Found. Comput. Math. (2003) 3:33–107

DOI: 10.1007/s102080010029

FOUNDATIONS OF COMPUTATIONAL MATHEMATICS
The Journal of the Society for the Foundations of Computational Mathematics

Nonlinear Methods of Approximation

V. N. Temlyakov

Department of Mathematics
University of South Carolina
Columbia, SC 29208, USA
temlyak@math.sc.edu

Abstract. Our main interest in this paper is 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. While the scope of this paper is mostly theoretical, we should note that this form of approximation appears in many numerical applications such as adaptive PDE solvers, compression of images and signals, statistical classification, and so on. The standard problem in this regard is the problem of m-term approximation where one fixes a basis and looks to approximate a target function by a linear combination of m terms of the basis. 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. We are interested in the quantitative aspects of this type of approximation. Namely, we want to understand the properties (usually smoothness) of the function which govern its rate of approximation in some given norm (or metric). We are also interested in stable algorithms for finding good or near best approximations using m terms. Some of our earlier work has introduced and analyzed such algorithms. More 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. Some types of approximation that fall into this general category are mathematical frames, adaptive pursuit (or greedy algorithms), and adaptive basis selection. Redundancy on the one hand offers much promise for greater efficiency in terms of approximation rate, but on the other hand gives rise to highly nontrivial theoretical and practical

Date received: May 1, 2001. Final version received: April 16, 2002. Communicated by Wolfgang Dahmen. Online publication: July 12, 2002.
AMS classification: 41A17, 41A25, 41A46, 41A65; 42A10, 42C10, 42C15; 46A35, 46C99, 46E35, 46N40; 65D15, 65J05.

problems. With this motivation, our recent work and the current activity focuses on nonlinear approximation both in the classical form of m-term approximation (where several important problems remain unsolved) and in the form of highly nonlinear approximation where a theory is only now emerging.

1. Introduction

The idea of replacing a complex object (target function) by a simpler one (approximant) is widely spread and successfully used in many areas of science, including computational mathematics. A specific feature of contemporary theoretical and practical problems is huge and unstructured data sets, which cannot be handled without replacing them by simpler objects. One more new feature, which is important for our motivation, is that now we cannot limit ourselves to the use of well-organized and well-structured approximation tools. For example, in signal processing the most popular means of approximation are wavelets and the system of Gabor functions $\{g_{a,b}(x-c), g_{a,b}(x) := e^{iax}e^{-bx^2}, a, c \in \mathbb{R}, b \in \mathbb{R}_+\}$. The Gabor system gives more flexibility in constructing an approximant but is not an orthogonal system. Moreover, it is a redundant (not minimal) system. Thus, in order to address the contemporary needs of computational mathematics we should consider a very general model. As such a model we choose a Banach space X with elements as target functions and an arbitrary system $\mathcal{D}$ of elements of this space as an approximating system. Clearly, in such generality this setting may cover more or less everything. However, in order to obtain meaningful results on approximation we need to impose some restrictions on the Banach space and on the approximating system. The next question is how should an approximant look and how should we measure its complexity? The present dominating approach is to form an approximant as a linear combination of m terms from a given system $\mathcal{D}$. Such an approximant is called an m-term approximant with regard (with respect) to $\mathcal{D}$. We assign a theoretical (idealized) complexity m to an m-term approximant. It is clear that in some cases the number of terms of an approximant reflects accurately the computational complexity of an approximant. For instance, this happens when all elements of $\mathcal{D}$ have approximately the same computational complexity. Another argument in favor of using m as a measure of complexity is that there is no other characteristic of an approximant at the level of generality we are working now. Introducing the concept of best m-term approximation

$$\sigma_m(f,\mathcal{D}) := \inf_{g_j\in\mathcal{D}, c_j;\, j=1,\dots,m} \left\| f - \sum_{j=1}^{m} c_j g_j \right\|, \tag{1.1}$$

we obtain the lower bound for the accuracy of any method providing m-term approximation. The fundamental problem is how to construct a good m-term approximant. It is known (see, for instance, [**13**]) that a problem of simultaneous optimization over many parameters (as in (1.1)) is a very difficult problem. Let us also note that even if we managed to solve (1.1) this solution has the following

disadvantage. The optimal elements, say, $g_1^m(f), \dots, g_m^m(f)$, may depend on m and, therefore, when we go from m to $m+1$ we need to recalculate all $m+1$ elements.

We would like to have an algorithm (see Remark 1.1 below, concerning the term algorithm) of constructing m-term approximants that adds at each step only one new element from $\mathcal{D}$ and keeps elements of $\mathcal{D}$ obtained at the previous steps. Clearly, we are looking for good algorithms which at a minimum 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). It turns out that there is one fundamental principal that allows us to build good algorithms both for arbitrary redundant systems and for very simple well-structured bases like the Haar basis. This principal is the use of a greedy step in searching for a new element to be added to a given m-term approximant. The common feature of all algorithms of m-term approximation discussed in this survey is the presence of a greedy step. By a greedy step, in choosing an mth element $g_m(f) \in \mathcal{D}$ to be used in an m-term approximant, we mean one which maximizes 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 (choosing coefficients of the linear combination) the m-term approximant from already found m elements of the dictionary.

We begin this survey in Section 2 by discussing the classical setting of linear approximation. This section will serve as a comparison when we discuss nonlinear methods. In Sections 3 and 4 we begin our discussion of nonlinear approximation and greedy algorithms. We initiate the discussion in the most general settings to illustrate how far the theory can be pushed when working in complete generality. In Section 3 we consider the case of Hilbert spaces and in Section 4 the case of Banach spaces. In particular, we discuss there some general convergence results for greedy-type algorithms. These results can be used in two ways. First, suppose we have a given system $\mathcal{D}$ and want to build m-term approximants with regard to it. Then if the system $\mathcal{D}$ is complicated enough to be treated as a general system we can use one of the greedy-type algorithms from Sections 3 and 4. The general theory guarantees convergence of those algorithms. Clearly, each system can be treated as a general system, but it does not make sense to do this when a system is simple and more specific and possibly more accurate methods can be used. Second, suppose we are studying convergence of some algorithm and we can identify this algorithm as a greedy-type algorithm with regard to some system $\mathcal{D}$. Then the general theory provides the convergence.

The way of presenting results in this survey is from the most general setting (Sections 3 and 4) to the less general setting of specific redundant systems (Sections 5 and 6) to the case of well-studied systems like bases and even further to concrete bases (wavelet bases, the trigonometric system) (Section 7). It is clear that for a narrower set of systems we can prove stronger results. The reader will see that tendency in this survey.

We believe this survey could be useful for people interested in computations. We know that some of the theoretical algorithms (Thresholding Greedy Algorithm (TGA), Pure Greedy Algorithm (PGA), Weak Greedy Algorithm (WGA)), discussed in this survey, have been implemented successfully in numerical problems of signal/image processing and statistics. We hope that this survey will promote further practical implementation of new ideas and methods developed in nonlinear approximation. With possible numerical applications in mind, we will clearly distinguish between constructive and nonconstructive methods of approximation. We will also address the questions of stability and simplicity of approximation methods.

This survey can be considered as a complement to the survey on nonlinear approximation written by R. A. DeVore [**14**]. We refer the reader interested in a detailed discussion of numerical applications of nonlinear approximation to the survey [**14**]. Our survey has a double purpose. On one hand, we try to present new ideas and concepts generated recently in nonlinear approximation in a way understandable to a wide audience of mathematicians. Section 2 is included for this purpose. Other sections also contain historical remarks that help to understand the motivation and the general spirit of results. Open problems, listed at the end of each section (2–11), may also serve this purpose. On the other hand, in addition to conceptual results, we have also included some recent results of a more technical nature which are addressed mostly to mathematicians working in approximation theory. In some cases these results are the first steps in solving important and difficult problems. We hope that this material combined with the list of open problems will stimulate further intensive development of nonlinear approximation.

Let us now proceed to a more systematic introduction of concepts of nonlinear approximation. We begin with the case where approximation takes place in a Banach space X equipped with a norm $\|\cdot\| := \|\cdot\|_X$. We formulate our approximation problem in the following general way. We say a set of functions $\mathcal{D}$ from X is a *dictionary* if each $g \in X$ has norm one ($\|g\|_X = 1$) and the closure of Span $\mathcal{D}$ coincides with X. 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

$$s = \sum_{g\in\Lambda} c_g g, \qquad \Lambda \subset \mathcal{D}, \quad \#\Lambda \le m,$$

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 $f \in X$ we define its approximation error

$$\sigma_m(f,\mathcal{D})_X := \inf_{s\in\Sigma_m(\mathcal{D})} \|f - s\|_X,$$

and for a function class F,

$$\sigma_m(F,\mathcal{D})_X := \sup_{f\in F} \sigma_m(f,\mathcal{D})_X.$$

The classical example of this type of approximation is the case $X = L_p([0, 2\pi])$ and $\mathcal{D} = \mathcal{B}$ is an orthogonal basis for X. In particular, $\mathcal{B}$ can be taken as the trigonometric system $\mathcal{T} := \{e^{ikx}, k \in \mathbb{Z}\}$ or the Haar system properly normalized. The first results on error estimates in m-term approximation showed an advantage of m-term approximation over approximation by polynomials of order m. R. S. Ismagilov [**35**] studied m-term trigonometric approximation of individual functions, namely, the Bernoulli kernels

$$F_r(x) = 2\sum_{k=1}^{\infty} k^{-r}\cos(kx - r\pi/2).$$

He proved that

$$\sigma_m(F_2, \mathcal{T})_{L_\infty} \le C_\varepsilon m^{-6/5+\varepsilon}$$

with arbitrary $\varepsilon > 0$. It is known that the best approximation $E_m(\cdot)_{L_\infty}$ by trigonometric polynomials of order m in the L_∞-norm has the asymptotic order $E_m(F_2)_{L_\infty} \asymp 1/m$. Further results in m-term trigonometric approximation showed the advantage of this type of nonlinear approximation over linear approximation. For many traditional pairs of function class F and orthogonal system $\mathcal{B}$ the orders of $\sigma_m(F, \mathcal{B})_X$ are now known. Investigation of the case $F = B^r_\theta(L_q)$ (standard Besov class), $\mathcal{B} = \mathcal{T}$, and $X = L_p$ was completed in [**17**]. This investigation required a new technique (see [**17**] and [**42**]) which uses deep results from finite-dimensional geometry. Thus this is an example of interaction between the theory of nonlinear m-term approximation and contemporary functional analysis. We discuss these results in Section 2.

In Sections 3 and 4 we discuss a theory of highly nonlinear m-term approximation. This theory is not complete yet, even in the case of a Hilbert space. We concentrate on an important problem of finding good methods of m-term approximation in the case of general dictionary $\mathcal{D}$ and on studying their efficiency. Let us begin this discussion in the special case of a Hilbert space with the inner product $\langle\cdot,\cdot\rangle$. We define first the Pure Greedy Algorithm (PGA) in Hilbert space H. We describe this algorithm for a general dictionary $\mathcal{D}$. If $f \in H$, we let $g(f) \in \mathcal{D}$ be an element from $\mathcal{D}$ which maximizes $|\langle f, g\rangle|$. We will assume for simplicity that such a maximizer exists; if not suitable modifications are necessary (see Weak Greedy Algorithm (WGA) in Section 3) in the algorithm that follows. We define

$$G(f, \mathcal{D}) := \langle f, g(f)\rangle g(f)$$

and

$$R(f, \mathcal{D}) := f - G(f, \mathcal{D}).$$

Pure Greedy Algorithm (PGA). *We define* $R_0(f, \mathcal{D}) := f$ *and* $G_0(f, \mathcal{D}) := 0$. *Then, for each* $m \ge 1$, *we inductively define*

$$\begin{aligned} G_m(f, \mathcal{D}) &:= G_{m-1}(f, \mathcal{D}) + G(R_{m-1}(f, \mathcal{D}), \mathcal{D}), \\ R_m(f, \mathcal{D}) &:= f - G_m(f, \mathcal{D}) = R(R_{m-1}(f, \mathcal{D}), \mathcal{D}). \end{aligned}$$

In Section 3 we consider the problem of efficiency of PGAs with regard to general dictionaries in Hilbert space. In spite of very general assumptions on the system $\mathcal{D}$, surprisingly, there are nontrivial convergence results (see Theorems 3.1 and 3.4) and also nontrivial estimates of the rate of convergence (see Subsection 3.2) of PGAs.

Remark 1.1. In this survey, we discuss only theoretical aspects of the efficiency of m-term approximation and possible ways to realize this efficiency. The above-defined "greedy algorithm" gives a procedure to construct an approximant which turns out to be a good approximant. The procedure of constructing a greedy approximant is not a numerical algorithm ready for computational implementation. Therefore, it would be more precise to call this procedure a "theoretical greedy algorithm" or "stepwise optimizing process." Keeping this remark in mind we, however, use the term "greedy algorithm" in this paper because it has been used in previous papers and has become a standard name for procedures like the above and for more general procedures of this type (see, for instance, [**14**], [**18**]). Following [**24**] we call an algorithm "incremental" if at step m we add at most one more element $\varphi_m \in \mathcal{D}$ and approximate by linear combination $c_1\varphi_1 + \cdots + c_m\varphi_m$. We use the term "greedy type" for an incremental algorithm with φ_m chosen to maximize a given functional $F(f_{m-1}, g)$ over $g \in \mathcal{D}$ with f_{m-1} the residual after the $(m-1)$th step of the algorithm. The form of $F(\cdot,\cdot)$ determines the kind of greedy algorithm. We use the term "weak greedy" for an incremental algorithm with φ_m satisfying a weaker condition than maximizing the given functional. For instance,

$$F(f_{m-1}^{\tau}, \varphi_m) \ge t_m \sup_{g\in\mathcal{D}} F(f_{m-1}^{\tau}, g), \qquad 0 \le t_m \le 1.$$

The sequence $\tau := \{t_k\}_{k=1}^{\infty}$ is called the "weakness" sequence.

We note that the PGA is known under other names in different areas of research. In statistics it was introduced in [**31**] for a special dictionary (ridge functions dictionary, see Section 6) under the name "projection pursuit regression." In signal processing PGA is known under the name "matching pursuit" (see [**57**], [**13**], [**8**]).

It is clear that for an orthonormal basis $\mathcal{B}$ of a Hilbert space H we have, for each f,

$$\|f - G_m(f, \mathcal{B})\| = \sigma_m(f, \mathcal{B}). \tag{1.2}$$

There is a nontrivial classical example of a redundant dictionary, having the same property: PGAs realize the best m-term approximation for each individual function. We describe that dictionary now. Let Π be a set of functions from $L_2([0,1]^2)$ of the form $u(x_1)v(x_2)$ with the unit L_2-norm. Then for this dictionary and $H = L_2([0,1]^2)$ we have, for each $f \in H$,

$$\|f - G_m(f, \Pi)\| = \sigma_m(f, \Pi). \tag{1.3}$$

This result and related results will be discussed in Section 5. In Section 6 we discuss one more classical redundant dictionary, namely, the dictionary $\mathcal{R}$ consisting of all normalized ridge functions, i.e., functions $w(x)$, $x \in \mathbb{R}^d$, that can be represented in the form $w(x) = g((x, e))$, where g is a univariate function and its argument (x, e) is the scalar product of x and a unit vector $e \in \mathbb{R}^d$. This dictionary and the PGA with regard to it play an important role in statistics (see [**31**], [**34**], [**36**]).

Much less is known about greedy algorithms with regard to general redundant dictionaries in the case of a general Banach space X. We discuss next two versions of generalization of PGAs from a Hilbert space H to a Banach space X (see Section 4 for details). The first one is a straightforward generalization of PGAs. We call it a PGA or X-Greedy Algorithm when we want to indicate a Banach space. For a given X and $\mathcal{D}$ we define $G(f, \mathcal{D}, X) := \alpha(f)g(f)$ where $\alpha(f) \in \mathbb{R}$ and $g(f) \in \mathcal{D}$ satisfy (we assume existence) the relation

$$\min_{\alpha\in\mathbb{R}, g\in\mathcal{D}} \|f - \alpha g\| = \|f - \alpha(f)g(f)\|.$$

X-Greedy Algorithm. *We define* $R_0(f, \mathcal{D}, X) := f$ *and* $G_0(f, \mathcal{D}, X) := 0$. *Then, for each* $m \geq 1$, *we inductively define*

$$\begin{aligned} R_m(f) := R_m(f, \mathcal{D}, X) &:= R_{m-1}(f) - G(R_{m-1}(f), \mathcal{D}, X), \\ G_m(f, \mathcal{D}, X) &:= G_{m-1}(f, \mathcal{D}, X) + G(R_{m-1}(f), \mathcal{D}, X). \end{aligned}$$

The second version of a PGA in a Banach space is based on the concept of a peak functional (norming functional). We call it the Dual Greedy Algorithm (DGA). Let a dictionary $\mathcal{D}$ in X be given. Take an element $f \in X$ and find a peak functional F_f, i.e., a functional such that $\|F_f\|_{X'} = 1$ and $F_f(f) = \|f\|_X$. The existence of such a functional follows from the Hahn–Banach theorem. Now the basic step of a PGA is modified to the following. Assume that there exists $g_f \in \mathcal{D}$ such that

$$|F_f(g_f)| = \max_{g\in\mathcal{D}} |F_f(g)|.$$

We take this g_f and solve one more optimization problem: find a number a such that

$$\|f - ag_f\|_X = \min_b \|f - bg_f\|_X.$$

We put

$$G^D(f, \mathcal{D}) := ag_f, \qquad R^D(f, \mathcal{D}) := f - ag_f.$$

Repeating this step m times we get $G^D_m(f, \mathcal{D})$ as an approximant and $R^D_m(f, \mathcal{D})$ as a residual. Some results on greedy algorithms in Banach spaces are presented in Section 4.

We discussed above best m-term approximation with regard to a dictionary $\mathcal{D}$ in a Banach space X. The sequence $\{\sigma_m(f, \mathcal{D})_X\}$ gives the lower estimates of

accuracy for any sequence of operators A_m that map X into $\Sigma_m(\mathcal{D})$ where, as above, $\Sigma_m(\mathcal{D})$ is the set of all functions in X which can be expressed as a linear combination of at most m elements from $\mathcal{D}$. Thus, the sequences $\{\sigma_m(f,\mathcal{D})_X\}$ and $\{\sigma_m(F,\mathcal{D})_X\}$ may serve as the target accuracies in constructing approximating operators A_m. It is clear that the best operator (if it exists) gives the error

$$\|f - A_m(f,\mathcal{D})\|_X = \sigma_m(f,\mathcal{D})_X, \tag{1.4}$$

(see, e.g., (1.2) and (1.3)). We identify a sequence of operators $\{A_m\}_{m=1}^{\infty}$ as an algorithm A. We call an algorithm $A := \{A_m\}_{m=1}^{\infty}$ near best or near best for individual functions if

$$\|f - A_m(f,\mathcal{D})\|_X \le C(\mathcal{D},X)\sigma_m(f,\mathcal{D})_X \tag{1.5}$$

for all $f \in X$ and all $m = 1, 2, \dots$. Similarly, we say that A is near best for a function class F if we have, for any $f \in F$,

$$\|f - A_m(f,\mathcal{D})\|_X \le C(F,\mathcal{D},X)\sigma_m(F,\mathcal{D})_X, \qquad m = 1, 2, \dots. \tag{1.6}$$

It is clear that an algorithm A_m satisfying (1.5) is excellent from the point of view of accuracy: it provides near best approximation for every individual function and, therefore, for any function class. The property (1.6) is weaker than (1.5) but is still very good. The corresponding results for nonlinear approximation with regard to a basis are discussed in Section 7.

Let a Banach space X with a normalized basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$, $\|\psi_k\| = 1$, $k = 1, 2, \dots$, be given. We consider the following theoretical greedy algorithm that we call the Thresholding Greedy Algorithm (TGA). For a given element $f \in X$ we consider the expansion

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

Let an element $f \in X$ be given. We call a permutation ρ, $\rho(j) = k_j$, $j = 1, 2, \dots$, of the positive integers decreasing and write $\rho \in D(f)$ if

$$|c_{k_1}(f)| \ge |c_{k_2}(f)| \ge \cdots.$$

In the case of strict inequalities, here $D(f)$ consists of only one permutation. We define the mth greedy approximant of f with regard to the basis Ψ corresponding to a permutation $\rho \in D(f)$ by the formula

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

This is a simple algorithm which describes a theoretical scheme (it is not computationally ready) for m-term approximation of an element f. We call a basis Ψ a

greedy basis if the TGA with regard to Ψ is a near best algorithm, i.e., it satisfies (1.5). In Section 7 we formulate a result (see Theorem 7.1) which says that a wavelet-type basis is a greedy basis for all L_p, $1 < p < \infty$. The following question arises naturally. Why should we use wavelet-type bases? In order to answer this question we consider an optimization problem in the spirit of the Kolmogorov width.

Let $\mathbb{D}$ be a collection of dictionaries. The classical example of $\mathbb{D}$ is $\mathbb{O}$ = {orthonormal bases on a given domain}. The optimization problem asks us to find (if possible), for a given pair of a collection of dictionaries $\mathbb{D}$ and a function class F, a dictionary $\mathcal{D}^* \in \mathbb{D}$ such that

$$\sigma_m(F, \mathcal{D}^*)_X \asymp \sigma_m(F, \mathbb{D})_X := \inf_{\mathcal{D} \in \mathbb{D}} \sigma_m(F, \mathcal{D})_X.$$

This problem is interesting and important for theoretical investigation and also for practical applications where we often want to have a dictionary $\mathcal{D}$ with a certain structure (from a collection $\mathbb{D}$) and do not want to stick to a particular one. In Section 10 we discuss only theoretical aspects of this problem for the classical example of $\mathbb{D} = \mathbb{O}$. We go even further (from optimal for a one class method to optimal for a collection of classes method) and ask the following question. What is a good basis for the multivariate approximation? We propose to use the concept of universality to answer the above question. A universal dictionary $\mathcal{D} \in \mathbb{D}$ is the one which is optimal for all F from a given collection $\mathcal{F}$ of function classes. We give a formal definition of this important concept.

Definition 1.1. Let two collections, $\mathcal{F}$ of function classes and $\mathbb{D}$ of dictionaries, be given. We say that $\mathcal{D} \in \mathbb{D}$ is universal for the pair $(\mathcal{F}, \mathbb{D})$ if there exists a constant C which may depend on $\mathcal{F}$, $\mathbb{D}$, and X such that for any $F \in \mathcal{F}$ we have

$$\sigma_m(F, \mathcal{D})_X \le C\sigma_m(F, \mathbb{D})_X, \qquad m = 1, 2, \dots.$$

This is a new concept in nonlinear approximation. The following observation motivates our interest in this setting. In practice we often do not know the exact smoothness class F where our input function (signal, image) comes from. Instead, we often know that our function comes from a class with a certain structure, for instance, an anisotropic Sobolev class. This is exactly the situation we are dealing with in the universal dictionary setting. So, if for a collection $\mathcal{F}$ there exists a universal dictionary $\mathcal{D} \in \mathbb{D}$, it is an ideal situation. We can use this universal dictionary $\mathcal{D}$ in all cases and we know that it adjusts automatically to the best smoothness class $F \in \mathcal{F}$ which contains the target function. Next, if a pair $(\mathcal{F}, \mathbb{D})$ does not allow a universal dictionary we have a trade-off between universality and accuracy. We discuss the universality results in Section 11.

Combining the ideas of near best approximation, the optimization problem for a function class and a universality concept for a collection of function classes, we describe a general way of finding a good basis. We suggest a three-step strategy to find a good basis (dictionary) for nonlinear m-term approximation. The first step

consists of solving an optimization problem for a given function class F, when we optimize over a collection $\mathbb{D}$ of bases (dictionaries). The second step is devoted to finding a universal basis (dictionary) $\mathcal{D}_u \in \mathbb{D}$ for a given pair $(\mathcal{F}, \mathbb{D})$ of collections: $\mathcal{F}$ of function classes and $\mathbb{D}$ of bases (dictionaries). The third step deals with constructing a theoretical algorithm that realizes near best m-term approximation with regard to $\mathcal{D}_u$ for function classes from $\mathcal{F}$. We worked this strategy out in the case of anisotropic function classes and the set of orthogonal bases (see [**93**]). The results are positive. We constructed a natural tensor-product-wavelet-type basis and proved that it is universal. Moreover, we proved that the TGA realizes near best m-term approximation with regard to this basis for all anisotropic function classes. We discuss these results in Section 11.

In Section 9 we discuss some results on how the entropy numbers $\varepsilon_n(F, X)$ can be used in estimating from below the quantities $\{\sigma_m(F, \Psi)\}$. The idea of estimating the Kolmogorov widths from below using the entropy numbers is well-known (see [**53**], [**7**], [**65**]). We used this idea in [**90**] for estimating nonlinear best m-term approximation. We proved that for good systems Ψ the estimate

$$\varepsilon_n(F, X) \gg n^{-a}(\log n)^b, \qquad a > 0, \quad b \in \mathbb{R},$$

for the entropy numbers implies the same estimate for best m-term approximation:

$$\sigma_m(F, \Psi)_X \gg m^{-a}(\log m)^b.$$

See Section 9 for more details.

Let us agree to denote by C various positive absolute constants and by C, with arguments or indexes ($C(q, p)$, C_r, and so on), positive numbers which depend on the arguments indicated. For two nonnegative sequences $a = \{a_n\}_{n=1}^{\infty}$ and $b = \{b_n\}_{n=1}^{\infty}$ the relation (order inequality) $a_n \ll b_n$ means that there is a number $C(a, b)$ such that for all n we have $a_n \le C(a, b)b_n$; and the relation $a_n \asymp b_n$ means that $a_n \ll b_n$ and $b_n \ll a_n$. The sign $\ll$ will be used for the sake of brevity in estimates of various characteristics of functions which belong to the class involved. In these cases we assume that constants in inequalities may depend on the class but not on the function considered.

Notations are introduced in the text.

2. Approximation by Linear Methods. Some Remarks

2.1. *Historical Remarks*

In order to give the reader some ideas for comparing the quality of approximation methods we now discuss some classical results in approximation of periodic functions. In this section we briefly discuss various classical approaches, created in linear approximation, for the estimation of the quality of a method of approximation. We will use and refine these approaches in nonlinear approximation. We

confine our discussion to the case of the approximation of periodic functions of a single variable. The two main parameters of a method of approximation are accuracy and complexity. These concepts may be treated in various ways depending on the particular problems involved. Here we will start from the classical idea about the approximation of functions by polynomials. After Fourier's article (1807) the representation of a 2π-periodic function by its Fourier series became natural. In other words, the function $f(x)$ is approximately represented by a partial sum $S_n(f, x)$ of its Fourier series.

We will be interested in the approximation of a function f by a polynomial $S_n(f)$ in some L_p-norm, $1 \le p \le \infty$. In the case $p = \infty$ we will assume that we are dealing with the uniform norm. As accuracy of the method of approximating a periodic function by its Fourier partial sum we will consider the quantity $\|f - S(f)\|_p$. The complexity of this method of approximation contains the two following characteristics. The order of the trigonometric polynomial $S_n(f)$ is the quantitative characteristic. The following observation gives us the qualitative characteristic. The coefficients of this polynomial are found by the Fourier formulas which means that the operator S_n is the orthogonal projection onto the subspace of trigonometric polynomials of order n.

In 1854 Chebyshev suggested representing a continuous function f by its polynomial of best approximation, namely by the polynomial $t_n(f)$ such that

$$\|f - t_n(f)\|_\infty = E_n(f)_\infty \stackrel{\text{def}}{=} \inf_{\alpha_k, \beta_k} \left\| f(x) - \sum_{k=0}^{n} (\alpha_k \cos kx + \beta_k \sin kx) \right\|_\infty .$$

He proved the existence and uniqueness of such a polynomial. We will consider this method of approximation not only in the uniform norm, but in all L_p-norms, $1 \le p < \infty$. The accuracy of the Chebyshev method can be easily compared with the accuracy of the Fourier method:

$$E_n(f)_p \le \|f - S_n(f)\|_p.$$

However, it is difficult to compare the complexities of these two methods. The quantitative characteristics coincide but the qualitative characteristics are different (e.g., it is not difficult to understand that for $p = \infty$ the mapping $f \to t_n(f)$ is not a linear operator). The Du Bois–Reymond example (1873) of a continuous function f such that $\|f - S_n(f)\|_\infty \to \infty$ when $n \to \infty$, and the Weierstrass theorem which says that for each continuous function f we have $E_n(f)_\infty \to 0$ as $n \to \infty$, showed the advantage of the Chebyshev method in comparison with the Fourier method from the point of view of accuracy. It is known that for each $f \in L_2(\mathbb{T})$ the approximation with the error $E_n(f)_2$ can be realized by the operator S_n of orthogonal projection onto the space of trigonometric polynomials of order n. The performance of operator S_n was studied thoroughly in all L_p spaces, $1 \le p \le \infty$. It was proved that S_n provides almost optimal or close to optimal approximation

for each $f \in L_p(\mathbb{T})$,

$$\|f - S_n(f)\|_p \le C(p)E_n(f)_p, \qquad 1 < p < \infty,$$
$$\|f - S_n(f)\|_p \le C \ln(n+2)E_n(f)_p, \qquad p = 1, \infty.$$

The desire to construct methods of approximation which have the advantages of the Fourier and Chebyshev methods led to the study of various methods of summation of the Fourier series. The most important among them from the point of view of approximation are the de la Vallée-Poussin, Fejér, and Jackson methods which were constructed early in the twentieth century. All these methods are linear. For example, the de la Vallée-Poussin method is the method of approximation of a function f by the polynomial

$$V_n(f) = \frac{1}{n}\sum_{l=n}^{2n-1} S_l(f)$$

of order $2n-1$.

From the point of view of accuracy this method is close to the Chebyshev method; de la Vallée-Poussin proved that

$$\|f - V_n(f)\|_p \le 4E_n(f)_p, \qquad 1 \le p \le \infty.$$

From the point of view of complexity it is close to the Fourier method, and the property of linearity essentially distinguishes it from the Chebyshev method.

We see that common to all these methods is the approximation by means of trigonometric polynomials; however, the ways of constructing these polynomials differ: orthogonal projections on the subspace of trigonometric polynomials of fixed order, the operator of best approximation, and linear operators.

The approximation of functions by algebraic polynomials was studied in parallel with that for trigonometric polynomials. We will now point out some results which determined the style of investigation of a number of problems in approximation theory. These problems are of interest even today.

De la Vallée-Poussin proved in 1908 that, for best approximations of the function $|x|$ in the uniform norm on $[-1, 1]$ by algebraic polynomials of degree n, the following upper estimate holds

$$e_n(|x|) \le C/n.$$

He raised the question of the possibility of an improvement of this estimate in the sense of order. Bernstein (1912) proved that this order estimate is sharp. Moreover, he then established the asymptotic behavior of the sequence $\{e_n(|x|)\}$:

$$e_n(|x|) = \mu/n + o(1/n), \qquad \mu = 0.282 \mp 0.004.$$

These results initiated a series of investigations into best approximations of individual functions which have special singularities.

At this stage of investigation the natural conjecture arose that the smoother a function, the more rapidly its sequence of best approximations decreases.

In 1911 Jackson proved the inequality

$$E_n(f)_\infty \le Cn^{-r}\omega(f^{(r)}, 1/n)_\infty,$$

where $\omega(g, t)_p$ is the modulus of continuity of g in the L_p-norm.

The relations which give upper estimates for the best approximations of a function in terms of its smoothness are now called Jackson inequalities, and in a wider sense such relations are called direct theorems of approximation theory.

As a result of Bernstein's (1912) and de la Vallée-Poussin's (1908, 1919) investigations we can formulate the following assertion which is now called the inverse theorem of approximation theory. If

$$E_n(f)_\infty \le Cn^{-r-\alpha}, \qquad 0 \le r \text{ integer}, \quad 0 < \alpha < 1,$$

then f has a continuous derivative of order r which belongs to the class Lip α, that is, $f \in W^r H^\alpha$. Thus, the results of Jackson, Bernstein, and de la Vallee-Poussin show that functions from the class $W^r H^\alpha$, $0 < \alpha < 1$, can be characterized by the order of decrease of its sequences of best approximations.

We remark that, at that time, classes similar to $W^r H^\alpha$ were used in other areas of mathematics for obtaining orders of decrease of various quantities. As an example we formulate a result of Fredholm. Let $f(x, y)$ be continuous on $[a, b] \times [a, b]$ and let

$$\max_{x,y} |f(x, y + t) - f(x, y)| \le C|t|^\alpha, \qquad 0 < \alpha \le 1.$$

Then for eigenvalues $\lambda(J_f)$ of the integral operator

$$(J_f\psi)(x) = \int_a^b f(x, y)\psi(y)\, dy$$

the following relation is valid for any $\rho > 2/(2\alpha + 1)$:

$$\sum_{n=1}^{\infty} |\lambda_n(J_f)|^\rho < \infty.$$

The investigation of upper bounds of errors of approximation of functions from a fixed class by some method of approximation began with an article by Lebesgue (1910). In particular, Lebesgue proved that

$$\sup_{f \in \text{Lip}\alpha} \|f - S_n(f)\|_\infty \asymp n^{-\alpha} \ln n.$$

In 1936 Kolmogorov introduced the concept of width $d_n(F, X)$ of a class F in a Banach space X (see Subsection 2.3 for details). This concept is designed to find, for a fixed n and for a class F, a subspace of dimension n, optimal with

respect to the construction of an approximating element as the element of best approximation. In other words, the Kolmogorov width gives the lower bound for the accuracy of Chebyshev methods, having the same quantitative characteristic of complexity (the dimension of the approximating subspace). In analogy to the concept of the Kolmogorov width, that is, to the problem concerning the best Chebyshev method, the problems concerning the best linear method and the best Fourier method were considered: Tikhomirov (1960) introduced the concept of linear width and Temlyakov (1982) introduced the concept of orthowidth (Fourier width). We discuss these widths in more detail later in this section. Here we remark that from the point of view of orthowidth the Fourier operator S_n is optimal (in the sense of order of approximation in the L_p-norm) for all Sobolev classes W_q^r with $1 \le q,\ p \le \infty$, with the exception of the two cases $q = p = 1$ and $q = p = \infty$.

Keeping in mind the primary question about the selection of a good method of approximation, we now draw some conclusions from this brief historical survey.

(1) The trigonometric polynomials were considered as a natural means of approximating periodic functions during the whole period of the development of approximation theory.

(2) In approximation theory (as well as in other fields of mathematics) it turned out that it is natural to unite functions with the same smoothness into a class.

(3) The subspaces of trigonometric polynomials have in many cases been obtained as a solution of (optimization) problems on the most precise Chebyshev method (Kolmogorov width), the linear method (linear width), and the Fourier method (orthowidth) for the classes of smooth functions.

In the above-mentioned results one can see two different approaches in approximation theory.

A. A method of approximation is fixed, for instance, the Fourier sums $S_n(\cdot)$, and we study this method. We compare it with the best for individual functions or for function classes (for more details, see Subsection 2.2). We look for a natural class of functions for this method of approximation.

B. A function class is given, for instance, $W^r H^\alpha$, and we study the approximation of functions in this class. We approximate functions from this class by trigonometric polynomials and consider the optimization problem of widths of this class. This results in a natural method of approximation of the given class.

Based on these remarks we may formulate the following general strategy for investigating approximation problems. We note that this strategy turns out to be most fruitful in those cases of linear approximation where we do not know a priori a natural method of approximation. First, we determine what kind of methods we are looking for (for instance, orthogonal projections on subspaces of fixed dimension). Second, we formulate the corresponding optimization problem for a function class which we are going to approximate (the orthowidth problem). Then we solve the problem for this class in the most simple case of approximation in a Hilbert space (L_2). After that we study a method obtained and apply it to approximation in other Banach spaces (L_p). This strategy has also been used in nonlinear approximation. Let us note that in addition to the above general strategy we will also need some

specific features of the problem under consideration. For instance, it is known that the multivariate approximation problems have some specific difficulties. One of them is that there is no natural ordering of the multivariate bases that are obtained as a tensor product of univariate bases, like the trigonometric system. Another feature is that multivariate function classes may be anisotropic, i.e., smoothness is characterized by a vector but not by a scalar as it was in the univariate case. Anisotropy of function classes raises a question of finding methods independent of anisotropy that are good for all anisotropic classes (see, for details, Section 11).

2.2. *Approximation of Individual Functions*

Let us consider a Banach space X with a basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$, $\|\psi_k\| = 1$, $k = 1, 2, \dots$. For a given element $f \in X$ we consider the expansion

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

and the corresponding partial sums

$$S_n(f, \Psi) := \sum_{k=1}^{n} c_k(f)\psi_k.$$

In order to understand the efficiency of approximating by S_n we introduce best approximations with regard to $\mathrm{Span}\{\psi_1, \dots, \psi_n\}$:

$$E_n(f, \Psi)_X := \inf_{a_k} \left\| f - \sum_{k=1}^{n} a_k\psi_k \right\|_X.$$

It is well-known (see [**49**]) that for a basis Ψ the operator S_n is bounded as an operator from X to X. Therefore, we have, for any $f, g \in X$,

$$\|S_n(f, \Psi) - S_n(g, \Psi)\|_X \le C(X, \Psi)\|f - g\|_X,$$

and, for any $f \in X$,

$$\|f - S_n(f, \Psi)\|_X \le C(X, \Psi)E_n(f, \Psi)_X.$$

This means that the partial sums method provides near best approximation for any individual f. Let us consider a classical example of $\Psi = \mathcal{T}$—the trigonometric system and $X = L_p$, $1 \le p \le \infty$. The basis $\mathcal{T}$ is an orthonormal basis and, therefore, the orthoprojector S_n realizes the best approximation in L_2. By the Riesz theorem (see [**105**]) we know that $\mathcal{T}$ is a basis for $1 < p < \infty$ and thus the Fourier sums realize near best trigonometric approximation in L_p, $1 < p < \infty$. It is well-known that $\mathcal{T}$ is not a basis for L_1 and L_∞. In this case, we have the Lebesgue inequality,

$$\|f - S_n(f, \mathcal{T})\|_p \le C \ln(n + 2)E_n(f, \mathcal{T})_p, \qquad p = 1, \infty.$$

An extra factor $\ln(2+n)$ is a slowly growing to infinity function on n but nonetheless there are different settings where an attempt to get rid of $\ln(2+n)$ was done. We will mention some of them. One can replace the partial sum $S_n(f,\mathcal{T})$ by the de la Vallée-Poussin operator

$$V_n(f,\mathcal{T}) := \frac{1}{n}\sum_{j=n}^{2n-1} S_j(f,\mathcal{T}).$$

This is not an orthoprojector anymore but one has the estimate

$$\|f - V_n(f,\mathcal{T})\|_p \le 4E_n(f,\mathcal{T})_p, \qquad p = 1,\infty,$$

that is good if $\{E_n(f,\mathcal{T})_p\}$ does not decrease fast (note that $V_n(f,\mathcal{T})$ is a trigonometric polynomial of degree $2n-1$). The following estimate was obtained by Oskolkov [**61**], for $p=\infty$,

$$\|f - S_n(f,\mathcal{T})\|_\infty \le C\sum_{j=n}^{2n}\frac{E_j(f,\mathcal{T})_\infty}{j-n+1}.$$

We also note that in the case of $p=\infty$ an extra $\ln(2+n)$ appears not only in the estimates for individual functions as above but also for function classes. We present here some well-known results for the Sobolev classes

$$W_q^r := \{f : f^{(r-1)}\text{-absolutely continuous}, \|f^{(r)}\|_q \le 1\}.$$

Kolmogorov proved that

$$\sup_{f\in W_\infty^r}\|f - S_n(f,\mathcal{T})\|_\infty = \frac{4}{\pi^2}(\ln n)n^{-r} + O(n^{-r}).$$

Favard, Akhiezer, and Krein (see [**100**]) proved the equality

$$\sup_{f\in W_\infty^r} E_n(f,\mathcal{T})_\infty = K_r(n+1)^{-r},$$

with K_r as a number depending on the number r.

We discuss an interplay between the approximation of individual functions and function classes. In this section we discuss certain aspects of the following question. Suppose that F is a function class and that $\{\delta_n(F)\}_{n=1}^\infty$ is a corresponding sequence of extremal quantities. In this section we take $\delta_n(F) := \sup_{f\in F}\delta_n(f)$ to be the supremum $e_n(F)$ or $E_n(F)$ of the best approximation in the uniform norm of functions in F by algebraic ($e_n(\cdot)$) or trigonometric ($E_n(\cdot)$) polynomials of order n. In Subsection 2.3 we will consider the case $\delta_n(F) = d_n(F)$—the sequence of the Kolmogorov widths of the class F. We discuss the question of the extent to which the sequence $\{\delta_n(F)\}_{n=1}^\infty$, which is connected with the whole function class

F, characterizes the corresponding properties of individual functions in F. In this section we discuss the question of the existence in F of a function f such that

$$\lim_{n\to\infty} \delta_n(f)/\delta_n(F) = 1.$$

The first result in this direction is apparently due to Lebesgue. In [**48**] he proved the equality

$$\sup_{\|f\|_\infty \le M} E_n(f)_\infty = M, \qquad n = 1, 2, \dots,$$

where sup is taken over continuous functions. This equality in combination with the Weierstrass theorem shows that in the class of all continuous functions bounded by the number M there is no asymptotically extremal function.

Let us make a historical remark due to Nikol'skii (see [**59**]). S. N. Bernstein discussed the role of function classes in constructive approximation in the opening session of his seminar on Approximation Theory (Moscow, Spring 1945). His general attitude to the role of studying the sequences of $E_n(F) := \sup_{f\in F} E_n(f)$ for a given function class F was skeptical. One of his arguments was that the sequence $\{E_n(F)\}$ may not reflect the behavior of $\{E_n(f)\}$ for any individual $f \in F$, because usually the extremal function that realizes $\sup_{f\in F} E_n(f)$ depends on n. He formulated a problem of studying

$$\sup_{f\in F} \limsup_{n\to\infty} \frac{E_n(f)}{E_n(F)} \qquad \text{and} \qquad \sup_{f\in F} \liminf_{n\to\infty} \frac{E_n(f)}{E_n(F)}$$

and their analogs for approximation by algebraic polynomials for some function classes. In particular, he thought that the function $|x|$ is an extremal function in the sense of the above quantities in the class $\mathrm{Lip}_1 1$ for approximation by algebraic polynomials in the uniform norm. However, it turned out not to be the case. S. M. Nikol'skii [**59**] proved in 1946 that for W^r_∞ classes there is a function $f \in W^r_\infty$ such that

$$\limsup_{n\to\infty} E_n(f)/E_n(W^r_\infty) = 1.$$

It was proved in [**73**], [**74**] that for the class W^r_∞ there exists a function $f \in W^r_\infty$ such that

$$\lim_{n\to\infty} E_n(f)/E_n(W^r_\infty) = 1.$$

Further results and some generalizations are obtained in [**76**], [**75**]. It is interesting to compare the above result with the following result of Oskolkov [**60**]

$$\max_{f\in W^1_\infty} \liminf_{n\to\infty} \left(\|f - S_n(f, \mathcal{T})\|_\infty \Big/ \sup_{f\in W^1_\infty} \|f - S_n(f, \mathcal{T})\|_\infty \right) = \tfrac{1}{2}.$$

2.3. *Approximation of Function Classes*

We now give the definitions of widths of a class F in a space X. The Kolmogorov width

$$d_n(F, X) = \inf_{\{\varphi_j\}_{j=1}^n} \sup_{f\in F} \inf_{\{c_j\}_{j=1}^n} \left\| f - \sum_{j=1}^n c_j\varphi_j \right\|_X .$$

The first result about widths, namely Kolmogorov's result (1936),

$$d_{2n+1}(W_2^r, L_2) = (n+1)^{-r},$$

showed that the best subspace of dimension $2n+1$ for the approximation of classes of periodic functions is the subspace of trigonometric polynomials of order n. This result confirmed that the approximation of functions in the class W_2^r by trigonometric polynomials is natural. Further estimates of the widths $d_{2n+1}(W_q^r, L_p)$, $1 \le q$, $p \le \infty$, some of which are discussed here, showed that for some values of the parameters q, p the subspace of trigonometric polynomials of order n is optimal (in the sense of order) but for other values of q, p this subspace is not optimal.

The Ismagilov [**35**] estimate for the quantity $d_n(W_1^r, L_\infty)$ gave the first example where the subspace of trigonometric polynomials of order n is not optimal. This phenomenon was thoroughly studied by Kashin [**39**].

The linear width

$$\lambda_n(F, X) = \inf_{A:\operatorname{rank} A\le n} \sup_{f\in F} \|f - Af\|_X .$$

The orthowidth (Fourier width)

$$\varphi_n(F, X) := d_n^\perp(F, X) := \inf_{\text{orthonormal system}\{u_i\}_{i=1}^n} \sup_{f\in F} \left\| f - \sum_{i=1}^n \langle f, u_i\rangle u_i \right\|_X .$$

All these widths have as a starting point a function class F. Thus in this setting we choose a priori a function class F and look for optimal subspaces for approximation of a given class. The following results are well-known [**86**]. We present these results for r positive integers. Similar results hold for any r greater than some $\alpha(q, p) \le 1$, which is defined below in Theorem 2.1. Positive integers satisfy the inequality $r > \alpha(q, p)$ for all $1 \le q, p \le \infty$, except $q = 1$, $p = \infty$, where we have $\alpha(1, \infty) = 1$. Thus in the case $q = 1$, $p = \infty$ we assume $r > 1$.

A. In the case $1 \le p \le q \le \infty$ or $1 \le q \le p \le 2$, one has

$$\varphi_n(W_q^r, L_p) \asymp \lambda_n(W_q^r, L_p) \asymp d_n(W_q^r, L_p) \asymp n^{-r+(1/q-1/p)_+}. \tag{2.1}$$

B. In the case $1 \le q < p \le \infty$, $p > 2$, one has

$$\begin{aligned} d_n(W_q^r, L_p) &\asymp n^{-r+(1/q-1/2)_+}, \\ \lambda_n(W_q^r, L_p) &\asymp n^{-r+\max(1/q-1/2,1/2-1/p)}, \\ \varphi_n(W_q^r, L_p) &\asymp n^{-r+1/q-1/p}. \end{aligned}$$

In Case A the classical trigonometric system provides the optimal orders for all widths, except φ_n for $q = p = 1, \infty$. Let us discuss a more interesting Case B for a particular choice of $q = 2$ and $p = \infty$. We have

$$d_n(W_2^r, L_\infty) \asymp n^{-r}, \tag{2.2}$$

$$\lambda_n(W_2^r, L_\infty) \asymp \varphi_n(W_2^r, L_\infty) \asymp n^{-r+1/2}. \tag{2.3}$$

These relations show that if we drop the linearity requirement for the approximation method we gain in accuracy a factor $n^{-1/2}$. However, there is a big difficulty in realization of the estimate (2.2). We know by Kashin's result that there exists a subspace realizing (2.2) but we do not know a way to construct it. Thus it is only an existence theorem for now.

Let us discuss one more special case $q = 1$ and $p = \infty$. In this case we have

$$d_n(W_1^r, L_\infty) \asymp \lambda_n(W_1^r, L_\infty) \asymp n^{-r+1/2} \tag{2.4}$$

and

$$\varphi_n(W_1^r, L_\infty) \asymp n^{-r+1}. \tag{2.5}$$

Therefore, by (2.4), the best possible approximation (in the sense of order) can be realized by a linear method, say, A_n. However, by (2.5), this linear method A_n is certainly not an orthogonal projector. Moreover, by [**86**], it cannot satisfy even the following much weaker restriction $\|A_n(e^{ikx})\|_2 \le C$, $k \in \mathbb{Z}$. This means that the optimal linear operator A_n is unstable. A small change in some of the Fourier coefficients of f may result in a big change of $\|A_n(f)\|_2$.

Let us make some conclusions now. In the Linear Approximation of W_q^r in L_p the bottom line is given by $\varphi_n(W_q^r, L_p)$ where the approximation method is the simplest—orthogonal projection. Partial sums with regard to classical systems provide an optimal error of approximation for this width. The trigonometric system works for all $1 \le q, p \le \infty$ except $(q, p) = (1, 1), (\infty, \infty)$. The wavelet systems (see [**1**]) work for all $1 \le q, p \le \infty$. In the example of the pair (W_1^r, L_∞) we have seen that we need to sacrifice important and convenient properties of the approximating operator in order to achieve better accuracy. In the example of (W_2^r, L_∞) we have seen that we need to pay even a bigger price for better accuracy in a form of proving only an existence theorem instead of providing a constructive method of approximation.

Let us continue the discussion from Subsection 2.2 on the interplay between the approximation of individual functions and function classes. Let us first try to associate with an individual function f a sequence of the Kolmogorov widths. It is clear that the choice $F[f] := \{f\}$ does not work because $d_1(F[f]) = 0$ for each f. The idea is to find a minimal reasonable class that contains f. In the periodic case it is natural to associate with $f(x)$ all translates $f(x - y)$. Thus define $F[f] := \{f(x - y), y \in \mathbb{T}\}$. All known classes of periodic functions are shift invariant. In such a case we have for $f \in F$ that $F[f] \subset F$ and $d_n(F[f], X) \le d_n(F, X)$. We

will present some results from [**78**]. For $r \in \mathbb{Z}_+$, $\alpha \in \mathbb{R}_+$, denote

$$W^r H_q^\alpha := \{f : f^{(r-1)} - \text{absolutely continuous}\},$$

$$\|f^{(r)}(x) - f^{(r)}(y)\|_q \le |x-y|^\alpha, x, y \in \mathbb{T}\}.$$

Theorem 2.1. *Let $1 \le q \le p \le \infty$ or $2 \le p \le q \le \infty$. Then each class $W^r H_q^\alpha$ with $0 < \alpha < 1$ and $r + \alpha \ge \alpha(q, p)$ contains a function f such that*

$$\liminf_{n\to\infty} d_n(F[f], L_p)/d_n(W^r H_q^\alpha, L_p) > 0.$$

We define here $\alpha(q, p) := (1/q - 1/p)_+$ for $1 \le q \le p \le 2$, $2 \le p \le q \le \infty$ and $\alpha(q, p) := \max(1/q, \frac{1}{2})$ for $1 \le q \le p \le \infty$, $p > 2$.

Let us consider one particular case, $q = p = \infty$, $\alpha = 1$, that is not covered by Theorem 2.1. As established by Tikhomirov [**99**], the values of the Kolmogorov width in this case are given by approximations by trigonometric polynomials. Results of Nikol'skii and this author mentioned in Subsection 2.2 show that each class W_∞^r contains a function asymptotically extremal for the best approximation by trigonometric polynomials. It turns out that the picture is different for the asymptotic behavior of the widths $d_n(F[f], L_\infty)$.

Theorem 2.2. *Any function $f \in W_\infty^r$, $r > \frac{1}{2}$, satisfies*

$$d_n(F[f], L_\infty) = o(d_n(W_\infty^r, L_\infty)).$$

It is interesting to note that for any periodic function $f \in L_p(\mathbb{T})$ we have

$$\sigma_m(f(x-y), \Pi)_{p,\infty} = d_m(F[f], L_p) \le \sigma_m(f, \mathcal{T})_p. \tag{2.6}$$

It is proved in [**78**] that for $1 \le q \le p \le \infty$ one has

$$\begin{aligned} d_m^{\text{ind}}(W^r H_q^\alpha, L_p) &:= \sup_{f \in W^r H_q^\alpha} d_m(F[f], L_p) \\ &\asymp d_m(W^r H_q^\alpha, L_p) \asymp m^{-r-\alpha+(1/q-\max(1/2,1/p))_+} \end{aligned} \tag{2.7}$$

provided $r + \alpha > \alpha(q, p)$ with $\alpha(q, p)$ defined in Theorem 2.1. We proved in [**17**] that

$$\sigma_m(W^r H_q^\alpha, \mathcal{T})_p \asymp m^{-r-\alpha+(1/q-\max(1/2,1/p))_+} \tag{2.8}$$

under the same assumption $r + \alpha > \alpha(q, p)$. Relations (2.7) and (2.8) show that for any pair of (q, p), $1 \le q \le p \le \infty$, and for each function $f \in W^r H_q^\alpha$, the trigonometric system $\mathcal{T}$ provides a subspace $\mathcal{T}(\Lambda)$, $\#\Lambda \le m$, spanned by $\{e^{ikx}\}$, $k \in \Lambda$, such that

$$d_m(F[f], L_p) \le \sup_{y\in\mathbb{T}} \inf_{t\in\mathcal{T}(\Lambda)} \|f(\cdot - y) - t(\cdot)\|_p \ll d_m^{\text{ind}}(W^r H_q^\alpha, L_p).$$

Open Problems

2.1. Construct a subspace realizing (2.2).

2.2. Does there exist $f \in W^r H^\alpha_\infty$, $0 < \alpha < 1$, such that

$$d_n(F[f], L_1) \gg n^{-r-\alpha}?$$

3. Greedy Algorithms in Hilbert Spaces

Perhaps the first example of m-term approximation with regard to a redundant dictionary was considered by E. Schmidt in 1907 [**69**] 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 with properties of the integral operator

$$J_f(g) := \int_0^1 f(x, y) g(y)\, dy$$

with kernel $f(x, y)$. E. Schmidt [**69**] gave an expansion (known as the Schmidt expansion)

$$f(x, y) = \sum_{j=1}^{\infty} s_j(J_f) \varphi_j(x) \psi_j(y), \tag{3.S}$$

where $\{s_j(J_f)\}$ is a nonincreasing sequence of singular numbers of J_f, i.e., $s_j(J_f) := \lambda_j(J_f^* J_f)^{1/2}$, $\{\lambda_j(A)\}$ is a sequence of eigenvalues of an operator A, and J_f^* is the adjoint operator to J_f. The two sequences $\{\varphi_j(x)\}$ 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

$$\left\| f(x, y) - \sum_{j=1}^{m} s_j(J_f) \varphi_j(x) \psi_j(y) \right\|_{L_2}$$
$$= \inf_{u_j, v_j \in L_2, j=1,\dots,m} \left\| f(x, y) - \sum_{j=1}^{m} u_j(x) v_j(y) \right\|_{L_2}.$$

It follows from the Schmidt expansion that the above best bilinear approximation can be realized by the PGA. This was observed and used in several papers (see [**66**] for a history).

Another problem of this type which is well-known in statistics is the projection pursuit regression problem. We formulate the related results in the function theory

language. The problem is to approximate in $L_2(\Omega)$, $\Omega \subset \mathbb{R}^d$ is a bounded domain, a given function $f \in L_2(\Omega)$ by a sum of ridge functions, i.e., by

$$\sum_{j=1}^{m} r_j(\omega_j \cdot x), \qquad x, \omega_j \in \mathbb{R}^d, \quad j = 1, \dots, m,$$

where r_j, $j = 1, \dots, m$, are univariate functions. The following greedy-type algorithm (projection pursuit regression) was proposed in [**31**]. Assume that functions $r_1, \dots, r_{m-1}$ and vectors $\omega_1, \dots, \omega_{m-1}$ have been determined after $m-1$ steps of the 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 the PGA. The PGA and some other versions of greedy-type algorithms have been intensively studied recently (see [**4**], [**24**], [**13**], [**28**], [**18**], [**19**], [**34**], [**36**], [**37**], [**67**], [**101**], [**88**]–[**98**]). In this section we discuss the PGA and some of its modifications which make them more ready for implementation. We call this new type of greedy algorithm a Weak Greedy Algorithm (WGA) (see the Introduction for the definition of a PGA). Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, be given. Following [**95**] we define a WGA.

Weak Greedy Algorithm (WGA). *We define* $f_0^\tau := f$. *Then for each* $m \ge 1$, *we inductively define*:

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

$$|\langle f_{m-1}^\tau, \varphi_m^\tau \rangle| \ge t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^\tau, g \rangle|;$$

(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 in a particular case $t_k = t$, $k = 1, 2, \dots$, this algorithm was considered in [**36**]. Thus, the WGA is a generalization of the PGA in the direction of 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), (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)

$$f \sim \sum_{j=1}^{\infty} c_j(f)\varphi_j^\tau, \qquad c_j(f) := \langle f_{j-1}^\tau, \varphi_j^\tau \rangle.$$

In general it is not an expansion into an orthogonal series 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_{j-1}^τ. It is easy to see that

$$\|f_m^\tau\|^2 = \|f_{m-1}^\tau\|^2 - |c_m(f)|^2.$$

We prove the convergence of a greedy expansion (see, for instance, Theorem 3.1) and, therefore, from the above equality we get for this expansion an analog of the Parseval formula for orthogonal expansions:

$$\|f\|^2 = \sum_{j=1}^{\infty} |c_j(f)|^2.$$

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. We let $g(f) \in \mathcal{D}$ be an element from $\mathcal{D}$ which maximizes $|\langle f, g\rangle|$. We shall assume for simplicity that such a maximizer exists; if not, suitable modifications are necessary (see Weak Orthogonal Greedy Algorithm (WOGA) below) in the algorithm that follows.

Orthogonal Greedy Algorithm (OGA). *We define* $R_0^o(f) := R_0^o(f, \mathcal{D}) := f$ *and* $G_0^o(f) := G_0^o(f, \mathcal{D}) := 0$. *Then, for each* $m \geq 1$, *we inductively define*

$$\begin{aligned} H_m &:= H_m(f) := \mathrm{Span}\{g(R_0^o(f)), \dots, g(R_{m-1}^o(f))\}, \\ G_m^o(f) &:= G_m^o(f, \mathcal{D}) := P_{H_m}(f), \\ R_m^o(f) &:= R_m^o(f, \mathcal{D}) := f - G_m^o(f). \end{aligned}$$

We remark that for each f we have

$$\|f - G_m^o(f, \mathcal{D})\| \leq \|R_{m-1}^o(f) - G_1(R_{m-1}^o(f), \mathcal{D})\|. \tag{3.1}$$

Let a sequence $\tau = \{t_k\}_{k=1}^\infty$, $0 \leq t_k \leq 1$, be given. Following [**95**] we define a WOGA.

Weak Orthogonal Greedy Algorithm (WOGA). *We define* $f_0^{o,\tau} := f$. *Then for each* $m \geq 1$ *we inductively define*:

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

$$|\langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau}\rangle| \geq t_m \sup_{g \in \mathcal{D}} |\langle f_{m-1}^{o,\tau}, g\rangle|;$$

(2)

$$G_m^{o,\tau}(f,\mathcal{D}) := P_{H_m^\tau}(f), \qquad \text{where} \quad H_m^\tau := \operatorname{Span}(\varphi_1^{o,\tau},\dots,\varphi_m^{o,\tau});$$

(3)

$$f_m^{o,\tau} := f - G_m^{o,\tau}(f,\mathcal{D}).$$

It is clear that G_m^τ and $G_m^{o,\tau}$ in the case $t_k = 1$, $k = 1, 2, \dots$, coincide with the PGA G_m and the Orthogonal Greedy Algorithm (OGA) G_m^o, respectively. It is also clear that the WGA and 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 Theorem 3.9 for $t_k = 1, k = 1, \dots$, with (3.11). Also, we discuss below some other greedy-type algorithms, where the greedy step of finding a maximizer (or weak maximizer) is replaced by a thresholding step. We believe that such a modification makes these algorithms easier to implement. However, we should note that these new algorithms work only for f in the class $A_1(\mathcal{D})$ (see definitions and a discussion below, after Theorem 3.9).

3.1. *Convergence*

Convergence is a fundamental question and we discuss it in detail for the most general setting. The convergence of the PGA and the WGA with $t_k = t, 0 < t < 1$, was established in [**36**] and [**68**]. The first sufficient condition on τ which includes sequences with $\liminf_{k\to\infty} t_k = 0$ was obtained in [**95**].

Theorem 3.1. *Assume that*

$$\sum_{k=1}^{\infty} \frac{t_k}{k} = \infty. \tag{3.2}$$

Then, for any dictionary $\mathcal{D}$ and any $f \in H$, we have

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

In [**95**] we reduced the proof of the convergence of the WGA with the weakness sequence τ to some properties of l_2-sequences with regard to τ. Theorem 3.1 was derived from the following two statements proved in [**95**].

Proposition 3.1. *Let τ be such that, for any $\{a_j\}_{j=1}^{\infty} \in l_2$, $a_j \geq 0$, $j = 1, 2, \ldots$, 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 3.2. *If τ satisfies condition* (3.2) *then τ satisfies the assumption of Proposition* 3.1.

The following simple necessary condition

$$\sum_{k=1}^{\infty} t_k^2 = \infty$$

was mentioned in [**95**]. The first nontrivial necessary conditions were obtained in [**51**]. We proved in [**51**] the following theorem:

Theorem 3.2. *In the class of monotone sequences $\tau = \{t_k\}_{k=1}^{\infty}$, $1 \geq t_1 \geq t_2 \geq \cdots \geq 0$, condition* (3.2) *is necessary and sufficient for the convergence of a WGA for each f and all Hilbert spaces H and dictionaries $\mathcal{D}$.*

The proof of this theorem is based on a special procedure which we called Equalizer. In [**51**] we gave an example of a class of sequences τ for which condition (3.2) is not a necessary condition for convergence. We also proved in [**51**] a theorem which covers Theorem 3.1.

Theorem 3.3. *Assume that*

$$\sum_{s=0}^{\infty} \left(2^{-s} \sum_{k=2^s}^{2^{s+1}-1} t_k^2 \right)^{1/2} = \infty.$$

Then, for any dictionary $\mathcal{D}$ and any $f \in H$, we have

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

We proved in [**96**] a criterion on τ for convergence of the WGA. To explain this we need some notation.

We define by $\mathcal{V}$ the class of sequences $x = \{x_k\}_{k=1}^{\infty}$, $x_k \geq 0$, $k = 1, 2, \ldots$, with the following property: there exists a sequence $0 = q_0 < q_1 < \cdots$ such that

$$\sum_{s=1}^{\infty} \frac{2^s}{\Delta q_s} < \infty \tag{3.3}$$

and

$$\sum_{s=1}^{\infty} 2^{-s} \sum_{k=1}^{q_s} x_k^2 < \infty, \tag{3.4}$$

where $\Delta q_s := q_s - q_{s-1}$.

Theorem 3.4. *The condition $\tau \notin \mathcal{V}$ is necessary and sufficient for convergence of a WGA with weakness sequence τ for each f and all Hilbert spaces H and dictionaries $\mathcal{D}$.*

The proof of the sufficient part of Theorem 3.4 is a refinement of the original proof of Jones [**36**]. The study of the behavior of sequences $a_n \sum_{j=1}^{n} a_j$ for $\{a_j\}_{j=1}^{\infty} \in l_2, a_j \ge 0, j = 1, 2, \dots$, plays an important role in the proof. It turns out that the class $\mathcal{V}$ appears naturally in the study of the above-mentioned sequences. We proved in [**96**] the following theorem:

Theorem 3.5. *The following two conditions are equivalent*

$$\tau \notin \mathcal{V}, \tag{C.1}$$

$$\forall \{a_j\}_{j=1}^{\infty} \in l_2, \quad a_j \ge 0, \qquad \liminf_{n\to\infty} a_n \sum_{j=1}^{n} a_j / t_n = 0. \tag{C.2}$$

We give a result on convergence of the WOGA now.

Theorem 3.6. *Assume*

$$\sum_{k=1}^{\infty} t_k^2 = \infty. \tag{3.5}$$

Then, for any dictionary $\mathcal{D}$ and any $f \in H$, we have

$$\lim_{m\to\infty} \|f - G_m^{o,\tau}(f, \mathcal{D})\| = 0. \tag{3.6}$$

Remark 3.1. It is easy to see that in the case $\mathcal{D} = \mathcal{B}$ - orthonormal basis the assumption (3.5) is also necessary for convergence (3.6) for all f.

Theorems 3.4 and 3.6 show that conditions on the weakness sequence for convergence of the WGA and WOGA are different.

3.2. *Rate of Convergence*

For a general dictionary $\mathcal{D}$ we define the class of functions

$$\mathcal{A}_1^o(\mathcal{D}, M) := \left\{ f \in H : f = \sum_{k\in\Lambda} c_k w_k,\ w_k \in \mathcal{D},\ \#\Lambda < \infty \text{ and } \sum_{k\in\Lambda} |c_k| \le M \right\}$$

and we define $\mathcal{A}_1(\mathcal{D}, M)$ as the closure (in H) of $\mathcal{A}_1^o(\mathcal{D}, M)$. Furthermore, we define $\mathcal{A}_1(\mathcal{D})$ as 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

$$|f|_{\mathcal{A}_1(\mathcal{D})}$$

as the smallest M such that $f \in \mathcal{A}_1(\mathcal{D}, M)$.

It was proved in [**18**] that for a general dictionary $\mathcal{D}$ the PGA provides the following estimate

$$\|f - G_m(f, \mathcal{D})\| \le |f|_{\mathcal{A}_1(\mathcal{D})} m^{-1/6}. \tag{3.7}$$

(In this and similar estimates we consider that the inequality holds for all possible choices of $\{G_m\}$.) Paper [**18**] also contains an example of a dictionary $\mathcal{D}$ and an element f such that (see Subsection 3.3 below)

$$\|f - G_m(f, \mathcal{D})\| > \tfrac{1}{2}|f|_{\mathcal{A}_1(\mathcal{D})} m^{-1/2}, \qquad m \ge 4. \tag{3.8}$$

We proved in [**45**] a new estimate

$$\|f - G_m(f, \mathcal{D})\| \le 4|f|_{\mathcal{A}_1(\mathcal{D})} m^{-11/62} \tag{3.9}$$

which improves a little the original one (see (3.7)).

E. Livshitz [**50**] proved that there exist $\delta > 0$, a dictionary $\mathcal{D}$, and an element $f \in H$, $f \ne 0$, such that

$$\|f - G_m(f, \mathcal{D})\| \ge C m^{-1/2+\delta} |f|_{\mathcal{A}_1(\mathcal{D})} \tag{3.10}$$

with a positive constant C. We developed and refined ideas from [**50**] in [**98**] and proved the following lower estimate. There exist a dictionary $\mathcal{D}$ and an element $f \in H$, $f \ne 0$, such that

$$\|f - G_m(f, \mathcal{D})\| \ge C m^{-1/3} |f|_{\mathcal{A}_1(\mathcal{D})} \tag{3.11}$$

with a positive constant C.

For the WGA we have the following estimate [**95**]:

Theorem 3.7. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Assume that $\tau := \{t_k\}_{k=1}^{\infty}$ is a nonincreasing sequence. Then, for $f \in \mathcal{A}_1(\mathcal{D}, M)$, we have*

$$\|f - G_m^{\tau}(f, \mathcal{D})\| \le M \left(1 + \sum_{k=1}^{m} t_k^2\right)^{-t_m/2(2+t_m)}. \tag{3.12}$$

In a particular case $\tau = t$ ($t_k = t$, $k = 1, 2, \ldots$), (3.12) gives

$$\|f - G_m^{t}(f, \mathcal{D})\| \le M(1 + mt^2)^{-t/(4+2t)}, \qquad 0 < t \le 1.$$

This estimate implies the following inequality

$$\|f - G_m^t(f, \mathcal{D})\| \le C_1 m^{-t/6} |f|_{\mathcal{A}_1(\mathcal{D})}, \tag{3.13}$$

with the exponent $t/6$ approaching 0 linearly in t. We proved in [**98**] that this exponent cannot decrease to 0 at a slower rate than linearly.

Theorem 3.8. *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*

$$\liminf_{m\to\infty} \|f_t - G_m^t(f_t, \mathcal{D}_t)\| m^{bt} / |f_t|_{\mathcal{A}_1(\mathcal{D}_t)} > 0.$$

We formulate one result for the WOGA from [**95**]. In the case of the OGA this theorem was proved in [**18**].

Theorem 3.9. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Then, for each $f \in \mathcal{A}_1(\mathcal{D}, M)$, we have*

$$\|f - G_m^{o,\tau}(f, \mathcal{D})\| \le M \left(1 + \sum_{k=1}^{m} t_k^2\right)^{-1/2}.$$

There is one more greedy-type algorithm which works well for functions from the convex hull $A_1(\mathcal{D}) := \{f : |f|_{\mathcal{A}_1(\mathcal{D})} \le 1\}$ of $\mathcal{D}^{\pm}$, where $\mathcal{D}^{\pm} := \{\pm g,\ g \in \mathcal{D}\}$.

There are several modifications of the Relaxed Greedy Algorithm (RGA) (see, for instance, [**4**], [**18**]). Before giving the definition of the Weak Relaxed Greedy Algorithm (WRGA) we make one remark which helps to motivate the corresponding definition. Assume that $G_{m-1} \in A_1(\mathcal{D})$ 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 $G_m := (1-a)G_{m-1} + ag$, $g \in \mathcal{D}^{\pm}$, $0 \le a \le 1$. This form guarantees that $G_m \in A_1(\mathcal{D})$. We now give the definition of two versions of the WRGA.

Weak Relaxed Greedy Algorithm (WRGA). *We define $f_0^{\tau,i} := f$ and $G_0^{\tau,i} := 0$ for $i = 1, 2$. Then for each $m \ge 1$ we inductively define*:

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

$$\langle f_{m-1}^{\tau,1}, \varphi_m^{\tau,1} - G_{m-1}^{\tau,1} \rangle \ge t_m \|f_{m-1}^{\tau,1}\|^2 \tag{3.14}$$

and

$$\|\varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\| \ge \|f_{m-1}^{\tau,1}\|. \tag{3.15}$$

$\varphi_m^{\tau,2} \in \mathcal{D}^{\pm}$ is any element satisfying

$$\langle f_{m-1}^{\tau,2}, \varphi_m^{\tau,2} - G_{m-1}^{\tau,2} \rangle \ge t_m \|f_{m-1}^{\tau,2}\|^2. \tag{3.16}$$

(2)

$$\begin{aligned} G_m^{\tau,1} &:= G_m^{\tau,1}(f,\mathcal{D}) := (1-\alpha_m)G_{m-1}^{\tau,1} + \alpha_m\varphi_m^{\tau,1}, \\ \alpha_m &:= \langle f_{m-1}^{\tau,1}, \varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\rangle \|\varphi_m^{\tau,1} - G_{m-1}^{\tau,1}\|^{-2}, \\ G_m^{\tau,2} &:= G_m^{\tau,2}(f,\mathcal{D}) := (1-\beta_m)G_{m-1}^{\tau,2} + \beta_m\varphi_m^{\tau,2}, \\ \beta_m &:= t_m\left(1+\sum_{k=1}^{m} t_k^2\right)^{-1} \qquad \textit{for} \quad m \geq 1. \end{aligned}$$

(3)

$$f_m^{\tau,i} := f - G_m^{\tau,i}, \qquad i = 1, 2.$$

We now formulate some theorems on convergence rates of greedy-type algorithms WRGA for functions from $\mathcal{A}_1(\mathcal{D}, M)$.

Theorem 3.10. *Let $\mathcal{D}$ be an arbitrary dictionary in H. Then, for each $f \in A_1(\mathcal{D})$, we have*

$$\|f - G_m^{\tau,i}(f,\mathcal{D})\| \leq 2\left(1+\sum_{k=1}^{m} t_k^2\right)^{-1/2}, \qquad i = 1, 2. \tag{3.17}$$

We note that in both versions of WRGAs the "greedy" steps (3.14) and (3.16) can be easily checked because of their thresholding nature. We now discuss a question of replacing "greedy" steps in WGAs and WOGAs by something similar to (3.14) and (3.16). Let us begin with a WOGA. Inspecting the proof of Theorem 3.9 (see [**95**, p. 222]) one realizes that the relation

$$|\langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau}\rangle| \geq t_m \sup_{g\in\mathcal{D}} |\langle f_{m-1}^{o,\tau}, g\rangle| \tag{3.18}$$

was used only for deriving the inequality

$$|\langle f_{m-1}^{o,\tau}, \varphi_m^{o,\tau}\rangle| \geq t_m \|f_{m-1}^{o,\tau}\|^2. \tag{3.19}$$

Thus if we define a Modified WOGA (MWOGA) by replacing (3.18) by (3.19) in Step (1) of the definition of a WOGA we get an analog of Theorem 3.9 for a MWOGA.

Let us proceed to a modification of a WGA. Let a sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \leq t_k \leq 1$, be given. We define a Modified Weak Greedy Algorithm (MWGA) for $f \in A_1(\mathcal{D})$.

Modified Weak Greedy Algorithm (MWGA). *We define $f_0^M := f$, $\varphi_1^M \in \mathcal{D}$ is any element satisfying*

$$|\langle f_0^M, \varphi_1^M\rangle| \geq t_1 \|f_0^M\|^2,$$

and we set

$$f_1^M := f_0^M - \langle f_0^M, \varphi_1^M \rangle \varphi_1^M.$$

Then for each $m > 1$ *we inductively define*:

(1) $\varphi_m^M \in \mathcal{D}$ *is any element satisfying*

$$|\langle f_{m-1}^M, \varphi_m^M \rangle| \geq t_m \|f_{m-1}^M\|^2 \left(1 + \sum_{k=1}^{m-1} |\langle f_{k-1}^M, \varphi_k^M \rangle|\right)^{-1};$$

(2)

$$f_m^M := f_{m-1}^M - \langle f_{m-1}^M, \varphi_m^M \rangle \varphi_m^M.$$

Proposition 3.3. *Let* $\tau = \{t_k\}$, $0 \leq t_k < 1$, $k = 1, 2, \ldots$, *be a nonincreasing sequence. For any* $f \in A_1(\mathcal{D})$ *there exists a realization of a MWGA. For any such realization we have*

$$\|f_m^M\| \leq \left(1 + \sum_{k=1}^{m} t_k^2\right)^{-t_m/(4+2t_m)}.$$

The proof of this proposition repeats the proof of Theorem 3.7 (see [**95**, Proof of Theorem 5.1]).

Let us make one more remark on the numerical implementation of greedy-type algorithms. We now know that after the modifications discussed above it is sufficient for the implementation of each of the greedy-type algorithms (WRGA, MWOGA, MWGA) to satisfy a thresholding-type inequality. However, the remaining problem is that we can only guarantee a realization of such algorithms under the assumption $f \in A_1(\mathcal{D})$. In some problems this assumption is satisfied automatically. In [**8**] the general procedure called "Basis Pursuit" was proposed for finding a representation of f with a minimal ℓ_1-norm of coefficients. It is pointed out in [**8**] that Basis Pursuit is a linear programming problem. Thus, one can use the following two-step implementation strategy:

(1) For a given f find (or estimate) $|f|_{\mathcal{A}_1(\mathcal{D})}$ using Basis Pursuit;
(2) Consider $f/|f|_{\mathcal{A}_1(\mathcal{D})} \in A_1(\mathcal{D})$ and use any of the above modified greedy algorithms.

We present some results from [**92**] on r-greedy dictionaries.

Definition 3.1. Let $r > 0$ be given. We call a dictionary $\mathcal{D}$ an r-greedy dictionary for H if $\mathcal{D}$ possesses the property (G): for any $f \in H$ such that

$$\sigma_m(f, \mathcal{D}) \leq m^{-r}, \qquad m = 1, 2, \ldots,$$

we have

$$\|f - G_m(f, \mathcal{D})\| \leq C(r, \mathcal{D}) m^{-r}, \qquad m = 1, 2, \ldots.$$

Definition 3.2. We say $\mathcal{D}$ is a λ-quasi-orthogonal dictionary if, for any $n \in \mathcal{N}$ and any $g_i \in \mathcal{D}, i = 1, \dots, n$, there exists a collection $\varphi_j \in \mathcal{D}$, $j = 1, \dots, M$, $M \le N := \lambda n$, with the properties

$$g_i \in X_M := \operatorname{Span}(\varphi_1, \dots, \varphi_M),$$

and for any $f \in X_M$ we have

$$\max_{1 \le j \le M} |\langle f, \varphi_j \rangle| \ge N^{-1/2} \|f\|.$$

Theorem 3.11. *Let a given dictionary $\mathcal{D}$ be λ-quasi-orthogonal and let $0 < r < (2\lambda)^{-1}$ be a real number. Then for any f such that*

$$\sigma_m(f, \mathcal{D}) \le m^{-r}, \qquad m = 1, 2, \dots,$$

we have

$$\|f - G_m(f, \mathcal{D})\| \le C(r, \lambda) m^{-r}, \qquad m = 1, 2, \dots.$$

Remark 3.2. It is clear that an orthonormal dictionary is a 1-quasi-orthogonal dictionary.

Remark 3.3. Theorem 3.11 holds if the assumption that $\mathcal{D}$ is λ-quasi-orthogonal is replaced by the assumption that $\mathcal{D}$ is asymptotically λ-quasi-orthogonal. In order to get the definition of an asymptotically λ-quasi-orthogonal dictionary we replace N in Definition 3.2 by $N(n)$ and instead of $N = \lambda n$ we require

$$\limsup_{n \to \infty} N(n)/n = \lambda.$$

Here are two examples of asymptotically λ-quasi-orthogonal dictionaries.

Example 3.1. The dictionary $\chi := \{f = |J|^{-1/2} \chi_J, J \subset [0, 1]\}$, where χ_J is the characteristic function of an interval J, is an asymptotically 2-quasi-orthogonal dictionary.

Example 3.2. The dictionary $\mathcal{P}(r)$ that consists of functions of the form $f = p\chi_J$, $\|f\| = 1$, where p is an algebraic polynomial of degree $r - 1$ and χ_J is the characteristic function of an interval J, is asymptotically $2r$-quasi-orthogonal.

Example 3.3. For given $\mu, \gamma \ge 1$ a dictionary $\mathcal{D}$ is called (μ, γ)-semistable if for any $g_i \in \mathcal{D}, i = 1, \dots, n$, there exist elements $h_j \in \mathcal{D}$, $j = 1, \dots, M \le \mu n$, such that

$$g_i \in \operatorname{Span}\{h_1, \dots, h_M\}$$

and for any $c_1, \dots, c_M$ we have

$$\left\| \sum_{j=1}^{M} c_j h_j \right\| \ge \gamma^{-1/2} \left(\sum_{j=1}^{M} c_j^2 \right)^{1/2}.$$

A (μ, γ)-semistable dictionary $\mathcal{D}$ is $\mu\gamma$-quasi-orthogonal.

3.3. *Saturation Property of the Pure Greedy Algorithm*

We consider in this subsection a generalization of the PGA. Take a fixed number $n \in \mathcal{N}$ and define the basic step of the n-Dimensional Greedy Algorithm as follows. Find an n-term polynomial

$$p_n(f) := p_n(f, \mathcal{D}) = \sum_{n=1}^{n} c_i g_i, \qquad g_i \in \mathcal{D}, \quad i = 1, \dots, n,$$

such that (we assume its existence)

$$\|f - p_n(f)\| = \sigma_n(f, \mathcal{D}).$$

Denote

$$G(n, f) := G(n, f, \mathcal{D}) := p_n(f), \qquad R(n, f) := R(n, f, \mathcal{D}) := f - p_n(f).$$

n-Dimensional Greedy Algorithm. *We define $R_0(n, f) := f$ and $G_0(n, f) := 0$. Then, for each $m \ge 1$, we inductively define*

$$\begin{aligned} G_m(n, f) &:= G_m(n, f, \mathcal{D}) := G_{m-1}(n, f) + G(n, R_{m-1}(n, f)), \\ R_m(n, f) &:= R_m(n, f, \mathcal{D}) := f - G_m(n, f) = R(n, R_{m-1}(n, f)). \end{aligned} \tag{3.20}$$

It is clear that a One-Dimensional Greedy Algorithm is a PGA.

For a general dictionary $\mathcal{D}$, and for any $0 < \beta \le 1$, we define the class of functions

$$\mathcal{A}_\beta^o(\mathcal{D}, M) := \left\{ f \in H : f = \sum_{k \in \Lambda} c_k w_k,\ w_k \in \mathcal{D},\ |\Lambda| < \infty \text{ and } \sum_{k \in \Lambda} |c_k|^\beta \le M^\beta \right\},$$

and we define $\mathcal{A}_\beta(\mathcal{D}, M)$ as the closure (in H) of $\mathcal{A}_\beta^o(\mathcal{D}, M)$. Furthermore, we define $\mathcal{A}_\beta(\mathcal{D})$ as the union of the classes $\mathcal{A}_\beta(\mathcal{D}, M)$ over all $M > 0$. For $f \in \mathcal{A}_\beta(\mathcal{D})$, we define the "quasi-norm"

$$|f|_{\mathcal{A}_\beta(\mathcal{D})}$$

as the smallest M such that $f \in \mathcal{A}_\beta(\mathcal{D}, M)$. The following general estimate for the error in the approximation of functions $f \in \mathcal{A}_\beta(\mathcal{D})$, $\beta \le 1$, was proved in [**18**].

Theorem 3.12. *If $f \in \mathcal{A}_\beta(\mathcal{D})$, $\beta \le 1$, then for $\alpha := 1/\beta - \frac{1}{2}$, we have*

$$\sigma_m(f, \mathcal{D}) \le C|f|_{\mathcal{A}_\beta(\mathcal{D})} m^{-\alpha}, \qquad m = 1, 2, \ldots, \tag{3.21}$$

where C depends on β if β is small.

In [**18**] we gave an example which showed that replacing a dictionary $\mathcal{B}$, given by an orthogonal basis, by a nonorthogonal redundant dictionary $\mathcal{D}$ may damage the efficiency of the PGA. The dictionary $\mathcal{D}$ in our example differs from the dictionary $\mathcal{B}$ by only one suitably chosen element g.

Let $\{h_k\}_{k=1}^\infty$ be an orthonormal basis in a Hilbert space H and let $\mathcal{B} = \{h_k\}_{k=1}^\infty$ be the corresponding dictionary. Consider the following element

$$g := Ah_1 + Ah_2 + aA \sum_{k \ge 3} (k(k+1))^{-1/2} h_k$$

with

$$A := (\tfrac{33}{89})^{1/2} \qquad \text{and} \qquad a := (\tfrac{23}{11})^{1/2}.$$

Then, $\|g\| = 1$. We define the dictionary $\mathcal{D} = \mathcal{B} \cup \{g\}$.

Theorem 3.13. *For the function*

$$f = h_1 + h_2$$

which is in each space $\mathcal{A}_\beta(\mathcal{D})$, $0 < \beta \le 1$, we have

$$\|f - G_m(f, \mathcal{D})\| \ge m^{-1/2}, \qquad m \ge 4. \tag{3.22}$$

We proved in [**92**] that the n-Dimensional Greedy Algorithm, like the PGA has a saturation property.

Theorem 3.14. *For a given n and any orthonormal basis $\{h_k\}_{k=1}^\infty$ there exists an element g such that for the dictionary $\mathcal{D} = g \cup \{h_k\}_{k=1}^\infty$ there is an element f which has the property, for any $0 < \beta \le 1$,*

$$\|f - G_m(n, f)\| / |f|_{\mathcal{A}_\beta(\mathcal{D})} \ge C(\beta) n^{-1/\beta} (m+2)^{-1/2}.$$

Open Problems

3.1. Find the order of decay of the sequence

$$\gamma(m) := \sup_{f, \mathcal{D}, \{G_m\}} (\|f - G_m(f, \mathcal{D})\| |f|^{-1}_{\mathcal{A}_1(\mathcal{D})}),$$

where sup is taken over all dictionaries $\mathcal{D}$, all elements $f \in \mathcal{A}_1(\mathcal{D}) \backslash \{0\}$, and all possible choices of $\{G_m\}$.

3.2. Is there a greedy-type algorithm realizing (3.21) for $0 < \beta < 1$?

4. Greedy Algorithms in Banach Spaces

In this section we present some results on greedy approximation with regard to redundant dictionaries in Banach spaces. These results are fragmentary and should be considered as an attempt to understand a role of redundancy and nonlinearity in the general setting for Banach spaces. There are no general results on the convergence of the X-Greedy Algorithm and the DGA. Some results about the performance of DGAs can be found in [**24**] and [**28**]. It is proved in [**24**] (see also [**28**]) that the assumption that X is a smooth Banach space is a necessary and sufficient condition for the sequence $\{\|R_m^D(f,\mathcal{D})\|_X\}$ to be strictly decreasing for each $f \in X$ and all dictionaries $\mathcal{D}$.

4.1. *Uniformly Smooth Banach Spaces*

Recently, we proved in [**97**] one general convergence result for the generalization of the WOGA to Banach spaces. We call this generalization a Weak Chebyshev Greedy Algorithm (WCGA). We will use the notation $\mathcal{D}^{\pm} := \{\pm g, g \in \mathcal{D}\}$ here. Let a weakness sequence $\tau = \{t_k\}_{k=1}^{\infty}$, $0 \le t_k \le 1$, be given.

Weak Chebyshev Greedy Algorithm (WCGA). *We define* $f_0^c := f_0^{c,\tau} := f$. *Then for each* $m \ge 1$ *we inductively define*:

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

$$F_{f_{m-1}^c}(\varphi_m^c) \ge t_m \sup_{g\in\mathcal{D}^{\pm}} F_{f_{m-1}^c}(g).$$

(2) *Define*

$$\Phi_m := \Phi_m^{\tau} := \mathrm{Span}\{\varphi_j^c\}_{j=1}^m,$$

and define $G_m^c := G_m^{c,\tau}$ *to be the best approximant to* f *from* Φ_m.

(3) *Denote*

$$f_m^c := f_m^{c,\tau} := f - G_m^c.$$

Let us give one more definition of a weak greedy-type algorithm. We will not present results on it here.

Weak Dual Greedy Algorithm (WDGA). *We define* $f_0^D := f_0^{D,\tau} := f$. *Then for each* $m \ge 1$ *we inductively define*:

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

$$F_{f_{m-1}^D}(\varphi_m^D) \ge t_m \sup_{g\in\mathcal{D}^{\pm}} F_{f_{m-1}^D}(g).$$

(2) *Define* a_m *as*

$$\|f_{m-1}^D - a_m\varphi_m^D\| = \min_{a\in\mathbb{R}} \|f_{m-1}^D - a_m\varphi_m^D\|.$$

(3) *Denote*

$$f_m^D := f_m^{D,\tau} := f_{m-1}^D - a_m\varphi_m^D.$$

We now define the generalization for Banach spaces of the WRGA studied in [**95**] in the case of Hilbert space.

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 inductively define*:

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

$$F_{f_{m-1}^r}(\varphi_m^r - G_{m-1}^r) \geq t_m \sup_{g\in\mathcal{D}^{\pm}} F_{f_{m-1}^r}(g - G_{m-1}^r).$$

(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) *Denote*

$$f_m^r := f_m^{r,\tau} := f - G_m^r.$$

Remark 4.1. It follows from the definition of a WCGA, a WDGA, and a WRGA that the sequences $\{\|f_m^c\|\}$, $\{\|f_m^D\|\}$, and $\{\|f_m^r\|\}$ are nonincreasing sequences.

The term "weak" in these definitions means that at Step (1) we do not shoot for the optimal element of the dictionary which realizes the corresponding sup but we are satisfied with a weaker property than being optimal. The obvious reason for this is that we do not know in general that the optimal element exists. Another, practical reason is that the weaker the assumption the easier it is to satisfy it and, therefore, easier to realize in practice. The WRGA provides incremental approximants discussed in [**24**]. In [**24**] they also impose weaker assumptions (ε-greedy) on an element of the dictionary rather than being optimal. For instance, for a given sequence $\{\varepsilon_n\}_{n=1}^{\infty}$, $\varepsilon_n > 0$, $n = 1, 2, \ldots$, they take $0 \leq \alpha_m \leq 1$ and $g_m \in \mathcal{D}$ satisfying

$$\|f - ((1-\alpha_m)G_{m-1} + \alpha_m g_m)\| \leq \inf_{0\leq\alpha\leq 1, g\in\mathcal{D}} \|f - ((1-\alpha)G_{m-1} + \alpha g)\| + \varepsilon_m$$

instead of trying to find optimal elements. Their approach is different from ours.

We discuss in this section the questions of convergence and the rate of convergence for the two above-defined methods of approximation: WCGA and WRGA. It is clear that in the case of WRGAs the assumption that f belongs to the closure of the convex hull of $\mathcal{D}^{\pm}$ is natural. We denote the closure of the convex hull of $\mathcal{D}^{\pm}$ by $A(\mathcal{D}) := A_1(\mathcal{D})$. It has been proven in [**95**] (see Theorems 3.9 and 3.10 from Section 3) that in the case of the Hilbert space both algorithms, WCGA and WRGA, give the approximation error for the class $A(\mathcal{D})$ of the order

$$\left(1 + \sum_{k=1}^{m} t_k^2\right)^{-1/2}.$$

We consider here approximation in uniformly smooth Banach spaces. For a Banach space X we define the modulus of smoothness

$$\rho(u) := \sup_{\|x\|=\|y\|=1} (\tfrac{1}{2}(\|x + uy\| + \|x - uy\|) - 1).$$

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

$$\max(0, u - 1) \le \rho(u) \le u, \qquad u \in (0, \infty). \tag{4.1}$$

It has been established in [**24**] that the approximation error of an algorithm analogous to our WRGA with $t_k = 1, k = 1, 2, \dots$, for the class $A(\mathcal{D})$ can be expressed in terms of a modulus of smoothness of a Banach space. Namely, if the modulus of smoothness ρ of X satisfies the inequality $\rho(u) \le \gamma u^q, q > 1$, then the error is of $O(m^{1/q-1})$. It has been proven in [**97**] that both algorithms, WCGA and WRGA, provide approximation for the class $A(\mathcal{D})$ in a Banach space X with modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$, of order

$$\left(1 + \sum_{k=1}^{m} t_k^p\right)^{-1/p}, \qquad p := \frac{q}{q-1}. \tag{4.2}$$

We also proved (see [**97**]) that the WCGA converges for any $f \in X$ and that the WRGA converges for any $f \in A(\mathcal{D})$ if τ satisfies the condition

$$\sum_{m=1}^{\infty} t_m \xi_m(\rho, \tau, \theta) = \infty. \tag{4.3}$$

The sequences $\{\xi_m(\rho, \tau, \theta)\}$ are defined as follows:

Definition 4.1. Let $\rho(u)$ be an even convex function on $(-\infty, \infty)$ with the property: $\rho(2) \geq 1$ and

$$\lim_{u \to 0} \rho(u)/u = 0.$$

For any $\tau = \{t_k\}_{k=1}^{\infty}$, $0 < t_k \leq 1$, and $0 < \theta \leq \frac{1}{2}$ we define $\xi_m := \xi_m(\rho, \tau, \theta)$ as a number u satisfying the equation

$$\rho(u) = \theta t_m u. \tag{4.4}$$

In a particular case of $\rho(u) \asymp u^q$, $1 < q \leq 2$, relation (4.3) is equivalent to

$$\sum_{k=1}^{m} t_k^p = \infty, \qquad p := \frac{q}{q-1}. \tag{4.5}$$

We gave in [**97**] an example which shows that (4.5) is a necessary condition for the convergence of the WCGA in Banach spaces with a modulus of smoothness of power type q for all $\mathcal{D}$ and $f \in X$.

It is well-known (see, for instance, [**24**, Lemma B.1]) that in the case $X = L_p$, $1 \leq p < \infty$, we have

$$\rho(u) \leq \begin{cases} u^p/p & \text{if } 1 \leq p \leq 2, \\ (p-1)u^2/2 & \text{if } 2 \leq p < \infty. \end{cases} \tag{4.6}$$

It is also known (see [**49**, p. 63]) that, for any X with $\dim X = \infty$, one has

$$\rho(u) \geq (1 + u^2)^{1/2} - 1$$

and, for every X, $\dim X \geq 2$,

$$\rho(u) \geq Cu^2, \qquad C > 0.$$

This limits power-type moduli of smoothness of nontrivial Banach spaces to the case $1 \leq q \leq 2$. The following theorem gives the rate of convergence of the WCGA for f in $A(\mathcal{D})$.

Theorem 4.1. *Let X be a uniformly smooth Banach space with the modulus of smoothness $\rho(u) \leq \gamma u^q$, $1 < q \leq 2$. Then, for a sequence $\tau := \{t_k\}_{k=1}^{\infty}$, $t_k \leq 1$, $k = 1, 2, \ldots$, we have for any $f \in A(\mathcal{D})$ that*

$$\|f - G_m^{c,\tau}(f, \mathcal{D})\| \leq C(q, \gamma) \left(1 + \sum_{k=1}^{m} t_k^p\right)^{-1/p}, \qquad p := \frac{q}{q-1},$$

with a constant $C(q, \gamma)$ which may depend only on q and γ.

4.2. *Finite-Dimensional Spaces*

We discuss some results from [**19**] on X-Greedy Algorithms in a particular case of finite-dimensional space $X = \mathbb{R}^n$, equipped with one of the standard norms ℓ_p. The reasons for our concentration on the finite-dimensional problems are as follows. It is well-known how one can apply the finite-dimensional results in studying the smoothness classes. Next, we are interested in understanding an interplay of several parameters including a parameter measuring the redundancy of a system $\mathcal{D}$. In this subsection it will be more convenient for us to use systems $\mathcal{D}$ that are not necessarily normalized. We note that the definition of an X-Greedy Algorithm does not depend on the normalization of a system.

We use the standard notation $\mathbb{R}^n$ for the n-dimensional space of real vectors and the ℓ_p-norm is defined as follows:

$$\|x\|_p := \left(\sum_{j=1}^n |x_j|^p\right)^{1/p}, \qquad 1 \le p < \infty,$$

$$\|x\|_\infty := \max_j |x_j|.$$

Let B_p^n denote the unit ℓ_p-ball of $\mathbb{R}^n$.

First we give two theorems from [**19**] about the m-term approximation in $\mathbb{R}^n$. In this subsection, we shall consider m-term approximation in the ℓ_p-norm of certain sets $F \subset \mathbb{R}^n$. In Theorem 4.2, we use ideas from [**42**] to give a lower estimate for m-term approximation in the ℓ_1-norm from a general dictionary to general sets $F \subset \mathbb{R}^n$. Lower estimates in the ℓ_1-norm automatically provide lower estimates in the other ℓ_q-norms, $q > 1$ (see Corollary 4.1).

We let $\mathrm{Vol}_n(S)$ denote the Euclidean n-dimensional volume of the set $S \subset \mathbb{R}^n$. We recall that the volume of the unit ball B_p^n, $1 \le p \le \infty$, in $\mathbb{R}^n$ can be estimated by

$$C_1^n n^{-n/p} \le \mathrm{Vol}_n(B_p^n) \le C_2^n n^{-n/p}, \tag{4.7}$$

with $C_1, C_2 > 0$ absolute constants.

Theorem 4.2. *If $F \subset B_2^n$ satisfies*

$$\mathrm{Vol}_n F \ge K^n \mathrm{Vol}_n B_2^n,$$

for some $0 < K \le 1$, then for any dictionary $\mathcal{D}$, $\#\mathcal{D} = N$, we have

$$\sigma_m(F, \mathcal{D})_1 \ge C K^2 n^{1/2} N^{-m/(n-m)}, \qquad m \le n/2.$$

with $C > 0$ an absolute constant.

Corollary 4.1. *Let F and $\mathcal{D}$ be as in Theorem* 4.2. *For any $1 \le q \le \infty$, we have*

$$\sigma_m(F, \mathcal{D})_q \ge C K^2 n^{1/q-1/2} N^{-m/(n-m)}, \qquad m \le n/2,$$

with C an absolute constant.

Corollary 4.2. *Let $\mathcal{D}$ be as in Theorem* 4.2. *For any $1 \le p, q \le \infty$, we have*

$$\sigma_m(B_p^n, \mathcal{D})_q \ge Cn^{1/q-1/p} N^{-m/(n-m)}, \qquad m \le n/2, \tag{4.8}$$

with C an absolute constant.

Remark 4.2. In the case $N = a^n$ and $p = q$, the lower bound in Corollary 4.2 can be replaced by Ca^{-2m}.

We shall next consider upper estimates for $\sigma_m(F, \mathcal{D})_p$. We begin with the following simple theorem:

Theorem 4.3. *Let X be any n-dimensional Banach space and let B be its unit ball. For any N there exists a system $\mathcal{D} \subset X$, $\#\mathcal{D} = N$, such that*

$$\sigma_m(B, \mathcal{D})_X \le \min(1, \varepsilon_N^m), \qquad \varepsilon_N := \frac{2}{N^{1/n} - 1}. \tag{4.9}$$

We now consider the ℓ_p-Greedy Algorithms, $1 \le p \le \infty$ (see the Introduction for the definition). In the case $p = 2$, the ℓ_p-Greedy Algorithm coincides with the PGA. Then, $G_m^p(x) := G_m^{\ell_p}(x, \mathcal{D})$ is an m-term approximation to x from $\mathcal{D}$ which we call the mth-greedy approximant. We note that the best approximation to $x \in \mathbb{R}^n$ from $\mathcal{D}$ is not necessarily unique and therefore $G_m^p(x)$ is not necessarily unique. We define

$$\bar{\gamma}_m^p(x, \mathcal{D})_q := \sup \|x - G_m^p(x, \mathcal{D})\|_q,$$

where the supremum is taken over all possible resulting $G_m^p(x, \mathcal{D})$. Similarly, we define

$$\underline{\gamma}_m^p(x, \mathcal{D})_q := \inf \|x - G_m^p(x, \mathcal{D})\|_q,$$

where the infimum is taken over all possible resulting $G_m^p(x, \mathcal{D})$. Thus, $\bar{\gamma}$ measures the worst possible error over all possible choices of best approximations in the greedy algorithm and $\underline{\gamma}$ represents the best possible error.

More generally, for a class $F \subset \mathbb{R}^n$, we define

$$\bar{\gamma}_m^p(F, \mathcal{D})_q := \sup_{f \in F} \bar{\gamma}_m^p(f, \mathcal{D})_q$$

with a similar definition for $\underline{\gamma}_m^p(F, \mathcal{D})_q$. In upper estimates for greedy approximation we would like to use $\bar{\gamma}$ and for lower estimates $\underline{\gamma}$.

Theorem 4.3 shows that for $p = q$ and for each $a > 1$ there exists a dictionary $\mathcal{D}$, $\#\mathcal{D} = b^n$, $b = 2a + 1$, such that

$$\bar{\gamma}_m^p(B_p^n, \mathcal{D})_p \le a^{-m}.$$

However, the dictionary $\mathcal{D}$ in that theorem is not very natural or easy to describe. This estimate and Remark 4.2 to Corollary 4.2 indicate that systems $\mathcal{D}$ with $\#\mathcal{D}$ of order C^n play an important role in m-term approximation in $\mathbb{R}^n$. We proceed now to study a natural family of such systems. We present results from [**19**].

Let $M \geq 3$ be an integer and consider the partition of $[-1, 1]$ into M disjoint intervals I_i of equal length: $|I_i| = 2/M$, $i = 1, \dots, M$. We let ξ_i denote the midpoint of the interval I_i, $i = 1, \dots, M$, and $\Xi := \{\xi_i\}_{i=1}^M$. We introduce the system

$$\mathcal{V}_M := \{x \in \mathbb{R}^n : x_j \in \Xi,\ j = 1, \dots, n\}.$$

Clearly $\#\mathcal{V}_M = M^n$. We shall study in this section the ℓ_∞-Greedy Algorithm for the systems $\mathcal{V}_M$.

Theorem 4.4. *For any $1 \leq q \leq \infty$, we have*

$$\bar{\gamma}_m^\infty(B_\infty^n, \mathcal{V}_M)_q \leq n^{1/q} M^{-m}, \qquad m = 1, 2, \dots. \tag{4.10}$$

We shall give results about the ℓ_1-Greedy Algorithm for the system $\mathcal{V}_3$. We consider this system in detail for the following reasons. It is a simple system which is easy to describe geometrically. Also, it is fairly easy to analyze the approximation properties of this system. Moreover, it turns out that this system gives a geometric order of approximation (see, e.g., Theorems 4.5 and 4.7) which we know is the best we can expect for general dictionairies (see Corollary 4.2).

Theorem 4.5. *We have the estimate*

$$\sigma_m(B_1^n, \mathcal{V}_3)_1 \leq \bar{\gamma}_m^1(B_1^n, \mathcal{V}_3)_1 \leq \left(1 - \frac{1}{k+1}\right)^m \tag{4.11}$$

where $k := [\log_2(n+1)]$.

The following lower estimate shows that (4.11) cannot be improved by replacing $\log_2(n+1)$ by a slower growing function.

Theorem 4.6. *Let $n = 4^k - 1$, with k a positive integer. For any $m \leq 3k/8$, we have*

$$\underline{\gamma}_m^1(B_1^n, \mathcal{V}_3)_1 \geq \tfrac{1}{2}.$$

We want to carry out an analysis similar to the above for the ℓ_2-Greedy Algorithm (PGA) and the dictionary $\mathcal{V}_3$.

Theorem 4.7. *Let $k := [\log_2 n]$. Then,*

$$\bar{\gamma}_m^2(B_2^n, \mathcal{V}_3)_2 \leq \left(1 - \frac{1}{k+1}\right)^{m/2}, \qquad m = 1, 2, \dots. \tag{4.12}$$

The following theorem shows that in a certain sense the estimates of Theorem 4.7 cannot be improved:

Theorem 4.8. *Let $n = 2^k$ for some positive integer k. For any $m \le k/2$, we have*

$$\underline{\gamma}_m^2(B_2^n, \mathcal{V}_3)_2 \ge \tfrac{1}{2}.$$

Theorem 4.3 gives the upper estimate for $\sigma_m(B_2^n, \mathcal{D})_2$. In the particular case $\#\mathcal{D} = C^n$, $C > 3$, this theorem guarantees the existence of $\mathcal{D}$ such that

$$\sigma_m(B_2^n, \mathcal{D})_2 \le \left(\frac{2}{C-1}\right)^m. \tag{4.13}$$

It was proposed in [**11**] to pick adaptively from a set of bases a single basis that is the "best basis." In this setting it is interesting to compare the estimate (4.13) with the following lower estimate in the problem of selection of optimal basis (see [**42**]). For given K there exists a positive $C(K)$ such that for any set of $S \le K^n$ bases $\mathcal{B}^j$, $j = 1, \dots, S$, in $\mathbb{R}^n$ we have, for each $m < n/2$,

$$\sup_{f \in B_2^n} \inf_j \sigma_m(f, \mathcal{B}^j)_2 \ge C(K).$$

Open Problems

4.1. Characterize Banach spaces X such that the X-Greedy Algorithm converges for all dictionaries $\mathcal{D}$ and each element f.

4.2. Characterize Banach spaces X such that the DGA converges for all dictionaries $\mathcal{D}$ and each element f.

4.3. (Conjecture). Prove that the DGA converges for all dictionaries $\mathcal{D}$ and each element $f \in X$ in uniformly smooth Banach spaces X with modulus of smoothness of fixed power type q, $1 < q \le 2$ ($\rho(u) \le \gamma u^q$).

4.4. Find the necessary and sufficient conditions on a weakness sequence τ to guarantee convergence of the WDGA in uniformly smooth Banach spaces X with modulus of smoothness of fixed power type q, $1 < q \le 2$ ($\rho(u) \le \gamma u^q$) for all dictionaries $\mathcal{D}$ and each element $f \in X$.

4.5. Find the correct (in both parameters n and m) order of decay of the quantities

$$\bar{\gamma}_m^p(B_p^n, \mathcal{V}_3)_p, \qquad \underline{\gamma}_m^p(B_p^n, \mathcal{V}_3)_p, \qquad p = 1, 2.$$

5. Bilinear Approximation

In this section we discuss one particular case of a dictionary. Denote by Π the system of functions of the form $u(x_1)v(x_2)$. It is clear that $\mathcal{T}^2 \subset \Pi$. It is also clear that Π is a very redundant system. We have already mentioned some results for this system in the Introduction and in Section 3. All of those results concerned approximation in Hilbert space $L_2([0, 1]^2)$ and it was convenient for us to normalize

elements of Π in L_2 (what made the system Π a dictionary in $L_2([0,1]^2)$). In this section we consider approximation by Π in all L_p, $1 \le p \le \infty$, spaces. In order to make the system Π a dictionary in L_p we need to normalize it in L_p. We will denote the normalized in the L_p system Π by Π_p. Most results of this section give estimates for best m-term approximation. These results do not depend on the normalization of Π and for convenience in such a case we will use notation Π without an index p. In this section we concentrate only on the approximation of bivariate functions from standard function classes. We note that the bilinear approximation is now a well-established area and many estimates are proved in a general setting: f is a function of $2d$ variables $x = (x_1, \dots, x_d)$, $y = (y_1, \dots, y_d)$; Π is replaced by $\Pi^d := \{u(x)v(y)\}$; L_p is replaced by L_{p_1,p_2}, where

$$\|f\|_{p_1,p_2} := \|\|f(\cdot, y)\|_{p_1}\|_{p_2}.$$

The key role in bilinear approximation is played by the Schmidt formula (see Section 3)

$$\sigma_m(f, \Pi)_2 = \left(\sum_{n=m+1}^{\infty} s_n(J_f)^2 \right)^{1/2}. \tag{5.1}$$

This formula implies in particular, for $a > 0$,

$$\sigma_m(f, \Pi)_2 \ll m^{-a} \quad \Leftrightarrow \quad s_n(J_f) \ll m^{-a-1/2}.$$

The following classes are well-known and important in studying integral operators. We say that J_f belongs to the Schatten v-class S_v if

$$\sum_n s_n(J_f)^v < \infty.$$

The Schmidt formula (5.1) allows us to prove the following result:

Theorem 5.1. *For any $v < 2$ we have*

$$J_f \in S_v \quad \Leftrightarrow \quad \sum_m (\sigma_m(f, \Pi)_2 m^{-1/2})^v < \infty.$$

This theorem is an analog of the following theorem (see [**18**]) for an orthonormal basis $\mathcal{B}$ for a Hilbert space H.

Theorem 5.2. *For any $\beta < 2$ and any orthonormal basis $\mathcal{B}$ we have*

$$f \in \mathcal{A}_\beta(\mathcal{B}) \quad \Leftrightarrow \quad \sum_m (\sigma_m(f, \mathcal{B}) m^{-1/2})^\beta < \infty.$$

Theorem 5.2 is a generalization of Stechkin's result [**71**] that corresponds to $\beta = 1$ in Theorem 5.2. Let us present some general results for approximation

in Banach spaces. As a corollary of these general results we will obtain error estimates for approximation by Π in L_p. We recall that $A_1(\mathcal{D})$ is a convex hull of $\mathcal{D}^{\pm}$. Similarly, to the definition of $\mathcal{A}_\beta(\mathcal{D})$ in Subsection 3.3, we define $\mathcal{A}_\beta(\mathcal{D})$ in a Banach space X with a dictionary $\mathcal{D}$. It is easy to derive (see an idea in [**18**, Theorem 3.3]) from Theorem 4.1 the following statement:

Theorem 5.3. *Let X be a uniformly smooth Banach space with the modulus of smoothness $\rho(u) \le \gamma u^q$, $1 < q \le 2$. Then, for any $f \in \mathcal{A}_\beta(\mathcal{D})$, $0 < \beta \le 1$, we have*

$$\sigma_m(f, \mathcal{D})_X \le C(X, \mathcal{D}) m^{1/q-1/\beta} |f|_{\mathcal{A}_\beta(\mathcal{D})}.$$

In a particular case, $X = L_p$, $1 < p < \infty$, $\mathcal{D} = \Pi_p$, Theorem 5.3 gives the estimate

$$\sigma_m(f, \Pi)_p \le C(p) m^{\max(1/p,1/2)-1/\beta} |f|_{\mathcal{A}_\beta(\Pi_p)}. \tag{5.2}$$

This inequality gives the error estimate of best m-term approximation in terms of $|f|_{\mathcal{A}_\beta(\Pi_p)}$ which is not well-studied for $p \ne 2$. We will present some results on estimates for $\sigma_m(f, \Pi)_p$ in terms of standard periodic Hölder–Nikol'skii classes $NH^{R_1,R_2}_{q_1,q_2}$ of functions of two variables. We define these classes in the following way. First of all, we define the vector L_{q_1,q_2}-norm as

$$\|f(x_1, x_2)\|_{q_1,q_2} := \|\|f(\cdot, x_2)\|_{q_1}\|_{q_2}.$$

The class $NH^{R_1,R_2}_{q_1,q_2}$ is the set of periodic functions $f \in L_{q_1,q_2}([0, 2\pi]^2)$ such that, for each $l_j = [R_j] + 1$, $j = 1, 2$, the following relations hold

$$\|f\|_{q_1,q_2} \le 1, \qquad \|\Delta_t^{l_j,j} f\|_{q_1,q_2} \le |t|^{R_j}, \qquad j = 1, 2,$$

where $\Delta_t^{l,j}$ is the lth difference with step t in the variable x_j. In the case $d = 1$, NH^R_q coincides with the standard Hölder class H^R_q.

The results from Section 2 (see (2.6)–(2.8)) indicate that the bilinear approximation of $f(x - y)$ is closely connected with the Kolmogorov widths $d_m(F[f], L_p)$ and the best m-term approximation of f with regard to the trigonometric system. If $f \in H^R_q$, then $f(x - y) \in NH^{(R,R)}_q$. We get from [**82**] that

$$\sigma_m(NH^{(R,R)}_q, \Pi)_p \ll m^{-R+(1/q-\max(1/p,1/2))_+} \tag{5.3}$$

for $1 \le q \le p \le \infty$ with $R > R(q, p)$, $R(q, p) = 2(1/q - 1/p)$ for $1 \le q \le p \le 2$, and $R(q, p) = 1/q + \max(1/q, \frac{1}{2})$ for $p > 2$. Comparing (5.3) with (2.7) we see that the upper estimates for the wider class $NH^{(R,R)}_q$ have the same order as for the class $\{f(x - y),\ f \in H^R_q\}$. Further results for anisotropic classes $NH^{(R_1,R_2)}_{q_1,q_2}$ and their $2d$-dimensional generalizations can be found in [**82**].

In the case $1 \le p \le q \le \infty$ we have

$$\sigma_m(NH^{(R,R)}_q, \Pi)_p \asymp m^{-R}. \tag{5.4}$$

A nontrivial estimate in (5.4) is the lower estimate for $p = 1, q = \infty$. This estimate and the generalizations of (5.4) are obtained in [**84**]. Let us now present the results in approximation in the L_2-norm for the general classes $NH^{(R_1,R_2)}_{q_1,q_2}$ (see [**82**] and [**85**]). We note that the study of $\sigma_m(NH^{(R_1,R_2)}_{q_1,q_2}, \Pi)_{p_1,p_2}$ is not complete. One of the open problems in this area is given in Open Problem 5.7. Known results can be found in [**82**] and [**85**]. Denote $\eta_i := (1/q_i - \frac{1}{2})_+$, $i = 1, 2$.

Theorem 5.4. *Let $R_1 \le R_2$ and $R_1 > \eta_1$, $R_2 > \eta_2(1 - \eta_1/R_1)^{-1}$. Then*

$$\sigma_m(NH^{(R_1,R_2)}_{q_1,q_2}, \Pi)_2 \asymp m^{-R_2(1-\eta_1/R_1)}, \qquad 1 \le q_1, q_2 \le \infty.$$

Theorem 5.5. *Let R_1, R_2 be as in Theorem* 5.4. *Then*

$$\sup_{f \in NH^{(R_1,R_2)}_{q_1,q_2}} s_m(J_f) \asymp m^{-R_2(1-\eta_1/R_1)-1/2}, \qquad 1 \le q_1, q_2 \le \infty.$$

Theorem 5.6. *Let $R_1 \ge R_2$, $R_2 > \eta_2$, $R_1 > \eta_1(1 - \eta_2/R_2)^{-1}$. Then*

$$\sigma_m(NH^{(R_1,R_2)}_{q_1,q_2}, \Pi)_2 \asymp m^{-R_1(1-\eta_2/R_2)+\eta_1-\eta_2}, \qquad 1 \le q_1, q_2 \le \infty.$$

Theorem 5.7. *Let R_1, R_2 be as in Theorem* 5.6. *Then*

$$\sup_{f \in NH^{(R_1,R_2)}_{q_1,q_2}} s_m(J_f) \asymp m^{-R_1(1-\eta_2/R_2)+\eta_1-\eta_2-1/2}, \qquad 1 \le q_1, q_2 \le \infty.$$

We now give some historical remarks on estimating the eigenvalues and singular numbers of integral operators. We begin with the following theorem that is a corollary to the Weyl Majorant Theorem (see [**32**, p. 41]).

Theorem 5.8. *Let A be a compact* (*completely continuous*) *operator in a Hilbert space H. Suppose that*

$$s_n(A) \ll n^{-r}, \qquad r > 0.$$

Then

$$|\lambda_n(A)| \ll n^{-r}.$$

Fredholm [**30**] proved that if the kernel $f(x, y)$ is a continuous function and satisfies the condition

$$\sup_{x,y} |f(x, y + t) - f(x, y)| \le C|t|^{\alpha}, \qquad 0 < \alpha \le 1,$$

then, for an arbitrary $\rho > 2/(2\alpha + 1)$, the series

$$\sum_{j=1}^{\infty} |\lambda_j(J_f)|^{\rho} < \infty$$

converges.

Starting with that article, smoothness conditions with respect to one variable were imposed on the kernel. Weyl [**102**] proved the estimate

$$\lambda_n(J_f) = o(n^{-r-1/2})$$

under the condition that the kernel $f(x, y)$ is symmetric and continuous and that $\partial^r f/\partial x^r$ is continuous. Let us introduce some more notation. Define $NH^{(R,0)}_{q_1,q_2}$ as follows: $f(x, y)$ belongs to this class if for all $y \in \mathbb{T}$ the function $f(\cdot, y)$ of x belongs to the class $H^R_{q_1}B(y)$, and $B(y)$ is such that $\|B(y)\|_{q_2} \le 1$. We use here the following notation. For a function class F and a number $B > 0$ we define $FB := \{f : f/B \in F\}$.

Hille and Tamarkin [**33**] achieved significant progress. They proved, in particular, that, for $1 < q \le 2$ and $R \ge 1$,

$$\sup_{f \in NH^{(R,0)}_{q,q'}} |\lambda_n(J_f)| \ll n^{-R-1+1/q}(\log n)^R, \qquad q' = q/(q-1),$$

and they conjectured that the extra logarithmic factor can be removed or even replaced by a logarithmic factor with a negative power.

The next important step was taken by Smithies [**70**]. He proved the estimate

$$\sup_{f \in NH^{(R,0)}_{q,2}} s_n(J_f) \ll n^{-R-1+1/q}, \qquad 1 < q \le 2, \quad R > 1/q - \tfrac{1}{2}. \tag{5.5}$$

Of later results we mention those of Gel'fond and M. G. Krein (see [**32**, Ch. III, S9.4]), Birman and Solomyak [**6**], and Cochran [**9**].

We proved in [**82**] the following estimate

$$\sigma_m(NH^{(R,0)}_{q_1,q_2}, \Pi)_{p_1,p_2} \asymp m^{-R+(1/q_1-\max(1/2,1/p_1)},$$

for $1 \le q_1 \le p_1 \le \infty$, $1 \le q_2 = p_2 \le \infty$, and $R > r(q_1, p_1)$. We denote here $r(q, p) := (1/q - 1/p)_+$ for $1 \le q \le p \le 2$ or $1 \le p \le q \le \infty$ and $r(q, p) := \max(\frac{1}{2}, 1/q)$ otherwise. This inequality implies in particular that (5.5) also holds for $q = 1$.

We now discuss an application of bilinear approximation to the theory of widths. As we know, the starting point of this theory is a function class, say, the function class W^r_q. This function class can be associated with one function, the Bernoulli kernel $F_r(x - y)$, with

$$F_r(t) := 2\sum_{k=1}^{\infty} k^{-r}\cos(kt - r\pi/2).$$

We have

$$W^r_q = \left\{ f : f(x) = \hat{f}(0) + (2\pi)^{-1}\int_0^{2\pi} F_r(x-y)\varphi(y)\,dy,\ \|\varphi\|_q \le 1 \right\}.$$

In the development of approximation by trigonometric polynomials it was understood that the rate of decay of $E_n(f)$ of individual functions, say $E_n(F_r)$, is governed by smoothness properties of the function. It turns out that we have similar phenomena on a much more general level.

For a function $g \in L_1(\mathbb{T}^2)$ define a function class

$$W_q^g := \left\{ f : f(x) = (2\pi)^{-1} \int_0^{2\pi} g(x, y)\varphi(y)\, dy, \|\varphi\|_q \le 1 \right\}.$$

We proved in [**80**] that $F_\rho(x-y)$ is a typical representative of the following class of functions. Denote by $MH_1^{r_1,r_2}B$ the class of functions $g(x, y)$ such that, $\|g\|_1 < \infty$,

$$\int_0^{2\pi} g(x, y)\, dx = \int_0^{2\pi} g(x, y)\, dy = 0$$

(this condition is imposed only for convenience), and

$$\|\Delta^l_{t_1,t_2} g(x, y)\|_1 \le B|t_1|^{r_1}|t_2|^{r_2}, \qquad r_1, r_2 > 0, \quad l := \max([r_1], [r_2]) + 1,$$

where $\Delta^l_{t_1,t_2}$ denotes the operator of the mixed difference of order l in each variable with step t_1 in x and step t_2 in y. We remark that the function $F_\rho(x - y)$ belongs to $MH_1^{r_1,r_2}B$ for any r_1, r_2 such that $r_1 + r_2 = \rho$. We proved in [**80**] the following statement:

Theorem 5.9. *For all* $1 \le q, p \le \infty$, *we have*

$$\sup_{g \in MH_1^{r_1,r_2}} d_m(W_q^g, L_p) \asymp d_m(W_q^{r_1+r_2}, L_p)$$

for $r_1 > 1$, $r_2 > 1 + \max(1/q, \frac{1}{2})$ *for* $2 \le q < p \le \infty$ *or* $1 \le q < 2 < p \le \infty$ *and* $r_2 > 1$ *otherwise.*

Open Problems

5.1. Find the necessary and sufficient conditions on a weakness sequence τ to guarantee convergence of the WGA with regard to Π_2 for each $f \in L_2$.

5.2. Does the L_p-Greedy Algorithm with regard to Π_p converge for each $f \in L_p$, $1 < p < \infty$?

5.3. Does the DGA with regard to Π_p converge for each $f \in L_p$, $1 < p < \infty$?

5.4. If the answer to Problem 5.3 is "yes" then find the necessary and sufficient conditions on a weakness sequence τ to guarantee convergence of the WDGA with regard to Π_p for each $f \in L_p$.

5.5. Find the necessary and sufficient conditions on a weakness sequence τ to guarantee convergence of the WCGA with regard to Π_p for each $f \in L_p$.

5.6. Let R_N be the Rudin–Shapiro polynomials (see Section 8). Prove that

$$\sigma_m(R_N(x - y), \Pi)_1 \gg N^{1/2}.$$

5.7. Find the order of the sequence

$$\sigma_m(NH_1^{(R_1,R_2)}, \Pi)_{p_1,\infty}, \quad m = 1, 2, \dots, \tag{5.6}$$

in the case $R_1 < R_2$, $2 < p_1 \le \infty$.

Comment. In the case $R_1 \ge R_2$ the order of (5.6) is known (see [**82**]).

5.8. Study the efficiency of the PGA (L_2-Greedy Algorithm) with regard to Π_2 for the approximation of function classes $NH_{q_1,q_2}^{(R_1,R_2)}$ in the L_{p_1,p_2}-norm.

6. Ridge Approximation

This section, similar to Section 5, is devoted to the approximation of functions of several variables. The results discussed here may be seen as one more (in addition to Section 5) example in the development of the following general approach in the multivariate approximation: approximate functions of several variables by univariate functions. This idea is interesting from a theoretical point of view and also looks reasonable from a computational point of view. There is a number of different realizations of this approach in approximation theory. We mention some of them for illustration. We begin with the simplest one. S. N. Bernstein (see [**5**]) suggested studying the following type of approximation of a continuous periodic function $f(x, y)$ of two variables

$$E_{n,\infty}(f) := \inf_{\{c_k(y)\}} \left\| f(x, y) - \sum_{|k|\le n} c_k(y) e^{ikx} \right\| \tag{6.1}$$

in the uniform norm $\|\cdot\| := \|\cdot\|_\infty$. The approximant in (6.1) is a linear combination of the products of univariate functions. The Bernstein setting of problem (6.1) is a variant of the classical problem of bilinear approximation which was discussed in Section 5. The important feature of the problem of bilinear approximation is that the approximating system $\{u(x)v(y)\}_{u,v\in L_2}$ is highly redundant. However, as we have seen in Section 5, the redundancy did not hinder the development of the nice theory to solve the problem of best bilinear approximation in the L_2-norm. What really allowed us to develop that theory is the structure of the system. In this section we discuss approximation by a redundant system with a quite different structure. We approximate by linear combinations of ridge functions, i.e., functions $G(x)$, $x \in \mathbb{R}^2$, which can be represented in the form

$$G(x) = g((x, e)), \tag{6.2}$$

where g is a univariate function and its argument (x, e) is the scalar product of x and a unit vector $e \in \mathbb{R}^2$. We denote the set of functions of the form (6.2) by $\mathcal{R}$ and call it the system of ridge functions. The above-mentioned approximation (approximation by ridge functions) also uses univariate functions, and the system $\mathcal{R}$ of all ridge functions is highly redundant. Unlike the bilinear approximation

problem we do not have a theory which provides (describes) the solution to the problem of best ridge approximation. In this section we confine ourselves to the case of the functions of two variables. We note that approximation by ridge functions received much attention recently for the following two reasons. The first is that a ridge function can be interpreted as a plane wave. This means that the problem of ridge approximation can be seen as a problem of representation of a general wave by plane waves. The second reason is that ridge approximation proved to be useful in neural networks approximation (see [**22**]).

Let $D := \{(x_1, x_2) : x_1^2 + x_2^2 \le 1\}$ be the unit disk and let $L_p(D)$, $1 \le p < \infty$, denote the Banach space with the norm

$$\|f\|_p := \|f\|_{L_p(D)} := \left(\frac{1}{\pi}\int_D |f(x)|^p\,dx\right)^{1/p}.$$

From this point on we denote by $\mathcal{R}_p$ the dictionary for $L_p(D)$ which consists of elements of the system $\mathcal{R}$ normalized in $L_p(D)$. Similar to the bilinear approximation we use the notation $\mathcal{R}$ instead of $\mathcal{R}_p$ when we talk about best m-term approximations. There are some general results on the approximation by linear combinations of elements of a redundant system in a Banach space (see Theorem 5.3). These results are expressed in terms of the $\mathcal{A}_\beta(\mathcal{D})$-quasi-norm determined by a dictionary $\mathcal{D}$. In a particular case, $X = L_p(D)$, $1 < p < \infty$, $\mathcal{D} = \mathcal{R}_p$, Theorem 5.3 gives the estimate

$$\sigma_m(f, \mathcal{R})_p \le C(p)m^{\max(1/p,1/2)-1/\beta}|f|_{\mathcal{A}_\beta(\mathcal{R}_p)}. \tag{6.3}$$

This inequality gives the error estimate of best m-term approximation in terms of $|f|_{\mathcal{A}_\beta(\mathcal{R}_p)}$ which is not well-studied. In order to use this general result we need to verify that a given function f can be approximated by functions which have a special representation (see the definition of $\mathcal{A}_\beta(\mathcal{D})$), that in turn could be a nontrivial problem. We will present some results on estimates for $\sigma_m(f, \mathcal{R})_p$ in terms of the standard classes of functions. In this section we deal with the function class which is defined in a way standard for constructive approximation. We define the class of functions $H_p^r(D)$ using the classical means of approximation, namely, algebraic polynomials. Let $\mathcal{P}(n, 2)$ denote the set of algebraic polynomials

$$\sum_{k+l\le n-1} c_{k,l}x_1^k x_2^l$$

of total degree $n-1$. Denote by $H_p^r(D)$, $r > 0$, the set of all functions $f \in L_p(D)$ which can be represented in the form

$$f = \sum_{n=1}^{\infty} p_n, \qquad p_n \in \mathcal{P}(2^n, 2), \quad n = 1, 2, \ldots,$$

with p_n satisfying the inequalities

$$\|p_n\|_p \le 2^{-rn}.$$

The following result (see [**52**]) gives the upper estimates for $\sigma_m(H_p^r(D), \mathcal{R})_p$ automatically.

Theorem 6.1. *For any algebraic polynomial $p \in \mathcal{P}(N, 2)$ there exist N univariate polynomials g^j, $j = 0, \dots, N-1$, of degree $N-1$ with the following property*

$$p(x) = \sum_{j=0}^{N-1} g^j((x, e_j^N)), \tag{6.4}$$

where $e_j^N := (\cos j\pi/N, \sin j\pi/N)$.

This gives the estimate

$$\sigma_m(H_p^r(D), \mathcal{R})_p \le C(r)m^{-r}. \tag{6.5}$$

It turns out that, in the case $p = 2$, the estimate (6.5) is sharp:

$$\sigma_m(H_2^r(D), \mathcal{R})_2 \ge C(r)m^{-r}. \tag{6.6}$$

The first result in this direction, which is a weaker version of (6.6), was obtained in [**87**]. Estimate (6.6) was proved in [**55**]. Estimate (6.6) also follows from the relation

$$\sigma_m(f, \mathcal{R})_2 \ge C \inf_{p \in \mathcal{P}(3m,2)} \|f - p\|_2 \tag{6.7}$$

established in [**62**] for radial functions f, $f(x_1, x_2) = h((x_1^2 + x_2^2)^{1/2})$, h is a univariate function.

We proved recently (see [**56**]) that estimate (6.5) in the case $p = 2$ can be realized by the PGA

$$\sup_{f \in H_2^r} \|f - G_m(f, \mathcal{R}_2)\|_2 \le C(r)m^{-r}. \tag{6.8}$$

Let us make some comments on (6.8). First of all this estimate shows that the PGA with regard to $\mathcal{R}_2$ is not saturated. Moreover, combining (6.8) with (6.7), we see that for radial functions f such that

$$\sigma_m(f, \mathcal{R})_2 \le C(r)m^{-r} \tag{6.9}$$

we have

$$\|f - G_m(f, \mathcal{R}_2)\|_2 \le C(r)m^{-r}.$$

This is a weaker analog of the r-greedy property for $\mathcal{R}_2$.

Open Problems

6.1. Find the necessary and sufficient conditions on a weakness sequence τ to guarantee convergence of the WGA with regard to $\mathcal{R}_2$ for each $f \in L_2$.

6.2. Does the L_p-Greedy Algorithm with regard to $\mathcal{R}_p$ converge for each $f \in L_p$, $1 < p < \infty$?

6.3. Does the DGA with regard to $\mathcal{R}_p$ converge for each $f \in L_p$, $1 < p < \infty$?

6.4. If the answer to Problem 6.3 is "yes" then find the necessary and sufficient conditions on a weakness sequence τ to guarantee convergence of the WDGA with regard to $\mathcal{R}_p$ for each $f \in L_p$.

6.5. Find the necessary and sufficient conditions on a weakness sequence τ to guarantee convergence of the WCGA with regard to $\mathcal{R}_p$ for each $f \in L_p$.

6.6. Find the order of the quantity

$$\sup_{f \in A_1(\mathcal{R}_2)} \|f - G_m(f, \mathcal{R}_2)\|_{L_2(\mathcal{D})}.$$

6.7. Could estimate (6.5) for $1 < p < \infty$ be realized by WCGA with $\tau = \{t\}$, $0 < t \le 1$?

7. Greedy Approximation with Regard to Bases

7.1. *Greedy Bases*

We will study the algorithms $G_m(f, \Psi, \rho)$ defined in the Introduction. In order to understand the efficiency of this algorithm we compare its accuracy with the best possible $\sigma_m(f, \Psi)$ when an approximant is a linear combination of m terms from Ψ. The best we can achieve with the algorithm G_m is

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

or a little weaker

$$\|f - G_m(f, \Psi, \rho)\| \le G\sigma_m(f, \Psi) \tag{7.1}$$

for all elements $f \in X$ with a constant $G = C(X, \Psi)$ independent of f and m.

Definition 7.1. We call a basis Ψ a greedy basis if for every $f \in X$ there exists a permutation $\rho \in D(f)$ such that (7.1) holds.

The following proposition has been proved in [**44**]:

Proposition 7.1. *If Ψ is a greedy basis then* (7.1) *holds for any permutation $\rho \in D(f)$.*

We will discuss the two most interesting cases of bases Ψ: the Haar basis $\mathcal{H}$ as a representative of wavelet-type bases and the trigonometric system $\mathcal{T}$ as a representative of uniformly bounded orthonormal bases.

Denote by $\mathcal{H}_p := \{H_k^p\}_{k=1}^\infty$ the Haar basis on $[0, 1)$ normalized in $L_p(0, 1)$: $H_1^p = 1$ on $[0, 1)$ and for $k = 2^n + l$, $l = 1, 2, \dots, 2^n$, $n = 0, 1, \dots$,

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

Denote by $\mathcal{T} := \{e^{ikx}\}_{k\in\mathbb{Z}}$ the univariate trigonometric system in the complex form and denote by $\mathcal{T}^d := \mathcal{T} \times \cdots \times \mathcal{T}$ the multivariate trigonometric system.

The following theorem (see [**88**]) establishes the existence of greedy bases for $L_p(0, 1)$, $1 < p < \infty$.

Theorem 7.1. *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*

$$\|f - G_m(f, \Psi, \rho)\|_{L_p} \le C(p, \Psi)\sigma_m(f, \Psi)_{L_p}$$

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

In this theorem we use the following definition of the L_p-equivalence. We say that $\Psi = \{\psi_k\}_{k=1}^\infty$ is L_p-equivalent to $\mathcal{H}_p = \{H_k^p\}_{k=1}^\infty$ if for any finite set Λ and any coefficients c_k, $k \in \Lambda$, we have

$$C_1(p, \Psi) \left\| \sum_{k\in\Lambda} c_k H_k^p \right\|_{L_p} \le \left\| \sum_{k\in\Lambda} c_k \psi_k \right\|_{L_p} \le C_2(p, \Psi) \left\| \sum_{k\in\Lambda} c_k H_k^p \right\|_{L_p}$$

with two positive constants $C_1(p, \Psi)$, $C_2(p, \Psi)$ which may depend on p and Ψ. For sufficient conditions on Ψ to be L_p-equivalent to $\mathcal{H}_p$, see [**29**] and [**21**].

Thus 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 in the case of a Hilbert space each orthonormal basis is a greedy basis with a constant $G = 1$ (see (7.1)).

We now give the definitions of unconditional and democratic bases.

Definition 7.2. 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 7.3. We say that a basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$ is a democratic basis if, for any two finite sets of indices P and Q with the same cardinality $\#P = \#Q$, we have

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

with a constant $D := D(X, \Psi)$ independent of P and Q.

We proved in [**44**] the following theorem:

Theorem 7.2. *A basis is greedy if and only if it is unconditional and democratic.*

The property of a basis to be a greedy basis for X is a very strong property. In such a case the greedy approximant $G_m(f, \Psi)$ realizes near best m-term approximation for any individual function f. There are different ways to weaken the greedy property of a basis. For instance, we can replace (7.1), that holds for individual functions, by its analog for some function classes. The following definition elaborates the above idea:

Definition 7.4. We call a basis Ψ an r-greedy basis for a Banach space X if for each $f \in X$ such that

$$\sigma_m(f, \Psi)_X \le m^{-r}, \qquad m = 1, 2, \dots,$$

we have, for every $\rho \in D(f)$,

$$\|f - G_m(f, \Psi, \rho)\| \le C(r, \Psi) m^{-r}, \qquad m = 1, 2, \dots.$$

It is clear that a greedy basis is r-greedy for all r. We now construct an example showing that the r-greedy property is weaker than the greedy property.

Example 7.1. There exist a Banach space X and a basis Ψ such that Ψ is an r-greedy basis for X for any $r > 0$ and Ψ is not an unconditional basis.

Proof. We use the construction from [**44**]. Let X be the set of all real sequences $x = (x_1, x_2, \dots) \in l_2$ such that

$$\|x\|' = \sup_{N \in \mathbf{N}} \left| \sum_{n=1}^{N} x_n / \sqrt{n} \right|$$

is finite. Clearly, X, equipped with the norm

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

is a Banach space. Let $\psi_k \in X$, $k = 1, 2, \ldots$, be defined as

$$(\psi_k)_n = \begin{cases} 1, & n = k, \\ 0, & n \neq k. \end{cases}$$

We take any $r > 0$ and prove that Ψ is the r-greedy basis for X. Indeed, the assumption $\sigma_m(f, \Psi)_X \leq m^{-r}$ implies $\sigma_m(f, \Psi)_{l_2} \leq m^{-r}$ and, therefore,

$$\|f - G_m(f, \Psi)\|_{l_2} \leq m^{-r}.$$

Let us prove a similar estimate for $\|\cdot\|'$. Let

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

Denote $Q_m(N) := [1, N] \backslash \Lambda_m$. Then

$$\|f - G_m(f, \Psi)\|' = \sup_N \left| \sum_{k \in Q_m(N)} c_k(f) k^{-1/2} \right| \leq \sum_{k=1}^{\infty} k^{-1/2}(m + k)^{-r-1/2} \ll m^{-r}.$$

This proves that Ψ is a r-greedy basis for X. It is proved in [**44**] that Ψ is not unconditional.

7.2. *The Trigonometric System*

Let us consider nonlinear approximation with regard to the trigonometric system $\mathcal{T}^d$. The existence of best m-term trigonometric approximation was proved in [**3**] (see also [**91**]). The method $G_m(f) := G_m(f, \mathcal{T}^d)$ has one more advantage over the traditional approximation by trigonometric polynomials in the case of the approximation of functions of several variables. In this case ($d > 1$) there is no natural order of trigonometric system and the use of G_m allows us to avoid the problem of finding natural subspaces of trigonometric polynomials for approximation purposes. We proved in [**91**] the following inequality:

Theorem 7.3. *For each $f \in L_p(\mathbb{T}^d)$ we have*

$$\|f - G_m(f)\|_p \leq (1 + 3m^{h(p)})\sigma_m(f)_p, \qquad 1 \leq p \leq \infty,$$

where $h(p) := |\frac{1}{2} - 1/p|$.

Remark 7.1. For all $1 \leq p \leq \infty$,

$$\|G_m(f)\|_p \leq m^{h(p)}\|f\|_p.$$

Remark 7.2. 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

$$\|G_m(f)\|_p \ge Cm^{h(p)}\|f\|_p. \tag{7.2}$$

The above results show that the trigonometric system is not a greedy basis for L_p, $p \ne 2$. This leads to a natural attempt to consider some other algorithms that may have some advantages over TGA in the case of $\mathcal{T}$. We discuss here the performance of WCGA (see Section 4) with regard to $\mathcal{T}$.

Let us compare the rate of approximation of TGA and WCGA for the class $A := A(\mathcal{RT})$ where $\mathcal{RT}$ denotes the real trigonometric system $\frac{1}{2}, \sin x, \cos x, \dots$. We need to switch to this system from the complex trigonometric system because the algorithm WCGA is defined for the real Banach space. We note that the system $\mathcal{RT}$ is not normalized in L_p but quasi-normalized: $C_1 \le \|t\|_p \le C_2$ for any $t \in \mathcal{RT}$ with absolute constants C_1, C_2, $1 \le p \le \infty$. This is sufficient for the application of the general methods developed in Section 4. For a sequence $\tau := \{t_k\}$ with $t_k = t$, $k = 1, 2, \dots$, we replace τ by t in the notation. Theorem 4.1 and (4.6) imply the following result:

Theorem 7.4. *Let $0 < t \le 1$. For $f \in A$ we have*

$$\|f - G_m^{c,t}(f, \mathcal{RT})\|_p \le C(p,t)m^{-1/2}, \qquad 2 \le p < \infty. \tag{7.3}$$

This estimate and Theorem 7.3 imply that for $f \in A$ we have

$$\|f - G_m(f, \mathcal{RT})\|_p \le C(p,t)m^{-1/p}, \qquad 2 \le p < \infty, \tag{7.4}$$

which is weaker than (7.3). It is proved in [**23**] that (7.4) cannot be improved. Thus the WCGA works better than the TGA for the class A. We note that the restriction $p < \infty$ in (7.3) is important. We now give a lower estimate for m-term approximation in L_∞.

Proposition 7.2. *For a given m define*

$$f := \sum_{k=0}^{2m} \cos 3^k x.$$

Then we have

$$\sigma_m(f, \mathcal{T})_\infty \ge m/4.$$

Proof. Consider the Riesz product

$$\Phi_0(x) := \prod_{j \in [0,2m]} (1 + \cos 3^j x) - 1.$$

This function has nonzero Fourier coefficients only with frequencies of the form

$$k(s) = \sum_{j=0}^{j(s)} s_j 3^j, \qquad s = (s_0, \dots, s_{2m}),$$

with $0 \le j(s) \le 2m$, $s_j = -1, 0, 1$ for $j < j(s)$, $s_{j(s)} = 1$, and $s_j = 0$ for $j(s) < j \le 2m$. It is clear that $k(s)$ is uniquely defined by s. Take any polynomial of the form

$$t(x) = \sum_{k \in \Lambda} a_k \cos kx, \qquad \#\Lambda = m.$$

Then for each $k \in \Lambda$ we look for an s such that $k = k(s)$. If we do not find such an s we have

$$\langle \cos kx, \Phi_0 \rangle = 0.$$

For those s that were found to satisfy $k(s) = k, k \in \Lambda$, we form a set J consisting of all $j(s)$ and define the new Riesz product

$$\Phi := \prod_{j \in [0,2m] \setminus J} (1 + \cos 3^j x) - 1.$$

Then we have

$$\langle t, \Phi \rangle = 0$$

and

$$m \le \langle f - t, \Phi \rangle \le \|f - t\|_\infty \|\Phi\|_1 \le 4\|f - t\|_\infty.$$

This implies

$$\sigma_m(f, \mathcal{T})_\infty \ge m/4.$$

7.3. *Greedy Bases. Direct and Inverse Theorems*

Theorem 7.1 points out the importance of bases that are L_p-equivalent to the Haar basis. We will now discuss necessary and sufficient conditions for f to have a prescribed decay of $\{\sigma_m(f, \Psi)_p\}$ under the assumption that Ψ is L_p-equivalent to the Haar basis $\mathcal{H}_p$, $1 < p < \infty$. We will express these conditions in terms of coefficients $\{f_n\}$ of the expansion

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

The following lemma from [**88**] plays the key role in this consideration.

Lemma 7.1. *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*

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

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 monotonically decreasing-to-zero sequence $\mathcal{E} = \{\varepsilon_k\}_{k=0}^{\infty}$ of positive numbers (we write $\mathcal{E} \in MDP$) we define inductively a sequence $\{N_s\}_{s=0}^{\infty}$ of nonnegative integers: $N_0 = 0$; N_s is the smallest satisfying

$$\varepsilon_{N_s} < 2^{-s}, \qquad n_s := \max(N_{s+1} - N_s, 1). \tag{7.5}$$

We are going to consider the following examples of sequences:

Example 7.2. Take $\varepsilon_0 = 1$ and $\varepsilon_k = k^{-r}$, $r > 0$, $k = 1, 2, \dots$. Then

$$N_s \asymp 2^{s/r} \qquad \text{and} \qquad n_s \asymp 2^{s/r}.$$

Example 7.3. Fix $0 < b < 1$ and take $\varepsilon_k = 2^{-k^b}$, $k = 0, 1, 2, \dots$. Then

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

Let $f \in L_p$. Rearrange the sequence $\|f_n \psi_n\|_p$ in decreasing order

$$\|f_{n_1}\psi_{n_1}\|_p \ge \|f_{n_2}\psi_{n_2}\|_p \ge \cdots$$

and denote

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

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

Lemma 7.2. *For any two positive integers $N < M$ we have*

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

Lemma 7.3. *For any sequence $m_0 < m_1 < m_2 < \cdots$ of nonnegative integers we have*

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

Theorem 7.5. *Assume that a given sequence $\mathcal{E} \in MDP$ satisfies the conditions*

$$\varepsilon_{N_s} \geq C_1 2^{-s}, \qquad n_{s+1} \leq C_2 n_s, \qquad s = 0, 1, 2, \ldots.$$

Then we have the equivalence

$$\sigma_n(f)_p \ll \varepsilon_n \quad \Leftrightarrow \quad a_{N_s}(f, p) \ll 2^{-s} n_s^{-1/p}.$$

Corollary 7.1. *Theorem* 7.5 *applied to Examples* 7.2 *and* 7.3 *gives the following relations*:

$$\sigma_m(f)_p \ll (m+1)^{-r} \quad \Leftrightarrow \quad a_n(f, p) \ll n^{-r-1/p}, \tag{7.6}$$

$$\sigma_m(f)_p \ll 2^{-m^b} \quad \Leftrightarrow \quad a_n(f, p) \ll 2^{-n^b} n^{(1-1/b)/p}. \tag{7.7}$$

Remark 7.3. Making use of Lemmas 7.2 and 7.3 we can prove a version of Corollary 7.1 with the sign $\ll$ replaced by $\asymp$.

Theorem 7.5 and Corollary 7.1 are in the spirit of classical Jackson–Bernstein direct and inverse theorems in linear approximation theory, where conditions on the corresponding sequences of approximating characteristics are imposed in the form

$$E_n(f)_p \ll \varepsilon_n \qquad \text{or} \qquad \|E_n(f)_p/\varepsilon_n\|_{l_\infty} < \infty. \tag{7.8}$$

It is well-known (see [**14**]) that in studying many questions of approximation theory it is convenient to consider, along with restriction (7.8), the following generalization

$$\|E_n(f)_p/\varepsilon_n\|_{l_q} < \infty. \tag{7.9}$$

Lemmas 7.2 and 7.3 are also useful in considering this more general case. For instance, in the particular case of Example 7.2 one gets the following statement:

Theorem 7.6. *Let $1 < p < \infty$ and $0 < q < \infty$. Then for any positive r we have the equivalence relation*

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

Remark 7.4. 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

$$\sum_n a_n(f, p)^\beta = \sum_n \|f_n \psi_n\|_p^\beta < \infty. \tag{7.10}$$

In the case $\Psi = \mathcal{H}_p$ condition (7.10) is equivalent to f being in the Besov space $B_\beta^r(L_\beta)$.

Corollary 7.2. *Theorem* 7.6 *implies the following relation*

$$\sum_m \sigma_m(f, \mathcal{H})_p^\beta m^{r\beta-1} < \infty \quad \Leftrightarrow \quad f \in B_\beta^r(L_\beta),$$

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

The statement similar to Corollary 7.2 for free knots spline approximation was proved by P. Petrushev [**64**]. Corollary 7.2 and further results in this direction can be found in [**16**] and [**20**]. We want to remark here that conditions in terms of $a_n(f, p)$ are convenient in applications. For instance, relation (7.6) can be rewritten using the idea of thresholding. For a given $f \in L_p$ denote

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

Then (7.6) is equivalent to

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

For further results in this direction see [**14**], [**10**], [**63**].

7.4. *Stability*

In this section we assume that a basis $\Psi = \{\psi_k\}_{k=1}^\infty$ is an unconditional normalized ($\|\psi_k\| = 1, k = 1, 2, \ldots$) basis for X (see Definition 7.2).

The uniform boundedness principle implies that the unconditional constant

$$K := K(X, \Psi) := \sup_\theta \|M_\theta\|$$

is finite.

The following theorem is a well-known fact about unconditional bases (see [**49**, p. 19]).

Theorem 7.7. *Let* Ψ *be an unconditional basis for* X. *Then, for every choice of bounded scalars* $\{\lambda_k\}_{k=1}^\infty$, *we have*

$$\left\| \sum_{k=1}^\infty \lambda_k a_k \psi_k \right\| \leq 2K \sup_k |\lambda_k| \left\| \sum_{k=1}^\infty a_k \psi_k \right\|$$

(*in the case of a real Banach space* X *we can take* K *instead of* $2K$).

In the numerical implementation of nonlinear m-term approximation one usually prefers to employ the strategy known as thresholding (see [**14**, S.7.8]) instead

of a greedy algorithm. We define and study here the soft thresholding (see [**27**]). Let a real function $v(x)$ defined for $x \geq 0$ satisfy the following relations:

$$v(x) = \begin{cases} 1, & \text{for } x \geq 1, \\ 0, & \text{for } 0 \leq x \leq \frac{1}{2}, \end{cases} \tag{7.11}$$

$$|v(x)| \leq A, \qquad x \in [0, 1], \tag{7.12}$$

there is a constant C_L such that for any $x, y \in [0, \infty)$ we have

$$|v(x) - v(y)| \leq C_L|x - y|. \tag{7.13}$$

Let

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

We define a soft thresholding mapping $T_{\varepsilon,v}$ as follows. Take $\varepsilon > 0$ and set

$$T_{\varepsilon,v}(f) := \sum_k v(|c_k(f)|/\varepsilon)c_k(f)\psi_k.$$

Theorem 7.7 implies that

$$\|T_{\varepsilon,v}(f)\| \leq 2KA\|f\|. \tag{7.14}$$

It was proved in [**93**] that the mapping $T_{\varepsilon,v}$ satisfies the Lipschitz condition with a constant independent of ε.

Theorem 7.8. *For any ε and any functions $f, g \in X$ we have*

$$\|T_{\varepsilon,v}(f) - T_{\varepsilon,v}(g)\| \leq (3A + 2C_L)2K\|f - g\|.$$

Open Problems

7.1. Does the inequality

$$\|f - G_m^{c,t}(f, \mathcal{RT})\|_p \leq C_1(p, t)\sigma_n(f, \mathcal{RT})_p$$

hold for any $f \in L_p(\mathbb{T})$, $1 < p < \infty$, with $m \leq C_2(p, t)n$?

7.2. Does the inequality

$$\|f - G_m^{c,t}(f, \mathcal{H}_p)\|_p \leq C_1(p, t)\sigma_n(f, \mathcal{H}_p)_p$$

hold for any $f \in L_p(0, 1)$, $1 < p < \infty$, with $m \leq C_2(p, t)n$?

7.3. Find the order of the quantity

$$\sup_{f \in W_p^r} \|f - G_m^{c,t}(f, \mathcal{RT})\|_p, \qquad 1 < p < \infty.$$

7.4. Find greedy-type algorithms realizing near best approximation in the L_p $([0, 1]^d)$, $1 < p < \infty$, $d \geq 2$, with regard to $\mathcal{H}_p^d$ for individual functions.

8. Some Convergence Results

In Section 7 we discussed greedy bases. That is justified from the point of view of efficient approximation. It follows from Proposition 7.1 that the inequality

$$\|G_m(f, \Psi, \rho)\| \le (G+1)\|f\| \tag{8.1}$$

holds for all m and all $f \in X$ for every $\rho \in D(f)$.

Definition 8.1. 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

$$\min_{k\in\Lambda} |c_k(f)| \ge \max_{k\notin\Lambda} |c_k(f)|, \tag{8.2}$$

we have

$$\|S_\Lambda(f, \Psi)\| = \left\| \sum_{k\in\Lambda} c_k(f)\psi_k \right\| \le C_Q\|f\|. \tag{8.3}$$

It is clear that the inequalities (8.1) and (8.3) are equivalent. P. Wojtaszczyk [**104**] proved that a basis Ψ is quasi-greedy if and only if the sequence $\{G_m(f, \Psi, \rho)\}$ converges to f for all $f \in X$ and any $\rho \in D(f)$. We constructed in [**44**] an example of a quasi-greedy basis that is not an unconditional basis (and, therefore, not a greedy basis). We have the following theorem for the trigonometric system.

Theorem 8.1. *The trigonometric system $\mathcal{T}$ is not a quasi-greedy basis for L_p if $p \ne 2$.*

This theorem has been proved in [**91**] and for $p < 2$ it has been proved independently and by a different method in [**12**]. We mention here that the method from [**91**] gives a little more than stated in Theorem 8.1.

Theorem 8.2. *There exists a continuous function f such that $G_m(f, \mathcal{T})$ does not converge to f in L_p for any $p > 2$.*

Theorem 8.3. *There exists a function f that belongs to any L_p, $p < 2$, such that $G_m(f, \mathcal{T})$ does not converge to f in measure.*

The proof of both theorems is based on two examples (one for $p > 2$ and the other for $p < 2$) constructed in [**91**, pp. 574–575]. We prove here only Theorem 8.3 where we use the example from [**91**] for $p < 2$.

Proof of Theorem 8.3. We use the Rudin–Shapiro polynomials (see [**41**])

$$R_N(x) = \sum_{k=0}^{N-1} \varepsilon_k e^{ikx}, \qquad \varepsilon_k = \pm 1, \quad x \in \mathbb{T},$$

that satisfy the inequality

$$\|R_N\|_\infty \le CN^{1/2}, \tag{8.4}$$

with an absolute constant C. Denote, for $s = \pm 1$,

$$\Lambda_s(N) := \{k : \hat{R}_N(k) = s\}.$$

Denote also

$$D_\Lambda(x) := \sum_{k\in\Lambda} e^{ikx}.$$

Then

$$R_N = D_{\Lambda_{+1}} - D_{\Lambda_{-1}}.$$

Inequality (8.4) implies

$$\|R_N\|_1 \ge C_1 N^{1/2}.$$

Using this inequality we prove that there exist two positive constants c_1 and c_2 such that for one of $s = \pm 1$ we have

$$m\{x : |D_{\Lambda_s(N)}(x)| \ge c_1 N^{1/2}\} \ge c_2. \tag{8.5}$$

We define a function f from Theorem 8.3 as follows:

$$f := \sum_{v=1}^{\infty} 2^{-v/2} e^{i2^v x}(D_{[0,2^v)} + s2^{-v}R_{2^v}).$$

Then for appropriately chosen m_1 and m_2 we get

$$G_{m_1}(f,\mathcal{T}) - G_{m_2}(f,\mathcal{T}) = 2^{-v/2}e^{i2^v x}(1+2^{-v})D_{\Lambda_s(2^v)}$$

and, by (8.5),

$$m\{x : |G_{m_1}(f) - G_{m_2}(f)| \ge c_1\} \ge c_2$$

which shows that $\{G_m(f,\mathcal{T})\}$ does not converge in measure. Further, for any $1 < p < 2$ we have

$$\|D_{[0,2^v)} + s2^{-v}R_{2^v}\|_p \le C2^{v(1-1/p)}$$

which implies that $f \in L_p$.

We also mention two interesting results on convergence almost everywhere. T. W. Körner answering a question raised by Carleson and Coifman constructed in [**46**] a function from L_2 and then in [**47**] a continuous function such that $\{G_m(f,\mathcal{T})\}$ diverges almost everywhere. T. Tao [**72**] proved that for the Haar system we have convergence: the sequence $\{G_m(f,\mathcal{H}_p)\}$ converges almost everywhere to f for any $f \in L_p$, $1 < p < \infty$.

Open Problems

8.1. Does the L_p-Greedy Algorithm with regard to $\mathcal{T}$ converge in L_p, $1 < p < \infty$, for each $f \in L_p(\mathbb{T})$?

8.2. Does the DGA with regard to $\mathcal{T}$ converge in L_p, $1 < p < \infty$, for each $f \in L_p(\mathbb{T})$?

8.3. Does the L_p-Greedy Algorithm with regard to $\mathcal{H}_p$ converge in L_p, $1 < p < \infty$, for each $f \in L_p(0, 1)$?

8.4. Does the DGA with regard to $\mathcal{H}_p$ converge in L_p, $1 < p < \infty$, for each $f \in L_p(0, 1)$?

9. Nonlinear m-Term Approximation and ε-Entropy

In this section, we want to bring out the connection between approximation from a dictionary and ε-entropy. We begin with covering numbers $N_\varepsilon(F, \ell_p)$ for a set $F \subset \mathbb{R}^n$ and recall their definition. For each $\varepsilon > 0$,

$$N_\varepsilon(F, l_p) := \min \left\{ N : F \subset \bigcup_{j=\ell}^{N} B_p^n(y^j, \varepsilon) \right\}$$

with the minimum taken over all sets $\{y^j\}_{j=1}^N$ of points from $\mathbb{R}^n$. Here $B_p^n(y, \varepsilon)$ denotes the ℓ_p-ball of radius ε with center y. By considering systems $\mathcal{D}$ consisting of the points y^j, we find

$$\inf_{\#\mathcal{D}=N_\varepsilon(F,\ell_p)} \sigma_1(F, \mathcal{D})_{\ell_p} \le \varepsilon. \tag{9.1}$$

In other words, the covering numbers immediately give estimates for 1-term approximation. We can extend the above observation to m-term approximation by using the concept of metric entropy. 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 spaces spanned by m elements of $\mathcal{D}$. For a linear space $L \subset X$, the ε-neighborhood $U_\varepsilon(L)$ 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

$$\varepsilon_{N,m}(F, X) := \inf_{\#\mathcal{D}=N} \inf \left\{ \varepsilon : F \subset \bigcup_{L \in \mathcal{L}_m(\mathcal{D})} U_\varepsilon(L) \right\}.$$

We can express $\sigma_m(F, \mathcal{D})$ as

$$\sigma_m(F, \mathcal{D}) = \inf \left\{ \varepsilon : F \subset \bigcup_{L \in \mathcal{L}_m(\mathcal{D})} U_\varepsilon(L) \right\}.$$

It follows therefore that

$$\inf_{\#\mathcal{D}=N} \sigma_m(F, \mathcal{D}) = \varepsilon_{N,m}(F, X).$$

In other words, finding best dictionaries for the m-term approximation of F is the same as finding sets $\mathcal{D}$ which attain the (N, m)-entropy numbers $\varepsilon_{N,m}(F, X)$. It is easy to see that $\varepsilon_{m,m}(F, X) = d_m(F, X)$. This establishes a connection between (N, m)-entropy numbers and the Kolmogorov widths.

The present section contains an attempt to generalize the concept of the classical Kolmogorov width in order to be used in estimating the best m-term approximation. For this purpose we introduce a nonlinear Kolmogorov (N, m)-width:

$$d_m(F, X, N) := \inf_{\Lambda_N, \#\Lambda_N \le N} \sup_{f \in F} \inf_{L \in \Lambda_N} \inf_{g \in L} \|f - g\|_X,$$

where Λ_N is a set of at most N m-dimensional subspaces L. It is clear that

$$d_m(F, X, 1) = d_m(F, X)$$

and

$$d_m\left(F, X, \binom{N}{m}\right) \le \varepsilon_{N,m}(F, X) \le \sigma_m(F, \mathcal{D})$$

for any $\mathcal{D}$ with $\#\mathcal{D} = N$. The new feature of $d_m(F, X, N)$ is that we allow a choice of subspace $L \in \Lambda_N$ depending on $f \in F$. It is clear that the larger N is, the more flexibility we have to approximate f. It turns out that from the point of view of our applications the following two cases:

(I)
$$N \asymp K^m,$$

where $K > 1$ is a constant, and

(II)
$$N \asymp m^{am},$$

where $a > 0$ is a fixed number,

play an important role.

We intend to use the (N, m)-widths to estimate from below the best m-term approximations. There are several general results (see [**53**], [**7**]) which give lower estimates of the Kolmogorov widths $d_n(F, X)$ in terms of the entropy numbers $\varepsilon_k(F, X)$. In [**90**] we generalized the following inequality due to Carl (see [**7**]): for any $r > 0$, we have

$$\max_{1 \le k \le n} k^r \varepsilon_k(F, X) \le C(r) \max_{1 \le m \le n} m^r d_{m-1}(F, X). \tag{9.2}$$

We denote here, for any positive integer k,

$$\varepsilon_k(F, X) := \inf\left\{\varepsilon : \exists f_1, \dots, f_{2^k} \in X : F \subset \bigcup_{j=1}^{2^k} (f_j + \varepsilon B(X))\right\},$$

where $B(X)$ is the unit ball of Banach space X. For noninteger k we set $\varepsilon_k(F, X) := \varepsilon_{[k]}(F, X)$ where $[k]$ is the integral part of number k. It is clear that

$$d_1(F, X, 2^n) \le \varepsilon_n(F, X).$$

In [**90**] we proved the inequality

$$\max_{1\le k\le n} k^r \varepsilon_k(F, X) \le C(r, K) \max_{1\le m\le n} m^r d_{m-1}(F, X, K^m), \tag{9.3}$$

where we denote

$$d_0(F, X, N) := \sup_{f\in F} \|f\|_X.$$

This inequality is a generalization of inequality (9.2). In [**90**] we also proved the following inequality

$$\max_{1\le k\le n} k^r \varepsilon_{(a+r)k\log k}(F, X) \le C(r, a) \max_{1\le m\le n} m^r d_{m-1}(F, X, m^{am}) \tag{9.4}$$

and gave an example showing that $k \log k$ in this inequality cannot be replaced by any more slowly growing function of k.

In [**90**] we applied inequalities (9.3) and (9.4) to estimate the best m-term trigonometric approximation from below. As a corollary to the following version of (9.3) (see Theorem 9.1 below) we gave a new proof (see [**17**]) for the estimate

$$\sigma_m(W^r_\infty, \mathcal{T})_1 \gg m^{-r},$$

where W^r_∞ is a standard Sobolev class (see Section 2) with the restriction imposed in the L_∞-norm.

Theorem 9.1. *For any positive constant K we have*

$$\max_{1\le k\le n} k^r \varepsilon_k(F, X) \le C(r, K) \max_{1\le m\le n} m^r d_{m-1}(F, X, (Kn/m)^m).$$

We used in [**90**] a version of (9.4) to get some new lower estimates of m-term trigonometric approximation in the L_1-norm of multivariate classes MW^r_∞ of functions with bounded mixed derivative. We proved in [**90**] that

$$\sigma_m(MW^r_\infty, \mathcal{T})_1 \gg m^{-r}(\log m)^{r(d-2)}. \tag{9.5}$$

Inequality (9.5) gives a new estimate for small r.

The above method can be applied to a general system Ψ instead of to the trigonometric system $\mathcal{T}$.

Assume a system $\Psi := \{\psi_j\}_{j=1}^\infty$ of elements in X satisfies the condition:

(VP) There exist three positive constants $A_i, i = 1, 2, 3$, and a sequence $\{n_k\}_{k=1}^\infty$, $n_{k+1} \le A_1 n_k$, $k = 1, 2, \dots$, such that there is a sequence of the de la Vallée-

Poussin-type operators V_k with the properties

$$V_k(\psi_j) = \lambda_{k,j}\psi_j, \tag{9.6}$$

$$\lambda_{k,j} = 1 \quad \text{for} \quad j = 1, \dots, n_k, \qquad \lambda_{k,j} = 0 \quad \text{for} \quad j > A_2 n_k,$$

$$\|V_k\|_{X\to X} \le A_3, \qquad k = 1, 2, \dots. \tag{9.7}$$

Theorem 9.2. *Assume that for some $a > 0$ and $b \in \mathbb{R}$ we have*

$$\varepsilon_m(F, X) \ge C_1 m^{-a}(\log m)^b, \qquad m = 1, 2, \dots.$$

Then if a system Ψ satisfies the condition (VP) and also satisfies the following condition:

$$E_n(F, \Psi) := \sup_{f\in F} \inf_{c_1,\dots,c_n} \left\| f - \sum_{j=1}^{n} c_j\psi_j \right\|_X \le C_2 n^{-a}(\log n)^b, \qquad n = 1, 2, \dots,$$

then we have

$$\sigma_m(F, \Psi)_X \gg m^{-a}(\log m)^b.$$

Open Problem

9.1. The correct order of the quantity $\sigma_m(MW^r_\infty, \mathcal{T})_1$ is unknown.

10. Optimal Methods in Nonlinear Approximation

In the widths problem of Linear Approximation we were looking for an optimal n-dimensional subspace for approximating a given function class. A nonlinear analog of this setting is the following. 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. Then, similar to the setting for the widths d_n, λ_n, φ_n, we get the optimization problems for m-term nonlinear approximation (see the Introduction). Let $\mathbb{B}$ be a collection of bases satisfying a given property.

I. Define an analog 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 analog of the orthowidth

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

We present here some results in the case $\mathbb{B} = \mathbb{O}$—the set of orthonormal bases, $F = W_q^r$, $X = L_p$, $1 \le q, p \le \infty$. First of all we formulate a result (see [**42**], [**94**]) that shows that in the case $p < 2$ we need some more restrictions on $\mathbb{B}$ in order to obtain meaningful results (lower bounds).

Proposition 10.1. *For any $1 \le p < 2$ there exists a complete in the $L_2(0,1)$ orthonormal system Φ such that for each $f \in L_p(0,1)$ we have $\sigma_1(f,\Phi)_p = 0$.*

Let us restrict our further discussion to the case $p \ge 2$. This case was also more interesting in the Linear Approximation discussion (see Section 2). Kashin [**40**] proved that

$$\sigma_m(W_\infty^r, \mathbb{O})_2 \gg m^{-r}. \tag{10.1}$$

We proved (see [**17**]) that

$$\sigma_m(W_2^r, \mathcal{T})_\infty \ll m^{-r}. \tag{10.2}$$

The estimates (10.1) and (10.2) imply that for $2 \le q, p \le \infty$ we have

$$\sigma_m(W_q^r, \mathbb{O})_p \asymp \sigma_m(W_q^r, \mathcal{T})_p \asymp m^{-r}. \tag{10.3}$$

Let us compare this relation with (2.2). We see that the best m-term trigonometric approximation provides the same accuracy as the best approximation from an optimal m-dimensional subspace. An advantage of nonlinear approximation here is that we use a natural basis instead of an existing but nonconstructive subspace. However, we should note that the estimate (10.2) was proved in [**17**] as an existence theorem. We did not give an algorithm to get (10.2) in [**17**] and still do not know the algorithm. The TGA does not provide the estimate (10.2). We have (see [**91**])

$$\sup_{f \in W_2^r} \|f - G_m(f, \mathcal{T})\|_\infty \asymp m^{-r+1/2}.$$

It is known from different results (see [**20**], [**14**], [**93**]) that wavelets are well-designed for nonlinear approximation. We present here one general result in this direction. We consider 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. Let $L_p := L_p(\Omega)$ with a normalized Lebesgue measure on Ω, $|\Omega| = 1$. First of all we assume that, for all $1 < q, p < \infty$, and $I \in D$, $D := D([0,1]^d)$ is the set of all dyadic intervals of $[0,1]^d$, we have

$$\|\psi_I\|_p \asymp \|\psi_I\|_q |I|^{1/p - 1/q}, \tag{10.4}$$

with constants independent of I. This property can easily be 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 $1 < p < \infty$,

$$\left\| \sum_{I \in D_s} c_I \psi_I \right\|_p^p \asymp \sum_{I \in D_s} \|c_I \psi_I\|_p^p, \tag{10.5}$$

where

$$D_s := \{I = I_1 \times \cdots \times I_d \in D : |I_j| = 2^{-s_j}, j = 1, \dots, d\}.$$

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 Littlewood–Paley inequality. This means the following. Let $1 < p < \infty$ and $f \in L_p$ has an expansion

$$f = \sum_I f_I \psi_I.$$

We assume that

$$\lim_{\min_j \mu_j \to \infty} \left\| f - \sum_{s_j \le \mu_j, j = 1, \dots, d} \sum_{I \in D_s} f_I \psi_I \right\|_p = 0, \tag{10.6}$$

and

$$\|f\|_p \asymp \left\| \left(\sum_s \left| \sum_{I \in D_s} f_I \psi_I \right|^2 \right)^{1/2} \right\|_p. \tag{10.7}$$

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

$$g(R) := \left(\sum_{j=1}^d R_j^{-1} \right)^{-1}.$$

For natural numbers l denote

$$\Psi(R, l) := \Psi(\mu), \qquad \mu_j = [g(R) l / R_j], \qquad j = 1, \dots, d.$$

We define the class $H_q^R(\Psi)$ as the set of functions $f \in L_q$ representable in the form

$$f = \sum_{l=1}^\infty t_l, \quad t_l \in \Psi(R, l), \qquad \|t_l\|_q \le 2^{-g(R) l}.$$

Theorem 10.1. *Let $1 < q, p < \infty$, and $g(R) > (1/q - 1/p)_+$. Then for Ψ satisfying* (10.4)–(10.7) *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 following basis $U^d := U \times \cdots \times U$ can be taken in place of Ψ in Theorem 10.1. We define the system $U := \{U_I\}$ in the univariate case. Denote

$$\begin{aligned} U_n^+(x) &:= \sum_{k=0}^{2^n-1} e^{ikx} = \frac{e^{i2^n x} - 1}{e^{ix} - 1}, && n = 0, 1, 2, \dots, \\ U_{n,k}^+(x) &:= e^{i2^n x} U_n^+(x - 2\pi k 2^{-n}), && k = 0, 1, \dots, 2^n - 1, \\ U_{n,k}^-(x) &:= e^{-i2^n x} U_n^+(-x + 2\pi k 2^{-n}), && k = 0, 1, \dots, 2^n - 1. \end{aligned}$$

We normalize the system of functions $\{U_{n,k}^+, U_{n,k}^-\}$ in L_2 and enumerate it by dyadic intervals. We write

$$\begin{aligned} U_I(x) &:= 2^{-n/2} U_{n,k}^+(x) \quad \text{with} \quad I = [(k + \tfrac{1}{2})2^{-n}, (k+1)2^{-n}), \\ U_I(x) &:= 2^{-n/2} U_{n,k}^-(x) \quad \text{with} \quad I = [k2^{-n}, (k + \tfrac{1}{2})2^{-n}), \end{aligned}$$

and

$$U_{[0,1)}(x) := 1.$$

It is well-known that $H_q^R(U^d)$ is equivalent to the standard anisotropic multivariate periodic Hölder–Nikol'skii classes NH_p^R. We define these classes in the following way. The class NH_p^R, $R = (R_1, \dots, R_d)$ 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

$$\|f\|_p \le 1, \qquad \|\Delta_t^{l_j, j} f\|_p \le |t|^{R_j}, \qquad j = 1, \dots, d, \tag{10.8}$$

where $\Delta_t^{l,j}$ is the lth difference with step t in the variable x_j. In the case $d = 1, NH_p^R$ coincides with the standard Hölder class H_p^R. Theorem 10.1 gives the following result:

Theorem 10.2. *Let $1 < q, p < \infty$; then for R such that $g(R) > (1/q - 1/p)_+$ we have*

$$\sup_{f \in NH_q^R} \|f - G_m^{L_p}(f, U^d)\|_p \ll m^{-g(R)}.$$

We also proved in [**93**] that the basis U^d is an optimal orthonormal basis for the approximation of classes NH_q^R in L_p:

$$\sigma_m(NH_q^R, \mathbb{O})_p \asymp \sigma_m(NH_q^R, U^d)_p \asymp m^{-g(R)} \tag{10.9}$$

for $1 < q < \infty$, $2 \le p < \infty$, $g(R) > (1/q - 1/p)_+$. It is important to remark that Theorem 10.2 guarantees that the estimate in (10.9) can be realized by TGA with regard to U^d.

Open Problem

10.1. Find a constructive proof of (10.2) (provide an algorithm).

11. Universality

In this section we discuss, in the model case of the anisotropic function classes, a general approach formulated in the Introduction of how to choose a good basis (dictionary) for approximation. This approach consists of several steps. We concentrate here on nonlinear approximation and compare realizations of this approach for linear and nonlinear approximations. The first step in this approach is an optimization problem. In both cases (linear and nonlinear) we begin with a function class F in a given Banach space X. A classical example of the optimization problem in the linear case is the problem of finding (estimating) the Kolmogorov width $d_m(F, X)$. This concept allows us to choose among various Chebyshev methods (best approximation) having the same dimensions of the approximating subspaces the one which has the best accuracy. The asymptotic behavior (in the sense of order) of the sequence $\{d_m(F, X)\}_{m=1}^{\infty}$ is known for a number of function classes and Banach spaces. It turns out that in many cases, for instance, when $F = W_p^r$ is a standard Sobolev class and $X = L_p$, the optimal (in the sense of order) m-dimensional subspaces are spanned by m elements from one orthogonal system. We describe this for the multivariate periodic Hölder–Nikol'skii classes NH_p^R. It is known (see, for instance, [**86**]) that

$$d_m(NH_p^R, L_p) \asymp m^{-g(R)}, \qquad 1 \le p \le \infty. \tag{11.1}$$

It is also known that the subspaces of trigonometric polynomials $\mathcal{T}(R, l)$ with frequencies k, satisfying the inequalities

$$|k_j| \le 2^{g(R)l/R_j}, \qquad j = 1, \dots, d,$$

can be chosen to realize (11.1). In this case l is set to be the largest integer satisfying $\dim \mathcal{T}(R, l) \le m$. We stress here that optimal (in the sense of order) subspaces $\mathcal{T}(R, l)$ are different for different R and formed from the same (trigonometric) system.

A nonlinear analog of the Kolmogorov m-width setting was discussed in Section 10. In this section we consider only the case $\mathbb{D} = \mathbb{O}$—the set of all orthogonal bases on a given domain. In Section 10 we mentioned that

$$\sigma_m(NH_q^R, \mathbb{O})_{L_p} \asymp m^{-g(R)} \tag{11.2}$$

for

$$1 < q < \infty, \quad 2 \le p < \infty, \qquad g(R) > (1/q - 1/p)_+.$$

It is important to remark that the basis U^d realizes (11.2) for all R (see the definition of U^d in Section 10).

The second step in our approach is to look for a universal basis (dictionary) for approximation. The above-mentioned result on the basis U^d means that U^d is universal for the pair $(\mathcal{F}_q([A, B]), \mathbb{O})$ and the space $X = L_p([0, 2\pi]^d)$ for $A, B \in \mathbb{Z}^d_+$ such that $g(A) > (1/q - 1/p)_+$, $1 < q < \infty$, $2 \le p < \infty$, where

$$\mathcal{F}_q([A, B]) := \{NH^R_q : 0 < A_j \le R_j \le B_j < \infty, j = 1, \dots, d\}.$$

It is interesting to compare this result on universal bases in the nonlinear approximation with the corresponding result in the linear setting. We define the index $\kappa(m, \mathcal{F}, X)$ of universality for a collection $\mathcal{F}$ with respect to the Kolmogorov width in X:

$$\kappa(m, \mathcal{F}, X) := L(m, \mathcal{F}, X)/m,$$

where $L(m, \mathcal{F}, X)$ is the smallest number among those L for which there is a system of functions $\{\varphi_i\}_{i=1}^L$ such that for each $F \in \mathcal{F}$ we have

$$\sup_{f \in F} \inf_{c_1, \dots, c_L} \left\| f - \sum_{i=1}^{L} c_i \varphi_i \right\| \le d_m(F, X).$$

It is proved in [**79**] (see also [**86**, Ch. 3, S.5]) that for any $A, B \in \mathbb{Z}^d_+$ such that $B_j > A_j$, $j = 1, \dots, d$, we have

$$\kappa(m, \mathcal{F}_p([A, B]), L_p) \gg (\log m)^{d-1}, \qquad 1 < p < \infty. \tag{11.3}$$

The estimate (11.3) says that there is no Chebyshev method that is universal for a nontrivial collection of anisotropic function classes. Thus, from the point of view of the existence of universal methods the nonlinear setting has an advantage over the linear setting.

After two steps of realizing our approach in the nonlinear approximation we get a universal dictionary $\mathcal{D}_u$ for a collection of function classes $\mathcal{F}$, say, U^d for $\mathcal{F}_q([A, B])$. This means that the dictionary $\mathcal{D}_u$ is well-designed for the best m-term approximation of functions from function classes in the given collection. The third step is to find an algorithm (theoretical first) to realize the best (near best) m-term approximation with regard to $\mathcal{D}_u$. It turns out that in the model case of $\mathcal{F}_q([A, B])$ and the basis U^d there is a simple algorithm which realizes near the best m-term approximation for classes NH^R_q. This is the TGA (see Theorem 10.2).

Thus we have established that in the above model case the basis U^d is optimal for nonlinear m-term approximation in a very strong sense. The following two features of U^d are the most important ones:

(1) U^d is the tensor product of the univariate basis U;
(2) the univariate basis U is a wavelet-type basis.

It is known [**103**] that U is L_p-equivalent, $1 < p < \infty$, to the Haar basis. Then, by Theorem 7.1, U is a greedy basis for L_p, $1 < p < \infty$. The tensor product structure of U^d is important in making U^d a universal basis for a collection of anisotropic Hölder–Nikol'skii classes. It would be ideal if U^d was a greedy basis for $L_p(\mathbb{T}^d)$, $1 < p < \infty$. Unfortunately, this is not the case. We have that, for $1 < p < \infty$,

$$\sup_{f \in L_p} \|f - G_m^p(f, U^d)\|_p / \sigma_m(f, U^d)_p \asymp (\log m)^{(d-1)|1/2-1/p|}. \tag{11.4}$$

This relation follows from its analog with U^d replaced by the multivariate Haar system $\mathcal{H}^d := \mathcal{H} \times \cdots \times \mathcal{H}$. The lower estimate in (11.4) for $\mathcal{H}^d$ was proved by R. Hochmuth; the upper estimate in (11.4) for $\mathcal{H}^d$ was proved in the case $d = 2$, $\frac{4}{3} \le p \le 4$, and was conjectured for all d, $1 < p < \infty$, in [**89**]. The conjecture was proved in [**104**].

Acknowledgments

The author would like to thank Ron DeVore and Wolfgang Dahmen for interesting and helpful discussions which resulted, in particular, in a better presentation of the material. Also, the author is grateful to Janice Long for polishing up the English. This research was supported by the National Science Foundation Grant DMS 9970326 and by ONR Grant N00014-91-J1343.

References

[1] A. V. Andrianov and V. N. Temlyakov, Best m-term approximation of functions from classes MW_q^r, *Approx. Theory* **IX**, 7–14.

[2] K. I. Babenko, Some problems in approximation theory and numerical analysis, *Russian Math. Surveys* **40** (1985), 1–30.

[3] B. M. Baishanski, Approximation by polynomials of given length, *Illinois J. Math.* **27** (1983), 449–458.

[4] A. R. Barron, Universal approximation bounds for superposition of n sigmoidal functions, *IEEE Trans. Inform. Theory* **39** (1993), 930–945.

[5] S. N. Bernstein, On the best approximation of functions of several variables by means of polynomials of trigonometric sums, *Trudy Mat. Inst. Steklov* **38** (1951), 24–29.

[6] M. Sh. Birman and M. Z. Solomyak, Estimates of singular numbers of integral operators, *Uspekhi Mat. Nauk* **32** (1977), 17–84; English transl. in *Russian Math. Surveys* **32** (1977).

[7] B. Carl, Entropy numbers, s-numbers, and eigenvalue problems, *J. Funct. Anal.* **41** (1981), 290–306.

[8] S. S. Chen, D. L. Donoho, and M. A. Saunders, Atomic decomposition by basis pursuit, *SIAM Rev.* **43** (2001), 129–159.

[9] J. A. Cochran, Composite integral operators and nuclearity, *Ark. Mat.* **15** (1977), 215–222.

[10] A. Cohen, R. A. DeVore, and R. Hochmuth, Restricted nonlinear approximation, *Constr. Approx.* **16** (2000), 85–113.

[11] R. R. Coifman and M. V. Wickerhauser, Entropy-based algorithms for best-basis selection, *IEEE Trans. Inform. Theory* **38** (1992), 713–718.

[12] A. Cordoba and P. Fernandez, Convergence and divergence of decreasing rearranged Fourier series, *SIAM J. Math. Anal.* **29** (1998), 1129–1139.
[13] G. Davis, S. Mallat, and M. Avellaneda, Adaptive greedy approximations, *Constr. Approx.* **13** (1997), 57–98.
[14] R. A. DeVore, Nonlinear approximation, *Acta Numerica* (1998), 51–150.
[15] R. A. DeVore and G. G. Lorenz, *Constructive Approximation*, Springer-Verlag, Berlin, 1993.
[16] R. A. DeVore and V. A. Popov, *Interpolation Spaces and Non-Linear Approximation*, Lecture Notes in Mathematics, Vol. 1302. Springer-Verlag, Berlin, 1988, pp. 191–205.
[17] R. A. DeVore and V. N. Temlyakov, Nonlinear approximation by trigonometric sums, *J. Fourier Anal. Appl.* **2** (1995), 29–48.
[18] R. A. DeVore and V. N. Temlyakov, Some remarks on greedy algorithms, *Adv. in Comput. Math.* **5** (1996), 173–187.
[19] R. A. DeVore and V. N. Temlyakov, Nonlinear approximation in finite-dimensional spaces, *J. Complexity* **13** (1997), 489–508.
[20] R. A. DeVore, B. Jawerth, and V. Popov, Compression of wavelet decompositions, *Amer. J. Math.* **114** (1992), 737–785.
[21] R. A. DeVore, S. V. Konyagin, and V. N. Temlyakov, Hyperbolic wavelet approximation, *Constr. Approx.* **14** (1998), 1–26.
[22] R. A. DeVore, K. I. Oskolkov, and P. P. Petrushev, Approximation by feedforward neural networks, *Ann. Numer. Math.* **4** (1997), 261–287.
[23] S. Dilworth, D. Kutzarova, and V. Temlyakov, Convergence of some greedy algorithms in Banach spaces, IMI-Preprint series **14** (2001), 1–21.
[24] M. Donahue, L. Gurvits, C. Darken, and E. Sontag, Rate of convex approximation in non-Hilbert spaces, *Constr. Approx.* **13** (1997), 187–220.
[25] D. L. Donoho, Unconditional bases are optimal bases for data compression and for statistical estimation, *Appl. Comput. Harmonic Anal.* **1** (1993), 100–115.
[26] D. L. Donoho, CART and best-ortho-basis: A connection, *Ann. Statist.* **25** (1997), 1870–1911.
[27] D. Donoho and I. Johnstone, Ideal spatial adaptation via wavelet shrinkage, *Biometrica* **81** (1994), 425–455.
[28] V. V. Dubinin, *Greedy algorithms and applications*, Ph.D. Thesis, University of South Carolina, 1997.
[29] M. Frazier and B. Jawerth, A discrete transform and decomposition of distribution spaces, *J. Funct. Anal.* **93** (1990), 34–170.
[30] I. Fredholm, Sur une classe d'équations fonctionalles, *Acta Math.* **27** (1903), 365–390.
[31] J. H. Friedman and W. Stuetzle, Projection pursuit regression, *J. Amer. Statist. Assoc.* **76** (1981), 817–823.
[32] I. C. Golberg and M. G. Krein, *Introduction to the Theory of Linear Nonselfadjoint Operators in Hilbert space*, American Mathematical Society, Providence, RI 02904, 1969.
[33] E. Hille and J. D. Tamarkin, On the characteristic values of linear integral equations, *Acta Math.* **57** (1931), 1–76.
[34] P. J. Huber, Projection pursuit, *Ann. Statist.* **13** (1985), 435–475.
[35] R. S. Ismagilov, Widths of sets in normed linear spaces and the approximation of functions by trigonometric polynomials, *Uspekhi Mat. Nauk* **29** (1974), 161–178; English transl. in *Russian Math. Surveys* **29** (1974).
[36] L. Jones, On a conjecture of Huber concerning the convergence of projection pursuit regression, *Ann. Statist.* **15** (1987), 880–882.
[37] L. Jones, A simple lemma on greedy approximation in Hilbert space and convergence rates for projection pursuit regression and neural network training, *Ann. Statist.* **20** (1992), 608–613.
[38] J. P. Kahane, *Series De Fourier Absolument Convergentes*, Springer-Verlag, Berlin, 1976.
[39] B. S. Kashin, Widths of certain finite-dimensional sets and classes of smooth functions, *Izv. Akad. Nauk SSSR, Ser. Mat.* **41** (1977), 334–351; English transl. in *Math. USSR-Izv.* **11** (1977).
[40] B. S. Kashin, On approximation properties of complete orthonormal systems, *Trudy Mat. Inst. Steklov* **172** (1985), 187–191; English transl. in *Proc. Steklov Inst. Math.* **1987**, no. 3, 207–211.

[41] B. S. Kashin and A. A. Saakyan, *Orthogonal Series*, American Mathematical Society, Providence, RI, 1989.
[42] B. S. Kashin and V. N. Temlyakov, On best m-term approximations and the entropy of sets in the space L^1, *Math. Notes* **56** (1994), 57–86.
[43] B. S. Kashin and V. N. Temlyakov, On estimating approximative characteristics of classes of functions with bounded mixed derivative, *Math. Notes* **58** (1995), 922–925.
[44] S. V. Konyagin and V. N. Temlyakov, A remark on greedy approximation in Banach spaces, *East J. Approx.* **5** (1999), 1–15.
[45] S. V. Konyagin and V. N. Temlyakov, Rate of convergence of pure greedy algorithm, *East J. Approx.* **5** (1999), 493–499.
[46] T. W. Körner, Divergence of decreasing rearranged Fourier series, *Ann. of Math.* **144** (1996), 167–180.
[47] T. W. Körner, Decreasing rearranged Fourier series, *J. Fourier Anal. Appl.* **5** (1999), 1–19.
[48] H. Lebesgue, Sur les intégrales singuliéres, *Ann. Fac. Sci. Univ. Toulouse* (3) **1** (1909), 25–117.
[49] J. Lindenstrauss and L. Tzafriri, *Classical Banach Spaces* I, Springer-Verlag, Berlin, 1977.
[50] E. D. Livshitz, On the rate of convergence of greedy algorithm, Manuscript (2000).
[51] E. D. Livshitz and V. N. Temlyakov, On convergence of weak greedy algorithms, IMI-Preprint **13** (2000), 1–9.
[52] B. F. Logan and L. A. Shepp, Optimal reconstruction of a function from its projections, *Duke Math. J.* **42** (1975), 645–659.
[53] G. G. Lorentz, Metric entropy and approximation, *Bull. Amer. Math. Soc.* **72** (1966), 903–937.
[54] G. G. Lorentz, M. von Golitschek, Yu. Makovoz, *Constructive Approximation*, Springer-Verlag, Berlin, 1996.
[55] V. E. Maiorov, On best approximation by ridge functions, Preprint, Technion, Israel **27** (1998).
[56] V. E. Maiorov, K. I. Oskolkov, and V. N. Temlyakov, Gridge approximations and radon compass, IMI-Preprint **9** (2000), 1–20.
[57] S. Mallat and Z. Zhang, Matching pursuit in a time-frequency dictionary, *IEEE Trans. Signal Proc.* **41** (1993), 3397–3415.
[58] S. N. Nikol'skii, *Approximation of functions of several variables and embedding theorems* (English translation of Russian, published by "Nauka," Moscow, 1969), Springer-Verlag, Berlin.
[59] S. M. Nikol'skii, On interpolation and best approximation of differentiable periodic functions by trigonometric polynomials, *Izv. Akad. Nauk SSSR Ser. Mat.* **10** (1946), 393–410.
[60] K. I. Oskolkov, An estimate in the approximation of continuous functions by subsequences of Fourier sums, *Proc. Steklov Inst. Math.* **134** (1975), 273–288.
[61] K. I. Oskolkov, On the Lebesgue inequality in the uniform metric and on a set of full measure, *Mat. Zametki* **18** (1975), 515–526.
[62] K. I. Oskolkov, Ridge approximation, Chebyshev–Fourier analysis and optimal quadrature formulas, *Proc. Steklov Inst. Math.* **219** (1997), 265–280.
[63] P. Oswald, Greedy algorithms and best m-term approximation with respect to biorthogonal systems, Preprint (2000), 1–22.
[64] P. Petrushev, *Direct and Converse Theorems for Spline and Rational Approximation and Besov Spaces*, Lecture Notes in Mathematics, Vol. 1302. Springer-Verlag, Berlin, 1988, pp. 363–377.
[65] G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry*, Cambridge University Press, Cambridge, 1989.
[66] V. V. Pospelov, On approximation of multivariate functions by products of univariate functions, *Inst. Appl. Math. Acad. Sci. USSR*, Preprint No. 32 (1978), 1–76.
[67] S. Qian and D. Chen, Signal representation using adaptive normalized Gaussian functions, *Signal Process.* **36** (1994), 329–355.
[68] L. Rejtö and G. G. Walter, Remarks on projection pursuit regression and density estimation, *Stochastic Anal. Appl.* **10** (1992), 213–222.
[69] E. Schmidt, Zur Theorie der linearen und nichtlinearen Integralgleichungen. I, *Math. Ann.* **63** (1906–1907), 433–476.

[70] F. Smithies, The eigenvalues and singular values of integral equations. *Proc. London Math. Soc.* (2) **43** (1937), 255–279.
[71] S. B. Stechkin, On absolute convergence of orthogonal series, *Dokl. Akad. Nauk SSSR* **102** (1955), 37–40.
[72] T. Tao, On the almost everywhere convergence of wavelet summation methods, *Appl. Comput. Harmonic Anal.* **3** (1996), 384–387.
[73] V. N. Temlyakov, On the asymptotic behavior of best approximations of continuous functions, *Soviet Math. Dokl.* **17** (1976), 739–743.
[74] V. N. Temlyakov, Asymptotic behavior of best approximations of continuous functions, *Math. USSR-Izv.* **11** (1977), 551–569.
[75] V. N. Temlyakov, Widths of some classes of functions of several variables, *Soviet Math. Dokl.* **26** (1982), 619–622.
[76] V. N. Temlyakov, Approximation of continuous functions by trigonometric polynomials, *Proc. Steklov Inst. Math.* (1983), 213–228.
[77] V. N. Temlyakov, On best bilinear approximations of periodic functions of several variables, *Soviet Math. Dokl.* **33** (1986), 96–99.
[78] V. N. Temlyakov, On the asymptotic behavior of best approximations of individual functions, *Proc. Steklov Inst. Math.* **3** (1987), 341–352.
[79] V. N. Temlyakov, Approximation by elements of a finite-dimensional subspace of functions from various Sobolev or Nikol'skii spaces, *Mat. Zametki* **43** (1988), 770–786; English transl. in *Math. Notes* **43** (1988), 444–454.
[80] V. N. Temlyakov, Estimates of the best bilinear approximations of functions of two variables and some of their applications, *Math. USSR-Sb.* **62** (1989), 95–109.
[81] V. Temlyakov, Approximation of functions with bounded mixed derivative, *Proc. Steklov Institute* (1989, Issue 1).
[82] V. N. Temlyakov, Bilinear approximation and applications, *Proc. Steklov Inst. Math.* **3** (1990), 221–248.
[83] V. N. Temlyakov, On estimates of approximation numbers and best bilinear approximation, *Constr. Approx.* **8** (1992), 23–33.
[84] V. N. Temlyakov, Estimates of best bilinear approximations of functions and approximation numbers of integral operators, *Math. Notes* **51** (1992), 510–517.
[85] V. N. Temlyakov, Bilinear approximation and related questions, *Proc. Steklov Inst. Math.* **4** (1993), 245–265.
[86] V. N. Temlyakov, *Approximation of periodic functions*, Nova Science, New York, 1993.
[87] V. N. Temlyakov, On approximation by ridge functions, Preprint, University of South Carolina, (1996), 1–12.
[88] V. N. Temlyakov, The best m-term approximation and greedy algorithms, *Adv. in Comput. Math.* **8** (1998), 249–265.
[89] V. N. Temlyakov, Nonlinear m-term approximation with regard to the multivariate Haar system, *East J. Approx.* **4** (1998), 87–106.
[90] V. N. Temlyakov, Nonlinear Kolmogorov's widths, *Mat. Zametki* **63** (1998), 891–902.
[91] V. N. Temlyakov, Greedy algorithm and m-term trigonometric approximation, *Constr. Approx.* **14** (1998), 569–587.
[92] V. N. Temlyakov, Greedy algorithms and m-term approximation with regard to redundant dictionaries, *J. Approx. Theory* **98** (1999), 117–145.
[93] V. N. Temlyakov, Universal bases and greedy algorithms, IMI-Preprint series **8** (1999), 1–20.
[94] V. N. Temlyakov, Greedy algorithms with regard to multivariate systems with special structure, *Constr. Approx.* **16** (2000), 399–425.
[95] V. N. Temlyakov, Weak greedy algorithms, *Adv. in Comput. Math.* **12** (2000), 213–227.
[96] V. N. Temlyakov, A criterion for convergence of weak greedy algorithms, IMI-Preprint series **21** (2000), 1–10.
[97] V. N. Temlyakov, Greedy algorithms in Banach spaces, *Adv. in Comput. Math.* **14** (2001), 277–292.

[98] V. N. Temlyakov, Two lower estimates in greedy approximation, IMI-Preprint series **07** (2001), 1–12.
[99] V. M. Tikhomirov, Widths of sets in function spaces and the theory of best approximations, *Uspekhi Mat. Nauk* **15** (1960), 81–120; English transl. in *Russian Math. Surveys* **15** (1960).
[100] A. F. Timan, *Theory of Approximation of Functions of a Real Variable*, Macmillan, New York, 1963.
[101] L. F. Villemoes, Best approximation with Walsh atoms, *Constr. Approx.* **13** (1997), 329–355.
[102] H. Weyl, Das asymptotische Verteilungsgesetz der Eiegenwerte linearer partieller Differentialgleichungen, *Math. Ann.* **71** (1911), 441–479.
[103] P. Wojtaszczyk, On unconditional polynomial bases in L_p and Bergman spaces, *Constr. Approx.* **13** (1997), 1–15.
[104] P. Wojtaszczyk, Greedy algorithms for general systems, *J. Approx. Theory* **107** (2000), 293–314.
[105] A. Zygmund, *Trigonometric Series*, Cambridge University Press, Cambridge, 1959.