CAMBRIDGE MONOGRAPHS ON APPLIED AND COMPUTATIONAL MATHEMATICS

20 Greedy Approximation

The *Cambridge Monographs on Applied and Computational Mathematics* series reflects the crucial role of mathematical and computational techniques in contemporary science. The series publishes expositions on all aspects of applicable and numerical mathematics, with an emphasis on new developments in this fast-moving area of research.

State-of-the-art methods and algorithms as well as modern mathematical descriptions of physical and mechanical ideas are presented in a manner suited to graduate research students and professionals alike. Sound pedagogical presentation is a prerequisite. It is intended that books in the series will serve to inform a new generation of researchers.

A complete list of books in the series can be found at
www.cambridge.org/mathematics
Recent titles include the following:

8. Schwarz–Christoffel mapping, *Tobin A. Driscoll & Lloyd N. Trefethen*
9. High-order methods for incompressible fluid flow, *M. O. Deville, P. F. Fischer & E. H. Mund*
10. Practical extrapolation methods, *Avram Sidi*
11. Generalized Riemann problems in computational fluid dynamics, *Matania Ben-Artzi & Joseph Falcovitz*
12. Radial basis functions, *Martin D. Buhmann*
13. Iterative Krylov methods for large linear systems, *Henk van der Vorst*
14. Simulating Hamiltonian dynamics, *Benedict Leimkuhler & Sebastian Reich*
15. Collocation methods for Volterra integral and related functional differential equations, *Hermann Brunner*
16. Topology for computing, *Afra J. Zomorodian*
17. Scattered data approximation, *Holger Wendland*
18. Modern computer arithmetic, *Richard Brent & Paul Zimmermann*
19. Matrix preconditioning techniques and applications, *Ke Chen*
20. Greedy approximation, *Vladimir Temlyakov*
21. Spectral methods for time-dependent problems, *Jan Hesthaven, Sigal Gottlieb & David Gottlieb*
22. The mathematical foundations of mixing, *Rob Sturman, Julio M. Ottino & Stephen Wiggins*
23. Curve and surface reconstruction, *Tamal K. Dey*
24. Learning theory, *Felipe Cucker & Ding Xuan Zhou*
25. Algebraic geometry and statistical learning theory, *Sumio Watanabe*
26. A practical guide to the invariant calculus, *Elizabeth Louise Mansfield*

Greedy Approximation

VLADIMIR TEMLYAKOV
University of South Carolina

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town,
Singapore, São Paulo, Delhi, Tokyo, Mexico City

Cambridge University Press
The Edinburgh Building, Cambridge CB2 8RU, UK

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org
Information on this title: www.cambridge.org/9781107003378

First published 2011

Printed in the United Kingdom at the University Press, Cambridge

A catalog record for this publication is available from the British Library

Library of Congress Cataloging in Publication data
Temlyakov, Vladimir, 1953–
Greedy approximation / Vladimir Temlyakov.
p. cm. – (Cambridge monographs on applied and computational mathematics ; 20)
ISBN 978-1-107-00337-8 (hardback)
1. Approximation theory. I. Title. II. Series.
QA221.T455 2011
518′.5–dc23
2011025053

ISBN 978-1-107-00337-8 Hardback

Contents

Preface

From the beginning of time, human beings have been trying to replace complicated with simpler things. From ancient shamans working magic upon clay figures to heal the sick, to Renaissance artists representing God Almighty as a nude painted onto a ceiling, the fundamental problem of representation of the only partially representable continues in contemporary applied mathematics. A generic problem of mathematical and numerical analysis is to represent a given function approximately. It is a classical problem that goes back to the first results on Taylor's and Fourier's expansions of a function.

The first step in solving the representation problem is to choose a representation system. Traditionally, a representation system has natural features such as minimality, orthogonality, simple structure and nice computational characteristics. The most typical representation systems are the trigonometric system $\{e^{ikx}\}$, the algebraic system $\{x^k\}$, the spline system, the wavelet system and their multivariate versions. In general we may speak of a basis $\Psi = \{\psi_k\}_{k=1}^{\infty}$ in a Banach space X.

The second step in solving the representation problem is to choose a form of an approximant that is built on the base of the chosen representation system Ψ. In a classical way that was used for centuries, an approximant a_m is a polynomial with respect to Ψ:

$$a_m := \sum_{k=1}^{m} c_k \psi_k. \tag{1}$$

The complexity of the approximant a_m is characterized by the order m of the polynomial. It is well known in approximation theory that approximation by polynomials is closely related to the smoothness properties of the

function being approximated. Approximation of this type is referred to as *linear approximation theory* because, for a fixed m, approximants come from a linear subspace spanned by $\psi_1, \dots, \psi_m$.

It is understood in numerical analysis and approximation theory that in many problems from signal/image processing it is more beneficial to use an m-term approximant with respect to Ψ than a polynomial of order m. This means that for $f \in X$ we look for an approximant of the form

$$a_m(f) := \sum_{k \in \Lambda(f)} c_k \psi_k, \tag{2}$$

where $\Lambda(f)$ is a set of m indices which is determined by f. The complexity of this approximant is characterized by the cardinality $|\Lambda(f)| = m$ of $\Lambda(f)$. Approximation of this type is referred to as *nonlinear approximation theory* because, for a fixed m, approximants $a_m(f)$ come from different linear subspaces spanned by ψ_k, $k \in \Lambda(f)$, which depend on f. The cardinality $|\Lambda(f)|$ is a fundamental characteristic of $a_m(f)$ called *sparsity* of $a_m(f)$ with respect to Ψ. It is now well understood that we need to study nonlinear sparse representations in order to increase significantly our ability to process (compress, denoise, etc.) large data sets. Sparse representations of a function are not only a powerful analytic tool, but also they are utilized in many applications in image/signal processing and numerical computation.

The third step in solving the representation problem is to choose a method of construction of an approximant of desired form. Let us begin with the linear theory. For example, the approximation method that picks the polynomial of degree m with respect to Ψ of best approximation of f in X as an approximant is an optimal method of approximation by polynomials of degree m. However, such an obvious optimal method of approximation may not be good from the point of view of practical implementation. Standard methods of approximation that are more practical than the above one are linear methods of approximation, in particular partial sums of the corresponding expansion of f with respect to the basis Ψ. Many books (see, for example, DeVore and Lorenz (1993)) discuss this classical topic of approximation theory.

An implementation of the third step in the nonlinear setting is not straightforward. It is clear that an analog of the best polynomial approximant of order m is the best m-term approximant. Nonlinearity brings about complications even at this stage. The existence of a best approximant from a finite dimensional subspace is well known; the existence of a best m-term approximant is a difficult problem. We discuss this problem in Chapter 1. Next, what are nonlinear analogs of, say, partial sums? We answer this question in Chapter 1.

It turns out that greedy approximants are natural substitutes for the partial sums.

We specify not only a form of an approximant, but also choose a specific method of approximation (for instance, the one that is known to be good in practical implementations). Now, we have a precise mathematical problem of studying the efficiency of our specific method of approximation. We discuss this problem in detail here. It turns out that a convenient and flexible way of measuring the efficiency of a specific approximation method is to prove the corresponding Lebesgue-type inequalities. We would like this method to work for all functions. Therefore, it should converge at least for each $f \in X$. Convergence is a fundamental theoretical problem. In this book we thoroughly discuss the problem of convergence of greedy algorithms.

The fundamental question of nonlinear approximation is how to devise good constructive methods (algorithms) of approximation. This problem has two levels of nonlinearity. The first level of nonlinearity (discussed above) is m-term approximation with regard to bases. In this problem one can use the unique function expansion with regard to a given basis to build an approximant. Nonlinearity comes in when we look for m-term approximants with terms (i.e. basis elements) that are allowed to depend on a given function. We discuss m-term approximation with regard to bases in detail in Chapter 1. On the second level of nonlinearity, we replace a basis by a more general system, which is not necessarily minimal (for example, redundant system, or dictionary). This setting is much more complicated than the first one (bases case); however, there is a solid justification of importance of redundant systems in both theoretical questions and in practical applications (see, for example, Donoho (2001), Huber (1985), Schmidt (1906)). In Chapters 2 and 6 we discuss approximation by linear combinations of elements that are taken from a redundant (overcomplete) system of elements. We briefly discuss the question: Why do we need redundant systems? Answering this question, we first of all mention three classical redundant systems that are used in different areas of mathematics.

Perhaps the first example of m-term approximation with regard to redundant dictionary was discussed by Schmidt (1906), who considered the approximation of functions $f(x, y)$ of two variables by bilinear forms $\sum_{i=1}^{m} u_i(x)v_i(y)$ in $L_2([0, 1]^2)$. This problem is closely connected with properties of the integral operator $J_f(g) := \int_0^1 f(x, y)g(y)dy$ with kernel $f(x, y)$.

Another example, which hails from statistics, is the projection pursuit regression problem. In the language of function theory, the problem is to approximate in $L_2(\Omega)$, where $\Omega \subset \mathbb{R}^d$ is a bounded domain, a given function

$f \in L_2(\Omega)$ by a sum of ridge functions, i.e. by $\sum_{j=1}^m r_j(\omega_j \cdot x), \quad x, \omega_j \in \mathbb{R}^d, \quad j = 1, \dots, m$, where r_j, $j = 1, \dots, m$, are univariate functions.

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

Thus, in order to address the contemporary needs of approximation theory and computational mathematics, a very general model of approximation with regard to a redundant system (dictionary) has been considered in many recent papers. As an example of 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 such that the closure of $\operatorname{span} \mathcal{D}$ coincides with X as an approximating system. We would like to have an 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. This requirement is an analog of *on-line* computation that is very desirable in practical algorithms. Clearly, we are looking for good algorithms which converge for each target function. It is not obvious that such an algorithm exists in a setting at the above level of generality (X, $\mathcal{D}$ are arbitrary).

The fundamental question is how to construct good methods (algorithms) of approximation. Recent results have established that greedy type algorithms are suitable methods of nonlinear approximation in both m-term approximation with regard to bases and m-term approximation with regard to redundant systems. It turns out that there is one fundamental principle that allows us to build good algorithms both for arbitrary redundant systems and for very simple well structured bases like the Haar basis. This principle is the use of a greedy step in searching for a new element to be added to a given m-term approximant. By a *greedy step*, 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 the already found m elements of the dictionary. In Chapters 2 and 6 we present different greedy type algorithms beginning with a very simple and very natural Pure Greedy Algorithm in a Hilbert space and ending with its rather complicated modifications in a Banach space. Different modifications aim to make the corresponding greedy

algorithms more ready for practical implementation. We discuss this issue in detail in Chapters 2 and 6.

It is known that in many numerical problems users are satisfied with a Hilbert space setting and do not consider a more general setting in a Banach space. We now make one remark that justifies our interest in Banach spaces. The first argument is an a-priori argument that the spaces L_p are very natural and should be studied along with the L_2 space. The second argument is an a-posteriori argument.

The study of greedy approximation in Banach spaces has revealed that the characteristic of a Banach space X that governs the behavior of greedy approximation is the *modulus of smoothness* $\rho(u)$ of X. It is known that the spaces L_p, $2 \le p < \infty$, have modulo of smoothness of the same order: u^2. Thus, many results that are known for the Hilbert space L_2 and proved using some special structure of a Hilbert space can be generalized to Banach spaces L_p, $2 \le p < \infty$. The new proofs use only the geometry of the unit sphere of the space expressed in the form $\rho(u) \le \gamma u^2$.

The theory of greedy approximation is developing rapidly, and results are spread over hundreds of papers by different authors. There are several surveys that discuss greedy approximation (see DeVore (1998), Konyagin and Temlyakov (2002), Temlyakov (2003a), Temlyakov (2006b), Temlyakov (2008b), Wojtaszczyk (2002a)). This is the first book on greedy approximation. This book is an extension of Temlyakov (2008b). The book provides a systematic presentation of fundamental results in greedy approximation. It also contains an introduction to two hot topics in numerical mathematics: learning theory and compressed sensing. This book possesses features of both a survey paper and a textbook. The majority of results are given with proofs. However, some important results with technically involved proofs are presented without proof. We included proofs of the most important and typical results; and we tried to include those proofs which demonstrate different ideas and are based on different techniques. In this sense the book has a feature of a survey – it tries to cover broad material. On the other hand, we limit ourselves to a systematic treatment of a specific topic rather than trying to give an overview of all related topics. In this sense the book is close to a textbook. There are many papers on theoretical and computational aspects of greedy approximation, learning theory and compressed sensing. We have chosen to cover the mathematical foundations of greedy approximation, learning theory and compressed sensing.

The book is addressed to researchers working in numerical mathematics, analysis, functional analysis and statistics. It quickly takes the reader from classical results to the frontier of the unknown, but is written at the level of a graduate course and does not require a broad background in order to understand the

topics. Graduate students working in different areas of numerical mathematics and analysis may find it useful to learn not only greedy approximation theory, but also theoretical foundations of learning theory and compressed sensing. The author taught three graduate courses, Greedy Approximation, Learning Theory and Compressed Sensing, based on the material of the book, at the University of South Carolina. All three courses were very well accepted by students. The book might be used for designing different graduate courses. It contains a number of important open problems which may assist in uncovering topics for dissertations and research papers.

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

Acknowledgements I am grateful to Jessica L. Nelson for help in the preparation of the book.