SIAM Symposium on Discrete Algorithms*, pp. 243–252.

S. Gogyan (2005), 'Greedy algorithm with regard to Haar subsystems', *East J. Approx.* **11**, 221–236.

S. Gogyan (2006), On convergence of weak thresholding greedy algorithm in $L^1(0,1)$. Manuscript.

R. Gribonval and M. Nielsen (2001*a*), 'Approximate weak greedy algorithms', *Adv. Comput. Math.* **14**, 361–368.

R. Gribonval and M. Nielsen (2001*b*), 'Some remarks on non-linear approximation with Schauder bases', *East J. Approx.* **7**, 267–285.

P. Habala, P. Hájek and V. Zizler (1996), *Introduction to Banach Spaces*, Vol. I, Matfyzpress, Univerzity Karlovy.

S. Heinrich, E. Novak, G. Wasilkowski and H. Wozniakowski (2001), 'The inverse of the star-discrepancy depends linearly on the dimension', *Acta Arith.* **96**, 279–302.

E. Hille and J. D. Tamarkin (1931), 'On the characteristic values of linear integral equations', *Acta Math.* **57**, 1–76.

P. J. Huber (1985), 'Projection pursuit', *Ann. Statist.* **13**, 435–475.

L. Jones (1987), 'On a conjecture of Huber concerning the convergence of projection pursuit regression', *Ann. Statist.* **15**, 880–882.

L. Jones (1992), 'A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training', *Ann. Statist.* **20**, 608–613.

N. J. Kalton, N. T. Beck and J. W. Roberts (1984), *An F-Space Sampler*, Vol. 5 of *London Math. Soc. Lecture Notes*, Cambridge University Press, Cambridge.

A. Kamont and V. N. Temlyakov (2004), 'Greedy approximation and the multivariate Haar system', *Studia Math.* **161**, 199–223.

B. S. Kashin (1977), 'Widths of certain finite-dimensional sets and classes of smooth functions', *Izv. Akad. Nauk SSSR, Ser. Mat.* **41**, 334–351. English translation in *Math. USSR IZV.* **11**.

B. S. Kashin (1985), 'On approximation properties of complete orthonormal systems', *Tr. Mat. Inst. Steklova* **172**, 187–191. English translation in *Proc. Steklov Inst. Math.* **3**, 207–211.

B. S. Kashin and V. N. Temlyakov (2007), A remark on compressed sensing. Manuscript.

G. Kerkyacharian and D. Picard (2004), 'Entropy, universal coding, approximation, and bases properties', *Constr. Approx.* **20**, 1–37.

G. Kerkyacharian, D. Picard and V. N. Temlyakov (2006), 'Some inequalities for the tensor product of greedy bases and weight-greedy bases', *East J. Approx.* **12**, 103–118.

S. V. Konyagin and M. A. Skopina (2001), 'Comparison of the L_1-norms of total and truncated exponential sums', *Mat. Zametki* **69**, 699–707.

S. V. Konyagin and V. N. Temlyakov (1999*a*), 'A remark on greedy approximation in Banach spaces', *East J. Approx.* **5**, 1–15.

S. V. Konyagin and V. N. Temlyakov (1999*b*), 'Rate of convergence of pure greedy algorithms', *East J. Approx.* **5**, 493–499.

S. V. Konyagin and V. N. Temlyakov (2002), 'Greedy approximation with regard to bases and general minimal systems', *Serdica Math. J.* **28**, 305–328.

S. V. Konyagin and V. N. Temlyakov (2003*a*), 'Convergence of greedy approximation I: General systems', *Studia Math.* **159**, 143–160.

S. V. Konyagin and V. N. Temlyakov (2003*b*), 'Convergence of greedy approximation II: The trigonometric system', *Studia Math.* **159**, 161–184.

S. V. Konyagin and V. N. Temlyakov (2004), Some error estimates in learning theory. In *Approximation Theory: A Volume Dedicated to Borislav Bojanov*, Marin Drinov Academy Publishing House, Sofia, pp. 126–144.

S. V. Konyagin and V. N. Temlyakov (2005), 'Convergence of greedy approximation for the trigonometric system', *Anal. Math.* **31**, 85–115.

S. V. Konyagin and V. N. Temlyakov (2007), 'The entropy in learning theory: Error estimates', *Constr. Approx.* **25**, 1–27.

T. W. Körner (1996), 'Divergence of decreasing rearranged Fourier series', *Ann. of Math.* **144**, 167–180.

T. W. Körner (1999), 'Decreasing rearranged Fourier series', *J. Fourier Anal. Appl.* **5**, 1–19.

H. Lebesgue (1909), 'Sur les intégrales singulières', *Ann. Fac. Sci. Univ. Toulouse* (3) **1**, 25–117.

W. S. Lee, P. L. Bartlett and R. C. Williamson (1996), 'Efficient agnostic learning of neural networks with bounded fan-in', *IEEE Trans. Inform. Theory* **42**, 2118–2132.

D. Leviatan and V. N. Temlyakov (2005), 'Simultaneous greedy approximation in Banach spaces', *J. Complexity* **21**, 275–293.

D. Leviatan and V. N. Temlyakov (2006), 'Simultaneous approximation by greedy algorithms', *Adv. Comput. Math.* **25**, 73–90.

J. Lindenstrauss and L. Tzafriri (1977), *Classical Banach Spaces*, Vol. I, Springer, Berlin.

E. D. Livshitz (2003), 'Convergence of greedy algorithms in Banach spaces', *Math. Notes* **73**, 342–368.

E. D. Livshitz (2006), 'On the recursive greedy algorithm', *Izv. RAN. Ser. Mat.* **70**, 95–116.

E. D. Livshitz (2007*a*), 'Optimality of the greedy algorithm for some function classes', *Mat. Sb.* **198**, 95–114.

E. D. Livshitz (2007*b*), On lower estimates of rate of convergence of greedy algorithms. Manuscript.

E. D. Livshitz and V. N. Temlyakov (2001), 'On the convergence of weak greedy algorithms', *Tr. Mat. Inst. Steklova* **232**, 236–247.

E. D. Livshitz and V. N. Temlyakov (2003), 'Two lower estimates in greedy approximation', *Constr. Approx.* **19**, 509–523.

A. Lutoborski and V. N. Temlyakov (2003), 'Vector greedy algorithms', *J. Complexity* **19**, 458–473.

V. E. Maiorov, K. I. Oskolkov and V. N. Temlyakov (2002), Gridge approximations and Radon compass. In *Approximation Theory: A Volume Dedicated to Blagovest Sendov*, DARBA, Sofia, pp. 284–309.

S. Mallat and Z. Zhang (1993), 'Matching pursuit in a time-frequency dictionary', *IEEE Trans. Signal Proc.* **41**, 3397–3415.

D. Needell and R. Vershynin (2007), Uniform uncertainty principle and signal recovery via regularized orthogonal matching pursuit. Manuscript.

H. Niederreiter, R. F. Tichy and G. Turnwald (1990), 'An inequality for differences of distribution functions', *Arch. Math.* **54**, 166–172.

M. Nielsen (2006), Trigonometric quasi-greedy bases for $L_p(\mathbb{T}; w)$. Manuscript.

S. N. Nikol'skii(1975), *Approximation of Functions of Several Variables and Embedding Theorems*, Springer.

P. Oswald (2001), 'Greedy algorithms and best m-term approximation with respect to biorthogonal systems', *J. Fourier Anal. Appl.* **7**, 325–341.

P. Petrushev (1988), Direct and converse theorems for spline and rational approximation and Besov spaces. In *Function Spaces and Applications*, Vol. 1302 of *Lecture Notes in Mathematics*, pp. 363–377.

T. Poggio and S. Smale (2003), 'The mathematics of learning: Dealing with data', *Notices Amer. Math. Soc.*, **50**, 537–544.

E. Schmidt (1906), 'Zur Theorie der linearen und nichtlinearen Integralgleichungen I', *Math. Annalen* **63**, 433–476.

A. V. Sil'nichenko (2004), 'Rate of convergence of greedy algorithms', *Mat. Zametki* **76**, 628–632.

F. Smithies (1937), 'The eigen-values and singular values of integral equations', *Proc. London Math. Soc.* (2) **43**, 255–279.

T. Stromberg and R. Heath Jr. (2003), 'Grassmannian frames with applications to coding and communications', *Appl. Comput. Harm. Anal.* **14**, 257–275.

V. N. Temlyakov (1988), 'Approximation by elements of a finite dimensional subspace of functions from various Sobolev or Nikol'skii spaces', *Matem. Zametki* **43**, 770–786. English translation in *Math. Notes* **43**, 444–454.

V. N. Temlyakov (1989*a*), 'Approximation of functions with bounded mixed derivative', *Proc. Steklov Inst.* **1**, 1–122.

V. N. Temlyakov (1989*b*) 'Estimates of the best bilinear approximations of functions of two variables and some of their applications', *Math. USSR Sb.* **62**, 95–109.

V. N. Temlyakov (1990), 'Bilinear approximation and applications', *Proc. Steklov Inst. Math.* **3**, 221–248.

V. N. Temlyakov (1992*a*), 'On estimates of approximation numbers and best bilinear approximation', *Constr. Approx.* **8**, 23–33.

V. N. Temlyakov (1992*b*), 'Estimates of best bilinear approximations of functions and approximation numbers of integral operators', *Math. Notes* **51**, 510–517.

V. N. Temlyakov (1993*b*), 'Bilinear approximation and related questions', *Proc. Steklov Inst. Math.* **4**, 245–265.

V. N. Temlyakov (1998*a*), 'The best m-term approximation and greedy algorithms', *Adv. Comput. Math.* **8**, 249–265.

V. N. Temlyakov (1998*b*), 'Nonlinear m-term approximation with regard to the multivariate Haar system', *East J. Approx.* **4**, 87–106.

V. N. Temlyakov (1998*c*), 'Greedy algorithm and m-term trigonometric approximation', *Constr. Approx.* **14**, 569–587.

V. N. Temlyakov (1999), 'Greedy algorithms and m-term approximation with regard to redundant dictionaries', *J. Approx. Theory* **98**, 117–145.

V. N. Temlyakov (2000*a*), 'Greedy algorithms with regard to multivariate systems with special structure', *Constr. Approx.* **16**, 399–425.

V. N. Temlyakov (2000*b*), 'Weak greedy algorithms', *Adv. Comput. Math.* **12**, 213–227.

V. N. Temlyakov (2001*a*), Lecture notes on approximation theory, Chapter I, University of South Carolina, pp. 1–20.

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

V. N. Temlyakov (2002*a*), 'Universal bases and greedy algorithms for anisotropic function classes', *Constr. Approx.* **18**, 529–550.

V. N. Temlyakov (2002*b*), 'A criterion for convergence of weak greedy algorithms', *Adv. Comput. Math.* **17**, 269–280.

V. N. Temlyakov (2002*c*), Nonlinear approximation with regard to bases. In *Approximation Theory X*, Vanderbilt University Press, Nashville, TN, pp. 373–402.

V. N. Temlyakov (2003*a*), 'Nonlinear methods of approximation', *Found. Comput. Math.* **3**, 33–107.

V. N. Temlyakov (2003*b*), 'Cubature formulas, discrepancy, and nonlinear approximation', *J. Complexity* **19**, 352–391.

V. N. Temlyakov (2004), 'A remark on simultaneous greedy approximation', *East J. Approx.* **10**, 17–25.

V. N. Temlyakov (2005*a*), 'Greedy type algorithms in Banach spaces and applications', *Constr. Approx.* **21**, 257–292.

V. N. Temlyakov (2005*b*), 'Greedy algorithms with restricted depth search', *Proc. Steklov Inst. Math.* **248**, 255–267.

V. N. Temlyakov (2005*c*), Approximation in learning theory. IMI Preprint 05, Department of Mathematics, University of South Carolina.

V. N. Temlyakov (2005*d*), On universal estimators in learning theory. IMI Preprint 17, Department of Mathematics, University of South Carolina.

V. N. Temlyakov (2006*a*), Greedy approximations. In *Foundations of Computational Mathematics: Santander 2005*, Vol. 331 of *London Mathematical Society Lecture Notes*, Cambridge University Press, pp. 371–394.

V. N. Temlyakov (2006*b*), Greedy approximations with regard to bases. In *Proc. International Congress of Mathematics* (Madrid 2006), Vol. 2, pp. 1479–1504.

V. N. Temlyakov (2006*c*), Relaxation in greedy approximation. IMI Preprint 03, Department of Mathematics, University of South Carolina.

V. N. Temlyakov (2006*d*), 'Optimal estimators in learning theory', *Approximation and Probability, Banach Center Publications*, **72**, 341–366.

V. N. Temlyakov (2007*a*), 'Greedy expansions in Banach spaces', *Adv. Comput. Math.* **26**, 431–449.

V. N. Temlyakov (2007*b*), 'Greedy algorithms with prescribed coefficients', *J. Fourier Anal. Appl.* **13**, 71–86.

J. F. Traub, G. W. Wasilkowski and H. Wozniakowski (1988), *Information-Based Complexity*, Academic Press, New York.

J. A. Tropp (2004), 'Greed is good: Algorithmic results for sparse approximation', *IEEE Trans. Inform. Theory* **50**, 2231–2242.

J. A. Tropp and A. C. Gilbert (2005), Signal recovery from partial information via orthogonal matching pursuit. Preprint, University of Michigan.

P. Wojtaszczyk (1997), 'On unconditional polynomial bases in L_p and Bergman spaces', *Constr. Approx.* **13**, 1–15.

P. Wojtaszczyk (2000), 'Greedy algorithms for general systems', *J. Approx. Theory* **107**, 293–314.

P. Wojtaszczyk (2002a), Greedy type bases in Banach spaces. In *Constructive Function Theory*, DARBA, Sofia, pp. 1–20.

P. Wojtaszczyk (2002b), 'Existence of best m-term approximation', *Functiones et Approximatio* **XXX**, 127–133.

P. Wojtaszczyk (2006), 'Greediness of the Haar system in rearrangement invariant spaces', *Banach Center Publications* **72**, 385–395.

H. Weyl (1911), 'Das asymptotische Verteilungsgesetz der Eigenwerte linearer partieller Differentialgleichungen', *Math. Ann.* **71**, 441–479.

A. Zygmund (1959), *Trigonometric Series*, Cambridge University Press.