Applied and Numerical Harmonic Analysis

$$\hat{f}(\gamma) = \int f(x)e^{-2\pi i x \gamma}\,dx$$

Radu Balan, Matthew J. Begué,
John J. Benedetto, Wojciech Czaja,
Kasso A. Okoudjou, Editors

Excursions in Harmonic Analysis, Volume 3

The February Fourier Talks at the Norbert Wiener Center

Birkhäuser

Birkhäuser

Applied and Numerical Harmonic Analysis

For further volumes:
http://www.springer.com/series/4968

Radu Balan • Matthew J. Begué • John J. Benedetto
Wojciech Czaja • Kasso A. Okoudjou
Editors

Excursions in Harmonic Analysis, Volume 3

The February Fourier Talks at the Norbert Wiener Center

Birkhäuser

Editors

Radu Balan
Department of Mathematics,
University of Maryland
Norbert Wiener Center
College Park
Maryland
USA

Matthew J. Begué
Department of Mathematics,
University of Maryland
Norbert Wiener Center
College Park
Maryland
USA

John J. Benedetto
Department of Mathematics,
University of Maryland
Norbert Wiener Center
College Park
Maryland
USA

Wojciech Czaja
Department of Mathematics,
University of Maryland
Norbert Wiener Center
College Park
Maryland
USA

Kasso A. Okoudjou
Department of Mathematics,
University of Maryland
Norbert Wiener Center
College Park
Maryland
USA

ISSN 2296-5009 ISSN 2296-5017 (electronic)
Applied and Numerical Harmonic Analysis
ISBN 978-3-319-13229-7 ISBN 978-3-319-13230-3 (eBook)
DOI 10.1007/978-3-319-13230-3

Library of Congress Control Number: 2012951313

Mathematics Subject Classification (2010): 26-XX, 35-XX, 40-XX, 41-XX, 42-XX, 43-XX, 44-XX,46-XX, 47-XX, 58-XX, 60-XX, 62-XX, 65-XX, 68-XX, 78-XX, 92-XX, 93-XX, 94-XX

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Dedicated to

Henry J. Landau,

who has set the standard for excellence and creativity in harmonic analysis and its applications.

ANHA Series Preface

The *Applied and Numerical Harmonic Analysis (ANHA)* book series aims to provide the engineering, mathematical, and scientific communities with significant developments in harmonic analysis, ranging from abstract harmonic analysis to basic applications. The title of the series reflects the importance of applications and numerical implementation, but richness and relevance of applications and implementation depend fundamentally on the structure and depth of theoretical underpinnings. Thus, from our point of view, the interleaving of theory and applications and their creative symbiotic evolution is axiomatic.

Harmonic analysis is a wellspring of ideas and applicability that has flourished, developed, and deepened over time within many disciplines and by means of creative cross-fertilization with diverse areas. The intricate and fundamental relationship between harmonic analysis and fields such as signal processing, partial differential equations (PDEs), and image processing is reflected in our state-of-the-art *ANHA* series.

Our vision of modern harmonic analysis includes mathematical areas such as wavelet theory, Banach algebras, classical Fourier analysis, time-frequency analysis, and fractal geometry, as well as the diverse topics that impinge on them.

For example, wavelet theory can be considered an appropriate tool to deal with some basic problems in digital signal processing, speech and image processing, geophysics, pattern recognition, biomedical engineering, and turbulence. These areas implement the latest technology from sampling methods on surfaces to fast algorithms and computer vision methods. The underlying mathematics of wavelet theory depends not only on classical Fourier analysis but also on ideas from abstract harmonic analysis, including von Neumann algebras and the affine group. This leads to a study of the Heisenberg group and its relationship to Gabor systems, and of the metaplectic group for a meaningful interaction of signal decomposition methods. The unifying influence of wavelet theory in the aforementioned topics illustrates the justification for providing a means for centralizing and disseminating information from the broader, but still focused, area of harmonic analysis. This will be a key role of *ANHA*. We intend to publish with the scope and interaction that such a host of issues demands.

Along with our commitment to publish mathematically significant works at the frontiers of harmonic analysis, we have a comparably strong commitment to publish major advances in the following applicable topics in which harmonic analysis plays a substantial role:

Antenna theory
Biomedical signal processing
Digital signal processing
Fast algorithms
Gabor theory and applications
Image processing
Numerical partial differential equations
Prediction theory
Radar applications
Sampling theory
Spectral estimation
Speech processing
Time-frequency and time-scale analysis
Wavelet theory

The above point of view for the *ANHA* book series is inspired by the history of Fourier analysis itself, whose tentacles reach into so many fields.

In the last two centuries Fourier analysis has had a major impact on the development of mathematics, on the understanding of many engineering and scientific phenomena, and on the solution of some of the most important problems in mathematics and the sciences. Historically, Fourier series were developed in the analysis of some of the classical PDEs of mathematical physics; these series were used to solve such equations. In order to understand Fourier series and the kinds of solutions they could represent, some of the most basic notions of analysis were defined, e.g., the concept of "function." Since the coefficients of Fourier series are integrals, it is no surprise that Riemann integrals were conceived to deal with uniqueness properties of trigonometric series. Cantor's set theory was also developed because of such uniqueness questions.

A basic problem in Fourier analysis is to show how complicated phenomena, such as sound waves, can be described in terms of elementary harmonics. There are two aspects of this problem: first, to find, or even define properly, the harmonics or spectrum of a given phenomenon, e.g., the spectroscopy problem in optics; second, to determine which phenomena can be constructed from given classes of harmonics, as done, for example, by the mechanical synthesizers in tidal analysis.

Fourier analysis is also the natural setting for many other problems in engineering, mathematics, and the sciences. For example, Wiener's Tauberian theorem in Fourier analysis not only characterizes the behavior of the prime numbers, but also provides the proper notion of spectrum for phenomena such as white light; this latter process leads to the Fourier analysis associated with correlation functions in filtering and prediction problems, and these problems, in turn, deal naturally with Hardy spaces in the theory of complex variables.

Nowadays, some of the theory of PDEs has given way to the study of Fourier integral operators. Problems in antenna theory are studied in terms of unimodular trigonometric polynomials. Applications of Fourier analysis abound in signal processing, whether with the fast Fourier transform (FFT), or filter design, or the adaptive modeling inherent in time-frequency-scale methods such as wavelet theory. The coherent states of mathematical physics are translated and modulated Fourier transforms, and these are used, in conjunction with the uncertainty principle, for dealing with signal reconstruction in communications theory. We are back to the raison d'être of the *ANHA* series!

University of Maryland
College Park

John J. Benedetto
Series Editor

Preface

The chapters in these Volumes 3 and 4 have at least one author who spoke at the February Fourier Talks during the period 2002–2013 or at the workshop on Phaseless Reconstruction that immediately followed the 2013 February Fourier Talks. Volumes 1 and 2 were limited to the period 2006–2011.

The February Fourier Talks (FFT)

The *FFT* were initiated in 2002 and 2003 as small meetings on harmonic analysis and applications, held at the University of Maryland, College Park. There were no *FFT*s in 2004 and 2005. The Norbert Wiener Center (NWC) for Harmonic Analysis and Applications was founded in 2004 in the Department of Mathematics at the university, and, since 2006, the *FFT* has been organized by the NWC. The *FFT* has developed into a major annual conference that brings together applied and pure harmonic analysts along with scientists and engineers from universities, industry, and government for an intense and enriching 2-day meeting.

The goals of the *FFT* are the following:

- To offer a forum for applied and pure harmonic analysts to present their latest cutting-edge research to scientists working not only in the academic community but also in industry and government agencies;
- To give harmonic analysts the opportunity to hear from government and industry scientists about the latest problems in need of mathematical formulation and solution;
- To provide government and industry scientists with exposure to the latest research in harmonic analysis;
- To introduce young mathematicians and scientists to applied and pure harmonic analysis;
- To build bridges between pure harmonic analysis and applications thereof.

These goals stem from our belief that many of the problems arising in engineering today are directly related to the process of making pure mathematics applicable. The Norbert Wiener Center sees the *FFT* as the ideal venue to enhance this process in a constructive and creative way. Furthermore, we believe that our vision is shared by the scientific community, as shown by the steady growth of the *FFT* over the years.

The *FFT* is formatted as a two-day single-track meeting consisting of 30-min talks as well as the following:

- Norbert Wiener Distinguished Lecturer Series;
- General Interest Keynote Address;
- Norbert Wiener Colloquium;
- Graduate and Postdoctoral Poster Session.

The talks are given by experts in applied and pure harmonic analysis, including academic researchers and invited scientists from industry and government agencies.

The Norbert Wiener Distinguished Lecture caps the technical talks of the first day. It is given by a senior harmonic analyst, whose vision and depth through the years have had profound impact on our field. In contrast to the highly technical day sessions, the Keynote Address is aimed at a general public audience and highlights the role of mathematics, in general, and harmonic analysis, in particular. Furthermore, this address can be seen as an opportunity for practitioners in a specific area to present mathematical problems that they encounter in their work. The concluding lecture of each *FFT*, our Norbert Wiener Colloquium, features a mathematical talk by a renowned applied or pure harmonic analyst. The objective of the Norbert Wiener Colloquium is to give an overview of a particular problem or a new challenge in the field. We include here a list of speakers for these three lectures.

Distinguished Lecturer

- Ronald Coifman
- Ingrid Daubechies
- Ronald DeVore
- Richard Kadison
- Peter Lax
- Elias Stein
- Gilbert Strang

Keynote Address

- Peter Carr
- Barry Cipra
- James Coddington
- Nathan Crone
- Mario Livio
- William Noel
- Steven Schiff
- Mark Stopfer
- Frederick Williams

Colloquium

- Rama Chellappa
- Margaret Cheney
- Charles Fefferman
- Robert Fefferman
- Gerald Folland
- Christopher Heil
- Peter Jones
- Thomas Strohmer
- Victor Wickerhauser

In 2013, the February Fourier Talks were followed by a workshop on phaseless reconstruction, also hosted by the Norbert Wiener Center and intellectually in the spirit of the *FFT*.

The Norbert Wiener Center

The Norbert Wiener Center for Harmonic Analysis and Applications provides a national focus for the broad area of mathematical engineering. Applied harmonic analysis and its theoretical underpinnings form the technological basis for this area. It can be confidently asserted that mathematical engineering will be to today's mathematics departments what mathematical physics was to those of a century ago. At that time, mathematical physics provided the impetus for tremendous advances within mathematics departments, with particular impact in fields such as differential equations, operator theory, and numerical analysis. Tools developed in these fields were essential in the advances of applied physics, e.g., the development of the solid state devices, which now enable our information economy.

Mathematical engineering impels the study of fundamental harmonic analysis issues in the theories and applications of topics such as signal and image processing, machine learning, data mining, waveform design, and dimension reduction into mathematics departments. The results will advance the technologies of this millennium.

The golden age of mathematical engineering is upon us. The Norbert Wiener Center reflects the importance of integrating new mathematical technologies and algorithms in the context of current industrial and academic needs and problems.

The Norbert Wiener Center has three goals:

- Research activities in harmonic analysis and applications;
- Education—undergraduate to postdoctoral;
- Interaction within the international harmonic analysis community.

We believe that educating the next generation of harmonic analysts, with a strong understanding of the foundations of the field and a grasp of the problems arising in applications, is important for a high level and productive industrial, government, and academic workforce.

The Norbert Wiener Center website: **www.norbertwiener.umd.edu.**

The Structure of the Volumes

To some extent the four parts for each of these volumes are artificial placeholders for all the diverse chapters. It is an organizational convenience that reflects major areas in harmonic analysis and its applications, and it is also a means to highlight significant modern thrusts in harmonic analysis. Each part includes an introduction that describes the chapters therein.

Volume 1

I Sampling Theory
II Remote Sensing
III Mathematics of Data Processing
IV Applications of Data Processing

Volume 2

V Measure Theory
VI Filtering
VII Operator Theory
VIII Biomathematics

Volume 3

IX Special Topics in Harmonic Analysis
X Applications and Algorithms in the Physical Sciences
XI Gabor Theory
XII RADAR and Communications: Design, Theory, and Applications

Volume 4

XIII Theoretical Harmonic Analysis
XIV Sparsity
XV Signal Processing and Sampling
XVI Spectral Analysis and Correlation

Acknowledgements

The editors of Volumes 3 and 4 gratefully acknowledge additional editorial assistance by Dr. Alfredo Nava-Tudela as well as the support of Danielle Walker, Associate Editor for Springer-Birkhäuser in New York.

The Norbert Wiener Center also gratefully acknowledges the indispensable support of the following groups: Birkhäuser and Springer Publishers, the IEEE Baltimore Section, MiMoCloud, Inc., Radyn, Inc., the SIAM Washington-Baltimore Section, and SR2 Group, LLC. One of the successes of the February Fourier Talks has been the dynamic participation of graduate student and postdoctoral engineers, mathematicians, and scientists. We have been fortunate to be able to provide travel and living expenses to this group due to continuing, significant grants from the National Science Foundation, which, along with the aforementioned organizations and companies, believes in and supports our vision of the FFT.

Contents

Part IX Special Topics in Harmonic Analysis

Part X Applications and Algorithms in the Physical Sciences

Part XI Gabor Theory

Part XII RADAR and Communications: Design, Theory, and Applications

Part IX
Special Topics in Harmonic Analysis

Part IX consists of two chapters, that are as distinctive as their *FFT* authors are distinguished. GILBERT STRANG spoke at *FFT 2012* and ROBERT S. STRICHARTZ spoke at *FFT 2002*, our first *FFT*; and their two chapters comprise Part I.

STRANG goes deeply into the analysis of the well-known *factorization* $A = LPU$ of an invertible $n \times n$ matrix A, where L is lower triangular, U is upper triangular, and P is a unique permutation matrix. Recall that an $n \times n$ permutation matrix is defined by the condition that it has the entry 1 in each row and in each column and is 0 for all other entries. Natural adjustments of $A = LPU$ lead to the Wiener–Hopf form $A = UPL$ and the Bruhat decomposition $A = U_1 \pi U_2$.

The Wiener–Hopf form is a natural matricial formation of Wiener–Hopf's method to solve systems of integral equations, as well as certain systems of partial differential equations arising in mathematical physics. The basic technique of Wiener and Hopf comes down to defining two complex functions Φ_+ and Φ_-, where Φ_+ (respectively, Φ_-) is analytic in the upper (respectivelylower) half-plane; and U (respectively, L) is the analogue of Φ_+ (respectively, Φ_-).

The factorization, $A = LPU$, for an $n \times n$ matrix A, can be viewed classically in terms of an interpretation of the Gaussian elimination method of solving n linear equations in n unknowns. Strang's main results deal with elimination on banded doubly infinite matrices. His insights are pivotal (sic) and magisterial, and his examples are truly enlightening.

BELLO, LI, AND STRICHARTZ outline a Hodge-de Rham theory of k-forms ($k = 0, 1, 2$) on a Sierpiǹski carpet. A wonderful aspect of this paper is that Strichartz' two co-authors were undergraduates at the time of the research, and were part of Strichartz' now famous Research Experience for Undergraduates (REU) program sponsored by the National Science Foundation (NSF).

The Sierpiǹski carpet is a fractal and the authors approximate it by a sequence of graphs, use classical Hodge-de Rham theory on each graph, and take the limit.

The interplay of mathematical and computational tools is labyrinthine and fascinating. The Sierpiǹski carpet, SC, itself is defined in terms of similarity maps, F_j, of contraction ratio 1/3. It is the analogue in $\mathbb{R}^2$ of the 1/3-Cantor set. However, SC is a connected set, as well as being compact with Lebesgue measure 0. Further, it cuts the plane into infinitely many disjoint parts. There is a natural way to associate a sequence of graphs, Γ_m, to $\{F_j\}_{j=1}^{\infty}$, and ultimately to define the associated de Rham complex for each m. Then, there is extensive experimentation and the definition of 0-, 1-, and 2-forms on the SC and the analysis of the corresponding Laplacians. For example, the Laplacian $-\Delta_0^{(m)}$ for the 0-form is exactly the graph Laplacian of Γ_m with specified weights on the vertices and edges—all very heady-stuff, quite like de Rham's challenges to the lofty Alps.

Part of analysis, some of which is open-ended and important, is the characterization of the spectra of the Laplacian on k-forms. Some of the fascination is the tantalizing possible relationship with the role of Laplacians on graphs with regard to current interest in dimension reduction as related to "big data."

The Algebra of Elimination

Gilbert Strang

Abstract Elimination with only the necessary row exchanges will produce the triangular factorization $A = LPU$, with the (unique) permutation P in the middle. The entries in L are reordered in comparison with the more familiar $\boldsymbol{A} = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{L}}\widehat{\boldsymbol{U}}$ (where $\widehat{\boldsymbol{P}}$ is not unique). Elimination with three other starting points 1, n and n, n and n, 1 produces three more factorizations of A, including the Wiener–Hopf form $\boldsymbol{UPL}$ and Bruhat's $U_1\,\pi\,U_2$ with two upper triangular factors.

All these starting points are useless for doubly infinite matrices. The matrix has no first or last entry. When A is banded and invertible, we look for a new way to establish $A = LPU$. The key is to locate the pivot rows (we also find the main diagonal of A). LPU was previously known in the periodic (block Toeplitz) case $A(i,j) = A(i+b, j+b)$, by factoring a matrix polynomial.

Keywords Factorization · Elimination · Banded matrix · Infinite matrix · Bruhat · Wiener · Hopf

1 Introduction

The "pedagogical" part of this chapter presents the LPU factorization of an invertible n by n matrix A :

$$A = LPU = (\text{lower triangular})\,(\text{permutation})\,(\text{upper triangular}).$$

The reader may feel that everything has been said about the algebra of elimination, which produces L, P, and U. This is potentially true. But who said it, and where, is not easy to discover. I hope you will feel that some of this is worth saying again.

G. Strang (✉)
Massachusetts Institute of Technology, Cambridge, MA, USA
e-mail: gilstrang@gmail.com

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_1

The LPU form that algebraists like best (with P in the middle instead of the more practical $\boldsymbol{A} = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{L}}\widehat{\boldsymbol{U}}$) is the least familiar within SIAM.

Once started in this direction, factorizations continue to appear. If elimination begins at the last entry A_{nn} and works upward, the result is $\boldsymbol{UPL}$. Those are new factors of A, and there must be relations to the original L, P, and U that we do not know. More inequivalent forms $A = U_1 \pi U_2$ and $A = L_1 \pi L_2$ come from starting elimination at A_{n1} and at A_{1n}. You may be surprised that the all-time favorite of algebraists is Bruhat's $U_1 \pi U_2$: hard to comprehend (but see Sect. 4).

$$\begin{tikzcd}
\text{(down and right) } A = LPU & A_{11} \arrow[l] & A_{1n} \arrow[r] & A = L_1\, \pi\, L_2 \text{ (down and left)} \\
\text{(up and right) } A = U_1\, \pi\, U_2 & A_{n1} \arrow[l] & A_{nn} \arrow[r] & A = \boldsymbol{UPL} \text{ (up and left)}
\end{tikzcd}$$

The more original part of this chapter extends $A = LPU$ to *banded doubly infinite matrices*. What makes this challenging is that elimination has no place to begin. A_{11} is deep in the middle of A, and algebra needs help from analysis. The choice of pivot appears to depend on infinitely many previous choices. The same difficulty arose for Wiener and Hopf, because they wanted $A = \boldsymbol{UL}$ and singly infinite matrices have no last entry A_{nn}. This was overcome in the periodic (block Toeplitz) case, and in Sect. 6 we go further.

2 The Uniqueness of P in $A = LPU$

Theorem 1. *The permutation P in $A = LPU$ is uniquely determined by A.*

Proof. Consider the s by t upper left submatrices of A and P. That part of the multiplication $A = LPU$ leads to $a = \ell pu$ for the submatrices, because L and U are triangular :

$$\begin{bmatrix} a & * \\ * & * \end{bmatrix} = \begin{bmatrix} \ell & 0 \\ * & * \end{bmatrix} \begin{bmatrix} p & * \\ * & * \end{bmatrix} \begin{bmatrix} u & * \\ 0 & * \end{bmatrix} \text{ gives } a = \ell pu. \tag{1}$$

The submatrix ℓ is s by s and u is t by t. Both have nonzero diagonals (therefore invertible) since they come from the invertible L and U. Then *p has the same rank as $a = \ell pu$.* The ranks of all upper left submatrices p are determined by A, so the whole permutation P is uniquely determined [7, 8, 14].

The number of 1's in p is its rank, since those 1's produce independent columns (they come from different rows of P). The rule is that $P_{ik} = 1$ exactly where the rank of the upper left submatrices a_{ik} of A increases :

$$\text{rank } a_{ik} = 1 + \text{rank } a_{i-1,k-1} = 1 + \text{rank } a_{i-1,k} = 1 + \text{rank } a_{i,k-1}. \tag{2}$$

In words, row i is dependent on previous rows until column k is included, and column k is dependent on previous columns until row i is included. When $A = LPU$ is constructed by elimination, a pivot will appear in this i, k position. The pivot row $i(k)$ for elimination in column k will be the first row (the smallest i) such that (2) becomes true. Since by convention rank $p_{i0} =$ rank $p_{0k} =$ rank $a_{i0} =$ rank $a_{0k} = 0$, the first nonzero in column 1 and in row 1 of A will determine $P_{i1} = 1$ and $P_{1k} = 1$.

In case the leading square submatrices a_{ii} are all nonsingular, which leads to rank $(a_{ik}) = \min\,(i, k)$, rule (2) puts all pivots on the diagonal : $P_{ii} = 1$. This is the case $P = I$ with no row exchanges and $A = LU$.

Elimination by columns produces the same pivot positions (in a different sequence) as elimination by rows. For elimination with different starting points, and also for infinite matrices, rule (2) is to be adjusted. Determining P so simply from (1) is all-important.

We describe below how P can jump when A changes smoothly.

3 The Algebra of Elimination : $A = LPU = \widehat{P}\widehat{L}\widehat{U}$

Suppose elimination starts with $A_{11} \neq 0$, and all leading submatrices a_{kk} are invertible. Then we reach $A = LU$ by familiar steps. For each $j > 1$, subtract a multiple ℓ_{j1} of row 1 from row j to produce zero in the $j, 1$ position. The next pivot position 2, 2 now contains the nonzero entry det $(a_{22})/$ det (a_{11}) : this is the second pivot.

Subtracting multiples ℓ_{j2} of that second row produces zeros below the pivot in column 2. For $k = 1, \ldots, n$, the kth pivot row becomes row k of U. The k, k pivot position contains the nonzero entry det $(a_{kk})/$ det $(a_{k-1,k-1})$. For lower rows $j > k$, subtracting a multiple ℓ_{jk} of row k from row j produces zero in the j, k position. Then the "magic of elimination" is that the matrix L of multipliers ℓ_{jk} times the matrix U of pivot rows equals the original matrix A. Suppose $n = 3$:

$$A = LU \qquad \begin{bmatrix} \text{row 1 of A} \\ \text{row 2 of A} \\ \text{row 3 of A} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ \ell_{21} & 1 & 0 \\ \ell_{31} & \ell_{32} & 1 \end{bmatrix} \begin{bmatrix} \text{row 1 of U} \\ \text{row 2 of U} \\ \text{row 3 of U} \end{bmatrix}. \tag{3}$$

The third row of that LU multiplication correctly states that

$$\text{row 3 of } U = (\text{row 3 of } A) - \ell_{31}(\text{row 1 of } U) - \ell_{32}(\text{row 2 of } U). \tag{4}$$

Now we face up to the possibility of zeros in one or more pivot positions. If a_{kk} is the first square upper left submatrix to be singular, the steps must change when elimination reaches column k. A lower row $i(k)$ must become the kth pivot row. Based on the current matrix, we have an algebraic choice and an algorithmic choice :

Algebraic

Choose the first row $i(k)$ that is not already a pivot row and has a nonzero entry in column k (to become the kth pivot). Subtract multiples of this pivot row $i(k)$ to produce zeros in column k of all lower nonpivot rows. This completes step k.

Note

For $A = LPU$, the pivot row $i(k)$ is *not moved immediately into row k* of the current matrix. It will indeed be row k of U, but it waits for the permutation P (with $P_{i(k),k} = 1$) to put it there.

Algorithmic

Choose any row $I(k)$ that is not already a pivot row and has a nonzero entry in column k. Our choice of $I(k)$ may maximize that pivot entry, or not. **Exchange** this new pivot row $I(k)$ with the current row k. Subtract multiples of the pivot row to produce zeros in column k of all later rows.

Note

This process normally starts immediately at column 1, by choosing the row $I(1)$ that maximizes the first pivot. Each pivot row $I(k)$ moves immediately into row k of the current matrix and also row k of U.

The algebraic choice will lead to $A = LPU$ and the algorithmic choice to $A = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{L}}\widehat{\boldsymbol{U}}$. If the choices coincide, so $I(k) = i(k)$, the multipliers will be the same numbers—but they appear in different positions in L and $\widehat{\boldsymbol{L}}$ because row $I(k)$ has been moved into row k. Then $\widehat{\boldsymbol{P}} = P$ and $\widehat{\boldsymbol{U}} = U$ and $\widehat{\boldsymbol{L}} = P^{-1}LP$ from the reordering of the rows. It is more than time for an example.

Example 1. The first pivot of A is in row $i(1) = 2$. The only elimination step is to subtract ℓ times that first pivot row from row 3. This reveals the second pivot in row $i(2) = 3$. The order of pivot rows is $2,3,1$ (and during LPU elimination they stay in that order!) :

$$A = \begin{bmatrix} 0 & 0 & 3 \\ 1 & a & b \\ \ell & \ell a+2 & \ell b+c \end{bmatrix} \overset{L^{-1}}{\rightarrow} \begin{bmatrix} 0 & 0 & 3 \\ 1 & a & b \\ 0 & 2 & c \end{bmatrix} = PU. \qquad (5)$$

The permutation P has 1's in the pivot positions. So its columns come from the identity matrix in the order 2, 3, 1 given by $i(k)$. Then U is upper triangular:

$$\begin{bmatrix} 0\,0\,3 \\ 1\,a\,b \\ 0\,2\,c \end{bmatrix} = \begin{bmatrix} 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \end{bmatrix} \begin{bmatrix} 1\,a\,b \\ 0\,2\,c \\ 0\,0\,3 \end{bmatrix} = PU. \tag{6}$$

The lower triangular L adds ℓ times row 2 of PU back to row 3 of PU. That entry $L_{32} = \ell$ recovers the original A from PU : the factorization is complete.

$$A = \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,\ell\,1 \end{bmatrix} \begin{bmatrix} 0\,0\,3 \\ 1\,a\,b \\ 0\,2\,c \end{bmatrix} = L(PU) = LPU. \tag{7}$$

Back to algebra.

Consider $A = \widehat{\boldsymbol{P}}\widehat{\boldsymbol{L}}\widehat{\boldsymbol{U}}$ with no extra row exchanges: $I(k) = i(k)$. Then $\widehat{\boldsymbol{P}}$ and $\widehat{\boldsymbol{U}}$ are the same as P and U in the original $A = LPU$. But the lower triangular $\widehat{\boldsymbol{L}}$ is different from L. In fact $P\widehat{\boldsymbol{L}} = LP$ tells us directly that $\widehat{\boldsymbol{L}} = P^{-1}LP$. *This reordered matrix* $\widehat{\boldsymbol{L}}$ *is still lower triangular.* It is this crucial property that uniquely identifies the specific L that is constructed by elimination. Other factors L can enter into $A = LPU$, but only the factor produced by elimination is "reduced from the left" with $P^{-1}LP$ also lower triangular.

The uniqueness of this particular L is illustrated by an example with many possible L's in $A = LPU$:

$$A = \begin{bmatrix} 0 & 1 \\ 1 & a \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \ell & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & u \\ 0 & 1 \end{bmatrix} \text{ provided } a = \ell + u. \tag{8}$$

Row 2 must be the first pivot row. There are no rows below that pivot row; the unique "reduced from the left" matrix is $L = I$ with $\ell = 0$. (And $P^{-1}IP = I$ is lower triangular as required.) To emphasize: All nonzero choices of ℓ are permitted in $A = LPU$ by choosing $u = a - \ell$. But that nonzero entry ℓ will appear *above* the diagonal in $P^{-1}LP$. Elimination produced $\ell = 0$ in the unique reduced factor L.

The difference between L and $\widehat{\boldsymbol{L}}$ in $A = LPU$ and $A = P\widehat{L}U$ can be seen in the 3 by 3 example. Both L and $\widehat{\boldsymbol{L}} = P^{-1}LP$ come from elimination, they contain the same entries, but these entries are moved around when P comes first in $A = P\widehat{L}U$.

Example (Continued). $\widehat{L}$ comes from elimination when the pivot rows of A are moved into 1, 2, 3 order in $\widehat{\boldsymbol{A}} = (\mathbf{inv}\boldsymbol{P})A$:

$$\widehat{\boldsymbol{A}} = \begin{bmatrix} 1 & a & b \\ \ell & \ell a + 2 & \ell b + c \\ 0 & 0 & 3 \end{bmatrix} \xrightarrow{\widehat{\boldsymbol{L}}^{-1}} \begin{bmatrix} 1\,a\,b \\ 0\,2\,c \\ 0\,0\,3 \end{bmatrix} = \boldsymbol{U}.$$

We subtracted ℓ times row 1 from row 2, and $\widehat{\boldsymbol{L}}$ adds it back :

$$\widehat{\boldsymbol{L}} = \begin{bmatrix} 1 & 0 & 0 \\ \ell & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

This agrees with (7) after the reordering $P^{-1}LP$. The nonzero entry is still below the diagonal, confirming that the L chosen earlier is "reduced from the left." No elimination steps were required to achieve zeros in the 3, 1 and 3, 2 positions, so $\widehat{L}_{31} = \widehat{L}_{32} = 0$. In terms of the original A rather than the reordered $\widehat{A}$, $\widehat{L}_{jk} = 0$ *when* $i(j) < i(k)$.

To summarize :

$A = LPU$ has a unique P, and a unique L reduced from the left. The permutation in $A = \widehat{P}\widehat{L}\widehat{U}$ is not unique. But if we exchange rows only when necessary to avoid zeros in the pivot positions, $\widehat{P}$ will agree with P and $\widehat{U} = U$. The lower triangular $\widehat{L}$ in this better known form is $P^{-1}LP$.

Elimination by Column Operations

To anticipate factorizations that are coming next, it is valuable (and satisfying) to recognize that "column elimination" is equally valid. In this brief digression, multiples of columns are subtracted from later columns. The result will be a *lower* triangular matrix L_c. Those column operations use *upper* triangular matrices multiplying from the right. The operations are inverted by an upper triangular matrix U_c. The quick way to see all steps is to transpose A, factor as usual by row operations, and transpose back.

When the pivot columns come in the natural order 1, 2, 3, elimination by columns produces $A = L_cU_c$. This is identical to $A = LU$ from row operations, except that the pivots now appear in L_c. When we factor out the diagonal matrix D of pivots, the uniqueness of L and U (from rows) establishes the simple link to L_c and U_c from columns :

$$A = L_cU_c = (L_cD^{-1})(DU_c) = LU. \tag{9}$$

In our 3 by 3 example, the first pivot (nonzero entry in row 1) is in column $k(1) = 3$. Then the second pivot (nonzero in the current row 2) is in column $k(2) = 1$. Column operations clear out row 2 in the last pivot column $k(3) = 2$:

$$A = \begin{bmatrix} 0 & 0 & 3 \\ 1 & a & b \\ \ell & \ell a + 2 & \ell b + c \end{bmatrix} \overset{U_c^{-1}}{\rightarrow} \begin{bmatrix} 0 & 0 & 3 \\ 1 & 0 & b \\ \ell & 2 & \ell b + c \end{bmatrix} = L_cP_c. \tag{10}$$

The permutation P_c has the *rows* of the identity matrix in the order 3, 1, 2 given by $k(i)$. Then L_c is lower triangular :

$$\begin{bmatrix} 0 & 0 & 3 \\ 1 & 0 & b \\ \ell & 2 & \ell b+c \end{bmatrix} = \begin{bmatrix} 3 & 0 & 0 \\ b & 1 & 0 \\ \ell b+c & \ell & 2 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = L_c P_c. \tag{11}$$

The constantly alert reader will recognize that $k(i)$ is the inverse of $i(k)$. The permutation P_c must agree with P by uniqueness. The factorization $A = L_c P_c U_c$ is completed when U_c undoes the column elimination by adding a times column 1 back to column 2 :

$$A = \begin{bmatrix} 0 & 0 & 3 \\ 1 & 0 & b \\ \ell & 2 & \ell b+c \end{bmatrix} \begin{bmatrix} 1 & a & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = (L_c P_c) U_c = L_c P_c U_c. \tag{12}$$

This matrix U_c is "reduced from the right" because $P_c U_c P_c^{-1}$ is still upper triangular. Under this condition the factors in $A = L_c P_c U_c$ are uniquely determined by A.

When P_c moves to the right, we can do extra column exchanges for the sake of numerical stability. If $|\, a \,| > 1$ in our example, columns 1 and 2 would be exchanged to keep the multiplier below 1.

In the 2 by 2 example, elimination using columns would move the nonzero entry from U (earlier) into L_c (now) :

$$\begin{bmatrix} 0 & 1 \\ 1 & a \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ a & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = L_c P_c U_c. \tag{13}$$

To summarize :

Column elimination can produce different triangular factors from row elimination, but L still comes before $\widehat{U}$. In production codes, the practical difference would come from the time to access columns instead of rows.

4 Bruhat Decomposition and Bruhat Order

Choosing the $1,1$ entry as the starting point of elimination seems natural. Probably the Chinese who first described the algorithm [13, 22] felt the same. A wonderful history [11] by Grcar describes the sources from antiquity and then Newton's "extermination" algorithm. (In lecture notes that he didn't want published, Newton anticipated Rolle and Gauss.) But an algebraist can prefer to start at $(n, 1)$, and a hint at the reason needs only a few words.

$A = LPU$ is built on two subgroups (lower triangular and upper triangular) of the group GL_n of invertible n by n real matrices. There is an underlying equivalence relation : $A \sim B$ if $A = LBU$ for some triangular L and U. Thus GL_n is partitioned into equivalence classes. Because P was unique in Theorem 1, each equivalence class contains exactly one permutation (from the symmetric group S_n of all permutations). Very satisfactory but not perfect.

Suppose the two subgroups are the same (say the invertible upper triangular matrices). Now $A \sim B$ means $A = U_1BU_2$ for some U_1 and U_2. Again GL_n is partioned into (new) equivalence classes, called "double cosets." Again there is a single permutation matrix π in each double coset from $A = U_1\,\pi\,U_2$. But now that the original subgroups are the same (here is the obscure hint, not to be developed further) we can multiply the double cosets and introduce an underlying algebra. The key point is that this "Bruhat decomposition" into double cosets $U\,\pi\,U$ succeeds for a large and important class of algebraic groups (not just GL_n).

Actually Bruhat did not prove this. His 1954 note [3] suggested the first ideas, which Harish-Chandra proved. Then Chevalley [5] uncovered the richness of the whole structure. George Lusztig gave more details of this (ongoing!) history in his lecture [16] at the Bruhat memorial conference in Paris.

One nice point, perhaps unsuspected by Bruhat, was the intrinsic partial order of the permutations π. Each π is shared by all the matrices $U_1\pi U_2$ in its double coset. We might expect the identity matrix $\pi = I$ to come first in the "Bruhat order" but instead it comes last. For a generic n by n matrix, the permutation in $A = U_1\pi U_2$ will be the reverse identity matrix $\pi = J$ corresponding to $(n,\ldots,1)$. Let me connect all these ideas to *upward elimination* starting with the n, 1 entry of A.

The first steps subtract multiples of row n from the rows above, to produce zeros in the first column (above the pivot A_{n1}). Assume that no zeros appear in the pivot positions along the reverse diagonal from n, 1 to 1, n. Then upward elimination ends with zeros above the reverse diagonal :

$$\begin{bmatrix} & & \circledast \\ & \circledast & * \\ \circledast & * & * \end{bmatrix} = \begin{bmatrix} & & 1 \\ & 1 & \\ 1 & & \end{bmatrix} \begin{bmatrix} \circledast & * & * \\ & \circledast & * \\ & & \circledast \end{bmatrix} = JU_2. \tag{14}$$

The upward elimination steps are taken by upper triangular matrices. Those are inverted by an upper triangular U_1 (containing all the multipliers). *This generic case has produced* $A = U_1JU_2$.

At stage k of Bruhat elimination, the pivot row is the *lowest* row that begins with exactly $k-1$ zeros. Then that stage produces zeros in column k for all other rows that began with $k-1$ zeros. These upward elimination steps end with a matrix $\pi\,U_2$, where the permutation π is decided by the order of the pivot rows. The steps are inverted by U_1, so the product $U_1\pi U_2$ recovers the original A and gives its Bruhat decomposition.

In the Bruhat partial order, the reverse identity J comes first and I comes last. The permutations P in $A = LPU$, from elimination that starts with A_{11}, fall naturally in the opposite order. These orders can be defined in many equivalent ways, and this is not the place for a full discussion. But one combinatorial definition fits perfectly with our "rank description" of the pivot positions in Eq. (2) :

In the Bruhat order for LPU decomposition (elimination starting at A_{11}), two permutations have $P \leq P'$ when all their upper left s by t submatrices have rank $(p_{st}) \geq$ rank (p'_{st}).

Example 2. $A_n = \begin{bmatrix} 1/n & 1 \\ 1 & 0 \end{bmatrix}$ has $P_n = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$ but in the limit $A_\infty = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} = P_\infty$. Here $P_n < P_\infty$.

The rank of the upper left 1 by 1 submatrix of A_n drops to zero in the limit A_∞. Our (small) point is that this semicontinuity is always true : ranks can drop but not rise. The rank of a limit matrix never exceeds the limit (or lim inf) of the ranks. The connection between rank and Bruhat order leads quickly to a known conclusion about the map $P(A)$ from A in GL_n to P in S_n :

Theorem 2. *Suppose $A_n = L_n P_n U_n$ approaches $A_\infty = L_\infty P_\infty U_\infty$ and the permutations P_n approach a limit P. Then $P \leq P_\infty$ in the Bruhat order for LPU (reverse of the usual Bruhat order for the π's in $U_1 \pi U_2$).*

Roughly speaking, A_∞ may need extra row exchanges because ranks can drop.

5 Singly Infinite Banded Matrices

Our first step toward new ideas is to allow infinite matrices. We add the requirement that *the bandwidth w is finite* : $A_{ij} = 0$ if $\mid i - j \mid > w$. Thus A is a "local" operator. Each row has at most $2w+1$ nonzeros. Each component in the product Ax needs at most $2w+1$ multiplications.

To start, assume that no finite combination of rows or of columns produces the zero vector (except the trivial combination). Elimination can begin at the $1,1$ position and proceed forever. The output is a factorization into $A = LPU$. *Those three factors are banded, but L and U are not necessarily bounded.*

An example will show how far we are from establishing that L and U are bounded. A is *block diagonal* and each block B_k of A factors into $L_k U_k$ with $P_k = I$:

$$B_k = \begin{bmatrix} \varepsilon_k & -1 \\ 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ \varepsilon_k^{-1} & 1 \end{bmatrix} \begin{bmatrix} \varepsilon_k & -1 \\ 0 & \varepsilon_k^{-1} \end{bmatrix} = L_k U_k. \tag{15}$$

If ε_k approaches zero in a sequence of blocks of A, the pivots ε_k and ε_k^{-1} approach zero and infinity. The block diagonal matrices L and U (with blocks L_k and U_k) are unbounded. At the same time A is bounded with bounded inverse :

$$\text{The blocks in } A^{-1} \text{ are } B_k^{-1} = \begin{bmatrix} 0 & 1 \\ 1 & -\varepsilon_k \end{bmatrix}.$$

To regain control, *assume in the rest of this section that A is* **Toeplitz** *or* **block Toeplitz**. This time invariance or shift invariance is expressed by $A_{ij} = A_{j-i}$. The scalars or square blocks A_k are repeated down the kth diagonal. It would be hard to overstate the importance of Toeplitz matrices. They can be finite or infinite—in many ways doubly infinite is the simplest of all.

Examples will bring out the intimate link between the matrix A and its symbol $a(z)$, the polynomial in z and z^{-1} with coefficients A_k. Suppose A is tridiagonal $(w=1)$:

$$A = \begin{bmatrix} 5 & -2 & & \\ -2 & 5 & -2 & \\ & -2 & 5 & \bullet \\ & & \bullet & \bullet \end{bmatrix} \quad \text{corresponds to } a(z) = -2z^{-1} + 5 - 2z.$$

With $z = e^{i\theta}$, the symbol $a(e^{i\theta})$ becomes $5 - 4\cos\theta$. This is positive for all θ, so A is positive definite. The symbol factors into $a(z) = (2-z)(2-z^{-1}) = u(z)\ell(z)$. The singly infinite matrix factors in the same way (and notice U before L):

$$\boldsymbol{A} = \begin{bmatrix} 2 & -1 & & \\ & 2 & -1 & \\ & & 2 & -1 \\ & & & \bullet \end{bmatrix} \begin{bmatrix} 2 & & & \\ -1 & 2 & & \\ & -1 & 2 & \\ & & -1 & \bullet \end{bmatrix} = \boldsymbol{UL}. \tag{16}$$

This was a **spectral factorization** of $a(z)$, and a **Wiener–Hopf factorization** $A = UL$.

When elimination produces $A = LU$ by starting in the $1,1$ position, the result is much less satisfying: L and U are not Toeplitz. They are asymptotically Toeplitz and their rows eventually approach the good factors UL.

One key point is that $A = UL$ does not come from straightforward elimination—because an infinite matrix has no corner entry A_{nn} to start upward elimination. We factored $a(z)$ instead.

Another key point concerns the location of the zeros of $u(z) = 2 - z$ and $\ell(z) = 2 - z^{-1}$. Those zeros $z = 2$ and $z = 1/2$ satisfy $|\, z \,| > 1$ and $|\, z \,| < 1$ respectively. Then L and U have bounded inverses, and those Toeplitz inverses correspond to $1/\ell(z)$ and $1/u(z) = 1/(2-z) = \frac{1}{2} + \frac{1}{4}z + \frac{1}{8}z^2 + \cdots$.

If we had chosen the factors badly, $u(z) = 1 - 2z$ and $\ell(z) = 1 - 2z^{-1}$ still produce $a = u\ell$ and $A = UL$:

$$\boldsymbol{A} = \begin{bmatrix} 1 & -2 & & \\ & 1 & -2 & \\ & & 1 & -2 \\ & & & \bullet \end{bmatrix} \begin{bmatrix} 1 & & & \\ -2 & 1 & & \\ & -2 & 1 & \\ & & -2 & \bullet \end{bmatrix} = \boldsymbol{UL}. \tag{17}$$

The formal inverses of U and L have $1, 2, 4, 8, \ldots$ on their diagonals, because the zeros of $u(z)$ and $\ell(z)$ are inside and outside the unit circle—the wrong places.

Nevertheless U in (17) is a useful example. It has $x = \left(1, \frac{1}{2}, \frac{1}{4}, \dots\right)$ in its nullspace : $Ux = 0$ because $u\left(\frac{1}{2}\right) = 0$. This is a *Fredholm matrix* because the nullspaces of U and U^T are finite-dimensional. Notice that $U^{\mathrm{T}} = L$ has nullspace $= \{0\}$. The *Fredholm index* is the difference in the two dimensions :

index (U) = dim (nullspace of U) $-$ dim (nullspace of U^{T})

index (L) $= 1 - 0$.

The index of L in (17) is -1; the two nullspaces are reversed. The index of the product $A = UL$ is $1 - 1 = 0$. In fact A is invertible, as the good factorization shows :

$$Ax = b \quad \text{is solved by} \quad x = A^{-1}b = L^{-1}(U^{-1}b). \tag{18}$$

The key to invertibility is $a(z) = u(z)\ell(z)$, with the correct location of zeros to make U and L and thus $A = UL$ invertible. The neat way to count zeros is to use the winding number of $a(z)$.

Theorem 3. *If $a(z) = \Sigma\, A_k z^k$ starts with $A_{-m} z^{-m}$ and ends with $A_M z^M$, we need M zeros with $\mid z \mid > 1$ and m zeros with $\mid z \mid < 1$ (and no zeros with $\mid z \mid = 1$). Then $a(z) = u(z)\,\ell(z)$ and $A = UL$ and those factors are invertible.*

The matrix case is harder. A is now *block Toeplitz*. The A_k that go down diagonal k are square matrices, say b by b. It is still true (and all-important) that complete information about the operator A is contained in the matrix polynomial $a(z) = \Sigma\, A_k z^k$. The factorization of $a(z)$ remains the crucial problem, leading as before to $A = UL$. Again this achieves "upward elimination without a starting point A_{nn}."

The appropriate form for a matrix factorization is a product $up\ell$:

$$a(z) = u(z)p(z)\ell(z) \text{with} p(z) = \operatorname{diag}\left(z^{k(1)}, \dots, z^{k(b)}\right).$$

The polynomial factor $u(z)$ gives the banded upper triangular block Toeplitz matrix U. The third factor $\ell(z)$ is a polynomial in z^{-1} and it produces L. The diagonal $p(z)$ yields a block Toeplitz matrix P. (It will be a permutation matrix in the doubly infinite case, and we reach $A = UPL$.) The diagonal entry $z^{k(j)}$ produces a 1 in the jth diagonal entry of the block P_k of P.

Example 3. Suppose the $up\ell$ factorization of $a(z)$ has $\ell(z) = I$:

$$a(z) = \begin{bmatrix} z^{-1} & 0 \\ 1 & z \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ z & 1 \end{bmatrix} \begin{bmatrix} z^{-1} & 0 \\ 0 & z \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}. \tag{19}$$

For *doubly* infinite block Toeplitz matrices, this gives $A = UPL$ with $L = I$. Then A is invertible. But for *singly* infinite matrices, the first row of UPL is zero. You see success in rows 3–4, 5–6, ... which are not affected by the truncation to this singly infinite UPL with $L = I$:

$$\begin{bmatrix} 1&0&0&0&&&& \\ 0&1&1&0&&&& \\ &&1&0&0&0&& \\ &&0&1&1&0&& \\ &&&&1&0&0&0 \\ &&&&0&1&1&0 \end{bmatrix} \begin{bmatrix} 0&0&0&0&&&& \\ 0&0&0&1&&&& \\ 1&0&0&0&0&0&& \\ 0&0&0&0&0&1&& \\ &&1&0&0&0&0&0 \\ &&0&0&0&0&0&1 \end{bmatrix}$$

$$= \begin{bmatrix} 0&0&0&0&&&& \\ 1&0&0&1&&&& \\ 1&0&0&0&0&0&& \\ 0&0&1&0&0&1&& \\ &&1&0&0&0&0&0 \\ &&0&0&1&0&0&1 \end{bmatrix} \begin{array}{l} \\ \\ \text{rows } 3-4 \text{ of A} \\ \\ \text{rows } 5-6 \text{ of A.} \\ \\ \end{array}$$

The missing nonzero in row 1 comes from the entry z^{-1} in $p(z)$. Invertibility of A in the singly infinite case requires all the exponents in $p(z)$ to be $k(j) = 0$. Those “partial indices” give the dimensions of the nullspaces of A and A^{T} (here 1 and 1). Invertibility in the doubly infinite case only requires $\Sigma\, k(j) = 0$. In both cases this sum is the Fredholm index of A (here 0), equal to the winding number of det $a(z)$.

The matrix factorization $a(z) = u(z)p(z)\,\ell(z)$ has a long and very distinguished history. The first success was by Plemelj [19] in 1908. Hilbert and G.D. Birkhoff contributed proofs. Wiener and Hopf found wide applications to convolution equations on a half-line, by factoring A into UL when $P = I$. The algebraic side was developed by Grothendieck, and the analytic side by the greatest matrix theorist of the twentieth century: Israel Gohberg. My favorite reference, for its clarity and its elementary constructive proof, is by Gohberg, Kaashoek, and Spitkovsky [10].

In the banded *doubly infinite* case, a bounded (and block Toeplitz) inverse only requires that $a(z)$ is invertible on the unit circle: det $a(z) \neq 0$ for $\mid z \mid = 1$. Then $a = up\ell$ and the reverse factorization into ℓpu give $A = UPL$ and $A = LPU$ with invertible block Toeplitz matrices. P and P are permutations of the integers.

All these are examples of triangular factorizations *when elimination has no starting point*. We presented them as the most important examples of their kind—when the periodicity of A reduced the problem to factorization of the matrix polynomial $a(z)$.

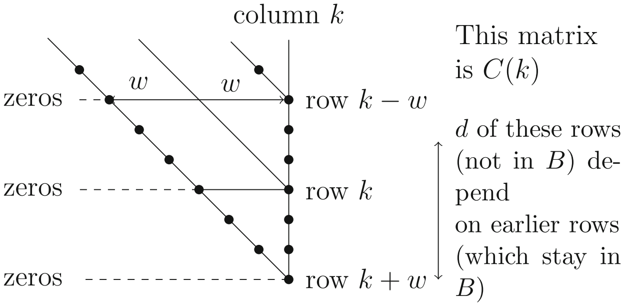
6 Elimination on Banded Doubly Infinite Matrices

We have reached the question that you knew was coming. *How can elimination get started on a doubly infinite matrix*? To produce zeros in column k, $-\infty < k < \infty$, we must identify the number $i(k)$ of the pivot row. When that row is ready for use, its entries before column k are all zero. Multiples ℓ_{ji} of this row are subtracted from lower rows $j > i$, to produce zeros below the pivot in column k of PU. The pivot row has become row $i(k)$ of PU, and it will be row k of U.

Clearly $i(k) \leq k+w$, since all lower rows of a matrix with bandwidth w are zero up to and including column k. So the submatrix $C(k)$ of A, containing all entries A_{ij} with $i \leq k+w$ and $j \leq k$, controls elimination through step k. Rows below $k+w$ and columns beyond k will not enter this key step : *the choice of pivot row* $i(k)$.

We want to establish these facts in Lemma 1 and Lemma 2 :

1. *The nullspaces* $N(C)$ *and* $N(C^{\mathrm{T}})$ *are finite-dimensional*: Infinite matrices with this Fredholm property behave in important ways like finite matrices.
2. The index $-d$ of $C(k)$, which is dim $N(C)$ – dim $N(C^{\mathrm{T}})$, is independent of k.
3. In the step from $C(k-1)$ to $C(k)$, the new kth column is independent of previous columns by the invertibility of A. (All nonzeros in column k of A are included in rows $k-w$ to $k+w$ of $C(k)$.) Since index $(C(k))$ = index $(C(k-1))$, the submatrix $C(k)$ must contain *exactly one row* $i(k)$ that is newly independent of the rows above it. Every integer i is eventually chosen as $i(k)$ for some k.
4. Let $B(k)$ be the submatrix of $C(k)$ formed from all pivot rows $i(j)$, $j \leq k$. Elimination can be described non-recursively, in terms of the original matrix. We have removed the lowest possible d rows of $C(k)$ to form this *invertible submatrix* $B(k)$. Those d nonpivot rows are combinations of the rows of $B(k)$. Elimination subtracts those same combinations of the rows of A to complete step k. (The example below shows how these combinations lead to L^{-1}, where recursive elimination using only the pivot row (and not all of B) leads directly to L.)
   The figure shows the submatrix $C(k)$. Removing the d dependent rows leaves the invertible submatrix $B(k)$.



5. When elimination is described recursively, the current row $i(k)$ has all zeros before column k. It is row $i(k)$ of PU. The multipliers ℓ_{ji} will go into column i of a lower triangular matrix L, with $L_{ii} = 1$. Then $A = LPU$ with $P_{k,i(k)} = 1$ in the doubly infinite permutation matrix P. The pivot row becomes row k of the upper triangular U.

We may regard "fact 5" as the execution of elimination, and "facts 1, 2, 3, 4" as the key steps in selecting the pivot rows. Our whole argument will rest on the stability of the index, not changing with k. This fact is familiar when A is a finite matrix! If C is an m by n submatrix of rank r, its index is $n-m$ (this is the difference in nullspace dimensions $n-r$ and $m-r$). When a new row and column increase m and n by 1, their difference is unchanged.

Lemma 1 extends that rule to banded infinite Fredholm matrices.

Lemma 1. *$C(k)$ is a Fredholm matrix and its index is independent of k.*

Proof. The invertible operator A is Fredholm with index $\dim N(A) - \dim N(A^{\mathrm{T}}) = 0-0$. We are assuming that A is invertible on the infinite sequence space $\ell_2(Z)$. Key point : Perturbation by a finite rank matrix like D, or by any compact operator, leaves index $= 0$. Therefore A and A^{T} have equal index 0:

$$A = \begin{bmatrix} C(k) & D(k) \\ 0 & E(k) \end{bmatrix} \quad \text{and} \quad A' = \begin{bmatrix} C(k) & 0 \\ 0 & E(k) \end{bmatrix}$$

For banded A, the submatrix $D(k)$ contains only finitely many nonzeros (thus $A-A'$ has finite rank). Now we can separate C from E :

$$A' = \begin{bmatrix} C(k) & 0 \\ 0 & I \end{bmatrix} \begin{bmatrix} I & 0 \\ 0 & E(k) \end{bmatrix} = \begin{bmatrix} I & 0 \\ 0 & E(k) \end{bmatrix} \begin{bmatrix} C(k) & 0 \\ 0 & I \end{bmatrix}.$$

These two commuting factors are Fredholm since A' is Fredholm [8]. The indices of the two factors are equal to the indices of $C(k)$ and $E(k)$. Those indices add to index $(A') =$ index $(A) = 0$.

Now change k. Since $C(k-1)$ comes from $C(k)$ by deleting one row and column, the index is the same. Strictly speaking, the last row and column of $C(k)$ are replaced by $(\ldots, 0, 0, 1)$. This is a finite rank perturbation of $C(k)$: no change in the index. And the index of this matrix diag $(C(k-1), 1)$ equals the index of $C(k-1)$.

Marko Lindner showed me this neat proof of Lemma 1, which he uses to define the "plus-index" and "minus-index" of the outgoing and incoming singly infinite submatrices A_+ and A_- of A. These indices are independent of the cutoff position (row and column k) between A_- and A_+. The rapidly growing theory of infinite matrices is described in [4, 15, 21].

Lemma 2. *There is a unique row number $i(k)$, with $\mid i-k \mid \leq w$, such that*

$$\begin{array}{ll} \text{row } i(k) \text{ of } C(k-1) & \text{is a combination of previous rows of } C(k-1) \\ \text{row } i(k) \text{ of } C(k) & \text{is not a combination of previous rows of } C(k). \end{array}$$

Proof. By Lemma 1, the submatrices $C(k)$ all share the same index $-d$. Each submatrix has nullspace $= \{0\}$, since $C(k)$ contains all nonzeros of all columns $\leq k$ of the invertible matrix A. With index $-d$, the nullspace of every $C(k)^{\mathrm{T}}$ has dimension d. This means that d rows of $C(k)$ are linear combinations of previous rows. Those d rows of $C(k)$ must be among rows $k-w+1,\ldots,k+w$ (since the earlier rows of $C(k)$ contain all nonzeros of the corresponding rows of the invertible matrix A).

$C(k)$ has one new row and column compared to $C(k-1)$. Since d is the same for both, there must be one row $i(k)$ that changes from dependent to independent when column k is included. In $C(k-1)$, that row was a combination of earlier pivot rows. In A, we can subtract that same combination of earlier rows from row $i(k)$. This leaves a row whose first nonzero is in column k. *This is the kth pivot row.*

Notice that this pivot row was not constructed recursively (the usual way). This row never changes again, it will be row $i(k)$ of the matrix PU when elimination ends, and it will be row k of U. The example below shows how the d dependencies lead to L^{-1}.

Let $A(k-1)$ denote the doubly infinite matrix after elimination is complete on columns $< k$ of A. Row $i(k)$ of $A(k-1)$ is that kth pivot row. By subtracting multiples ℓ_{ji} of this row from later non-pivot rows, we complete step k and reach $A(k)$. This matrix has zero in columns $\leq k$ of all d rows that are combinations of earlier pivot rows. The multipliers are $\ell_{ji} = 0$ for all rows $j > k+w$, since those rows (not in $C(k)$) are and remain zero in all columns $\leq k$.

Each row is eventually chosen as a pivot row, because row $k-w$ of $C(k)$ has all the nonzeros of row $k-w$ of A. That row cannot be a combination of previous rows when we reach step k; it already was or now is a pivot row. The bandwidth w of the permutation P (associated with the ordering $i(k)$ of the integers) is confirmed.

This completes the proof of facts 1, 2, 3, 4 and $A = LPU$.

Theorem 4. *Each banded invertible doubly infinite matrix factors into $A = LPU$.*

Toeplitz example with diagonals -2, 5, -2 (now doubly infinite). The correct choice of pivot rows is $i(k) = k$ for all k. The invertible upper left submatrix $B(k-1)$ has 5 along its diagonal. The matrix $C(k-1)$ includes also the dependent row k below (here $w = 1$ and $d = 1$). To see the dependency, multiply rows $k-1, k-2, k-3, \ldots$ by $\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \ldots$ and add to row k:

$$\left[\begin{array}{ccc|cc} 5 & -2 & & & \\ -2 & 5 & -2 & & \\ & -2 & 5 & -2 & \\ \hline 0 & 0 & -2 & 5 & -2 \end{array}\right] \begin{array}{c} \\ \to \\ k-1 \\ k \end{array} \left[\begin{array}{ccc|cc} 5 & -2 & & & \\ -2 & 5 & -2 & & \\ & -2 & 5 & -2 & \\ \hline 0 & 0 & 0 & 4 & -2 \end{array}\right] \tag{20}$$

Row k of A has become row k of PU (also row k of U, since $P = I$). The matrix L^{-1} that multiplies A to produce PU has those coefficients $1, \frac{1}{2}, \frac{1}{4}, \ldots$ leftward along each row. Then its inverse, which is L, has $1, -\frac{1}{2}, 0, 0, \ldots$ down each column.

This was nonrecursive elimination. It produced the pivot row $\ldots,\ 0, 4, -2, 0, \ldots$ by solving one infinite system. We can see normal recursive elimination by using this pivot row to remove the -2 that still lies below the pivot 4. The multiplier in L is $-\frac{1}{2}$.

Suppose we make the *incorrect pivot choice* $i(k) = k-1$ for all k. That gives $P =$ doubly infinite shift. It leads to an *LPU* factorization of A that we don't want, with $L = (A)(\text{inverseshift})$ and $P = (\text{shift})$ and $U = I$. This lower triangular L has -2, 5, -2 down each column. (To maintain the convention $L_{ii} = 1$, divide this L by -2 and compensate with $U = -2I$.)

Recursively, this looks innocent. We are using the -2's above the diagonal to eliminate each 5 and -2 below them. But when the singly infinite submatrix in (20) loses its last row $\ldots$, -2, 5 (and becomes lower triangular with -2 on its diagonal instead of 5), it is *no longer invertible*. The vector $\left(\ldots, \frac{1}{4}, \frac{1}{2}, 1\right)$ is in its nullspace. The correct choice had bidiagonal L and U as in (16).

In the language of Sect. 5, this lower triangular matrix has roots at 2 and $\frac{1}{2}$. It cannot have a bounded inverse. The misplaced root produced that vector in the nullspace.

Theorem 5. *The nonzero entries of P, L, U lie in bands of width* $2w$ *:*

$$\begin{aligned} P_{ik} &= 0 \ \text{ if } \mid i-k \mid > w \\ L_{ik} &= 0 \ \text{ if } i-k > 2w \quad (\text{and if } i < k) \\ U_{ik} &= 0 \ \text{ if } k-i > 2w \quad (\text{and if } k < i). \end{aligned}$$

Proof. For finite matrices, the rank conditions (2) assure that $P_{ik} = 1$ cannot happen outside the diagonal band $\mid i-k \mid \leq w$ containing all nonzeros of A. Then

$$A = LPU \quad \text{gives} \quad L = AU^{-1}P^{-1} = AU^{-1}P^{\mathrm{T}}.$$

The factor $U^{-1}P^{\mathrm{T}}$ cannot have nonzeros below subdiagonal w, since U^{-1} is upper triangular. Then L cannot have nonzeros below subdiagonal $2w$.

Similarly the matrices $P^{\mathrm{T}}L^{-1}$ and A are zero above superdiagonal w. So their product $U = P^{\mathrm{T}}L^{-1}A$ is zero above superdiagonal $2w$.

For infinite matrices, the choice of row $i(k)$ as pivot row in Lemma 2 satisfies $\mid i-k \mid \leq w$. Thus P again has bandwidth w. The entries ℓ_{ji} multiply this pivot row when it is subtracted from lower rows of $C(k)$. Since row $k+w$ is the last row of $C(k)$, its distance from the pivot row cannot exceed $2w$.

Pivot rows cannot have more than $2w$ nonzeros beyond the pivot. So when they move into U with the pivot on the diagonal, U cannot have nonzeros above superdiagonal $2w$.

The extreme cases are matrices with all nonzeros on subdiagonal and superdiagonal w. These show that the bands allowed by Theorem 5 can be attained.

7 Applications of $A = LPU$

In this informal final section, we comment on the doubly infinite $A = LPU$ and a few of its applications.

7.1 If A is a block Toeplitz matrix, so that $A(i, j) = A(i+b, j+b)$ for all i and j, then L, P, and U will have the same block Toeplitz property. The multiplication $A = LPU$ of doubly infinite matrices translates into a multiplication $a(z) = \ell(z)\, p(z)\, u(z)$ of b by b matrix polynomials. Our result can be regarded as providing a new proof of that classical factorization.

This new proof is non-constructive because the steps from original rows (of A) to pivot rows (of PU) require the solution of singly-infinite systems with matrices $B(k)$. The constructive solution of those systems would require the Wiener-Hopf idea that is itself based on $a(z) = u(z)\, p(z)\, \ell(z)$: a vicious circle.

7.2 **Infinite Gram–Schmidt**. From the columns a_1, …, a_n of an invertible matrix A we can produce the orthonormal columns q_1, …, q_n of Q. If each q_k is a combination of $q_1, \ldots, q_{k-1}, a_k$, then each a_k is a combination of $q_1, \ldots, q_k$. This means that A is factored into Q times an upper triangular matrix R. The question is how to start the process when A is doubly infinite.

Notice that $Q^{\mathrm{T}}Q = I$ leads to $A^{\mathrm{T}}A = (QR)^{\mathrm{T}}(QR) = R^{\mathrm{T}}R$. This is a special LU factorization (Cholesky factorization) of the symmetric positive definite matrix $A^{\mathrm{T}}A$. The factors R^{T} and R will have the same main diagonal, containing the square roots of the pivots of $A^{\mathrm{T}}A$ (which are all positive).

If A is doubly infinite and banded, so is $A^{\mathrm{T}}A$. Then its factorization in the section "Elimination on Banded Doubly Infinite Matrices" produces $R^{\mathrm{T}}R$. The invertible submatrices $B(k)$ in the proof share the main diagonal of $A^{\mathrm{T}}A$. All their inverses are bounded by $\| (A^{\mathrm{T}}A)^{-1} \|$. No permutation P is needed to reach the triangular factor R.

Now $Q = AR^{-1}$ has orthonormal columns q_k. Each q_k is a combination of the original a_j, $j \leq k$. Q is banded below its main diagonal but not above—apart from the exceptional cases when R has a banded inverse.

7.3 Theorem 1 came from the observation that the upper left submatrices of A, L, P, U satisfy $a = \ell pu$. With doubly infinite matrices and singly infinite submatrices, this remains true. The ranks of diagonal blocks A_+ and A_- are now infinite, so we change to nullities. But as the block diagonal example in Sect. 5 made clear, L and U and their inverses may not be bounded operators. At this point the uniqueness of P comes from its construction (during elimination) and not from Theorem 1.

7.4 In recent papers we studied the group of banded matrices *with banded inverses* [23–25]. These very special matrices are products $A = F_1 \ldots F_N$ of block diagonal invertible matrices. Our main result was that $A = F_1F_2$ if we allow blocks of size $2w$, and then $N \leq Cw^2$ when the blocks have size ≤ 2. The key point is that the number N of block diagonal factors is controlled by w and not by the size of A. The proof uses elimination and A can be singly infinite.

We have no proof yet when A is doubly infinite. It is remarked in [23] that $A = LPU$ reduces the problem to banded triangular matrices L and U with banded inverses. We mention Panova's neat factorization [18] of P (whose inverse is P^{T}). With bandwidth w, a singly infinite P is the product of $N < 2w$ parallel exchanges of neighbors (block diagonal permutations with block size ≤ 2).

A doubly infinite P will require a power of the infinite shift matrix S, in addition to $F_1 \ldots F_N$. This power $s(P)$ is the "shifting index" of P and $\mid s \mid \leq w$. The main diagonal is not defined for doubly infinite matrices, until the shifting index $s(A) = s(P)$ tells us where it ought to be. This agrees with the main diagonal located by de Boor $[bi:006]$.

7.5 For singly infinite Fredholm matrices the main diagonal is well defined. It is located by the Fredholm index of A. When the index is zero, the main diagonal is in the right place. (Still A may or may not be invertible. For a block Toeplitz matrix invertibility requires all partial indices $k(j)$ to be zero, not just their sum.)

The proof of Lemma 1 showed why the Fredholm indices of the incoming A_- and outgoing A_+ are independent of the cutoff position (row and column k). When A is invertible, that "minus-index" and "plus-index" add to zero. The connection to the shifting index was included in [23].

Theorem 6. *The shifting index of a banded invertible matrix A (and of its permutation P) equals the Fredholm index of A_+ (the plus-index).*

Check when A is the doubly infinite shift matrix S with nonzero entries $S_{i,i+1} = 1$. Then P coincides with S and has shifting index 1 *(one S in its factorization into bandwidth* 1 *matrices). The outgoing submatrix A_+ is a singly infinite shift with* $(1,0,0,\ldots)$ *in its nullspace. Then $A_+^{\mathrm{T}}x = 0$ only for $x = 0$, so the Fredholm index of A_+ is also* 1.

A deep result from the theory of infinite matrices [20, 21] concerns the Fredholm indices of the limit operators *of A.*

7.6 I would like to end with a frightening example. It shows that the associative law $A(Bx) = (AB)x$ can easily fail for infinite matrices. I always regarded this as the most fundamental and essential law! It defines AB (by composition), and it is the key to so many short and important proofs that I push my linear algebra classes to recognize and even anticipate a "proof by moving the parentheses."

The example has $Bx = 0$ but $AB = I$. And $0 = A(Bx) = (AB)x = x$ is false.

$$A = \begin{bmatrix} 1 & 1 & 1 & \bullet \\ 0 & 1 & 1 & \bullet \\ 0 & 0 & 1 & \bullet \\ 0 & 0 & 0 & \bullet \end{bmatrix} \quad B = \begin{bmatrix} 1 & -1 & 0 & \bullet \\ 0 & 1 & -1 & \bullet \\ 0 & 0 & 1 & \bullet \\ 0 & 0 & 0 & \bullet \end{bmatrix} \quad x = \begin{bmatrix} 1 \\ 1 \\ 1 \\ \bullet \end{bmatrix} \tag{21}$$

This is like the integral of the derivative of a constant. A is an unbounded operator, the source of unbounded difficulty. A direct proof of the law $A(Bx) = (AB)x$

would involve rearranging series. Riemann showed us that without absolute convergence, which is absent here, all sums are possible if $a_n \to 0$.

This example has led me to realize that grievous errors are all too possible with infinite matrices. I hope this paper is free of error. But when elimination has no starting point (and operator theory is not developed in detail), it is wise to be prepared for the worst.

Acknowledgements This chapter was begun on a visit to the beautiful Hong Kong University of Science and Technology. We thank its faculty and its president Tony Chan for their warm welcome. On my final morning, George Lusztig happened to walk by—and he clarified the history of Bruhat. The books of Marko Lindner (and our email correspondence) introduced me to the limit analysis of non-Toeplitz infinite matrices.

References

1. Albert C, Li C-K, Strang G, Yu G. Permutations as product of parallel transpositions. SIAM J Discret Math. 2011;25(3):1412–17.
2. Asplund E. Inverses of matrices $\{a_{ij}\}$ which satisfy $a_{ij} = 0$ for $j > i + p$. Math Scand. 1959;7:57–60.
3. Bruhat F. Representations induites des groups de Lie semisimples complexes. Comptes Rendus Acad Sci Paris. 1954;238:437–9.
4. Chandler-Wilde SN, Lindner M. Limit operators, collective compactness, and the spectral theory of infinite matrices. Providence: American Mathematical Society; 2011. (American Mathematical Society Memoirs, vol. 210, no. 989).
5. Chevalley C. Sur certains groupes simples. J Tôhoku Math. 1955;7(2):14–66.
6. de Boor, Carl. What is the main diagonal of a bi-infinite band matrix? Quantitative approximation (Proc. Internat. Sympos., Bonn, 1979), pp. 11–23, Academic Press, New York-London, 1980.
7. Elsner L. On some algebraic problems in connection with general eigenvalue algorithms. Lin Alg Appl 1979;26:123–38.
8. Gohberg I, Goldberg S. Finite dimensional Wiener-Hopf equations and factorizations of matrices. Lin Alg Appl. 1982;48:219–36.
9. Gohberg, Israel; Goldberg, Seymour; Kaashoek, Marinus A. Basic classes of linear operators. Birkhäuser Verlag, Basel, 2003. xviii+423 pp. ISBN: 3-7643-6930-2
10. Gohberg, I.; Kaashoek, M. A.; Spitkovsky, I. M. An overview of matrix factorization theory and operator applications. Factorization and integrable systems (Faro, 2000), 1–102, Oper. Theory Adv. Appl., 141, Birkhäuser, Basel, 2003
11. Grcar, Joseph F. How ordinary elimination became Gaussian elimination. Historia Math. 2011;38(2):163–218.
12. Harish-Chandra. On a lemma of Bruhat. J Math Pures Appl. 1956;35:203–10.
13. Hart, Roger. The Chinese roots of linear algebra. Johns Hopkins University Press, Baltimore, MD, 2011. xiv+286 pp. ISBN: 978-0-8018-9755-9
14. Kolotilina L Yu, Yeremin A Yu. Bruhat decomposition and solution of sparse linear algebra systems. Sov J Numer Anal Math Model. 1987;2:421–36.
15. Lindner, Marko. Infinite matrices and their finite sections. An introduction to the limit operator method. Frontiers in Mathematics. Birkhäuser Verlag, Basel, 2006. xv+191 pp. ISBN: 978-3-7643-7766-3; 3-7643-7766-6
16. Lusztig G. Bruhat decomposition and applications. math.mit.edu/∼gyuri.

17. Olshevsky V, Zhlobich P, Strang G. Green's matrices. Lin Alg Appl. 2010;432:218–41.
18. Panova G. Factorization of banded permutations. Proc Am Math Soc. 2012;140:3805–12.
19. Plemelj J. Riemannsche Funktionenscharen mit gegebener Monodromiegruppe. Monat Math Phys. 1908;19:211–45.
20. Rabinovich VS, Roch S, Roe J. Fredholm indices of band-dominated operators. Integral Equ Oper Theory. 2004;49:221–38.
21. Rabinovich VS, Roch S, Silbermann B. The finite section approach to the index formula for band-dominated operators. Oper Theory. 2008;187:185–93.
22. Shen, Kangshen; Crossley, John N.; Lun, Anthony W.-C. The nine chapters on the mathematical art. Companion and commentary. With forewords by Wentsn Wu and Ho Peng Yoke. Oxford University Press, New York; Science Press, Beijing, 1999. xiv+596 pp. ISBN: 0-19-853936–3
23. Strang G. Fast transforms: Banded matrices with banded inverses. Proc Natl Acad Sci. 2010;107:12413–16.
24. Strang G. Banded matrices with banded inverses and $A = LPU$, International Congress of Chinese Mathematicians, Beijing, 2010. Proceedings to appear.
25. Strang, Gilbert. Groups of banded matrices with banded inverses. Proc. Amer Math Soc. 2011;139(12):4255–64.

Hodge-de Rham Theory of K-Forms on Carpet Type Fractals

Jason Bello*, Yiran Li, and Robert S. Strichartz†

Abstract We outline a Hodge-de Rham theory of k-forms (for k = 0,1,2) on two fractals: the Sierpinski Carpet (SC) and a new fractal that we call the Magic Carpet (MC), obtained by a construction similar to that of SC modified by sewing up the edges whenever a square is removed. Our method is to approximate the fractals by a sequence of graphs, use a standard Hodge-de Rham theory on each graph, and then pass to the limit. While we are not able to prove the existence of the limits, we give overwhelming experimental evidence of their existence, and we compute approximations to basic objects of the theory, such as eigenvalues and eigenforms of the Laplacian in each dimension, and harmonic 1-forms dual to generators of 1-dimensional homology cycles. On MC we observe a Poincare type duality between the Laplacian on 0-forms and 2-forms. On the other hand, on SC the Laplacian on 2-forms appears to be an operator with continuous (as opposed to discrete) spectrum.
2010 *Mathematics Subject Classification.* Primary: 28A80

Keywords Analysis on fractals · Hodge-de Rham theory · k-forms · Harmonic 1-forms · Sierpinski carpet · Magic carpet

J. Bello (✉)
Mathematics Department, Ohio State University, Columbus, OH 43210, USA
e-mail: jbello01@yahoo.com

Y. Li
Department of Mathematics, University of Maryland, College Park, Maryland, MD 20742, USA
e-mail: yl534@math.umd.edu

R. S. Strichartz
Mathematics Department, Malott Hall, Cornell University, Ithaca, NY 14853, USA
e-mail: str@math.cornell.edu

* Research supported by the National Science Foundation through the Research Experiences for Undergraduates Program at Cornell,grant DMS- 1156350

† Research supported by the National Science Foundation, grant DMS - 1162045

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_2

1 Introduction

There have been several approaches to developing an analogue of the Hodge-de Rham theory of k-forms on the Sierpinski gasket (SG) and other post-critically finite (pcf) fractals ([1, 8–14, 16, 17]). In this chapter we extend the approach in [1] to the Sierpinski carpet (SC) and a related fractal that we call the *magic carpet* (MC). These fractals are not finitely ramified, and this creates technical difficulties in proving that the conjectured theoretical framework is valid. On the other hand, the structure of "2-dimensional" cells intersecting along "1-dimensional" edges allows for a nontrivial theory of 2-forms. Our results are largely experimental, but they lead to a conjectured theory that is more coherent than for SG.

The approach in [1] is to approximate the fractal by graphs, define k-forms and the associated d, δ, Δ operators on them, and then pass to the limit. In the case of SC there is a natural choice of graphs. Figure 1 shows the graphs on levels 0, 1, and 2.

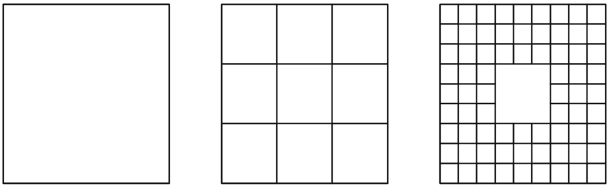


Fig. 1 The graphs approximating SC on levels 0,1, and 2

SC is defined by the self-similar identity:

$$SC = \bigcup_{j=1}^{\infty} F_j(SC)$$

where F_j is the similarity map of contraction ratio $1/3$ from the unit square to one of the eight of the nine subsquares (all except the center square) after tic-tac-toe subdivision. We define the sequence of graphs

$$\Gamma_m = \bigcup_{j=1}^{\infty} F_j(\Gamma_{m-1})$$

with the appropriate identification of vertices in $F_j(\Gamma_{m-1})$ and $F_k(\Gamma_{m-1})$. Note that a hole in SC on level m does not become visible on the graph until level $m+1$, but it will influence the definition of 2-cells. We denote by $E_0^{(m)}$ the vertices of Γ_m. A 0-form on level m is just a real-valued function $f_0^{(m)}(e_0^{(m)})$ defined on $e_0^{(m)} \in E_0^{(m)}$. We denote the vector space of 0-forms by $\Lambda_0^{(m)}$. The edges $E_1^{(m)}$ of Γ_m exist in opposite orientations $e_1^{(m)}$ and $-e_1^{(m)}$, and a 1-form (element of $\Lambda_1^{(m)}$) is a function on $E_1^{(m)}$

satisfying

$$f_1^{(m)}(-e_1^{(m)}) = -f_1^{(m)}(e_1^{(m)}).$$

By convention we take vertical edges oriented upward and horizontal edges oriented to the right. We denote by E_2^m the squares in Γ_m that bound a cell $F_\omega(SC)$, where $\omega = (\omega_1, ..., \omega_m)$ is a word of length m, $\omega_j = 1,2,...,8$ and $F_\omega = F_{\omega_1} \circ F_{\omega_2} \circ ... \circ F_{\omega_m}$. Thus an element e_2^m of E_2^m consists of the subgraph of Γ_m consisting of the four vertices $\{F_\omega(e_0^0) : e_0^0 \in E_0^{(0)}\}$. In particular, there are eight elements of $E_2^{(1)}$, even thought the central square is a subgraph of the same type. In general $\#E_2^{(m)} = 8^m$, and we will denote squares by the word ω that generates them. A 2-form is defined to be a function $f_2^{(m)}(\omega)$ on $E_2^{(m)}$.

The boundary of a square consists of the four edges in counterclockwise orientation. With our orientation convention the bottom and right edges will have a plus sign and the top and left edges will have minus sign. We build a signum function to do the bookkeeping: if $e_1^{(m)} \subseteq e_2^{(m)}$ then

$$sgn(e_1^{(m)}, e_2^{(m)}) = \begin{cases} +1 & \text{top and right} \\ -1 & \text{bottom and left} \\ 0 & e_1^{(m)} \text{ is not a boundary edge of } e_2^{(m)}. \end{cases}$$

It is convenient to define $sgn(e_1^{(m)}, e_2^{(m)}) = 0$ if $e_1^{(m)}$ is not a boundary edge of $e_2^{(m)}$. Similarly, if $e_1^{(m)}$ is an edge containing the vertex $e_0^{(m)}$, define

$$sgn(e_0^{(m)}, e_1^{(m)}) = \begin{cases} +1 & e_0^{(m)} \text{ is top and right} \\ -1 & e_0^{(m)} \text{is bottom and left} \\ 0 & e_0^{(m)} \text{is not an endpoint of } e_1^{(m)}. \end{cases}$$

It is easy to check the consistency condition

$$\sum_{e_1^{(m)} \in E_1^{(m)}} sgn(e_0^{(m)}, e_1^{(m)}) sgn(e_1^{(m)}, e_2^{(m)}) = 0 \tag{1}$$

for any fixed $e_0^{(m)}$ and $e_2^{(m)}$, since for $e_0^{(m)} \in e_2^{(m)}$ there are only two nonzero summands, one +1 and the other -1.

We may define the de Rham complex

$$0 \to \Lambda_0^{(m)} \xrightarrow{d_0^{(m)}} \Lambda_1^{(m)} \xrightarrow{d_1^{(m)}} \Lambda_2^{(m)} \to 0$$

with the operators

$$d_0^{(m)} f_0^{(m)}(e_1^{(m)}) = \sum_{e_0^{(m)} \in E_0^{(m)}} sgn(e_0^{(m)}, e_1^{(m)}) f_0^{(m)}(e_0^{(m)}) \tag{2}$$

(only two nonzero terms) and

$$d_1^{(m)} f_1^{(m)}(e_2^{(m)}) = \sum_{e_1^{(m)} \in E_1^{(m)}} sgn(e_1^{(m)}, e_2^{(m)}) f_1^{(m)}(e_1^{(m)}) \tag{3}$$

(only four nonzero terms). The relation

$$d_1^{(m)} \circ d_0^{(m)} \equiv 0 \tag{4}$$

is an immediate consequence of (1).

To describe the δ operators and the dual de Rham complex we need to choose inner products on the spaces $\Lambda_0^{(m)}$, $\Lambda_1^{(m)}$, $\Lambda_2^{(m)}$, or what is the same thing, to choose weights on $E_0^{(m)}$, $E_1^{(m)}$, $E_2^{(m)}$. The most direct choice is to weight each square $e_2^{(m)}$ equally, say

$$\mu_2(e_2^{(m)}) = \frac{1}{8^m},$$

making μ_2 a probability measure on $E_2^{(m)}$. Note that we might decide to renormalize by multiplying by a constant, depending on m, when we examine the question of the limiting behavior as $m \to \infty$. For the weighting on edges we may imagine that each square passes on a quarter of its weight to each boundary edge. Some edges bound one square and some bound two squares, so we choose

$$\mu_1(e_1^{(m)}) = \begin{cases} \dfrac{1}{4 \cdot 8^m} & \text{if } e_1^{(m)} \text{ bounds one square} \\ \dfrac{1}{2 \cdot 8^m} & \text{if } e_1^{(m)} \text{ bounds two squares.} \end{cases} \tag{5}$$

For vertices we may again imagine the weight of each square being split evenly among its vertices. A vertex may belong to 1, 2, 3, or 4 squares, so

$$\mu_0(e_0^{(m)}) = \frac{k}{4 \cdot 8^m} \tag{6}$$

if $e_0^{(m)}$ lies in k squares.

The dual de Rham complex

$$0 \leftarrow \Lambda_0^{(m)} \xleftarrow{\delta_1^{(m)}} \Lambda_1^{(m)} \xleftarrow{\delta_2^{(m)}} \Lambda_2^{(m)} \leftarrow 0$$

is defined abstractly by $\delta_1^{(m)} = d_0^{(m)*}$ and $\delta_2^{(m)} = d_1^{(m)*}$ where the adjoints are defined in terms of the inner products induced by the weights, or concretely as

$$\delta_1^{(m)} f_1^{(m)}(e_0^{(m)}) = \sum_{e_1^{(m)} \in E_1^{(m)}} \frac{\mu_1(e_1)}{\mu_0(e_0)} sgn(e_0^{(m)}, e_1^{(m)}) f_1^{(m)}(e_1^{(m)}), \tag{7}$$

$$\delta_2^{(m)} f_2^{(m)}(e_1^{(m)}) = \sum_{e_2^{(m)} \in E_2^{(m)}} \frac{\mu_2(e_2)}{\mu_1(e_1)} sgn(e_1^{(m)}, e_2^{(m)}) f_2^{(m)}(e_2^{(m)}). \tag{8}$$

There are one or two nonzero terms in (8) depending on whether $e_1^{(m)}$ bounds one or two squares. In (7) there may be 2, 3, or 4 nonzero terms, depending on the number of edges that meet at the vertex $e_0^{(m)}$. The condition

$$\delta_1^{(m)} \circ \delta_2^{(m)} = 0 \tag{9}$$

is the dual of (4).

We may then define the Laplacian

$$\begin{aligned} -\Delta_0^{(m)} &= \delta_1^{(m)} d_0^{(m)} \\ -\Delta_1^{(m)} &= \delta_2^{(m)} d_1^{(m)} + d_0^{(m)} \delta_1^{(m)} \\ -\Delta_2^{(m)} &= d_1^{(m)} \delta_2^{(m)} \end{aligned} \tag{10}$$

as usual. These are nonnegative self-adjoint operators on the associated L^2 spaces, and so have a discrete nonnegative spectrum. We will be examining the spectrum (and associated eigenfunctions) carefully to try to understand what could be said in the limit as $m \to \infty$. Also of particular interest are the harmonic 1-forms $\mathscr{H}_1^{(m)}$, solutions of $-\Delta_1^{(m)} h_1^{(m)} = 0$. As usual these can be characterized by the two equations

$$d_1^{(m)} h_1^{(m)} = 0 \qquad\qquad \delta_1^{(m)} h_1^{(m)} = 0,$$

and can be put into cohomology/homology duality with the homology generating cycles in Γ_m. The Hodge decomposition

$$\Lambda_1^{(m)} = d_0^{(m)} \Lambda_0^{(m)} \oplus \delta_2^{(m)} \Lambda_2^{(m)} \oplus \mathscr{H}_1^{(m)} \tag{11}$$

shows that the eigenfunctions of $-\Delta_1^{(m)}$ with $\lambda \neq 0$ are either $d_0^{(m)} f_0^{(m)}$ for $f_0^{(m)}$ an eigenfunction of $-\Delta_0^{(m)}$, or $\delta_2^{(m)} f_2^{(m)}$ for $f_2^{(m)}$ an eigenfunction of $-\Delta_2^{(m)}$ with the same eigenvalue. Thus the nonzero spectrum of $-\Delta_1^{(m)}$ is just the union of the $-\Delta_0^{(m)}$ and $-\Delta_2^{(m)}$ spectrum.

The irregular nature of the adjacency of squares in Γ_m, with the associated variability of the weights in (6) and (5) , leads to a number of complications in the behavior of $-\Delta_2^{(m)}$. To overcome these complications we have invented the fractal MC, that is obtained from SC by making identifications to eliminate boundaries. On the outer boundary of SC we identify the opposite pairs of edges with the same orientation, turning the full square containing SC into a torus. Each time we delete a small square in the construction of SC we identify the opposite edges of the deleted square with the same orientation. We may think of MC as a limit of closed surfaces of genus $g = 1 + (1 + 8 + ... + 8^{m-1})$, because each time we delete and "sew up" we add on a handle to the torus. This surface carries a flat metric with singularities at the corners of each deleted square (all four corners are identified). It is not difficult to show that the limit exists as a topological space, because the identifications made at level m are all points that are very close to each other, distance 3^{-m} in the Euclidean metric. The metric structure of MC is more complicated. If we consider the geodesic metric, the approximation on level m has diameter $(2/3)^m$, because we can use diagonal zig-zag lines to take advantage of the identifications that lead to the singular points to connect all vertices with a sequence of 2^m edges of size 3^{-m}. Thus it would appear that we should renormalize the metric at level m by multiplying by $(3/2)^m$ in order to obtain a metric structure in the limit. The exact nature of this metric structure remains to be investigated. Whether or not the analytic structures (energy, Laplacian, Brownian motion) on SC can be transferred to MC remains to be investigated. Our results give overwhelming evidence that this is the case.

We pass from the graphs Γ_m approximating SC to graphs $\tilde{\Gamma}_m$ approximating MC by making the same identifications of vertices and edges. The graph $\tilde{\Gamma}_1$ is shown in Fig. 2.

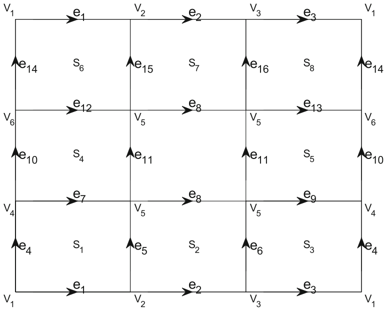


Fig. 2 $\tilde{\Gamma}_1$ with 6 vertices labeled v_j, 16 edges labeled e_j, and 8 squares labeled s_j

Each square has exactly four neighbors (not necessarily distinct) with each edge separating two squares. For example, in Fig. 2, we see that s_2 has neighbors s_1, s_3, and s_7 twice, as e_2 and e_8 both separate s_2 and s_7. There are two types of vertices, that we call *nonsingular* and *singular*. The nonsingular vertices (all except v_5 in Fig. 2) belong to exactly four distinct squares and four distinct edges (two incoming and two outgoing according to our orientation choice). For example, v_1 belongs to squares s_1, s_3, s_6, and s_8 and has incoming edges e_3 and e_{14} and outgoing edges e_1 and e_4. Singular vertices belong to 12 squares (with double counting) and 12 edges (some of which may be loops). In Fig. 2 there is only one singular vertex, v_5. It belongs to squares s_1, s_3, s_6, and s_8 counted once and s_2, s_4, s_5, and s_7 counted twice. In Fig. 3 we show a neighborhood of this vertex in $\tilde{\Gamma}_2$, with the incident squares and edges shown.

The definition of the de Rham complex for the graphs $\tilde{\Gamma}_m$ approximating MC is exactly the same as for Γ_m approximating SC. The difference is in the dual de Rham

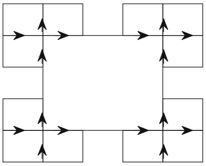


Fig. 3 A neighborhood in $\tilde{\Gamma}_2$ of vertex v_5 from Fig. 2

complex, because the weights are different. We take $\tilde{\mu}_2(e_2^{(m)}) = 1/8^m$ as before, but now $\tilde{\mu}_1(e_1^{(m)}) = \frac{1}{2\cdot 8^m}$ because every edge bounds two squares. Finally

$$\tilde{\mu}_0(e_0^{(m)}) = \begin{cases} \dfrac{1}{8^m} & \text{if } e_0^{(m)} \text{ is nonsingular} \\ \dfrac{3}{8^m} & \text{if } e_0^{(m)} \text{ is singular} \end{cases}$$

because $e_0^{(m)}$ belongs to four squares in the first case and 12 squares in the second case. After that the definitions are the same using the new weights.

Explicitly, we have

$$-\tilde{\Delta}_2^{(m)} f_2^{(m)}(e_2^{(m)}) = \left(2(\sum_{e_2^{(m)\prime} \sim e_2^{(m)}} f_2^{(m)}(e_2^{(m)}) - f_2^{(m)}(e_2^{(m)\prime})) \right) \tag{12}$$

(exactly four terms in the sum), the factor 2 coming from $\tilde{\mu}_2(e_2^{(m)})/\tilde{\mu}_1(e_1^{(m)})$. Except for the factor 2 this is exactly the graph Laplacian on the 4-regular graph whose vertices are the squares in $\tilde{E}_2^{(m)}$ and whose edge relation is $e_2^{(m)\prime} \sim e_2^{(m)}$ if they have an edge in common (double count if there are two edges in common).

The explicit expression for $-\Delta_0^{(m)}$ is almost as simple:

$$-\tilde{\Delta}_0^{(m)} f_0^{(m)}(e_0^{(m)}) = \begin{cases} \dfrac{1}{2} \displaystyle\sum_{e_0^{(m)\prime} \sim e_0^{(m)}} (f_0^{(m)}(e_0^{(m)}) - f_0^{(m)}(e_0^{(m)\prime})) & \text{if } e_0^{(m)} \text{ is nonsingular} \\ \dfrac{1}{6} \displaystyle\sum_{e_0^{(m)\prime} \sim e_0^{(m)}} (f_0^{(m)}(e_0^{(m)}) - f_0^{(m)}(e_0^{(m)\prime})) & \text{if } e_0^{(m)} \text{ is singular.} \end{cases}$$

Note that there are four summands in the first case and 12 summands in the second case (some may be zero if there is a loop connecting $e_0^{(m)}$ to itself in the singular case). We expect that the spectra of these two Laplacians will be closely related, aside from the multiplicative factor of 4. We may define Hodge star operators from

$\Lambda_0^{(m)}$ to $\Lambda_2^{(m)}$ and from $\Lambda_2^{(m)}$ to $\Lambda_0^{(m)}$ by

$$*f_0^{(m)}(e_2^{(m)}) = \frac{1}{4} \sum_{e_0^{(m)} \subseteq e_2^{(m)}} f_0^{(m)}(e_0^{(m)}) \tag{13}$$

and

$$*f_2^{(m)}(e_0^{(m)}) = \begin{cases} \dfrac{1}{4} \displaystyle\sum_{e_2^{(m)} \supseteq e_0^{(m)}} f_2^{(m)}(e_2^{(m)}) & \text{if } e_0^{(m)} \text{ is nonsingular} \\ \dfrac{1}{12} \displaystyle\sum_{e_2^{(m)} \supseteq e_0^{(m)}} f_2^{(m)}(e_2^{(m)}) & \text{if } e_0^{(m)} \text{ is singular.} \end{cases} \tag{14}$$

Note that we do not have the inverse relation that $**$ is equal to the identity in either order. Nor is it true that the star operators conjugate the two Laplacians. However, they are approximately valid, so we can hope that in the appropriate limit there will be a complete duality between 0-forms and 2-forms with identical Laplacians. Nothing remotely like this valid for SC.

It is also easy to describe explicitly the equations for harmonic 1-forms. The condition $d_1^{(m)} h_1^{(m)}(e_2^{(m)}) = 0$ is simply the condition that the sum of the values $h_1^{(m)}(e_1^{(m)})$ over the four edges of the square is zero (with appropriate signs). Similarly the condition $\delta_1^{(m)} h_1^{(m)}(e_0^{(m)}) = 0$ means the sum over the incoming edges equals the sum over the outgoing edges at $e_0^{(m)}$. Those equations have two redundancies, since the sums $\sum_{e_2^{(m)} \in E_2^{(m)}} d_1^{(m)} f_1^{(m)}(e_2^{(m)})$ and $\sum_{e_0^{(m)} \in E_0^{(m)}} \delta_1^{(m)} f_1^{(m)}(e_0^{(m)})$ are automatically zero for any 1-form $f_1^{(m)}$. Thus in $\tilde{\Lambda}_1^{(1)}$ there is a 4-dimensional space of harmonic 1-forms, and in general the dimension is $2g$, which is exactly the rank of the homology group for a surface of genus g. It is easy to identify the homology generating cycles as the edges that are identified.

The remainder of this chapter is organized as follows: In Sects. 2, 3, and 4 we give the results of our computations on SC for 0-forms, 1-forms, and 2-forms. In Sect. 5 we give the results for 0-forms and 2-forms on MC. In Sect. 6 we give the results for 1-forms on MC. We conclude with a discussion in Sect. 7 of all the results and their implications. The website [25] gives much more data than we have been able to include in this chapter, and also contains all the programs used to generate the data.

2 0-Forms on the Sierpinski Carpet

The 0-forms on SC will simply be continuous functions on SC, and we can restrict them to the vertices of Γ_m to obtain 0-forms on Γ_m. The Laplacian $-\Delta_0^{(m)}$ is exactly the graph Laplacian of Γ_m with weights on vertices and edges given by (6) and (5).

Thus

$$-\Delta_0^{(m)} f_0^{(m)}(x) = \sum_{y\sim x} c(x,y)(f(x)-f(y)) \tag{15}$$

with coefficients show in Fig. 4

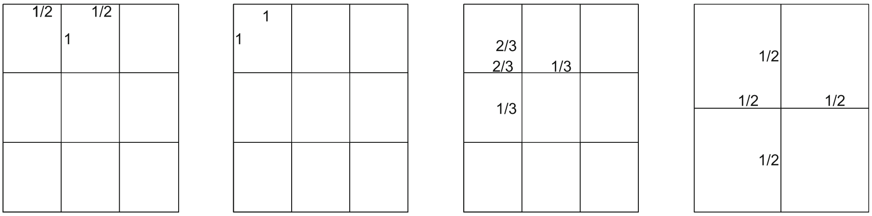


Fig. 4 Coefficients in (15)

The sequence of renormalized Laplacians

$$\{-r^m \Delta_0^{(m)}\} \tag{16}$$

for $r \approx 10.01$ converges to the Laplacian on functions [3, 5–7, 19].

In Table 1 we give the beginning of the spectrum $\{\lambda_j^{(m)}\}$ for $m = 2, 3, 4$ and the ratios $\lambda_j^{(3)}/\lambda_j^{(2)}$ and $\lambda_j^{(4)}/\lambda_j^{(3)}$. The results are in close agreement with the computations in [7] and [6], and suggest the convergence of (16).

Table 1 Eigenvalues of $-\Delta_0^{(m)}$ for $m = 2, 3, 4$ and ratios

$m=2$	Multiplicity	$m=3$	Multiplicity	$m=4$	Multiplicity	$\lambda_j^{(3)}/\lambda_j^{(2)}$	$\lambda_j^{(4)}/\lambda_j^{(3)}$
0.0000	1	0.0000	1	0.0000	1		
0.0414	2	0.0041	2	0.0004	2	0.1001	0.0999
0.1069	1	0.0109	1	0.0011	1	0.1024	0.1002
0.2006	1	0.0204	1	0.0020	1	0.1017	0.0999
0.2635	2	0.0272	2	0.0027	2	0.1031	0.1002
0.2720	1	0.0284	1	0.0028	1	0.1044	0.1003
0.3927	2	0.0414	2	0.0041	2	0.1053	0.1001
0.4260	1	0.0449	1	0.0045	1	0.1055	0.1002
0.4490	1	0.0472	1	0.0047	1	0.1051	0.1001
0.5276	1	0.0560	1	0.0056	1	0.1062	0.1004
0.6375	2	0.0673	2	0.0067	2	0.1055	0.1002
0.6700	1	0.0696	1	0.0069	1	0.1038	0.0997
0.8405	2	0.0976	2	0.0099	2	0.1161	0.1014
0.8713	1	0.1009	1	0.0102	1	0.1158	0.1008
0.9102	1	0.1069	1	0.0109	1	0.1175	0.1024
0.9336	1	0.1103	1	0.0112	1	0.1181	0.1019

The convergence of (16) would imply that $\lim_{m\to\infty} r^m\lambda_j^{(m)} = \lambda_j$ gives the spectrum of the limit Laplacian $-\Delta_0$ on 0-forms on SC. In particular the values $\{r^m\lambda_j^{(m)}\}$ for small values of j (depending on m) would give a reasonable approximation of some lower portion of the spectrum of $-\Delta_0$. To visualize this portion of the spectrum we compute the eigenvalue counting function $N(t) = \#\{\lambda_j \leq t\} \approx \#\{r^m\lambda_j^{(m)} \leq t\}$ and the Weyl ratio $W(t) = \frac{N(t)}{t^\alpha}$. In Figure we display the graphs of the Weyl ratio using the $m = 1,2,3,4$ approximations, with value α determined from the data to get a function that is approximately constant. As explained in [6], we expect $\alpha = \frac{\log 8}{\log r} \approx 0.9026$, which is close to the experimentally determined values. This is explained by the phenomenon called *miniaturization* as described in [7]. Every eigenfunction $u_j^{(m)}$ of $\Delta_0^{(m)}$ reappears in miniaturized form $u_k^{(m+1)}$ of $\Delta_0^{(m+1)}$ with the same eigenvalue $\lambda_k^{(m+1)} = \lambda_j^{(m)}$, so in terms of $\Delta_0^{(m)}$ we have $r^{m+1}\lambda_k^{(m+1)} = r(r^m\lambda_j^{(m)})$. We can in fact see this in Table 1. In passing from level m to level $m+1$, the number of eigenvalues is multiplied by 8, so we expect to have $N(rt) \approx 8N(t)$, and this explains why $\alpha = \frac{\log 8}{\log r}$ is the predicted power growth factor of $N(t)$. We also expect to see an approximate multiplicative periodicity in $W(t)$, namely $W(rt) \approx W(t)$. It is difficult to observe this in our data, however. We also mention that miniaturization is valid for all k-forms (k =0, 1, 2) on SC and MC. This is most interesting for 0-forms and 2-forms on MC as discussed in Sect. 5.

In Fig. 5, 6, and 7 we show graphs of selected eigenfunctions on levels 2, 3, 4. We only display those whose eigenspaces have multiplicity one. The convergence is visually evident. To quantify the rate of convergence we give the values of $\|f_j^{(m)}{}_{|_{E_0^{(m-1)}}} - f_j^{(m-1)}\|_2^2$ in Table 2. Here the L^2 norm on $E_0^{(m-1)}$ is defined by

$$\|f\|_2^2 = \sum_{e_0^{(m-1)} \in E_0^{(m-1)}} \mu_0(e_0^{(m-1)})|f(e_0^{(m-1)})|^2$$

and we normalize the eigenfunctions so that $\|f_j^{(m-1)}\|_2^2$ and $\|f_j^{(m)}{}_{|_{E_0^{(m-1)}}}\|_2 = 1$. In Fig. 8 we show the graph of the weyl ratio of eigenvalues of the 0 forms on different levels.

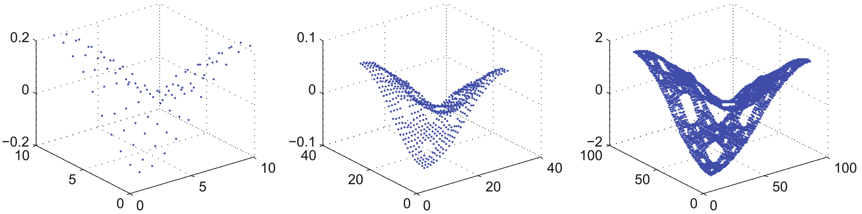


Fig. 5 Graph of eigenfunction of 4th eigenvalue on 0 forms of level 2, 3, and 4

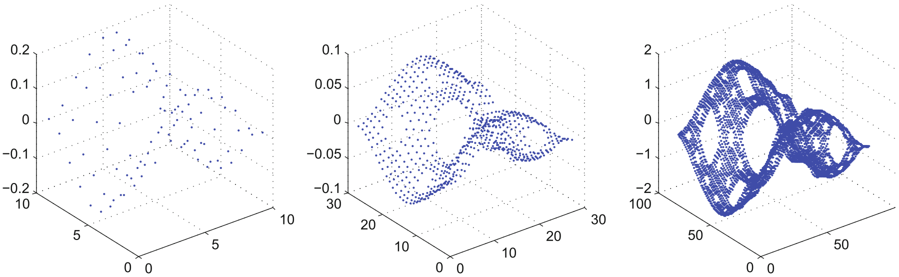


Fig. 6 Graph of eigenfunction of 5th eigenvalue on 0 forms of level 2, 3, and 4

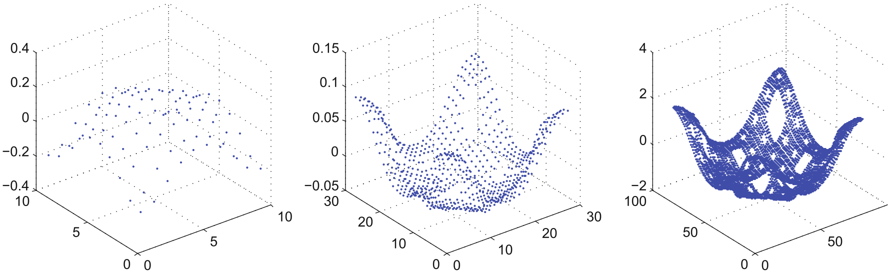


Fig. 7 Graph of eigenfunction of 8th eigenvalue on 0 forms of level 2, 3, and 4

Table 2 Values of $\|f_j^{(m)}|_{E_0^{(m-1)}} - f_j^{(m-1)}\|_2^2$, from level 3 to level 2 and level 4 to level 3 for eigenspaces of multiplicity one

Number of Eigenvalue	$m = 2$	$m = 3$	$m = 4$	Level 3 to level 2	Level 4 to level 3
1	0.0000	0.0000	0.0000	0.0000	0.0000
4	0.1069	0.0109	0.0011	0.0001	0.0000
5	0.2006	0.0204	0.0020	0.0006	0.0001
8	0.2720	0.0284	0.0028	0.0003	0.0001
11	0.4260	0.0449	0.0045	0.0009	0.0001
12	0.4490	0.0472	0.0047	0.0013	0.0001
13	0.5276	0.0560	0.0056	0.00	0.0001
16	0.6700	0.0696	0.0069	0.0161	0.0004
19	0.8713	0.1009	0.0102	0.8818	0.0004
20	0.9102	0.1069	0.0109	0.8755	0.0001
21	0.9336	0.1103	0.0112	0.9226	0.0003

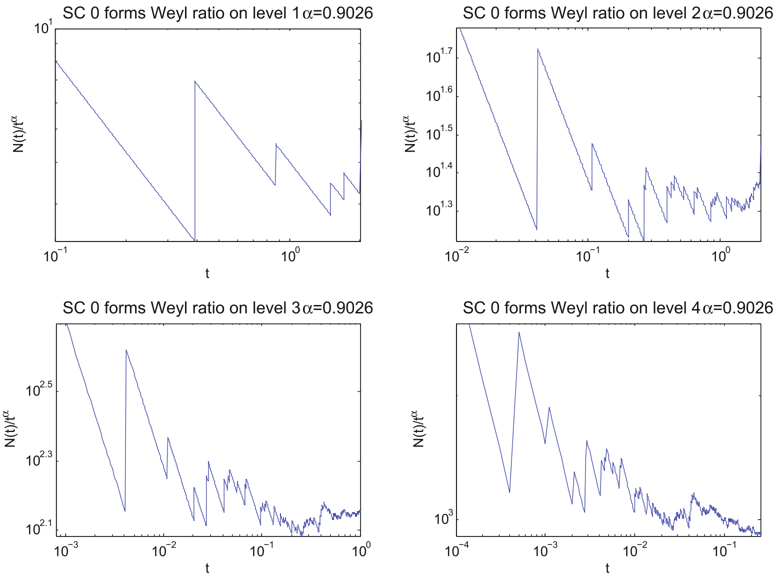


Fig. 8 Weyl ratio of SC 0-forms on level 1, 2, 3, 4 with $\alpha = 0.9026$

3 1-Forms on the Sierpinski Carpet

The 1-forms on Γ_m are functions on the edges of Γ_m, with $f_1^{(m)}(-e_1^{(m)}) = -f_1^{(m)}(e_1^{(m)})$ if $-e_1^{(m)}$ denotes the edge $e_1^{(m)}$ with opposite orientation. If L denotes any oriented path made up of edges, we may integrate $f_1^{(m)}$ over L by summing:

$$\int_L df_1^{(m)} = \sum_{e_1^{(m)} \subseteq L} f_1^{(m)}(e_1^{(m)}). \tag{17}$$

In particular, if $f_1^{(m)} = df_0^{(m)}$ then

$$\int_L df_0^{(m)} = f_0^{(m)}(b) - f_0^{(m)}(a) \tag{18}$$

where b and a denote the endpoints of L.

The Hodge decomposition splits this space $\Lambda_1^{(m)}$ of 1-forms into three orthogonal pieces.

$$\Lambda_1^{(m)} = d_0^{(m)}\Lambda_0^{(m)} \oplus \delta_2^{(m)}\Lambda_2^{(m)} \oplus \mathscr{H}_1^{(m)} \tag{19}$$

The map $d_0^{(m)} : \Lambda_0^{(m)} \to \Lambda_1^{(m)}$ has the 1-dimensional kernel consisting of constants, and the map $\delta_2^{(m)} : \Lambda_2^{(m)} \to \Lambda_1^{(m)}$ has zero kernel. A dimension count shows

$$dim\mathscr{H}_1^{(m)} = 1+8+...+8^{m-1} = \frac{8^m-1}{7}. \tag{20}$$

The Laplacian $-\Delta_1^{(m)}$ respects the decomposition, with

$$\begin{aligned} &-\Delta_1^{(m)}(d_0^{(m)}f_0^{(m)}) = d_0^{(m)}(-\Delta_0^{(m)}f_0^{(m)}) = d_0^{(m)}\delta_1^{(m)}(d_0^{(m)}f_0^{(m)}) \\ &-\Delta_1^{(m)}(\delta_2^{(m)}f_2^{(m)}) = \delta_2^{(m)}(-\Delta_2^{(m)}f_1^{(m)}) = \delta_2^{(m)}d_1^{(m)}(\delta_2^{(m)}f_2^{(m)}) \\ &-\Delta_1^{(m)}\mid_{\mathscr{H}_1^{(m)}} = 0. \end{aligned} \tag{21}$$

Although we may write $-\Delta_1^{(m)} = d_0^{(m)}\delta_1^{(m)} + \delta_2^{(m)}d_1^{(m)}$, in fact (21) is more informative. In particular, it shows that the spectrum of $-\Delta_1^{(m)}$ is just a union of the nonzero eigenvalues of $-\Delta_0^{(m)}$ and the eigenvalues of $-\Delta_2^{(m)}$, together with 0 with multiplicity given by (20). As such, it has no independent interest, and we will not present a table of its values. Also, when we discuss later the question of renormalizing in order to pass to the limit as $m \to \infty$, we will want to use a different factor for the two terms.

The main object of interest is the space $\mathscr{H}_1^{(m)}$ of harmonic 1-forms. We note that the dimension given by (20) is exactly equal to the number of homology generating cycles, one for each square deleted in the construction of SC up to level m. Thus the integrals

$$\int_{\gamma_j} h_1^{(m)} \tag{22}$$

as γ_j varies over the cycles and $h_1^{(m)}$ varies over a basis of $\mathscr{H}_1^{(m)}$ give a cohomology/homology pairing.

Conjecture 3.1 1 *For each m, the matrix* (22) *is invertible.*

If this conjecture is valid (we have verified it for $m \leq 4$) then we may define a canonical basis $h_k^{(m)}$ of $\mathscr{H}_1^{(m)}$ by the conditions

$$\int_{\gamma_j} h_k^{(m)} = \delta_{jk}. \tag{23}$$

We are particularly interested in the consistency among these harmonic 1-forms as m varies. The cycles at level m contain all the cycles from previous levels and, in addition, the $8^{(m-1)}$ cycles around the level m deleted squares. Thus we can order the cycles consistently from level to level. Also, a 1-form $f_1^{(m)}$ in $\Lambda^{(m)}$ can be restricted to a 1-form in $\Lambda^{(m-1)}$ by defining

$$Rf_1^{(m)}(e_1^{(m-1)}) = \int_{e_1^{(m-1)}} f_1^{(m)},$$

in other words summing the values of $f_1^{(m)}$ on the three level m edges that make up $e_1^{(m-1)}$. Thus we may compare the $\Lambda_1^{(m-1)}$ 1-forms $h_k^{(m-1)}$ and $Rh_k^{(m)}$ for values of k where γ_k is a level $m-1$ cycle. If these are close, we may hope to define a harmonic 1-form on SC by $\lim\limits_{m\to\infty} h_k^{(m)}$. Note that $Rh_k^{(m)}$ will not be a harmonic 1-form in $\mathscr{H}_1^{(m-1)}$. It is easy to see that the equation $d_1^{(m)} h_k^{(m)} = 0$ and the fact that (22) is zero for all the level m cycles implies (by addition) $d_1^{(m-1)} Rh_k^{(m)} = 0$. However, there is no reason to believe that $\delta_1^{(m-1)} Rh_k^{(m)}$ should be zero.

For a graphical display of the numerical data we computed for some of the functions $h_k^{(m)}$ with $k=1$ and their restrictions see website [25]. (We rounded the decimal expansions and multiplied by 10^4 so that all values are integers.) The condition $d_1 h_k^{(m)}(e_2^{(m)}) = 0$ says that the sum of $h_k^{(m)}$ on the four edges of the square (with appropriate ± signs) vanishes, or equivalently, the sum on the bottom and right edge equals the sum on top and left edge. A similar condition gives (23). The condition $\delta_1 h_k^{(m)}(e_0^{(m)}) = 0$ says that the weighted sum of $h_k^{(m)}$ on the incoming edges at the vertex $e_0^{(m)}$ equals the weighted sum on the outgoing edges, with weights given in Fig. 4 .

To quantify the rate of convergence, we give in Table 3 the values of $\|h_k^{(m-1)} - Rh_k^{(m)}\|_2$, where we use an L^2 norm on $E_1^{(m-1)}$.

Table 3 Values of $\|h_k^{(m-1)} - Rh_k^{(m)}\|_2$

k	1	2	3	4	5	6	7	8	9
$\|\|h_k^{(2)} - Rh_k^{(1)}\|\|$	0.0303								
$\|\|h_k^{(3)} - Rh_k^{(2)}\|\|$	0.0197	0.0292	0.0383	0.0292	0.0383	0.0383	0.0292	0.0383	0.0197
$\|\|h_k^{(4)} - Rh_k^{(3)}\|\|$	0.0140	0.0206	0.0224	0.0206	0.0224	0.0224	0.0206	0.0224	0.0206

Another observation is that the size of $h_k^{(m)}$ tends to fall off as the edge moves away from the cycle γ_k. This is not a very rapid decay, however.

Since 1-forms are functions on edges, it appears difficult to display the data graphically. However, there is another point of view, of independent interest, that would enable us to "see" harmonic 1-forms graphically. Note that if $f_0^{(m)}$ is a harmonic function, then $d_0 f_0^{(m)}$ is a harmonic 1-form. This is not interesting globally, since the only harmonic functions are constant. But it is interesting locally, and we can obtain harmonic 1-forms by gluing together 1-forms $d_0 f_0^{(m)}$ for different functions $f_0^{(m)}$ that are locally harmonic. We consider *harmonic mappings* taking values in the circle $\mathbb{R}/\mathbb{Z}$. Such a mapping is represented locally by a harmonic function $f_0^{(m)}$, but when we piece the local representations globally the values may change by an additive integer constant. The additive constant will not change $d_0 f_0^{(m)}$, so this will be a global harmonic 1-form. Again, adding a global constant to $f_0^{(m)}$ will have no effect on $d_0 f_0^{(m)}$. Thus, for each basis element $h_k^{(m)}$ we can construct a harmonic mapping $f_k^{(m)}$ by setting it equal to 0 at the lower left corner of SG, and then integrating using (18) to successively extend its values to vertices at the end of an edge where the value at the other endpoint has been determined. For each of the cycles γ_k we can find "cut line" so that the function $f_k^{(m)}$ is single valued away from the cut line, and only satisfies the equation $\delta_1^{(m)} d_0^{(m)} f_k^{(m)}(e_0^{(m)}) = 0$ at the vertices $e_0^{(m)}$ along the cut line if we use different values across the cut line.

Another fundamental question is the behavior of the restriction of a harmonic 1-form to a line segment. Suppose that L is a horizontal or vertical line segment of length one in SC (of course shorter line segments are also of interest, but the answers are expected to be the same). We regard $h_k^{(m)}$ on L as a signed measure on L via (17). Do we obtain a measure in the limit as $m \to \infty$? If so, is it absolutely continuous with respect to Lebesgue measure on L? In terms of the restriction of the harmonic mapping $f_k^{(m)}$ to L (we choose L to avoid the cut line), these questions become: is it of bounded variation, and if so is it an absolutely continuous function? Of course we can't answer these questions about the limit, but we can get a good sense by observing the approximations. In Fig. 9 we graph the restrictions of $h_k^{(m)}$ to L in some cases as m varies. For a quantitative approach to the first question we compute the total variation of the approximations. The results are show in Table 4. Note that in this figure and all subsequent graphs of restrictions of 1-forms to lines, we are displaying the graphs of running totals starting at the left end of the interval. Thus in the limit we would hope to get a function of bounded variation (or perhaps even an absolutely continuous function) whose derivative is a measure on the line.

Of course the same questions are of interest for 1-forms that are eigenfunctions of the Laplacian. In Fig. 10, 11 and Table 5 we give the analogous results for $d_0^{(m)} f_0^{(m)}$ and $\delta_2^{(m)} f_2^{(m)}$ where $f_0^{(m)}$ and $f_2^{(m)}$ are eigenforms for the Laplacians $-\Delta_0^{(m)}$ and $-\Delta_2^{(m)}$.

4 2-Forms on the Sierpinski Carpet

In principle, we should think of 2-forms on SC simply as measures. In the Γ_m approximation we have 8^m squares in $E_2^{(m)}$, and our 2-forms $f_2^{(m)}$ assign values to these squares, which may be identified with the m-cells in SC that lie in these squares. It is not clear a priori what class of measures we should consider; the simplest choice is the set of measures absolutely continuous with respect to the standard self-similar measure μ.

The Laplacian $-\Delta_2^{(m)}$ on $\Lambda_2^{(m)}$ is quite different from other Laplacians considered here or elsewhere. In fact, it is the sum of a difference operator and diagonal operator. For a fixed square $e_2^{(m)} \in E_2^{(m)}$ we have

$$d_1^{(m)} \delta_2^{(m)} f_2^{(m)}(e_2^{(m)}) = \sum_{e_1^{(m)} \subseteq e_2^{(m)}} sgn(e_1^{(m)}, e_2^{(m)}) \delta_2^{(m)} f_2^{(m)}(e_1^{(m)})$$

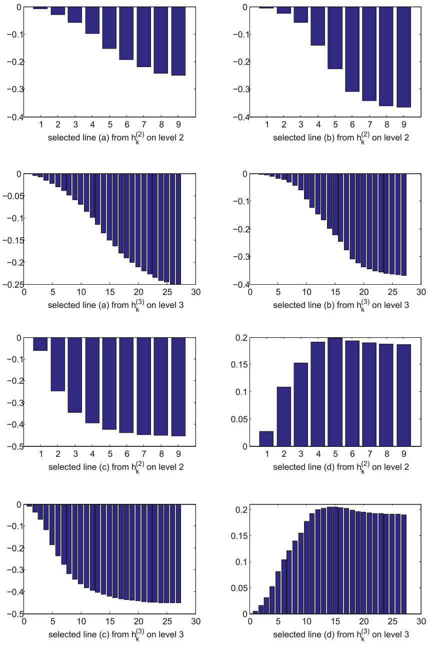


Fig. 9 Restrictions of $h_k^{(m)}$ to L with $m = 2,3$

Table 4 Total variation of SC harmonic one forms along horizontal lines at level 1, level 2 and level 3

		Level 1	Level 2	Level 3
Line 1	Form 1	0.00	0.00	0.00
	Form 2	0.4522	0.4522	0.4522
	Form 3	0.1442	0.1442	0.1442
	Form 4	0.0478	0.0478	0.0478
Line 2	Form 1	0.3678	0.3678	0.3678
	Form 2	0.2057	0.2188	0.2188
	Form 3	0.82	0.82	0.82
	Form 4	0.0741	0.0741	0.0741
Line 3	Form 1	0.3678	0.3678	0.3678
	Form 2	0.0741	0.0741	0.0741
	Form 3	0.82	0.82	0.82
	Form 4	0.2057	0.2188	0.2188
Line 4	Form 1	0.00	0.00	0.00
	Form 2	0.0478	0.0478	0.0478
	Form 3	0.1442	0.1442	0.1442
	Form 4	0.4522	0.4522	0.4522

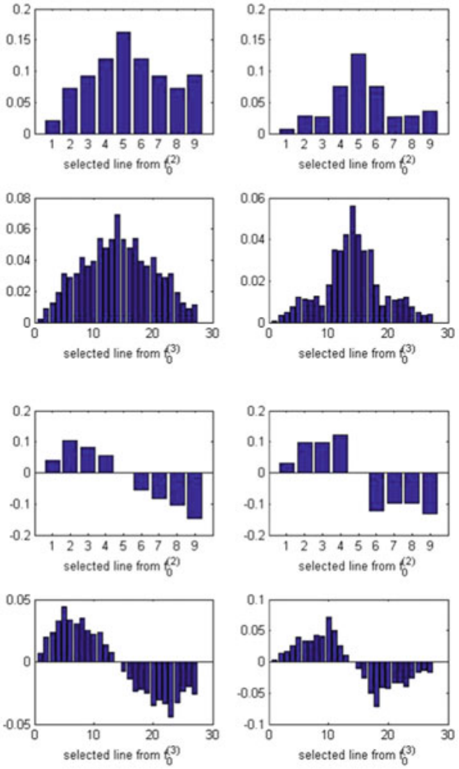


Fig. 10 Restrictions of $d_0^{(m)} f_0^{(m)}$ to L

where the sum is over the four edges of the square. However, $\delta_2^{(m)} f_2^{(m)}(e_2^{(m)})$ depends on the nature of the edge $e_1^{(m)}$, which may bound one or two squares:

$$\delta_2^{(m)} f_2^{(m)}(e_1^{(m)}) = \begin{cases} 4sgn(e_1^{(m)}, \tilde{e}_2^{(m)}) f_2^{(m)}(\tilde{e}_2^{(m)}) \\ \text{if } e_1^{(m)} \text{ bounds only } \tilde{e}_2^{(m)} \\ 2\left(sgn(e_1^{(m)}, \tilde{e}_2^{(m)}) f_2^{(m)}(\tilde{e}_2^{(m)}) + sgn(e_1^{(m)}, \tilde{\tilde{e}}_2^{(m)}) f_2^{(m)}(\tilde{\tilde{e}}_2^{(m)})\right) \\ \text{if } e_1^{(m)} \text{ bounds } \tilde{e}_2^{(m)} \text{ and } \tilde{\tilde{e}}_2^{(m)}. \end{cases} \tag{24}$$

Let $N(e_2^{(m)})$ denote the number of squares adjacent to $e_2^{(m)}$ in Γ_m (2, 3, or 4). Then

$$\begin{aligned} -\Delta_2^{(m)} f_2^{(m)}(e_2^{(m)}) &= d_1^{(m)} \delta_2^{(m)} f_2^{(m)}(e_2^{(m)}) \\ &= 4\left(4 - N(e_2^{(m)})\right) f_2^{(m)}(e_2^{(m)}) + 2 \sum_{\tilde{e}_2^{(m)} \sim e_2^{(m)}} \left(f_2^{(m)}(e_2^{(m)}) - f_2^{(m)}(\tilde{e}_2^{(m)})\right) \end{aligned} \tag{25}$$

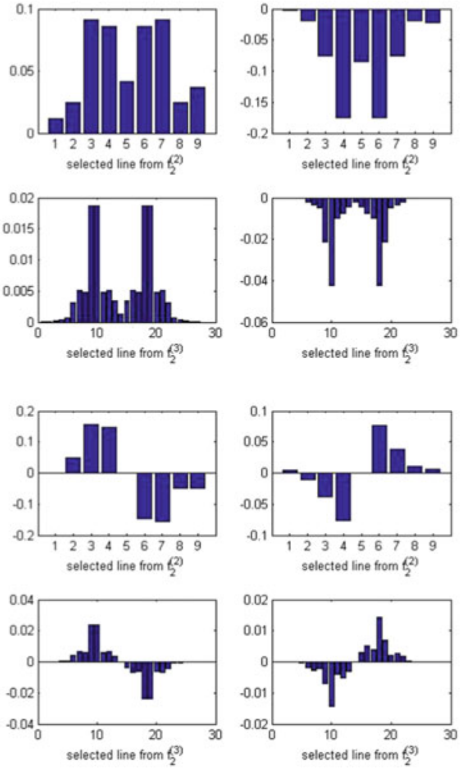


Fig. 11 Restrictions of $\delta_2^{(m)} f_2^{(m)}$ to L

Table 5 Total variation of $d_0^{(m)} f_0^{(m)}$ and $\delta_2^{(m)} f_2^{(m)}$ along horizontal lines at level 2 and level 3

D		Level 1	Level 2	Level 3	DEL		Level 1	Level 2	Level 3
Line 1	Form 1	0.8845	0.8845	0.8845	Line 1	Form 1	0.1319	0.1319	0.1319
	Form 2	0.5010	0.6201	0.6348		Form 2	0.0663	0.1232	0.1289
	Form 3	0.6758	0.8093	0.8203		Form 3	0.1701	0.1701	0.1701
	Form 4	0.7538	0.7538	0.7826		Form 4	0.0849	0.1584	0.1660
Line 2	Form 1	0.4588	0.4588	0.4588	Line 2	Form 1	0.1960	0.1960	0.1960
	Form 2	0.5010	0.7970	0.8178		Form 2	0.0658	0.1849	0.19
	Form 3	0.3688	0.4376	0.4484		Form 3	0.0861	0.0861	0.0861
	Form 4	0.5387	0.5387	0.5454		Form 4	0.0284	0.0754	0.0819
Line 3	Form 1	0.4588	0.4588	0.4588	Line 3	Form 1	0.1960	0.1960	0.1960
	Form 2	0.5010	0.7970	0.8178		Form 2	0.0658	0.1849	0.19
	Form 3	0.3688	0.4376	0.4484		Form 3	0.0861	0.0861	0.0861
	Form 4	0.5387	0.5387	0.5454		Form 4	0.0284	0.0754	0.0819
Line 4	Form 1	0.8845	0.8845	0.8845	Line 4	Form 1	0.1319	0.1319	0.1319
	Form 2	0.5010	0.6201	0.6348		Form 2	0.0663	0.1232	0.1289
	Form 3	0.6758	0.8093	0.8203		Form 3	0.1701	0.1701	0.1701
	Form 4	0.7538	0.7538	0.7826		Form 4	0.0849	0.1584	0.1660

Note that
$$-D_2^{(m)} f_2^{(m)}(e_2^{(m)}) = \sum_{\tilde{e}_2^{(m)} \sim e_2^{(m)}} \left(f_2^{(m)}(e_2^{(m)}) - f_2^{(m)}(\tilde{e}_2^{(m)}) \right)$$

is exactly the graph Laplacian for the cell graph on level m of SC. This Laplacian was first studied in [19], and it was proved in [5] that when appropriately normalized it converges to the essentially unique self-similar Laplacian on SC. This was used as the basis for extensive numerical investigations in [6]. In other words, if $f_2^{(m)}(e_2^{(m)}) = \int_{e_2^{(m)}} f_0 d\mu$ for some function f_0 in the domain of the Laplacian on SC, then $\lim_{m\to\infty} r^m D_2^{(m)} f_2^{(m)} = c(\Delta f_0) d\mu$ for $r \approx 10.01$.

However, according to (25) we have
$$-\Delta_2^{(m)} = M^{(m)} - D_2^{(m)}, \tag{26}$$

where $M^{(m)}$ is the operator of multiplication by the function ϕ_m given by
$$\phi_m = 4\left(4 - N(e_2^{(m)})\right).$$

Note that ϕ_m takes on values 0, 4, 8. If we were to renormalize $M^{(m)}$ by multiplying by r^m the result would surely diverge. In other words, if there is any hope of obtaining a limit for a class of measures, it would have to be $\lim_{m\to\infty}(-\Delta_2^{(m)})$ without renormalization. Of course the sequence of functions $\{\phi_m\}$ does not converge, so it seems unlikely that we could make sense of $\lim_{m\to\infty}(M^{(m)})$. Thus, although $M^{(m)}$ is clearly the major contributor to the sum (26), both operators must play a role if the limit is to exist.

We compute the distribution of the three values of $N(e_2^{(m)})$ as $e_2^{(m)}$ varies over the 8^m squares of level m. Let $n_3^{(m)}$ and $n_4^{(m)}$ denote the number of squares with N value 3 and 4, and $n_{2a}^{(m)}$ and $n_{2b}^{(m)}$ denote the number with $N = 2$, with neighbors on opposite sides ($n_{2a}^{(m)}$) or adjacent sides ($n_{2b}^{(m)}$). In fact, $n_{2b}^{(m)} = 4$ since this case only occurs at the four corners of SC. If $e_2^{(m-1)}$ is in one of those cases we may compute the N values on the eight subsquares as shown in Fig. 12 .

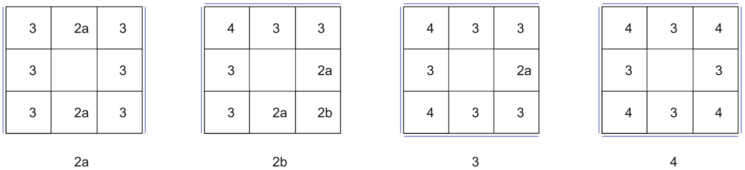


Fig. 12 Values of N on subsquares. The neighboring edges are marked by a *double line*

This gives us the recursion relation

$$\begin{pmatrix} n_{2a}^{(m)} \\ n_{2b}^{(m)} \\ n_3^{(m)} \\ n_4^{(m)} \end{pmatrix} = A \begin{pmatrix} n_{2a}^{(m-1)} \\ n_{2b}^{(m-1)} \\ n_3^{(m-1)} \\ n_4^{(m-1)} \end{pmatrix} \text{ for } A = \begin{pmatrix} 2 & 2 & 1 & 0 \\ 0 & 1 & 0 & 0 \\ 6 & 4 & 5 & 4 \\ 0 & 1 & 2 & 4 \end{pmatrix}$$

and so

$$\begin{pmatrix} n_{2a}^{(m)} \\ n_{2b}^{(m)} \\ n_3^{(m)} \\ n_4^{(m)} \end{pmatrix} = A^{m-1} \begin{pmatrix} 4 \\ 4 \\ 0 \\ 0 \end{pmatrix}.$$

We also note that (1111) is a left eigenvector of A with eigenvalue 8 (in other words column sums are 8), and this implies

$$n_{2a}^{(m)} + n_{2b}^{(m)} + n_3^{(m)} + n_4^{(m)} = 8^m$$

as required. We also observe that $\left(1063\right)^T$ is the right eigenvector with eigenvalue 8, so asymptotically

$$\begin{pmatrix} n_{2a}^{(m)} \\ n_{2b}^{(m)} \\ n_3^{(m)} \\ n_4^{(m)} \end{pmatrix} \sim 8^m \begin{pmatrix} .1 \\ 0 \\ .6 \\ .3 \end{pmatrix} \text{ as } m \to \infty.$$

Thus the operator $M^{(m)}$ has eigenvalues 0, 4, 8 with multiplicities approximately $\frac{3}{10}8^m$, $\frac{6}{10}8^m$, $\frac{1}{10}8^m$.

The spectrum of the operator $-\Delta_2^{(m)}$ is quite different. Table 6 shows the entire spectrum for $m = 1, 2$ and the beginning of the spectrum for $m = 3, 4$.
In Fig. 13 we give a graphical display of these spectra. In Fig. 14 we show the graphs of some of the early eigenfuction on levels 2, 3, 4.

The data suggests that there may be a limit

$$-\Delta_2 = \lim_{m\to\infty} (-\Delta_2^{(m)}) \tag{27}$$

for a class of measures, perhaps $L^2(d\mu)$. The limit operator would be a bounded, self-adjoint operator that is bounded away from zero, hence invertible. The spectrum would be continuous (or a mix of discrete and continuous) with support on a Cantor set.

We can give some explanations as to why we might expect the following types of limits:

$$\delta_2 = \lim_{m\to\infty} \left(\frac{8}{3}\right)^m \delta_2^{(m)} \tag{28}$$

Table 6 Eigenvalues of $-\Delta_2^{(m)}$ for m =1, 2, 3, 4

Number of eigenvalues	$m = 1$	$m = 2$	$m = 3$	$m = 4$
1	2.0000	0.9248	0.5101	0.4505
2	2.2929	0.9497	0.5102	0.4505
3	2.2929	0.9497	0.5102	0.4505
4	3.0000	0.9789	0.5104	0.4505
5	3.0000	1.3068	0.5290	0.4511
6	3.7071	1.3506	0.5291	0.4511
7	3.7071	1.3506	0.5291	0.4511
8	4.0000	1.3989	0.5292	0.4511
9		1.5764	0.61	0.4604
10		1.6355	0.64	0.4604
11		1.6355	0.64	0.4604
12		1.7134	0.67	0.4604
13		1.8573	0.6554	0.4614
14		1.9559	0.6557	0.4614
15		1.9559	0.6557	0.4614
16		2.0000	0.6561	0.4614
17		2.0593	0.9248	0.5101
18		2.0593	0.9280	0.5102
19		2.0786	0.9280	0.5102
20		2.1187	0.9332	0.5102
21		2.1910	0.9395	0.5102
22		2.2301	0.9443	0.5102
23		2.2929	0.9450	0.5102
24		2.2929	0.9450	0.5102
		2.3272	0.9497	0.5102
26		2.3272	0.9497	0.5102
27		2.3990	0.9527	0.5103
28		2.4422	0.9544	0.5103
29		2.5000	0.9623	0.5103
30		2.5000	0.9690	0.5103
31		2.5391	0.9690	0.5103
32		2.5391	0.9755	0.5104

Table 7 Values of $\|f_j^{(m)}|_{E_2^{(m-1)}} - f_j^{(m-1)}\|_2^2$, from level 3 to level 2 and level 4 to level 3 for eigenspaces of multiplicity one

number	$m = 2$	$m = 3$	$m = 4$	3to2	4to3
1	0.9248	0.5101	0.4505	0.0726	0.0196
4	0.9789	0.5104	0.4505	0.0532	0.0190
5	1.3068	0.5290	0.4511	0.11	0.0114
8	1.3989	0.5292	0.4511	0.0670	0.0106
9	1.5764	0.61	0.4604	0.1709	0.0334
12	1.7134	0.67	0.4604	0.1591	0.0317
13	1.8573	0.6554	0.4614	0.2796	0.0243

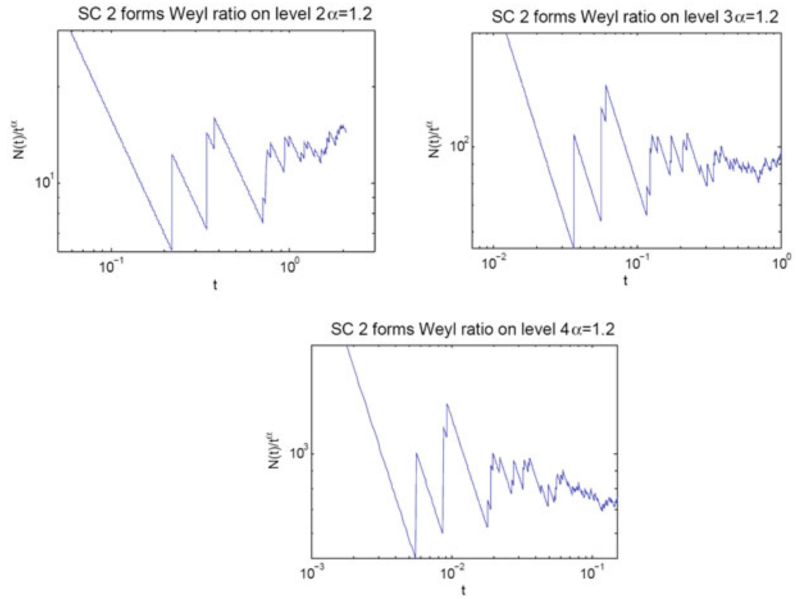


Fig. 13 Weyl ratio for spectra of SC 2 forms with α=1.2

$$d_1 = \lim_{m\to\infty} \left(\frac{3}{8}\right)^m d_1^{(m)}$$

which are consistent with (27). Suppose $f_2 = f d\mu$ for a reasonable function f, and define $f_2^{(m)}(e_2^{(m)}) = \int_{e_2^{(m)}} f d\mu$. Then $f_2^{(m)}(e_2^{(m)})$ is on the order of 8^{-m}. In the definition of $\delta_2^{(m)} f_2^{(m)}$ in (24) we note that when $e_1^{(m)}$ bounds only one square, $\delta_2^{(m)} f_2^{(m)}(e_1^{(m)})$ is also on the order of 8^{-m}, so multiplying by $(8/3)^m$ gives a value on the order of 3^{-m}, which is reasonable for a measure on a line segment L containing $e_1^{(m)}$. On the other hand, if $e_1^{(m)}$ bounds two squares, then the *sgn* function has opposite signs so $(8/3)^m \delta_2^{(m)} f_2^{(m)}(e_1^{(m)})$ is close to zero. If we assume the function f is continuous then the limit in (28) will give a measure on L equal to $4f|_L dt$ on the portion of L with squares on only one side, and zero on the portion of L with squares on both sides.

Next suppose the f_1 is a measure on each line L in SC that has

$$|f_1(e_1^{(m)})| \le c3^{-m}. \tag{29}$$

Fix a square $e_2^{(m)}$ on level n, and write it as a union of 8^{m-n} squares on level m. We want to define

$$d_1 f_1(e_2^{(n)}) = \lim_{m\to\infty} \left(\frac{3}{8}\right)^m \sum_{e_2^{(m)} \subseteq e_2^{(n)}} d_1^{(m)} f_1^{(m)}(e_2^{(m)}). \tag{30}$$

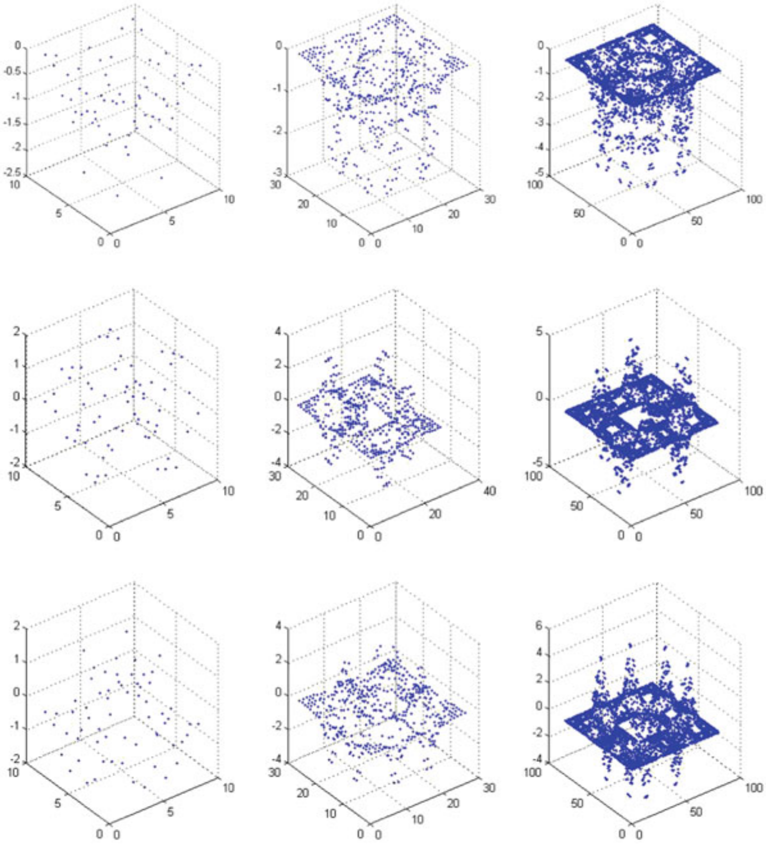


Fig. 14 Graphs of early eigenfunctions of SC 2 forms on level 2, 3, 4

Does this make sense? Because of the cancellation from the *sgn* function on opposite sides of an edge, $\sum_{e_2^{(m)} \subseteq e_2^{(n)}} d_1^{(m)} f_1^{(m)}(e_2^{(m)})$ is just the measure of the boundary of $e_2^{(n)}$ decomposed to level m. This boundary is the union of the four edges $e_1^{(n)}$ on the outside of the square, the four edges around the inner deleted square on level $n+1$, and in general the $4 \cdot 8^{k-1}$ edges around the 8^k deleted squares on level $n+k$, for $k \leq m-n$.(see Fig. 15 for $m = n+2$)
Using the estimate (29) we have

$$\left| \left(\frac{3}{8}\right)^m \sum_{e_2^{(m)} \subseteq e_2^{(n)}} d_1^{(m)} f_1^{(m)}(e_2^{(m)}) \right| \leq 4c \left(\frac{3}{8}\right)^m \left(\frac{1}{3^n} + \frac{8}{3^{n+1}} + \frac{8^2}{3^{n+2}} + \ldots + \frac{8^m}{3^m} \right)$$
$$\leq c.$$

Thus the terms on the right side of (30) are uniformly bounded, so it is not unreasonable to hope that the limit exists.

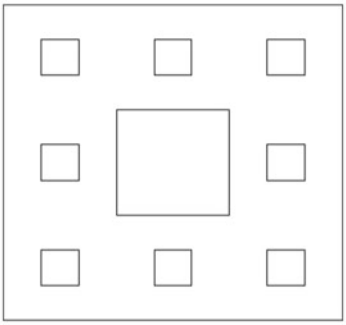


Fig. 15 Edges in sum for $m = n + 2$

5 0-Forms and 2-Forms on the Magic Carpet

The vertices $\tilde{E}_0^{(m)}$ of $\tilde{\Gamma}_m$ split into singular $\tilde{E}_{0s}^{(m)}$ and nonsingular vertices $\tilde{E}_{0n}^{(m)}$. Let $V_m = \#\tilde{E}_0^{(m)}$, $S_m = \#\tilde{E}_{0s}^{(m)}$, and $N_m = \#\tilde{E}_{0n}^{(m)}$. Then $S_m = 1 + 8 + 8^2 + \ldots + 8^{m-1} = (8^m - 1)/7$, since each time we remove a square and identify its boundaries we create a single singular vertex. To compute the other two counts, we note that each of the 8^m squares has 4 vertices, and singular vertices arise in 12 different ways, while nonsingular vertices arise in 4 different ways. Thus

$$12S_m + 4N_m = 8^m$$

Solving for N_m we obtain

$$N_m = \frac{4 * 8^m + 3}{7}, V_m = \frac{5 * 8^m + 2}{7}.$$

Asymptotically, one-fifth of all vertices are singular.

The Laplacian $-\tilde{\Delta}_0^{(m)}$ given by (1) has V_m eigenvalues, starting at $\lambda_0 = 0$ corresponding to the constants. In Table 8 we give the beginning of the spectrum for $m = 1, 2, 3, 4$ along with the ratios from levels 3 to 2 and 4 to 3.

We note that the ratios are around $r = 6....$, so we expect

$$-\tilde{\Delta}_0 = \lim_{m\to\infty} r^m \left(-\tilde{\Delta}_0^{(m)}\right)$$

to define a Laplacian on MC. At present there is no proof that MC has a Laplacian, so our data is strong experimental evidence that $-\tilde{\Delta}_0$ exists. What is striking is that

Table 8 Beginning of the spectrum of $-\tilde{\Delta}_0^{(m)}$ and ratios

λ_1	λ_2	λ_3	λ_4	λ_2/λ_3	λ_3/λ_4
0.0000	0.0000	0.0000	0.0000		
1.5000	0.2877	0.0458	0.0071	6.2809	6.4099
1.7047	0.4458	0.0734	0.0114	6.0699	6.4402
2.5000	0.4458	0.0734	0.0114	6.0699	6.4402
2.5000	0.4496	0.0764	0.0120	5.8860	6.3642
3.1287	0.8333	0.1449	0.0230	5.7521	6.2957
	0.8595	0.1538	0.0244	5.5870	6.2971
	0.8595	0.1538	0.0244	5.5870	6.2971
	0.9862	0.1645	0.07	5.9938	6.4006
	1.0967	0.1854	0.0289	5.9150	6.4218
	1.1838	0.2175	0.0346	5.4441	6.2913
	1.3014	0.2386	0.0372	5.4555	6.4138
	1.3014	0.2386	0.0372	5.4555	6.4138
	1.5000	0.2684	0.0420	5.5881	6.3862
	1.5000	0.2684	0.0420	5.5881	6.3862
	1.5000	0.2701	0.0428	5.5541	6.3056
	1.5793	0.2849	0.0454	5.5438	6.2771
	1.5793	0.2877	0.0458	5.4898	6.2809
	1.6912	0.3407	0.0552	4.9636	6.1695
	1.7047	0.3869	0.0626	4.4058	6.1772
	1.8770	0.3922	0.0633	4.7861	6.1940
	1.8913	0.3922	0.0633	4.8228	6.1940
	1.9661	0.4169	0.0685	4.7164	6.0849
	1.9661	0.43	0.0705	4.5462	6.1319

$r < 8$, in contrast to the factor of ≈ 10.01 for SC. Since the measure renormalization factor is 8^m for both carpets, we conclude that the energy renormalization factor for MC would have to be less than one. In Euclidean spaces or manifolds, this happens in dimensions greater than two. Note that the unrenormalized energy on level m would be

$$\tilde{E}^{(m)}(f_0) = \frac{1}{2 * 8^m} \sum_{e_1^{(m)} \in \tilde{E}_1^{(m)}} |d_0 f_0(e_1^{(m)})|^2,$$

so $\widetilde{\mathscr{E}}^{(m)} = r^m \tilde{E}^{(m)}$ means that the graph energy $\sum_{e_1^{(m)} \in \tilde{E}_1^{(m)}} |d_0 f_0(e_1^{(m)})|^2$ is multiplied by $(r/8)^m$ before taking the limit.

In Fig. 16 we show the graphs of selected eigenfunctions on levels 2, 3, 4. Again we quantify the rate of convergence as in the case of SC by giving in Table 9 the values of $\|f_j^{(m)}|_{\tilde{E}_0^{(m-1)}} - f_j^{(m-1)}\|_2^2$.

We give similar data for the spectrum of $-\tilde{\Delta}_2^{(m)}$ defined by (12). In order to make the comparison with Table 8 clear, we give Table 10 the eigenvalues of $-\tilde{\Delta}_2^{(m)}$ multiplied by 0.3221 and ratios. In Table 11 we show quantitative rates of convergence.

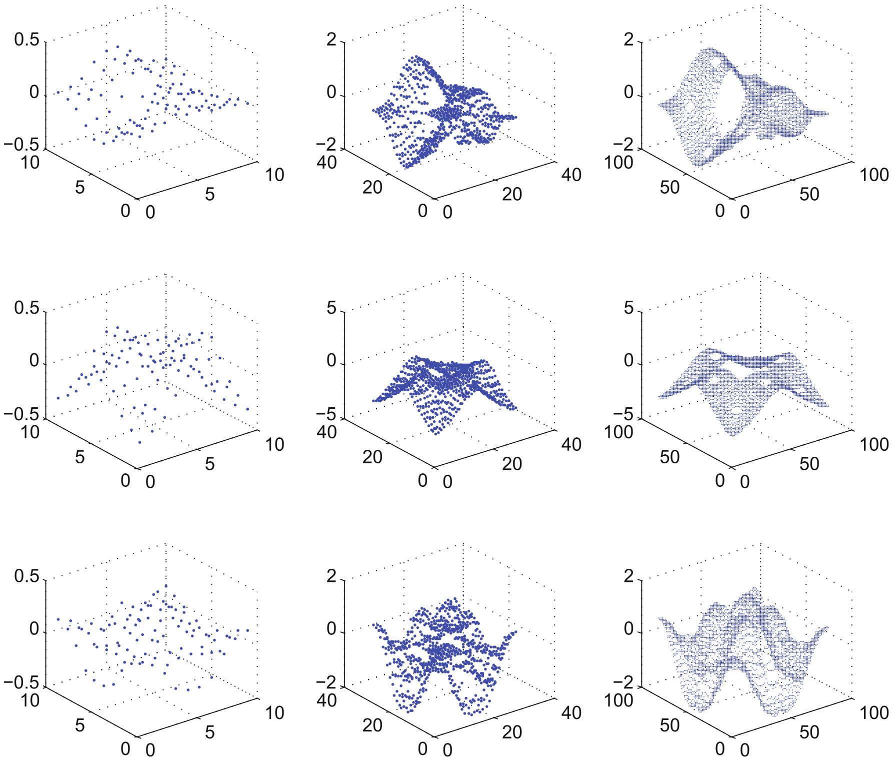


Fig. 16 Graphs of selected eigenfunctions of MC 0 forms on levels 2, 3, 4

Table 9 Values of $\|f_j^{(m)}|_{\tilde{E}_0^{(m-1)}} - f_j^{(m-1)}\|_2^2$

number of eigenvalues	$m=2$	$m=3$	$m=4$	3 to 2	4 to 3
1	0.0000	0.0000	0.0000	0.0000	0.0000
2	0.2877	0.0458	0.0071	0.0000	0.0000
5	0.4496	0.0764	0.0120	0.0001	0.0000
6	0.8333	0.1449	0.0230	0.0002	0.0000
9	0.9862	0.1645	0.07	0.0004	0.0000
10	1.0967	0.1854	0.0289	0.0008	0.0000

In Fig. 17 we show graphs of eigenfunctions. In Fig. 18 we show graphs of $*f_0^{(m)}$ when $f_0^{(m)}$ is an eigenfunction in Fig. 16, and in Fig. 19 we show graphs of $*f_2^{(m)}$ when $f_2^{(m)}$ is an eigenfunction in Fig. 17. In Fig. 20 and 21, we give the Weyl ratios of the eigenvalues of the 0 forms and 2 forms.

Because miniaturization holds we expect a value $\alpha = \frac{\log 8}{\log r} \approx \frac{\log 8}{\log 6}$. We have renormalized the eigenvalues as $\{r^m \lambda_j^{(m)}\}$, with the expectation that in the limit as $m \to \infty$ we obtain eigenvalues of $-\tilde{\Delta}_0$ and $-\tilde{\Delta}_2$ on MC.

Table 10 Eigenvalues of $-\tilde{\Delta}_2^{(m)}$ multiplied by 0.3221 and ratios

$m=1$	$m=2$	$m=3$	$m=4$	λ_2/λ_3	λ_3/λ_4
0.0000	0.0000	0.0000	0.0000		
1.2885	0.2841	0.0463	0.0071	6.1409	6.4749
2.1054	0.4449	0.0721	0.0112	6.1710	6.4514
2.1054	0.4449	0.0721	0.0112	6.1710	6.4514
2.5770	0.4889	0.0776	0.0119	6.2970	6.5040
3.8655	0.9193	0.1504	0.0232	6.1139	6.4904
4.3371	0.9553	0.1580	0.0244	6.0455	6.4640
4.3371	0.9553	0.1580	0.0244	6.0455	6.4640
	0.9653	0.1623	0.03	5.9494	6.4158
	1.0110	0.1792	0.0283	5.6430	6.3393
	1.2077	0.2192	0.0345	5.5097	6.3554
	1.2135	0.2226	0.0356	5.4505	6.32
	1.2135	0.2226	0.0356	5.4505	6.32
	1.2885	0.2679	0.0418	4.8091	6.4058
	1.5007	0.2679	0.0418	5.6010	6.4058
	1.5562	0.2699	0.04	5.7662	6.3573
	1.5562	0.2841	0.0452	5.4771	6.2923
	1.6334	0.2878	0.0463	5.6757	6.2202
	1.7959	0.3541	0.0560	5.0715	6.3241
	1.9528	0.3946	0.0628	4.9493	6.2829
	1.9528	0.3969	0.0630	4.9204	6.3040
	2.0288	0.3969	0.0630	5.1118	6.3040
	2.0840	0.4358	0.0693	4.78	6.2886
	2.1054	0.4410	0.0712	4.7737	6.1952
	2.1054	0.4410	0.0712	4.7737	6.1952
	2.1700	0.4449	0.0721	4.8774	6.1710
	2.1700	0.4449	0.0721	4.8774	6.1710
	2.3521	0.4529	0.0728	5.1929	6.2232
	2.3579	0.4847	0.0770	4.8648	6.2915
	2.4038	0.4889	0.0776	4.9172	6.2970
	2.4038	0.4971	0.0796	4.8356	6.2415
	2.4760	0.4971	0.0796	4.9807	6.2415
	2.4760	0.5364	0.0870	4.6156	6.1657

Table 11 Values of $\|f_j^{(m)}|_{\tilde{E}_2^{(m-1)}} - f_j^{(m-1)}\|_2^2$

Number of eigenvalues	$m=2$	$m=3$	$m=4$	3 to 2	4 to 3
1	0	0	0	0	0
2	0.2205	0.0359	0.0055	0.0006	0.0001
5	0.3794	0.0602	0.0093	0.0012	0.0002
6	0.7135	0.1167	0.0180	0.0051	0.0003
9	0.7492	0.19	0.0196	0.0056	0.0004
10	0.7846	0.1391	0.0219	0.0018	0.0004
11	0.9373	0.1701	0.0268	0.0350	0.0011

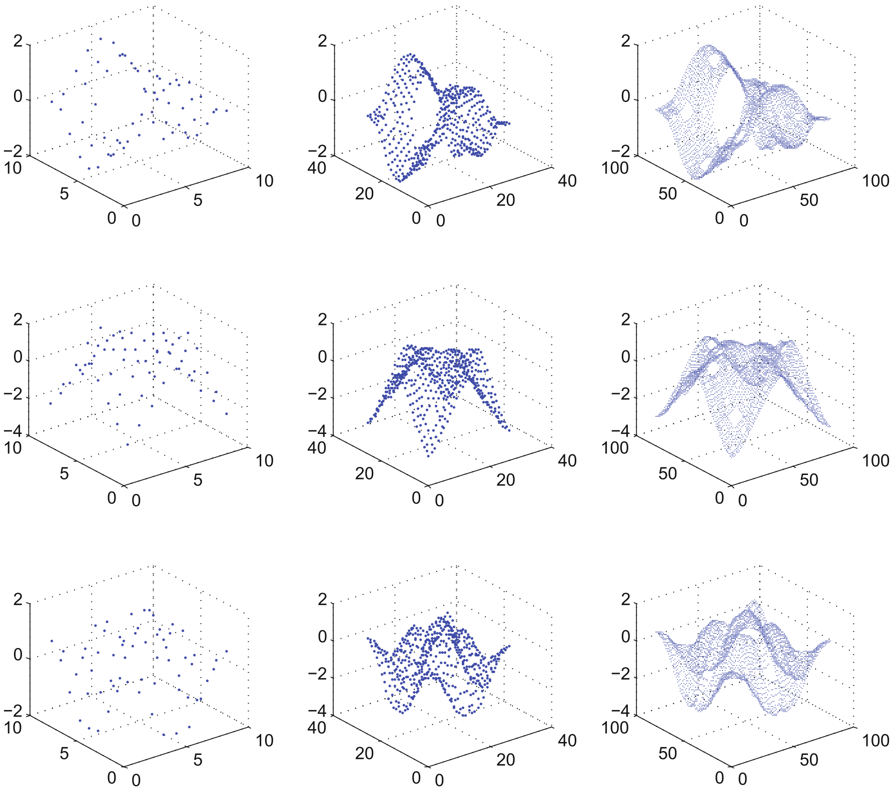


Fig. 17 Graphs of selected eigenfunctions of MC 2 forms on levels 2, 3, 4

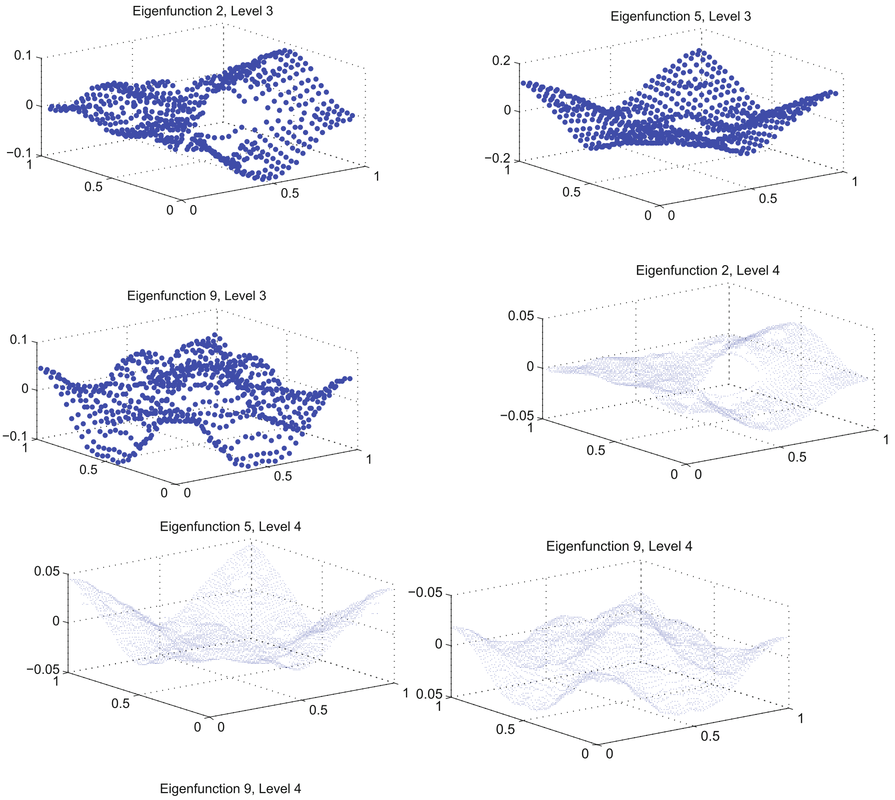


Fig. 18 Graphs of $*f_0^{(m)}$ when $f_0^{(m)}$ is an eigenfunction in Fig. 16

6 1-Forms on the Magic Carpet

We indicate briefly how the theory differs from SC. We have the analogs of (17), (18) and (19), but the dimension count is different because $\tilde{\delta}_2^{(m)}$ also has a 1-dimensional kernel, namely the constants. There are exactly $2\cdot 8^m$ edges, since each edge is the boundary of exactly two squares, so

$$dim\mathscr{H}_1^{(m)} = 2\cdot 8^m - 8^m - \frac{5\cdot 8^m+1}{7} + 2 = \frac{2\cdot 8^m+12}{7}.$$

On the other hand, the surface approximating MC on level m has genus $g = (8^m + 6)/7$, and the cycles generating the homology are in one-to-one correspondence with the horizontal and vertical identified edges. The analog of Conjecture 3.1 is true, in fact it is a well-known result in topology.

In Fig. 22, 23, and 24 we show the values of $h_1^{(2)}$, $h_1^{(3)}$ and the restriction of $h_1^{(3)}$ to $\tilde{E}_1^{(2)}$ where γ_1 is the top and bottom horizontal line. There are some surprising

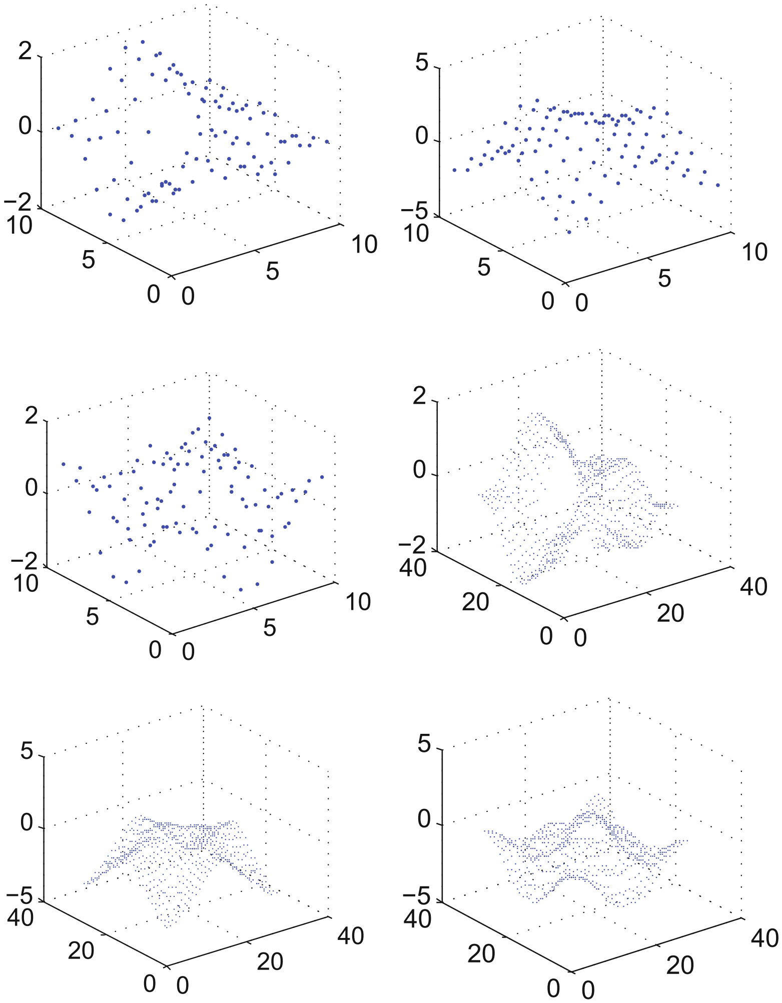


Fig. 19 Graphs of $*f_2^{(m)}$ when $f_2^{(m)}$ is an eigenfunction in Fig. 17

features of these 1-forms that may be explained by symmetry. Let R_H denote the horizontal reflection and R_V the vertical reflection about the center. Note that $R_H\gamma_1 = -\gamma_1$ because the orientation is reversed, while $R_V\gamma_1 = \gamma_1$. Since $-h_1^{(m)}(R_H x)$ and $h_1^{(m)}(R_V x)$ are harmonic 1-forms with the same integrals around cycles as $h_1^{(m)}$, it follows by uniqueness that $-h_1^{(m)}(R_H x) = h_1^{(m)}(x)$ and $h_1^{(m)}(R_V x) = h_1^{(m)}(x)$. This implies that $h_1^{(m)}$ vanishes identically along the cycle consisting of the vertical edges of the large square, and indeed any square that is symmetric with respect to R_H. Certain other vanishings of $h_1^{(m)}$ are accidental to the level m, and do not persist

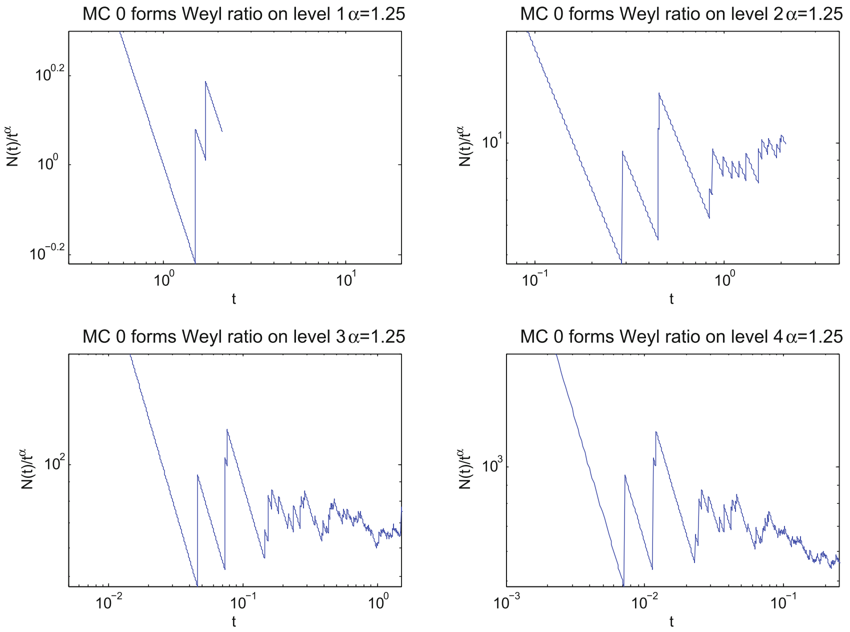


Fig. 20 Weyl ratio of MC 0 forms eigenvalues

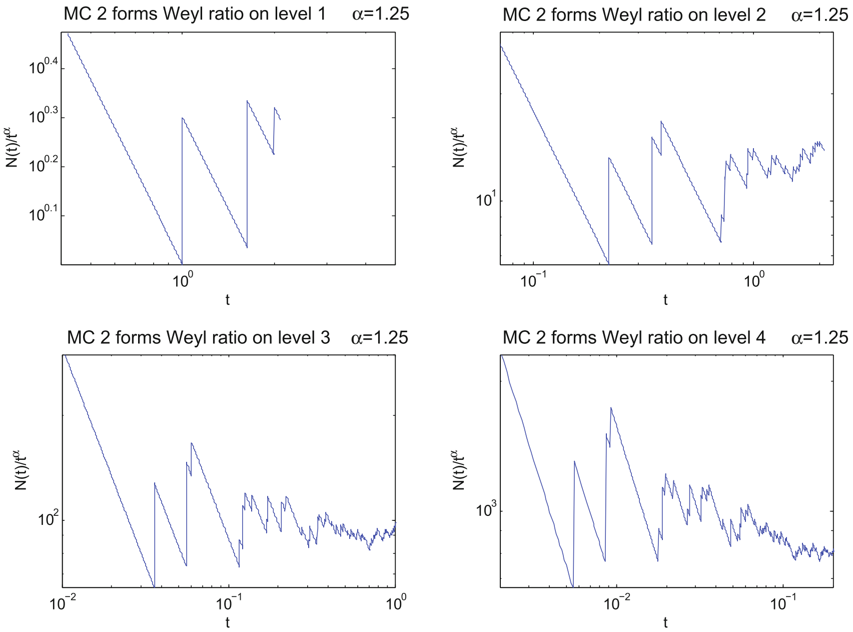


Fig. 21 Weyl ratio of MC 2 forms eigenvalues

when m increases. For example, every cycle of level m contains just a single edge, so vanishing of the integral forces vanishing on the edge. Sometimes this has a ripple effect. For example, consider the region in the top center of the level 2 graph show in Fig. 25, where the horizontal symmetry has been used in labeling edges. We obtain the equation $2a+b=0$ from the equation $\tilde{d}_1^{(2)}\tilde{h}_1^{(2)}=0$ on the small square, and the equation $2c+b=0$ from $\int_\gamma \tilde{h}_1^{(2)}=0$ on the cycle along the top of the big square. Finally, the equation $a+c-b=0$ comes from $\tilde{\delta}_1^{(2)}\tilde{h}_1^{(2)}=0$ at the indicated point. These yield $a=b=c=0$ that we see in Fig. 22, 23 and 24.

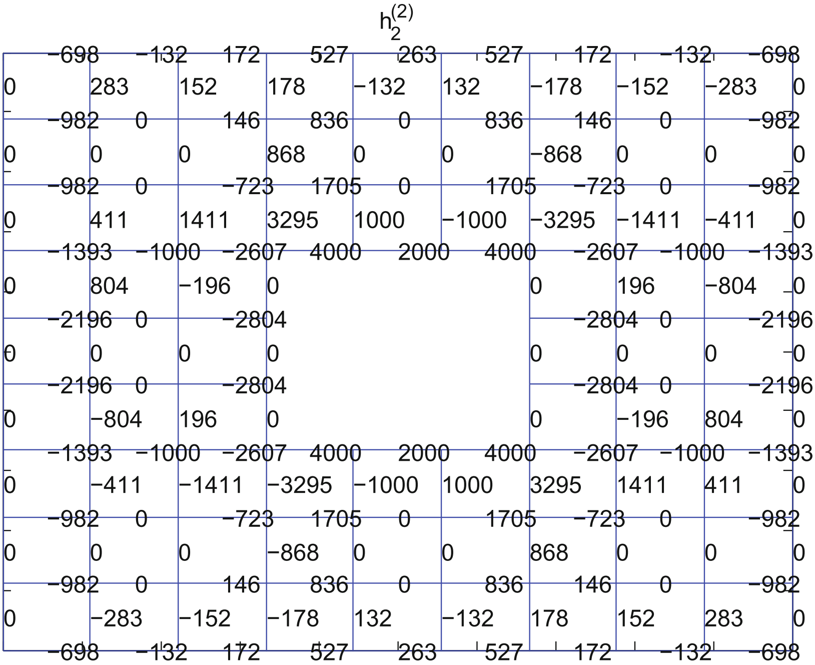
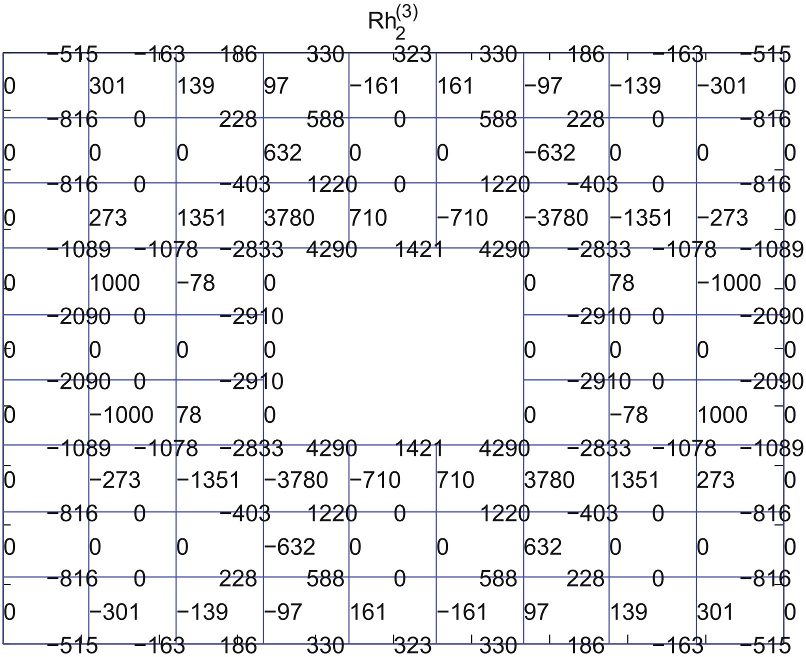


Fig. 22 Values of $h_2^{(2)}$

To quantify the rate of convergence we give in Table 12 the values of $\|h_k^{(m)} - Rh_k^{(m)}\|_2$ analogous to Table 3.

7 Discussion

We consider first SC. In Sect. 1 we defined the operators $d_0^{(m)}$, $d_1^{(m)}$, $\delta_1^{(m)}$, $\delta_2^{(m)}$, $\Delta_0^{(m)}$, $\Delta_1^{(m)}$, $\Delta_2^{(m)}$ on the graphs Γ_m approximating SC. In light of the computational results



$Rh_2^{(3)}$

−515 −163 186 330 323 330 186 −163 −515
0 301 139 97 −161 161 −97 −139 −301 0
−816 0 228 588 0 588 228 0 −816
0 0 0 632 0 0 −632 0 0 0
−816 0 −403 1220 0 1220 −403 0 −816
0 273 1351 3780 710 −710 −3780 −1351 −273 0
−1089 −1078 −2833 4290 1421 4290 −2833 −1078 −1089
0 1000 −78 0 0 78 −1000 0
−2090 0 −2910 −2910 0 −2090
0 0 0 0 0 0 0 0
−2090 0 −2910 −2910 0 −2090
0 −1000 78 0 0 −78 1000 0
−1089 −1078 −2833 4290 1421 4290 −2833 −1078 −1089
0 −273 −1351 −3780 −710 710 3780 1351 273 0
−816 0 −403 1220 0 1220 −403 0 −816
0 0 0 −632 0 0 632 0 0 0
−816 0 228 588 0 588 228 0 −816
0 −301 −139 −97 161 −161 97 139 301 0
−515 −163 186 330 323 330 186 −163 −515

Fig. 23 The restriction of $Rh_2^{(3)}$ to $\tilde{E}_1^{(1)}$ where γ_1 is the *top* and *bottom* horizontal line

reported in Sects. 2, 3, and 4 we may speculate on how it may be possible to pass to the limit as $m \to \infty$ to obtain corresponding operators on SC.

The simplest case to consider is d_0. We note that (2) is consistent from level to level without renormalization. In other words, if f_0 is any continuous function on SC whose restriction to any line segment in SC is of bounded variation, then

$$d_0 f_0(L_{ab}) = f_0(b) - f_(a), \tag{31}$$

where L_{ab} is a horizontal or vertical line segment joining a to b, defines a measure on all such line segments with

$$d_0 f_0(e_1^{(m)}) = d_0^{(m)} f_0^{(m)}(e_1^{(m)}) \tag{32}$$

where $f_0^{(m)} = f_0$ restricted to $E_0^{(m)}$. This does not settle the question of what is the most suitable domain for the space of 0-forms f_0, and then what the corresponding space of 1-forms will be the range under d_0.

The renormalization of $\Delta_0^{(m)}$ suggested in (16) requires that we renormalize $\delta_1^{(m)}$ also by the factor of r^m, and so we would like to define

$$\delta_1 f_1(x) = \lim_{m\to\infty} r^m \delta_1^{(m)} f_1^{(m)}(x) \tag{33}$$

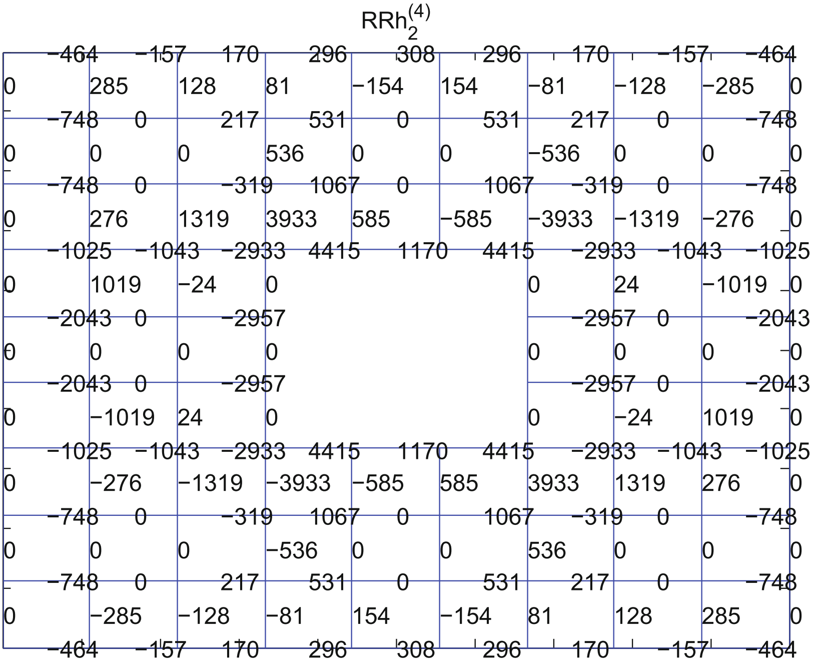


Fig. 24 The restriction of $RRh_2^{(4)}$ to $\tilde{E}_1^{(1)}$ where γ_1 is the *top* and *bottom* horizontal line

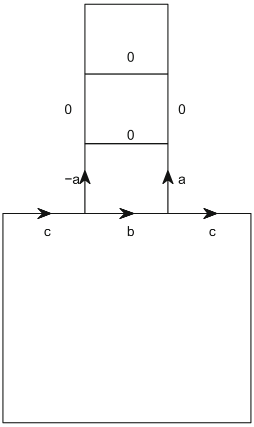


Fig. 25 A figure of the region in the top center of the level 2 graph where the horizontal symmetry has been used in labeling edges

Table 12 Values of $\|h_k^{(m)} - Rh_k^{(m)}\|_2$

Number	1	2	3	4	5	6	7	8	9	10
$\|\|h_k^{(2)} - Rh_k^{(1)}\|\|$	0.1549	0.4647								
$\|\|h_k^{(3)} - Rh_k^{(2)}\|\|$	0.1476	0.1581	0.2419	0.2039	0.2419	0.2669	0.2669	0.2419	0.2039	0.2419
$\|\|h_k^{(4)} - Rh_k^{(3)}\|\|$	0.1485	0.1349	0.1452	0.1468	0.1452	0.1458	0.1458	0.1452	0.1468	0.1452

for $x \in \bigcup_m E_0^{(m)}$, with $f_1^{(m)} = f_1$ restricted to $E_1^{(m)}$ and $\delta_1^{(m)}$ given by $\dot{\text{T}}$o make sense of this we need to find a precise domain for the 1-forms f_1 so that the limit exists and is a continuous function on the dense set $\bigcup_m E_0^{(m)}$, so that $\delta_1 f_1$ extends to a function on SC. Then we would want to characterize the range of δ_1. We would specifically like to have the domain of δ_1 contain the range of d_0, so that

$$\Delta_0 = \delta_1 d_0$$

might be defined directly without limits. The results of [5] would then imply that this Laplacian must agree, up to a constant multiple, with any other symmetric Laplacian defined on functions on SC.

The numerical evidence reported in Sect. 2 gives very strong support for the existence of $-\Delta_0$ as the limit in (16). The spectrum behaves well, and Table 2 indicates convergence of the eigenfunctions (note that we only computed changes in eigenfunctions coming from eigenvalues of multiplicity one because it is easier to make sure we are comparing the same eigenfunctions; once we normalize the eigenfunction to have L^2 norm equal to one, it is uniquely determined up to a ± sign). So we have strong evidence that the limit in (33) exists when $f_1 = d_1 f_0$ for f_0 an eigenfunction, and hence any linear combination of eigenfunctions. This does not, of course, give any "generic" space of 1-forms for which $\delta_1 f_1$ exists. Because the factors of 8^m in the definitions 6 and 5 of the weights in (7), only the sizes of $f_1^{(m)}(e_1^{(m)})$ can create decay, and under the assumption that the measures comprising f_1 are absolutely continuous, this would only create a decay rate of 3^{-m}, not nearly enough to compensate for r^m in (33). In other words, it is only cancellation of positive and negative terms that can lead to the existence of the limit in (33).

The situation for 2-forms is discussed in Sect. 4 The evidence is not as clearcut, but the suggestion is that $-\Delta_2$ may exist as a nonrenormalized limit, while δ_2 and d_1 require renormalization factors $8/3$ and $3/8$. We would then still have $d_1 d_0 = 0$ and $\delta_1 \delta_0 = 0$ because (4) and (9) hold on each level, not because of limits of nonzero terms tending to zero. If this is correct, the spectrum of $-\Delta_2$ would not be discrete, and most likely would be entirely continuous.

For 1-forms, we would expect the Hodge decomposition (11) on the approximate levels to converge to a Hodge decomposition

$$\Lambda_1 = d_0 \Lambda_0 \oplus \delta_2 \Lambda_2 \oplus \mathscr{H}_1 \tag{34}$$

for the appropriate spaces $\Lambda_0, \Lambda_1, \Lambda_2$ of forms on SC, where $\mathscr{H}_1$ denotes the harmonic 1-forms, with

$$-\Delta_1 = \begin{cases} d_0\delta_1 & \text{on } d_0\Lambda_0 \text{ or an appropriate subspace} \\ \delta_2 d_1 & \text{on } \delta_2\Lambda_2 \text{ or an appropriate subspace} \\ 0 & \text{on } \mathscr{H}_1. \end{cases}$$

For example, any L^2 0-form may be written

$$f_0 = \sum_k c_k (f_k)_0$$

where $\{(f_k)_0\}$ is an orthonormal basis of 0-eigenforms with

$$-\Delta_0 (f_k)_0 = \lambda_k (f_k)_0,$$

where

$$\sum_k |c_k|^2 = \|f_0\|_2^2 < \infty. \tag{35}$$

However, to define $-\Delta_0 f_0$ in L^2 via

$$-\Delta_0 f_0 = \sum_k \lambda_k c_k (f_k)_0$$

we need the condition

$$\sum_k \lambda_k^2 |c_k|^2 < \infty, \tag{36}$$

which is stronger than (35).

On the other hand, $\{\lambda_k^{-\frac{1}{2}} d_0 (f_k)_0\}$ is an orthonormal set of 1-forms, since

$$< d_0(f_k)_0, d_0(f_k)_0 >_1 = < (f_k)_0, \delta_1 d_0 (f_k)_0 >_0 = \lambda_k,$$

and this might be considered an orthonormal basis for the $d_0\Lambda_0$ part of the Hodge decomposition. But if

$$f_1 = \sum_k c_k \lambda_k^{-\frac{1}{2}} d_0 (f_k)_0,$$

then to define $\delta_1 f_1$ via

$$\delta_1 f_1 = \sum_k c_k \lambda_k^{-\frac{1}{2}} \delta_1 d_0 (f_k)_0 = \sum_k c_k \lambda_k^{\frac{1}{2}} (f_k)_0$$

we would require

$$\sum_k |c_k|^2 \lambda_k < \infty,$$

a condition intermediate between (35) and (36). This would also allow us to define

$$-\Delta_1 f_1 = d_0 \delta_1 f_1 = \sum_k c_k \lambda_k^{\frac{1}{2}} d_0(f_k)_0$$

as an L^2 1-form.

The situation for the $\delta_2 \Lambda_2$ part of the Hodge decomposition would be similar, except that we would have a spectral resolution rather than an infinite series of eigenfunctions, and because the operator $-\Delta_2$ is presumably bounded, we would not encounter different conditions to define $\delta_2 f_2$ and $-\Delta_2 f_2$.

What can be said about $\mathscr{H}_1$? The evidence presented in Sect. 3, in particular Table 3, suggests that limits of sequences of 1-forms in $\mathscr{H}_1^{(m)}$ will exist and give us 1-forms in $\mathscr{H}_1$, namely

$$h_k = \lim_{m \to \infty} h_k^{(m)}$$

where $h_k^{(m)}$ is defined by the cohomology/homology duality (23), keeping the cycle γ_k the same at all large enough levels m. In particular we would have

$$\int_{\gamma_j} h_k = \delta_{jk}$$

as γ_j ranges over all homology generating cycles at all levels. By taking finite linear combinations we would get 1-forms with prescribed integrals over cycles, provided all but a finite number of integrals are zero. A more challenging question is whether there is a reasonable class of harmonic 1-forms expressible as infinite linear combinations of $\{h_k\}$.

The evidence presented in Figs. 9, 10 and 11 and Tables 4 and 5 supports the conjecture that all our 1-forms restrict to measures on line segments. For harmonic 1-forms it seems almost certain that the measures are absolutely continuous. This appears to hold in all cases.

Next we consider MC. The results reported in Sect. 5 strongly support the conjecture that Laplacians $-\tilde{\Delta}_0$ and $-\tilde{\Delta}_2$ exist in the limit, and have the identical spectrum (up to a constant). Unfortunately, we are not able to get a good estimate of the Laplacian renormalization factor r, other than to say it is approximately equal to 6. The definition (31) and (32) for d_0 is the same as for SC, and similarly (33) would presumably define δ_1 with the MC value of r. In this case we would also define d_1 without renormalization,

$$d_1 f_1(e_2^{(k)}) = \lim_{m \to \infty} d_1^{(m)} f_1^{(m)}(e_2^{(k)}),$$

where $d_1^{(m)}$ is defined by (3), and $f_1^{(m)}$ is the restriction to $E_1^{(m)}$ of f_1. Indeed the limit in (33) exists trivially because $d_1^{(m)} f_1^{(m)}(e_2^{(k)})$ is constant for $m \geq k$. This follows by induction from the identity

$$d_1^{(k+1)} f_1^{(k+1)}(e_2^{(k)}) = d_1^{(k)} f_1^{(k)}(e_2^{(k)}). \tag{37}$$

To see why (37) holds we note that $e_2^{(k)}$ splits into 8 $(k+1)$-cells. Now the edges of these smaller cells come in two varieties, those that partition the edges bounding $e_2^{(k)}$ and pairs of interior edges. But each pair includes opposite orientations, so the sum of $f_1^{(k+1)}$ over the pair is zero. This of course uses the fact that the central edges in MC are identified in pairs, so nothing like this is true for SC.

The renormalization factor for $-\tilde{\Delta}_2$ would then have to be the renormalization factor defining δ_2,

$$\delta_2 f_2(e_1^{(k)}) = \lim_{m\to\infty} r^m \delta_2^{(m)} f_2^{(m)}(e_1^{(k)}). \tag{38}$$

Since $\delta_2^{(m)} f_2^{(m)}(e_1^{(k)})$ involves the difference between the f_2 measure of "thickenings" of the edge on either side, it seems difficult to explain why the limit (38) should exist, or for what class of 2-forms.

It is rather simple to understand how to obtain the Hodge *-operators on 0-forms and 2-forms on MC as renormalized limits of 13 and 14. Indeed we should define

$$*f_0 = \lim_{m\to\infty} *8^m f_0^{(m)}$$

and

$$*f_2 = \lim_{m\to\infty} *8^m f_2^{(m)}$$

If f_0 is continuous on MC then $*f_0 = f_0 d\mu$, where μ is the standard measure on MC with $\mu(e_2^{(m)}) = 8^{-m}$ for every m-cell $e_2^{(m)}$, since $*f_2^{(m)}(e_2^{(m)})$ simply averages f_0 over the four vertices of $e_2^{(m)}$. Similarly, if $f_2 = g d\mu$ for some continuous function g, then $*f_2 = g$ because $*f_2^{(m)}(e_0^{(m)})$ averages g over a small neighborhood of $e_0^{(m)}$. In particular we have ** equal to the identity in both directions.

It seems plausible that the *-operations should conjugate the two Laplacians, at least up to a constant multiple. Our data strongly supports this conjecture. The eigenvalues match well when multiplied by an experimentally determined constant. We don't have any intuitive explanation for this particular constant, however. The graphs of the eigenfunctions seem qualitatively similar, and so do the graphs of the *-operators at level m applied to eigenfunctions. Thus we have evidence for Poincaré duality on MC. Since MC is a limit of surfaces without boundary (but with singular points for the geometry) this is perhaps not surprising.

The apparent fact that the Laplacian renormalization factor r is small than 8 (the measure renormalization factor) has important implications. In Euclidean space this property only appears in dimensions higher than two. In particular, if it were possible to define an energy ε on MC so that $-\tilde{\Delta}_0$ is derived from ε and μ by the weak formulation

$$\int (-\tilde{\Delta}_0 f) g d\mu = \mathscr{E}(f,g),$$

then functions of finite energy would not necessarily be continuous. Also the value of α in the Weyl ratio will be greater than one.

For 1-forms on MC we also expect a Hodge decomposition (34). The operator d_0 intertwines $-\tilde{\Delta}_0$ on Λ_0 with $-\tilde{\Delta}_1$ on the $d_0\Lambda_0$ portion of Λ_1, and similarly δ_2 intertwines $-\tilde{\Delta}_2$ on Λ_2 with $-\tilde{\Delta}_1$ on the $\delta_2\Lambda_2$ portion of Λ_1. Note that d_0 annihilates constants and δ_2 annihilates μ, so the spectrum of $-\tilde{\Delta}_1$ consists exactly of the nonzero portions of the spectra of $-\tilde{\Delta}_0$ and $-\tilde{\Delta}_2$ together with the 0-eigenspace $\mathscr{H}_1$. Because the eigenvalues of $-\tilde{\Delta}_2$ and $-\tilde{\Delta}_1$ are proportional by a nontrivial factor, we do not see the eigenspaces doubling in multiplicity.

The story for the space of harmonic 1-forms $\mathscr{H}_1$ on MC is almost identical to SC, except the homology is different so the total number arising at each level is different.

A different approach to onstructing a Laplacian on functions on MC by using a Peano curve is given in [20].

It should be possible to define a Hodge *-operator from 1-forms to 1-forms as a limit of such operators on the graph approximations. We have not been successful in carrying this through.

References

1. Arron S, Conn Z, Strichartz R, Yu H. Hodge-de Rham theory on fractal graphs and fractals. Preprint 2012.
2. Aougab T, Dong CY, Strichartz R. Laplacians on a family of Julia sets II. comm Pure Appl Anal. 2013;12:1–58.
3. Barlow M, Bass R. On the resistance of the sierpinski carpet. Proc R Soc Lond A 1990;431:345–60.
4. Barlow M, Bass RF. coupling and Harnack inequalities for Sierpinski carpets. Bull Am Math Soc. 1993;29:208–12.
5. Barlow MT, Bass RF, Kumagai T, Teplyaev A. Uniqueness of Brownian motion on Sierpinski carpets. J Eur Math Soc. (JEMS) 2010;12:655701.
6. Begué M, Kalloiatis T, Strichartz R. Harmonic functions and the spectrum of the Laplacian on the Sierpinski carpet. Fractals. 2013;21(1).
7. Berry T, Heilman S, Strichartz RS. Outer Approximation of the Spectum of a Fractal Laplacian. Exp Math. 2009;18(4):449–80.
8. Cipriani F. Diriclet forms on noncommutative spaces, L.N.M. In: Franz U, Schurmann M, eds. ‘Quantum Potential Thoery’, 1954. New York: Springer-Verlag; 2008; pp. 161–72.
9. Cipriani F, Sauvageot J. Derivations as square roots of Dirichlet forms. J Funct Anal. 2003;201:78–120.
10. Cipriani F, Guido D, Isola T, Sauvageot J. Differential 1-forms, their integral and potential theory on the Sierpinski gasket. 2011. arXiv: 1105.1995.
11. Cipriani F, Guido D, Isola T, Sauvageot J. Spectral triples on the Sierpinski gasket, AMS Meeting ‘Analysis, Probability and Mathematical Physics on Fractals’, Cornell U.; 2011.
12. Guido D, Isola T. Singular traces on semi-finite von Neumann algebras. J Funct Anal. 1995;134:451–85.
13. Guido D, Isol T. Dimensions and singular traces for spectral triples, with applications to fractals. J Funct Anal. 2003;203:362–400.

14. Guido D, Isola T. Dimensions and singular traces for spectral triples for fractals in R^N, Advances in Operator Algebras and Mathematical Physics. In: Boca F, Bratteli O, Longo R, Siedentop H, Editors. Proceedings of the Conferene held in Sinaia, Romania, June 2003. Theta Series in Advanced Mathematics, Bucharest; 2005.
15. Heilman S, Strichartz RS. Homotopies of eigenfunctions and the spectrum of the Laplacian on the Sierpinski carpet. Fractals. 2010;18(1):1–34.
16. Hinz M. Limit chains on the Sierpinski gasket. Indiana U. Math. J., to appear.
17. Ionescu M, Rogers LG, and Teplyaev A. Derivations and Dirichlet forms on fractals. J Functional Analysis 2012;263(8):2141–69.
18. Kigami J. Analysis on fractals, Cambridge Tracts in Mathematics. Vol. 143. Cambridge: Cambridge University Press; 2001.
19. Kusuoka S, Zhou XY. Dirichlet form on fractals: poincare constant and resistance. Probab Theory Relat Fields. 2003;93:169–96.
20. Molitor D, Ott N, Strichartz RS. Using Peano curves to define Laplacians on fractals. preprint.
21. Oberlin R, Street B, Strichartz RS. Sampling on the Sierpinski gasket. Exp Math. 2003;12:403–18.
22. Strichartz R. Fractafolds based on the Sierpinski Gasket and their spectra. Trans Am Math Soc. 2003;355(10):4019–43.
23. Strichartz R. Differential equations on fractals: a tutorial. Princeton University Press; 2006.
24. Strichartz R. Laplacians on fractals with spectral gaps have nicer Fourier series. Math Res Lett. 2005;12:269–74.
25. Li, Y. “Data and Programs For Hodge DeRham Theory of K-Forms on Carpet Type Fractals” www.math.cornell.edu/∼yl534 (2013)

Part X
Applications and Algorithms in the Physical Sciences

One of the many ways in which harmonic analysis distinguishes itself among other areas of modern mathematics is through the emphasis placed on algorithm development and the connections that it builds with applied sciences. This phenomenon goes back to Joseph Fourier, whose main motivation for introducing what we know as the Fourier transform, was his work on heat flow and thermal conduction. Other prominent examples of these interactions include the role of Radon transform in Magnetic Resonance Imaging, the impact of Kaczmarz algorithm on Computed Tomography, and the role played by the Phase Problem in X-ray crystallography—all rewarded with Nobel Prizes. The continuation of these trends is certain, as is illustrated by the selection of four excellent chapters devoted to state-of-the-art applications of recent developments in harmonic analysis.

An example of such a fundamental connection to applied sciences is the concept of Laplace transform, which is treated in the first chapter. NAIL A. GUMEROV and RAMANI DURAISWAMI present a detailed analysis of the spherical harmonic rotation coefficients, together with new, fast, and stable recursive algorithms for their computation. Spherical harmonics form an orthonormal basis for the space of square integrable functions on the unit sphere. As such, they have many practical applications, ranging from computation of electron configurations in quantum mechanics, providing solutions for many fundamental equations in mathematical physics, to applications in geostatistics and astrophysics. Detailed description of the involved algorithms is provided and illustrated with numerical examples.

BRIAN O'DONNELL, ALEXANDER MAURER, and ANTONIA PAPANDREOU-SUPPAPPOLA give an excellent overview of the role of modern time-frequency signal processing techniques in molecular biology. Highly localized waveform analysis and parameter estimation are the main tools used to detect and analyze variations in the profiles of antibodies to discriminate between pathogens. When combined with the recent developments in measuring expression levels for large numbers of genes, proteins, or peptides, these methods become a powerful tool with such possible applications, as diagnosis of infectious diseases before they become symptomatic.

Harmonic analysis inspired representations of 3D objects are treated in the third chapter of this part. Efficient visualization of complex 3D phenomena plays an important role in such diverse fields as computer graphics, X-ray crystallography, or magnetic resonance imaging. DAVID A. SCHUG, GLENN R. EASLEY, and DIANNE P. O'LEARY analyze novel geometric multiscale representation systems called shearlets, and demonstrate the advantages arising from including directional information in the multiresolution analysis, over classical wavelet techniques. Resulting 3D edge detection algorithms are carefully studied and compared with traditional 2D methods.

In the final chapter, SHERRY E. SCOTT introduces the readers to the rich field of fluid dynamics and its many interactions with wavelet theory. This gives us a better understanding of the role played by novel mathematical models in analysis and monitoring of Earth's climate and weather. The key notion in this presentation is the concept of ergodicity defect—a value that captures the deviation of a system from ergodicity, and which can serve as a diagnostic tool in a variety of geoscience applications.

Biosequence Time–Frequency Processing: Pathogen Detection and Identification

Brian O'Donnell, Alexander Maurer, and Antonia Papandreou-Suppappola

Abstract Diagnostic information obtained from antibodies binding to random peptide sequences is now feasible using immunosignaturing, a recently developed microarray technology. The success of this technology is highly dependent upon the use of advanced algorithms to analyze the random sequence peptide arrays and to process variations in antibody profiles to discriminate between pathogens. This work presents the use of time–frequency signal processing methods for immunosignaturing. In particular, highly-localized waveforms and their parameters are used to uniquely map random peptide sequences and their properties in the time–frequency plane. Advanced time–frequency signal processing techniques are then applied for estimating antigenic determinants or epitope candidates for detecting and identifying potential pathogens.

Keywords Random-sequence peptide microarray · Epitope · Pathogen · Immunosignaturing · Time–frequency processing · Detection · Identification

1 Introduction

1.1 Signal Processing of Biological Sequences

The area of bioinformatics is mainly involved with the management of biological information using computer technology and statistics. Signal processing for molec-

B. O'Donnell (✉) · A. Maurer · A. Papandreou-Suppappola
School of Electrical, Computer and Energy Engineering, Arizona State University,
Tempe, AZ, USA
e-mail: bnodonnell@gmail.com,

A. Maurer
e-mail: ajmaurer@asu.edu,

A. Papandreou-Suppappola
e-mail: papandreou@asu.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_3

ular biology, on the other hand, encompasses the development of algorithms and methodologies for extracting, processing, and interpreting information from biological sequences [1–6]. Intelligent use of signal processing algorithms can provide invaluable insight into the structure, functioning, and evolution of biological systems. For example, complex assays to determine functional activities of analytes or peptide chips to manifest key residues for protein binding can provide a wealth of information on underlying biological systems. However, in each of these cases, appropriately designed processing is required to robustly extract the most relevant information. Images of array fluorescence are enhanced to improve the estimation of gene reactivity, while gene expression classification performance is increased by including biological and experimental variability in the algorithm design [4].

Genomics and proteomics, in general terms, study the functions and structures of genomes and proteomes, respectively. Genomes, which are genetic material of organisms encoded in deoxyribonucleic acid (DNA) or ribonucleic acid (RNA), and proteomes, which are expressed proteins in given organisms, provide discrete information, represented in sequences of unique elements [7, 8]. More specifically, DNA are biomolecules that are represented as letter sequences of precise orderings of four nucleobases; the different orderings correspond to patterns that influence the formation and development of different organisms. Similarly, proteins are biomolecules represented as sequences of unique orderings of 20 linked amino acids, with each amino acid represented by a letter of the alphabet. DNA and protein sequence analysis requires significant processing of the discrete gene orderings in order to identify intrinsic common features or find gene variations such as mutations [9, 10]. One genome analysis application is gene sequence periodicity as regions of genetic repetition have been shown to correlate with functionally important genes [11, 12]. Gene periodicity has been analyzed using spectral methods [13–16]; such methods have also been used to estimate variations in base pair frequencies between organisms as they can indicate phylogenic origin from the species genome. Time–frequency signal processing methods such as wavelet transforms have also been used in gene sequencing such as to characterize long range correlations or identify irregularities in DNA sequences [14, 17, 18].

Signal processing methods have also been used for sequence alignment, or arranging sequences to identify regions of similarity due to functional, structural, or evolutionary relationships between the sequences [19, 20]. As thousands of organisms have been sequenced completely, and many more have been partially sequenced, searching for these similarities requires a vast number of computations. There are many algorithms designed to perform these searches including dynamic programming algorithms such as Smith–Waterman and BLAST, correlation based methods, Bayesian approaches, and time–frequency (TF) based methods [10, 21–28]. Computational alignment tools based on dynamic programming such as the Smith–Waterman algorithm is guaranteed to find all similarity matches, but it runs slowly [21]. Other tools, such as BLAST [22, 23], are widely made available for database similarity searching as they were developed to provide a fast approach of approximating the complete alignment found by dynamic programming algorithms. BLAST runs very quickly, around an order of magnitude faster than the complete alignment algorithms, and finds most significant alignments under most circumstances. However, it tends to miss alignments for queries with repetitive segments. Correlation based methods map DNA or amino acid sequences to real or complex numbered sequences and use sequence correlation to achieve a match in similarity [26]. The algorithm can be implemented fast using the fast Fourier transform; however, errors increase when aligning sequences of longer lengths. We have recently developed a TF based method that first uniquely maps sequences to highly-localized Gaussian waveforms in the TF plane and then uses the matching pursuit decomposition (MPD) algorithm to perform alignment [28–30]. The alignment approach is compared to other approaches and shown to perform well with repetitive segments in real time without preprocessing.

In addition to gene sequencing, microarray analysis has also played a significant role in the extraction and interpretation of genomic information. Microarrays can provide measurements of expression levels of large numbers of genes. For example, peptide microarrays have been used to study binding properties and functionality of different types of protein–protein interactions and provide insight into specific pathogens [31–35]. Peptide microarrays are a relative new application for biological signal processing. The technology to create assays using single peptide chains has been around for a while in the form of the enzyme linked immunosorbent assay (ELISA) [36]. In recent years, as the cost of printing many peptide clusters onto a single substrate has been dropping, tens or hundreds of thousands of peptide clusters can be reasonably printed on a single array. In addition to being able to construct large-scale peptide arrays to detect specific diseases, another important aspect is the robust interpretation and analysis of the extracted data in order to establish relationships between peptide sequences and binding strengths. Some recent analysis approaches include support vector machine (SVM) modeling methods [37], computational alignment approaches [38], and statistical tools such as t-test and analysis of variance linear regression [39–41].

1.2 Signal Processing Challenges: Random-Sequence Microarrays

The recently developed immunosignaturing technology uses microarrays with random-sequence peptides to associate antibodies to a pathogen or infectious agent, in a patients blood sample [32, 42–49]. The immunosignatures can potentially provide pre-symptomatic diagnosis for infectious diseases [35, 44, 48]. The large number of peptide sequences on each microarray, and the attraction of the ability to diagnose as many pathological ailments as possible, renders a challenging problem in signal processing. This is further complicated by the fact that, in general, training data is not available. Current processing methods include statistical tests [45] and supervised classification and learning methods such as support vector machines [43, 48]. Recently, we have developed adaptive learning methodologies for unsupervised clustering integrated with immunosignature feature extraction approaches [50–52].

This work develops new algorithms for analyzing and processing random peptide sequences in the TF plane in order to recognize pathogens from variations in antibody profiles without any prior information. Given immunosignaturing random-sequence peptide microarray data for an individual, the task is to detect and identify the binding sites of antibodies for target antigens. These binding sites, or linear epitopes, are short continuous subsequences of the peptide sequence that correspond to the part of an antigen that is recognized by the antibodies [32, 53, 54]. Detecting which peptides bind to which antibodies by identifying the corresponding antibody subsequence binding sites using immunosignaturing data is very useful as one dataset contains localized information on multiple pathogens [52]. As a result, the detection and identification algorithms can be used to characterize antibody specificity for the molecular recognition of the immune system or for deciphering molecular mechanisms for various diseases.

2 Random Sequence Peptide Microarrays

Immunosignaturing is a microarray-based technology that uses random peptide sequences to provide a comprehensive profiling of a person's antibodies [42, 44, 55]. It has been shown that a person's antibody profile, about 109 different antibodies in the blood at a given time, is a sensitive indicator of the person's health status. Part of the body's response to a foreign pathogen is to create antibodies which identify and aid in the destruction of that pathogen. Pathogen detection and determination is possible due to a uniquely identifying amino acid sequence on its exterior called an antigen. The antibodies created in response to the pathogen are designed to only bind to that specific antigen sequence, or one that is very similar. As the antibodies amplify when the host is exposed to an infectious agent, the amplified antibody response enables monitoring a disease upon infectivity. Rather than trying to identify an antibody by designing a microarray specific to a pathogen, the concept of immunosignaturing is to identify an entire immune response. This is achieved by printing an array with many different random peptides, so that small subsets of peptide sequences are similar enough to antigen sequences of specific pathogen antibodies to bind to them.

The immunosignaturing technology has been developed by the director and researchers of the Center for Innovations in Medicine (CIM) at the Biodesign Institute at Arizona State University [56]. In particular, the random-peptide microarray data used for algorithm demonstration in this work was provided by CIM. Information on the technology, such as a description of the equipment, arrays, and a technological overview, can be found at `http://www.immunosignature.com`. The immunosignaturing technology currently employs slides spotted with peptides, resulting in microarrays with 330,000 peptide sequences (330k chip). The peptides sequences are 20 amino acids long and a random number generator is used to generate the specific peptide sequences. Other than cysteine that is used as the C-terminal amino acid, all natural amino acids are included in the peptide sequence generation. As a result, the peptide sequences are random and not related to any naturally occurring peptide sequence; however, the sequence on each spot on the slide is known.

This ensures that the peptide array is not designed to monitor one specific disease or a set of diseases. Array peptides are designed to fluoresce in proportion to an antibody binding strength when light of a specific frequency is shined on them. Peptides with attached antibodies are expected to fluoresce more brightly than the those not related to the antigen. Multiple identical peptides are printed within a predefined circular area on the array; after the sample is given sufficient time to bind, the array is washed and then illuminated. An image of the fluorescing array is taken, and the median fluorescence value of the pixels in each circular area is calculated and recorded. The resulting data used for analysis is the peptide sequence of amino acids and its corresponding median fluorescence value at each array spot. Using the data from the whole array, the problem is to detect the highly fluorescing peptides and identify the corresponding underlying pathogens.

This is not a simple detection and identification problem; processing can be complicated by the fact that there are additional macromolecules in blood samples that can also bind to peptides due to hydrogen bonding, electrostatic interactions, and van der Waals forces [57]. The concept of adding a large number of random peptides on the array is novel as more pathogens can be detected on a single patient. However, the large number of sequences to process also increases the number of sequences that are close in structure to more than one pathogen's antigen. Antibodies bind with enough variability that trends across multiple peptides must be used. However, a significant difficulty in finding these trends is that only a subsequence of the peptide which binds to the antibody is responsible for the binding, and within that subsequence there can be one or two peptides which have little or no effect on the binding strength. Determining which peptide subsequences are responsible for that binding must be determined using multiple peptides with similar sub-sequences [44, 45].

3 Time–Frequency Processing of Peptide Sequences

The novel signal processing algorithms presented in this work aim to improve pathogen detection and identification performance when using immunosignaturing random peptide sequences. Toward this end, advanced signal processing methodologies are exploited to first map amino acid sequences to unique and highly TF localized waveforms and then use matched TF representations to identify specific peptide sub-sequences.

3.1 Mapping Peptide Sequences to Time–frequency Waveforms

The biosequence-to-waveform mapping considered must provide a unique waveform in the TF plane for each peptide sequence. When deciding on appropriate waveforms to use in the mapping, the waveform parameters and properties must be selected to ensure uniqueness in peptide representation and robustness in matched correlation-based processing, respectively. Following the scheme we adopted in [28], Gaussian waveforms are selected for mapping as they are the most localized waveforms in both time and frequency [58]. A basic Gaussian waveform $g(t)$ is first obtained as

$$g(t) = \frac{1}{(\pi\sigma^2)^{1/4}}\, e^{-t^2/(2\sigma^2)},\ \ t \in (-T_g/2, T_g/2)\,, \tag{1}$$

with unit energy and centered at the origin in the TF plane. The parameter σ^2 affects the waveform's duration T_g and spread in frequency. When this waveform is time-shifted by nT and frequency-shifted by kF,

$$g_{n,k}(t) = g(t-nT)\, e^{j2\pi kF(t-nT)},\ \ t \in (nT - T_g/2, nT + T_g/2) \tag{2}$$

for integer n and k, the resulting Gaussian waveform is highly-localized at the TF point (nT, kF). Note that the time shift step $T > T_g$ and the frequency shift step F, and thus σ^2 in (1), are chosen to ensure that the spacing between the time–frequency shifted Gaussian waveforms is compact and the waveforms are nonoverlapping.

For the biosequence-to-waveform mapping, the time shift and frequency shift are used to uniquely represent properties of the amino acids in the peptide sequence. Each of the 20 possible different amino acids in a peptide sequence can be characterized by a unique one-letter code, as shown in the first two columns of Table 1. For the mapping, 20 possible frequency shifts kF, $k=1,\ldots,20$, in Eq. (2) are used to represent the 20 different types of amino acids, as shown in the third column of Table 1. The position of the amino acid in the peptide sequence is mapped to the time shift parameter nT in (2). Considering a peptide sequence of length $N=20$ amino acids, N time shifts are needed to represent the peptide sequence; the number of time shifts is the same as the length of the sequence. A TF representation of all possible Gaussian waveforms needed to map peptide sequences of length $N=10$ amino acids is demonstrated in Fig. 1a.

Considering a peptide sequence $p[n]=\alpha_n$, $n=1,\ldots,N$, of N amino acids $\alpha_1, \alpha_2, \ldots, \alpha_{N-1}, \alpha_N$, the mapping function $f[\{\alpha_n\}]=k$ is used to identify the one-letter code representing the amino acid α_n and its corresponding frequency shift kF from Table 1. Note that the range of the mapping function $f[\{\cdot\}]$ is the set of positive integers, $k=1,\ldots,20$; the domain of the function consists of the one-letter codes from Table 1. Using this mapping function, the resulting waveform that is used to map peptide sequence $p[n]$ is given by

$$g_{\mathrm{pept}}(t) = \sum_{n=1}^{N} g_{n,f[\{\alpha_n\}]}(t;p) = \sum_{n=1}^{N} g(t-nT)\, e^{j2\pi f[\{\alpha_n\}]F(t-nT)}\,. \tag{3}$$

The duration of the overall waveform $g_{\mathrm{pept}}(t)$ is $NT + T_g$.

An example of a peptide sequence of length $N=10$ is given by `ARVHHKHVVE`; its corresponding TF representation is shown in Fig. 1b. The waveform in (3) used to map this sequence is a linear combination of ten TF-shifted Gaussian waveforms. Ten unique time shifts are used in the mapping; the frequency shifts are not unique

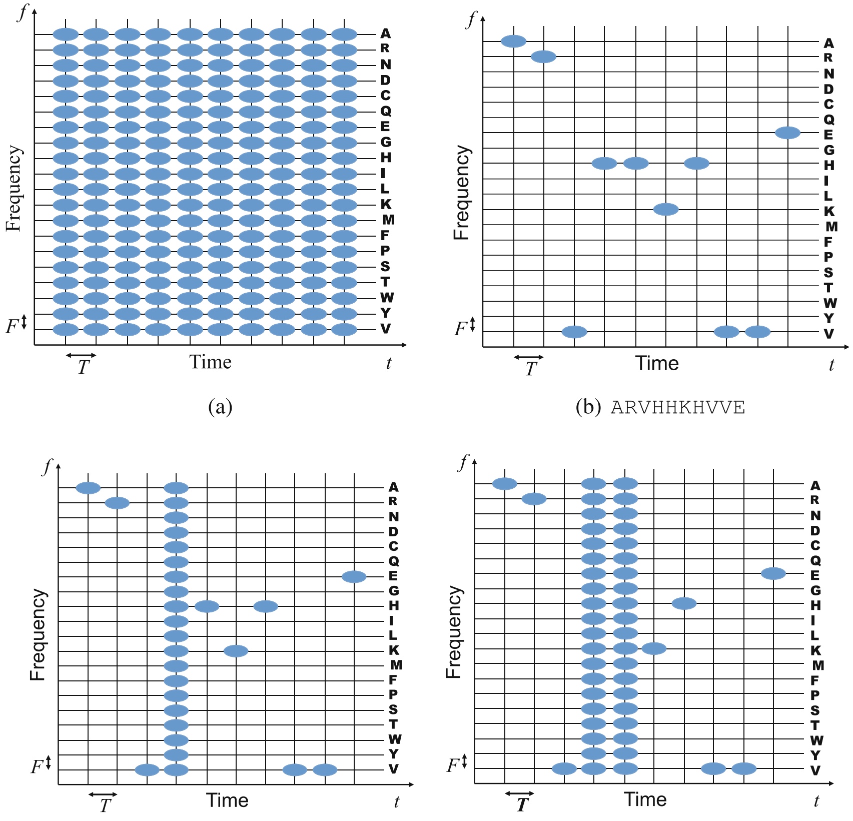
Table 1 Frequency mapping of 20 amino acids

Amino acid	**One-letter code**	**Mapped frequency**
Alanine	A	$20F$
Arginine	R	$19F$
Asparagine	N	$18F$
Aspartic acid	D	$17F$
Cysteine	C	$16F$
Glutamic acid	E	$15F$
Glutamine	Q	$14F$
Glycine	G	$13F$
Histidine	H	$12F$
Isoleucine	I	$11F$
Leucine	L	$10F$
Lysine	K	$9F$
Methionine	M	$8F$
Phenylalanine	F	$7F$
Proline	P	$6F$
Serine	S	$5F$
Threonine	T	$4F$
Tryptophan	W	$3F$
Tyrosine	Y	$2F$
Valine	V	F

since the same amino acid can occur multiple times in a peptide sequence. It follows that there is only one Gaussian waveform at each time shift but (possibly) multiple Gaussian waveforms at different frequency shifts; a peptide sequences does not necessarily consist of all possible 20 amino acids. For the example of the length ten peptide sequence ARVHHKHVVE in Fig. 1b, and using the mapping in (3) and the information from Table 1, $\alpha_1\,\alpha_2\,\alpha_3\,\alpha_4\,\alpha_5\,\alpha_6\,\alpha_7\,\alpha_8\,\alpha_9\,\alpha_{10} = \mathrm{ARVHHKHVVE}$ and $f[\{\alpha_1\}]$=20, $f[\{\alpha_2\}]$=19, $f[\{\alpha_3\}]$=$f[\{\alpha_8\}]$=$f[\{\alpha_9\}]$=1, $f[\{\alpha_4\}]$=$f[\{\alpha_5\}]$=$f[\{\alpha_7\}]$ =12, $f[\{\alpha_6\}]$=9, and $f[\{\alpha_{10}\}]$= 15. Specifically, the three histidine (H) amino acids in the sequence are represented in Fig. 1(b) by the three Gaussian waveforms at the same frequency shift $12F$ and different time shifts, $4T$, $5T$, and $7T$, respectively.

3.2 Processing Waveforms of Mapped Peptide Sequences

The peptide sequence mapping in Eq. (3) results in a linear combination of nonoverlapping Gaussian signals in the TF plane. A linear epitope or small continuous segment of the peptide sequence can be used as an antigenic determinant to a pathogen's antibodies. Identifying epitopes can be seen as searching for potential subsequences that are either repeated very often or are frequently repeated with significant binding strength on the microarray. After waveform mapping, the detection and identification problem of epitopes or repeated subsequences over a large number of peptide



(a)

(b) ARVHHKHVVE

(c) ARV-HKHVVE

(d) ARV--KHVVE

Fig. 1 Time–frequency representation of Gaussian mapped waveforms **a** for peptide sequences of 10 amino acids in length; **b** for amino acid sequence ARVHHKHVVE; **c** for the same sequence with any substitution in the 4th amino acid position, ARV-HKHVVE; **d** for the same sequence with any substitutions in the 4th and 5th amino acid positions, ARV--KHVVE

sequences on a microarray becomes an estimation problem of the matched Gaussian waveform parameters representing the amino acids in the epitope. As a result, epitope waveform parameter (EpiWP) estimation can be performed using matched signal expansion algorithms, such as the MPD [59]. Specifically, identifying repetitions in the sequences maps to estimating matched parameters in the waveforms.

The MPD is an iterative algorithm that can decompose a waveform into a linear combination of weighted dictionary waveforms. The dictionary waveforms are formed by TF shifting a basis waveform that is selected to be well-matched to the analysis waveform. The MPD can be applied to the EpiWP estimation problem using the Gaussian waveform in (1) as the dictionary basis signal. Then the epitope mapped waveform to be decomposed and the MPD dictionary waveforms are in the form of (3). In particular, mapped epitope candidate (MEpiC) waveforms are formed by considering possible amino acid subsequences from the peptide sequences. Using the MPD to decompose an MEpiC waveform results in a small set of MPD features that uniquely characterize the MEpiC waveforms; those features are then searched over all mapped peptide sequences on the microarray. A suitably derived metric for the number of peptide sequences identified to have the matched MEpiC waveform can then be used to indicate whether the epitope candidate could be related to an epitope of the antibodies of a particular pathogen.

Assuming an epitope of length L, the MEpiC waveform $g_{\mathrm{epit}}(t)$ is given by Equation (3) with N replaced by L, the length of the epitope.[1] At each iteration, the MPD identifies a single TF shifted Gaussian waveform from the MEpiC waveform. This is accomplished by finding the best match between each of the mapped amino acids forming the MEpiC waveform $g_{\mathrm{epit}}(t)$ and possible mapped amino acids $g_{\mathrm{pept},n,f[\{\alpha_n\}]}(t)$ forming the peptide waveform. The MPD requires L iterations to find a match of the MEpiC waveform within the mapped peptide waveforms. At the start of the MPD algorithm, the best matched dictionary waveform between the MEpiC waveform and the mapped peptide amino acid waveforms is obtained as

$$g^{(1)}_{n_1,f[\{\alpha_{n_1}\}]}(t) = \underset{n}{\operatorname{argmax}} \int g_{\mathrm{epit}}(t)\, g_{\mathrm{pept},n,f[\{\alpha_n\}]}(t)\, dt\,, \tag{4}$$

where $g^{(1)}_{n_1,f[\{\alpha_{n_1}\}]}(t)$ is a Gaussian waveform centered at time shift $n_1 T$ and frequency shift $f[\{\alpha_{n_1}\}]F$, and n_1 is the value of n that yields the maximum correlation value in (4) after the first iteration. At the ℓth iteration, $\ell = 2,\ldots,L$, the residual MEpiC waveform is given by

$$r^{(\ell)}_{\mathrm{epit}}(t) = g_{\mathrm{epit}}(t) - \sum_{m=1}^{\ell-1} g^{(m)}_{n_m,f[\{\alpha_{n_m}\}]}(t)\,.$$

The best matched dictionary waveform between the residual MEpiC waveform and the mapped peptide waveform is given by

$$g^{(\ell)}_{n_\ell,f[\{\alpha_{n_\ell}\}]}(t) = \underset{n}{\operatorname{argmax}} \int r^{(\ell)}_{\mathrm{epit}}(t)\, g_{\mathrm{pept},n,f[\{\alpha_n\}]}(t)\, dt\,. \tag{5}$$

The discrete value n_ℓ is the sequence position index n that yields the maximum correlation value in (5) at the ℓth iteration. Note that, there are no correlation coefficients to consider in the expansion as the Gaussian waveforms are normalized to have unit energy. The algorithm iteratively continues until L iterations, when there are no more matches left between the MEpiC waveform and the mapped peptide

[1] Note that the same notation, α_n, is used to denote amino acids in peptide sequences and amino acids in epitope sequences, which are subsequences of the peptide sequences. The specific type of sequence, peptide or epitope, is differentiated, when needed, using the notation $g_{\mathrm{pept}}(t)$ and $g_{\mathrm{epit}}(t)$, respectively.

waveform. After L iterations, the decomposed mapped peptide waveform is given by

$$\tilde{g}(t) = \sum_{\ell=1}^{L} g^{(\ell)}_{n_\ell, f[\{\alpha_{n_\ell}\}]}(t) + r^{(L+1)}_{\text{epit}}(t)\,. \tag{6}$$

The matched MEpiC waveform components are given by the summation term in the right-hand side of Eq. (6); the unmatched ones are in the residue $r^{(L+1)}_{\text{epit}}(t)$. The Gaussian waveform matching can then be used to obtain an epitope identification metric in terms of the energy of the decomposed Gaussian waveform components. The metric, for a candidate epitope, is given by

$$\mathrm{s}_{\text{epit}} = \int \Big| \sum_{\ell=1}^{L} g^{(\ell)}_{n_\ell, f[\{\alpha_{n_\ell}\}]}(t) \Big|^2 dt. \tag{7}$$

Note that, each mismatch between the MEpiC waveform and the matched peptide waveform decreases the matching metric by one as the energy of the decomposed term also decreases by one.

The MPD algorithm can also be used for matching MEpiC waveforms which model biologically relevant substitutions. The matching is performed using the same MPD algorithm with a modification to the MEpiC waveforms. This is demonstrated in Fig. 1c and d for the length ten sequence `ARVHHKHVVE` represented in the TF plane in Fig. 1b. In Fig. 1c, the same sequence is considered but with a substitution allowed by any amino acid in the forth position. The effect on the MEpiC waveform is to include a Gaussian waveform at each frequency shift at the forth position (or time shift); this implies that any mapped peptide waveform is matched to the MEpiC waveform at the forth time-shift. The same sequence but with two substitutions in the forth and fifth amino acid positions is demonstrated in Fig. 1d.

4 Epitope Waveform Parameter Estimation

The epitope waveform parameter estimation algorithm consists of three main steps. During the first step, the candidate epitope and peptide amino acid sequences are mapped to Gaussian waveforms, following the discussion in Section 3.1. During the second step, the peptide sequences are down selected by first preprocessing the peptide sequences, and then applying some selection criteria and thresholding; the reduced number of peptides after selection are the ones most likely to have been bound to by antibodies. The third step performs the epitope waveform parameter estimation using the MPD-based matching approach discussed in Section 3.2. The steps are summarized in Fig. 2.

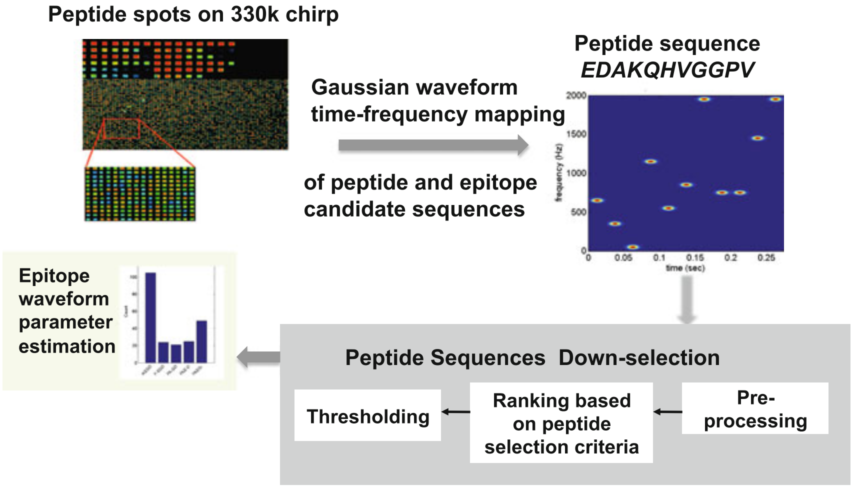


Fig. 2 Block diagram depicting the algorithm for epitope waveform parameter (EpiWP) estimation

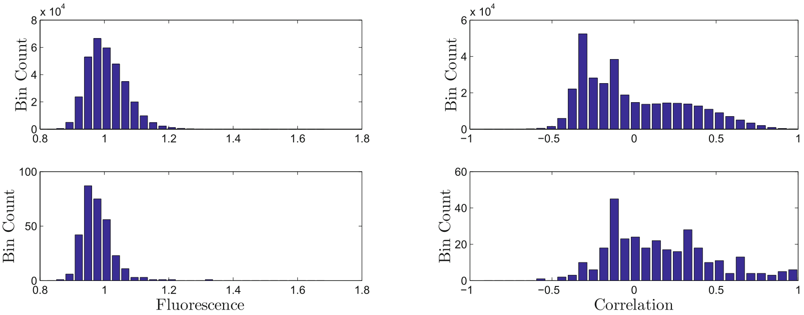
4.1 Peptide Selection Method

Preprocessing

Peptide array data from individual disease samples are median normalized to account for the different binding times required. Some samples require longer time to bind fully to the array before the sample solution is rinsed off.

Ranking Based on Peptide Selection Criteria

As the number of microarray random peptide sequences, M_p, is very large for efficient processing, the peptide sequences need to be ranked according to some peptide selection criteria, and then a selected smaller number of peptides, M_s, can be used as input to the EpiWP estimation algorithm. One peptide selection criterion is based on fluorescent intensity levels; the peptides with the highest fluorescent intensity levels, or levels above some background fluorescent intensity threshold are selected as they correspond to the peptides that bind to the antibodies. For some datasets, it is possible for antibodies to bind weakly to peptides that do not have the highest fluorescence values. As a result, a different peptide selection criterion is needed for these datasets. The criterion is based on finding correlations or dependence between multiple datasets; the fluorescent intensity levels from multiple datasets can be compared in order to select peptides with high fluorescence values relative to the comparison data.



(a) (b)

Fig. 3 Histogram plots for the monoclonal antibody `mAb1` using values of **a** fluorescence and **b** correlation. *Top* plots process all peptide sequences; *bottom* plots process only peptides with the exact epitope subsequence

The second peptide selection criterion is applied using Pearson's correlation coefficient between the fluorescent intensity levels of the array peptides and a binary indicator vector. Assuming D microarray datasets for comparison, with M_p peptides per microarray dataset, the correlation coefficient for the mth peptide, $m = 1, \ldots, M_p$, at microarray $\tilde{d}$, is computed as

$$r_{\tilde{d},m} = \frac{\sum_{d=1}^{D} \left(\mathrm{fl}_{d,m} - \bar{\mathrm{fl}}_m\right)\left(b_{\tilde{d},d} - (1/D)\right)}{\left(\sum_{d=1}^{D} \left(\mathrm{fl}_{d,m} - \bar{\mathrm{fl}}_m\right)^2\right)^{1/2} \left(\sum_{d=1}^{D} \left(b_{\tilde{d},d} - (1/D)\right)^2\right)^{1/2}}, \tag{8}$$

where $\mathrm{fl}_{d,m}$ is the fluorescent intensity of the mth peptide of the dth array,

$$\bar{\mathrm{fl}}_m = \frac{1}{D} \sum_{d=1}^{D} \mathrm{fl}_{d,m},$$

is the fluorescence sample mean of the mth peptide across all D microarray datasets, and $b_{\tilde{d},d}$ is 1 if $\tilde{d} = d$ and 0 otherwise.

One example of a monoclonal antibody for which a different peptide selection criterion can give different estimation results is `mAb1`. For this monoclonal antibody, using the fluorescent intensity peptide selection criterion demonstrates that the antibodies bind weakly to the peptides relative to background binding and cannot be detected at any threshold. In particular, there are many peptides with high fluorescent intensity that do not bind to the `mAb1` antibody. This is illustrated in Fig. 3a, where the figure on the top is a histogram of all of the 330k fluorescent intensity levels on the array, while the figure on the bottom is a histogram of just the fluorescent intensity levels of peptides which contain a subsequence of the `mAb1` epitope. For this dataset, as there are peptides with higher fluorescent intensities than most of the peptides with the mAb epitope, the fluorescent intensity selection criterion fails to provide correct epitope estimates. On the other hand, if the correlation value peptide selection criterion is applied, epitope estimation performance is improved. This is demonstrated in Fig. 3b, where there are a cluster of epitope sequence peptides with correlation values approaching 1. These sequences will be kept after thresholding, and there will be a higher percentage of peptides with epitope subsequences when correlation is used as the ranking metric compared to when fluorescence is used as the ranking metric. Note that an epitope subsequence is four or more contiguous amino acids from the epitope and that the fluorescent intensity levels in Fig. 3a were logarithmically transformed to improve visualization.

Thresholding

Depending on the peptide selection criterion, thresholding is used to keep $M_s \ll M_p$ peptide sequences as input to the epitope estimation. A background fluorescent intensity threshold is used with the fluorescent intensity criterion. With the correlation coefficient criterion, the correlation coefficient values $r_{\tilde{d},m}$ of the mth peptide, $m = 1,\ldots,M_p$, on the $\tilde{d}$ array in (8) are first ranked in descending order and then compared to some threshold.

4.2 Epitope Estimation Algorigthms

The epitope candidate sequences are derived from the remaining P peptide array sequences obtained after applying the selection method in Section 4.2. There are three different methods considered for epitope waveform parameter (EpiWP) estimation, resulting in the detection and identification of the epitope candidate sequences. The epitope candidate sequences for the EpiWP-1 estimation method include all possible subsequences of length L adopted from the peptide microarray sequences. The epitope candidate sequences for the EpiWP-2 estimation method include all subsequences of length L from the peptide array sequences, together with the subsequences formed by allowing for a single amino acid substitution (by any other type of amino acid). The epitope candidate sequences for the EpiWP-3 estimation method include all subsequences of length L from the peptide array sequences, together with the subsequences formed by allowing for two adjacent amino acid substitutions.

The main steps of the estimation algorithm are summarized in Algorithm 1. Using the down-selected M_s peptide sequences, the MPD is used to compare the peptide sequences to M_e epitope candidate sequences. The overall matching score uses the metric in (7) and the number of peptide sequences that include each of the M_e epitope candidate sequences. Algorithm 2 finds the maximum match between two sequences with dissimilar lengths. The two sequences are peptide array sequence of length N and the epitope candidate sequence of length L, where $N > L$. The algorithm first maps both sequences to Gaussian waveforms and then uses the

MPD to perform the matching. The number of maximum matches found using Algorithm 1 is recorded, and epitope candidate sequences are sorted in descending order according to the number of peptides they were found in. The epitope candidate sequences that occur most frequently are the top epitope estimates.

Algorithm 1 Matches Between Peptides and Epitope Candidate Sequences

```
for $i = 1$ to $M_s$ do
    for $j = 1$ to $M_e$ do
        Run Algorithm 2 on the $i$th peptide and $j$th epitope candidate sequences
        Record the number of maximum matches for each epitope candidate sequence
    end for
end for
```

Algorithm 2 Maximum Match Between Two Sequences

```
for $n = 1$ to $N - L + 1$ do
    Map peptide sequence $p[m]$, $m = n, \ldots, n + L - 1$, onto TF waveforms $g_{\mathrm{pept}}(t)$
    Map epitope candidate sequence $e[l]$, $l = 1, \ldots, L$, onto TF waveforms $g_{\mathrm{epit}}(t)$
    Perform MPD using $g_{\mathrm{pept}}(t)$ and $g_{\mathrm{epit}}(t)$ to find score $s_{\mathrm{epit}}$
end for
```

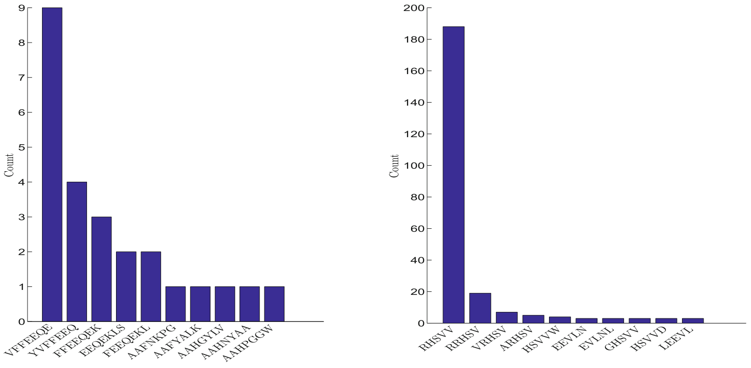
4.3 Evaluation of Epitope Estimation

In order to evaluate the performance of the random-sequence peptide microarray with the EpiWP estimation method for identifying antibody epitopes, data sets for eight monoclonal antibodies (mAbs) were acquired. The mAbs used have known epitopes that were used to probe the microarray. Monoclonal antibodies are used in the evaluation, instead of blood samples from patients, as the mAbs bind to a single linear epitope selected for high specificity for the antigen [44, 60–63]. On the contrary, epitopes for most diseases are not known; even if the epitope for a single strain of a disease is known, it may not be known for the specific strain of the analyzing sample. The mAb random-sequence peptide microarray data were provided by CIM [56]; each microarray sample consists of 330,000 peptide sequences (330k chip). Although this is a large number of sequences on the array, only a small percentage of the sequences bind to different mAbs.

Table 2 provides a list of the eight mAbs used to demonstrate epitope waveform parameter estimation. The known epitope of each mAb is provided in the second column of this table. The last column provides the estimated epitopes with varying lengths. The EpiWP estimation method performed well for all but the monoclonal antibody `mAb5` epitope. Based on this result, the mAb epitope estimation performance is about 88 % accurate.

Table 2 Epitope estimates for eight monoclonal antibodies

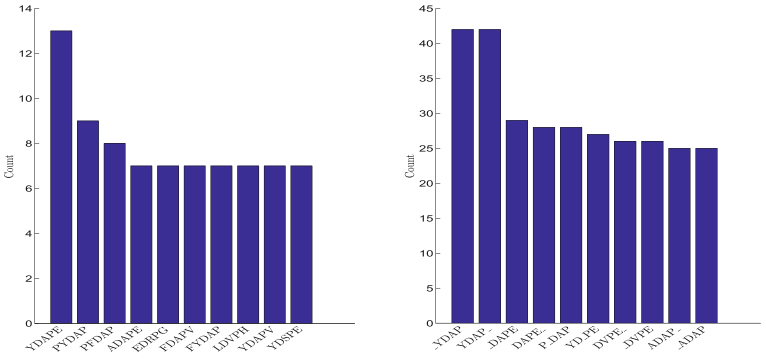
Name	Full epitope	Estimated epitope
mAb1	NAHYYVFFEEQE	YVFFEEQE
mAb2	QAFDSH	AFDSH
mAb3	EEDFRV	EDFRV
mAb4	RHSVV	RHSVV
mAb5	SDLWKL	–
mAb6	AALEKD	ALEKD
mAb7	DYKDDDDK	KDGD
mAb8	YPYDVPDYA	YDAPE



(a) (b)

Fig. 4 Top epitope estimates using estimation method EpiWP-1 for antibody **a** mAb1 (true epitope NAHYYVFFEEQE) and **b** mAb4 (true epitope RHSVV)

For most of the monoclonal antibody samples, the EpiWP-1 estimation method performs well in estimating the true epitope. Two examples of this are for estimating the epitopes of mAb1 and mAb4. The true epitope of mAb1 is NAHYYVFFEEQE; using EpiWP-1 finds YVFFEEQE as the epitope. Similarly, the epitope for mAb4 is RHSVV; EpiWP-1 estimates epitope RHSVV. The top results for these two epitopes are shown in Fig. 4a and b. In some of monoclonal antibodies, an obvious substitution but not the exact epitope is found. This is demonstrated in Fig. 5a and b for mAb8 with true epitope YPYDVPDYA. The EpiWP-1 estimation method results in candidate epitopes YDAPE and PYDAP. Allowing one amino acid substitutions as in estimation method EpiWP-2, the candidate epitopes are -YDAP and YDAP-. One way to interpret this is that YDAP is the most important portion of the epitope to binding.



(a) (b)

Fig. 5 Top epitope estimates for `mAb8` (true epitope `YPYDVPDYA`) using estimation method **a** EpiWP-1 and EpiWP-2

4.4 Comparison with Existing Epitope Identification Methods

Existing sequence alignment approaches [10, 21–26, 64] can potentially be used for the epitope estimation problem, in order to find similarities between peptide and epitope sequences. However, most of these approaches were optimized for very long amino acid sequencers and not for short-length peptide sequences [65]. An approach for finding a motif or pattern among the peptides is a direct sequence analysis approach that compares peptide sequences to epitope sequences based on their primary structure. This was demonstrated in [65] using data obtained using phage display technology; the scoring used for this approach is similarity between the sequences. Other approaches use pattern graphs, combinatorics for motif finding, exhaustive length and substitution analysis, and optimization methods to find motifs by maximizing scoring functions [66–74]. A most recent statistical based approach arranges peptides in position specific scoring matrices and computes their mean value for each position; a threshold value is then used to identify positions where the mean differs significantly [75].

Directly applied to immunosignaturing, a method called GuiTope was presented in [47] for mapping random-sequence peptides to protein sequences. The method is based on using a scoring matrix and a local alignment approach that compares similarity results using a score threshold. Using GuiTope, monoclonal antibody epitopes were estimated with about 74–81 % accuracy when aligning to a limited protein library.

5 Efficient Implementation of Epitope Estimation

The aforementioned pathogen detection and identification methods need to be repeated tens to hundreds of thousands of times to scan through all necessary peptide sequences when estimating a single epitope. Therefore, for this method to be useful, it is very important to decrease the runtime of the epitope estimation algorithm. When implemented, the algorithm spends most of its time computing the multiplication in Eq. (5). The inner product computational step involves sample multiplication and summation of all products. Reducing the number of multiplications can drastically decrease the algorithm's runtime, as discussed next.

Reducing Number of Multiplications

To increase code efficiency, the time domain waveforms described in Section 3.1 can be constructed by selecting relevant parameters T, F, and σ^2 so that the Gaussian waveforms are close together in TF but are nonoverlapping. While the Gaussian waveforms are theoretically nonzero across all time, setting $T = 3\sigma^2$ and fixing the time–bandwidth product to be $T F = 1$, is sufficient for the accuracy required in this application. The resulting Gaussian waveforms can also be sampled at Nyquist to minimize the number of samples needed to uniquely represent each frequency shift.

Frequency Domain Implementation of Epitope Estimation Method

Even after taking steps to reduce the number of time domain multiplications, it is still more efficient to represent the waveforms in the frequency domain, where each waveform is sampled once at the location of all the frequency shifts. Because the Gaussian waveforms in the dictionary are nonoverlapping, each of the frequency domain samples will either be a 1 or a 0.

Eliminating all Multiplications

For the EpiWP-1 estimation method, the multiplications in Eq. (5) can be eliminated simply by counting the number of Gaussian waveforms, in each epitope amino acid and peptide amino acid waveform pairs, that occur at the same TF location. When matched in the TF plane, a maximum matching score is obtained when all waveform pairs share the same TF support. The frequency domain implementation should still be used as this is the implementation with the smallest number of samples.

6 Conclusions

This work presented advanced signal processing approaches to analyze immunosignature biosequences. Immunosignaturing technology uses random sequence peptide microarrays to assess health status by associating antibodies from a biological sample to immune responses. The immunosignature processing requires the detection and identification of antibody epitopes from the microarray peptide sequences to discriminate between pathogens and diagnose diseases. This is achieved by first mapping characteristics of peptide and epitope sequences to parameters of highly-localized Gaussian waveforms in the time–frequency plane. After down-selecting the large number of sequences from a microarray, time–frequency based matching methods are used to estimate epitope candidates corresponding to specific pathogens. The performance of the novel epitope detection and identification method is demonstrated using eight monoclonal antibodies. The candidate sequences that resulted in a stronger response for one antibody over the others, corresponded well with the actual epitope sequences that generated the monoclonal antibodies.

Acknowledgements The authors gratefully acknowledge Dr. Stephen A. Johnston, Dr. Phillip Stafford and Joshua Richer (Director, Associate Professor Research, and Graduate Research Assistant, respectively, of the Center for Innovations in Medicine, Biodesign Institute, Arizona State University) for providing the immunosignaturing data used in this work and for valuable discussions on the epitope estimation algorithm. This work was supported by the Defense Threat Agency Reduction Contract HDTRA1-12-C-0058.

References

1. Anastassiou D. Genomic signal processing. IEEE Signal Process Mag. 2001;18(4):8–20.
2. Zhang XY, Chen F, Zhang Y-T, Agner SC, Akay M, Lu Z-H, Waye MMY, Tsui SK-W. Signal processing techniques in genomic engineering. Proc IEEE. 2002;90(12):1822–33.
3. Vaidyanathan PP. Genomics and proteomics: a signal processor's tour. IEEE Circuits Syst Mag. 2004;4(4):6–29.
4. Dougherty ER, Datta A, Sima C. Research issues in genomic signal processing. IEEE Signal Process Mag. 2005;22(6):46–68.
5. Aydin Z, Altunbasak Y. A signal processing application in genomic research: protein secondary structure prediction. IEEE Signal Process Mag. 2006;23(4):128–31.
6. Schonfeld D, Goutsias J, Shmulevich I, Tabus I, Tewfik AH. Introduction to the issue on genomic and proteomic signal processing. IEEE J Sel Top Signal Process. 2008;2.
7. Rowen L, Mahairas G, Hood L. Sequencing the human genome. Science. 1997;278(5338):605–7.
8. Schuster SC. Next-generation sequencing transforms today's biology. Nature Methods. 2008;5:16–8.
9. Durbin RM, Eddy SR, Krogh A, Mitchison G. Biological sequence analysis: probabilistic models of proteins and nucleic acids. Cambridge, England: Cambridge University Press; 1998.
10. D'Avenio G, Grigioni M, Orefici G, Creti R. SWIFT (sequence-wide investigation with Fourier transform): a software tool for identifying proteins of a given class from the unannotated genome sequence. Bioinformatics. 2005;21(13):2943–9.
11. Herzel H, Große I. Correlations in DNA sequences: the role of protein coding segments. Phys Rev E. 1997;55(1):800–10.
12. Herzel H, Trifonov EN, Weiss O, Große I. Interpreting correlations in biosequences. Phys A Stat Mech Appl. 1998;249(1):449–59.
13. Anastassiou D. Frequency-domain analysis of biomolecular sequences. Bioinformatics. 2000;16(12):1073–81.
14. Dodin G, Vandergheynst P, Levoir P, Cordier C, Marcourt L. Fourier and wavelet transform analysis, a tool for visualizing regular patterns in DNA sequences. J Theor Biol. 2000;206(3):323–6.

15. Berger JA, Mitra SK, Astola J. Power spectrum analysis for DNA sequences. IEEE Trans Signal Process. 2003;2:29–32.
16. Sussillo D, Kundaje A, Anastassiou D. Spectrogram analysis of genomes. EURASIP J Adv Signal Process. 2004; 2004(1):29–42.
17. Altaiski M, Mornev O, Polozov R. Wavelet analysis of DNA sequences. Genet Anal Biomol Eng. 1996;98(3):165–8.
18. Meng T, Soliman AT, Shyu M, Yang Y, Chen S, Iyengar SS, Yordy JS, Iyengar P. Wavelet analysis in current cancer genome research: a survey. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(6):1442–59.
19. Rockwood AL, Crockett DK, Oliphant JR, Elenitoba-Johnson KSJ. Sequence alignment by cross-correlation. J Biomol Tech. 2005;16:453–8.
20. Rosenberg MS, Editor. Sequence alignment: methods, models, concepts, and strategies. Oakland, CA: University of California Press; 2009.
21. Smith TF, Waterman MS. Identification of common molecular subsequences. J Mol Biol. 1981;147(1):195–7.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215(3):403–10.
23. Kent WJ. BLAT–The BLAST-like alignment tool. Genome Res. 2002;12(4):656–64.
24. Zhu J, Liu JS, Lawrence CE. Bayesian adaptive sequence alignment algorithms. Bioinformatics. 1998;14(1):25–39.
25. Wang W, Johnson DH. Computing linear transforms of symbolic signals. IEEE Trans Signal Process. 2002;50(3):628–34.
26. Brodzik AK. A comparative study of cross-correlation methods for alignment of DNA sequences containing repetitive patterns. European Signal Processing Conference, 2005.
27. Brodzik AK. Phase-only filtering for the masses (of DNA data): a new approach to DNA sequence alignment. IEEE Trans Signal Process. 2006;54(6):2456–66.
28. Ravichandran L, Papandreou-Suppappola A, Spanias A, Lacroix Z, Legendre C. Waveform mapping and time-frequency processing of DNA and protein sequences. IEEE Trans Signal Process. 2011;59(9):4210–24.
29. Ravichandran L. Waveform mapping and time-frequency processing of biological sequences and structures, Ph.D., Arizona State University, Tempe, Arizona, 2011.
30. O'Donnell B, Maurer A, Papandreou-Suppappola A. Waveform processing for protein multialignment by mapping locational, structural and functional attributes. Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, November 2013.
31. Delehanty JB, Ligler FS. A microarray immunoassay for simultaneous detection of proteins and bacteria. Anal Chem. 2002;74(21):5681–7.
32. Reineke U, Ivascu C, Schlief M, Landgraf C, Gericke S, Zahn G, Herzel H, Volkmer-Engert R, Schneider-Mergener J. Identification of distinct antibody epitopes and mimotopes from a peptide array of 5520 randomly generated sequences. J Immunol Methods. 2002;267(1): 37–51.
33. Duburcq X, Olivier C, Malingue F, Desmet R, Bouzidi A, Zhou F, Auriault C, Gras-Masse H, Melnyk O. Peptide-protein microarrays for the simultaneous detection of pathogen infections. Bioconjugate Chem. 2004;15(2):307–16.
34. Breitling F, Nesterov A, Stadler V, Felgenhauer T, Bischoff FR. High-density peptide arrays. Mol BioSyst. 2009;5(3):224–34.
35. Price JV, Tangsombatvisit S, Xu G, Yu J, Levy D, Baechler EC, Gozani O, Varma M, Utz PJ, Liu CL. On silico peptide microarrays for high-resolution mapping of antibody epitopes and diverse protein-protein interactions. Nature Med. 2012;18(9):1434–40.
36. Lequin RM. Enzyme immunoassay (EIA)/enzyme-linked immunosorbent assay (ELISA). Clin Chem. 2005;51(12):2415–8.
37. Chen G, Zuo Z, Zhu Q, Hong A, Zhou X, Gao X, Li T. Qualitative and quantitative analysis of peptide microarray binding experiments using SVM-PEPARRAY. Methods Mol Biol. 2009;570:403–11.

38. Renard BY, Lower M, Kuhne Y, Reimer U, Rothermel A, Tureci O, Castle JC, Sahin U. rapmad: robust analysis of peptide microarray data. BMC Bioinform. 2011;12(324):1–10.
39. Wolfinger RD, Gibson G, Wolfinger ED, Bennett L, Hamadeh H, Bushel P, Afshari C, Paules RS. Assessing gene significance from cDNA microarray expression data via mixed models. J Comput Biol. 2001;8(6):625–37.
40. Haan JR, Bauerschmidt S, van Schaik R, Piek E, Buydens L, Wehrens R. Robust ANOVA for microarray data. Chemom Intell Lab Syst. 2009;98:38–44.
41. Hirakawa A, Hamada C, Yoshimura I. Sample size calculation for a regularized t-statistic in microarray experiments. Stat Probab Lett. 2011;81:870–5.
42. Legutki JB, Magee DM, Stafford P, Johnston SA. A general method for characterization of humoral immunity induced by a vaccine or infection. Vaccine. 2010;28(28):4529–37.
43. Restrepo L, Stafford P, Magee DM, Johnston SA. Application of immunosignatures to the assessment of Alzheimer's disease. Am Neurol Assoc. 2011;70:286–95.
44. Halperin RF, Stafford P, Johnston SA. Exploring antibody recognition of sequence space through random-sequence peptide microarrays. Mol Cell Proteomics. 2011;10(3):e101230–236.
45. Brown JR, Stafford P, Johnston SA, Dinu V. Statistical methods for analyzing immunosignatures. BMC Bioinform. 2011;12(349):1–15.
46. Stafford P, Halperin R, Legutki JB, Magee DM, Galgiani J, Johnston SA. Physical characterization of the immunosignaturing effect. Mol Cell Proteomics. 2012;11(4):011 593–1–14.
47. Halperin RF, Stafford P, Emery JS, Navalkar KA, Johnston SA. GuiTope: an application for mapping random-sequence peptides to protein sequences. BMC Bioinform. 2012;13(1):1.
48. Kukreja M, Johnston SA, Stafford P. Comparative study of classification algorithms for immunosignaturing data. BMC Bioinform. 2012;13(1):139–52.
49. Legutki JB, Johnston SA. Immunosignatures can predict vaccine efficacy. Proc Natl Acad Sci. 2013;110(46):18614–9.
50. Malin A, Kovvali N, Zhang JJ, Chakraborty B, Papandreou-Suppappola A, Johnston SA, Stafford P. Adaptive learning of immunosignaturing peptide array features for biothreat detection and classification. Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, November 2011, pp. 1883–7.
51. Malin A, Kovvali N, Papandreou-Suppappola A, Zhang JJ, Johnston SA, Stafford P. Beta process based adaptive learning of immunosignaturing peptide-antibody factors. Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, November 2012, pp. 1651–5.
52. Malin A, Kovvali N, Papandreou-Suppappola A. Adaptive learning of immunosignaturing features for multi-disease pathologies. Asilomar Conference on Signals, Systems and Computers, Pacific Grove, CA, November 2013.
53. Mierendorf RC, Hammer B, Novy RE. A method for the rapid identification of epitopes and other functional peptide domains. Mol Diagn Infect Dis. 1997;13:107–23.
54. Yua X, Owensa GP, Gilden DH. Rapid and efficient identification of epitopes/mimotopes from random peptide libraries. J Immunol Methods. 2006;316:67–74.
55. Sykes KF, Legutki JB, Stafford P. Immunosignaturing: a critical review. Trends Biotechnol. 2013;31:45–51.
56. [Online]. Available: http://www.biodesign.asu.edu/research/research-centers/innovations-in-medicine/
57. Emery JS. Computational modeling of peptide-protein binding, Ph.D., Arizona State University, Tempe, Arizona, 2010.
58. Cohen L. Time-frequency analysis. Edgewood Cliffs, NJ: Prentice-Hall; 1995.
59. Mallat SG, Zhang Z. Matching pursuits with time-frequency dictionaries. IEEE Trans Signal Process. 1993;41:3397–415.
60. Petersen G, Song D, Hugle-Dorr B, Oldenburg I, Bautz EK. Mapping of linear epitopes recognized by monoclonal antibodies with gene-fragment phage display libraries. Mol & Gen Genet. 1995;249(4):425–31.
61. Goding JW. Monoclonal antibodies: principles and practice. 3rd ed. San Diego, CA: Academic Press; 1996.

62. Yip YL, Ward RL. Epitope discovery using monoclonal antibodies and phage peptide libraries. Comb Chem High Throughput Screen. 1999;2(3):125–38.
63. Clementi N, Mancini N, Castelli M, Clementi M, Burioni R. Characterization of epitopes recognized by monoclonal antibodies: experimental approaches supported by freely accessible bioinformatic tools. Drug Discov Today. 2012;18(9–10):1–8.
64. Randic M, Zupan J, Vikic-Topic D, Plavsic D. A novel unexpected use of a graphical representation of DNA: graphical alignment of DNA sequences. Elsevier Chem Phys Lett. 2006;431:375–9.
65. Mandava S, Makowski L, Devarapalli S, Uzubell J, Rodi DJ. RELIC—a bioinformatics server for combinatorial peptide analysis and identification of protein-ligand interaction site. Proteomics. 2004;4:1439–60.
66. I. Jonassen. Efficient discovery of conserved patterns using a pattern graph. Comput Appl Biosci. 1997;13:509–22.
67. Rigoutsos I, Floratos A. Combinatorial pattern discovery in biological sequences: the TEIRESIAS algorithm. Bioinformatics. 1998;14:55–67.
68. Alix AJ. Predictive estimation of protein linear epitopes by using the program PEOPLE. Vaccine. 1999;18:311–4.
69. Bastas G, Sompuram SR, Pierce B, Vani K, Bogen SA. Bioinformatic requirements for protein database searching using predicted epitopes from disease-associated antibodies. Mol Cell Proteomics. 2008;7(2):247–56.
70. Buus S, Rockberg J, Forsstrom B, Nilsson P, Uhlen M, Schafer-Nielsen C. High-resolution mapping of linear antibody epitopes using ultrahigh-density peptide microarrays. Mol Cell Proteomics. 2012;11:1790–1800.
71. Bailey TL, Williams N, Misleh C, Li WW. MEME: discovering and analyzing DNA and protein sequence motifs. Nucl Acids Res. 2006;34:369–73.
72. Bailey TL, Elkan C. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings / ... International Conference on Intelligent Systems for Molecular Biology; ISMB. International Conference on Intelligent Systems for Molecular Biology, vol. 2, pp. 28–36; 1994.
73. Andreatta M, Schafer-Nielsen C, Lund O, Buus S, Nielsen M. NNAlign: a web-based prediction method allowing non-expert end-user discovery of sequence motifs in quantitative peptide data. PLoS ONE. 2011;6(11):1–11.
74. Meskin N, Nounou H, Nounou M, Datta A. Parameter estimation of biological phenomena: an unscented Kalman filter approach. IEEE/ACM Trans Comput Biol Bioinform. 2013;10(2):537–43.
75. Hansen LB, Buus S, Schafer-Nielsen C. Identification and mapping of linear antibody epitopes in human serum albumin using high-density peptide arrays. PLoS ONE. 2013;8(7):1–10.

Wavelet–Shearlet Edge Detection and Thresholding Methods in 3D

David A. Schug, Glenn R. Easley and Dianne P. O'Leary

Abstract Edge detection in images is well studied, but using full three-dimensional information to display volumes from 3D-imaging devices or to perform pixel-level tracking of moving objects from a sequence of frames in a movie is less well understood. In this work, we study the interplay between 3D wavelet–shearlet edge detection and thresholding in accomplishing these two tasks. We find that a simple thresholding algorithm, modeling the edge image as a sum of two distributions, is very effective.

Keywords Wavelet edge detector · Shearlet edge detector · Image thresholding · 3D volume display · 3D tracking

1 Introduction

The identification of distributed discontinuities, such as edges or surface boundaries, is an important problem in computer vision and image processing. Edge classification is based on estimating the gradient norm at each pixel, but the complication comes in filtering noise. One of the most well-known and successful methods for identifying edges is due to Canny [2], but this is a single-scale algorithm.

D. A. Schug (✉)
NAWCAD, Patuxent River, MD, USA
e-mail: david.schug@navy.mil

G. R. Easley
The MITRE Corporation, 7515 Colshire Drive, McLean, VA 22102, USA
e-mail: geasley@mitre.org

D. P. O. Leary
Computer Science Department and Institute for Advanced Computer Studies,
University of Maryland, College Park, MD 20742, USA
e-mail: oleary@cs.umd.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_4

The advantages of using a multiscale approach for edge detection have widely been acknowledged since [5] and [6]. It has been shown that extending such methods to include multiple directions on multiple scales can greatly improve results [1, 7, 10]. A particular approach using the multiscale and multidirectional representation known as the shearlet representation has shown particular promise [11].

In this work, we continue our investigation [8] of extending some of these multiscale and multidirectional methods into three dimensions (3D). In particular, we focus on wavelet, shearlet, and hybrid combinations. The hybrid combinations improve efficiency by realizing that many 3D datasets encountered in practice may not exhibit complex curvilinear discontinuity structures in all three dimensions, and thus the full generality of 3D shearlets may not be needed.

One important application of edge detection on 3D objects is to better represent or visualize the data, for example, data from X-ray or MRI scans. Yet our original motivation for these 3D extensions has been to apply these techniques to motion video by viewing the third dimension as the component obtained by stacking the individual images. Using the edge/surface detections in this case, we can use this information to precisely track objects within a few pixels of their true position [8].

An important component in these algorithms is to threshold, retaining only nominated edge points that are above a magnitude considered to be background noise. Sometimes these thresholds can be preset to values for a given class of expected images and noise levels. However, in this work we will extend common 2D thresholding techniques and propose a simple new method to find the appropriate thresholding value based directly on the data .

This chapter is organized as follows. In the Section 2, the 3D edge detection problem and the basics of the proposed multiscale and multidirectional methods are given. Demonstrations of the advantages of a shearlet approach over a wavelet approach are also shown. The Section 3 describes three thresholding algorithms and shows the experimental results based on using them on a sequence of moving targets. Concluding remarks follow in the Section 4.

2 Three-Dimensional Edge Detection

Detecting changes such as edges or surface changes in 3D data has many important applications. One of these is the ability of the collection of edge intensities to be used to visualize the content of a given image data $I := [0,1]^3 \to [0,1]$. Specifically, we may loosely define the collection of edges as

$$\mathscr{E} = \left\{ t \in [0,1]^3 : |\nabla I(t)| \geq h \right\}, \tag{1}$$

the set of points for which the magnitude of the gradient of I is above a scalar threshold $h \in (0,1]$. This characterization of edges, however, is only suitable for noise free images I. To deal with noise, one solution is to prefilter the image to remove the noise before using this characterization. This prefiltering can be done by applying a Gaussian filter $g_a = \exp(-(x^2+y^2+z^2)/2a^2)$ dependent on a which determines the filter's noise dampening properties. This methodology is highly successful if the optimal parameter a can be found for the particular image of interest.

By framing this methodology in the form of a wavelet transform, Mallat et al. [5, 6] related a to the scale parameter of the transform. Specifically, the *continuous wavelet transform* of an image I is given by

$$W_\psi I(a,\tau) = \langle I, \psi_{M_a,\tau}\rangle\,, \tag{2}$$

where $M_a = aI_3$, I_3 is the 3×3 identity matrix, $\tau \in \mathbb{R}^3$, and $a > 0$. The analysis functions

$$\psi_{M_a,\tau}(t) = |\det M|^{-\frac{1}{2}}\,\psi({M_a}^{-1}(t-\tau)) \quad \tau \in \mathbb{R}^3 \tag{3}$$

are well localized waveforms that can decompose images $I \in L^2(\mathbb{R}^3)$ so that

$$I = \int_{\mathbb{R}^3} \langle I, \psi_{M_a,\tau}\rangle\, \psi_{M_a,\tau}\, d\tau. \tag{4}$$

By setting ψ to be ∇g_1, the *first derivative of a Gaussian wavelet*, the above edge detection methodology corresponds to the detection of the local maxima of the wavelet transform of I for a particular scale a. Further more, this framework allows one to develop an efficient and effective detection scheme by knowing how the magnitude of the wavelet transform behaves at location points corresponding to edges (see [8] for details).

We have found this approach to be very successful for many image data sets, as we shall demonstrate. However, when the data has sharp curvilinear elements or edges that change with complicated orientations, the wavelet transform is not effective in isolating such features. For such cases, a multidirectional representation is needed. To deal with these problems, we have developed an edge detection scheme using the shearlet representation.

2.1 The Shearlet Representation

The shearlet representation is essentially a multidirectional extension of the wavelet representation. Its unique spatial frequency tiling is achieved through the action of *shearing matrices* that give the transform its name. Shearlets are constructed by first restricting the subspace of $L^2(\mathbb{R}^3)$ to be $L^2(\mathscr{P}_1)^\vee = \{f \in L^2(\mathbb{R}^3) : \mathrm{supp}\widehat{f} \subset \mathscr{P}_1\}$, where $\mathscr{P}_1$ is the horizontal pyramidal region in the frequency plane:

$$\mathscr{P}_1 = \{(\nu_1,\nu_2,\nu_2) \in \mathbb{R}^3 : |\nu_1| \geq 2, \left|\frac{\nu_2}{\nu_1}\right| \leq 1 \text{ and } \left|\frac{\nu_3}{\nu_1}\right| \leq 1\}.$$

We define the shearlet group to be

$$\Lambda_1 = \left\{ (M^{(1)}_{as_1s_2,\tau}) : 0 \leq a \leq \frac{1}{4}, -\frac{3}{2} \leq s_1 \leq \frac{3}{2}, -\frac{3}{2} \leq s_2 \leq \frac{3}{2}, \tau \in \mathbb{R}^2 \right\}$$

where

$$M^{(1)}_{as_1s_2} = \begin{pmatrix} a & -a^{1/2}s_1 & -a^{-1/2}s_2 \\ 0 & a^{1/2} & 0 \\ 0 & 0 & a^{1/2} \end{pmatrix}.$$

The shearlet analyzing functions defined on $L^2(\mathscr{P}_1)^\vee$ are given by

$$\psi^{(1)}_{as_1s_2\tau}(t) = |\det M^{(1)}_{as_1s_2}|^{-\frac{1}{2}} \psi^{(1)}((M^{(1)}_{as_1s_2})^{-1}(t-\tau)). \tag{5}$$

In order for any function in $L^2(\mathscr{P}_1)^\vee$ to be decomposed by these analyzing functions, the following conditions and assumptions on $\psi^{(1)}$ need to be satisfied [4]. For $v = (v_1, v_2, v_3) \in \mathbb{R}^3, v_1 \neq 0$, the function $\psi^{(1)}$ should be such that

$$\hat{\psi}^{(1)}(v) = \hat{\psi}^{(1)}(v_1, v_2, v_3) = \hat{\psi}_1(v_1)\hat{\psi}_2\left(\frac{v_2}{v_1}\right)\hat{\psi}_2\left(\frac{v_3}{v_1}\right). \tag{6}$$

The function $\psi_1 \in L^2(\mathbb{R})$ should satisfy the Calderón condition

$$\int_0^\infty |\hat{\psi}_1(av)|^2 \frac{da}{a} = 1 \quad \text{for a.e.} \quad v \in \mathbb{R}$$

with supp $\hat{\psi}_1 \subset \left[-2, -\frac{1}{2}\right] \cup \left[\frac{1}{2}, 2\right]$ and $\|\psi_2\|_{L^2} = 1$ with supp $\hat{\psi}_2 \subset \left[-\frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{4}\right]$. If these assumptions are met, then

$$f(\mathbf{t}) = \int_{\mathbb{R}^3} \int_{-\frac{3}{2}}^{\frac{3}{2}} \int_{-\frac{3}{2}}^{\frac{3}{2}} \int_0^{\frac{1}{4}} \langle f, \psi^{(1)}_{as_1s_2\tau} \rangle \psi^{(1)}_{as_1s_2\tau}(\mathbf{t}) \frac{da}{a^4}\, ds_1\, ds_2\, d\tau \tag{7}$$

for all $f \in L^2(\mathscr{P}_1)^\vee$.

The shearlet analyzing functions $\psi^{(1)}_{as_1s_2\tau}$ in the frequency domain are given by

$$\hat{\psi}^{(1)}_{as_1s_2\tau}(v_1, v_2, v_3) = a\hat{\psi}_1(av_1)\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{v_2}{v_1} - s_1))\hat{\psi}_2(a^{-\frac{1}{2}}(\frac{v_3}{v_1} - s_2))e^{-2\pi i v \cdot \tau}. \tag{8}$$

Because of the particular support constraints given, this means each function $\hat{\psi}^{(1)}_{as_1s_2\tau}$ has support described by the elements (v_1, v_2, v_3) such that for a given s_1, s_2, and $v_1 \in \left[-\frac{2}{a}, -\frac{1}{2a}\right] \cup \left[\frac{1}{2a}, \frac{2}{a}\right]$, v_2 needs to satisfy $\left|\frac{v_2}{v_1} - s_1\right| \leq \frac{\sqrt{2}}{4}a^{\frac{1}{2}}$, and v_3 needs to satisfy $\left|\frac{v_3}{v_1} - s_2\right| \leq \frac{\sqrt{2}}{4}a^{\frac{1}{2}}$. Such points end up describing a pair of hyper-trapezoids that are symmetric with respect to the origin with orientation determined by slope parameters s_1 and s_2. These hyper-trapezoids become elongated as $a \to 0$.

Since the analyzing functions $\psi^{(1)}_{as_1s_2\tau}$ only decompose elements in $L^2(\mathscr{P}_1)^\vee$, we form complementary analyzing functions supported on the complementary pyramidal regions. Specifically, we define

$$\mathscr{P}_2 = \{(\nu_1,\nu_2,\nu_2) \in \mathbb{R}^3 : |\nu_2| \geq 2, \left|\frac{\nu_1}{\nu_2}\right| \leq 1 \text{ and } \left|\frac{\nu_3}{\nu_2}\right| \leq 1\}$$

and

$$\mathscr{P}_3 = \{(\nu_1,\nu_2,\nu_2) \in \mathbb{R}^3 : |\nu_3| \geq 2, \left|\frac{\nu_2}{\nu_3}\right| \leq 1 \text{ and } \left|\frac{\nu_1}{\nu_3}\right| \leq 1\}.$$

We also define

$$\Lambda_2 = \left\{(M^{(2)}_{as_1s_2,\tau}) : 0 \leq a \leq \frac{1}{4}, -\frac{3}{2} \leq s_1 \leq \frac{3}{2}, -\frac{3}{2} \leq s_2 \leq \frac{3}{2}, \tau \in \mathbb{R}^2\right\}$$

where,

$$M^{(2)}_{as_1s_2} = \begin{pmatrix} a^{1/2} & 0 & 0 \\ -a^{1/2}s_1 & a & -a^{-1/2}s_2 \\ 0 & 0 & a^{1/2} \end{pmatrix}.$$

Likewise, we define

$$\Lambda_3 = \left\{(M^{(3)}_{as_1s_2,\tau}) : 0 \leq a \leq \frac{1}{4}, -\frac{3}{2} \leq s_1 \leq \frac{3}{2}, -\frac{3}{2} \leq s_2 \leq \frac{3}{2}, \tau \in \mathbb{R}^2\right\}$$

where,

$$M^{(3)}_{as_1s_2} = \begin{pmatrix} a^{1/2} & 0 & 0 \\ 0 & a^{1/2} & 0 \\ -a^{1/2}s_1 & -a^{1/2}s_2 & a \end{pmatrix}.$$

By defining $\psi^{(2)}$ and $\psi^{(3)}$ as $\hat{\psi}^{(2)}(\nu) = \hat{\psi}^{(2)}(\nu_1,\nu_2,\nu_3) = \hat{\psi}_1(\nu_2)\hat{\psi}_2\left(\frac{\nu_1}{\nu_2}\right)\hat{\psi}_2\left(\frac{\nu_3}{\nu_2}\right)$ and $\hat{\psi}^{(3)}(\nu) = \hat{\psi}^{(3)}(\nu_1,\nu_2,\nu_3) = \hat{\psi}_1(\nu_3)\hat{\psi}_2\left(\frac{\nu_1}{\nu_3}\right)\hat{\psi}_2\left(\frac{\nu_2}{\nu_3}\right)$, the analyzing functions

$$\psi^{(j)}_{as_1s_2\tau}(t) = |\det M^{(j)}_{as_1s_2}|^{-\frac{1}{2}}\psi^{(j)}((M^{(j)}_{as_1s_2})^{-1}(t-\tau)), \tag{9}$$

for $j = 2,3$ decompose the subspaces $L^2(\mathscr{P}_2)^\vee$ and $L^2(\mathscr{P}_3)^\vee$, respectively. Since the union of $L^2(\mathscr{P}_1)^\vee$, $L^2(\mathscr{P}_2)^\vee$, and $L^2(\mathscr{P}_3)^\vee$ forms the space $L^2(\mathbb{R}^3)$ minus the functions whose frequency supports are contained in $[-2,2]^3$, we can obtain a complete decomposition of $L^2(\mathbb{R}^3)$ by adding analyzing functions that can decompose these elements. This is done by using an appropriate bandlimited window function φ and forming the analyzing functions $\varphi_\tau(t) = \varphi(t-\tau)$.

The shearlet representation consists of the collection of analyzing functions $\{\psi^{(j)}_{as_1s_2\tau}\}^3_{j=1}$ and φ restricted to the appropriate groups, but for simplicity of notation we will drop the superscript. Examples of the spatial frequency hypertrapezoidal regions for these atoms are shown in Fig. 1.

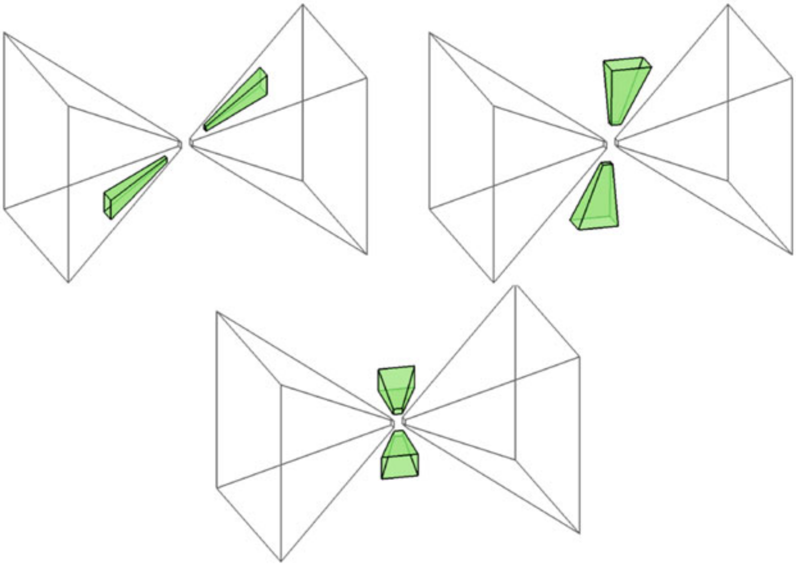


Fig. 1 Three-dimensional spatial frequency representations (four disjoint domains for each) of shearlet atoms for three different choices of scales and shearing parameters

We denote $\mathscr{S}\mathscr{H}_{\psi}f(a,s_1,s_2,\tau)$ to be $\langle f,\psi_{as_1s_2\tau}\rangle$. We are able to design an edge detection routine by using the following result [4] that characterizes the asymptotic decay as $a \to 0$ for edge point locations.

Theorem 1. [4] *Let Ω be a bounded region in $\mathbb{R}^3$ with boundary $\partial\Omega$ and define the function B to be the characteristic function over Ω. Assume that $\partial\Omega$ is a piecewise smooth two-dimensional manifold. Let γ_j, $j = 1,2,\ldots,m$ be the separating curves of $\partial\Omega$. Then we have*

1. *If $\tau \notin \partial\Omega$, then*
$$\lim_{a\to 0^+} a^{-N}\mathscr{S}\mathscr{H}_{\psi}B(a,s_1,s_2,\tau) = 0 \quad \textit{for all } N > 0.$$
2. *If $\tau \notin \partial\Omega \setminus \cup_{j=1}^{m}\gamma_j$ and (s_1,s_2) does not correspond to the normal direction of $\partial\Omega$ at τ, then*
$$\lim_{a\to 0^+} a^{-N}\mathscr{S}\mathscr{H}_{\psi}B(a,s_1,s_2,\tau) = 0 \quad \textit{for all } N > 0.$$
3. *If $\tau \notin \partial\Omega \setminus \cup_{j=1}^{m}\gamma_j$ and (s_1,s_2) corresponds to the normal direction of $\partial\Omega$ at τ or $\tau \in \cup_{j=1}^{m}\gamma_j$ and (s_1,s_2) corresponds to one of the two normal directions of $\partial\Omega$ at p, then*
$$\lim_{a\to 0^+} a^{-1}\mathscr{S}\mathscr{H}_{\psi}B(a,s_1,s_2,\tau) \neq 0.$$

4. *If* $\tau \in \gamma_j$ *and* (s_1, s_2) *does not correspond to the normal directions of* $\partial\Omega$ *at* τ*, then*

$$|\mathscr{S}\mathscr{H}_{\psi}B(a,s_1,s_2,\tau)| \leq Ca^{\frac{3}{2}}.$$

This result allows us to use the concepts developed in [11] for 2D to be extended to obtain a 3D shearlet edge detection algorithm. Details of the algorithm are given in [8].

2.2 Visualization

An edge map of a 3D dataset is particularly useful for visualizing complex objects by taking advantage of the ability of an alpha map to give some transparency to the detected surfaces [9]. In this section, we demonstrate the capability of the wavelet and shearlet edge detection schemes to be used for visualization.

For implementation, we use the thresholding technique called hysteresis to determine the true edge intensity magnitudes. Specifically, hysteresis uses two diferent thresholds t_{low} and t_{high} to help distinguish true edges, even if the magnitude of the gradient is somewhat below the value h specified in (1). A pixel is identified as a strong edge pixel if its intensity gradient magnitude is greater than t_{high}. A pixel is also marked as part of an edge if it is connected to a strong edge and its gradient magnitude is larger than t_{low} and larger than the magnitude of each of its two neighbors in at least one of the compass directions (N–S, E–W, NE–SW, NW–SE). This 2D hysteresis is applied to the 3D intensity magnitudes M on a slice by slice basis.

Assuming the magnitudes M of the gradient intensities are normalized, t_{high} is given as $\eta\overline{M}$ for some $\eta > 0$ where $\overline{M}$ denotes the mean of M and t_{low} is given as ρt_{high} for some $\rho < 1$.

We illustrate these techniques on two examples.

In the first example, we added white Gaussian noise with a standard deviation of 0.1 to the 3D Shepp-Logan Phantom dataset. In this case, we set $\eta = 4.1$ and $\rho = 0.45$. Figures 1, 2, 3, and 4 show the results. The 3D shearlet edge detector gives a higher quality rendering of edge information.

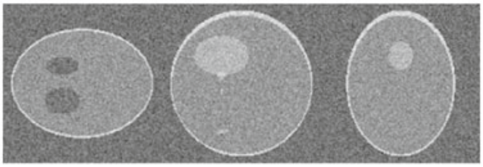


Fig. 2 Data for the 3D phantom experiment: images of slices through main axis

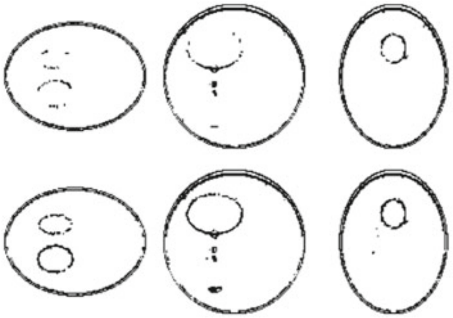


Fig. 3 Results for the 3D phantom experiment with hysteresis filtering : images of slices. *Top*: results of wavelet-based edge detection. *Bottom*: corresponding results for shearlet-based edge detection

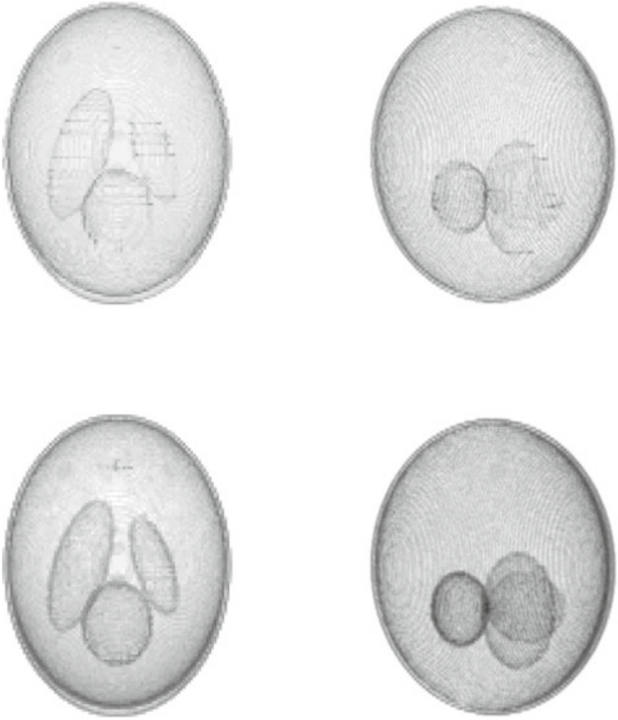


Fig. 4 Results for the 3D phantom experiment with hysteresis filtering: 3D display. *Top*: results of wavelet-based edge detection. *Bottom*: corresponding results for shearlet-based edge detection

In the second example, we added a white Gaussian noise with a standard deviation of 0.2 to a spherical harmonic of order 2 and degree 6 with a shading applied. In this case, we set $\eta = 4.3$ and $\rho = 0.45$. Figures 5, 6, and 7 show the results. Again, the shearlet-based edge detector gives a better reconstruction.

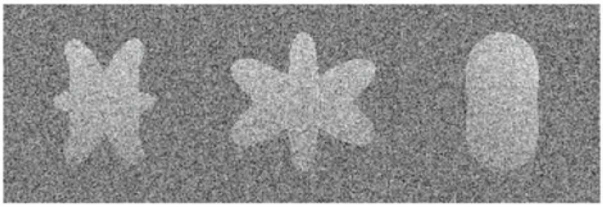


Fig. 5 Data for the 3D spherical harmonic experiment: images of slices through main axis

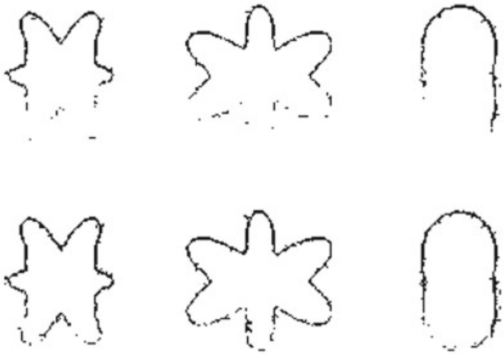


Fig. 6 Results for the 3D spherical harmonic experiment with hysteresis filtering: images of slices through main axis. *Top*: results of wavelet-based edge detection. *Bottom*: corresponding results for shearlet-based edge detection

3 Thresholding the Results of Edge Detectors

In this section, we study the use of various edge detection algorithms and various thresholding algorithms in determining edges in a noisy sequence of images.

3.1 The Eight Edge Detection Algorithms

We used several edge detection algorithms: 2D and 3D versions of the Canny, wavelet, and shearlet edge detectors, as well as hybrid wavelet (shearlet) edge detectors that combine the results of 2D slices in the xy, xt, and yt directions. We used

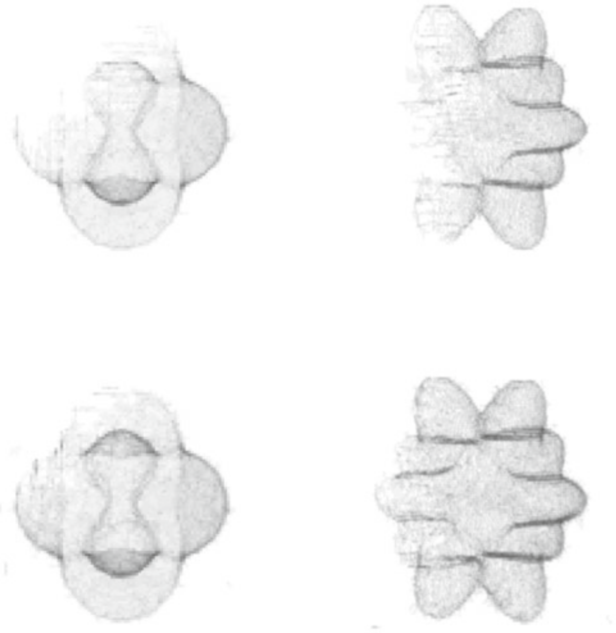


Fig. 7 Results for the 3D spherical harmonic experiment with hysteresis filtering: 3D display. *Top*: results of wavelet-based edge detection. *Bottom*: corresponding results for shearlet-based edge detection

MATLAB's implementation of the 2D Canny algorithm in `edge.m`. The other algorithms are described in detail in [8] and the MATLAB implementations are available at `https://www.cs.umd.edu/users/oleary/software/`.

Note that the Canny algorithm returns a 0-1 image, so thresholding cannot be applied.

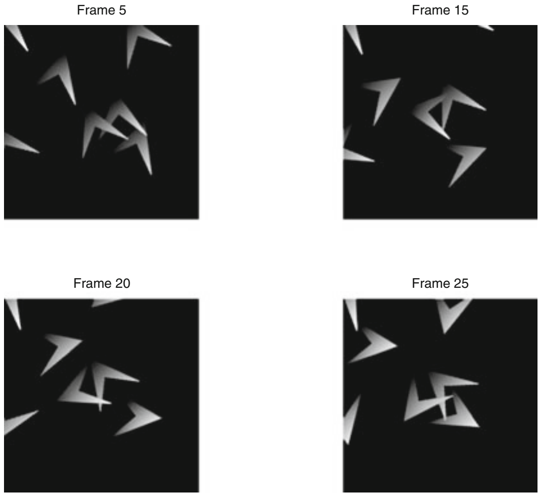
3.2 The Three Thresholding Algorithms

The thresholding algorithms were one taken from Gonzales [3, p. 406], Otsu's method as implemented in MATLAB's `graythreshold.m`, and a method that we developed.

Gonzales determines the threshold iteratively. He sets the threshold halfway between the mean of the pixels currently labeled "black" and those labeled "white." The iteration terminates when the change in mean is less than a specified tolerance. In the figures, this method is referred to as the "global" method.

Otsu's method aims to choose the threshold to minimize the sum of the variance among pixels labeled "black" and the variance among pixels labeled "white," weighted by the proportions of pixels in each group. We applied it frame-by-frame to the norm-squared of the gradient estimate from the edge detector.

Our method assumes that in an edge image, an overwhelming number of pixels should be labeled "black." Therefore, we set the threshold to three times the standard deviation of the pixel values. We applied it separately to the three components of the gradient estimate from our edge detector, and it is referred to as the "stat" method in the figures.


Frame 5

Frame 15

Frame 20

Frame 25

Fig. 8 Frames 5, 15, 20, and 25 from the movie with shading and no noise

3.3 Tests on Moving Objects

To test our algorithms, we generated a 30-frame movie containing seven wedges translating and rotating at different velocities. We added shading and noise to make the problem more difficult. Several frames of the movie (with shading but no noise) are shown in Fig. 8. We display the results of our algorithms on frame 20.

No noise, no shading: Figs. 9, 10, and 11

In this case, edge detection is rather easy, and all of the algorithms do well, although the 3D wavelet version tends to broaden the edges due to motion.

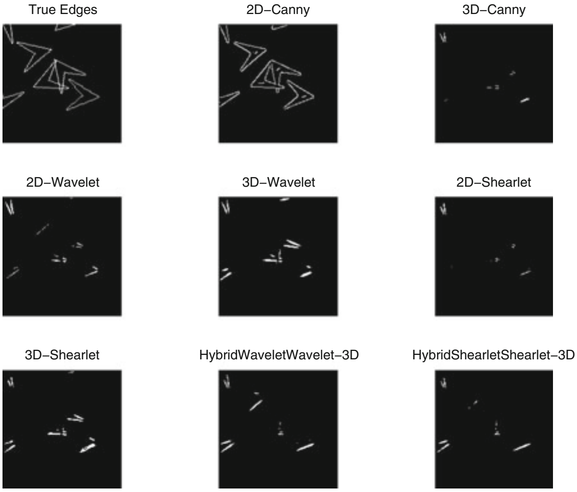


Fig. 9 Results of edge detection on movie with no noise using the *global* thresholding algorithm

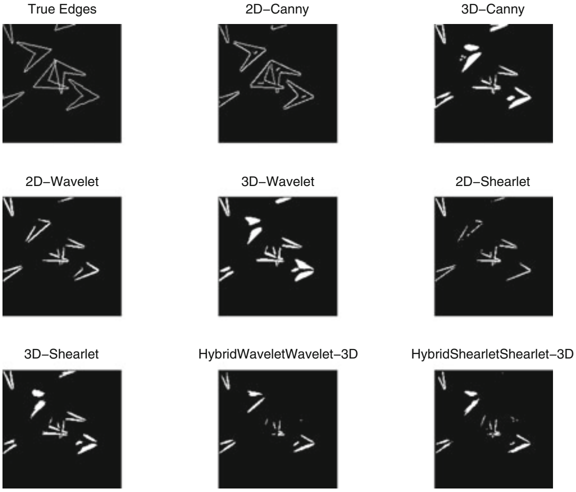


Fig. 10 Results of edge detection on movie with no noise using the *Otsu* thresholding algorithm

Noise with shading: Figs. 12, 13, and 14

The Canny algorithm breaks down when noise (standard deviation of 2) is added (using MATLAB default parameters). The Gonzales threshold is again too small for the wavelet algorithms, but with the other two threshold algorithms, the 2D and 3D wavelets continue find all seven wedges. Although the 3D has considerable broadening, it finds all of the wedge edges more reliably.

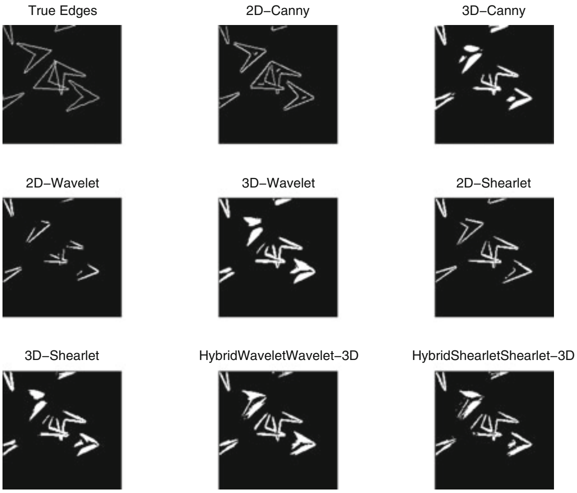


Fig. 11 Results of edge detection on movie with no noise using the *stat* thresholding algorithm

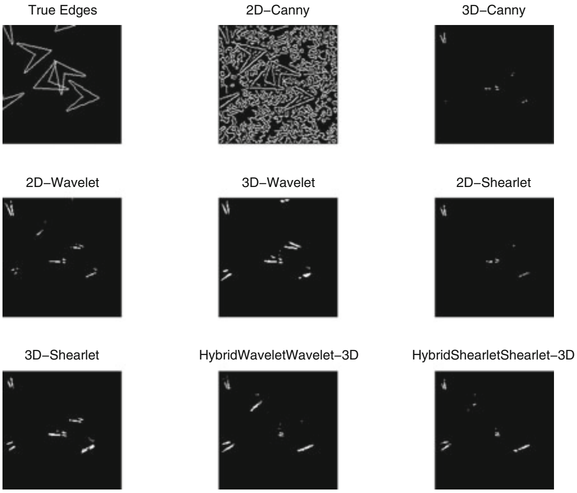


Fig. 12 Results of edge detection on movie using the *global* thresholding algorithm. Standard deviation of the noise is equal to 0.2 relative to *white* pixels

Increased noise with shading: Figs. 15, 16, and 17

When the standard deviation of the noise is increased to 4, it is hard to find the wedges in the output of the 2D wavelet, but the seven wedges are somewhat visible in the 3D wavelet-based results.

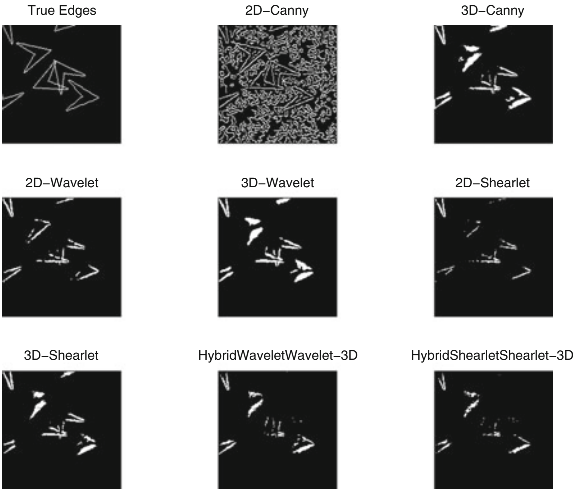


Fig. 13 Results of edge detection on movie using the *Otsu* thresholding algorithm. Standard deviation of the noise is equal to 0.2 relative to *white* pixels

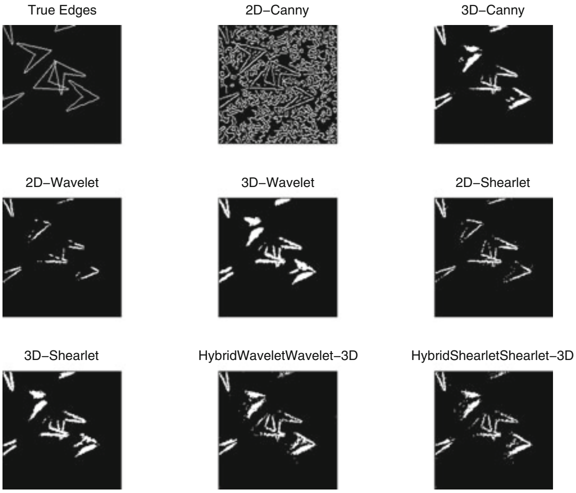


Fig. 14 Results of edge detection on movie using the *stat* thresholding algorithm. Standard deviation of the noise is equal to 0.2 relative to *white* pixels

Results for the spherical harmonic example: Figs. 18 and 19

For comparison with Figs. 6 and 7, we applied our algorithms to the results for the spherical harmonic example. As seen in Figs. 18 and 19, the trends persist. The *Otsu* algorithm does not produce good results. The *global* algorithm allows too much noise. The *stat* algorithm is a little too conservative in declaring edges but produces good results.

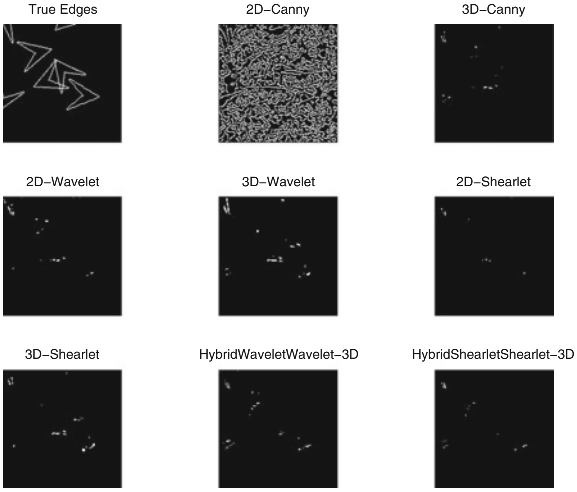


Fig. 15 Results of edge detection on movie using the *global* thresholding algorithm. Standard deviation of the noise is equal to 0.4 relative to *white* pixels

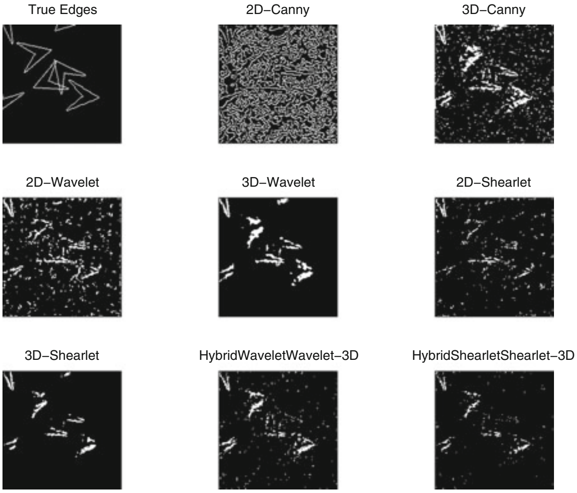


Fig. 16 Results of edge detection on movie using the *Otsu* thresholding algorithm. Standard deviation of the noise is equal to 0.4 relative to *white* pixels

4 Conclusion

We have developed 2D and 3D wavelet, shearlet, and hybrid combination based edge detection algorithms. Our investigation focused on combining edge detectors with different thresholding methods to improve the ability to differentiate image features from the background in the presence of noise. The value of a particular method is dependent on the nature of the problem being considered. For rigid sta-

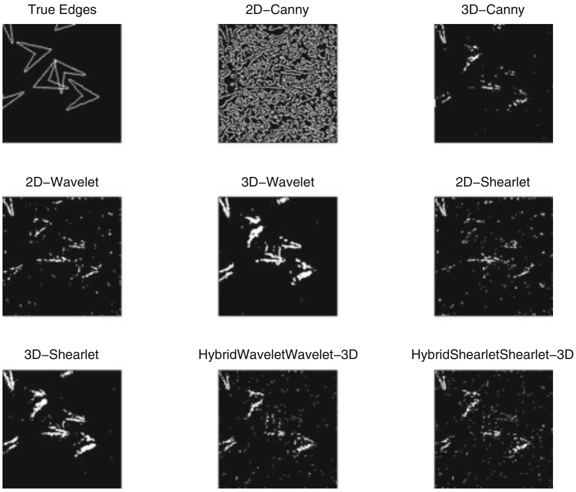


Fig. 17 Results of edge detection on movie using the stat thresholding algorithm. Standard deviation of the noise is equal to 0.4 relative to *white* pixels

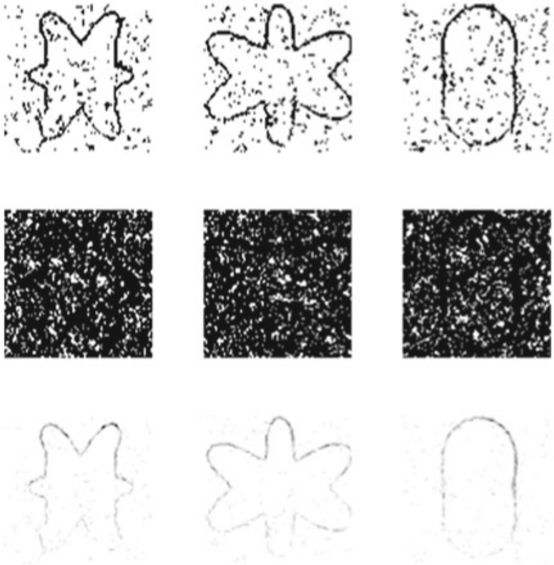


Fig. 18 Results of edge detection on shearlet results from spherical harmonic example. Top: *global* thresholding. Middle: *Otsu* algorithm. Bottom: *stat* thresholding. Compare with the *bottom row* of Fig. 6

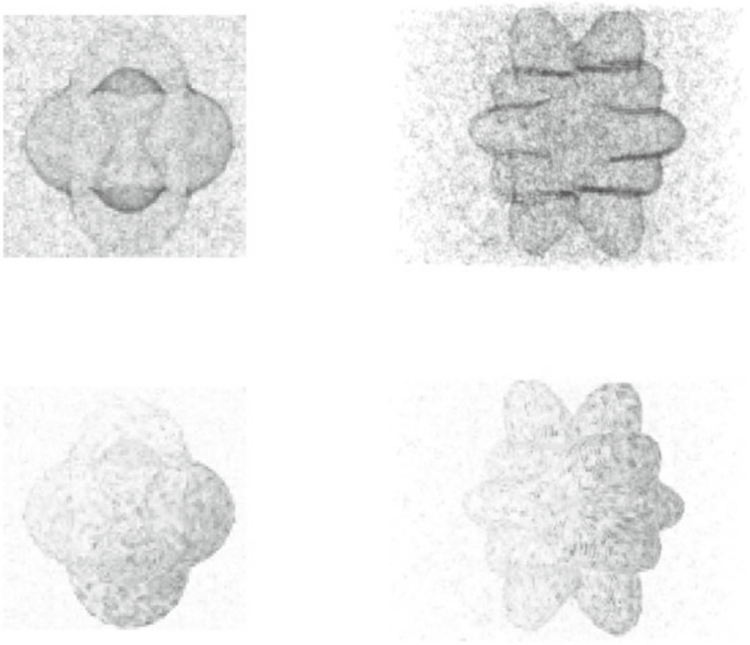


Fig. 19 Results of edge detection using the 3D shearlet version on the spherical harmonic example. *Top*: *global* thresholding. *Bottom*: *stat* thresholding. Otsu results are unusable in this case. Compare with Fig. 7.

tionary image features, the 2D and 3D methods perform equally well. When features change scale and orientation rapidly with a curvilinear description, the 3D shearlet method has been shown to perform better as seen in the visualization of the phantom and spherical harmonic examples. For dynamic image features that change location and orientation rapidly over time, 3D methods have a distinct advantage. Each 3D method incorporates significant horizontal, vertical, and time gradient information over a fixed window of neighboring image slices. Visually, this means that each edge image feature has additional thickness because neighboring directional gradients are accumulated. In terms of tracking, this feature is in fact beneficial. As the noise level increases, 2D methods do not account for the neighboring intensity changes and show a degraded performance.

The proposed 3D thresholding methods are efficient to implement so that they can be used as components to a tracking algorithm. On average our *stat* method performed the best on most cases. For shaded stationary objects the wavelet methods combined with *Otsu* are adequate. However, for shaded dynamic image features, the 3D Shearlet detection combined with *stat* thresholding seemed to perform the best.

References

1. Aydin T, Yemez Y, Anarim E, Sankur B. Multidirectional and multiscale edge detection via m-band wavelet transform. IEEE Trans Image Process. 1996;5(9):1370–7.
2. Canny JF. A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell. 1986; PAMI-8(6):679–98.
3. Gonzalez RC, Woods RE. Digital image processing. Saddle River: Prentice Hall; 2002.
4. Guo K, Labate D. Analysis and detection of surface discontinuities using the 3D continuous shearlet transform. Appl Comput Harmon Anal. 2010;30(2):231–49.
5. Mallat SA, Hwang WL. Singularity detection and processing with wavelets. IEEE Trans Inf Theory. 1992;38(2):617–643.
6. Mallat SA, Zhong S. Characterization of signals from multiscale edges. IEEE Trans Pattern Anal Mach Intell. 1992;14(7):710–732.
7. Pietro P. Steerable-scalable kernels for edge detection and junction analysis. Image Vision Comput. 1992;10(10):663–72.
8. Schug DA, Easley GR, O'Leary DP. Tracking objects using three dimensional edge detection, SIAMCSE, 2013.
9. Toklu C, Murat Tekalp A, Tanju Erdem A. Simultaneous alpha map generation and 2-D mesh tracking for multimedia applications. In Image Processing, 1997. Proceedings., International Conference on, volume 1, pp. 113–116. IEEE, 1997.
10. Tu C-L, Wen-Liang H, Ho J. Analysis of singularities from modulus maxima of complex wavelets. IEEE Trans Inform Theory. 2005;51(3):1049–62.
11. Yi S, Labate D, Easley GR, Krim H. A shearlet approach to edge analysis and detection. IEEE Trans Image Process. 2009;18(5):929–41.

Recursive Computation of Spherical Harmonic Rotation Coefficients of Large Degree

Nail A. Gumerov and Ramani Duraiswami

Abstract Computation of the spherical harmonic rotation coefficients or elements of Wigner's d-matrix is important in a number of quantum mechanics and mathematical physics applications. Particularly, this is important for the fast multipole methods in three dimensions for the Helmholtz, Laplace, and related equations, if rotation-based decomposition of translation operators is used. In these and related problems related to representation of functions on a sphere via spherical harmonic expansions computation of the rotation coefficients of large degree n (of the order of thousands and more) may be necessary. Existing algorithms for their computation, based on recursions, are usually unstable, and do not extend to n. We develop a new recursion and study its behavior for large degrees, via computational and asymptotic analyses. Stability of this recursion was studied based on a novel application of the Courant-Friedrichs-Lewy condition and the von Neumann method for stability of finite-difference schemes for solution of PDEs. A recursive algorithm of minimal complexity $O\left(n^2\right)$ for degree n and FFT-based algorithms of complexity $O\left(n^2 \log n\right)$ suitable for computation of rotation coefficients of large degrees are proposed, studied numerically, and cross-validated. It is shown that the latter algorithm can be used for $n \lesssim 10^3$ in double precision, while the former algorithm was tested for large n (up to 10^4 in our experiments) and demonstrated better performance and accuracy compared to the FFT-based algorithm.

Keywords SO(3) · Spherical harmonics · Recursions · Wigner d-matrix · Rotation

N. A. Gumerov (✉)
Institute for Advanced Computer Studies, University of Maryland, College Park, USA
e-mail: gumerov@umiacs.umd.edu

R. Duraiswami
Department of Computer Science and Institute for Advanced Computer Studies,
University of Maryland, College Park, USA
e-mail: ramani@umiacs.umd.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_5

1 Introduction

Spherical harmonics form an orthogonal basis for the space of square integrable functions defined over the unit sphere, S_u, and have important application for a number of problems of mathematical physics, interpolation, approximation, and Fourier analysis on the sphere. Particularly, they are eigenfunctions of the Beltrami operator on the sphere, and play a key role in the solution of Laplace, Helmholtz, and related equations (polyharmonic, Stokes, Maxwell, Schroedinger, etc.) in spherical coordinates. Expansions of solutions of these equations via spherical basis functions, whose angular part are the spherical harmonics (multipole and local expansions), are important in the fast multipole methods (FMM) [1–5].

The FMM for the Helmholtz equation as well as other applications in geostatistics require operations with expansions which involve large numerical values of the maximum degree of the expansion, p. This may reach several thousands, and the expansions, which have p^2 terms, will have millions of coefficients. For the FMM for the Helmholtz equation, such expansions arise when the domain has size of the order of $M \sim 100$ wavelengths, and for convergence we need $p \sim O(M)$. These large expansions need to be translated (change of the origin of the reference frame). Translation operators for truncated expansions of degree p (p^2 terms) can be represented by dense matrices of size $\left(p^2\right)^2 = p^4$ and, the translation can be performed via matrix vector product with cost $O\left(p^4\right)$. Decomposition of the translation operators into rotation and coaxial translation parts (the RCR-decomposition: rotation-coaxial translation-back rotation) [5, 6] reduces this cost to $O\left(p^3\right)$. While translation of expansions can be done with asymptotic complexity $O\left(p^2\right)$ using diagonal forms of the translation operators [2], use of such forms in the multilevel FMM requires additional operations, namely, interpolation and anterpolation, or filtering of spherical harmonic expansions, which can be performed for $O\left(p^2 \log p\right)$ operations. The practical complexity of such filtering has large asymptotic constants, so that $\left(p^3\right)$ methods are competitive with asymptotically faster methods for p up to several hundreds [7]. So, wideband FMM for the Helmholtz equation for such p can be realized in different ways, including [8] formally scaled as $O\left(p^2 \log p\right)$ and [9], formally scaled as $O\left(p^3\right)$, but with comparable or better performance for the asymptotically slower method.

Rotation of spherical harmonic expansions is needed in several other applications (e.g., [25], and is interesting from a mathematical point of view, and has deep links with group theory [19]. Formally, expansion of degree p can be rotated for the expense of p^3 operations, and, in fact, there is a constructive proof that this can be done for the expense of $O\left(p^2 \log p\right)$ operations [5]. The latter is related to the fact that the rotation operator (matrix) can be decomposed into the product of diagonal and Toeplitz/Hankel matrices, where the matrix-vector multiplication involving the Toeplitz/Hankel matrices can be performed for $O\left(p^2 \log p\right)$ operations using the FFT. There are two issues which cause difficulty with the practical realization of such an algorithm. First, the matrix-vector products should be done for $O(p)$ matrices of sizes $O\left(p \times p\right)$ each, so for $p \sim 10^2 - 10^3$ the efficiency of the Toeplitz matrix-vector multiplication of formal complexity $O(p \log p)$ per matrix involving two FFTs is not so great compared to a direct matrix-vector product, and so the practical complexity is comparable with $O\left(p^2\right)$ brute-force multiplication due to large enough asymptotic constant for the FFT. Second, the decomposition shows poor scaling of the Toeplitz matrix (similar to Pascal matrices), for which renormalization can be done for some range of p, but is also algorithmically costly [10].

Hence, from a practical point of view $O\left(p^3\right)$ methods of rotation of expansions are of interest. Efficient $O\left(p^3\right)$ methods are usually based on direct application of the rotation matrix to each rotationally invariant subspace, where the the rotation coefficients are computed via recurrence relations. There are numerous recursions, which can be used for computation of the rotation coefficients (e.g. [11–14], see also the review in [15]), and some of them were successfully applied for solution of problems with relatively small p ($p \lesssim 100$). However, attempts to compute rotation coefficients for large p using these recursions face numerical instabilities. An $O\left(p^3\right)$ method for rotation of spherical expansion, based on pseudospectral projection, which does not involve explicit computation of the rotation coefficients was proposed and tested in [15]. The rotation coefficient values are however needed in some applications. For example for the finite set of fixed angle rotations encountered in the FMM, the rotation coefficients can be precomputed and stored. In this case the algorithm which simply uses the precomputed rotation coefficients is faster than the method proposed in [15], since brute force matrix-vector multiplications do not require additional overheads related to spherical harmonic evaluations and Fourier transforms [15], and are well optimized on hardware.

We note that almost all studies related to computation of the rotation coefficients that advertise themselves as “fast and stable”, in fact, do not provide an actual stability analysis. “Stability” then is rather a reflection of the results of numerical experiments conducted for some limited range of degrees n. Strictly speaking, all algorithms that we are aware of for this problem are not proven stable in the strict sense—that the error in computations is not increasing with increasing n. While there are certainly unstable schemes, which “blow up” due to exponential error growth, there are some unstable schemes with slow error growth rate $O\left(n^{\alpha}\right)$ at large n.

In the present study, we investigate the behavior of the rotation coefficients of large degree and propose a “fast and stable” $O\left(p^3\right)$ recursive method for their computation, which numerically is much more stable than other algorithms based on recursions used in the previous studies. We found regions where the recursive processes used in the present scheme are unstable as they violate a Courant-Friedrichs-Lewy (CFL) stability condition [16]. The proposed algorithm manages this. We also show that in the regions which satisfy the CFL condition the recursive computations despite being formally unstable have a slow error growth rate. Such conclusion comes partially from the well-known von Neumann stability analysis [17] combined with the analysis of linear one-dimensional recursions and partially from the numerical experiments on noise amplification when using the recursive algorithm. The proposed algorithm was tested for computation of rotation coefficients of degrees up to $n = 10^4$, without substantial constraints preventing their use for larger

n. We also proposed and tested a non-recursive FFT-based algorithm of complexity $O\left(p^3 \log p\right)$, which despite larger complexity and higher errors than the recursive algorithm, shows good results for $n \lesssim 10^3$ and can be used for validation (as we did) and other purposes.

2 Preliminaries

2.1 Spherical Harmonic Expansion

Cartesian coordinates of points on the unit sphere are related to the angles of spherical coordinates as

$$\mathbf{s} = (x, y, z) = (\sin\theta\cos\varphi, \sin\theta\sin\varphi, \cos\theta), \tag{1}$$

We consider functions $f \in L_2(S_u)$ of bandwidth p, means expansion of $f(\theta,\varphi)$ over spherical harmonic basis can be written in the form

$$f(\theta,\varphi) = \sum_{n=0}^{p-1} \sum_{m=-n}^{n} C_n^m Y_n^m(\theta,\varphi), \tag{2}$$

where orthonormal spherical harmonics of degree n and order m are defined as

$$Y_n^m(\theta,\varphi) = (-1)^m \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m|)!}{(n+|m|)!}} P_n^{|m|}(\cos\theta) e^{im\varphi}, \tag{3}$$
$$n = 0, 1, 2, \ldots; \quad m = -n, \ldots, n.$$

Here $P_n^m(\mu)$ are the associated Legendre functions, which are related to the Legendre polynomials, and can be defined by the Rodrigues' formula

$$P_n^m(\mu) = (-1)^m \left(1-\mu^2\right)^{m/2} \frac{d^m P_n(\mu)}{d\mu^m}, \; n = 0, 1, 2, \ldots, \; m = 0, 1, 2, \ldots \tag{4}$$
$$P_n(\mu) = \frac{1}{2^n n!} \frac{d^n}{d\mu^n} \left(\mu^2 - 1\right)^n, \quad n = 0, 1, 2, \ldots$$

The banwidth p can be arbitrary, and the fact that we consider finite number of harmonics relates only to computations. We also note that different authors use slightly different definitions and normalizations for the spherical harmonics. Our definition is consistent with that of [3]. Some discussion on definitions of spherical harmonics and their impact on translation relations can also be found there. Particularly, for spherical harmonics defined as

$$\widetilde{Y}_n^m(\theta,\varphi) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} P_n^{|m|}(\cos\theta)e^{im\varphi}, \tag{5}$$

$$n = 0,1,2,...; \qquad m = -n,...,n,$$

$$Y_n^m(\theta,\varphi) = \varepsilon_m \widetilde{Y}_n^m(\theta,\varphi), \tag{6}$$

where

$$\varepsilon_m = \begin{cases} (-1)^m, & m \geqslant 0, \\ 1, & m < 0. \end{cases} \tag{7}$$

Hence one can expect appearance of factors ε_m in relations used by different authors.

2.2 Rotations

There are two points of view on rotations, active (alibi), where vectors are rotated in a fixed reference frame, and passive (alias), where vectors are invariant objects but the reference frame rotates and so the coordinates of vectors change. In the present chapter we use the latter point of view, while it is not difficult to map the relations to the active view by replacing rotation matrices by their transposes (or inverses).

An arbitrary rotation transform can be specified by three Euler angles of rotation. We slightly modify these angles to be consistent with the rotation angles α,β,γ defined in [5]. Let $\mathbf{i}_x,\mathbf{i}_y$, and $\mathbf{i}_z$ be the Cartesian basis vectors of the original reference frame, while $\widehat{\mathbf{i}}_x,\widehat{\mathbf{i}}_y$, and $\widehat{\mathbf{i}}_z$ be the respective basis vectors of the rotated reference frame. Cartesian coordinates of $\widehat{\mathbf{i}}_z$ in the original reference frame and $\mathbf{i}_z$ in the rotated reference frame can be written as

$$\begin{aligned} \widehat{\mathbf{i}}_z &= (\sin\beta\cos\alpha, \sin\beta\sin\alpha, \cos\beta), \\ \mathbf{i}_z &= (\sin\beta\cos\gamma, \sin\beta\sin\gamma, \cos\beta). \end{aligned} \tag{8}$$

Figure 1 illustrates the rotation angles and reference frames. Note that

$$Q^{-1}(\alpha,\beta,\gamma) = Q^T(\alpha,\beta,\gamma) = Q(\gamma,\beta,\alpha), \tag{9}$$

where Q is the rotation matrix and superscript T denotes transposition. We also note that as β is related to spherical angle θ, so its range can be limited by half period $\beta \in [0,\pi]$, while for α and γ we have full periods $\alpha \in [0,2\pi)$, $\gamma \in [0,2\pi)$.

Let us introduce the Euler rotation angles, α_E, β_E, and γ_E, where general rotation is defined as rotation around original z axis by angle α_E followed by rotation around about the new y axis by angle β_E and, finally, by rotation around the new z axis by

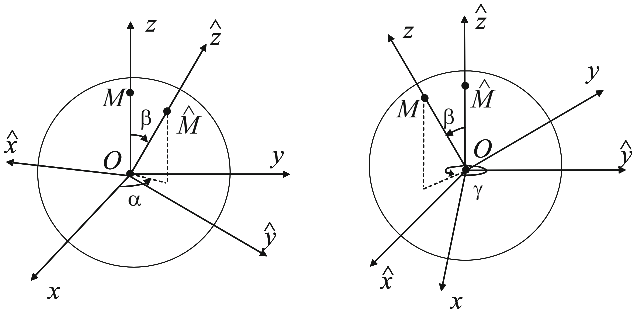


Fig. 1 Rotation angles α, β, and γ defined as spherical angles (β, α) of the rotated z-axis in the original reference frame and sperical angles (β, γ) of the original z-axis in the rotated reference frame

angle γ_E, then angles α, β, γ are simply related to that as

$$\alpha = \alpha_E, \quad \beta = \beta_E, \quad \gamma = \pi - \gamma_E. \tag{10}$$

Note that in [5], the Euler angles were introduced differently ($\alpha = \pi - \alpha_E, \quad \beta = \beta_E, \quad \gamma = \gamma_E$), so formulae obtained via such decomposition should be modified if the present work is to be combined with those relations. Elementary rotation matrices about axes z and y are

$$Q_z(\alpha_E) = \begin{pmatrix} \cos\alpha_E & \sin\alpha_E & 0 \\ -\sin\alpha_E & \cos\alpha_E & 0 \\ 0 & 0 & 1 \end{pmatrix}, \quad Q_y(\beta_E) = \begin{pmatrix} \cos\beta_E & 0 & -\sin\beta_E \\ 0 & 1 & 0 \\ \sin\beta_E & 0 & \cos\beta_E \end{pmatrix}. \tag{11}$$

The standard Euler rotation matrix decomposition $Q = Q_z(\gamma_E) Q_y(\beta_E) Q_z(\alpha_E)$ turns to

$$Q = Q_z(\pi - \gamma) Q_y(\beta) Q_z(\alpha). \tag{12}$$

More symmetric forms with respect to angle β can be obtained, if we introduce elementary matrices A and B as follows

$$A(\gamma) = \begin{pmatrix} \sin\gamma & \cos\gamma & 0 \\ -\cos\gamma & \sin\gamma & 0 \\ 0 & 0 & 1 \end{pmatrix} = Q_z\left(\frac{\pi}{2} - \gamma\right), \tag{13}$$

$$B(\beta) = \begin{pmatrix} -1 & 0 & 0 \\ 0 & -\cos\beta & \sin\beta \\ 0 & \sin\beta & \cos\beta \end{pmatrix} = Q_z\left(\frac{\pi}{2}\right) Q_y(\beta) Q_z\left(\frac{\pi}{2}\right),$$

which results in decomposition

$$Q(\alpha, \beta, \gamma) = A(\gamma) B(\beta) A^T(\alpha). \tag{14}$$

The property of rotation transform is that any subspace of degree n is transformed independently of a subspace of a different degree. Also rotation around the z axis results in diagonal rotation transform operator, as it is seen from the definition provided by Eq. (3). So, the rotation transform can be described as

$$\widehat{C}_n^{m'} = \sum_{m=-n}^{n} T_n^{m'm}(\alpha,\beta,\gamma)\,C_n^m, \quad T_n^{m'm} = e^{-im'\gamma}H_n^{m'm}(\beta)\,e^{im\alpha}, \tag{15}$$

where $\widehat{C}_n^{m'}$ are the expansion coefficients of function f over the spherical harmonic basis in the rotated reference frame,

$$f(\theta,\varphi) = \sum_{n=0}^{p-1}\sum_{m=-n}^{n} C_n^m Y_n^m(\theta,\varphi) = \sum_{n=0}^{p-1}\sum_{m'=-n}^{n} \widehat{C}_n^{m'} Y_n^{m'}\left(\widehat{\theta},\widehat{\varphi}\right) = \widehat{f}\left(\widehat{\theta},\widehat{\varphi}\right). \tag{16}$$

Particularly, if $C_n^m = \delta_{\nu m}$, where $\delta_{\nu m}$ is Kronecker's delta, we have from Eqs (15) and (16) for each subspace

$$Y_n^{\nu}(\theta,\varphi) = e^{i\nu\alpha}\sum_{m'=-n}^{n} H_n^{m'\nu}(\beta)\,e^{-im'\gamma}Y_n^{m'}\left(\widehat{\theta},\widehat{\varphi}\right), \quad \nu = -n,\dots,n. \tag{17}$$

It should be mentioned then that the matrix with elements $T_n^{m'm}(\alpha,\beta,\gamma)$, which we denote as $\mathbf{Rot}(Q(\alpha,\beta,\gamma))$ and its invariant subspaces as $\mathbf{Rot}_n(Q(\alpha,\beta,\gamma)$, is the Wigner D-matrix in an irreducible representation of the group of rigid body rotations SO(3) [18], [19] (with slight modifications presented below). Particularly, we have decompositions

$$\begin{aligned}\mathbf{Rot}_n(Q(\alpha,\beta,\gamma)) &= \mathbf{Rot}_n(Q_z(\pi-\gamma))\,\mathbf{Rot}_n(Q_y(\beta))\,\mathbf{Rot}_n(Q_z(\alpha)) \\ &= \mathbf{Rot}_n(A(\gamma))\,\mathbf{Rot}_n(B(\beta))\,\mathbf{Rot}_n(A(-\alpha)).\end{aligned} \tag{18}$$

Since $Q_y(0)$ and $Q_z(0)$ are identity matrices (see Eq. (11)), then corresponding matrices $\mathbf{Rot}(Q_y(0))$ and $\mathbf{Rot}(Q_y(0))$ are also identity matrices. So, from Eqs (18) and (15) we obtain

$$\begin{aligned}\mathbf{Rot}_n(Q_z(\alpha)) &= \mathbf{Rot}_n(Q(\alpha,0,\pi)) = \left\{(-1)^{m'}H_n^{m'm}(0)\,e^{im\alpha}\right\}, \\ \mathbf{Rot}_n(Q_y(\beta)) &= \mathbf{Rot}_n(Q(0,\beta,\pi)) = \left\{(-1)^{m'}H_n^{m'm}(\beta)\right\},\end{aligned} \tag{19}$$

where in the figure brackets we show the respective elements of the matrices. Since $\mathbf{Rot}_n(Q_y(0))$ is the identity matrix, the latter equation provides

$$H_n^{m'm}(0) = (-1)^{m'}\,\delta_{m'm}. \tag{20}$$

Being used in the first equation (19) this results in

$$\mathbf{Rot}_n(Q_z(\alpha)) = \mathbf{Rot}_n(Q(\alpha,0,\pi)) = \left\{e^{im\alpha}\delta_{m'm}\right\}, \tag{21}$$

which shows that

$$\mathbf{Rot}_n\left(A\left(\gamma\right)\right) = \mathbf{Rot}_n\left(Q_z\left(\frac{\pi}{2}-\gamma\right)\right) = \left\{e^{im\pi/2}e^{-im\gamma}\delta_{m'm}\right\}. \tag{22}$$

We also have immediately from Eq. (15)

$$\mathbf{Rot}_n\left(B\left(\beta\right)\right) = \mathbf{Rot}_n\left(Q\left(0,\beta,0\right)\right) = \left\{H_n^{m'm}\left(\beta\right)\right\}. \tag{23}$$

The rotation coefficients $H_n^{m'm}\left(\beta\right)$ are real and are simply related to the Wigner's (small) d-matrix elements ($d_n^{m'm}\left(\beta\right)$, elsewhere, e.g. [20]),

$$\begin{gathered} d_n^{m'm}\left(\beta\right) = \left(-1\right)^{m'-m}\rho_n^{m'm}\times \\ \sum_{\sigma=\max(0,-(m'-m))}^{\min\left(n-m',n+m\right)} \frac{\left(-1\right)^{\sigma}\cos^{2n-2\sigma+m-m'}\frac{1}{2}\beta\sin^{2\sigma+m'-m}\frac{1}{2}\beta}{\sigma!\left(n+m-\sigma\right)!\left(n-m'-\sigma\right)!\left(m'-m+\sigma\right)!}, \end{gathered} \tag{24}$$

but slightly different due to the difference in definition of spherical harmonics and rotation matrix. In this expression we defined $\rho_n^{m'm}$ as

$$\rho_n^{m'm} = \left[(n+m)!(n-m)!(n+m')!(n-m')!\right]^{1/2}. \tag{25}$$

Explicit expression for coefficients $H_n^{m'm}\left(\beta\right)$ can be obtained from Wigner's formula,

$$H_n^{m'm}\left(\beta\right) = \varepsilon_{m'}\varepsilon_m\rho_n^{m'm}\sum_{\sigma=\max(0,-(m'+m))}^{\min\left(n-m',n-m\right)}\left(-1\right)^{n-\sigma}h_n^{m'm\sigma}\left(\beta\right), \tag{26}$$

where

$$h_n^{m'm\sigma}\left(\beta\right) = \frac{\cos^{2\sigma+m+m'}\frac{1}{2}\beta\sin^{2n-2\sigma-m-m'}\frac{1}{2}\beta}{\sigma!\left(n-m'-\sigma\right)!\left(n-m-\sigma\right)!\left(m'+m+\sigma\right)!}, \tag{27}$$

and symbol $\varepsilon_{m'}$ is defined by Eq. (7). Note that summation limits in Eq. (26) can look a bit complicated, but this can be avoided, if we simply define $1/(-n)! = 0$ for $n = 1,2,...$ (which is consistent with the limit of $1/\Gamma(-n)$ for $n = 0,1,...$, where Γ is the gamma-function), so

$$H_n^{m'm}\left(\beta\right) = \varepsilon_{m'}\varepsilon_m\rho_n^{m'm}\sum_{\sigma=-\infty}^{\infty}\left(-1\right)^{n-\sigma}h_n^{m'm\sigma}\left(\beta\right). \tag{28}$$

The Wigner's (small) d-matrix elements are related to $H_n^{m'm}\left(\beta\right)$ coefficients as

$$d_n^{m'm}\left(\beta\right) = \varepsilon_{m'}\varepsilon_{-m}H_n^{m'm}\left(\beta\right), \tag{29}$$

which can be checked directly using Eqs (24) and (26), and symmetries (30).

2.3 Symmetries

There are several symmetries of the rotation coefficients, from which the following are important for the proposed algorithm

$$\begin{aligned} H_n^{m'm}(\beta) &= H_n^{mm'}(\beta), \\ H_n^{m'm}(\beta) &= H_n^{-m',-m}(\beta), \\ H_n^{m'm}(\pi-\beta) &= (-1)^{n+m'+m} H_n^{-m'm}(\beta), \\ H_n^{m'm}(-\beta) &= (-1)^{m'+m} H_n^{m'm}(\beta). \end{aligned} \tag{30}$$

The first symmetry follows trivially from Eq. (26), which is symmetric with respect to m' and m.

The second symmetry also can be proved using Eq. (26) or its analog (28). This can be checked straightforward using $\varepsilon_{-m'} = (-1)^{m'}\varepsilon_{m'}$ and replacement $\sigma = \sigma' - m' - m$ in the sum. The third symmetry (30) can be also obtained from Eq. (28) using $\sin\frac{1}{2}(\pi-\beta) = \cos\frac{1}{2}\beta$, $\varepsilon_{-m'} = (-1)^{m'}\varepsilon_{m'}$ and replacement $\sigma = n - \sigma' - m$ in the sum. The fourth symmetry follows simply from Eq. (28). It is not needed for $\beta \in [0,\pi]$ but we list it here for completeness, as full period change of β sometimes may be needed.

2.4 Particular Values

For some values of m', m, and n coefficients $H_n^{m'm}$ can be computed with minimal cost, which does not require summation (26). For example, the addition theorem for spherical harmonics can be written in the form

$$P_n(\cos\theta) = \frac{4\pi}{2n+1}\sum_{m'=-n}^{n} Y_n^{-m'}(\theta_2,\varphi_2)\, Y_n^{m'}(\theta_1,\varphi_1), \tag{31}$$

where θ is the angle between points on a unit sphere with spherical coordinates (θ_1,φ_1) and (θ_2,φ_2). From definition of rotation angles α,β,γ, we can see then that if $(\theta_1,\varphi_1) = \left(\widehat{\theta},\widehat{\varphi}\right)$ are coordinates of the point in the rotated reference frame, which coordinates in the original reference frame are (θ,φ) then $(\theta_2,\varphi_2) = (\beta,\gamma)$, since the scalar product of radius-vectors pointed to $\left(\widehat{\theta},\widehat{\varphi}\right)$ and (β,γ) (the $\mathbf{i}_z$ in the rotated frame, Eq. (8)) will be $\cos\theta$ (the z-coordinate of the same point in the original reference frame). Comparing this with Eq. (17) for $\nu = 0$ and using definition of spherical harmonics (3) and (4), we obtain

$$H_n^{m'0}(\beta) = (-1)^{m'}\sqrt{\frac{(n-|m'|)!}{(n+|m'|)!}}P_n^{|m'|}(\cos\beta), \tag{32}$$

since the relation is valid for arbitrary point on the sphere. Note that computation of normalized associated Legendre function can be done using well-known stable recursions and standard library routines are available. This results in $O\left(p^2\right)$ cost of computation of all $H_n^{m'0}(\beta)$ for $n = 0,...,p-1$, $m' = -n,...,n$.

Another set of easily and accurately computable values comes from Wigner's formula (26), where the sum reduces to a single term ($\sigma = 0$) at $n = m$. We have in this case

$$H_n^{m'n}(\beta) = \varepsilon_{m'}\left[\frac{(2n)!}{(n-m')!\,(n+m')!}\right]^{1/2} \cos^{n+m'}\frac{1}{2}\beta \sin^{n-m'}\frac{1}{2}\beta. \tag{33}$$

2.5 Axis Flip Transform

There exist more expressions for $H_n^{m'm}(\beta)$ via finite sums and the present algorithm uses one of those. This relation can be obtained from consideration of the axis flip transform. A composition of rotations which puts axis y in position of axis z and then performs rotation about the z axis, which is described by diagonal matrix followed by the inverse transform is well-known and used in some algorithms. The flip transform can be described by the following formula

$$Q_y(\beta) = Q_z\left(-\frac{\pi}{2}\right) Q_y\left(-\frac{\pi}{2}\right) Q_z(\beta)\, Q_y\left(\frac{\pi}{2}\right) Q_z\left(\frac{\pi}{2}\right), \tag{34}$$

which meaning is rather obvious from geometry.

From Eq. (13) we have

$$\begin{aligned} Q_y\left(\frac{\pi}{2}\right) &= Q_z\left(-\frac{\pi}{2}\right) B\left(\frac{\pi}{2}\right) Q_z\left(-\frac{\pi}{2}\right), \\ Q_y\left(-\frac{\pi}{2}\right) &= Q_y^T\left(\frac{\pi}{2}\right) = Q_z\left(\frac{\pi}{2}\right) B\left(\frac{\pi}{2}\right) Q_z\left(\frac{\pi}{2}\right). \end{aligned} \tag{35}$$

Using the same equation one can express $Q_y(\beta)$ via $B(\beta)$ and obtain from Eqs (34) and (35)

$$B(\beta) = Q_z\left(\frac{\pi}{2}\right) B\left(\frac{\pi}{2}\right) Q_z(\beta)\, B\left(\frac{\pi}{2}\right) Q_z\left(\frac{\pi}{2}\right). \tag{36}$$

The representation of this decomposition for each subspace n results in

$$\begin{aligned} \mathbf{Rot}_n(B(\beta)) = {} & \mathbf{Rot}_n\left(Q_z\left(\frac{\pi}{2}\right)\right) \mathbf{Rot}_n\left(B\left(\frac{\pi}{2}\right)\right) \times \\ & \mathbf{Rot}_n(Q_z(\beta))\, \mathbf{Rot}_n\left(B\left(\frac{\pi}{2}\right)\right) \mathbf{Rot}_n\left(Q_z\left(\frac{\pi}{2}\right)\right). \end{aligned} \tag{37}$$

Using expressions (21) and (23), we can rewrite this relation in terms of matrix elements

$$H_n^{m'm}(\beta) = \sum_{\nu=-n}^{n} e^{im'\pi/2} H_n^{m'\nu}\left(\frac{\pi}{2}\right) e^{i\nu\beta} H_n^{\nu m}\left(\frac{\pi}{2}\right) e^{im\pi/2}. \tag{38}$$

Since $H_n^{m'm}(\beta)$ is real, we can take the real part of the right hand side of this relation, to obtain

$$H_n^{m'm}(\beta) = \sum_{\nu=-n}^{n} H_n^{m'\nu}\left(\frac{\pi}{2}\right) H_n^{m\nu}\left(\frac{\pi}{2}\right) \cos\left(\nu\beta + \frac{\pi}{2}\left(m'+m\right)\right), \tag{39}$$

where we used the first symmetry (30). Note also that the third symmetry (30) applied to $H_n^{m'm}(\pi/2)$ results in

$$\begin{aligned} H_n^{m'm}(\beta) = {} & H_n^{m'0}\left(\frac{\pi}{2}\right) H_n^{m0}\left(\frac{\pi}{2}\right) \cos\left(\frac{\pi}{2}\left(m'+m\right)\right) + \\ & 2\sum_{\nu=1}^{n} H_n^{m'\nu}\left(\frac{\pi}{2}\right) H_n^{m\nu}\left(\frac{\pi}{2}\right) \cos\left(\nu\beta + \frac{\pi}{2}\left(m'+m\right)\right). \end{aligned} \tag{40}$$

2.6 Recursions

Several recursions for computation of coefficients $H_n^{m'm}(\beta)$ were derived from the invariancy of differential operator ∇ [13], including the following one, which was suggested for computing all $H_n^{m'm}(\beta)$

$$\begin{aligned} b_n^m H_{n-1}^{m',m+1} = {} & \frac{1-\cos\beta}{2} b_n^{-m'-1} H_n^{m'+1,m} - \\ & \frac{1+\cos\beta}{2} b_n^{m'-1} H_n^{m'-1,m} - \sin\beta\, a_{n-1}^{m'} H_n^{m'm}, \end{aligned} \tag{41}$$

where $n = 2,3,\ldots, \quad m' = -n+1,\ldots,n-1, \quad m = 0,\ldots,n-2$, and

$$\begin{aligned} a_n^m &= a_n^{-m} = \sqrt{\frac{(n+1+m)(n+1-m)}{(2n+1)(2n+3)}}, \quad n \geqslant |m|. \\ a_n^m &= 0, \quad n < |m|, \end{aligned} \tag{42}$$

$$b_n^m = \operatorname{sgn}(m)\sqrt{\frac{(n-m-1)(n-m)}{(2n-1)(2n+1)}}, \quad n \geqslant |m| \qquad n < |m|, \tag{43}$$

$$\operatorname{sgn}(m) = \begin{cases} 1, & m \geqslant 0 \\ -1, & m < 0 \end{cases}. \tag{44}$$

This recursion allows one to get $\left\{H_n^{m',m+1}\right\}$ from $\left\{H_n^{m'm}\right\}$. Once the value for $m = 0$, $H_n^{m'0}$, is known from Eq. (32) for any n and m' all coefficients can be computed. The cost of the procedure is $O\left(p^3\right)$ as n is limited by $p-1$ (to get that $H_n^{m'0}$ should be computed up to $n = 2p-2$). The relation was extensively tested and used in the FMM (e.g. [21]), however, it showed numerical instability $p \gtrsim 100$.

From the commutativity of rotations around the axis y,

$$Q_y(\beta)\,Q_y'(0) = Q_y'(0)\,Q_y(\beta), \quad Q_y'(\beta) = \frac{dQ_y}{d\beta}. \tag{45}$$

one may derive a recurrence relation [5]. For the nth invariant rotation transform subspace, the relation (45) gives

$$\mathbf{Rot}_n\left(Q_y(\beta)\right)\mathbf{Rot}_n\left(Q_y'(0)\right) = \mathbf{Rot}_n\left(Q_y'(0)\right)\mathbf{Rot}_n\left(Q_y(\beta)\right). \tag{46}$$

Differentiating the r.h.s. of (28) w.r.t β and evaluating at $\beta = 0$ we get

$$\left.\frac{dH_n^{m'm}(\beta)}{d\beta}\right|_{\beta=0} = c_n^{m'-1}\delta_{m,m'-1} + c_n^{m'}\delta_{m,m'+1}, \tag{47}$$

$$c_n^m = \frac{1}{2}(-1)^m \operatorname{sgn}(m)\left[(n-m)(n+m+1)\right]^{1/2}, \quad m = -n-1,...,n. \tag{48}$$

Using (19), relation (46) can be rewritten as

$$\sum_{\nu=-n}^{n} H_n^{m'\nu}(\beta)(-1)^\nu \left.\frac{dH_n^{\nu m}(\beta)}{d\beta}\right|_{\beta=0} = \sum_{\nu=-n}^{n} \left.\frac{dH_n^{m'\nu}(\beta)}{d\beta}\right|_{\beta=0}(-1)^\nu H_n^{\nu m}(\beta). \tag{49}$$

Using Eq. (47) we obtain the recurrence relation for $H_n^{m',m}(\beta)$

$$d_n^{m-1}H_n^{m',m-1} - d_n^m H_n^{m',m+1} = d_n^{m'-1}H_n^{m'-1,m} - d_n^{m'}H_n^{m'+1,m}, \tag{50}$$

$$d_n^m = \frac{\operatorname{sgn}(m)}{2}\left[(n-m)(n+m+1)\right]^{1/2}, \quad m = -n-1,...,n. \tag{51}$$

This recurrence is used as the basis of the stable algorithm obtained in this paper. In contrast to the recurrence (41), (50) relates values of rotation coefficients $H_n^{m'm}$ *within the same subspace n*. Thus, if boundary values for a subspace are provided, all other coefficients can be found.

3 Bounds for Rotation Coefficients

It should be noticed that the $(2n+1)\times(2n+1)$ matix $\mathbf{H}_n(\beta) = \mathbf{Rot}_n(B(\beta))$ with entries $H_n^{m'm}(\beta)$, $m', m = -n,...,n$ is real, unitary, and self-adjoint (Hermitian). This follows from

$$[\mathbf{H}_n(\beta)]^2 = \mathbf{I}_n, \quad \mathbf{H}_n(\beta) = \mathbf{H}_n^T(\beta), \tag{52}$$

where $\mathbf{I}_n$ is $(2n+1) \times (2n+1)$ identity matrix. Particularly, this shows that the norm of matrix $\mathbf{H}_n(\beta)$ is unity and

$$\left| H_n^{m'm}(\beta) \right| \leqslant 1, \quad n = 0, 1, \ldots, \quad m', m = -n, \ldots, n. \tag{53}$$

While bound (53) is applicable for any values of β, n, m and m', we note that in certain regions of the parameter space it can be improved.

3.1 General Bound

To get such a bound we note that due to the third symmetry (30) it is sufficient to consider β in range $0 \leqslant \beta \leqslant \pi/2$. The first and the second symmetries (30) provide that only nonnegative m can be considered, $m \geqslant 0$, and also m' from the range $|m'| \leqslant m$. The latter provides $m + m' \geqslant 0$, $m' \leqslant m$, and $n - m' \geqslant n - m$. Hence in this range Eq. (26) can be written in the form

$$H_n^{m'm}(\beta) = \varepsilon_{m'} \varepsilon_m \rho_n^{m'm} \sum_{\sigma=0}^{n-m} (-1)^{n-\sigma} h_n^{m'm\sigma}(\beta), \tag{54}$$

where $h_n^{m'm\sigma}(\beta) \geqslant 0$, and we can bound $\left| H_n^{m'm}(\beta) \right|$ as

$$\begin{aligned} \left| H_n^{m'm}(\beta) \right| &\leqslant \rho_n^{m'm} \sum_{\sigma=0}^{n-m} h_n^{m'm\sigma}(\beta) \\ &\leqslant \rho_n^{m'm} (n-m+1) h_n^{m'ms}(\beta), \end{aligned} \tag{55}$$

where

$$h_n^{m'ms}(\beta) = \max_{0 \leqslant \sigma \leqslant n-m} h_n^{m'm\sigma}(\beta). \tag{56}$$

In other words s is the value of σ at which $h_n^{m'm\sigma}(\beta)$ achieves its maximum at given β, n, m and m'.

Note then that for $n = m$ we have only one term, $\sigma = 0$, the sum (54), so $s = 0$ and in this case $\left| H_n^{m'm}(\beta) \right|$ reaches the bound (55), as it also follows from Eqs (27) and (33). Another simple case is realized for $\beta = 0$. In this case, again, the sum has only one nonzero term at $\sigma = n - m$ and

$$h_n^{m'm,n-m}(0) = \frac{\delta_{m'm}}{(n-m)!(n+m)!}. \tag{57}$$

While exact value in this case is $\left|H_n^{m'm}(0)\right| = \delta_{m'm}$ (see also Eq. (20)), bound (55) at $s = n-m$ provides $\left|H_n^{m'm}(\beta)\right| \leqslant (n-m+1)\,\delta_{m'm}$, which is correct, but is not tight.

To find the maximum of $h_n^{m'm\sigma}(\beta)$ for $\beta \neq 0$ and $n-m > 0$ (so, also $n-m' > 0$) we can consider the ratio of the consequent terms in the sum

$$r_n^{m'm\sigma}(\beta) = \frac{h_n^{m'm,\sigma+1}(\beta)}{h_n^{m'm\sigma}(\beta)} = \frac{(n-m'-\sigma)(n-m-\sigma)}{t^2(\sigma+1)(m'+m+\sigma+1)}, \quad t = \tan\frac{\beta}{2}, \tag{58}$$

where parameter t is varying in the range $0 < t \leqslant 1$, as we consider $0 < \beta \leqslant \pi/2$. This ratio considered as a function of σ monotonously decay from its value at $\sigma = 0$ to zero at $\sigma = n-m$ (as the numerator is a decaying function, while the denominator is a growing function ($n-m' \geqslant n-m$). So, if $r_n^{m'm0}(\beta) \leqslant 1$ then the maximum of $h_n^{m'm\sigma}(\beta)$ is reached at $\sigma = 0$, otherwise it can be found from simultaneous equations $r_n^{m'ms}(\beta) > 1$, $r_n^{m'm,s+1}(\beta) \leqslant 1$. To treat both cases, we consider roots of equation $r_n^{m'm\sigma}(\beta) = 1$, which turns to a quadratic equation with respect to σ,

$$\left(n-m'-\sigma\right)(n-m-\sigma) = t^2(\sigma+1)\left(m'+m+\sigma+1\right). \tag{59}$$

If there is no real nonnegative roots in range $0 \leqslant \sigma \leqslant n-m$ then $s = 0$, otherwise s should be the integer part of the smallest root of Eq. (59), as there may exist only one root of equation $r_n^{m'm\sigma}(\beta) = 1$ in range $0 \leqslant \sigma \leqslant n-m$ (the largest root in this case is at $\sigma > n-m' \geqslant n-m$). So, we have

$$\begin{aligned}
\sigma_1 &= \frac{2n+2t^2-\left(1-t^2\right)(m+m')-\sqrt{D}}{2\left(1-t^2\right)}, \\
D &= \left(m-m'\right)^2 + t^2\left(2\left(2n^2-m^2-(m')^2\right)+t^2(m+m')^2+4(2n+1)\right).
\end{aligned} \tag{60}$$

This shows that $D \geqslant 0$, so the root is anyway real. Note also that $t = 1$ $(\beta = \pi/2)$ is a special case, since at this value equation (59) degenerates to a linear equation, which has root

$$\sigma_1 = \frac{(n-m')(n-m)-(m+m'+1)}{2(n+1)}, \quad (t=1). \tag{61}$$

Summarizing, we obtain the following expression for s

$$s = \begin{cases} [\sigma_1], & \sigma_1 \geqslant 0 \\ 0, & \sigma_1 < 0, \end{cases} \tag{62}$$

where [] denotes the integer part, $\sigma_1(n,m,m',t)$ for $t \neq 1$ is provided by Eq. (60), and its limiting value at $t = 1$ is given by Eq. (61). Equations (55) and (27) then yield

$$\left|H_n^{m'm}(\beta)\right| \leqslant \frac{\rho_n^{m'm}(n-m+1)\cos^{2s+m+m'}\frac{\beta}{2}\sin^{2n-2s-m-m'}\frac{\beta}{2}}{s!\,(n-m'-s)!\,(n-m-s)!\,(m+m'+s)!}. \tag{63}$$

3.2 Asymptotic Behavior

The bounds for the magnitude of the rotation coefficients are important for study of their behavior at large n, as at large enough n recursions demonstrate instabilities, while direct computations using the sums becomes difficult due to factorials of large numbers. So, we are going to obtain asymptotics of expression (63) for $\left|H_n^{m'm}(\beta)\right|$ at $n \to \infty$. We note then that for this purpose, we consider scaling of parameters m, m', and s, i.e. we introduce new variables

$$\mu = \frac{m}{n}, \quad \mu' = \frac{m'}{n}, \quad \xi = \frac{s}{n}, \tag{64}$$

which, as follows from the above consideration, are in the range $0 \leqslant \mu \leqslant 1$, $-\mu \leqslant \mu' \leqslant \mu$, $0 \leqslant \xi \leqslant 1-\mu$. The asymptotics can be constructed assuming that these parameters are fixed, while $n \to \infty$.

Note now that Eq. (63) can be written in the form

$$\left|H_n^{m'm}(\beta)\right| \leqslant (n-m+1)\left[C_{n-m}^{s} C_{n-m'}^{s} C_{n+m}^{m+m'+s} C_{n+m'}^{m+m'+s}\right]^{1/2} \times \\ \cos^{2s+m+m'}\frac{1}{2}\beta \sin^{2n-2s-m-m'}\frac{1}{2}\beta, \tag{65}$$

where

$$C_q^l = \frac{q!}{l!\,(q-l)!}, \tag{66}$$

are the binomial coefficients. Consider asymptotics of C_{an}^{bn}, where a and b are fixed, $0 < b < a$, and $n \to \infty$. Using the inequality, valid for $x > 0$, (e.g. see [22])

$$\sqrt{2\pi}\exp\left(\frac{2x+1}{2}\ln x - x\right) < x! < \sqrt{2\pi}\exp\left(\frac{2x+1}{2}\ln x - x + \frac{1}{12x}\right). \tag{67}$$

We can find that

$$C_{an}^{bn} < C_{ab} n^{-1/2} e^{\lambda_{ab} n}, \tag{68}$$

where

$$\lambda_{ab} = a\ln a - b\ln b - (a-b)\ln(a-b), \tag{69}$$

and the constant C_{ab} is bounded as

$$C_{ab} \leqslant \sqrt{\frac{a}{2\pi b(a-b)}}\,e^{1/12}. \tag{70}$$

Using this bound in Eq. (65) and definition (64), we obtain

$$\left|H_n^{m'm}(\beta)\right| \leqslant C e^{\lambda n}, \tag{71}$$

where C is some constant depending on μ, μ',and ξ, while for λ we have the following expression.

$$\begin{aligned}\lambda = &\frac{1-\mu}{2}\ln(1-\mu)+\frac{1-\mu'}{2}\ln\left(1-\mu'\right)+\frac{1+\mu}{2}\ln(1+\mu)+ \\ &\frac{1+\mu'}{2}\ln\left(1+\mu'\right)-\xi\ln\xi-\left(\mu+\mu'+\xi\right)\ln\left(\mu+\mu'+\xi\right)- \\ &(1-\mu-\xi)\ln(1-\mu-\xi)-\left(1-\mu'-\xi\right)\ln\left(1-\mu'-\xi\right)+ \\ &\left(\mu+\mu'+2\xi\right)\ln\cos\frac{1}{2}\beta+\left(2-\mu-\mu'-2\xi\right)\ln\sin\frac{1}{2}\beta.\end{aligned} \tag{72}$$

Relation between ξ and other parameters for $n \to \infty$ follows from Eqs (60)–(62),

$$\begin{aligned}\xi &= \frac{2-\left(1-t^2\right)\left(\mu+\mu'\right)-\sqrt{\Delta}}{2\left(1-t^2\right)}+O\left(n^{-1}\right), \quad t \neq 1, \\ \xi &= \frac{1}{2}\left(1-\mu'\right)(1-\mu)+O\left(n^{-1}\right), \quad t=1,\end{aligned} \tag{73}$$

where the discriminant can be written in the form

$$\Delta = \left(2-\left(1-t^2\right)\left(\mu+\mu'\right)\right)^2-4\left(1-t^2\right)(1-\mu)\left(1-\mu'\right). \tag{74}$$

Since $\Delta \geqslant 0$ and $4\left(1-t^2\right)(1-\mu)\left(1-\mu'\right) \geqslant 0$ this results that the principal term $\xi \geqslant 0$. The residual $O\left(n^{-1}\right)$ does not affect bound as the asymptotic constant can be corrected, while the principal term can be used in λ in Eq. (72). So, λ then is a function of three parameters, $\lambda = \lambda\left(\mu, \mu'; \beta\right)$.

Bound (71) is tighter than (53) when $\lambda < 0$, as it shows that for $n \to \infty$ the rotation coefficients in parameter region $\lambda\left(\mu, \mu', \beta\right) < 0$ become exponentially small. This region of exponentially small $\left|H_n^{m'm}(\beta)\right|$ at fixed β is bounded by curve

$$\lambda\left(\mu, \mu'; \beta\right) = 0. \tag{75}$$

In Fig. 2 computations of log $\left|H_n^{m'm}(\beta)\right|$ at different β and large enough n $(n = 100)$ are shown. Here also the boundary curve (75) is plotted (the curve is extended by symmetry for all β and μ and μ', so it becomes a closed curve). It is seen, that it agrees with the computations and indeed, $\left|H_n^{m'm}(\beta)\right|$ decays exponentially.

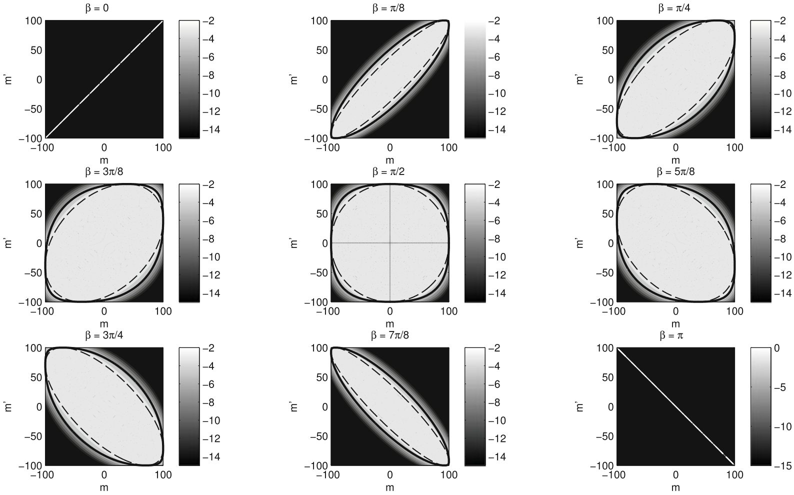


Fig. 2 Magnitude of rotation coefficients, $\log_{10}\left|H_n^{m'm}(\beta)\right|$ at $n = 100$ and different β. The solid *bold curves* show analytical bounds of exponentially decaying region determined by Eq. (75). The *dashed curves* plot the ellipse, Eq. (92), obtained from asymptotic analysis of recursions for large n

4 Asymptotic Behavior of Recursion

Consider now asymptotic behavior of recursion (50), where coefficients of recursion d_n^m do not depend on β. First, we note that this relation can be written in the form

$$k_n^{m'}\left(H_n^{m'+1,m} - H_n^{m'-1,m}\right) - l_n^{m'}\left(H_n^{m'+1,m} + H_n^{m'-1,m}\right) = \\ k_n^{m}\left(H_n^{m',m+1} - H_n^{m',m-1}\right) - l_n^{m}\left(H_n^{m',m-1} + H_n^{m',m+1}\right), \tag{76}$$

where

$$k_n^m = \frac{1}{2}\left(d_n^{m-1} + d_n^m\right), \quad l_n^m = \frac{1}{2}\left(d_n^{m-1} - d_n^m\right). \tag{77}$$

At $m \neq 0$ and large n and $n - |m|$, $m = n\mu$, asymptotics of coefficients d_n^m can be obtained from Eq. (51),

$$k_n^m = \operatorname{sgn}(\mu)\left(1-\mu^2\right)^{1/2}\frac{1}{2n^{-1}}\left[1+O(n^{-1})\right], \\ l_n^m = \frac{1}{4}\operatorname{sgn}(\mu)\frac{\mu}{\left(1-\mu^2\right)^{1/2}}\left[1+O(n^{-1})\right]. \tag{78}$$

Hence, asymptotically relation (76) turns into

$$\frac{(1-\mu'^2)^{1/2}\left(H_n^{m'+1,m}-H_n^{m'-1,m}\right)}{2h}-\frac{\mu'\left(H_n^{m'+1,m}+H_n^{m'-1,m}\right)}{4\left(1-\mu'^2\right)^{1/2}} \tag{79}$$
$$=\frac{\operatorname{sgn}(\mu)}{\operatorname{sgn}(\mu')}\left[\frac{(1-\mu^2)^{1/2}\left(H_n^{m',m+1}-H_n^{m',m-1}\right)}{2h}-\frac{\mu\left(H_n^{m',m+1}+H_n^{m',m-1}\right)}{4\left(1-\mu^2\right)^{1/2}}\right],$$

where $h=1/n$. Let us interpret now $H_n^{m'm}$ as samples of differentiable function $H_n(\mu',\mu)$ on a $(2n+1)\times(2n+1)$ grid on the square $(\mu',\mu)\in[-1,1]\times[-1,1]$ with step h in each direction, $H_n(m'/n,m/n)=H_n^{m'm}$. In this case relation (79) corresponds to a central difference scheme for the hyperbolic PDE

$$\operatorname{sgn}(\mu)\left(1-\mu^2\right)^{1/2}\frac{\partial H_n}{\partial\mu}-\operatorname{sgn}\left(\mu'\right)\left(1-\mu'^2\right)^{1/2}\frac{\partial H_n}{\partial\mu'}- \tag{80}$$
$$\frac{1}{2}\left[\operatorname{sgn}(\mu)\frac{\mu}{\left(1-\mu^2\right)^{1/2}}-\operatorname{sgn}\left(\mu'\right)\frac{\mu'}{\left(1-\mu'^2\right)^{1/2}}\right]H_n=0,$$

where H_n is approximated to $O(h)$ via its values at neighbouring grid points in each direction,

$$H_n\left(\mu',\mu\right)=\frac{1}{2}\left(H_n\left(\mu'-h,\mu\right)+H_n\left(\mu'+h,\mu\right)+O(h)\right)=$$
$$\frac{1}{2}\left(H_n\left(\mu',\mu-h\right)+H_n\left(\mu',\mu+h\right)+O(h)\right). \tag{81}$$

Note then $K_n(\mu',\mu)$ defined as

$$K_n\left(\mu',\mu\right)=\left(1-\mu'^2\right)^{1/4}\left(1-\mu^2\right)^{1/4}H_n\left(\mu',\mu\right), \tag{82}$$

satisfies

$$\operatorname{sgn}(\mu)\left(1-\mu^2\right)^{1/2}\frac{\partial K_n}{\partial\mu}-\operatorname{sgn}\left(\mu'\right)\left(1-\mu'^2\right)^{1/2}\frac{\partial K_n}{\partial\mu'}=0. \tag{83}$$

Let us introduce the variables $\psi=\arcsin\mu$ and $\psi'=\arcsin\mu'$, and

$$G_n(\psi',\psi)=K_n(\mu',\mu),\quad -\frac{\pi}{2}\leqslant\psi,\psi'\leqslant\frac{\pi}{2}. \tag{84}$$

In this case $\mu=\sin\psi$, $(1-\mu^2)^{1/2}=\cos\psi$ and similarly, $\mu'=\sin\psi'$, $(1-\mu'^2)^{1/2}=\cos\psi'$. So in these variables Eq. (78) turns into

$$\operatorname{sgn}\left(\psi'\right)\frac{\partial G_n}{\partial\psi'}-\operatorname{sgn}(\psi)\frac{\partial G_n}{\partial\psi}=0. \tag{85}$$

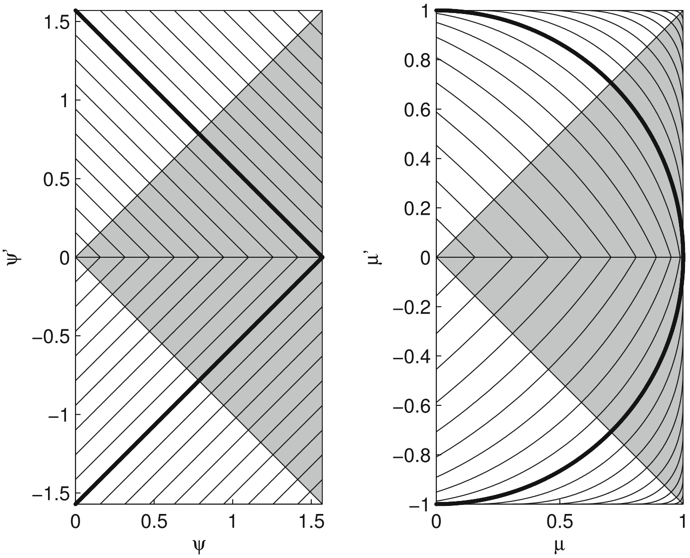


Fig. 3 The characteristics of equations (85) and (80). The shaded area shows the region for which the rotation coefficients should be computed as symmetry relations (30) can be applied to obtain the values in the other regions

Due to symmetry relations (30), we can always constrain the region with $\mu \geqslant 0$ $(0 \leqslant \psi \leqslant \pi/2)$, so $\operatorname{sgn}(\psi) = 1$. So, we have two families of characteristics of this equation in this region,

$$\psi' + \psi = C^+, \quad 0 \leqslant \psi' \leqslant \frac{\pi}{2}; \quad \psi' - \psi = C^-, \quad -\frac{\pi}{2} \leqslant \psi' \leqslant 0, \tag{86}$$

which are also characteristics of equation (80) for $\psi = \arcsin\mu$, $\psi' = \arcsin\mu'$. Figure 3 illustrates these characteristics in the (ψ, ψ') and the (μ, μ') planes. It is interesting to compare these characteristics with the results shown in Fig. 2. One can expect that the curves separating the regions of exponentially small values of the rotation coefficients should follow, at least qualitatively, some of the characteristic curves. Indeed, while $H_n(\mu', \mu)$ according to Eq. (82) can change along the characteristics (the value of $K_n(\mu', \mu)$ should be constant), such a change is rather weak (proportional to $\left(1-\mu'^2\right)^{1/4}\left(1-\mu^2\right)^{1/4}$, which cannot explain the exponential decay). However, we can see that the boundaries of the regions plotted at different β partly coincide with the characteristics (e.g. at $\beta = \pi/4$ this curve qualitatively close to characteristic family C^- at $\mu' < 0$, but qualitatively different from the characteristic family C^+ for $\mu' > 0$; similarly the curve $\beta = 3\pi/4$ coincides with one of characteristics C^+ for $\mu' > 0$, while characteristics of family C^- are rather orthogonal to the curve at $\mu' < 0$). As Eq. (85) should be valid for any β this creates a puzzle, the solution of which can be explained as follows.

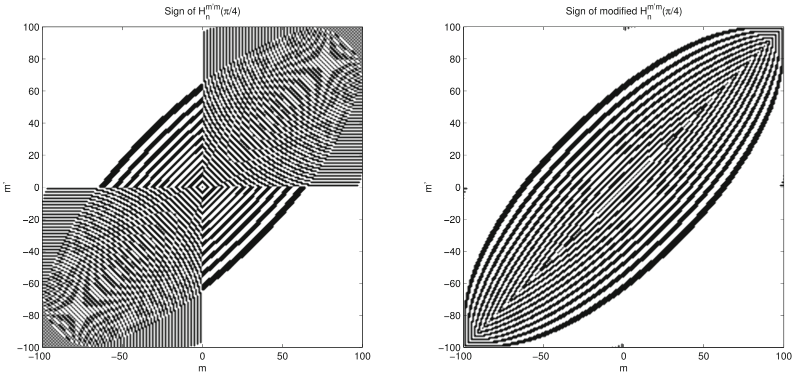


Fig. 4 The sign of coefficients $H_n^{m'm}$ and modified coefficients $\widehat{H}_n^{m'm}$, Eq. (87), at $\beta = \pi/4$. Positive sign (*white*) is assigned to coefficients which magnitude is below 10^{-13}

$H_n^{m'm}(\beta)$ considered as a function of m' oscillates with some local frequency $\omega_n(\mu', \mu, \beta)$. In the regions where this frequency is small, the transition from the discrete relation (79) to the PDE (80) is justified. However at frequencies $\omega_n(\mu', \mu, \beta) \sim 1$ the PDE is not valid. Such regions exist, e.g., if $H_n^{m'm}(\beta) = (-1)^{m'} \widehat{H}_n^{m'm}(\beta)$, where $\widehat{H}_n^{m'm}$ oscillates with a low frequency. For example, Eqs (33) and (7) show that for $0 < \beta < \pi$ the boundary value $H_n^{m'n}(\beta)$ is a smooth function of μ' for $\mu' < 0$, while it cannot be considered as differentiable function of μ' for $\mu' > 0$. On the other hand, this example shows that the function $\widehat{H}_n^{m'n}(\beta)$ has smooth behavior for $\mu' > 0$ and a differential equation can be considered for this function. Equations (76) and (79) show that for function $\widehat{H}_n^{m'm}$ one obtains the same equation, but the sign of $\mathrm{sgn}(\mu')$ should be changed. Particularly, this means that if this is the case then characteristics of the family C^+ can be extended to the region $-\pi/2 \leqslant \psi' \leqslant 0$, while the characteristics of the family C^- can be continued to the region $0 \leqslant \psi' \leqslant \pi/2$. Of course, such an extension must be done carefully, based on the analysis, and this also depends on the values of β, which plays the role of a parameter. Figure 4 illustrates the signs of function $H_n^{m'm}(\beta)$ and

$$\widehat{H}_n^{m'm}(\beta) = \begin{cases} \varepsilon_{m'} \varepsilon_{-m} H_n^{m'm}(\beta), & m < m' \\ \varepsilon_{-m'} \varepsilon_m H_n^{m'm}(\beta), & m \geqslant m' \end{cases}. \tag{87}$$

It is seen that $\widehat{H}_n^{m'm}$ is a “smoother” function of m' and m (for $0 < \beta < \pi/2$; this is not the case for $\pi/2 < \beta < \pi$; for those values we use the third symmetry relation (30). Also note that for $m \geqslant m'$ the function $\widehat{H}_n^{m'm}$ coincides with $d_n^{m'm}$ due to relation (29).

This enables determination of the boundary curve separating oscillatory and exponentially decaying regions of $\widehat{H}_n^{m'm}$. Indeed, at $m' = 0$, $m \geqslant 0$ the boundary value (32) and the first symmetry (30) provides that $\widehat{H}_n^{0m}(\beta)$ is proportional to the associated Legendre function $P_n^m(\cos\beta)$. Function $P_n^m(x)$ satisfies differential equation

$$\left(1-x^2\right)\frac{d^2w}{dx^2}-2x\frac{dw}{dx}+\left[n(n+1)-\frac{m^2}{1-x^2}\right]w=0, \tag{88}$$

which for

$$y(x)=\left(1-x^2\right)^{1/2}w(x), \tag{89}$$

at large n and $m=\mu n$ turns into

$$\frac{d^2y}{dx^2}=n^2q(x)y, \quad q(x)=-\frac{1-x^2-\mu^2}{\left(1-x^2\right)^2}, \tag{90}$$

An accurate asymptotic study can be done based on the Liouville-Green or WKB-approximation [23], while here we limit ourselves with the qualitative observation, that $q(x_\mu)=0$ at $1-x_\mu^2-\mu^2=0$, which is a “turning” point, such that at $\mu^2>1-x_\mu^2$ we have $q(x)>0$ which corresponds to asymptotically growing/decaying regions of y (the decaying solution corresponds to the associated Legendre functions of the first kind, which is our case). Region $\mu^2<1-x_\mu^2$ corresponds to $q(x)<0$ and to the oscillatory region. The vicinity of the turning point can be studied separately (using the Airy functions) [23], but it should be noticed immediately that the local frequency $\omega\sim n\sqrt{-q}$ is much smaller than n at $|q|\ll 1$, so function $\widehat{H}_n^{m'm}$ is relatively smooth on the grid with step $h=n^{-1}$ in the vicinity of this turning point. Hence, for the characteristic C^- passing through the turning point $\mu=\sqrt{1-x_\mu^2}=\sqrt{1-\cos^2\beta}=\sin\beta$ at $\mu'=0$ $(\psi'=0)$

$$C^- = \psi'-\psi\big|_{\psi'=0}=-\psi=-\arcsin\mu=-\beta, \quad 0\leqslant\beta\leqslant\pi/2. \tag{91}$$

(similarly, characteristic C^+ can be considered, which provides the boundary curve for the case $\pi/2\leqslant\beta\leqslant\pi$). We note now that curve $\psi'-\psi=-\beta$ in (μ,μ') space describes a piece of ellipse. Using symmetries (30) we can write equation for this ellipse in the form

$$\frac{(\mu+\mu')^2}{4\cos^2\frac{1}{2}\beta}+\frac{(\mu-\mu')^2}{4\sin^2\frac{1}{2}\beta}=1. \tag{92}$$

The ellipse has semiaxes $\cos\frac{1}{2}\beta$ and $\sin\frac{1}{2}\beta$, which are turned to $\pi/4$ angles in the (μ,μ'). This ellipse is also shown in Fig. 2, and it is seen that it approximates the regions of decay obtained from the analysis of bounds of the rotation coefficients.

More accurate consideration and asymptotic behavior of the rotation coefficients at large n can be obtained using the PDE and the boundary values of the coefficients (32) and (33). Such analysis, however, deserves a separate paper and is not presented here, as the present goal is to provide a qualitative picture and develop a stable numerical procedure.

5 Stability of Recursions

Now we consider stability of recursion (50), which can be used to determine the rotation coefficients at $m \geqslant 0$ and $|m'| \leqslant m$ for $0 \leqslant \beta \leqslant \pi$ as for all other values of m and m' symmetries (30) can be used. This recursion is two dimensional and, in principle, one can resolve it with respect to any of its terms, e.g. $H_n^{m'+1,m}$, and propagate it in the direction of increasing m' if the initial and boundary values are known. Several steps of the recursion can be performed anyway and there is no stability question for relatively small n. However, at large n stability becomes critical, and, so the asymptotic analysis and behavior plays an important role for establishing of stability conditions.

5.1 Courant-Friedrichs-Lewy (CFL) Condition

Without any regard to a finite difference approximation of a PDE recursion (50) can be written in the form (76), which for large n takes the form (79). The principal term of (76) for $n \to \infty$ here can be written as

$$H_n^{m'+1,m} - H_n^{m'-1,m} = c\left(H_n^{m',m+1} - H_n^{m',m-1}\right), \quad c = \frac{k_n^m}{k_n^{m'}}. \tag{93}$$

An analysis of a similar recursion, appearing from the two-wave equation is provided in [16], which can be also applied to the one-wave equation approximated by the central difference scheme. If we treat here m' as an analog of time, m as an analog of a spatial variable, and c as the wave speed (the grid in both variables has the same step $\Delta m = \Delta m' = 1$), then the Courant-Friedrichs-Lewy (CFL) stability condition becomes

$$|c| \leqslant 1. \tag{94}$$

Note now that from definitions (77) and (51) we have

$$k_n^0 = 0, \quad k_n^{-m} = -k_n^m, \quad k_n^m \geqslant k_n^{m+1}, \quad m = 1, ..., n-1. \tag{95}$$

The CFL condition is satisfied for any $m \geqslant |m'|$, $m' \neq 0$. This is also consistent with the asymptotic behavior of the recursion coefficients (78), as

$$c^2 \sim \frac{1-\mu^2}{1-\mu'^2}, \tag{96}$$

which shows that in region $\mu'^2 \leqslant \mu^2$ we have Eq. (94). Note that the CFL condition for the central difference scheme includes only the absolute value of c so, independent of which variable $H_n^{m'+1,m}$ or $H_n^{m'-1,m}$ the recursion (93) is resolved the scheme satisfies the necessary stability condition (CFL). This means that within the region $m \geqslant |m'|$ the scheme can be applied in the forward or backward directions, while some care may be needed for passing the value $m' = 0$.

This analysis shows also that if recursion (93) will be resolved with respect to $H_n^{m',m+1}$ or $H_n^{m',m-1}$ in region $m \geqslant |m'|$ then the recursion will be absolutely unstable, as it does not satisfy the necessary condition, which in this case will be $|1/c| \leqslant 1$ (as the recursion is symmetric with respect to m and m'). Figure 5 (chart on the left) shows the stable and unstable directions of propagation. Under "stable" we mean here conditional or neutral stability, as the CFL criterium is only a necessary, not sufficient, condition.

5.2 Von Neumann Stability Analysis

The von Neumann, or Fourier, stability analysis is a usual tool for investigation of finite difference schemes of linear equations with constant coefficients (original publication [17], various applications can be found elsewhere). Despite the recursion we study is linear; it has variable coefficients. So, we can speculate that only in some region, where such variability can be neglected, and we can perform several recursive steps with quasiconstant coefficients, can such an analysis give us some insight on the overall stability. As the recursion (50) can be written in the form (76), the asymptotic behavior of the recursion coefficients (78) shows that the assumption that these coefficients are quasiconstant indeed is possible in a sense that many grid points can be handled with the same value of coefficients, as they are functions of "slow" variables μ and μ', but regions $\mu \to 1$ and $\mu' \to 1$ are not treatable with this approach, as either the coefficients or their derivatives become unbounded. Hence, we apply the fon Neumann analysis for $1-\mu$ and $1-\mu'$ treated as quantities of the order of the unity.

Equation (79) then can be written in the form

$$H_n^{m'+1,m} - H_n^{m'-1,m} - \frac{a}{n}\left(H_n^{m'+1,m} + H_n^{m'-1,m}\right) = \tag{97}$$
$$c\left(H_n^{m',m+1} - H_n^{m',m-1}\right) - \frac{b}{n}\left(H_n^{m',m+1} + H_n^{m',m-1}\right),$$

where a, b, and c are coefficients depending on μ and μ' and, formally, not depending on m and m' (separation to "slow" and "fast" variables typical for multiscale analysis), for $\mu \geqslant 0$

$$a = \frac{\mu'}{2\left(1-\mu'^2\right)}, \quad b = \frac{\operatorname{sgn}\left(\mu'\right)\mu}{2\left(1-\mu'^2\right)^{1/2}\left(1-\mu^2\right)^{1/2}}, \quad c = \operatorname{sgn}\left(\mu'\right)\left(\frac{1-\mu^2}{1-\mu'^2}\right)^{1/2}. \tag{98}$$

Now we consider perturbation $\eta_n^{m',m}$ of coefficients $H_n^{m',m}$ and their propagation within the conditionally stable scheme above. As the true values of $H_n^{m',m}$ satisfy Eq. (97), the perturbation satisfies the same equation. Let $\widehat{\eta}_n^{m'}(k)$ be the Fourier transform of $\eta_n^{m',m}$ with respect to m at layer m', where k is the wavenumber (the kth harmonic of $\eta_n^{m',m}$ is $\widehat{\eta}_n^{m'}(k)\, e^{ikm}$). Then Eq. (97) for the k-th harmonic takes the

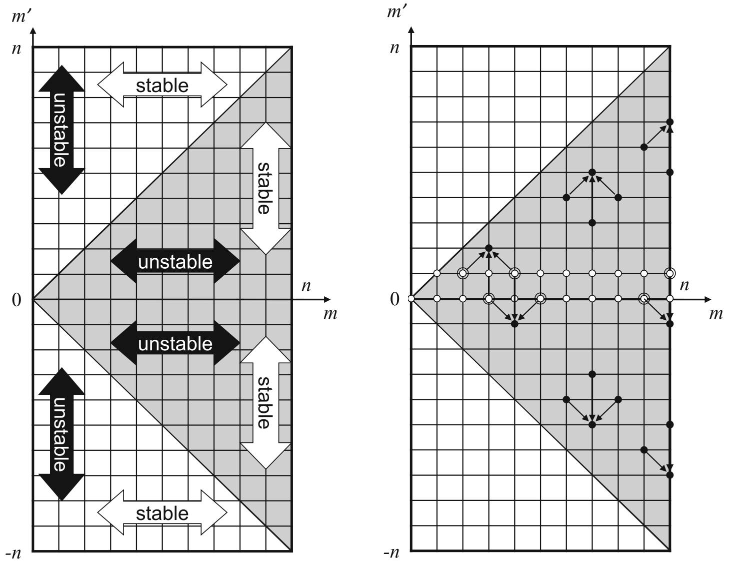


Fig. 5 On the *left* are shown unstable and conditionally stable directions of propagation for the recursion (93) based on the Courant-Friedrichs-Lewy (CFL) criterion. On the *right* are shown the stencils for the recursive algorithm in the shaded region. The nodes with white discs are the initial values of the recursion computed using Eqs (32) and (105)

form

$$\widehat{\eta}_n^{m'+1} - \widehat{\eta}_n^{m'-1} - \frac{b}{n}\left(\widehat{\eta}_n^{m'+1} + \widehat{\eta}_n^{m'-1}\right) = \left(2ic\sin k - 2\frac{b}{n}\cos k\right)\widehat{\eta}_n^{m'}. \tag{99}$$

This is a one-dimensional recurrence relation, with well established stability analysis. Particularly, one can consider solutions of type $\widehat{\eta}_n^{m'} = (\lambda_n)^{m'}$, which after insertion into Eq. (99) results in the characteristic equation

$$\left(1 - \frac{a}{n}\right)\lambda_n^2 - \left(2ic\sin k - 2\frac{b}{n}\cos k\right)\lambda_n - \left(1 + \frac{a}{n}\right) = 0, \tag{100}$$

with roots

$$\lambda_n^{\pm} = \frac{ic\sin k - \frac{b}{n}\cos k \pm \sqrt{\left(ic\sin k - \frac{b}{n}\cos k\right)^2 + 1 - \left(\frac{a}{n}\right)^2}}{1 - \frac{a}{n}}. \tag{101}$$

If $|c\sin k| < 1$ then for $n \to \infty$ we have

$$\begin{aligned}|\lambda_n^{\pm}| &\sim \left(1+\frac{a}{n}\right)\left[1\mp\frac{2b}{n}\cos k\left(1-c^2\sin^2 k+\frac{c^2\sin^2 k\cos k}{\sqrt{1-c^2\sin^2 k}}\right)\right]^{1/2} \\ &\lesssim \left(1+\frac{1}{n}\left[a+|b|\left(1+\frac{c^2}{4\sqrt{1-c^2}}\right)\right]\right).\end{aligned} \tag{102}$$

In the asymptotic region near $|c\sin k| = 1$, since $|c| \leqslant 1$, we also have $|\cos k| \ll 1$. So, denoting $|c\sin k| = 1-n^{-1}c'$ and expanding $\lambda_n^{\pm}$ from Eq. (101) at $n \to \infty$ we obtain

$$|\lambda_n^{\pm}| \sim \left(1+\frac{1}{n}\left(a+c'\right)\right). \tag{103}$$

Note then that for certain k the recursion appears to be unstable, since $|a| \leqslant |b|$ in region $|\mu'| < \mu$. However, in both cases described by Eq. (102) and (103) the growth rate is close to one. So if we have some initial perturbation of magnitude ε_0 we have after n steps (which is the maximum number of steps for propagation from $m' = 0$ to $m' = n$) error ε satisfies

$$\varepsilon \leqslant \varepsilon_0 |\lambda_n|^n \sim \varepsilon_0\left(1+\frac{C}{n}\right)^n \sim \varepsilon_0 e^C. \tag{104}$$

Here C is some constant of order of 1, which does not depend on n, so despite the instability the error should not grow more than a certain finite value, $\varepsilon/\varepsilon_0 \lesssim e^C$.

6 Algorithms for Computation of Rotation Coefficients

We present below two different and novel algorithms for computation of the rotation coefficients $H_n^{m',m}$. The recursive algorithm is more practical (faster), while the Fast Fourier Transform (FFT) based algorithm has an advantage that it does not use any recursion and so it is free from recursion related instabilities. Availability of an alternative independent method enables cross-validation and error/performance studies.

6.1 Recursive Algorithm

The analysis presented above allows us to propose an algorithm for computation of the rotation coefficients based on recursion (50). Note that this recursion, in a shortened form, is also valid for the boundary points, i.e. it holds at $m = n$ where one should set $H_n^{m',n+1} = 0$ (this appears automatically as also $d_n^n = 0$). Using this observation one can avoid some extra work of direct computation of the boundary values (33), which, however, is also not critical for the overall algorithm complexity.

In the algorithm coefficients $H_n^{m'm}$ are computed for each subspace n independently for m and m' located inside a triangle, i.e. for values $m = 0,\ldots,n$, $m' = -m,\ldots,m$. Angle β can take any value from 0 to π.

1: If $n = 0$ set $H_0^{00} = 1$. For other $n = 1,\ldots,p-1$ consider the rest of the algorithm.

2: Compute values $H_n^{0,m}(\beta)$ for $m = 0,\ldots,n$ and $H_{n+1}^{0,m}(\beta)$ for $m = 0,\ldots,n+1$ using Eq. (32) (one can replace there m' with m due to symmetry). It is instructive to compute these values using a stable standard routine for computation of the normalized associated Legendre functions (usually based on recursions), which avoids computation of factorials of large numbers. A standard Matlab function serves as an example of such a routine.

3: Use relation (41) to compute $H_n^{1,m}(\beta)$, $m = 1,\ldots,n$. Using symmetry and shift of the indices this relation can be written as

$$\begin{aligned} b_{n+1}^{0} H_n^{1,m} = {} & \frac{b_{n+1}^{-m-1}(1-\cos\beta)}{2} H_{n+1}^{0,m+1} \\ & - \frac{b_{n+1}^{m-1}(1+\cos\beta)}{2} H_{n+1}^{0,m-1} - a_n^m \sin\beta H_{n+1}^{0,m}. \end{aligned} \tag{105}$$

4: Recursively compute $H_n^{m'+1,m}(\beta)$ for $m' = 1,\ldots,n-1$, $m = m',\ldots,n$ using relation (50) resolved with respect to $H_n^{m'+1,m}$

$$d_n^{m'} H_n^{m'+1,m} = d_n^{m'-1} H_n^{m'-1,m} - d_n^{m-1} H_n^{m',m-1} + d_n^m H_n^{m',m+1}, \tag{106}$$

which for $m = n$ turns into

$$H_n^{m'+1,m} = \frac{1}{d_n^{m'}} \left(d_n^{m'-1} H_n^{m'-1,m} - d_n^{m-1} H_n^{m',m-1} \right). \tag{107}$$

5: Recursively compute $H_n^{m'-1,m}(\beta)$ for $m' = -1,\ldots,-n+1$, $m = -m',\ldots,n$ using relation (50) resolved with respect to $H_n^{m'-1,m}$

$$d_n^{m'-1} H_n^{m'-1,m} = d_n^{m'} H_n^{m'+1,m} + d_n^{m-1} H_n^{m',m-1} - d_n^m H_n^{m',m+1}, \tag{108}$$

which for $m = n$ turns into

$$H_n^{m'-1,m} = \frac{1}{d_n^{m'-1}} \left(d_n^{m'} H_n^{m'+1,m} + d_n^{m-1} H_n^{m',m-1} \right). \tag{109}$$

6: Apply the first and the second symmetry relations (30) to obtain all other values $H_n^{m'm}$ outside the computational triangle $m = 0,\ldots,n$, $m' = -m,\ldots,m$.

Figure 5 (right) illustrates this algorithm. It is clear that the algorithm needs $O(1)$ operations per value of $H_n^{m',m}$. It also can be applied to each subspace independently, and is parallelizable. So, the complexity for a single subspace of degree n is $O\left(n^2\right)$, and the cost to compute all the rotation coefficients for p subspaces ($n = 0,\ldots,p-1$) is $O\left(p^3\right)$. It also can be noticed that for computation of rotation coefficients for all subspaces $n = 0,\ldots,p-1$ the algorithm can be simplified, as instead of computation of $H_n^{0,m}$and $H_{n+1}^{0,m}$ for each subspace in step 2, $H_n^{0,m}$ can be computed for all $n =$

$1,\ldots,p$ $(m=0,\ldots,n)$ and stored. Then the required initial values can be retrieved at the time of processing of the nth subspace.

6.2 FFT Based Algorithms

6.2.1 Basic Algorithm

We propose this algorithm based on Eq. (17), which for $\alpha=0$ and $\gamma=0$ takes the form

$$\begin{aligned} f_n^m\left(\widehat{\varphi};\beta,\widehat{\theta}\right) &= \sum_{m'=-n}^{n} F_n^{m'm}\left(\beta,\widehat{\theta}\right)e^{im'\widehat{\varphi}}, \qquad (110)\\ f_n^m\left(\widehat{\varphi};\beta,\widehat{\theta}\right) &= Y_n^m\left(\theta\left(\widehat{\varphi},\beta,\widehat{\theta}\right),\varphi\left(\widehat{\varphi},\beta,\widehat{\theta}\right)\right),\\ F_n^{m'm}\left(\beta,\widehat{\theta}\right) &= (-1)^{m'}\sqrt{\frac{2n+1}{4\pi}\frac{(n-|m'|)!}{(n+|m'|)!}}P_n^{|m'|}(\cos\widehat{\theta})H_n^{m'm}(\beta)\,. \end{aligned}$$

Here we used the definition of the spherical harmonics (3); $\theta\left(\widehat{\varphi},\beta,\widehat{\theta}\right)$ and $\varphi\left(\widehat{\varphi},\beta,\widehat{\theta}\right)$ are determined by the rotation transform (13) and (14), where we set $\alpha=0$ and $\gamma=0$. The rotation matrix Q is symmetric (see Eq. (9))

$$\begin{pmatrix} x\\ y\\ z\end{pmatrix} = \begin{pmatrix} -\cos\beta & 0 & \sin\beta\\ 0 & -1 & 0\\ \sin\beta & 0 & \cos\beta\end{pmatrix}\begin{pmatrix}\widehat{x}\\ \widehat{y}\\ \widehat{z}\end{pmatrix}. \qquad (111)$$

Using relation between the Cartesian and spherical coordinates (1), we obtain

$$\begin{aligned} \sin\theta\cos\varphi &= -\cos\beta\sin\widehat{\theta}\cos\widehat{\varphi}+\sin\beta\cos\widehat{\theta}, \qquad (112)\\ \sin\theta\sin\varphi &= -\sin\widehat{\theta}\sin\widehat{\varphi},\\ \cos\theta &= \sin\beta\sin\widehat{\theta}\cos\widehat{\varphi}+\cos\beta\cos\widehat{\theta}. \end{aligned}$$

This specifies functions $\varphi\left(\widehat{\varphi},\beta,\widehat{\theta}\right)$ and $\theta\left(\widehat{\varphi},\beta,\widehat{\theta}\right)$, $0\leqslant\varphi<2\pi$, $0\leqslant\theta\leqslant\pi$.

Now, let us fix some $\widehat{\theta}$, such that $\cos\widehat{\theta}$ is not a zero of the associated Legendre function $P_n^m(x)$ at any $m=0,\ldots,n$. Then for a given β function $f_n^m\left(\widehat{\varphi};\beta,\widehat{\theta}\right)$ is completely defined as a function of $\widehat{\varphi}$, while $\beta,\widehat{\theta},m$, and n play a role of parameters. The first equation shows then that this function has a finite Fourier spectrum ($2n+1$ harmonics). The problem then is to find this spectrum ($F_n^{m'm}$), which can be done via the FFT, and from that determine $H_n^{m'm}(\beta)$ using the last relation (110). The complexity of the algorithm for subspace n is, obviously, $O\left(n^2\log n\right)$ and for all subspaces $n=0,1,\ldots,p-1$ we have complexity $O\left(p^3\log p\right)$.

6.2.2 Modified Algorithm

The problem with this algorithm is that at large n the associated Legendre functions (even the normalized ones) are poorly scaled. Analysis of Eq. (90) shows that to have coefficients $F_n^{m'm}$ of the order of unity parameter $\widehat{\theta}$ should be selected as close to $\pi/2$ as possible. On the other hand, this cannot be exactly $\pi/2$ as in this case $P_n^{|m'|}(0) = 0$ for odd values of $n + |m'|$. The following trick can be proposed to fix this.

Consider two functions $g_n^{(1)m}(\widehat{\varphi};\beta) = f_n^m(\widehat{\varphi};\beta,\pi/2)$ and $g_n^{(2)m}(\widehat{\varphi};\beta) = \partial f_n^m\left(\widehat{\varphi};\beta,\widehat{\theta}\right)/\partial\widehat{\theta}\Big|_{\widehat{\theta}=\pi/2}$. The first function has spectrum $\left\{F_n^{m'm}(\beta,\pi/2)\right\}$, while the second function $\left\{\partial F_n^{m'm}\left(\beta,\widehat{\theta}\right)/\partial\widehat{\theta}\Big|_{\widehat{\theta}=\pi/2}\right\}$. Note that $P_n^{|m'|}(\cos\widehat{\theta})$ is an even function of $\widehat{x} = \cos\widehat{\theta}$ for even $n + |m'|$, and an odd function of $\widehat{x} = \cos\widehat{\theta}$ for odd values of $n + |m'|$. In the latter case $x = 0$ is a single zero and $\partial F_n^{m'm}\left(\beta,\widehat{\theta}\right)/\partial\widehat{\theta}\Big|_{\widehat{\theta}=\pi/2}$ is not zero for odd $n + |m'|$ (its absolute value reaches the maximum at $\widehat{x} = 0$), while it is zero for even $n + |m'|$. Hence,

$$\begin{aligned} g_n^m(\widehat{\varphi};\beta) &= g_n^{(1)m}(\widehat{\varphi};\beta) + \gamma_n^m g_n^{(2)m}(\widehat{\varphi};\beta) \\ &= \left[f_n^m(\widehat{\varphi};\beta,\widehat{\theta}) + \gamma_n^m \frac{\partial}{\partial\widehat{\theta}} f_n^m(\widehat{\varphi};\beta,\widehat{\theta})\right]_{\widehat{\theta}=\pi/2}, \end{aligned} \tag{113}$$

where $\gamma_n^m \neq 0$ is an arbitrary number, has Fourier spectrum

$$G_n^{m'm}(\beta) = H_n^{m'm}(\beta) K_n^{m'm}, \tag{114}$$

where, for $m' = 2k - n, \quad k = 0,...,n,$

$$K_n^{m'm} = (-1)^{m'} \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m'|)!}{(n+|m'|)!}} P_n^{|m'|}(0),$$

while, for $\quad m' = 2k - n - 1, \quad k = 1,...,n$

$$K_n^{m'm} = -(-1)^{m'} \sqrt{\frac{2n+1}{4\pi}\frac{(n-|m'|)!}{(n+|m'|)!}} \gamma_n^m (n + |m'|)(n - |m'| + 1) P_n^{|m'|-1}(0),$$

The values for odd $n + |m'|$ come from the well-known recursion for the associated Legendre functions (see [22]),

$$\frac{d}{dx} P_n^m(x) = \frac{(n+m)(n-m+1)}{\sqrt{1-x^2}} P_n^{m-1}(x) + \frac{mx}{1-x^2} P_n^m(x), \tag{115}$$

(note $P_n^{-1}(x) = -P_n^1(x)/(n(n+1))$), which is evaluated at $\widehat{x} = 0$:

$$\left.\frac{d}{d\widehat{\theta}} P_n^{|m'|}(\cos\widehat{\theta})\right|_{\widehat{\theta}=\pi/2} = -\left.\frac{d}{d\widehat{x}} P_n^{|m'|}(\widehat{x})\right|_{\widehat{x}=0}$$
$$= -(n+|m'|)(n-|m'|+1)P_n^{|m'|-1}(0). \qquad (116)$$

Note also that $P_n^{|m'|}(0)$ can be simply expressed via the gamma-function (see [22]) and, so $K_n^{m'm}$ can be computed without use of the associated Legendre functions. The magnitude of arbitrary constant γ_n^m can be selected based on the following observation. As coefficients $H_n^{m'm}(\beta)$ at fixed β large n asymptotically behave as functions of m/n and m'/n we can try to have odd and even coefficients $K_n^{m'm}$ and $K_n^{m'+1,m}$ to be of the same order of magnitude. We can write this condition and the result as

$$\sqrt{\frac{(n-|m'|)!}{(n+|m'|)!}} \sim \sqrt{\frac{(n-|m'|-1)!}{(n+|m'|+1)!}} \gamma_n^m (n+|m'|+1)(n-|m'|), \qquad (117)$$

and $\gamma_n^m \sim 1/n$. Now, we can simplify expression (113) for g_n^m. It is sufficient to consider only positive m, since for negative values we can use symmetry (30), while for $m = 0$ we do not need Fourier transform, as we already have Eq. (32). Using definitions (110) and (3), we obtain

$$\frac{\partial f_n^m}{\partial \widehat{\theta}} = (-1)^m \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}} e^{im\varphi} \times$$
$$\left[\frac{dP_n^m(x)}{dx}\frac{\partial \cos\theta}{\partial \widehat{\theta}} + imP_n^m(x)\frac{\partial \varphi}{\partial \widehat{\theta}}\right]_{x=\cos\theta}. \qquad (118)$$

Differentiating (112) w.r.t. $\widehat{\theta}$ and taking values at $\widehat{\theta} = \pi/2$

$$\left.\frac{\partial \cos\theta}{\partial \widehat{\theta}}\right|_{\widehat{\theta}=\pi/2} = -\cos\beta, \quad \left.\frac{\partial \varphi}{\partial \widehat{\theta}}\right|_{\widehat{\theta}=\pi/2} = \frac{\sin\beta\sin\varphi}{\sin\theta}. \qquad (119)$$

We also have from relations (112) at $\widehat{\theta} = \pi/2$

$$x = \cos\theta = \sin\beta\cos\widehat{\varphi}, \ \cos\varphi = -\frac{\cos\beta\cos\widehat{\varphi}}{\sqrt{1-x^2}}, \ \sin\varphi = -\frac{\sin\widehat{\varphi}}{\sqrt{1-x^2}}. \qquad (120)$$

Using these relations and identity (115), we can write

$$\left[\frac{dP_n^m(x)}{dx}\frac{\partial \cos\theta}{\partial \widehat{\theta}} + imP_n^m(x)\frac{\partial \varphi}{\partial \widehat{\theta}}\right]_{x=\cos\theta} = \qquad (121)$$
$$-(n+m)(n-m+1)\cos\beta\frac{P_n^{m-1}(x)}{\sqrt{1-x^2}} + me^{i\varphi}\sin\beta\frac{P_n^m(x)}{\sqrt{1-x^2}}.$$

Hence, function $g_n^m(\widehat{\varphi};\beta)$ introduced by Eq. (113) can be written as

$$g_n^m(\widehat{\varphi};\beta) = (-1)^m \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}\, e^{im\varphi} \times \tag{122}$$
$$\left[P_n^m(x) - \gamma_n^m \left((n+m)(n-m+1)\cos\beta\,\frac{P_n^{m-1}(x)}{\sqrt{1-x^2}} - me^{i\varphi}\sin\beta\,\frac{P_n^m(x)}{\sqrt{1-x^2}}\right)\right],$$

It may appear that $x = \pm 1$ can be potentially singular, but this is not the case. Indeed, these values can be achieved only when $\beta = \pi/2$ (see the first equation (120)). But in this case, we can simplify Eq. (122), as we have $\cos\beta = 0$, $\sqrt{1-x^2} = |\sin\widehat{\varphi}|$, and so

$$\begin{aligned} e^{i\varphi} &= \cos\varphi + i\sin\varphi = -\frac{\cos\beta\cos\widehat{\varphi}}{\sqrt{1-x^2}} - i\frac{\sin\widehat{\varphi}}{\sqrt{1-x^2}} = -i\,\mathrm{sgn}(\sin\widehat{\varphi}), \\ e^{im\varphi} &= (-i\,\mathrm{sgn}(\sin\widehat{\varphi}))^m, \end{aligned} \tag{123}$$

and Eq. (122) takes the form

$$g_n^m\left(\widehat{\varphi};\frac{\pi}{2}\right) = \sqrt{\frac{2n+1}{4\pi}\frac{(n-m)!}{(n+m)!}}\,[i\,\mathrm{sgn}(\sin\widehat{\varphi})]^m \left(1 - \frac{i\gamma_n^m m}{\sin\widehat{\varphi}}\right) P_n^m(\cos\widehat{\varphi}). \tag{124}$$

Note that this expression has a removable singularity at $\sin\widehat{\varphi} = 0$. Indeed for $m \geqslant 2$ we have $P_n^m(\cos\widehat{\varphi}) \sim \sin^m\widehat{\varphi}$, while for $m = 1$ we have

$$\left.\frac{P_n^1(x)}{\sqrt{1-x^2}}\right|_{x\to\pm 1} = -\frac{dP_n}{dx}(\pm 1) = -n(n+1)\varepsilon_{\pm n}, \tag{125}$$

where symbol ε_m is defined by Eq. (7). So,

$$g_n^m\left(\pi k;\frac{\pi}{2}\right) = \begin{cases} \gamma_n^1(-1)^{k+1}\frac{1}{2}\sqrt{\frac{2n+1}{4\pi}n(n+1)}, & m = 1, \\ 0, & m \geqslant 2. \end{cases} \tag{126}$$

Hence, the modified algorithm is based on the equation

$$g_n^m(\widehat{\varphi};\beta) = \sum_{m'=-n}^{n} G_n^{m'm}(\beta)\, e^{im'\widehat{\varphi}}, \tag{127}$$

where function $g_n^m(\widehat{\varphi};\beta)$ can be computed for equispaced values of $\widehat{\varphi}$ sampling the full period. The FFT produces coefficients $G_n^{m'm}(\beta)$ from which $H_n^{m'm}(\beta)$ can be found using Eq. (114).

7 Numerical Experiments

The algorithms were implemented in Matlab and tested for $n = 0, \ldots, 10000$.

7.1 Test for Recursion Stability

First we conducted numerical tests of the algorithm stability. Note that the dependence on β comes only through the initial values, which are values of coefficients for layers $m' = 0$ and $m' = 1$. Hence, if instead of these values we put some arbitrary function (noise) then we can measure the growth of the magnitude of this noise as the recursive algorithm is completed. For stable algorithms it is expected that the noise will not amplify, while amplification of the noise can be measured and some conclusions about practical value of the algorithm can be made. Formally the amplitude of the noise can be arbitrary (due to the linearity of recursions), however, to reduce the influence of roundoff errors we selected it to be of the order of unity.

Two models of noise were selected for the test. In the first model perturbations $\eta_n^{m',m}$ at the layers $m' = 0$ and $m' = 1$ were specified as random numbers distributed uniformly between -1 and 1. In the second model perturbations were selected more coherently. Namely, at $m' = 0$ the random numbers were non-negative (distributed between 0 and 1). At $m' = 1$ such a random distribution was pointwise multiplied by factor $(-1)^m$. The reason for this factor is that effectively this brings some symmetry for resulting distributions of $\eta_n^{m',m}$ for $m' > 0$ and $m' < 0$. Figure 6 shows that in the first noise model the overall error (in the L^∞ norm) grows as $\sim n^{1/4}$, while for the second noise model the numerical data at large enough n are well approximated by $\varepsilon = \varepsilon_0 n^{1/2}$. Note that the data points shown on this figure were obtained by taking the maximum of 10 random realizations per each data point. On the right hand side of Fig. 6 are shown error distributions for some random realization and some n ($n = 100$, the qualitative picture does not depend on n). It is seen that for the first noise model the magnitude of $\eta_n^{m',m}$ is distributed approximately evenly (with higher values in the central region and diagonals $m' = \pm m$). For the second noise model the distribution is substantially different. The highest values are observed in the boundary regions $m' \approx \pm n$ and $m \approx \pm n$ with the highest amplitudes near the corners of the computational square in the (m, m') space.

The behavior observed in the second noise model can be anticipated, as the scheme is formally unstable, the absolute values of coefficients a and b (see Eq. (98)) grow near the boundaries of the computational domain and a fast change of these coefficients near the boundaries requires some other technique for investigation of instabilities than the method used. Smaller errors and their distribution observed in the first noise model are more puzzling, and we can speculate about some cancellation effects for random quantities with zero mean appearing near the boundaries, and to the variability of coefficients a, b, and c in Eq. (98), so that the stability analysis is only approximate. What is important that in all our tests with dif-

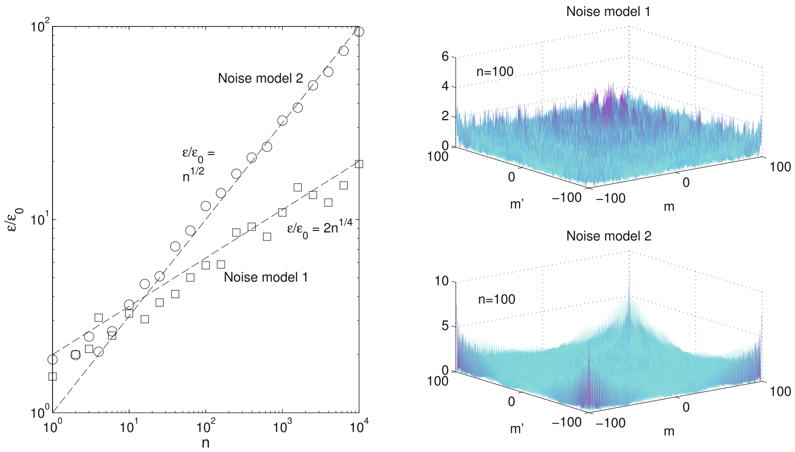


Fig. 6 The chart on the *left* shows amplification of the noise in the proposed recursive algorithms for two noise models. The charts on the *right* show distributions of the noise amplitude for some random realization at $n = 100$

ferent distributions of initial values of $\eta_n^{m',m}$ we never observed exponential growth. The maximum growth rate behaved at large n as n^{α}, $\alpha \approx 1/2$. Hence, for $n \sim 10^4$ one can expect the errors in the domain of two orders of magnitude larger than the errors in the initial conditions, which makes the algorithm practical. Indeed, in double precision, which provides errors $\sim 10^{-15}$ in the initial values of the recursions, then for $n = 10^4$ one can expect errors $\sim 10^{-13}$, which is acceptable for many practical problems. Of course, if desired the level of the error can be reduced, if needed, using e.g. quadruple precision, etc.

7.2 Error and Performance Tests

The next error tests were performed for actual computations of $H_n^{m',m}$. For small enough n ($n \sim 10$) one can use an exact expression (26) as an alternative method to figure out the errors of the present algorithm. Such tests were performed and absolute errors of the order of 10^{-15}, which are consistent with double precision roundoff errors were observed. The problem with sum (26) is that at large n it requires computation of factorials of large numbers, which creates numerical difficulties. While computation of factorials and their summation when the terms have the same sign is not so difficult (e.g. using controlled accuracy asymptotic expansion), the problem appears in the sums with large positive and negative terms. In this case to avoid the loss of information special techniques of working with large integers (say, with thousand digits) should be employed. This goes beyond the present study, and we used different methods for validation than comparing with these values.

Another way is to compute $H_n^{m',m}(\beta)$ is based on the flip decomposition, i.e. to use one of the equations (38)–(40). In this decomposition all coefficients are "good" in terms that complex exponents or cosines can be computed accurately. These formulae require the flip rotation coefficients, $H_n^{m'm}(\pi/2)$, which should be computed and stored to get $H_n^{m',m}(\beta)$ for arbitrary β. Note that these relations also hold for $\beta = \pi/2$, which provides a self-consistency test for $H_n^{m'm}(\pi/2)$. Despite the summations in (38)–(40) requires $O(n^3)$ operations per subspace n and are much slower than the algorithms proposed above, we compared the results obtained using the recursive algorithm for consistency with Eq. (38) and found a good agreement (up to the numerical errors reported below) for n up to 5000 and different β including $\beta = \pi/2$.

One more test was used to validate computations, which involve both recursions and relation (40). This algorithm with complexity $O\left(n^2\right)$ per subspace was proposed and tested in [24]. Their coefficients $H_n^{m'm}(\pi/2)$ were found using the present recursion scheme, which then were used only to compute the diagonal coefficients $H_n^{mm}(\beta)$ and $H_n^{m,m+1}(\beta)$ at arbitrary β. The recursion then was applied to obtain all other coefficients, but with propagation from the diagonal values, not from $H_n^{0,m}(\beta)$ and $H_n^{1,m}(\beta)$. Motivation for this was heuristic, based on the observation that $O(1)$ magnitudes are achieved on the diagonals and then they can decay exponentially (see Fig. 2), so it is expected that the errors will also decay. The tests for n up to 5000 showed that such a complication of the algorithm is not necessary, and both the present and the cited algorithms provide approximately the same errors.

In the present study as we have two alternative ways to compute $H_n^{m'm}(\beta)$ using the recursive algorithm and the FFT-based algorithm, we can use self-consistency and cross validation tests to estimate the errors of both methods.

Self consistency tests can be based on relation (52). In this case computed $H_n^{m'm}(\beta)$ were used to estimate the following error

$$\varepsilon_n^{(0)}(\beta) = \max_{m,m'} \left| \sum_{\nu=-n}^{n} H_n^{m'\nu}(\beta) H_n^{\nu m}(\beta) - \delta_{mm'} \right|, \quad n = 0, 1, \ldots \tag{128}$$

We also found the maximum of $\varepsilon_n^{(0)}(\beta)$ over five values of $\beta = 0, \pi/4, \pi/2, 3\pi/4, \pi$ for the recursive algorithm and for two versions of the FFT-based algorithms. For the basic FFT-based algorithm we used $\widehat{\theta} = \pi/2 - \xi/n$, where ξ was some random number between 0 and 1. For the modified algorithm, which has some arbitrary coefficient γ_n^m we used $\gamma_n^m = 1/n$, which, as we found provides smaller errors than $\gamma_n^m = 1$ or $\gamma_n^m = 1/n^2$ and consistent with the consideration of magnitude of the odd and even normalization coefficients (see Eq. (117)). The results of this test are presented in Fig. 7. It is seen that while at small n the error is of the order of the double precision roundoff error, at larger n it grows as some power of n. The error growth rate at large n for the recursive algorithm is smaller and approximately $\varepsilon_n^{(0)} \sim n^{1/2}$, which is in a good agreement with the error growth in the noise model #2 discussed above. For the FFT-based algorithms the error grows approximately as

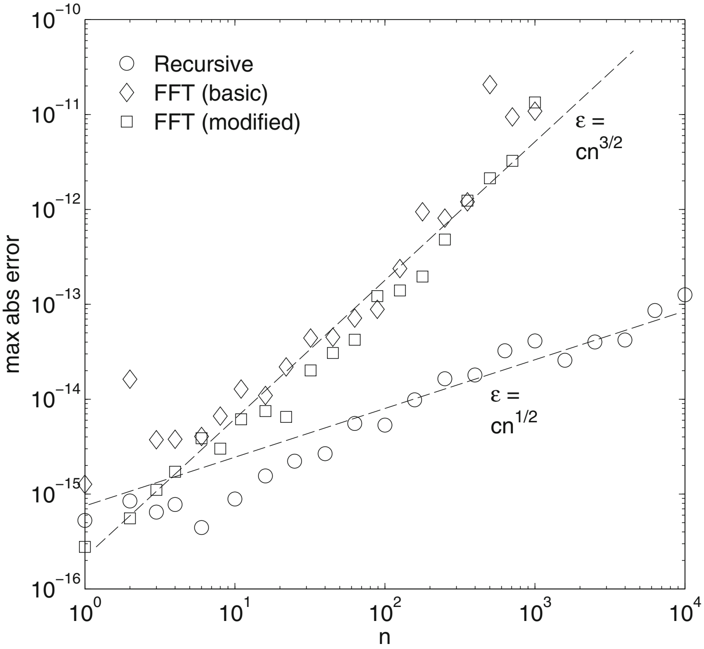


Fig. 7 Self-consistency error test of the recursive and FFT-based algorithms validating that the symmetric matrix of rotation coefficients is unitary

$\varepsilon_n^{(0)} \sim n^{3/2}$, so it can be orders of magnitude larger than in the recursive algorithm, while still acceptable for some practical purposes up to $n \sim 10^3$. Such growth can be related to summation of coefficients of different magnitude in the FFT, which results in the loss of information. Comparison of the basic and the modified FFT-based algorithms show that the errors are approximately the same for the both versions, while the error in the basic algorithm can behave more irregularly than that in the modified algorithm. This can be related to the fact that used values of $\widehat{\theta}$ at some n were close to zeros of the associated Legendre functions, and if this algorithm should be selected for some practical use then more regular way for selection of $\widehat{\theta}$ should be worked out.

The second test we performed is a cross-validation test. In this case we computed the difference

$$\varepsilon_n^{(1)}(\beta) = \max_{m,m'} \left| H_n^{m'm(FFT)}(\beta) - H_n^{m'm(rec)}(\beta) \right|, \quad n = 0, 1, ... \qquad (129)$$

We also measured and compared the wall clock times for execution of the algorithms (standard Matlab and its FFT library on a standard personal computer). Results of these tests are presented in Fig. 8. First we note that both FFT-based algorithms showed large errors for n ˜ 2300 and failed to produce results for larger n. This can be related to the loss of information in summation of terms of different magnitude,

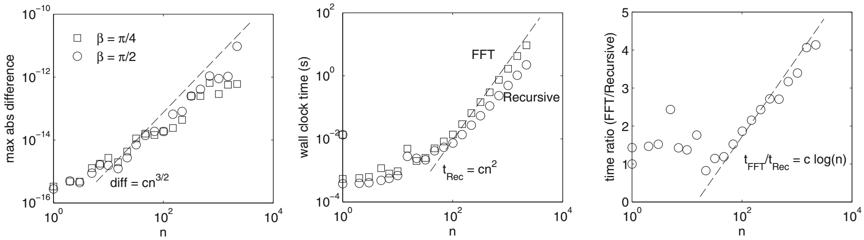


Fig. 8 The maximum of absolute difference in the rotation coefficients $H_n^{m'm}$ computed using the recursive and the FFT-based (modified) algorithms as a function of n at two values of β (*left*). The *center* and the *right* plots show wall-clock times for these algorithms and the ratio of these times, respectively (standard PC, Matlab)

as it was mentioned above. On the other hand the recursive algorithm was producing reasonable results up to $n = 10000$ and there were no indications that it may not run for larger n (our constraint was the memory available on the PC used for the tests). So the tests presented in the figure were performed for $n \leqslant 2200$. It is seen that the difference is small (which cross-validates the results in this range), while $\varepsilon_n^{(1)}(\beta)$ grows approximately at the same rate as the error $\varepsilon_n^{(0)}$ for the FFT-based algorithms shown in Fig. 7. Taking into account this fact and numerical instability of the FFT-based algorithms we relate it rather to the errors in that algorithms, not in the recursive algorithm. We also noticed that the FFT-based algorithm at large n produces somehow larger errors for $\beta = \pi/2$ than for other values tried in the tests. In terms of performance, it is seen that both, the recursive and the FFT-based algorithm are well-scaled and $O\left(n^2\right)$ scaling for large enough n is acieved by the recursive algorithm nicely. The FFT-based algorithm in the range of tested n can be several times slower. The ratio of the wall-clock times at $n > 100$ is well approximated by a straight line in semi-logarithmic plots, which indicates $\log n$ behavior of this quantity, as expected.

8 Conclusion

This chapter first presented a study of the asymptotic behavior of the rotation coefficients $H_n^{m'm}(\beta)$ for large degrees n. Based on this study, we proposed a recursive algorithm for computation of these coefficients, which can be applied independently for each subspace n (with cost $O\left(n^2\right)$) or to p subspaces ($n = 0, \ldots, p-1$) (cost $O\left(p^3\right)$). A theoretical and numerical analysis of the stability of the algorithm shows that while the algorithm is weakly unstable, the growth rate of perturbations is small enough, which makes it practical for computations for relatively large n (the tests were performed up to $n = 10^4$, but the scaling of the error indicates that even larger n can be computed). Alternative FFT-based algorithms of complexity $O\left(n^2 \log n\right)$ per subspace n were also developed and studied. Both the FFT-based and recursive computations produce consistent results for $n \lesssim 10^3$. In this range the recursive algorithm is faster and produces smaller errors than the FFT-based algorithm.

References

1. Greengard L., Rokhlin V. A fast algorithm for particle simulations. J Comput Phys. 1987;73:325–348.
2. Rokhlin V. Diagonal forms of translation operators for the Helmholtz equation in three dimensions. Appl Comput Harmon Anal. 1993;1:82–93.
3. Epton M.A., Dembart B. Multipole translation theory for the three-dimensional Laplace and Helmholtz equations." SIAM J Sci Comput. 1995;16(4):865–897.
4. Chew W.C., Jin J.-M., Michielssen E., Song J. Fast and efficient algorithms in computational electromagnetics. Boston: Artech House; 2001.
5. Gumerov N.A., Duraiswami R. Fast multipole methods for the Helmholtz equation in three dimensions. Oxford: Elsevier; 2005.
6. White C.A., Head-Gordon M. Rotation around the quartic angular momentum barrier in fast multipole method calculations. J Chem Phys. 1996;105(12):5061–5067.
7. Suda R., Takami M. A fast spherical harmonics transform algorithm. Math Comput. 2002;71(238):703–715.
8. Cheng H., Crutchfield W.Y., Gimbutas Z., Greengard L., Ethridge F., Huang J., Rokhlin V., Yarvin N., Zhao J. A wideband fast multipole method for the Helmholtz equation in three dimensions, J Comput Phys. 2006;216:300–325.
9. Gumerov N.A., Duraiswami R. A broadband fast multipole accelerated boundary element method for the three dimensional Helmholtz equation. J Acoust Soc Am. 2009;125:191–205.
10. Tang Z., Duraiswami R., Gumerov N.A. Fast algorithms to compute matrix-vector products for Pascal matrices. UMIACS TR 2004-08 and Computer Science Technical Report CS-TR 4563, University of Maryland, College Park; 2004.
11. Biedenharn L.C., Louck J.D. Angular momentum in quantum physics: theory and application. Reading: Addison-Wesley; 1981.
12. Ivanic J., Ruedenberg K. Rotation matrices for real spherical harmonics. Direct determination by recursion, J Phys Chem. 1996;100(15):6342–6347.
13. Gumerov N.A., Duraiswami R. Recursions for the computation of multipole translation and rotation coefficients for the 3-D Helmholtz equation. SIAM J Sci Comput. 2003;25(4):1344–1381.
14. Dachsel H. Fast and accurate determination of the Wigner rotation matrices in the fast multipole method. J Chem Phys. 2006;124:144115–144121.
15. Gimbutas Z., Greengard L. A fast and stable method for rotating spherical harmonic expansions. J Comput Phys. 2009;228:5621–5627.
16. Courant R., Friedrichs K., Lewy,H. On the partial difference equations of mathematical physics. IBM J Res Dev. 1967;11(2):215–234. (English translation of original (in German) work "Uber die partiellen Differenzengleichungen der mathematischen Physik," Mathematische Annalen, 100(1), 32–74, 1928).
17. Charney J.G., Fjortoft R., von Neumann J. Numerical intergation of the barotropic vorticity equation. Quarterly J Geophys, 1950;2(4):237–254.
18. Wigner E.P. Gruppentheorie und ihre Anwendungen auf die Quantenmechanik der Atomspektren. Braunchweig: Vieweg; 1931.
19. Vilenkin N.J. Special functions and the theory of group representations. USA: American Methematical Society; 1968. (Translated from the original Russian edition of 1965).
20. Stein S. Addition theorems for spherical wave functions. Quart Appl Math. 1961;19:15–24.
21. Gumerov N.A., Duraiswami R. Computation of scattering from clusters of spheres using the fast multipole method. J Acoust Soc Am. 2005;117(4):1744–1761.

22. Abramowitz M., Stegun I.A. Handbook of mathematical functions.Washington, DC: National Bureau of Standards; 1972.
23. Nayfeh A.H. Perturbation methods. New-York: Wiley; 1973.
24. Gumerov N.A., Berlin K., Fushman D., Duraiswami R. A hierarchical algorithm for fast Debye summation with applications to small angle scattering. J Comput Chem. 2012; 33:1981–1996.
25. Park W., Leibon G., Rockmore D.N., Chirikjian G.S. Accurate image rotation using Hermite expansions. IEEE Trans Image Proc. 2009;18:1988–2003.

Analyzing Fluid Flows via the Ergodicity Defect

Sherry E. Scott

Abstract The ergodicity defect, a relatively new approach for analyzing fluid flows, is presented. The technique combines tools and concepts from ergodic theory and wavelet theory, and we briefly consider this theoretical background. The ergodicity defect provides a measure of the complexity of individual fluid particle trajectories and this measurement is used to identify Lagrangian coherent structures in the flow. Results for both idealized and realistic ocean flows are compared and contrasted with other methods for identifying coherent structures. Other possible applications for the technique are also discussed.

Keywords Lagrangian coherent structures (LCS) · Ergodicity · Dynamical systems · Fluid flows · Wavelets

1 Introduction

Coherent structures such as gyres and eddies play an important role in the behavior and dynamics of ocean flows. For example, these structures can either aid or prevent the transport of biomass and enhance or impede the exchange and mixing of materials. Much of our understanding of these key structures (and the overall ocean) comes from data collected along the path of devices such as floats and drifters that move with the ocean flow, i.e., from Lagrangian data. Dynamical systems tools and theory allow us to use these trajectory paths to delineate the coherent structures that develop in the flow. Coherent structures determined in this manner are known as Lagrangian coherent structures (LCS) and although several methods for detecting LCS are known, no one method addresses all cases and issues such as the quantifi-

S. E. Scott (✉)
Department of Mathematics, Statistics and Computer Science, Marquette University,
P.O. Box 1881, Milwaukee, WI 53201-1881, USA
e-mail: sherry.scott@mu.edu, sscott1008@google.com

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_6

cation of the transport [5]. In this correspondence we consider a technique called the ergodicity defect [21, 24] for analyzing the dynamics of the flow and identifying the LCS. We also discuss other possible uses of the technique in geosciences related problems.

1.1 Background and Notation

Given a map T on a probability space (M,β,μ), a set $A \in \beta$ is T-invariant if $A = T^{-1}(A)$ and the map T is said to be measure μ preserving if for all $A \in \beta$, $\mu(T^{-1}(A)) = \mu(A)$. Such a system is ergodic if the only T-invariant subsets A are such that $\mu(A) = 0 \,\text{or}\, \mu(A) = 1$. Throughout the paper, a fluid flow is treated as a dynamical system (M,β,μ,T) where T is μ-preserving and μ is the Lebesgue measure. More specifically, we are interested in trajectories of a dynamical system (M,β,μ,T) in discrete time, i.e.,

$$\mathbf{x}_{i+1} = T(\mathbf{x}_i), \tag{1}$$

where $\mathbf{x}$ denotes a vector in $\mathbb{R}^d$. We work primarily in $\mathbb{R}^1$ and $\mathbb{R}^2$ and assume a periodic domain. Thus, in $\mathbb{R}^1$, M is a unit circle, μ is the length function, and T is a length preserving map and in $\mathbb{R}^2$, M is a torus, μ is the area function, and T is an area preserving map. We also assume that for a given fluid flow $\mathbf{u}(\mathbf{x},t)$, the fluid particle trajectories with initial position $\mathbf{x}_0$ at time t_0, denoted as $\mathbf{x}(t;\mathbf{x}_0,t_0)$ satisfy

$$\frac{d\mathbf{x}}{dt} = \mathbf{u}(\mathbf{x},t)\,. \tag{2}$$

We use the standard notation $\|f\|$ to denote the L^2 norm of a function f, i.e., $\|f(\mathbf{x})\|^2 = \int_M f^2(\mathbf{x})d\mathbf{x}$ and we denote the inner product by $\langle,\rangle$, i.e., $< f,g >= \int fgd\mathbf{x}$ is the inner product of functions f and g. The indicator function on a set A (also known as the characteristic function) is given as $\chi_A(\mathbf{x})$. For example, the indicator function on the unit interval $[0,1)$ can be given as follows

$$\chi_{[0,1)}(x) = \begin{cases} 1 & \text{if } x \in [0,1) \\ 0 & \text{else} \ . \end{cases}$$

Proving that any given system is ergodic is usually a very difficult task, however there are several definitions that may make the task more tractable [27]. One standard way of viewing ergodicity is in terms of Birkhoff's characterization which, as did much of the work on ergodicity, originated in Boltzmann's hypothesis concerning the equality of time averages and space averages [19]. By Birkhoff's characterization, if T is a measure-preserving map on a probability space (M,β,μ), then T is ergodic, if and only if, for all $f \in L^1(\mu)$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{r=1}^{n} f(T^r(\mathbf{x})) = \int_M f(\mathbf{x})d\mu \quad a.e., \tag{3}$$

where the left hand side of (3) gives the time average of f and the right hand side is the space average of f. In this way, ergodic maps T are characterized via the equality of the time average (along the trajectories of T) and the space average of integrable functions f on the space [2, 9, 13, 30]. Throughout this chapter, for any function f and map T on M, the space average of f is denoted as $\overline{f}$, i.e.,

$$\overline{f}(\mathbf{x}) = \int_M f(\mathbf{x}) d\mathbf{x},$$

and the time average is indicated as f^*, so that

$$f^*(\mathbf{x},T) = \lim_{N\to\infty} \frac{1}{N} \sum_{r=1}^{N} f(T^r \mathbf{x}),$$

gives the time average of f along the trajectory $T^r(\mathbf{x})$ starting at the point $\mathbf{x}$.

In the definition of ergodicity defect we use Birkhoff's characterization of ergodicity and we view the function f as the way in which the underlying system is being observed or analyzed. In this work, we choose wavelets as the analyzing functions and we take the time average of wavelets along a (fluid particle) trajectory of the flow. More specifically, as an intuition building step, we use the Haar father wavelets as the analyzing functions and define the ergodicity defect as the difference between the time average and the space average. Wavelets allow us to analyze in terms of how the trajectories sample the space when considered at different spatial resolutions. However, in this presentation, this scaling analysis aspect is not the focus, and instead, we average over a range of scales of interest. The use of wavelets in our fluid flow analysis follows a rich history of wavelet theory in the geosciences and ocean related issues, e.g., see [18].

1.2 Motivation

The ergodicity defect utilizes the premise that in order to investigate the complexity of flows, one can consider the complexity of their trajectories, where this complexity is measured in terms of ergodicity, i.e., how the trajectory samples the region of interest. In [24], the desire is to distinguish maps and flow fields in terms of their deviation from ergodicity; however, here, the goal is to analyze the coherent structures in a given fluid flow. This idea of analyzing a flow via chaotic advection dates back several decades, e.g., see [1]. It is known that standard approaches based on an Eulerian (fixed time) perspective such as eyeballing the vector field cannot distinguish cases of differing trajectory complexity [8] and tracking these features by satellite altimetry gives ambiguous results [5]. But, the importance of Lagrangian studies in the ocean suggest that an approach based on analyzing the complexity of flows and their trajectories in terms of ergodicity may be more appropriate.

For a given flow, every initial point has a trajectory and every trajectory has an ergodicity defect value. The ergodocity defect value indicates the complexity of the trajectory and distinguishing the different trajectories according to these complexities helps to identify structures in the flow.

To fix ideas, let us consider the conceptual illustration given in Fig. 1 that shows a flow with a stable (attracting) manifold, and unstable (repelling) manifold, a hyperbolic trajectory and two other trajectories that veer away from the hyperbolic trajectory and lie on different sides of the stable manifold. The areas of attracting or repelling material have trajectories whose complexity (ergodicity defect value) is similar to the hyperbolic trajectory, and trajectories on different sides of the manifolds have different behavior and hence different defect values. In this manner, a color or grayscale map of the relative complexity of the trajectories, i.e., a map of more complex versus less complex trajectories, reveals different regions of distinct behavior and the coherent structures which accompany these regions.

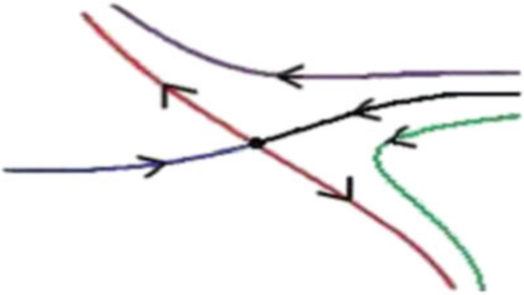


Fig. 1 Trajectories for stable and unstable manifold intersecting at the hyperbolic trajectory

Other schemes such as the Mix-Norm and the finite time Lyapunov exponents (FTLE) approach are aimed at addressing similar issues of analyzing LCS and in [5] an operator theory method is used. Hyperbolic and stable/unstable manifolds are key in determining the LCS barriers and transport pathways, and the FTLE method provides numerical estimates of the stable/unstable manifolds by measuing the maximum rate of separation between a fluid particle trajectory and its nearby neighbors [7]. There are also other uses for such measures and diagnostics, e.g., the authors of [14] propose an algorithm for the design of an optimal sampling strategy using ergodicity and Fourier basis functions instead of wavelets. Similarly the ergodicity defect can be applied in these settings.

The discussion is organized as follows. First, a brief background and motivation are considered. In the Section 2, the ergodicity defect is discussed and then in Section 3, the application of the ergodicity defect technique is applied to a few examples and the ergodicity defect results are compared to other methods. In the last section, other possible ocean flow related uses of the ergodicity defect are proposed.

2 Ergodicity Defect (ED)

Since an ergodic system satisfies the requirement that the time average is equal to the space average, we consider the difference between the time average and the space average to obtain a value that captures the deviation of a system from ergodicity. We call this value the ergodicity defect of the system. In this section, we first give general definitions for the ergodicity defect of a flow. These general definitions are intended for use in comparing flows according to their deviation from ergodicity when the identity map is taken as the worse case scenario. That is, as the identity map does not move any points, the technique is normalized so that the identity map is the "least ergodic."

Definition 1. The *ergodicity defect of a map T with respect to a function f is given as*

$$d(f,T) = \omega(f)[\int (f^*(\mathbf{x},T) - \overline{f})^2 d\mathbf{x}], \tag{4}$$

where $\omega(f)$ is a normalization factor chosen such that $d(f,I) = 1$, where I is the identity.

Note that with this normalization, the defect values range from 0 to 1 with 1 denoting a flow that deviates the most from being ergodic—like the identity map—and 0 denoting an ergodic map, i.e., a map for which the time averages and space averages are equal.

If we have a finite family of functions $F = \{f_n\}_{n=1}^N$, then we can consider the ergodicity defect of T with respect to the family F as follows.

Definition 2. The *ergodicity defect of a map T with respect to a family of functions $F = \{f_n\}_{n=1}^N$* is expressed as:

$$d(F,T) = \omega(F)[\sum_{n=1}^{N} \int (f_n^*(\mathbf{x},T) - \overline{f_n})^2 d\mathbf{x}], \tag{5}$$

where $\omega(F)$ is the normalization factor such that $d(F,I) = 1$.

We choose a basis as our analyzing functions in order to obtain a good representation of the overall behavior of the flow. More specifically, we take a wavelet basis and we start with the Haar father wavelet. However, note that the Haar mother wavelet or other wavelets can also be used, e.g., see [23].

In $\mathbb{R}^1$, the Haar father wavelet, or Haar scaling function, is the indicator function on $[0,1)$, i.e., $\chi_{[0,1)}(x)$ and the Haar mother wavelet ψ is given as a difference of two indicator functions, i.e., $\psi(x) = \chi_{[1/2,1)}(x) - \chi_{[0,1/2)}(x)$. If we denote the Haar scaling function as ϕ, we have that $\phi(x) = \chi_{[0,1)}(x)$ and $\psi(x) = \phi(2x-1) - \phi(2x)$. By taking the dyadic dilations and translations of ϕ, we obtain the Haar scaling functions at scale s as follows.

$$\phi_j^{(s)}(x) = \phi(2^s x - (j-1)) \quad j = 1, ..., 2^s. \tag{6}$$

If we think in terms of the phase space or the domain of interest, then the Haar scaling functions at scale s correspond to a dyadic equipartition of the phase space in which the support of each wavelet $\phi_j^{(s)}$ is a set in the partition—specifically an interval of length $1/2^s$. Thus, the time average $\phi_j^{(s),*}$ gives the average time the trajectory resides in the jth interval/set in the partition and the space average is the measure of that set, i.e., the space average $\overline{\phi_j^{(s)}}$ is $1/2^s$.

One-Dimensional (1D) Haar Ergodicity Defect

Using the family of dilations and translations of the Haar scaling function at a fixed scale s, $\{\phi_j^{(s)}\}$, and $j = 1, ...2^s$, the Haar ergodicity defect at scale s of the map T is given as

$$d_\chi(s,T) = \frac{2^s}{2^s - 1} \sum_{j=1}^{2^s} [\int (\phi_j^{(s),*}(x,T) - \frac{1}{2^s})^2 dx], \tag{7}$$

where the index χ denotes that the scaling function ϕ is $\chi_{[0,1)}$ and the normalization factor $\omega(s)$ is $\frac{2^s}{2^s-1}$.

If the domain of interest is in $\mathbb{R}^d$, $d \geq 2$, then there is an appropriate collection of mother wavelets and the wavelet basis is usually denoted as $\{\psi_{m,n}\}_{(m,n)\in Z\times Z}$ where m is the dilation index and n is the translation index. For a torus in $\mathbb{R}^2$, the scaling functions of the Haar partition at scale s can be given as

$$\phi_{i_1 i_2}^{(s)}(x,y) = \phi_{i_1}^{(s)}(x)\phi_{i_2}^{(s)}(y), \quad i_1, i_2 = 1 \dots 2^s,$$

where $\phi_i^{(s)}$ is given by (6). In 2D, the Haar scaling functions at scale s correspond to an equipartition of the phase space into 2^{2s} squares of area $1/2^{2s}$ and we have the following Haar defect in 2D.

Two-Dimensional (2D) Haar Ergodicity Defect

Using the family of dilations and translations of the Haar scaling function at fixed scale s, $\phi_{i_1 i_2}^{(s)}(x,y) = \phi_{i_1}^{(s)}(x)\phi_{i_2}^{(s)}(y)$ and $i_1, i_2 = 1 \dots 2^s$, the Haar ergodicity defect at scale s of a map T (on a domain in $\mathbb{R}^2$), is given as

$$d_\chi(s,T) = \frac{4^s}{4^s-1}\sum_{i_1,i_2=1}^{2^s}\Big[\int(\phi_{i_1 i_2}^{(s),*}(\mathbf{x},T) - \frac{1}{2^{2s}})^2 d\mathbf{x}\Big], \tag{8}$$

where the normalization factor $\omega(s)$ is $\frac{4^s}{4^s-1}$.

The ergodicity defect at scale s is computed numerically by taking the domain of interest and mapping it onto a unit square and then partitioning the square into 2^{2s} squares of edge length 2^{-s} and area 2^{-2s}. Note that in this computation the domain of interest can be an arbitrary region in space, which is then mapped to the unit square. A sample of points $\{x_i\}$ are taken from a trajectory $\mathbf{x}(t,(\mathbf{x}_0,t_0))$ where $(\mathbf{x}_0,t_0)$ are the initial conditions, so that the time average $\phi_{i_1 i_2}^{(s),*}$ is the number average of points N_{i_1,i_2} from $\{x_i\}$ that lie in the support of $\phi_{i_1 i_2}^{(s)}$.

Ergodicity Defect for an Individual Trajectory

The ergodicity defect formulas given thus far are designed to capture the deviation of systems or flows from ergodicity. In order to determine the LCS in the flow [21], we apply the ergodicity defect to each individual trajectory $\mathbf{x}(t,(\mathbf{x}_0,t_0))$ as follows

$$d_\chi(s;\mathbf{x}_0,t_0) = \sum_{i_1,i_2=1}^{2^s}\Big[(\phi_{i_1 i_2}^{(s),*}(\mathbf{x}_0,T) - \frac{1}{2^{2s}})^2\Big]. \tag{9}$$

Or computationally we have

$$d_\chi(s;\mathbf{x}_0,t_0) = \sum_{j=1}^{2^{2s}}\Big[(\frac{N_j(s)}{N} - \frac{1}{2^{2s}})^2\Big], \tag{10}$$

where N is the total number of sample points taken from the trajectory and $N_j(s)$ is the number of points that lie inside the jth square that serves as the support for the jth wavelet.

3 Some Results for ED and Other Similar Metrics

As mentioned in the background section, there are a considerable number of other metrics for identifying LCS. We consider in this section two such methods—the finite time Lyapunov exponent (FTLE) and the correlation dimension. FTLE is a frequently used method that is based on the separation rate of nearby trajectories. The correlation dimension is a fractal dimension method which is more commonly used in a time series setting. In [6], the correlation dimension of a set of points

$\{x_i\}$ is computed by covering the domain with adjacent squares of edge length s and counting the number of points N_j that lie inside the j-th square, so that for a total of N sample points, the distribution function $F(s)$ is estimated as $F(s) = \frac{1}{N^2}\sum_j [N_j(s)]^2$. Then the correlation dimension c is computed as the slope of the log-log plot of $F(s)$ versus s for small s and large N. The correlation dimension takes values varying from 0 to 2 with 0 corresponding to a stationary point, 1 corresponding to a smooth 1D curve and 2 corresponding to a curve that fills the 2D area. Note that the correlation dimension values are opposite the values of the ergodicity defect, which is 0 for a curve that samples well and 1 for a stationary point. Also, for the correlation dimension it is necessary to assume that the distribution function $F(s)$ obeys a power law of the form $F(s) \cong s^c$, but the ergodicity defect requires no such assumption.

For both the ergodicity defect and the correlation dimension computations, we seed the domain and track the path of each particle trajectory. Then we map the sampled domain to the unit square and partition the square into squares of edge length 2^{-s} for each s. Next we compute a complexity value for each trajectory and assign the value to its initial position. Finally we grayscale or color code the domain using the complexity values assigned to each initial position. The correlation dimension gives a single value for a range of scales s while the ergodicity defect gives a value at each scale s. In order to compare the two techniques, we average the defect values over a range of scales and we refer to this value as d_{mean}.

For our first flow example, we consider a quasiperiodic Duffing oscillator flow as an illustration of a double gyre flow. This flow is a 2D time-dependent fluid flow which satisfies (2) and is given by the equations

$$u = -y \tag{11}$$

$$v = -x - \varepsilon x(cos(\nu_1 t) + cos(\nu_2 t)) + \frac{x^3}{a^2} \tag{12}$$

where $a = 1$, $\nu_1 = \frac{3\pi}{2}$, $\nu_2 = \nu_1\frac{\sqrt{5}-1}{2}$, and $\varepsilon = 0.25$. The Duffing oscillator has two elliptic regions and one hyperbolic trajectory at the origin. There is a pair of stable and unstable manifolds emanating from the origin. The velocity field at $t = 0$ is given in the top panel of Fig. 2 with the stable manifold indicated by attracting arrows and the unstable manifold indicated by repelling arrows. In the bottom panel of Fig. 2, are grayscale results of computations of the correlation dimension c (left) and mean ergodicity defect d_{mean} (right) for the right half ($x > 0$) of the domain. By using trajectories computed in forward time with integration time $T_{int} = \frac{8\pi}{\nu_2}$, the stable manifold is correctly identified by the maximizing ridge of the ergodicity defect d_{mean} field and the minimizing ridge of the correlation dimension c-field. For reference, the stable manifold is also highlighted here with a dotted curve obtained from a direct evolution method. Similarly the unstable manifold can be identified using trajectories computed in backward time. Also, observe that the stationary center for the gyre is also detected by both methods as a compact solid color area with small c-values and large d_{mean} values.

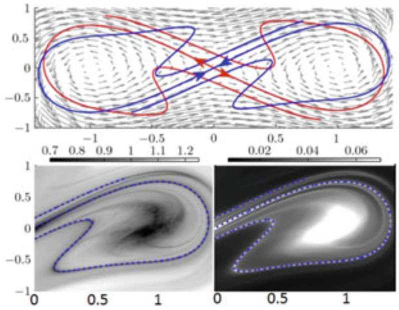


Fig. 2 Duffing oscillator flow; *bottom panel*: c on *left*, d_{mean} on *right*

In a numerical example with the flow field generated by the Regional Ocean Modeling System (ROMS) [26], see Fig. 3, the simulated eddy is captured by the correlation dimension (left) and ergodicity defect (right) as abrupt shade changes on both sides of the manifold. Similarly, the stationary center is detected as a cluster of small c-values and large d_{mean} values.

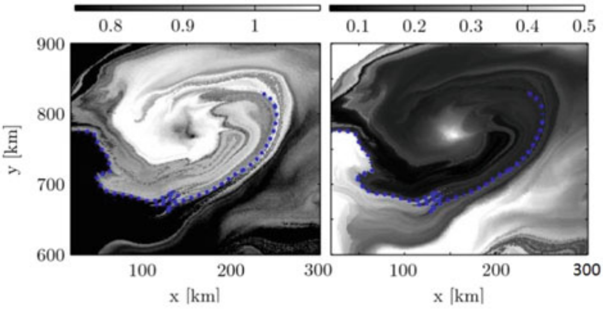


Fig. 3 Numerically generated ROMS flow c on *left* and d_{mean} on *right*

Finally, if we compare the ergodicity defect and the FTLE method for simulated trajectories that are randomly spaced, Fig. 4 suggests the possible advantages that the defect (on the left) may have over the commonly used FTLE method (middle and right panels). For example, the defect results are more clear and less noisy than the FTLE results [7, 25]. In this case, the FTLE computation of the separation rates is more difficult since the trajectories are not evenly spaced. This example illustrates the challenges of the FTLE method with data that is not numerically simulated but from in situ drifters, which do not move in a uniform arrangement or with drifter data that may be sparse in some areas. On the other hand, for each computation the ergodicity defect uses the individual trajectory and does not require trajectories in a uniform or particular arrangement.

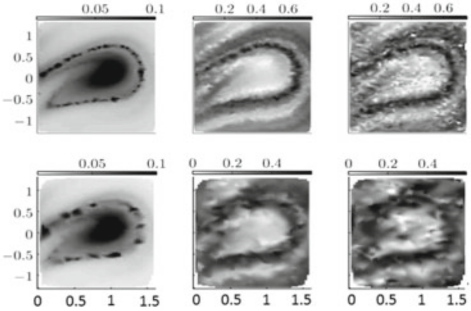


Fig. 4 LCS from randomly spaced trajectories—(*left panel*) d (*middle*) FTLE using the Lekien and Ross (2010) method (*right*) conventional FTL (darkest color = stable manifold)

4 Future Work, Other Fluid Flow Aspects

In this last section, we briefly explore some future 3D work for flows and a few other possible applications for the ergodicity defect as a diagnostic for signals and flows. So far we have only considered 2D examples, but a natural 3D analog of the Haar ergodicity defect for individual trajectories can be given as:

$$d_{\chi}(s;\mathbf{x}_0,t_0)=\sum_{j=1}^{2^{3s}}[(\frac{N_j(s)}{N}-\frac{1}{2^{3s}})^2], \tag{13}$$

where the domain is partitioned into cubes of side length s. As with the 2D case, time snapshots are taken of the flow, and in the 3D case, this is done for each initial fixed depth (z) level. With this 3D formula for the ergodicity defect, we aim to investigate the 3D structures in ocean flows. For example, in ongoing work with D. Rivas at CICESE on an upwelling flow off the coast of Oregon, preliminary results indicate a definite difference between the 2D and 3D defect analysis along the coast. These results suggest that the 3D defect reveals structure or behavior in the flow that the 2D defect misses and this agrees with the idea that there is significant vertical movement along the coast that produces more 3D sampling in that area than can be observed from only the 2D analysis. More and finer resolution work is in progress to investigate the underlying structures and their effects.

Other possible applications of the ergodicity defect can be used to inform (1) data assimilation techniques, (2) deployment or sampling strategies for data collecting devices and (3) optimal use of data for estimating flow properties. For example, the ergodicity defect could possibly be used as a diagnostic for determining when to stop assimilating data from the trajectory of a particular device and some preliminary results with E. Spiller at Marquette University using the particle filter data assimilation method indicate the feasibility of this idea. For deployment or sampling strategies, the idea is to use the ergodicity defect and similar measures to decide where to deploy drifters for optimal sampling of the flow, e.g., see [14]. The ergodicity defect as a diagnostic for how to best use the data for estimating flow properties can be illustrated by a toy double gyre example having a cold core and a warm core. Trajectories that get caught inside these cores give biased readings, e.g., too low or too high, however using the defect to distinguish these trajectories and properly weigh their measurements can help offset this biased estimation of the temperature from the data, see [23].

5 Summary

Drawing from ergodic and wavelet theory, the ergodicity defect technique captures the complexity of flows and distinguishes their trajectories in terms of deviation from ergodicity. The ergodicity defect measurements of a flow's trajectories are used to identify the Lagrangian coherent structures in the flow. These structures impact the flow's overall behavior especially with regards to the transport and exchange of materials. The ergodicity defect allows for a choice of analyzing functions to be used and because each computation only requires an individual trajectory the defect appears to be more amenable to realistic, nonuniform, and sparse data. In this manner, the erogodicity defect offers advantages in a practical or applications setting. The ergodicity technique can also serve as a diagnostic tool in a variety of other applications in geosciences related problems.

References

1. Aref H. Stirring by chaotic advection. J Fluid Mech. 1984;143:1–21.
2. Birkhoff G. Proof of the ergodic theorem. Proc Natl Acad Sci. 1931b;17:656–60.
3. Daubechies I. Ten Lectures on Wavelets CBMS-NSF Reg. Conf. Ser. Appl. Math. Soc. Ind. Appl. Math, Philadelphia; 1992.
4. Debnath, L. Wavelet transforms and their applications. Boston: Birkhauser; 2002.
5. Froyland G, Padberg K, England M, Treguier A. Detection of coherent oceanic structures via transfer operators. Phys Rev Lett. 2007;98:1–4.
6. Grassberger P, Procaccia I. Measure the strangeness of strange attractors. Phys D. 1983;9:189–208.
7. Haller G, Sapsis T. Lagrangian coherent structures and the smallest finite-time Lyapunov exponent. Chaos. 2011;2:1–7.
8. Kuznetsov L, Jones C, Toner M, Kirwan AD Jr. Assessing coherent feature kinematics in ocean models. Phys D. 2004;191:81–105.
9. Lasota A, Mackey M. Chaos, fractals and noise. Applied Mathematical Science 97. Berlin:Springer; 1994.
10. Lekien F, Shadden SC, Marsden JE. Lagrangian coherent structure in *n*-dimensional systems. J Math Phys. 2007;48:1–19.
11. Madrid JA, Mancho AM. Distinguished trajectories in time dependent vector fields. Chaos. 2009;19:1–18.
12. Malhotra N, Mezić I, Wiggins S. Patchiness: a new diagnostic for Lagrangian trajectory analysis in fluid flows. J Bifurc Chaos. 1998;8:1053–94.
13. Mane R. Ergodic theory and differentiable dynamics. New York: Springer-Verlag; 1987.

14. Mathew G, Mezić I. Metrics for ergodicity and design of ergodic dynamics for multi-agent systems. Phys D. 2011;240:432–42.
15. Mathew G, Mezić I, Petzold L. A multiscale measure for mixing and its applications. Phys D. 2005;211:23–46.
16. Mezić I. Caltech PhD Thesis; 1994
17. Mezić I, Wiggins S. A method for visualization of invariant sets of dynamical systems based on the ergodic partition. Chaos. 1999;9:213–8.
18. Oliver M. Special issue on applications of wavelets in the geosciences. Math Geosci. 2009;41:609–10.
19. Petersen K. Ergodic theory. Cambridge: Cambridge University Press; 1983.
20. Poje AC, Haller G, Mezić I. The geometry and statistics of mixing in aperiodic flows. Phys Fluids. 1999;11:2963–8.
21. Rypina I, Scott S, Pratt LJ, Brown MG. Investigating the connection between complexity of isolated trajectories and Lagrangian coherent structures. Nonlinear Proc Geophys. 2011;18:977–87.
22. Scott SE. Different perspectives and formulas for capturing deviation from ergodicity. SIAM J Appl Dyn Syst. 2014;124:1889–947.
23. Scott SE. Characterizing and capturing ergodic properties of random processes with a focus on geoscience applications. Submitted to IEEE Transactions on Information Theory
24. Scott SE, Redd C, Kutznetsov L, Mezić I, Jones C. Capturing deviation from ergodicity at different scales. Phys D Nonlinear Phenom. 2009;238:1668–79.
25. Shadden SC, Lekien F, Marsden JE. Definition and properties of Lagrangian coherent structures from finite-time Lyapunov exponents in two-dimensional aperiodic flows. Phys D. 2005;212:271–304.
26. Shchepetkin AF, McWilliams JC. The regional ocean modeling system: a split-explicit, free-surface, topography following coordinates ocean model. Ocean Model. 2005;9:347–404.
27. Sturman R, Ottino J, Wiggins S. The mathematical foundations of mixing. Cambridge: Cambridge University Press; 2006.
28. Tupper PF. Ergodicity and the numerical simulation of Hamiltonian systems. Submitted
29. Walnut D. An introduction to wavelet analysis. Boston: Birkhauser; 2002.
30. Walters P. Introduction to ergodic theory. New York: Springer; 1982.

Part XI
Gabor Theory

Part XI is devoted to three fundamentals, one might say elementary-to-state, and devilishly difficult problems in Gabor or time-frequency analysis.

In the first chapter of this part, HEIL AND SPEEGLE give the current state of the HRT conjecture. The conjecture was posed by Heil, Ramanathan, and Topiwala in 1996, and, as noted by Bourgain, working on HRT could become "addictive." The HRT conjecture asserts that finite Gabor systems in $L^2(\mathbb{R})$ are linearly independent. What could be more straightforward than verifying (or not) that a finite set, $\{M_bT_ag\}$, of translates T_a and modulates $M_b(x) = e^{2\pi ibx}$ of a given $g \in L^2(\mathbb{R}) \setminus \{0\}$, is linearly independent! In fact, before getting into the subtleties of HRT, the authors tease the reader by pointing out that finite wavelet systems in $L^2(\mathbb{R})$ are often linearly dependent. Along with the subtleties of HRT, the authors give an expert exposition of known results about HRT. This is all fascinating, and the addiction continues.

Actually, a significant portion of the chapter (Sect. 5) deals with the "relation (or lack thereof)" between the HRT conjecture and the comparably tantilizing Zero Divisor Conjecture.

As basic as the HRT conjecture is, in the next chapter, SALIANI begins with a comparably innocent looking topic: the study of linear independence of $\{T_k\psi_j : k \in \mathbb{Z}, j = 1,...,m\}$. Intricacies arise almost immediately. First if $\psi \in L^2(\mathbb{R})$, then $\{T_k\psi\}$ is linearly independent in $L^2(\mathbb{R})$. Then there is a big step forward with the theorem: $\{T_k\psi\}$ is ℓ^2-linearly independent if and only if the Fourier transform periodicity function,

$$p_\psi(\xi) = \sum_{k\in\mathbb{Z}} |\widehat{\psi}(\xi - k)|^2,$$

is positive a.e. The "only if" part is based on a theorem by Kislyakov, whose proof relies on a theorem of Vinogradov. And at the heart of it is the celebrated theorem of Menchoff (1942): Every measurable function becomes a function with uniformly convergent Fourier series after a modification on a set of arbitrary small measure. Saliani has established herself as an international authority in this area; and her chapter exhibits this expertise in her description of her own results as well as her perspective on some recent contributions by others.

Besides the deepest results about principal shift invariant subspaces (one ψ), a significant part of her chapter is devoted to analogous problems in dealing with finite sets, $\{\psi_1,...,\psi_m\}$. It is a marvelous state of the art exposition.

In the final chapter of this part, DAI AND SUN outline their spectacular solution of the *abc*-problem for Gabor frames. This problem is not to be confused with the *abc*-conjecture in arithemetic-geometry, that deals with finding integer or rational solutions to multivariable polynomial equations. Dai and Sun consider the Gabor system,

$$\mathcal{G} = \mathcal{G}(\chi_{[0,c)}, a\mathbb{Z} \times b\mathbb{Z}) = \{e^{-2\pi inbt}\chi_{[0,c)}(t - ma) : (m,n) \in \mathbb{Z} \times \mathbb{Z}\}.$$

The *abc*-problem is to characterize the triples (a,b,c) of positive numbers a,b,c for which $\mathcal{G}$ is a frame for $L^2(\mathbb{R})$. Dai and Sun have solved this problem! Their solution is not for the faint of heart, *but* their presentation herein is an ideal summary and outline of their systematic high-level series of steps. As a preliminary step, the problem can be reformulated to deal only with a and c. The analysis begins with a characterization of all pairs (a,c) in terms of infinite matrices depending on said pairs and used by others to address the problem. And then the fun begins going from the infinite matrix characterization to the specific values of a and c for which the Gabor system is a frame. The ride is intricate and worth the price! An exciting epilogue is a new sampling theorem depending on these ideas.

The HRT Conjecture and the Zero Divisor Conjecture for the Heisenberg Group

Christopher Heil and Darrin Speegle*

Abstract This chapter reports on the current status of the HRT Conjecture (also known as the linear independence of time–frequency shifts conjecture), and discusses its relationship with a longstanding conjecture in algebra known as the zero divisor conjecture.

Keywords Gabor systems · Heisenberg group · HRT Conjecture · Indicable group · Linear Independence of Time–Frequency Shifts Conjecture · Time–frequency analysis · Wavelet systems · Zero Divisor Conjecture

1 Introduction

The building blocks of Gabor and wavelet systems are *translations* (or *time shifts*):

$$T_a g(x) \;=\; g(x-a), \qquad \text{where } a \in \mathbb{R};$$

modulations (or *frequency shifts*):

$$M_b g(x) \;=\; e^{2\pi i b x}\, g(x), \qquad \text{where } b \in \mathbb{R};$$

C. Heil (✉)
School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA
e-mail: heil@math.gatech.edu

D. Speegle
Department of Mathematics, Saint Louis University St. Louis, MO 63103, USA
e-mail: speegled@slu.edu

* This work was partially supported by grants from the Simons Foundation (#229035 Heil; #244953 Speegle).

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_7

and *dilations*:

$$D_r g(x) = r^{1/2} g(rx), \qquad \text{where } r > 0.$$

Gabor systems employ the compositions

$$M_b T_a g(x) = e^{2\pi i b x} g(x-a),$$

which are called *time–frequency shifts*, while wavelet systems use the compositions

$$D_r T_a g(x) = r^{1/2} g(rx-a),$$

which are called *time-scale shifts*.

Explicitly, a Gabor system has the form

$$\mathscr{G}(g,\Lambda) = \{M_b T_a g\}_{(a,b)\in\Lambda},$$

where Λ is an index set contained in $\mathbb{R}^2$, while a wavelet system is

$$\mathscr{W}(g,\Gamma) = \{D_r T_a g\}_{(a,r)\in\Gamma},$$

where $\Gamma \subset \mathbb{R} \times (0,\infty)$. Typically we are interested in constructing Gabor systems or wavelet systems that are bases or frames for $L^2(\mathbb{R})$. In this case Λ or Γ will be countable index sets. However, the "local" properties of Gabor and wavelet systems (by which we mean the properties of finite subsets) are often of key interest. In particular, given any set of vectors in a vector space, one of the first and most fundamental questions we can ask about this set is whether it is finitely linearly independent, i.e., whether every finite subset of the collection is linearly independent. It is known that there exist nontrivial functions $g \in L^2(\mathbb{R})$ and finite sets Γ such that the finite wavelet system $\mathscr{W}(g,\Gamma)$ is linearly dependent. It is *not known* whether there exist nontrivial functions $g \in L^2(\mathbb{R})$ and finite sets Λ such that the finite Gabor system $\mathscr{G}(g,\Lambda)$ is linearly dependent. The HRT Conjecture is that finite Gabor systems are indeed linearly independent.

In this chapter we will give a short report on the current status of the HRT conjecture, and also comment on its relation to a longstanding conjecture in algebra known as the *Zero Divisor Conjecture*. We begin in Section 2 with some examples that illustrate that finite wavelet systems can be linearly dependent. We formulate the HRT Conjecture in Section 3, and in Section 4 we review some of the main partial results related to linear independence of time–frequency shifts that are currently known. Finally, in Section 5 we discuss the relationship between the HRT conjecture and the zero divisor conjecture.

2 Linear Dependence of Time-Scale Shifts

The fact that wavelet systems *can be linearly dependent* is the starting point for the construction of compactly supported wavelet bases through a multiresolution analysis. The first step in the construction of a compactly supported wavelet via this method is actually the construction of a compactly supported *scaling function*. A scaling function is a function $\varphi \in L^2(\mathbb{R})$ that satisfies a *refinement equation* of the form

$$\varphi(x) \;=\; \sum_{k=0}^{N} c_k\, \varphi(2x-k)$$

and additionally is such that $\{\varphi(x-k)\}_{k\in\mathbb{Z}}$ is an orthonormal system in $L^2(\mathbb{R})$. While the integer translates of a scaling function are orthonormal, and hence linearly independent, if we rewrite the refinement equation as

$$\varphi \;=\; \sum_{k=0}^{N} 2^{-1/2} c_k D_2 T_k \varphi,$$

then we see that the refinement equation is a statement that the collection of time-scale shifts

$$\{D_r T_a \varphi\}_{(a,r)\in\Gamma}$$

with

$$\Gamma \;=\; \{(0,1)\} \,\cup\, \{(k,2) : k = 0,\dots,N\}$$

is linearly dependent. Here is an example.

Example 1. The *Haar wavelet* is

$$\psi \;=\; \chi_{[0,\frac{1}{2}]} \;-\; \chi_{[\frac{1}{2},1]}.$$

Haar proved directly in [13] that

$$\bigl\{D_{2^n} T_k \psi\bigr\}_{k,n\in\mathbb{Z}},$$

which today we call the *Haar system*, is an orthonormal basis for $L^2(\mathbb{R})$. Of course, since the Haar system is a collection of orthonormal functions, it is finitely linearly independent. However, if we wish to construct the Haar system using the modern framework of multiresolution analysis, we first begin by constructing the corresponding scaling function. For the Haar system, the scaling function is the *box function*

$$\varphi \;=\; \chi_{[0,1]}.$$

This function satisfies the refinement equation

$$\varphi(x) \;=\; \varphi(2x) \;+\; \varphi(2x-1).$$

Consequently, the set of three functions

$$\bigl\{\varphi, D_2\varphi, D_2T_1\varphi\bigr\}$$

is linearly dependent. Once we have the scaling function, the machinery of multiresolution analysis tells us that the wavelet

$$\psi(x) \;=\; \varphi(2x) \;-\; \varphi(2x-1)$$

can be used to generate an orthonormal basis for $L^2(\mathbb{R})$.

For a detailed description of what a multiresolution analysis is and how it leads to a wavelet orthonormal basis, we refer to [15, Chap. 12] or [7].

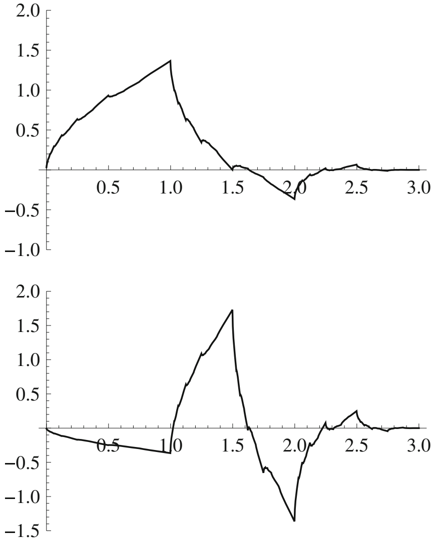


Fig. 1 The Daubechies $\mathbf{D}_4$ scaling function (*top*), and the corresponding wavelet $\mathbf{W}_4$ (*bottom*)

Here is a modern example of a compactly supported scaling function.

Example 2. The *Daubechies* $\mathbf{D}_4$ *scaling function* is the function that satisfies the refinement equation

$$\begin{aligned} \mathbf{D}_4(x) &= \tfrac{1+\sqrt{3}}{4}\,\mathbf{D}_4(2x) + \tfrac{3+\sqrt{3}}{4}\,\mathbf{D}_4(2x-1) \\ &\qquad + \tfrac{3-\sqrt{3}}{4}\,\mathbf{D}_4(2x-2) + \tfrac{1-\sqrt{3}}{4}\,\mathbf{D}_4(2x-3). \end{aligned} \tag{1}$$

It can be shown that there is a unique (up to scale) compactly supported function $\mathbf{D}_4 \in L^2(\mathbb{R})$ that satisfies this refinement equation. This function, which is continuous and supported in the interval $[0,3]$ is illustrated in Fig. 1. It is not obvious from the picture, but it is true that the integer translates $\{\mathbf{D}_4(x-k)\}_{k\in\mathbb{Z}}$ are orthonormal. The scaling function $\mathbf{D}_4$ generates a multiresolution analysis, and because of this it follows that the wavelet

$$\begin{aligned} \mathbf{W}_4(x) &= \tfrac{1-\sqrt{3}}{4}\,\mathbf{D}_4(2x) - \tfrac{3-\sqrt{3}}{4}\,\mathbf{D}_4(2x-1) \\ &\qquad + \tfrac{3+\sqrt{3}}{4}\,\mathbf{D}_4(2x-2) - \tfrac{1+\sqrt{3}}{4}\,\mathbf{D}_4(2x-3) \end{aligned}$$

can be used to generate an orthonormal basis for $L^2(\mathbb{R})$. Specifically, the wavelet system

$$\{D_{2^n} T_k \mathbf{W}_4\}_{k,n\in\mathbb{Z}}$$

is an orthonormal basis for $L^2(\mathbb{R})$. However, the point we are making here is that the refinement Eq. (1) implies that the finite collection of time-scale shifts

$$\{\mathbf{D}_4, D_2\mathbf{D}_4, D_2T_1\mathbf{D}_4, D_2T_2\mathbf{D}_4, D_2T_3\mathbf{D}_4\}$$

is *linearly dependent*.

3 Gabor Systems and the HRT Conjecture

Now we turn to Gabor systems. We will need to employ the Fourier transform, which for an integrable function g we normalize as

$$\widehat{g}(\xi) \;=\; \int_{-\infty}^{\infty} g(x)\, e^{-2\pi i \xi x}\, dx.$$

The Fourier transform extends to a unitary operator that maps $L^2(\mathbb{R})$ onto itself. We have the following relations between translation, modulation, and the Fourier transform:

$$(T_a g)^\wedge \;=\; M_{-a}\widehat{g} \qquad \text{and} \qquad (M_b g)^\wedge(\xi) \;=\; T_b\widehat{g}.$$

Although translations and modulations do not commute, we have

$$M_bT_ag(x) \;=\; e^{2\pi i bx} g(x-a) \qquad \text{and} \qquad T_aM_bg(x) \;=\; e^{2\pi i b(x-a)} g(x-a),$$

and therefore

$$T_aM_bg \;=\; e^{-2\pi i ab} M_bT_ag. \tag{2}$$

We can easily show that if we only consider translations alone, then we will always obtain linearly independent collections. To see why, let $g \in L^2(\mathbb{R})$ be nontrivial (not zero a.e.) and let $a_1,\dots,a_N$ be any distinct points in $\mathbb{R}$. If

$$\sum_{k=1}^{N} c_k\, g(x-a_k) \;=\; 0 \text{ a.e.},$$

then by applying the Fourier transform we obtain

$$\sum_{k=1}^{N} c_k\, e^{-2\pi i a_k \xi}\, \widehat{g}(\xi) \;=\; 0 \text{ a.e.}$$

However, $\widehat{g}$ is not the zero function, so this implies that

$$\sum_{k=1}^{N} c_k\, e^{-2\pi i a_k \xi} \;=\; 0 \text{ a.e.}$$

Multiplying through by $e^{2\pi i a_1 \xi}$ we obtain

$$m(\xi) \;=\; c_1 \;+\; \sum_{k=2}^{N} c_k\, e^{-2\pi i (a_k - a_1)\xi} \;=\; 0 \text{ a.e.}$$

But m is a nonharmonic trigonometric polynomial, and therefore can be extended to an analytic function on the complex plane. An analytic function cannot vanish on any set that has an accumulation point without being identically zero. Hence $m(\xi) = 0$ for every ξ, which implies that $c_1 = 0$. Iterating, we obtain $c_2 = \cdots = c_N = 0$.

Similarly, any finite set of modulations of g is linearly independent. However, as soon as we combine translations with modulations, the situation becomes much less clear. In particular, if we fix a finite Gabor system

$$\bigl\{M_{b_k} T_{a_k} g : k = 1, \dots, N\bigr\}$$

and we assume that

$$\sum_{k=1}^{N} c_k\, M_{b_k} T_{a_k} g \;=\; 0 \text{ a.e.}, \tag{3}$$

then all that we obtain by applying the Fourier transform is that

$$\sum_{k=1}^{N} c_k\, T_{b_k} M_{-a_k} \widehat{g} \;=\; 0 \text{ a.e.}$$

In view of the commutation relation in Eq. (2), we can rewrite this as

$$\sum_{k=1}^{N} c_k e^{2\pi i a_k b_k}\, M_{-a_k} T_{b_k} \widehat{g} \;=\; 0 \text{ a.e.},$$

which is an equation of exactly the same nature as Eq. (3). The Fourier transform yields no simplification here.

At first glance, this may seem to be only a minor stumbling block—surely another approach, perhaps another transform, will show that Gabor systems are finitely linearly independent. Yet this most basic question about a set of vectors in $L^2(\mathbb{R})$ remains unanswered today. The following conjecture, today known as the *Linear Independence of Time–Frequency Translates Conjecture* or the *HRT Conjecture*, first appeared in print in [18].

Conjecture 1 (HRT Conjecture). If g in $L^2(\mathbb{R})$ is not the zero function and $\Lambda = \bigl\{(a_k, b_k) : k = 1, \dots, N\bigr\}$ is a set of finitely many distinct points in $\mathbb{R}^2$, then

$$\mathscr{G}(g, \Lambda) \;=\; \bigl\{M_{b_k} T_{a_k} g : k = 1, \dots, N\bigr\}$$

is linearly independent.

Various partial results on this conjecture are known (we will discuss these in Section 4). However, even the following very restricted version of the conjecture is open as of the time of writing. Here $\mathscr{S}(\mathbb{R})$ denotes the Schwartz class of all infinitely differentiable functions which, along with all of their derivatives, have faster than polynomial decay at infinity.

Conjecture 2 (HRT Subconjecture). If $g \in \mathscr{S}(\mathbb{R})\backslash\{0\}$, then

$$\bigl\{g(x),\, g(x-1),\, e^{2\pi ix}g(x),\, e^{2\pi i\sqrt{2}x}g(x-\sqrt{2})\bigr\} \tag{4}$$

is linearly independent.

We observe that the set of functions that appears in Eq. (4) is the Gabor system $\mathscr{G}(g,\Lambda)$ where

$$\Lambda \;=\; \bigl\{(0,0),\, (1,0),\, (0,1),\, (\sqrt{2},\sqrt{2})\bigr\}. \tag{5}$$

4 Partial Results

We will briefly list some of the main partial results that are currently available on the HRT Conjecture. These are approximately ordered chronologically, but due to vastly differing time lengths from research to publication, this particular ordering should not be interpreted as anything other than a convenience for presentation purposes. Further, no attempt has been made to state all partial results from every paper, nor to state them precisely; this list simply serves as a brief summary of the literature on Conjecture 1. A few of the results below extend to systems in $L^2(\mathbb{R}^d)$, but for the most part there are substantial obstacles in moving to higher dimensions. For a discussion of why this is so, we refer to the survey paper [14] (that paper also presents context, motivation, and related results that are not discussed here).

The paper [18] which originally presented the HRT conjecture included the following results.

- $\mathscr{G}(g,\Lambda)$ is independent if g is compactly supported, or just supported within a half-line $[a,\infty)$ or $(-\infty,a]$.
- $\mathscr{G}(g,\Lambda)$ is independent if $g(x) = p(x)e^{-x^2}$, where p is a polynomial.
- $\mathscr{G}(g,\Lambda)$ is independent if $N \leq 3$.
- If A is a 2×2 invertible matrix with $|\det(A)| = 1$, and $z \in \mathbb{R}^2$, then $\mathscr{G}(g,\Lambda)$ is independent for all nontrivial g if and only if $\mathscr{G}(g,A(\Lambda)+z)$ is independent for all nontrivial g.
- If $\mathscr{G}(g,\Lambda)$ is independent, then there exists an $\varepsilon > 0$ such that $\mathscr{G}(g,\Lambda')$ is independent for all $\Lambda' = \bigl\{(a'_k,b'_k) : k = 1,\ldots,N\bigr\}$ such that $|a_k - a'_k| < \varepsilon$ and $|b_k - b'_k| < \varepsilon$ for $k = 1,\ldots,N$.

- If $\mathscr{G}(g,\Lambda)$ is independent, then there exists an $\varepsilon > 0$ such that $\mathscr{G}(h,\Lambda)$ is independent for all $\|h-g\|_2 < \varepsilon$.

Since the set of compactly supported functions is dense in $L^2(\mathbb{R})$, if Λ is a fixed finite set then by applying the final perturbation result listed above it follows that there exists an open, dense subset $\mathscr{U}$ of $L^2(\mathbb{R})$ such that $\mathscr{G}(g,\Lambda)$ is linearly independent for all $g \in \mathscr{U}$.

Linnell proved in [22] that

- $\mathscr{G}(g,\Lambda)$ is independent if Λ is a finite subset of a translate of a full-rank lattice in $\mathbb{R}^2$, i.e., if $\Lambda \subset A(\mathbb{Z}^2)+z$ where A is an invertible 2×2 matrix (with any nonzero determinant) and $z \in \mathbb{R}^2$.

In particular, any set of three points in $\mathbb{R}^2$ is contained in a translate of a full-rank lattice, so this gives another proof that $\mathscr{G}(g,\Lambda)$ is independent if $N \leq 3$. However, four points in $\mathbb{R}^2$ need not be contained in a translate of a full-rank lattice. In particular, the set of points Λ given in Eq. (5) is not contained in any translate of any full-rank lattice.

In [6], Christensen and Lindner obtain

- estimates of the frame bounds of a finite Gabor system $\mathscr{G}(g,\Lambda)$.

In principle, sufficiently strong estimates of the frame bounds of finite Gabor systems could be combined with the perturbation theorems in [18] to yield a proof of the HRT conjecture. Unfortunately, the results of [6] do not seem to be conducive for advancing this type of approach. A short introduction to frame theory can be found in [16].

Rzeszotnik, in an unpublished work, proved that

- $\mathscr{G}(g,\Lambda)$ is independent if

$$\Lambda = \bigl\{(0,0),\, (1,0),\, (0,1),\, (\sqrt{2},0)\bigr\}.$$

This set Λ, like the one given in Eq. (5), is not contained in any translate of a full-rank lattice in $\mathbb{R}^2$. On the other hand, the points in this set Λ lie on two parallel lines, while the four points in Eq. (5) do not.

An equivalent formulation of the HRT conjecture is that if $c_1,\dots,c_N$ are not all zero, then the kernel of the operator

$$T = \sum_{k=1}^{N} c_k M_{b_k} T_{a_k}$$

is $\{0\}$, or, in other words, 0 is not an eigenvalue of T. This suggests that it would be interesting to investigate the spectrum of such a linear combination of time–frequency shift operators. Now, if T had a nonzero eigenvalue λ, then for some nonzero function g we would have

$$Tg = \sum_{k=1}^{N} c_k M_{b_k} T_{a_k} g = \lambda g.$$

Setting $a_0 = b_0 = 0$ and $c_0 = -\lambda$, it follows that

$$\sum_{k=0}^{N} c_k M_{b_k} T_{a_k} g \;=\; 0,$$

and hence 0 is an eigenvalue of the operator $S = \sum_{k=0}^{N} c_k M_{b_k} T_{a_k}$, which is simply another finite linear combination of time–frequency shift operators. Consequently, we can restate the HRT conjecture in terms of eigenvalues as follows.

Conjecture 3 (Spectral Version of the HRT Conjecture). If

$$\Lambda \;=\; \{(a_k, b_k) : k = 1, \dots, N\}$$

is a set of finitely many distinct points in $\mathbb{R}^2$ and $c_1, \dots, c_N$ are not all zero, then the point spectrum of

$$T \;=\; \sum_{k=1}^{N} c_k M_{b_k} T_{a_k}$$

is empty (i.e., T has no eigenvalues).

Balan proved in [1] that

- the operator T cannot have an isolated eigenvalue of finite multiplicity.

Consequently, *if* the point spectrum of T is not empty, then it can only contain eigenvalues of infinite multiplicity, or eigenvalues that also belong to the continuous spectrum of T. Further results on the spectral properties of T have been obtained by Balan and Krishtal in [2]. In particular, a consequence of the results of that paper is that if Λ is the set of four points specified in Eq. (5), then the corresponding operator T has no isolated eigenvalues.

Linnell's proof that $\mathscr{G}(g,\Lambda)$ is linearly independent if Λ is a finite subset of a shift of a full-rank lattice was obtained through the machinery of von Neumann algebras. In [4],

- a different proof of Linnell's result was derived through time–frequency methods.

Interestingly, Demeter and Gautam obtained in [9]

- yet another proof of Linnell's result, this time based on the spectral theory of random Schrödinger operators.

Demeter in [8], and Demeter and Zaharescu in [10], focused on the case where Λ contains four distinct points. These two papers prove that

- $\mathscr{G}(g,\Lambda)$ is independent if $\#\Lambda = 4$ and Λ is a subset of two parallel lines in $\mathbb{R}^2$.

In particular, this recovers the result by Rzeszotnik that was stated earlier.

One of the results obtained in [18] is that $\mathscr{G}(g,\Lambda)$ is linearly independent if g has the particular form $g(x) = p(x)e^{-x^2}$ where p is a polynomial. No other results related to decay conditions seem to have been obtained until quite recently. In [5] it is proved that

- $\mathscr{G}(g,\Lambda)$ is linearly independent if

$$\lim_{x\to\infty} |g(x)|\, e^{cx^2} \;=\; 0 \;\text{ for all } c>0,$$

and

- $\mathscr{G}(g,\Lambda)$ is linearly independent if

$$\lim_{x\to\infty} |g(x)|\, e^{cx\log x} \;=\; 0 \;\text{ for all } c>0.$$

However, the case where $|g(x)|\,e^{cx} \to 0$ for all $c>0$ remains open. Note that while functions in the Schwartz class $\mathscr{S}(\mathbb{R})$ have extremely rapid decay, they need not decay at an exponential rate.

Benedetto and Bourouihiya also obtained partial results related to decay. They proved in their paper [3] that

- $\mathscr{G}(g,\Lambda)$ is linearly independent if g is ultimately positive and $b_1,\dots,b_N$ are independent over $\mathbb{Q}$,

and

- $\mathscr{G}(g,\Lambda)$ is linearly independent if $\#\Lambda = 4$, g is ultimately positive, and $g(x)$ and $g(-x)$ are ultimately decreasing.

The most recent paper related to the HRT Conjecture of which we are aware is [12] by Gröchenig. He proves that

- the lower Riesz bound of a finite section of a Gabor frame that is not a Riesz basis converges to zero, and in many cases this convergence is super-fast.

A Gabor frame that is not a Riesz basis is “globally redundant” in some sense. Even so, if the HRT Conjecture is true then every finite subset of such a frame must be linearly independent. Gröchenig’s result implies that, from a numerical point of view, such finite subsets rapidly become “nearly dependent” as their size increases. To quote Gröchenig, this “illustrates the spectacular difference between a conjectured mathematical truth and a computationally observable truth.”

Finally, although they do not obtain results directly about Conjecture 1, we mention that the papers of Kutyniok [20] and Rosenblatt [24] consider some generalizations of the conjecture.

5 The Zero Divisor Conjecture

In this section we discuss the relation (or lack thereof) between the HRT conjecture and the zero divisor conjecture.

First we review some terminology. Let G be a group (with the group operation written multiplicatively). The *complex group algebra* of G is the set of all formal finite linear combinations of elements of G. We write this as

$$\mathbb{C}G = \left\{ \sum_{g\in G} c_g g : c_g \in \mathbb{C} \text{ with only finitely many } c_g \neq 0 \right\}.$$

The natural operation of addition in $\mathbb{C}G$ is defined by

$$\sum_{g\in G} c_g g + \sum_{g\in G} d_g g = \sum_{g\in G} (c_g + d_g) g,$$

and we define multiplication in $\mathbb{C}G$ by

$$\left(\sum_{g\in G} c_g g\right)\left(\sum_{h\in G} d_h g\right) = \sum_{g\in G}\sum_{h\in G} c_g d_h gh = \sum_{g\in G}\left(\sum_{h\in G} c_{gh^{-1}} d_h\right) g.$$

All of the above sums are well-defined since only finitely many terms in any sum are nonzero. More generally, the field $\mathbb{C}$ can be replaced by other fields, but we will restrict our attention here to the complex field.

Suppose that g is an element of G that has finite order $n > 1$. If we let e denote the identity element of G and set

$$\alpha = g - e \qquad \text{and} \qquad \beta = g^{n-1} + \cdots + g + e,$$

then

$$\begin{aligned} \alpha\beta &= (g-e)(g^{n-1} + \cdots + g + e) \\ &= (g^n + \cdots + g^2 + g) - (g^{n-1} + \cdots + g + e) \\ &= g^n - e \\ &= 0. \end{aligned}$$

Thus, if G has any nontrivial elements of finite order then $\mathbb{C}G$ has zero divisors. What happens if there are no nontrivial elements of G that have finite order? In general the answer is unknown, and this is the context of the following *Zero Divisor Conjecture*.

Conjecture 4 (Zero Divisor Conjecture). Let G be a torsion-free group (i.e., G contains no elements of finite order other than the identity). If $\alpha, \beta \in \mathbb{C}G$, then

$$\alpha \neq 0 \text{ and } \beta \neq 0 \implies \alpha\beta \neq 0. \qquad \diamondsuit$$

This conjecture is sometimes attributed to Kaplansky (for example, this is the attribution stated by Wikipedia in the article "Group Ring"). Variants of the conjecture seem to have appeared in the literature over time. Higman proved one version of the zero divisor conjecture for "locally indicable" groups in 1940 [19]. Two surveys of the conjecture published in 1977 are the paper [25] by Snider and Chapter 13 of Passman's text [23]. According to Snider, Higman's result was "essentially all that was known until 1974, when Formanek proved the conjecture for supersolvable groups" (see [11]). Since then various results have been obtained, but the conjecture remains open in the generality stated. A survey by Linnell of *analytic versions of the zero divisor conjecture* can be found in [21].

There is a natural group associated with time–frequency analysis, the *Heisenberg group* $\mathbf{H}$, and therefore we can consider the zero divisor conjecture for the special case that $G = \mathbf{H}$. There are several versions of the Heisenberg group and many isomorphic definitions of each of these. For our purposes, it is simplest to consider the *reduced Heisenberg group* defined by

$$\mathbf{H} \;=\; \bigl\{zM_bT_a : z \in \mathbb{T}, a, b \in \mathbb{R}\bigr\},$$

where $\mathbb{T}$ is the unit circle in the complex plane, i.e.,

$$\mathbb{T} \;=\; \{z \in \mathbb{C} : |z| = 1\}.$$

In other words, with this definition $\mathbf{H}$ is the set of all unit modulus scalar multiples of time–frequency operators. The group operation is simply composition of operators, which by Eq. (2) follows the rule

$$(zM_bT_a)\,(wM_dT_c) \;=\; zwe^{-2\pi iad}M_{b+d}T_{a+c}.$$

The Heisenberg group is noncommutative (but even so, as a locally compact group it turns out that left and right Haar measure on $\mathbf{H}$ coincide, and therefore $\mathbf{H}$ is *unimodular*). If we fix any element zM_bT_a of $\mathbf{H}$, then the nth power of this element has the form

$$(zM_bT_a)^n \;=\; w_nM_{nb}T_{na},$$

where w_n is a scalar that has unit modulus. Therefore no element of $\mathbf{H}$ other than the identity $I = M_0T_0$ has finite order. Consequently, $\mathbf{H}$ is torsion-free.

Since $\mathbf{H}$ consists of scalar multiples of time–frequency shift operators, its group algebra $\mathbb{C}\mathbf{H}$ is the vector space of all finite linear combinations of time–frequency shift operators:

$$\mathbb{C}\mathbf{H} \;=\; \left\{\sum_{k=1}^{N} c_kM_{b_k}T_{a_k} : N > 0, c_k \in \mathbb{C}, a_k \in \mathbb{R}, b_k \in \mathbb{R}\right\}.$$

The question we wish to answer is whether $\mathbb{C}\mathbf{H}$ has any zero divisors. So, suppose that α and β are nonzero elements of $\mathbb{C}\mathbf{H}$ such that $\alpha\beta = 0$. Since α and β belong to $\mathbb{C}\mathbf{H}$, we can write

$$\alpha \;=\; \sum_{j=1}^{M} z_jM_{b_j}T_{a_j} \qquad \text{and} \qquad \beta \;=\; \sum_{k=1}^{N} w_kM_{d_k}T_{c_k},$$

with $(a_j,b_j) \neq (a_{j'},b_{j'})$ whenever $j \neq j'$, and similarly for (c_k,d_k). Since α and β are each nonzero, we can assume without loss of generality that z_j and w_k are nonzero complex scalars for every j and k. For simplicity of notation, let t_{jk} be the nonzero scalar

$$t_{jk} \;=\; z_j w_k e^{-2\pi i a_j d_k}.$$

Then we have

$$\alpha\beta \;=\; \sum_{j=1}^{M}\sum_{k=1}^{N} t_{jk} M_{b_j+d_k} T_{a_j+c_k} \;=\; 0. \tag{6}$$

Many of the values of a_j+c_k or b_j+d_k in Eq. (6) may coincide. For convenience, and without loss of generality, we can assume that the a_j and c_k are ordered (possibly with duplicates). That is, we can assume that

$$a_1 \le a_2 \le \cdots \le a_M \qquad \text{and} \qquad c_1 \le c_2 \le \cdots \le c_N.$$

Let

$$I \;=\; \big\{(j,k) : a_j+c_k = a_M+c_N\big\}.$$

Since $\alpha\beta = 0$, for every $f \in L^2(\mathbb{R})$ we have

$$\sum_{(j,k)\notin I} t_{jk} M_{b_j+d_k} T_{a_j+c_k} f(x) \;+\; \sum_{(j,k)\in I} t_{jk} M_{b_j+d_k} T_{a_M+c_N} f(x) \;=\; 0 \ \text{ a.e.} \tag{7}$$

Now, if $(j,k) \notin I$, then $a_j+c_k < a_M+c_N$. Since there are only finitely many choices, if we set

$$r \;=\; \max\{a_j+c_k : (j,k)\notin I\} \qquad \text{and} \qquad s \;=\; a_M+c_N,$$

then we have $r<s$.

Let $f \in L^2(\mathbb{R})$ be any function that is nonzero everywhere on $(-\infty,0)$ and zero on $[0,\infty)$. For example, we could take

$$f(x) \;=\; e^x \chi_{(-\infty,0)}(x).$$

If $r<x<s$, then for each $(j,k)\notin I$ we have $x-a_j-c_k \ge 0$ and therefore $f(x-a_j-c_k)=0$. Thus for such x the first summation in Eq. (7) is zero. Consequently, only the second summation in Eq. (7) remains for such x. Simplifying, it follows that for a.e. x in the interval (r,s) we have

$$\begin{aligned} 0 \;&=\; \sum_{(j,k)\in I} t_{jk} M_{b_j+d_k} T_{a_M+c_N} f(x) \\ &=\; \sum_{(j,k)\in I} t_{jk} e^{2\pi i (b_j+d_k)x} f(x-s) \\ &=\; p(x)\, f(x-s), \end{aligned}$$

where

$$p(x) \;=\; \sum_{(j,k)\in I} t_{jk} e^{2\pi i (b_j+d_k)x}. \tag{8}$$

All of the sums above are finite and $f(x-s) \neq 0$ for $r < x < s$, so this implies that

$$p \;=\; 0 \text{ a.e. on } (r,s).$$

But p is a nonharmonic trigonometric polynomial and therefore cannot vanish on any set of positive measure (indeed, p can have at most countably many zeros). Therefore, we must have $p(x) = 0$ for every $x \in \mathbb{R}$.

We are tempted to conclude from this that every t_{jk} is zero, but we cannot do this because some of the values $b_j + d_k$ may coincide. To deal with this, note that if $(j,k) \in I$ then $a_j = a_M$ and $c_k = c_N$. Therefore, if we choose two distinct points (j,k) and (j',k') in I, then $a_j = a_M = a_{j'}$. Since (a_j,b_j) must be distinct from $(a_{j'},b_{j'})$, this implies that $b_j \neq b_{j'}$. Consequently there is a *unique* j_0 such that

$$b_{j_0} \;=\; \max\{b_j : (j,k) \in I\}.$$

Similarly, there is a unique k_0 such that

$$d_{k_0} \;=\; \max\{d_k : (j,k) \in I\}.$$

Therefore Eq. (8) can be rewritten as

$$p(x) \;=\; t_{j_0k_0} e^{2\pi i (b_{j_0}+d_{k_0})x} \;+\; \sum_{(j,k)\in I, b_j+d_k<b_{j_0}+d_{k_0}} t_{jk} e^{2\pi i (b_j+d_k)x}.$$

As p is identically zero, this implies that $t_{j_0k_0} = 0$. However, this contradicts the fact that every t_{jk} is nonzero.

In summary, we have shown that the zero divisor conjecture is true when we take G to be the Heisenberg group $\mathbf{H}$. We state this formally as a theorem.

Theorem 1. *If α and β are nonzero elements of $\mathbb{C}\mathbf{H}$, then $\alpha\beta \neq 0$.*

The proof of Theorem 1 uses an ordering, or indexing, of the elements in the Heisenberg group. First we ordered the translations and examined the largest pair, then we ordered the modulations and again examined the largest pair. It will not be surprising, then, to see that there are general arguments which imply that the Heisenberg group satisfies the Zero Divisor Conjecture. In fact, we will see below that Higman's original work already implies that the Heisenberg group satisfies the zero divisor conjecture.

A group is said to be *locally indicable* if each of its nonidentity finitely generated subgroups maps homomorphically onto $\mathbb{Z}$. A locally indicable group has no nontrivial elements of finite order. Indeed, if $g \in G$ is an element of finite order, then the subgroup $\langle g\rangle$ generated by g is finite, so there cannot exist a homomorphism from $\langle g\rangle$ onto $\mathbb{Z}$.

Given a homomorphism $\gamma : H \to \mathbb{Z}$, we define the *degree* of $h \in H$ (relative to γ) to be $\gamma(h)$. The degree of an element depends on the particular homomorphism chosen, but in what follows we will not explicitly indicate this dependence. We say an element of the group ring $\mathbb{C}H$ is *homogeneous of degree a* (relative to γ) if it can be put in the form $\sum_{j=1}^{k} c_j e_j$, where for each $1 \leq j \leq k$ we have $\gamma(e_j) = a$. Any element of the group ring $\mathbb{C}H$ can be written as

$$P_1 + \cdots + P_p, \tag{9}$$

where each P_i is homogeneous of order a_i and $a_1 < a_2 < \cdots < a_p$.

Here is Higman's result for locally indicable groups, proved in [19].

Theorem 2 (Higman). *If G is a locally indicable group, and if α and β are nonzero elements of $\mathbb{C}G$, then $\alpha\beta \neq 0$.*

Proof. Write $\alpha = m_1 g_1 + \cdots + m_K g_K$ and $\beta = n_1 h_1 + \cdots + n_L h_L$, where $m_1, \ldots, m_K$ and $n_1, \ldots, n_K$ are nonzero complex numbers and $g_1, \ldots, g_K$ and $h_1, \ldots, h_K$ are in G.

We proceed by induction on $K + L$. If $K + L = 2$, then $\alpha\beta = m_1 n_1 g_1 h_1$, which is not zero since $m_1 n_1 \neq 0$.

Let $n > 2$ and assume that $\alpha\beta \neq 0$ whenever the sum of the number of terms of α and β is less than n; that is when $K + L < n$. We will show that $\alpha\beta \neq 0$ whenever $K + L = n$.

Note that if $\alpha\beta = 0$, then $g_1^{-1} \alpha\beta h_1^{-1} = 0$ as well, so we may assume without loss of generality that g_1 and h_1 are the group identity element. Let H be the subgroup of G generated by

$$\{g_1, \ldots, g_K, h_1, \ldots, h_L\},$$

and let ϕ be a homomorphism from H onto $\mathbb{Z}$. Write

$$\alpha = \sum_{i=1}^{r} P_i \qquad \text{and} \qquad \beta = \sum_{j=1}^{s} Q_j$$

as in Eq. (9), where the degrees of $P_1, \ldots, P_r$ are $a_1 < \cdots < a_r$ and the degrees of $Q_1, \ldots, Q_s$ are $b_1 < \cdots < b_s$. Since we have assumed that g_1 and h_1 are the identity, we have that $\phi(g_1) = \phi(h_1) = 0$. Moreover, since ϕ is onto, it cannot map every element of H to zero, and therefore at least one of r or s must exceed one. The product $\alpha\beta$ has the form

$$\alpha\beta = P_1 Q_1 + \cdots + P_r Q_s,$$

where $P_1 Q_1$ is homogeneous of degree $a_1 + b_1$, $P_r Q_s$ is homogeneous of degree $a_r + b_s$, and the terms not listed have degrees strictly between $a_1 + b_1$ and $a_r + b_s$. Since $\alpha\beta \neq P_1 Q_1$, it follows that the sum of the number of terms in P_1 and Q_1 is less than n. Therefore, by the induction hypothesis, $P_1 Q_1 \neq 0$ and hence $\alpha\beta \neq 0$. □

An easy first example of a group which admits a homomorphism onto $\mathbb{Z}$ is $\mathbb{Z}^N$; one choice of homomorphism is $\phi(m_1, \ldots, m_N) = m_1$. Since every finitely generated subgroup of an Abelian group is isomorphic to $\mathbb{Z}^N \bigoplus_{i=1}^K \mathbb{Z}_{a_i}$, it follows that every torsion-free Abelian group is locally indicable. In particular, $(\mathbb{R},+)$, the real line under addition, is locally indicable.

We provide a direct proof that the Heisenberg group is locally indicable. For this argument, it will be most convenient to represent the Heisenberg group as

$$\mathbf{H} = \{(a,b,c) : a,b,c \in \mathbb{R}\},$$

with product

$$(a,b,c)\cdot(x,y,z) = (a+x, b+y, c+z+ay).$$

We gather some basic facts about $\mathbf{H}$ in the following lemma.

Lemma 1. *Let* $\mathbf{H}$ *denote the Heisenberg group.*

(a) $(x,y,z)^{-1} = (-x,-y,-z+xy)$.
(b) *The commutator* $\mathbf{H}' = \{aba^{-1}b^{-1} : a,b \in \mathbf{H}\}$ *is* $\{(0,0,z) : z \in \mathbb{R}\}$, *which is isomorphic to* $(\mathbb{R},+)$.
(c) $\mathbf{H}'$ *is the center of* $\mathbf{H}$ *and is a normal subgroup of* $\mathbf{H}$.
(d) $\mathbf{H}/\mathbf{H}'$ *is isomorphic to* $(\mathbb{R}^2,+)$.

Proposition 1. *The Heisenberg group* $\mathbf{H}$ *is locally indicable.*

Proof. Let G be any finitely generated subgroup of $\mathbf{H}$. We must show that there exists a homomorphism that maps G onto $\mathbb{Z}$.

Case I: $G \subset \mathbf{H}'$. Since $\mathbf{H}'$ is isomorphic to $(\mathbb{R},+)$, this case follows from the local indicability of $(\mathbb{R},+)$.

Case II: $G \not\subset \mathbf{H}'$. Note that $G\mathbf{H}'/\mathbf{H}'$ is a normal subgroup of $\mathbf{H}/\mathbf{H}'$, and that $\mathbf{H}/\mathbf{H}'$ is isomorphic to $\mathbb{R}^2$. By the Second Isomorphism Theorem, there exists an isomorphism

$$\eta : G/(G \cap \mathbf{H}') \to G\mathbf{H}'/\mathbf{H}'.$$

Since G is finitely generated, so is $G/(G\cap\mathbf{H}')$, and therefore $G\mathbf{H}'/\mathbf{H}'$ is finitely generated as well. Therefore, since $\mathbb{R}^2$ is locally indicable, there is a homomorphism ϕ from $G\mathbf{H}'/\mathbf{H}'$ onto $\mathbb{Z}$. Consequently, if we let

$$\psi : G \to G/(G\cap\mathbf{H}')$$

be the natural onto homomorphism, then the composition $\phi \circ \eta \circ \psi$ is a surjective homomorphism of G onto $\mathbb{Z}$. □

Corollary 1. *If* α *and* β *are nonzero elements of* $\mathbb{C}\mathbf{H}$, *then* $\alpha\beta \neq 0$.

A group is said to be *indicable throughout* if *every* subgroup admits a homomorphism onto $\mathbb{Z}$. Since $\mathbf{H}$ contains $\mathbb{R}$ as a subgroup, and there is no homomorphism from $\mathbb{R}$ onto $\mathbb{Z}$, the Heisenberg group is not indicable throughout.

We remark that arguments similar to the ones above can be used to show that the affine group also has no nontrivial zero divisors. Even so, as we described in Section 2, time-scale shifts of functions in $L^2(\mathbb{R})$ are not necessarily linearly independent.

We have shown that no product of two nontrivial finite linear combinations of time–frequency shifts operators can be the zero operator. We close with a related, but simpler, observation.

Lemma 2. *The set of time–frequency shift operators*

$$\{M_bT_a : a,b \in \mathbb{R}\}$$

is a finitely linearly independent set in $\mathbb{C}\mathbf{H}$. *That is, if*

$$\Lambda = \{(a_k,b_k) : k = 1,\dots,N\}$$

is a set of finitely many distinct points in $\mathbb{R}^2$ *and*

$$\sum_{k=1}^{N} c_k M_{b_k} T_{a_k} = 0,$$

then $c_1 = \cdots = c_N = 0$.

Proof. If $\sum_{k=1}^{N} c_k M_{b_k} T_{a_k}$ is the zero operator, then

$$\sum_{k=1}^{N} c_k M_{b_k} T_{a_k} f = 0 \text{ a.e.} \tag{10}$$

for *every* function $f \in L^2(\mathbb{R})$. Yet we know that there are at least *some* functions that have linearly independent time–frequency translates. For example, by the partial results reviewed earlier this is true for every compactly supported function, and for the Gaussian function. Taking f to be one of these functions, Eq. (10) implies that $c_1 = \cdots = c_N = 0$. □

References

1. Balan R. The noncommutative Wiener lemma, linear independence, and spectral properties of the algebra of time–frequency shift operators. Trans Am Math Soc. 2008;360:3921–41.
2. Balan R, Krishtal I. An almost periodic noncommutative Wiener's lemma. J Math Anal Appl. 2010;370:339–49.
3. Benedetto JJ, Bourouihiya A. Linear independence of finite Gabor systems determined by behavior at infinity. J Geom Anal.2014; to appear.
4. Bownik M, Speegle D. Linear independence of Parseval wavelets. Illinois J Math. 2010;54:771–85.
5. Bownik M, Speegle D. Linear independence of time-frequency translates of functions with faster than exponential decay. Bull Lond Math Soc. 2013;45:554–66.
6. Christensen O, Lindner AM. Lower bounds for finite wavelet and Gabor systems. Approx Theory Appl (N.S.). 2001;17:18–29.

7. Daubechies I. Ten lectures on wavelets. Philadelphia: SIAM;1992.
8. Demeter C. Linear independence of time frequency translates for special configurations. Math Res Lett. 2010;17:761–79.
9. Demeter C, Gautam SZ. On the finite linear independence of lattice Gabor systems. Proc Am Math Soc. 2013;141:1735–47.
10. Demeter C, Zaharescu A. Proof of the HRT conjecture for $(2,2)$ configurations. J Math Anal Appl. 2012;388:151–9.
11. Formanek E. The zero divisor question for supersolvable groups. Bull Aust Math Soc. 1973;9:69–71.
12. Gröchenig K. Linear independence of time-frequency shifts? Preprint. 2014.
13. Haar A. Zur Theorie der orthogonalen Funktionensysteme. Math Ann. 1910;69:331–71; English translation in [17].
14. Heil C. Linear independence of finite Gabor systems, In: Heil C, Editor. Harmonic analysis and applications. Boston: Birkhäuser; 2006. p. 171–206.
15. Heil C. A basis theory primer, expanded Edition. Boston: Birkhäuser; 2011.
16. Heil C. WHAT IS ... a frame? Notices Am Math Soc. 2013;60:748–50.
17. Heil C, Walnut DF, Editors. Fundamental papers in wavelet theory. Princeton: Princeton University Press; 2006.
18. Heil C, Ramanathan J, Topiwala P. Linear independence of time-frequency translates. Proc Am Math Soc. 1996;124:2787–95.
19. Higman G. The units of group-rings. Proc Lond Math Soc. 1940;46(2):231–48.
20. Kutyniok G. Linear independence of time–frequency shifts under a generalized Schrödinger representation. Arch Math. (Basel) 2002;78:135–44.
21. Linnell PA. Analytic versions of the zero divisor conjecture. In: Kropholler PH, Niblo GA, Stöhr R, Editors. Geometry and cohomology in group theory. London Mathematical Society Lecture Note Ser., Vol. 252. Cambridge: Cambridge University Press; 1998. p. 209–48.
22. Linnell PA. Von Neumann algebras and linear independence of translates. Proc Am Math Soc. 1999;127:3269–77.
23. Passman DS. The algebraic structure of group rings. New York: Wiley-Interscience; 1977.
24. Rosenblatt J. Linear independence of translations. Int J Pure Appl Math. 2008;45:463–73.
25. Snider RL. The zero divisor conjecture. In: McDonald BR, Morris RA, Editors. Ring theory, II. Lecture notes in pure and applied mathematics. Vol. 26. New York : Dekker; 1977. p. 261–95.

The *abc*-Problem for Gabor Systems and Uniform Sampling in Shift-Invariant Spaces

Xin-Rong Dai and Qiyu Sun

Abstract In this chapter, we identify ideal window functions χ_I on finite intervals I and time–frequency shift lattices $a\mathbb{Z}\times b\mathbb{Z}$ such that the corresponding Gabor systems

$$\mathscr{G}(\chi_I, a\mathbb{Z}\times b\mathbb{Z}) := \{e^{-2\pi inbt}\chi_I(t-ma):\ (m,n)\in\mathbb{Z}\times\mathbb{Z}\}$$

are frames for $L^2(\mathbb{R})$. Also we consider a stable recovery of rectangular signals f in a shift-invariant space

$$V_2(\chi_I, b\mathbb{Z}) := \Big\{\sum_{\lambda\in b\mathbb{Z}} d(\lambda)\chi_I(t-\lambda):\ \sum_{\lambda\in b\mathbb{Z}} |d(\lambda)|^2<\infty\Big\}$$

from their equally-spaced samples $f(t_0+\mu), \mu\in a\mathbb{Z}$, for arbitrary initial sampling position t_0.

Keywords Gabor frame · Infinite matrix · Uniform stability · Sampling · Shift-invariant space

1 Introduction

Denote by $L^2 := L^2(\mathbb{R})$, the space of all square-integrable functions on the real line $\mathbb{R}$ with inner product $\langle\cdot,\cdot\rangle$ and norm $\|\cdot\|_2$. Let $\mathscr{G}(\phi, a\mathbb{Z}\times b\mathbb{Z})$ be the *Gabor system*

X.-R. Dai (✉)
School of Mathematical and Computational Science, Sun Yat-Sen University,
510275 Guangzhou, P. R. China
e-mail: daixr@mail.sysu.edu.cn

Q. Sun
Department of Mathematics, University of Central Florida, Orlando, FL 32816, USA
e-mail: qiyu.sun@ucf.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_8

associated with a window function $\phi \in L^2$ and a rectangular lattice $a\mathbb{Z} \times b\mathbb{Z}$,

$$\mathscr{G}(\phi, a\mathbb{Z} \times b\mathbb{Z}) := \{e^{-2\pi inbt}\phi(t-ma) : \ (m,n) \in \mathbb{Z} \times \mathbb{Z}\}.$$

We say that $\mathscr{G}(\phi, a\mathbb{Z} \times b\mathbb{Z})$ is a *Gabor frame* for L^2, if there exist positive constants A and B such that

$$A\|f\|_2 \le \Big(\sum_{m,n\in\mathbb{Z}} |\langle f, e^{-2\pi inb\cdot}\phi(\cdot - ma)\rangle|^2\Big)^{1/2} \le B\|f\|_2, \ f \in L^2$$

[6, 8, 13]. Gabor frames have been shown to be important and useful in many mathematical and engineering fields [5, 8, 16, 17, 20–22, 25]. The history of Gabor theory could date back to the completeness claim in 1932 by von Neumann [31, p. 406], and the expansion conjecture in 1946 by Gabor [15, Eq. 1.29]. It has been widely studied in the past three decades, see the landmark paper by Daubechies, Grossmann and Meyer [12], the textbook by Gröchenig [16], and the survey by Janssen [25] and Heil [22].

Given a window function ϕ, one of fundamental problems in Gabor theory is to find the range $\mathscr{R}(\phi)$ of pairs (a,b) such that $\mathscr{G}(\phi, a\mathbb{Z} \times b\mathbb{Z})$ is a Gabor frame for L^2. It has been shown that the range $\mathscr{R}(\phi)$ is contained in the equilateral hyperbola for arbitrary window function ϕ,

$$\mathscr{R}(\phi) \subset \{(a,b) : ab \le 1\}$$

[4, 10, 24, 29, 32], and it is an open domain for a window function ϕ in Feichtinger's algebra [14]. The range $\mathscr{R}(\phi)$ is fully known unexpectedly only for very few families of window functions ϕ, including Gaussian windows, hyperbolic secant windows, two-sided exponential windows, one-sided exponential windows, and totally positive windows [11, 18, 27, 28, 30, 34, 35].

The Gabor system generated by the ideal window χ_I (the characteristic function) on an interval I has received special attention in Gabor theory. The range $\mathscr{R}(\chi_I)$ for the ideal window χ_I could be arbitrarily complicated as the famous Janssen tie suggests [19, 26], and it is not open and path-connected [9]. It has been a long standing problem to find the range $\mathscr{R}(\chi_I)$ and study its algebraic and topological properties.

Recall that for any given interval I, $\mathscr{G}(\chi_I, a\mathbb{Z} \times b\mathbb{Z})$ is a Gabor frame if and only if $\mathscr{G}(\chi_{I+d}, a\mathbb{Z} \times b\mathbb{Z})$ is a Gabor frame for every $d \in \mathbb{R}$. Due to the above shift-invariance, we may assume that $I = [0,c)$, where c is the length of the interval I. Then, the range problem for the ideal window on an interval reduces to the so-called *abc-problem for Gabor systems*: given a triple (a,b,c) of positive numbers, determine whether $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z} \times b\mathbb{Z})$ is a Gabor frame [7]. In the first part of this chapter, we report our complete answer to the above *abc*-problem for Gabor systems in [9].

The second topic of this chapter is to identify generators ϕ, particularly rectangular signal generator χ_I, and sampling-shift lattices $a\mathbb{Z}\times b\mathbb{Z}$ such that

$$A\|f\|_2 \le \Big(\sum_{\mu\in a\mathbb{Z}} |f(t_0+\mu)|^2\Big)^{1/2} \le B\|f\|_2,\ f\in V_2(\phi, b\mathbb{Z}) \tag{1}$$

are satisfied for all $t_0\in\mathbb{R}$, where A and B are positive constants, and

$$V_2(\phi, b\mathbb{Z}) := \Big\{\sum_{\lambda\in b\mathbb{Z}} d(\lambda)\phi(t-\lambda):\ \sum_{\lambda\in b\mathbb{Z}} |d(\lambda)|^2<\infty\Big\}$$

is the shift-invariant space generated by ϕ. In other words, signals f in the shift-invariant space $V_2(\phi, b\mathbb{Z})$ could be stably recovered from their uniform samples $f(t_0+\mu), \mu\in a\mathbb{Z}$, for arbitrary initial sampling position t_0. For fixed initial sampling position t_0, there are lots of literatures on the stability requirement (1) and robust recovery algorithm, see for instance [2, 3, 37–39, 41]. But for arbitrary initial sampling position, it is known only for few generators ϕ, including bandlimited signals and spline signals [1, 36, 38]. In the second part of this chapter, we consider the almost equivalence between frame property of the Gabor system $\mathscr{G}(\chi_I, a\mathbb{Z}\times\mathbb{Z}/b)$ and stability requirement (1) for uniform sampling signals in the shift-invariant space $V_2(\chi_I, b\mathbb{Z})$.

Notation: For a real number s, we let $s_+=\max(s,0)$, $s_-=\min(s,0)$, $\lfloor s\rfloor$ be the largest integer not greater than s, $\mathrm{sgn}(s)$ be the sign of s, and $\mathbf{s}:=(\cdots,s,s,s,\cdots)^T$ be the column vector whose entries take value s. Specially for the window size parameter c, we let $c_0:=c-\lfloor c\rfloor$ be the fractional part of the window size. For a set E, we denote by χ_E the characteristic function on it, by $|E|$ its Lebesgue measure, and by $\#(E)$ its cardinality, respectively. We also denote by $\gcd(s,t)$ the greatest common divisor such that $s/\gcd(s,t), t/\gcd(s,t)\in\mathbb{Z}$ for any given s and t in a lattice $r\mathbb{Z}$ with $r>0$. In this chapter, we let

$$\mathscr{B}^0 := \big\{(\mathbf{x}(\lambda))_{\lambda\in\mathbb{Z}}:\ \ \mathbf{x}(0)=1 \text{ and } \mathbf{x}(\lambda)\in\{0,1\} \text{ for all } \lambda\in\mathbb{Z}\big\}$$

contain all binary column vectors taking value one at the origin; $\ell^2:=\ell^2(\Lambda)$ be the space of all square-summable vectors $\mathbf{z}:=(\mathbf{z}(\lambda))_{\lambda\in\Lambda}$ with standard norm $\|\cdot\|_2:=\|\cdot\|_{\ell^2(\Lambda)}$, where Λ is a given index set; and let shift-operators $\tau_{\nu'}, \nu'\in\alpha\mathbb{Z}$, on sequence spaces be defined by

$$\tau_{\nu'}\mathbf{z} := (\mathbf{z}(\nu+\nu'))_{\nu\in\alpha\mathbb{Z}}$$

for $\mathbf{z}:=(\mathbf{z}(\nu))_{\nu\in\alpha\mathbb{Z}}$ and $\alpha>0$.

2 Gabor Frames and Infinite Matrices

Given a triple (a,b,c) of positive numbers, it is obvious that $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times b\mathbb{Z})$ is a Gabor frame if and only if $\mathscr{G}(\chi_{[0,bc)}, (ab)\mathbb{Z}\times\mathbb{Z})$ is. By the above dilation-invariance, the frequency-spacing parameter b can be normalized to 1 and then the abc-problem for Gabor systems reduces to finding out *all pairs* (a,c) *of positive numbers such that* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *are Gabor frames.*

For pairs (a,c) satisfying either $a \ge 1$ or $c \le 1$, it is known that the Gabor system $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ is a Gabor frame if and only if $c=1$ and $0<a\le 1$, see for instance [12, 19, 26] and also Theorem 8. Thus, it remains to consider the abc-problem for Gabor systems with triples (a,b,c) with

$$0<a<1<c \quad \text{and} \quad b=1.$$

Define infinite matrices $\mathbf{M}_{a,c}(t), t\in\mathbb{R}$, by

$$\mathbf{M}_{a,c}(t) := \big(\chi_{[0,c)}(t-\mu+\lambda)\big)_{\mu\in a\mathbb{Z},\lambda\in\mathbb{Z}},\ t\in\mathbb{R}. \tag{2}$$

For fixed $t\in\mathbb{R}$, the infinite matrix $\mathbf{M}_{a,c}(t)$ in (2) has its rows (respectively columns) containing $\lfloor c\rfloor+\{0,1\}$ (respectively $\lfloor c/a\rfloor+\{0,1\}$) consecutive ones, and its rows are obtained by shifting one (or zero) unit of the previous row with possible reduction or expansion by one unit. The above observations could be illustrated from the example below:

$$\mathbf{M}_{a,c}(0) = \left(\begin{array}{ccccccccccccccccc}
\ddots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & & & & & & & & & \\
& 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & & & & & & & & \\
& & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & & & & & & & \\
& & & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & & & & & & \\
& & & & 0 & 1 & 1 & 1 & 1 & 1 & 0 & & & & & & \\
& & & & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & & & & & \\
& & & & & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & & & & \\
& & & & & & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & & & \\
& & & & & & & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & & \\
& & & & & & & & 0 & 1 & 1 & 1 & 1 & 1 & 0 & & \\
& & & & & & & & 0 & 1 & 1 & 1 & 1 & 1 & 1 & 0 & \\
& & & & & & & & & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \ddots
\end{array}\right)$$

where $(a,c)=(\pi/4, 23-11\pi/2)$.

Infinite matrices $\mathbf{M}_{a,c}(t), t\in\mathbb{R}$ in (2) have been used by Ron and Shen [33] to characterize the frame property for the Gabor system $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ via their uniform stability on ℓ^2,

$$0<\inf_{t\in\mathbb{R}}\inf_{\|\mathbf{z}\|_2=1}\|\mathbf{M}_{a,c}(t)\mathbf{z}\|_2 \le \sup_{t\in\mathbb{R}}\sup_{\|\mathbf{z}\|_2=1}\|\mathbf{M}_{a,c}(t)\mathbf{z}\|_2<\infty. \tag{3}$$

In this chapter, we report a **new** characterization of frame property for the Gabor system $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ via nonexistence of binary solutions $\mathbf{x}\in\mathscr{B}^0$ of the following infinite-dimensional linear system

$$\mathbf{M}_{a,c}(t)\mathbf{x} = \mathbf{2},\ t \in \mathbb{R}, \tag{4}$$

or equivalently $\mathscr{D}_{a,c} = \emptyset$, where

$$\mathscr{D}_{a,c} := \left\{ t \in \mathbb{R} :\ \mathbf{M}_{a,c}(t)\mathbf{x} = \mathbf{2} \text{ for some binary vectors } \mathbf{x} \in \mathscr{B}^0 \right\} \tag{5}$$

is the set of real numbers t such that there exists a binary solution $\mathbf{x} \in \mathscr{B}^0$ to the linear system (4).

Theorem 1. ([9]) *Let* $0 < a < 1 < c$. *Then,* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z} \times \mathbb{Z})$ *is a Gabor frame if and only if* $\mathscr{D}_{a,c} = \emptyset$.

The necessity in the above theorem has been implicitly used in [19, 26] for their classifications. For pairs (a,c) of positive numbers satisfying

$$\text{either } \ c_0 := c - \lfloor c \rfloor \leq 1 - a \ \text{ or } \ c_0 \geq a,$$

the set $\mathscr{D}_{a,c}$ in (5) could be constructed explicitly, cf. Theorem 9.

For any $t \in \mathbb{R}$ and $\mathbf{x} \in \mathscr{B}$, define

$$Q_{a,c}(t,\mathbf{x}) := \begin{cases} 0 & \text{if } K(t,\mathbf{x}) = \emptyset \\ \sup \big\{ n \in \mathbf{N} \big| \ [\mu, \mu + na) \cap a\mathbf{Z} & \\ \qquad\qquad \subset K(t,\mathbf{x}) \text{ for some } \mu \in a\mathbb{Z} \big\} & \text{otherwise,} \end{cases}$$

where $K(t,\mathbf{x}) := \left\{ \mu \in a\mathbb{Z} \big| \ \mathbf{M}_{a,c}(t)\mathbf{x}(\mu) = 2 \right\}$. An equivalent formulation of the empty set property $\mathscr{D}_{a,c} \neq \emptyset$ is

$$\mathbf{2} \notin \mathbf{M}_{a,c}(t)\mathscr{B}^0 \quad \text{for all} \quad t \in \mathbb{R}.$$

A quantitative version of the above equivalent formulation is the maximal length $Q_{a,c}(t)$ of consecutive twos for vectors in range spaces $\mathbf{M}_{a,c}(t)\mathscr{B}^0, t \in \mathbb{R}$, where

$$Q_{a,c}(t) = \sup_{\mathbf{x} \in \mathscr{B}^0} Q_{a,c}(t,\mathbf{x}).$$

Clearly $Q_{a,c}(t) = +\infty$ for any $t \in \mathscr{D}_{a,c}$. In [9], we further show that $\mathscr{D}_{a,c} = \emptyset$ if and only if $Q_{a,c} := \sup_{t \in \mathbb{R}} Q_{a,c}(t)$ is finite. This together with Theorem 1 leads to the equivalence between frame property of Gabor systems $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z} \times \mathbb{Z})$ and boundedness of the maximal length $Q_{a,c}$ of consecutive twos in range spaces $\mathbf{M}_{a,c}(t)\mathscr{B}^0, t \in \mathbb{R}$ of infinite matrices. More importantly, the quantity $Q_{a,c}$ can be used to estimate Gabor frame bounds of the Gabor system $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z} \times \mathbb{Z})$,

$$\begin{aligned} \frac{a^2 \lfloor 1/a \rfloor}{4c^2(aQ_{a,c} + 2a + c + 1)^2} \|f\|_2 \leq & \Big(\sum_{\phi \in \mathscr{G}(\chi_{[0,c)}, a\mathbb{Z} \times \mathbb{Z})} |\langle f, \phi \rangle|^2 \Big)^{1/2} \\ & \leq (\lfloor c \rfloor + 1)(\lfloor c/a \rfloor + 1)\|f\|_2,\ f \in L^2. \end{aligned}$$

3 Maximal Invariant Sets

Consider binary solutions $\mathbf{x} \in \mathscr{B}^0$ of the following infinite-dimensional linear system

$$\mathbf{M}_{a,c}(t)\mathbf{x} = \mathbf{1} \tag{6}$$

for $t \in \mathbb{R}$, and let

$$\mathscr{S}_{a,c} := \big\{t \in \mathbb{R} : \ \mathbf{M}_{a,c}(t)\mathbf{x} = \mathbf{1} \text{ for some vector } \mathbf{x} \in \mathscr{B}^0\big\}. \tag{7}$$

Given $t \in \mathscr{D}_{a,c}$ and binary vector $\mathbf{x} \in \mathscr{B}^0$ satisfying $\mathbf{M}_{a,c}(t)\mathbf{x} = \mathbf{2}$, let K be the set of all $\lambda \in \mathbb{Z}$ with $\mathbf{x}(\lambda) = 1$, write $K = \{\lambda_j : j \in \mathbb{Z}\}$ for a strictly increasing sequence $\{\lambda_j\}_{j=-\infty}^{\infty}$ with $\lambda_0 = 0$, and define $\mathbf{x}^* := (x^*(\lambda))_{\lambda \in \mathbb{Z}}$ by $\mathbf{x}^*(\lambda) = 1$ if $\lambda = \lambda_{2j}$ for some $j \in \mathbb{Z}$ and $\mathbf{x}^*(\lambda) = 0$ otherwise. One may easily verify that $\mathbf{x}^*$ is a binary vector in $\mathscr{B}^0$ that satisfies (6). Thus $\mathscr{S}_{a,c}$ is a supset of $\mathscr{D}_{a,c}$, i.e.,

$$\mathscr{D}_{a,c} \subset \mathscr{S}_{a,c}.$$

Conversely, it is shown in [9] that $\mathscr{D}_{a,c}$ can be obtained from $\mathscr{S}_{a,c}$ via certain set operations,

$$\begin{aligned}\mathscr{D}_{a,c} = {} & \big(\mathscr{S}_{a,c} \cap (\cup_{\lambda=1}^{\lfloor c\rfloor-1}(\mathscr{S}_{a,c} - \lambda))\big) \\ & \cup \big(\mathscr{S}_{a,c} \cap ([0, (c_0 + a - 1)_+) + a\mathbb{Z}) \cap (\mathscr{S}_{a,c} - \lfloor c\rfloor)\big).\end{aligned}$$

A **pivotal** observation for the set $\mathscr{S}_{a,c}$ in (7) is that for any $t \in \mathscr{S}_{a,c}$, there is a unique binary solution $\mathbf{x}_t := (\mathbf{x}_t(\lambda))_{\lambda \in \mathbb{Z}} \in \mathscr{B}^0$ for the linear system (6). Let $\lambda_{a,c}(t)$ be the smallest positive integer such that $\mathbf{x}_t(\lambda_{a,c}(t)) = 1$ and similarly let $\tilde{\lambda}_{a,c}(t)$ be the largest negative integer such that $\mathbf{x}_t(\tilde{\lambda}_{a,c}(t)) = 1$. Then

$$\tau_{\lambda_{a,c}(t)}\mathbf{x}_t, \tau_{\tilde{\lambda}_{a,c}(t)}\mathbf{x}_t \in \mathscr{B}^0$$

and

$$\mathbf{M}_{a,c}(t + \lambda_{a,c}(t))\tau_{\lambda_{a,c}}\mathbf{x}_t = \mathbf{M}_{a,c}(t + \tilde{\lambda}_{a,c}(t))\tau_{\tilde{\lambda}_{a,c}}\mathbf{x}_t\mathbf{M}_{a,c}(t)\mathbf{x}_t = \mathbf{1}$$

by the frequency-shift property

$$\mathbf{M}_{a,c}(t - \lambda')\mathbf{z} = \mathbf{M}_{a,c}(t)\tau_{\lambda'}\mathbf{z} \ \text{ for all } \lambda' \in \mathbb{Z}$$

for infinite matrices $\mathbf{M}_{a,c}(t), t \in \mathbb{R}$. This defines maps

$$\mathscr{S}_{a,c} \ni t \longrightarrow t + \lambda_{a,c}(t) \in \mathscr{S}_{a,c} \quad \text{and} \quad \mathscr{S}_{a,c} \ni t \longrightarrow t + \tilde{\lambda}_{a,c}(t) \in \mathscr{S}_{a,c}$$

on the set $\mathscr{S}_{a,c}$. Our inspection shows that the above two maps on $\mathscr{S}_{a,c}$ can be extended to piecewise linear transformations $R_{a,c}$ and $\tilde{R}_{a,c}$ on the line $\mathbb{R}$, respectively, where

$$R_{a,c}(t) := \begin{cases} t+\lfloor c \rfloor & \text{if } t \in [(c_0-a)_-, 0) + a\mathbb{Z} \\ t+\lfloor c \rfloor + 1 & \text{if } t \in [0, (c_0+a-1)_+) + a\mathbb{Z} \\ t & \text{if } t \in [(c_0+a-1)_+, (c_0-a)_- + a) + a\mathbb{Z} \end{cases} \tag{8}$$

and

$$\tilde{R}_{a,c}(t) = \begin{cases} t-\lfloor c \rfloor - 1 & \text{if } t \in [c-(c_0+a-1)_+, c) + a\mathbb{Z} \\ t-\lfloor c \rfloor & \text{if } t \in [c, c-(c_0-a)_-) + a\mathbb{Z} \\ t & \text{if } t \in [c-(c_0-a)_-, c+a-(c_0+a-1)_+) + a\mathbb{Z}. \end{cases} \tag{9}$$

So $\mathscr{S}_{a,c}$ is an invariant set under transformations $R_{a,c}$ and $\tilde{R}_{a,c}$,

$$R_{a,c}\mathscr{S}_{a,c} = \mathscr{S}_{a,c} \quad \text{and} \quad \tilde{R}_{a,c}\mathscr{S}_{a,c} = \mathscr{S}_{a,c}, \tag{10}$$

and it has empty intersection with their *black holes*,

$$\begin{cases} \mathscr{S}_{a,c} \cap ([(c_0+a-1)_+, a+(c_0-a)_-) + a\mathbb{Z}) = \emptyset \\ \mathscr{S}_{a,c} \cap ([c-(c_0-a)_-, c+a-(c_0+a-1)_+) + a\mathbb{Z}) = \emptyset. \end{cases} \tag{11}$$

More importantly, it is a **maximal** set that is invariant under the transformation $R_{a,c}$ and has empty intersection with its black hole.

Theorem 2. ([9]) *Let $0 < a < 1 < c$. Then any set E satisfying $R_{a,c}E = E$ and having empty intersection with the black hole $[(c_0+a-1)_+, a+(c_0-a)_-) + a\mathbb{Z}$ of the transformation $R_{a,c}$ is contained in $\mathscr{S}_{a,c}$.*

The maximal invariance property for the set $\mathscr{S}_{a,c}$ is crucial in our study. So we call $\mathscr{S}_{a,c}$ the *maximal invariant set*. We remark that the set $\mathscr{D}_{a,c}$ in (5) is also invariant under transformations $R_{a,c}$ and $\tilde{R}_{a,c}$ and has empty intersection with their black holes.

Set $c_1 := \lfloor c \rfloor - \lfloor (\lfloor c \rfloor / a) \rfloor a$. For pairs (a,c) of positive numbers satisfying

$$\text{either } \lfloor c \rfloor = 1 \text{ or } c_1 \ge 2a-1 \text{ or } c_1 = 0,$$

we can apply the maximality in Theorem 2 to construct the set $\mathscr{S}_{a,c}$ explicitly, and then we can determine whether the corresponding Gabor system $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z} \times \mathbb{Z})$ is a frame for L^2, see Theorem 10.

4 Piecewise Linear Transformations

The transformations $R_{a,c}$ and $\tilde{R}_{a,c}$ in (8) and (9) are well-defined as $(c_0+a-1)_+ \le (c_0-a)_- + a$. They are not measure-preserving on the whole line, but they are measure-preserving outside their black holes,

$$\begin{cases} |R_{a,c}(E)| = |E| & \text{if } E \cap ([(c_0+a-1)_+, (c_0-a)_- + a) + a\mathbb{Z}) = \emptyset \\ |\tilde{R}_{a,c}(E)| = |E| & \text{if } E \cap ([c-(c_0-a)_-, c+a-(c_0+a-1)_+) + a\mathbb{Z}) = \emptyset. \end{cases} \tag{12}$$

In [9], we consider ergodicity of the transformation $R_{a,c}$, see [40] for introduction to ergodic theory of various dynamic systems.

Theorem 3. *Let $0 < a < 1 < c$ and F be a continuous periodic function with period a. Then*

$$F(t) := \lim_{n\to\infty} \frac{\sum_{k=0}^{n-1} f((R_{a,c})^k(t))}{n}$$

is well-defined for all $t \in \mathbb{R}$. Moreover

$$F(t) = \begin{cases} \frac{1}{|\mathscr{S}_{a,c}\cap[0,a)|}\int_{\mathscr{S}_{a,c}\cap[0,a)} f(s)ds & \text{if } t \in \mathscr{S}_{a,c} \text{ and } a \notin \mathbb{Q} \\ \frac{1}{D+1}\sum_{k=0}^{D} f((R_{a,c})^k(t)) & \text{if } t \in \mathscr{S}_{a,c} \text{ and } a \in \mathbb{Q} \\ f(t_0) & \text{if } t \notin \mathscr{S}_{a,c}, \end{cases}$$

where $D \ge 0$ is a nonnegative integer independent on $t \in \mathscr{S}_{a,c}$, and $t_0 \in [(c_0+a-1)_+, (c_0-a)_- + a) + a\mathbb{Z}$ is the limit of $(R_{a,c})^n(t)$ as $n \to \infty$ for $t \notin \mathscr{S}_{a,c}$.

Due to the measure-preserving property (12), the transformation $R_{a,c}$ is noncontractive and its maximal invariant set $\mathscr{S}_{a,c}$ does not directly follow from the Hutchinson's remarkable construction [23]. We observe from the invariance property (10) and the empty intersection property (11) that

$$\mathscr{S}_{a,c} \subset \cap_{n=0}^{\infty}(R_{a,c})^n(\mathbb{R}) \backslash ([(c_0+a-1)_+, (c_0-a)_- + a) + a\mathbb{Z}).$$

In the case that $\mathscr{S}_{a,c} \neq \emptyset$, we show in [9] that the infinite intersection in the above inclusion can be replaced by **finite** intersection and the inclusion is indeed an equality.

Theorem 4. *Let $0 < a < 1 < c$. Assume that $\mathscr{S}_{a,c} \neq \emptyset$. Then, there exists a nonnegative integer D such that*

$$\mathscr{S}_{a,c} = (R_{a,c})^L(\mathbb{R}) \backslash ([(c_0+a-1)_+, (c_0-a)_- + a) + a\mathbb{Z}) \quad \text{for all } L \ge D.$$

Applying Theorem 4 to pairs $(\pi/4, 23-11\pi/2)$ and $(13/17, 77/17)$ leads to the explicit expression of their corresponding maximal invariant set $\mathscr{S}_{a,c}$:

$$\mathscr{S}_{a,c} = \big[18 - \frac{23\pi}{4}, 11 - \frac{7\pi}{2}\big) \cup \big[12 - \frac{15\pi}{4}, 5 - \frac{3\pi}{2}\big) \cup \big[6 - \frac{7\pi}{4}, 17 - \frac{21\pi}{4}\big) + \frac{\pi}{4}\mathbb{Z}$$

for the pair $(a,c) = (\pi/4, 23-11\pi/2)$, and

$$\mathscr{S}_{a,c} = \frac{1}{17}([2,3) \cup [9,10) \cup [12,13)) + \frac{13}{17}\mathbb{Z}$$

for the pair $(a,c) = (13/17, 77/17)$.

From Theorem 4, we see that the maximal invariant set $\mathscr{S}_{a,c}$ has its complement composed by finitely many left-closed right-open intervals, called *holes*, on one period a (and hence the maximal invariant set $\mathscr{S}_{a,c}$ is measurable). Therefore, we may squeeze out those holes and then reconnect their endpoints. This holes-removal surgery could be mathematically described by the map

$$Y_{a,c}(t) := \operatorname{sgn}(t)|[t_-, t_+) \cap \mathscr{S}_{a,c}|, \ t \in \mathbb{R} \tag{13}$$

on the line in the sense that it is an isomorphism from the maximal invariant set $\mathscr{S}_{a,c}$ to the *line with marks* (image of the holes). In Fig. 1 below, we illustrate the performance of the holes-removal surgery via

$$a\mathbb{T} \ni a\exp(2\pi it/a) \longmapsto Y_{a,c}(a)\exp\big(-2\pi i Y_{a,c}(t)/Y_{a,c}(a)\big) \in Y_{a,c}(a)\mathbb{T},$$

where $\big(\pi/4, 23-11\pi/2\big)$ and $(13/17, 77/17)$ are used as pairs (a,c) in the left and right subfigures, respectively. After performing the holes-removal surgery, it is shown in [9] that the restriction of the piecewise linear transformation $R_{a,c}$ onto the maximal invariant set $\mathscr{S}_{a,c}$ becomes a **linear** transformation on a line with marks.

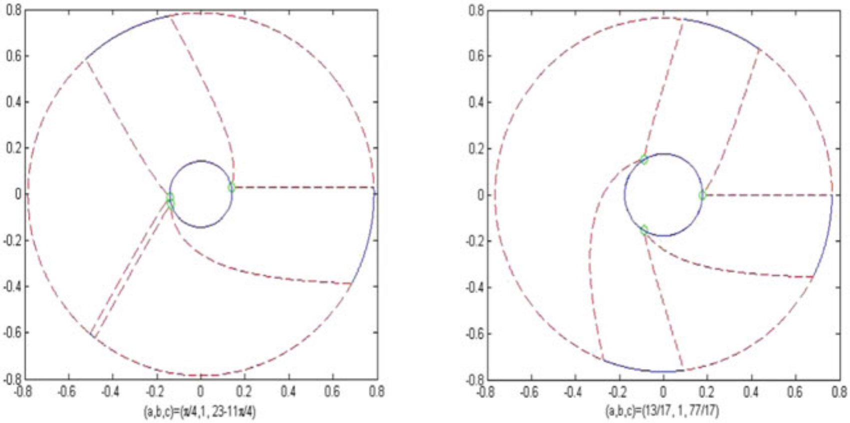


Fig. 1 The set $a\exp(2\pi i\mathscr{S}_{a,c}/a)$ contains *blue arcs* in the *big circle*. The set $a\exp(2\pi i(\mathbb{R}\backslash\mathscr{S}_{a,c})/a)$ is composed of *red dashed arcs* in the *big circle*. The image $Y_{a,c}(a)\exp\big(2\pi i Y_{a,c}(\mathbb{R})/Y_{a,c}(a)\big)$ of the map $Y_{a,c}$ is the *small circle*. The set $Y_{a,c}(a)\exp\big(2\pi i\mathscr{K}_{a,c}/Y_{a,c}(a)\big)$ is marked with *tiny circle*, where $\mathscr{K}_{a,c}$ is the set of marks on the line

Theorem 5. *Let* $0<a<1<c$. *Assume that* $\mathscr{S}_{a,c} \neq \emptyset$. *Then, the following diagram commutes,*

$$\begin{tikzcd}
\mathscr{S}_{a,c} \arrow[r, "R_{a,c}"] \arrow[d, "Y_{a,c}"'] & \mathscr{S}_{a,c} \arrow[d, "Y_{a,c}"] \\
\mathbb{R}/(Y_{a,c}(a)\mathbb{Z}) \arrow[r, "S(\theta_{a,c})"'] & \mathbb{R}/(Y_{a,c}(a)\mathbb{Z})
\end{tikzcd} \tag{14}$$

where $\theta_{a,c} = Y_{a,c}(\lfloor c \rfloor + 1)$ *and*

$$S(\theta_{a,c})(z + Y_{a,c}(a)\mathbb{Z}) = \theta_{a,c} + z + Y_{a,c}(a)\mathbb{Z}, \quad z \in \mathbb{R}/(Y_{a,c}(a)\mathbb{Z}).$$

5 Parameterizations of Maximal Invariant Sets

For parameterization of maximal invariant sets $\mathscr{S}_{a,c}$, we need more detailed information on their holes for pairs (a,c) satisfying

$$0 < a < 1 < c, 1 - a < c_0 < a, \lfloor c \rfloor \ge 2 \text{ and } 0 < c_1 < 2a - 1. \tag{15}$$

So in this section, we always assume that the pair (a,c) satisfies the above condition. One may verify that the transformation $\tilde{R}_{a,c}$ is the left-inverse of the transformation $R_{a,c}$ outside its black hole and vice versa (hence, the transformations $R_{a,c}$ and $\tilde{R}_{a,c}$ are one-to-one outside their black holes), i.e.,

$$\begin{cases} \tilde{R}_{a,c}(R_{a,c}(t)) = t \text{ if } t \notin [c_0 + a - 1, c_0) + a\mathbb{Z} \\ R_{a,c}(\tilde{R}_{a,c}(t)) = t \text{ if } t \notin [c - c_0, c - c_0 + 1 - a) + a\mathbb{Z}. \end{cases}$$

This, together with the invariance property (10) and the empty intersection property (11) for the set $\mathscr{S}_{a,c}$, implies that holes

$$\mathscr{A}_n := (R_{a,c})^n([c - c_0, c - c_0 + 1 - a) + a\mathbb{Z}),\ n \ge 0,$$

obtained from applying the transformation $R_{a,c}$ to the black hole $[c - c_0, c - c_0 + 1 - a) + a\mathbb{Z}$ of the transformation $\tilde{R}_{a,c}$ have empty intersection with the maximal invariant set $\mathscr{S}_{a,c}$, and that their mutual intersections are contained in the black hole $[c_0 + a - 1, c_0) + a\mathbb{Z}$ of the transformation $R_{a,c}$. For the case that $\mathscr{S}_{a,c} \ne \emptyset$, we further show in [9] that those holes will **eventually** become the black hole $[(c_0 + a - 1)_+, a + (c_0 - a)_-) + a\mathbb{Z}$ of the transformation $R_{a,c}$ and hence, the black hole $[c_0 + a - 1, c_0) + a\mathbb{Z}$ of the transformation $R_{a,c}$ and the black hole $[c - c_0, c - c_0 + 1 - a) + a\mathbb{Z}$ of the transformation $\tilde{R}_{a,c}$ are transformable through periodic holes $\mathscr{A}_n, 0 \le n \le D$, in **finite** steps. Thus

$$\mathscr{S}_{a,c} = \mathbb{R} \backslash \big(\cup_{n=0}^{D} (R_{a,c})^n([c - c_0, c - c_0 + 1 - a) + a\mathbb{Z}) \big)$$

by the maximal invariance given in Theorem 2, cf. Theorem 4.

For irrational time-spacing parameter a, $\mathscr{A}_n = (R_{a,c})^n(c - c_0 + 1 - a) + [a - 1, 0) + a\mathbb{Z}, 0 \le n \le D$, and they have their closure being mutually disjoint. This leads to a one-to-one correspondence between the maximal invariant set $\mathscr{S}_{a,c}$ and the set $\mathscr{K}_{a,c}$ of marks on the line. By the commutative diagram (14) for the transformation

$R_{a,c}$, the set of marks is given by

$$\begin{aligned}\mathscr{K}_{a,c} &= \{Y_{a,c}((R_{a,c})^n(c-c_0+1-a)),\ 0\le n\le D\}+Y_{a,c}(a)\mathbb{Z}\\ &= \{nY_{a,c}(c-c_0+1-a),\ 1\le n\le D+1\}+Y_{a,c}(a)\mathbb{Z}.\end{aligned}$$

Thus, the set of marks is completely determined by the number of marks on one period $[0,Y_{a,c}(a))$ and the locations $Y_{a,c}(c-c_0+1-a)+Y_{a,c}(a)\mathbb{Z}$ and $Y_{a,c}(c_0)+Y_{a,c}(a)\mathbb{Z}$ of two marks associated with black holes $[c-c_0,c-c_0+1-a)+a\mathbb{Z}$ and $[c_0+a-1,c_0)+a\mathbb{Z}$ of transformations $\tilde{R}_{a,c}$ and $R_{a,c}$, respectively. Using the above conclusion, we may fully classify the maximal invariant set $\mathscr{S}_{a,c}$ by two parameters d_1 and d_2, the numbers of holes in $[0,c_0+a-1)$ and $[c_0,a)$, respectively.

Theorem 6. *Let (a,c) satisfy (15) and $a\notin\mathbb{Q}$. Then, $\mathscr{S}_{a,c}\neq\emptyset$ if and only if there exist nonnegative integers d_1 and d_2 such that*

$$(d_1+d_2+1)c_1-c_0+(d_1+1)(1-a)\in a\mathbb{Z}, \tag{16}$$

$$(d_1+1)(1-a)<c_0<1-(d_2+1)(1-a),$$

and

$$\#E_{a,c}=d_1,$$

where

$$m=\frac{(d_1+d_2+1)c_1-c_0+(d_1+1)(1-a)}{a}$$

and

$$\begin{aligned}E_{a,c}=\big\{n\in[1,d_1+d_2+1]\big|\ & n(c_1-m(1-a))\\ &\in[0,c_0-(d_1+1)(1-a))+(a-(d_1+d_2+1)(1-a))\mathbb{Z}\big\}.\end{aligned} \tag{17}$$

The nonnegative integers d_1 and d_2 in Theorem 6 are uniquely determined by the pair (a,c) of positive numbers by (16) and the assumptions that $\lfloor c\rfloor\ge 2$ and $a\notin\mathbb{Q}$. The nonnegative integer parameters d_1 and d_2 in Theorem 6 are indeed the numbers of holes contained in $[0,c_0+a-1)$ and $[c_0,a)$, respectively, and the set of marks is given by

$$\mathscr{K}_{a,c}=\big\{n(c_1-m(1-a))\big\}_{n=1}^{d_1+d_2+1}+(a-(d_1+d_2+1)(1-a))\mathbb{Z}.$$

Finally, we consider the case that the time-spacing parameter a is rational. Recall for $c\notin\gcd(a,1)\mathbb{Z}$, $\mathscr{G}(\chi_{[0,c)},a\mathbb{Z}\times\mathbb{Z})$ is a Gabor frame if and only if both $\mathscr{G}(\chi_{[0,\tilde{c})},a\mathbb{Z}\times\mathbb{Z})$ and $\mathscr{G}(\chi_{[0,\tilde{c}+\gcd(a,1))},a\mathbb{Z}\times\mathbb{Z})$ are Gabor frames, where $\tilde{c}=\lfloor c/\gcd(a,1)\rfloor\gcd(a,1)\in\gcd(a,1)\mathbb{Z}$ [26]. Therefore, we may consider that

$$c\in\gcd(a,1)\mathbb{Z}.$$

In this case, the set $\mathscr{S}_{a,c}$ is finite union of intervals of length $\gcd(a,1)$, and it is completely determined by its restriction to $\gcd(a,1)\mathbb{Z}$. In particular,

$$\mathscr{S}_{a,c} = \mathscr{S}_{a,c} \cap \gcd(a,1)\mathbb{Z} + [0, \gcd(a,1)),$$

because infinite matrices $\mathbf{M}_{a,c}(t), t \in \mathbb{R}$ in (2) satisfy

$$\mathbf{M}_{a,c}(t) = \mathbf{M}_{a,c}(\lfloor t/\gcd(a,1)\rfloor \gcd(a,1)).$$

Let $[\delta', 0) \subset [c_0 - a, 0)$ and $[0, \delta) \subset [0, c_0 + a - 1)$ be maximal intervals contained in the complement $\mathbb{R}\backslash\mathscr{S}_{a,c}$ of the maximal invariant set $\mathscr{S}_{a,c}$. In the case that $\mathscr{S}_{a,c} \neq \emptyset$, at least one of them is equal to zero,

$$\delta\delta' = 0,$$

and there exist nonnegative integers $N \le D$ such that

$$(R_{a,c})^N([c - c_0 + \delta', c - c_0 + 1 - a + \delta) + a\mathbb{Z}) = [c_0 + a - 1 - \delta, c_0 - \delta') + a\mathbb{Z} \quad (18)$$

and

$$(R_{a,c})^{D+1}(c - c_0 + 1 - a + \delta) \in c - c_0 + 1 - a + \delta + a\mathbb{Z} \quad (19)$$

[9]. Without loss of generality, let N and D be minimal nonnegative integers satisfying (18) and (19). In [9], we show that the maximal invariant set $\mathscr{S}_{a,c}$ has its complement consisting of periodic holes of two different sizes, particularly, these holes are $(R_{a,c})^n([c - c_0 + \delta', c - c_0 + 1 - a + \delta) + a\mathbb{Z}), 0 \le n \le N$, of length $1 - a + \delta - \delta'$, and $(R_{a,c})^m([c_0 + a - 1 - \delta, c_0 - \delta')\backslash[c_0 + a - 1, c_0) + a\mathbb{Z}), 1 \le m \le D - N$, of length $\delta - \delta'$. Taking holes-removal surgery described by the map in (13) leads to a line with marks. More interestingly for the case that one of δ, δ' is nonzero, the maximal invariant set $\mathscr{S}_{a,c}$ is the union of mutually disjoint intervals of same size,

$$\mathscr{S}_{a,c} = \mathscr{G}_{a,c} + \big[0, Y_{a,c}(a)/(D+1)\big),$$

and the set of marks $\mathscr{K}_{a,c}$ is a **cyclic group**,

$$\mathscr{K}_{a,c} = \gcd\big(Y_{a,c}(c - c_0 + 1 - a), Y_{a,c}(a)\big)\mathbb{Z} = \frac{Y_{a,c}(a)}{D+1}\mathbb{Z},$$

where

$$\mathscr{G}_{a,c} := \{(R_{a,c})^n(c - c_0 + 1 - a + \delta)\}_{n=0}^{D} + a\mathbb{Z}.$$

Therefore, the maximal invariant set $\mathscr{S}_{a,c}$ can be recovered from the real line by putting marks at appropriate positions and then inserting holes of appropriate sizes at marked positions, even though that augmentation operation is much more delicate and complicated than the hole-removal surgery. Using the augmentation operation, we can parameterize maximal invariant sets $\mathscr{S}_{a,c}$ via four nonnegative integer parameters $d_i, 1 \le i \le 4$, for rational time-spacing parameter a.

Theorem 7. *Let (a,c) satisfy (15), $a \in \mathbb{Q}$ and $c \in \gcd(a,1)\mathbb{Z}$. Then $\mathscr{S}_{a,c} \neq \emptyset$ if and only if the pair (a,c) of positive numbers is one of the following three types:*

(1)$c_0 < \gcd(c_1,a)$.
(2)$1-c_0 < \gcd(c_1+1,a)$.
(3)There exist nonnegative integers d_1,d_2,d_3,d_4 such that

$$0 < B_d := a-(d_1+d_2+1)(1-a) \in (D+1)\gcd(a,1)\mathbb{Z},$$

$$(D+1)c_1+(d_1+d_3+1)(1-a) \in a\mathbb{Z},$$

$$(d_1+d_2+1)((D+1)c_1+(d_1+d_3+1)(1-a))-(d_1+d_3+1)a \in (D+1)a\mathbb{Z},$$

$$\gcd((D+1)c_1+(d_1+d_3+1)(1-a),(D+1)a) = a,$$

$$c_0 = (d_1+1)(1-a)+(d_1+d_3+1)B_d/(D+1)+\gamma$$

for some $\gamma \in (-\min(B_d/(D+1),a-c_0),\min(B_d/(D+1),c_0+1-a))$, and

$$\#E^d_{a,c} = d_1, \tag{20}$$

where $D = d_1+d_2+d_3+d_4+1$ and

$$E^d_{a,c} = \big\{n \in [1,d_1+d_2+1] \big| \, n((D+1)c_1+(d_1+d_3+1)(1-a)) \\ \in (0,(d_1+d_3+1)a)+(D+1)a\mathbb{Z}\big\}. \tag{21}$$

In the above theorem, a hole of large size at the origin is created for the first two cases, while a hole of small size is inserted at the origin for the third case. For the first two cases, no holes of small size have been inserted at any location of marks and the size of holes inserted is always c_0 for the first case and $1-c_0$ for the second case. For the third case, the nonnegative integer parameters d_1,d_2 are indeed the numbers of gaps of size $1-a+|\gamma|$ inserted in $[0,c_0+a-1)$ and $[c_0,a)$, respectively, and the nonnegative integer parameters d_3,d_4 are the numbers of gaps of size $|\gamma|$ inserted in $[0,c_0+a-1)$ and $[c_0,a)$, excluding the one inserted at the origin, respectively.

Using the parametrization of maximal invariant sets $\mathscr{S}_{a,c}$ in Theorems 6 and 7, we can provide full classification of pairs (a,c) satisfying (15) such that the corresponding Gabor system $\mathscr{G}(\chi_{[0,c)},a\mathbb{Z}\times\mathbb{Z})$ is a frame for L^2, see Theorem 11.

6 The *abc*-Problem for Gabor Systems

In this section, we provide a complete answer to the *abc*-problem for Gabor system in [9], or equivalently full classification of all pairs (a,c) of positive numbers such that $\mathscr{G}(\chi_{[0,c)},a\mathbb{Z}\times\mathbb{Z})$ are Gabor frames for L^2.

We start from recalling some known classifications, see for instance [12, 19, 26].

Theorem 8. *Let $a > 0$ and $c > 0$. Then, the following statements hold.*

(I) *If* $a > c$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is not a Gabor frame.*
(II) *If* $a = c$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is a Gabor frame if and only if* $a \le 1$.
(III) *If* $a < c$ *and* $a \ge 1$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is not a Gabor frame.*
(IV) *If* $a < c$ *and* $c \le 1$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is a Gabor frame.*

By Theorem 8, it remains to consider $0 < a < 1 < c$. Applying the new characterization of Gabor frame property in Theorem 1, we take one step forward in our way to solve the *abc*-problem for Gabor systems.

Theorem 9. *Let* $0 < a < 1 < c$ *and set* $c_0 = c - \lfloor c\rfloor$. *Then the following statements hold.*

(V) *If* $c_0 \ge a$ *and* $c_0 \le 1-a$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is a Gabor frame.*
(VI) *If* $c_0 \ge a$ *and* $c_0 > 1-a$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is not a Gabor frame if and only if* $a \in \mathbb{Q}$, *and either*

(1)$c_0 > 1 - \gcd(\lfloor c\rfloor+1, a)$ *and* $\gcd(\lfloor c\rfloor+1, a) \ne (\lfloor c\rfloor+1)\gcd(a,1)$
(2)$c_0 > 1 - \gcd(\lfloor c\rfloor+1, a) + \gcd(a,1)$ *and* $\gcd(\lfloor c\rfloor+1, a) = (\lfloor c\rfloor+1)\gcd(a,1)$.

(VII) *If* $c_0 < a$ *and* $c_0 \le 1-a$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is not a Gabor frame if and only if either*

(3)$c_0 = 0$
(4)$a \in \mathbb{Q}$, $0 < c_0 < \gcd(\lfloor c\rfloor, a)$ *and* $\gcd(\lfloor c\rfloor, a) \ne \lfloor c\rfloor\gcd(a,1)$
(5)$a \in \mathbb{Q}$, $0 < c_0 < \gcd(\lfloor c\rfloor, a) - \gcd(a,1)$ *and* $\gcd(\lfloor c\rfloor, a) = \lfloor c\rfloor\gcd(a,1)$

The statement (V) in the above theorem is given in [26, Section 3.3.3.2]. By Theorems 8 and 9, we may assume that $0 < a < 1 < c$ and $1-a < c_0 < a$ hereafter. Applying maximality of the set $\mathscr{S}_{a,c}$ in Theorem 2, we can move one more step close to answer the *abc*-problem for Gabor systems.

Theorem 10. *Let* $0 < a < 1 < c$ *and* $1-a < c_0 < a$. *Set* $c_1 := c - c_0 - \lfloor((c-c_0)/a)\rfloor a$. *Then the following statements hold.*

(VIII) *If* $\lfloor c\rfloor = 1$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is a Gabor frame.*
(IX) *If* $\lfloor c\rfloor \ge 2$ *and* $c_1 > 2a-1$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is a Gabor frame.*
(X) *If* $\lfloor c\rfloor \ge 2$ *and* $c_1 = 2a-1$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is not a Gabor frame if and only if* $a \in \mathbb{Q}$, $c_0 \le 1 - a + \gcd(a,1)$ *and* $a = (\lfloor c\rfloor+1)\gcd(a,1)$.
(XI) *If* $\lfloor c\rfloor \ge 2$ *and* $c_1 = 0$, *then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is not a Gabor frame if and only if* $a \in \mathbb{Q}$, $c_0 \ge a - \gcd(a,1)$ *and* $a = \lfloor c\rfloor\gcd(a,1)$.

The statement (VIII) in the above theorem can be found in [19, 26]. Using the parameterization of the maximal invariant set $\mathscr{S}_{a,c}$ in Theorems 6 and 7, we make last extremely hard move to solve the *abc*-problem for Gabor systems.

Theorem 11. *Let* (a,c) *satisfy* $0 < a < 1 < c, 1-a < c_0 < a, \lfloor c\rfloor \ge 2$ *and* $0 < c_1 < 2a-1$. *Then the following statement holds.*

(XII) *If* $a \notin \mathbb{Q}$, *then the Gabor system* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z}\times\mathbb{Z})$ *is not a Gabor frame if and only if there exist nonnegative integers* d_1 *and* d_2 *such that*

(a) $a \neq c-(d_1+1)(\lfloor c\rfloor+1)(1-a)-(d_2+1)\lfloor c\rfloor(1-a) \in a\mathbb{Z}$;
(b) $\lfloor c\rfloor+(d_1+1)(1-a) < c < \lfloor c\rfloor+1-(d_2+1)(1-a)$; *and*
(c) $\#E_{a,c}=d_1$, *where* $m=((d_1+d_2+1)c_1-c_0+(d_1+1)(1-a))/a$ *and* $E_{a,c}$ *is given in* (17).

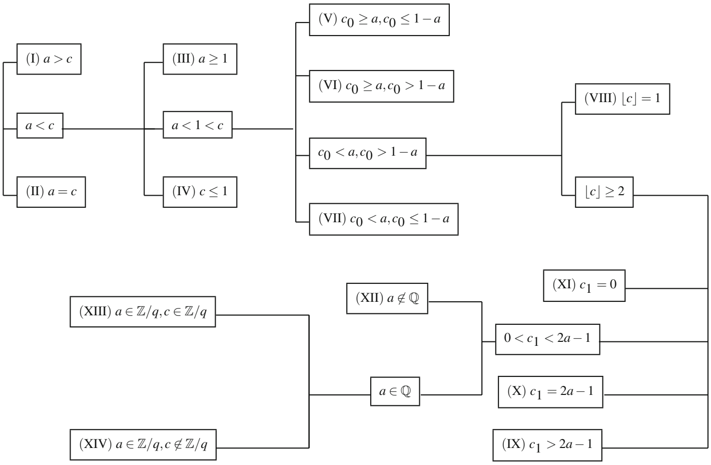
(XIII) If $a\in\mathbb{Q}$ *and* $c\in\gcd(a,1)\mathbb{Z}$, *then* $\mathscr{G}(\chi_{[0,c)},a\mathbb{Z}\times\mathbb{Z})$ *is not a Gabor frame if and only if the pair* (a,c) *satisfies one of the following three conditions:*

(1)$c_0<\gcd(a,c_1)$ *and* $\lfloor c\rfloor(\gcd(a,c_1)-c_0)\neq\gcd(a,c_1)$
(2)$1-c_0<\gcd(a,c_1+1)$ *and* $(\lfloor c\rfloor+1)(\gcd(a,c_1+1)+c_0-1)\neq\gcd(a,c_1+1)$
(3)There exist nonnegative integers d_1,d_2,d_3,d_4 *such that (a)* $0<a-(d_1+d_2+1)(1-a)\in(D+1)\gcd(a,1)\mathbb{Z}$; *(b)* $(D+1)c_1+(d_1+d_3+1)(1-a)\in a\mathbb{Z}$; *(c)* $(d_1+d_2+1)((D+1)c_1+(d_1+d_3+1)(1-a))-(d_1+d_3+1)a\in(D+1)a\mathbb{Z}$; *(d)* $\gcd((D+1)c_1+(d_1+d_3+1)(1-a),(D+1)a)=a$; *(e)* $\#E^d_{a,c}=d_1$; *(f)* $c_0=(d_1+1)(1-a)+(d_1+d_3+1)B_d/(D+1)+\gamma$ *for some some* $\gamma\in(-\min((a-(d_1+d_2+1)(1-a))/(D+1),a-c_0),\min((a-(d_1+d_2+1)(1-a))/(D+1),c_0+1-a))$; *and (g)* $|\gamma|+a/((D+1)\lfloor c\rfloor+(d_1+d_3+1))\neq(a-(d_1+d_2+1)(a-1))/(D+1)$, *where* $D:=d_1+d_2+d_3+d_4+1$ *and* $E^d_{a,c}$ *is defined by* (21).

(XIV) If $a\in\mathbb{Q}$ *and* $c\notin\gcd(a,1)\mathbb{Z}$, *then* $\mathscr{G}(\chi_{[0,c)},a\mathbb{Z}\times\mathbb{Z})$ *is a Gabor frame if and only if both* $\mathscr{G}(\chi_{[0,\tilde{c})},a\mathbb{Z}\times\mathbb{Z})$ *and* $\mathscr{G}(\chi_{[0,\tilde{c}+\gcd(a,1))},a\mathbb{Z}\times\mathbb{Z})$ *are Gabor frames, where* $\tilde{c}=\lfloor c/\gcd(a,1)\rfloor\gcd(a,1)$.

In the Conclusion (XII) of the above theorem, we insert d_1 and d_2 holes contained in intervals $[0,c_0+a-1)$ and $[c_0,a)$, respectively, and put marks at $\cup_{n=1}^{d_1+d_2+1}(n(c_1-m(1-a))+(a-(d_1+d_2+1)(1-a))\mathbb{Z})$. In Case (6) of the Conclusion (XIII), the set $\mathscr{K}_{a,c}$ of marks is $(\gcd(a,c_1)-c_0)\mathbb{Z}$ and holes inserted at marked positions have same length c_0. In Case (7) of the Conclusion (XIII), the set of marks is given by $K_{a,c}=(\gcd(a,c_1+1)+c_0-1)\mathbb{Z}$ and holes inserted are of size $1-c_0$. In Case (8) of the Conclusion (XIII), $K_{a,c}=Y_{a,c}(a)\mathbb{Z}/(D+1)$ and holes inserted at marked positions $lmh+Y_{a,c}(a)\mathbb{Z}, 1\le l\le N$, have size $1-a+|\gamma|$ for $1\le l\le d_1+d_2+1$ and $|\gamma|$ for $d_1+d_2+2\le l\le D$, where $D=d_1+d_2+d_3+d_4+1, Y_{a,c}(a)=(a-(d_1+d_2+1)(1-a))-(D+1)|\gamma|, m=((D+1)c_1+(d_1+d_3+1)(1-a))/a$ and $\gamma=c_0-(d_1+1)(1-a)-(d_1+d_3+1)(a-(d_1+d_2+1)(1-a))/(D+1)$. The Conclusion (XIII) can be found in [26].

Combining Theorems 8–11 gives a complete answer to the *abc*-problem for Gabor systems. The classification diagram of pairs (a,c) in Theorems 8–11 is presented below:


(I) $a > c$
$a < c$
(II) $a = c$
(III) $a \geq 1$
$a < 1 < c$
(IV) $c \leq 1$
(V) $c_0 \geq a, c_0 \leq 1-a$
(VI) $c_0 \geq a, c_0 > 1-a$
$c_0 < a, c_0 > 1-a$
(VII) $c_0 < a, c_0 \leq 1-a$
(VIII) $\lfloor c \rfloor = 1$
$\lfloor c \rfloor \geq 2$
(XI) $c_1 = 0$
$0 < c_1 < 2a-1$
(X) $c_1 = 2a-1$
(IX) $c_1 > 2a-1$
(XII) $a \notin \mathbb{Q}$
$a \in \mathbb{Q}$
(XIII) $a \in \mathbb{Z}/q, c \in \mathbb{Z}/q$
(XIV) $a \in \mathbb{Z}/q, c \notin \mathbb{Z}/q$

From Classifications (V)–(IX) and (XII) in Theorems 9–11, it confirms a conjecture in [26, Section 3.3.5]: *If* $ab < 1 < bc, ab \notin \mathbb{Q}$ *and* $c \notin a\mathbb{Q} + \mathbb{Q}/b$*, then* $\mathscr{G}(\chi_{[0,c)}, a\mathbb{Z} \times b\mathbb{Z})$ *is a Gabor frame for* L^2. This, together with Classification (IV) in Theorem 8 and the shift-invariance, implies that the range of density parameters a, b such that $\mathscr{G}(\chi_I, a\mathbb{Z} \times b\mathbb{Z})$ is a Gabor frame is a dense subset of the open region $\mathscr{U}_c := \{(a,b) : 0 < a < \max(1/b, c)\}$, where c is the length of the interval I.

7 Uniform Sampling of Signals in a Shift-Invariant Space

We say that $\{\phi(\cdot - \lambda) : \ \lambda \in b\mathbb{Z}\}$ generates a *Riesz basis* if there exist positive constants A and B such that

$$A\Big(\sum_{\lambda \in b\mathbb{Z}} |\mathbf{z}(\lambda)|^2\Big)^{1/2} \leq \Big\| \sum_{\lambda \in b\mathbb{Z}} \mathbf{z}(\lambda)\phi(\cdot - \lambda)\Big\|_2 \leq B\Big(\sum_{\lambda \in b\mathbb{Z}} |\mathbf{z}(\lambda)|^2\Big)^{1/2}$$

for all square-summable sequences $(\mathbf{z}(\lambda))_{\lambda \in b\mathbb{Z}}$. For the generator χ_I on an interval $I = [d, c+d)$, $\{\chi_I(\cdot - \lambda) : \lambda \in b\mathbb{Z}\}$ always generates a Riesz basis except that $2 \leq c/b \in \mathbb{Z}$. Thus, except that $2 \leq c/b \in \mathbb{Z}$, any signal f in the shift-invariant space $V_2(\chi_I, b\mathbb{Z})$ can be stably recovered from its equally-spaced samples $f(t_0 + \mu), \mu \in a\mathbb{Z}$, for any initial sampling position t_0 if and only if infinite matrices $\mathbf{M}_{a/b, c/b}(t), t \in \mathbb{R}$, in (2) satisfy the uniform stability (3) on ℓ^2, c.f. [2, 39, 41]. Hence, we have the following almost equivalence between our sampling problem associated with the box generator χ_I and the *abc*-problem for Gabor systems.

Theorem 12. *Let $a,b>0$ and I be an interval with length $c>0$. Then, except that $2\le c/b\in\mathbb{Z}$, the stability requirement (1) holds if and only if $\mathscr{G}(\chi_I, a\mathbb{Z}\times(\mathbb{Z}/b))$ is a Gabor frame for L^2.*

For the case that $I=[d,c+d)$ with $2\le c/b\in\mathbb{Z}$, the shift-invariant space $V_2(\chi_I, b\mathbb{Z})$ is not closed in L^2 and its closure is the shift-invariant space $V_2(\chi_{I'}, b\mathbb{Z})$, where $I'=[d,b+d)$ has length b. Therefore, for the case that $I=[d,c+d)$ with $2\le c/b\in\mathbb{Z}$, any signal f in $V_2(\chi_I, b\mathbb{Z})$ can be stably recovered from equally-spaced samples $f(t_0+\mu), \mu\in a\mathbb{Z}$, for any initial sampling position $t_0\in\mathbb{R}$ if and only if $a\le b$. On the other hand, $\mathscr{G}(\chi_{I/b}, (a/b)\mathbb{Z}\times\mathbb{Z})$ is not a Gabor frame for L^2 by Theorems 8 and 9. This indicates that the equivalence in Theorem 12 does not hold for the case that $2\le c/b\in\mathbb{Z}$ and $a\le b$.

Acknowledgement The research is partially supported by the National Natural Science Foundation of China #11371383 and #10871180, and the National Science Foundation DMS-1109063.

References

1. Aldroubi A, Gröchenig K. Beurling-Landau-type theorems for non-uniform sampling in shift invariant spline spaces. J Fourier Anal Appl. 2000;6:93–103.
2. Aldroubi A, Gröchenig K. Nonuniform sampling and reconstruction in shift-invariant space. SIAM Rev. 2001;43:585–620.
3. Aldroubi A, Sun Q, Tang W-S. Convolution, average sampling, and a Calderon resolution of the identity for shift-invariant spaces. J Fourier Anal Appl. 2005;22:215–44.
4. Baggett LW. Processing a radar signal and representations of the discrete Heisenberg group. Colloq Math. 1990;60/61:195–203.
5. Borichev A, Gröchenig K, Lyubarskii T. Frame constants of Gabor frames near the critical density. J Math Pures Appl. 2010;94:170–82.
6. Casazza P. The art of frame theory. Taiwanese J Math. 2000;4:129–201.
7. Casazza P, Kalton NJ. Roots of complex polynomials and Weyl–Heinsberg frame sets. Proc Am Math Soc. 2002;130:2313–8.
8. Christensen O. An introduction to Frames and Riesz Bases. Boston:Birkhäuser; 2002.
9. Dai X-R, Sun Q, The *abc*-problem for Gabor system, arXiv:1304.7750.
10. Daubechies I. The wavelet transform, time-frequency localization and signal analysis. IEEE Trans Inform Theory. 1990;36:961–1005.
11. Daubechies I, Grossmann A. Frames in the Bargmann space of entire functions. Comm Pure Appl Math. 1988;41:151–64.
12. Daubechies I, Grossmann A, Meyer Y. Painless nonorthogonal expansions. J Math Phys. 1986;27:1271–83.
13. Duffin RJ, Schaeffer AC. A class of nonharmonic Fourier series. Trans Am Math Soc. 1952;72:341–66.
14. Feichtinger HG, Kaiblinger N. Varying the time-frequency lattice of Gabor frames. Trans Am Math Soc. 2004;356:2001–23.
15. Gabor D. Theory of communications. J Inst Electr Eng (London). 1946;93:429–57.
16. Gröchenig K. Foundations of time–frequency analysis. Boston:Birkhäuser; 2001.
17. Gröchenig K, Leinert M. Wiener's lemma for twisted convolution and Gabor frames. J Am Math Soc. 2003;17:1–18.

18. Gröchenig K, Stöckler J. Gabor frames and totally positive functions. Duke Math J. 2013;162:1003–31.
19. Gu Q, Han D. When a characteristic function generates a Gabor frame. Appl Comput Harmonic Anal. 2008;24:290–309.
20. Han D, Wang Y. Lattice tiling and the Weyl–Heisenberg frames. Geom Funct Anal. 2001;11:742–58.
21. He X-G, Lau K-S. On the Weyl–Heisenberg frames generated by simple functions. J Funct Anal. 2011;261:1010–27.
22. Heil C. History and evolution of the density theorem for Gabor frames. J Fourier Anal Appl. 2007;13:113–166.
23. Hutchinson JE. Fractals and self similarity. Indiana Univ Math J. 1981;30(5):713–747.
24. Janssen AJEM. Signal analytic proofs of two basic results on lattice expansion. Appl Comp Harmonic Anal. 1994;1:350–354.
25. Janssen AJEM. Representations of Gabor frame operators, In: Byrnes JS, editor. Twentieth Century Harmonic Analysis–A Celebration, NATO Sci. Ser. II, Math. Phys. Chem., Vol. 33. Dordrecht: Kluwer Academic; 2001, pp. 73–101.
26. Janssen AJEM, Zak transforms with few zeros and the tie In: Feichtinger HG, Strohmer T, editor. Advances in Gabor analysis. Boston: Birkhäuser; 2003, 31–70.
27. Janssen AJEM. On generating tight Gabor frames at critical density. J Fourier Anal Appl. 2003;9:175–214.
28. Janssen AJEM, Strohmer T. Hyperbolic secants yields Gabor frames. Appl Comput Harmonic Anal. 2002;12:259–67.
29. Landau H. On the density of phase space expansions. IEEE Trans Inform Theory. 1993;39:1152–6.
30. Lyubarskii Yu. I. Frames in the Bargmann space of entire functions, In Entire and Subharmonic Functions, Amer. Math. Soc., Providences, RI; 1992, pp. 167–80.
31. Neumann J von. Mathematische Grundlagen der Quantenmechanik. Berlin:Springer; 1932.
32. Rieffel MA. Von Neumann algebras associated with pairs of lattices in Lie groups. Math Ann. 1980;257:403-18.
33. Ron A, Shen Z. Weyl-Heisenberg systems and Riesz bases in $L^2(\mathbb{R}^d)$. Duke Math J. 1997;89:237–82.
34. Seip K. Density theorems for sampling and interpolation in the Bargmann-Fock space I. J Reine Angew Math. 1992;429:91–106.
35. Seip K, Wallstén R. Density theorems for sampling and interpolation in the Bargmann-Fock space II. J Reine Angew Math. 1992;429:107–113.
36. Sun Q. Local reconstruction for sampling in shift-invariant space. Adv Computat Math. 2010;32:335–352.
37. Sun Q, Xian J. Rate of innovation for (non-)periodic signals and optimal lower stability bound for filtering. J Fourier Anal Appl. 2014;20:119–134.
38. Sun W, Zhou X. Characterization of local sampling sequences for spline subspaces. Adv Computat Math. 2009;30:153–75.
39. Unser M. Sampling—50 years after Shannon. Proc IEEE. 2000;88:569–87.
40. Walters P. An introduction to Ergodic theory. Graduate Texts in Mathematics, Vol. 79. New York: Springer; 1982.
41. Walter GG. A sampling theorem for wavelet subspaces. IEEE Trans Inform Theory. 1992;38:881–4.

On Various Levels of Linear Independence for Integer Translates of a Finite Number of Functions

Sandra Saliani

Abstract Systems of integer translates arise in the context of approximation theory, wavelet analysis, and in the theory of shift invariant spaces. After a review of the known properties for integer translates of one square summable function, we explore various levels of linear independence of integer translates of a finite number of functions in terms of properties of the associated Gramian matrix. In some cases, the results are not a straightforward generalization of the case when a single function is considered.

Keywords Linear independence · Integer translates · Gramian · Shift invariant spaces

1 Introduction

Let $\psi_1,\dots,\psi_m \in L^2(\mathbb{R})$. We denote by $\mathscr{B}_m$ the system of all integer translates of $\psi_1,\dots,\psi_m$,

$$\mathscr{B}_m = \{T_k\psi_j, k \in \mathbb{Z}, j = 1,\dots,m\}, \tag{1}$$

where T_k is the translation operator, i.e., $T_k f(x) = f(x-k)$, $x \in \mathbb{R}$.

The study of linear independence for a system of integer translates has crossed various theories in analysis over time. Several works in approximation theory focus on linear independence of compactly supported distributions, such as those of Ron [12], and Jia and Micchelli [6], just to name a few.

A great impulse to research on this subject is due to the theory of shift invariant spaces, which investigates closed subspaces of $L^2(\mathbb{R})$, closed under integer transla-

S. Saliani (✉)
Dipartimento di Matematica, Informatica ed Economia, Università degli Studi della Basilicata, viale dell'ateneo lucano 10, 85100 Potenza, Italia
e-mail: sandra.saliani@unibas.it

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_9

tions. A closed subspace $V \subset L^2(\mathbb{R})$ is shift invariant if

$$f \in V \Longrightarrow T_k f \in V, \quad \text{for all } k \in \mathbb{Z}.$$

Shift invariant spaces are always generated by integer translates of some (at most) countable number of functions, meaning that there is a subset $\Psi \subset L^2(\mathbb{R})$ such that $V = S(\Psi) = \overline{\text{span}}\{T_k\psi,\ \psi \in \Psi,\ k \in \mathbb{Z}\}$. If the set Ψ is finite, say $\Psi = \{\psi_1, \ldots, \psi_m\}$, one simply writes $S(\psi_1, \ldots, \psi_m)$, called finitely generated shift invariant space. If $m = 1$, $S(\psi)$ is a called a principal shift invariant space.

Several works are devoted to the quest for generating sets with desirable properties, including linear independence. Among the first, papers by de Boor, DeVore and Ron [1], and Ron and Shen [13].

All these studies reveal that many properties of a system of translates depend on properties of an invariant associated to it: the periodization function for principal shift invariant spaces, the Gramian matrix for nonprincipal. We premise that we use the Fourier transform defined, for $f \in L^1(\mathbb{R})$, as $\hat{f}(\xi) = \int_{\mathbb{R}} f(x)\, e^{-2\pi ix}\, dx$.

Definition 1. The periodization function associated to $\psi \in L^2(\mathbb{R})$ is the $L^1(\mathbb{T})$ function defined, for a.e. $\xi \in \mathbb{T}$, as $p_\psi(\xi) = \sum_{k\in\mathbb{Z}} |\hat{\psi}(\xi - k)|^2$.

Definition 2. The Gramian associated to functions $\psi_1, \ldots, \psi_m \in L^2(\mathbb{R})$ is the $m \times m$ matrix $G(\xi)$ with entries defined, for a.e. $\xi \in \mathbb{T}$, as

$$(G(\xi))_{i,j} = \sum_{k\in\mathbb{Z}} \hat{\psi}_i(\xi - k)\overline{\hat{\psi}_j(\xi - k)} = [\hat{\psi}_i, \hat{\psi}_j](\xi), \quad i, j = 1, \ldots, m. \tag{2}$$

($[\hat{\psi}_i, \hat{\psi}_j]$ is called the bracket of $\hat{\psi}_i$ and $\hat{\psi}_j$). Note that each entry is in $L^1(\mathbb{T})$. The Gramian is a.e. Hermitian and positive semidefinite.

Linear independence is crucial in frame theory. Frames were introduced by Duffin and Schaeffer [4].

Definition 3. A sequence $(e_n)_{n\in\mathbb{Z}}$ in a Hilbert space H is called a frame for H if there exist two real constants $A, B > 0$ such that, for any $x \in H$

$$A\|x\|^2 \leq \sum_{n\in\mathbb{Z}} |\langle x, e_n\rangle|^2 \leq B\|x\|^2.$$

If only the right hand inequality holds, $(e_n)_{n\in\mathbb{Z}}$ is called a Bessel sequence with Bessel bound B. If $A = B$, we call it an $A-$tight frame and if $A = B = 1$, we call it a Parseval frame.

Frames of translates have been intensely studied since the birth of wavelet theory. The main feature that one looks for in a frame is redundancy, which means linear dependence, responsible of accurate reconstruction algorithms.

In recent years, different levels of linear independence of translates have been investigated following general definitions in a separable Hilbert space.

Definition 4. Let $(e_n)_{n\in\mathbb{Z}}$ be a sequence in a separable Hilbert space H. We say that

(i) $(e_n)_{n\in\mathbb{Z}}$ is linearly independent if every finite subsequence of $(e_n)_{n\in\mathbb{Z}}$ is linearly independent.

(ii)$(e_n)_{n\in\mathbb{Z}}$ is ℓ^2- linearly independent if whenever the series $\sum\limits_{n\in\mathbb{Z}} c_n e_n$ is convergent and equal to zero for some coefficients $(c_n)_{n\in\mathbb{Z}} \in \ell^2(\mathbb{Z})$, then necessarily $c_n = 0$ for all $n \in \mathbb{Z}$.

(iii) $(e_n)_{n\in\mathbb{Z}}$ is ω-independent if whenever the series $\sum\limits_{n\in\mathbb{Z}} c_n e_n$ is convergent and equal to zero for some scalar coefficients $(c_n)_{n\in\mathbb{Z}}$, then necessarily $c_n = 0$ for all $n \in \mathbb{Z}$.

(iv) $(e_n)_{n\in\mathbb{Z}}$ is minimal if for all $k \in \mathbb{Z}$, $e_k \notin \overline{\text{span}}\{e_n,\, n \neq k\}$.

Since we will not always be dealing with unconditionally convergent series, we order $\mathbb{Z} = \{0, 1, -1, 2, -2, \dots\}$ as it is usually done with the Fourier series. For example, convergence of $\sum_{k\in\mathbb{Z}} c_k\, e_k$ in H means convergence of the sequence $(\sum_{|k|\leq n} c_k\, e_k)_{n\in\mathbb{Z}}$, in H.

In this context, we can place works by Nielsen and Šikić [9], Šikić and Speegle [16], Weiss and alt. [5], Paluszyński [10], Šikić and Slamić [15], and Saliani [14]. All these papers deal with a system generated by one function, its properties linked to those of the periodization function and its zero set. They show, also, as many questions on various types of linear independence are tightly connected to classical themes of Fourier analysis, so as to fascinate besides the mere utility.

This work continues the analysis in this direction. After a summary of known results for the system consisting of one function in Sect. 2, we extend the same type of results to a finite number of functions in Sect. 3. We explore linear independence and we show that, differently from the one function case, it is not always achieved.

We then pass to minimality, where a natural extension exists. Finally, we provide sufficient and necessary conditions for the ℓ^2-linear independence, which reduces to the known characterization for one function, even if it is not a complete characterization for more functions. All these results are matched with examples. Many questions remain open, and so the picture is not complete.

2 Known Result for Translates of One Function

Let $\psi \in L^2(\mathbb{R})$. We denote by $\mathscr{B}_\psi$ the system of all integer translates of ψ, i.e.,

$$\mathscr{B}_\psi = \{T_k\psi,\ k \in \mathbb{Z}\}.$$

In this section we recall known results about various type of linear independence for $\mathscr{B}_\psi$. We refer the reader to [5] for many of the proofs, if not stated otherwise.

Theorem 1. *Let $\psi \in L^2(\mathbb{R})$, then the system $\mathscr{B}_\psi$ is linear independent.*

Theorem 2. *The weighted Hilbert space $L^2(\mathbb{T}, p_\psi)$ of all 1-periodic functions $f : \mathbb{T} \to \mathbb{C}$ satisfying*

$$\int_{\mathbb{T}} |f(\xi)|^2 \, p_\psi(\xi) \, d\xi \, < \infty,$$

is isometric to the principal shift invariant space $S(\psi)$.

Theorem 3. *The system $\mathscr{B}_\psi$ is minimal if and only if $\frac{1}{p_\psi} \in L^1(\mathbb{T})$.*

Passing to ℓ^2-linear independence we have

Theorem 4 ([5],[14]). *$\mathscr{B}$ is ℓ^2-linearly independent if and only if $p_\psi(\xi) > 0$ a.e. $\xi \in \mathbb{T}$.*

The proof of the "only if" part in the above result is based on a theorem by Kislyakov [7, Theorem 4], whose proof relies upon a result by Vinogradov [18]. It extend the Correction Theorem of Menchoff,

Theorem 5 ([8]). *Every measurable function becomes a function with uniformly convergent Fourier series after a modification on a set of arbitrarily small measure.*

Let us recall first some basic notations: U^∞ denotes the space of functions $f \in L^\infty(\mathbb{T})$ for which the following norm is finite

$$\|f\|_{U^\infty} = \sup\left\{ \left| \sum_{n \le k \le m} \hat{f}(k)\xi^k \right|, \, n, m \in \mathbb{Z}, \, n \le m, \, \xi \in \mathbb{T} \right\}.$$

Theorem 6. *For every $f \in L^\infty(\mathbb{T})$ with $\|f\|_\infty \le 1$ and every $0 < \varepsilon \le 1$ there exists a function $g \in U^\infty$ with the following properties: $|g| + |f - g| = |f|$; $|\{\xi \in \mathbb{T}, f(\xi) \neq g(\xi)\}| \le \varepsilon \|f\|_1$; $\|g\|_{U^\infty} \le \mathrm{const}(1 + \log(\varepsilon^{-1}))$.*

We shall see that the above result is also crucial in the case of more functions (see Theorem 12).

A further natural refinement of the concept of linear independence is obtained by ℓ^p-linear independence, $1 \le p < 2$. We provide, after the definition, a result by Šikić and Slamić which characterizes ℓ^p-linear independence in case the periodization function is bounded. In this result, we can appreciate a further connection between the properties of the system of integer translates and classical concepts of Fourier analysis, namely ℓ^p-sets of uniqueness.

Definition 5. We say that a sequence $(e_n)_{n\in\mathbb{Z}}$ in a Hilbert space H is ℓ^p-linearly independent, $1 \le p < 2$, if whenever the series $\sum_{n\in\mathbb{Z}} c_n e_n$ is convergent and equal to zero for some coefficients $(c_n)_{n\in\mathbb{Z}} \in \ell^p(\mathbb{Z})$, then necessarily $c_n = 0$ for all $n \in \mathbb{Z}$.

ℓ^2-linear independence implies ℓ^p-linear independence for $1 \le p < 2$. Moreover the latter is linked with the notion of ℓ^p-set of uniqueness.

Definition 6. We call a Lebesgue measurable set $A \subset \mathbb{T}$ an ℓ^p-set of uniqueness, $1 \le p \le 2$, if no nonzero function $f \in L^2(\mathbb{T})$, vanishing almost everywhere in the complement of A, satisfies the condition $(\hat{f}(n))_{n\in\mathbb{Z}} \in \ell^p(\mathbb{Z})$.

ℓ^2-sets of uniqueness coincides with those of measure zero. So it is to be expected that ℓ^p-sets of uniqueness are linked to ℓ^p-linear independence as showed in the following theorem. Actually in the original formulation the authors assume $p_\psi \in L^\infty(\mathbb{T})$, when $1 < p < 2$, but it can be showed that this hypothesis is unnecessary in one direction.

Theorem 7 ([15]). *Let $\mathscr{Z}_\psi = \{\xi,\ p_\psi(\xi) = 0\}$ be the set of zeros of p_ψ. Then the following hold:*

1. *$\mathscr{B}_\psi$ is ℓ^1-linearly independent if and only if the set $\mathscr{Z}_\psi$ is an ℓ^1-set of uniqueness;*
2. *Let $1 < p < 2$. If the set $\mathscr{Z}_\psi$ is an ℓ^p-set of uniqueness, then $\mathscr{B}_\psi$ is ℓ^p-linearly independent;*
3. *Assume $p_\psi \in L^\infty(\mathbb{T})$, and $1 < p < 2$. If $\mathscr{B}_\psi$ is ℓ^p-linearly independent, then the set $\mathscr{Z}_\psi$ is an ℓ^p-set of uniqueness.*

We can summarize in the following scheme all the relations between the various types of independence for $\mathscr{B}_\psi$ and properties of the periodization function:

$$\boxed{\begin{array}{ccl}
\mathscr{B}_\psi \text{ is minimal} & \Longleftrightarrow & \dfrac{1}{p_\psi} \in L^1(\mathbb{T}) \\
\Downarrow & & \\
\mathscr{B}_\psi \text{ is } \omega\text{-independent} & & \\
\Downarrow & & \\
\mathscr{B}_\psi \text{ is } \ell^2\text{-linearly independent} & \Longleftrightarrow & p_\psi(\xi) > 0 \text{ a.e.} \\
\Downarrow & & \\
\mathscr{B}_\psi \text{ is } \ell^p\text{-linearly independent, } 1 < p < 2 & \Longleftarrow & \mathscr{Z}_\psi \text{ is a set of } \ell^p - \text{uniqueness} \\
\Downarrow & & \\
\mathscr{B}_\psi \text{ is } \ell^1\text{-linearly independent} & \Longleftrightarrow & \mathscr{Z}_\psi \text{ is a set of } \ell^1 - \text{uniqueness} \\
\Downarrow & & \\
\mathscr{B}_\psi \text{ is linearly independent} & & \text{Always true}
\end{array}}$$

3 Translates of a Finite Number of Functions

Let $\psi_1, \ldots, \psi_m \in L^2(\mathbb{R})$, and let $\mathscr{B}_m$ the system of all integer translates of $\psi_1, \ldots, \psi_m$, as defined in (1). This section is devoted to understanding how various levels of linear independence for $\mathscr{B}_m$ are reflected on the properties of the Gramian matrix $G(\xi)$ defined in (2). The role of the zero set $\mathscr{Z}_\psi$ will be played by the set where the Gramian is not positive definite.

Since we shall deal with eigenvalues and eigenvectors defined for a.e. $\xi \in \mathbb{T}$, it is convenient to recall the following result:

Theorem 8. *Suppose that D is a measurable set in $\mathbb{R}^k$ and that $A(t)$ is an $m \times m$ Hermitian matrix for $t \in D$. If each element of the matrix is measurable on D, then* *the eigenvalues* $\lambda_j(t)$, $j=1,\dots,m$ *(arranged in increasing order* $\lambda_1(t) \le \lambda_2(t) \le \dots \lambda_m(t)$*), are measurable on D, and corresponding eigenvectors* $Y^j(t)$ *may be chosen so that each* $Y^j(t)$ *is measurable on D.*

Let $A(t)$ be an $m \times m$ Hermitian matrix, defined in a measurable set $t \in D \subset \mathbb{R}$, and $j = 1,\dots,m$ fixed. Any (column) vector measurable function $Y^j(t)$, of eigenvectors corresponding to the eigenvalue $\lambda_j(t)$, will be called an eigenvector function (corresponding to the eigenvalue $\lambda_j(t)$.)

In general, we shall say that a matrix (vector) function is measurable, continuous, polynomial, etc., when each entry of the matrix (vector) possesses the specified property.

The following identity will be useful in the sequel.

Let $c_{j,k}$, $j = 1,\dots,m$, $k \in \mathbb{Z}$, be a sequence of complex numbers. Set, for any $n \in \mathbb{Z}$,

$$X_j^n(\xi) = \sum_{|k|\le n} c_{j,k}\, e^{-2\pi ik\xi}, \quad \xi \in \mathbb{T},$$

and consider the (column) vector $X^n(\xi) \in \mathbb{C}^m$ with entries $X_j^n(\xi)$, $j = 1,\dots,m$.

Then, by a usual periodization technique,

$$\int_0^1 X^n(\xi)^* G(\xi) X^n(\xi)\, d\xi = \| \sum_{j=1}^m \sum_{|k|\le n} c_{j,k} T_k \psi_j \|^2_{L^2(\mathbb{R})}. \tag{3}$$

($X^n(\xi)^*$ means the conjugate transpose of $X^n(\xi)$).

In contrast with $m = 1$, linear independence of translates of more functions is not always assured. The equivalent condition stated in the following theorem is obviously verified when $m = 1$, since positive definiteness means nonzero.

Theorem 9. *The system* $\mathscr{B}_m$ *of all integer translates of* $\psi_1,\dots,\psi_m$*, is linear independent if and only if either the set where* $G(\xi)$ *is positive definite has positive measure or none trigonometric polynomial is an eigenvector for G corresponding to the zero eigenvalue.*

Proof. Consider a finite set of non-zero coefficients $c_{j,k} \in \mathbb{C}$, $|k| \le n$, $j = 1,\dots,m$, such that

$$\| \sum_{j=1}^m \sum_{|k|\le n} c_{j,k} T_k \psi_j \|^2_{L^2(\mathbb{R})} = 0.$$

Then, by (3), the vector $X^n(\xi) \in \mathbb{C}^m$ with polynomial entries

$$X_j^n(\xi) = \sum_{|k|\le n} c_{j,k}\, e^{-2\pi ik\xi}, \quad \xi \in \mathbb{T},$$

verifies

$$\int_0^1 X^n(\xi)^* G(\xi) X^n(\xi)\, d\xi = 0,$$

i.e., $X^n(\xi)^* G(\xi) X^n(\xi) = 0$ a.e. $\xi \in \mathbb{T}$.

The latter is equivalent to say that the minimal eigenvalue of $G(\xi)$ is zero a.e. and $X(\xi)$ is an eigenvector corresponding to it. □

In the following example we show that there exists a system of translates whose Gramian has determinant zero a.e., it is infinitely differentiable, but each normed eigenvector function is not even continuous. It is a modification of an analogue example in [11].

Example 1. Let S be a perfect, nowhere dense on $[0,1]$ that contains $\xi = 0,1$, and which has positive Lebesgue measure. Denote by $I_j = (a_j, b_j)$, $j \in \mathbb{N}$, the infinite sequence of disjoint intervals whose union is the complement of S in $[0,1]$, S^c. For $\xi \in I_j$ let $\theta(\xi) = (\xi - a_j)^{-1} + (b_j - \xi)^{-1}$. Let $\psi_1, \psi_2 \in L^2(\mathbb{R})$ such that

$$\widehat{\psi_1}(\xi) = e^{-\theta(\xi)^2} \sin\theta(\xi)\,\chi_{S^c}\,\chi_{[0,1]}, \quad \widehat{\psi_2}(\xi) = e^{-\theta(\xi)^2} \cos\theta(\xi)\,\chi_{S^c}\,\chi_{[0,1]},$$

where, for any set A, χ_A denotes the characteristic function of A. Then the associated Gramian is $G(\xi) = 0$ for $\xi \in S$, and

$$G(\xi) = e^{-2\theta(\xi)^2} \begin{pmatrix} \sin^2\theta(\xi) & \sin\theta(\xi)\cos\theta(\xi) \\ \sin\theta(\xi)\cos\theta(\xi) & \cos^2\theta(\xi) \end{pmatrix}, \quad \text{for } \xi \in S^c.$$

The smallest eigenvalue of $G(\xi)$ is $\lambda_1(\xi) \equiv 0$ for all $\xi \in \mathbb{T}$.

As shown in [11], $G(\xi)$ is infinitely differentiable on $\mathbb{T}$. Moreover, for $\xi \in S^c$, any normed eigenvector function corresponding to $\lambda_1(\xi)$ is of the form $X_1(\xi) = (c\cos\theta(\xi), -c\sin\theta(\xi))$, where the scalar c has modulus 1. So, no matter what choice of $X_1(\xi)$ is made for $\eta \in S$, the resulting eigenvector function X_1 is discontinuous at η.

By the above result it follows that the system is linear independent.

Example 2. Let $\psi_1, \psi_2 \in L^2(\mathbb{R})$ such that

$$\widehat{\psi_1}(\xi) = \sin(2\pi\xi)\,\chi_{[0,1]}, \quad \widehat{\psi_2}(\xi) = \cos(2\pi\xi)\,\chi_{[0,1]}.$$

Then the Gramian is

$$G(\xi) = \begin{pmatrix} \sin^2(2\pi\xi) & \sin(2\pi\xi)\cos(2\pi\xi) \\ \sin(2\pi\xi)\cos(2\pi\xi) & \cos^2(2\pi\xi) \end{pmatrix}, \quad \text{for } \xi \in \mathbb{T}.$$

The smallest eigenvalue of $G(\xi)$ is $\lambda_1(\xi) \equiv 0$ for all $\xi \in \mathbb{T}$, and any normed eigenvector function corresponding to $\lambda_1(\xi)$ is a trigonometric polynomial of the form $X_1(\xi) = (c\cos t(2\pi\xi), -c\sin(2\pi\xi))$, where the scalar c has modulus 1. Hence the system is not linear independent.

We now pass to show that the nonprincipal shift invariant space $S(\psi_1, \dots, \psi_m)$ is isometric to a Hilbert space of square summable (vector) functions, (compare with Theorem 2).

Definition 7. The vector-valued weighted Hilbert space $L^2(G)$ is defined as the space of all vector valued 1-periodic measurable functions $X : \mathbb{T} \to \mathbb{C}^m$ with the norm

$$\|X\|_{L^2(G)} = \left(\int_0^1 X(\xi)^* G(\xi) X(\xi)\, dx \right)^{1/2} = \left(\int_0^1 \|G(\xi)^{1/2} X(\xi)\|^2_{\mathbb{C}^m}\, dx \right)^{1/2},$$

where $G(\xi)^{1/2}$ denotes the square root of the positive semidefinite Hermitian matrix $G(\xi)$. (Of course we should factorize over the subspace of functions of zero norm).

Lemma 1 ([1]). *Let $g \in L^2(\mathbb{R})$. Then $g \in S(\psi_1, \psi_2, \dots, \psi_m)^\perp$ if and only if $[\hat{g}, \widehat{\psi_j}] = 0$ for all $j = 1, \dots, m$.*

Theorem 10. *The space $L^2(G)$ is isometric to the space $S(\psi_1, \psi_2, \dots, \psi_m)$.*

Proof. Consider the map $J_m : L^2(G) \to L^2(\mathbb{R})$ defined by

$$J_m(X) = \Big(\sum_{j=1}^m X_j(\xi) \widehat{\psi_j}(\xi)\Big)^\vee, \quad X \in L^2(G),$$

where each $X_j(\xi)$ is the j-th entry of $X(\xi)$.

The equality

$$\|X\|^2_{L^2(G)} = \| \sum_{j=1}^m X_j(\xi) \widehat{\psi_j}(\xi) \|^2_{L^2(\mathbb{R})},$$

shows that J_m is well defined and an isometry.

To show that J_m is onto $S(\psi_1, \dots, \psi_m)$, we first consider g in the orthogonal space $S(\psi_1, \psi_2, \dots, \psi_m)^\perp$. If $X \in L^2(G)$, we have, by Lemma 1

$$[\hat{g}, \widehat{J_m(X)}](\xi) = \sum_k \hat{g}(\xi - k) \overline{\sum_{j=1}^m X_j(\xi)\, \widehat{\psi_j}(\xi - k)} = \sum_{j=1}^m \overline{X_j(\xi)}\, [\hat{g}, \widehat{\psi_j}](\xi) = 0.$$

This means that $J_m(L^2(G)) \perp S(g)$, i.e. $J_m(L^2(G)) \subset S(\psi_1, \psi_2, \dots, \psi_m)$.

Conversely, assume there exists a nonzero $g \in S(\psi_1, \psi_2, \dots, \psi_m)$ such that, for all $X \in L^2(G)$,

$$0 = \langle g, J_m(X) \rangle = \langle \hat{g}, \widehat{J_m(X)} \rangle = \int_0^1 \sum_{j=1}^m X_j(\xi)\, [\hat{g}, \widehat{\psi_j}](\xi)\, d\xi.$$

In particular, it is always possible to take X, with bounded entries X_j such that $X_j(\xi)[\hat{g}, \widehat{\psi_j}](\xi) = |[\hat{g}, \widehat{\psi_j}](\xi)|$, so that X belongs to $L^2(G)$, and

$$0 = \int_0^1 \sum_{j=1}^m |[\hat{g}, \widehat{\psi_j}](\xi)|\, d\xi.$$

So $g \in S(\psi_1, \psi_2, \dots, \psi_m)^\perp$ and we get the contradiction $g \equiv 0$. □

Definition 8. Two sequences $(e_n)_{n\in\mathbb{Z}}, (v_n)_{n\in\mathbb{Z}}$, in a separable Hilbert space H are said bi-orthogonal sequences if $\langle e_n, v_m \rangle = \delta_{n,m}$ for all $n, m \in \mathbb{Z}$. In that case $(v_n)_{n\in\mathbb{Z}}$ is called a dual sequence to $(e_n)_{n\in\mathbb{Z}}$, and vice versa.

In general, if a dual sequence exists, it is not necessarily unique. When the Hilbert space is $H = S(\psi_1, \psi_2, \ldots, \psi_m)$, if functions $\tilde{\psi}_1, \ldots, \tilde{\psi}_m \in H$ are such that $\langle \tilde{\psi}_l, T_k\psi_j \rangle = \delta_{j,j}\,\delta_{k,0}$, then, by a change of variable, the system $\widetilde{\mathscr{B}}_m = \{T_k\tilde{\psi}_i, k \in \mathbb{Z}, i = 1, \ldots, m\}$, is biorthogonal to $\mathscr{B}_m$. Since H is generated by all $T_k\psi_j$, if it exists, this dual sequence is unique and it is called the canonical dual to $\mathscr{B}_m$.

It is a known fact, see [17], that the existence of a dual sequence is equivalent to minimality . We can state another equivalent condition which involves the properties of the Gramian matrix (compare with Theorem 3).

Theorem 11. *$\mathscr{B}_m$ has a canonical dual if and only if $G(\xi)$ is a.e. positive definite and all elements in the main diagonal of $G^{-1}(\xi)$ are in $L^1(\mathbb{T})$.*

Proof. Assume first $\tilde{\psi}_1, \ldots, \tilde{\psi}_m \in S(\psi_1, \psi_2, \ldots, \psi_m)$ is a a canonical dual of $\mathscr{B}_m$, and let $X^l \in L^2(G)$ be the unique vector function such that

$$\widehat{\tilde{\psi}_l}(\xi) = \widehat{J_m(X^l)}(\xi) = \sum_{p=1}^{m} X_p^l(\xi)\widehat{\psi_p}(\xi), \quad a.e..$$

By biorthogonality, for all $h \in \mathbb{Z}$, and $l = 1, \ldots, m$,

$$\delta_{h,0}\,\delta_{l,j} = \langle \widetilde{\psi}_l, T_h\psi_j \rangle = \int_0^1 \sum_{p=1}^{m} X_p^l(\xi)\,[\widehat{\psi_p}, \widehat{\psi_j}](\xi)\,e^{2\pi i h\xi}\,d\xi,$$

hence the following identity holds a.e.

$$X^l(\xi)^* G(\xi) \equiv (0, \ldots, \underbrace{1}_{l-\text{th entry}}, \ldots, 0) = e_l^*.$$

The matrix $A(\xi)$ with rows $X^l(\xi)$ verifies $A(\xi)G(\xi) = I$ a.e., the identity matrix, and so $G(\xi)$ is a.e. positive definite.

Finally, from the equality $G^{-1}(\xi)\,e_l = X^l(\xi)$, we get

$$\int_0^1 e_l^* G^{-1}(\xi) e_l\,d\xi = \int_0^1 X^l(\xi)^* G(\xi) X^l(\xi)\,d\xi < +\infty,$$

and so each element in the main diagonal of G^{-1} is in $L^1(\mathbb{T})$.

Conversely, the hypotheses guarantee that the rows of G^{-1}, $X^l(\xi)^* = e_l^*\,G^{-1}(\xi)$, $l = 1, \ldots, m$, are in $L^2(G)$, and so we can define

$$\tilde{\psi}_l = J_m(X^l) \in S(\psi_1, \ldots, \psi_m).$$

The biorthogonality then follows by

$$\begin{aligned}\langle \tilde{\psi}_l, T_k \psi_j \rangle &= \int_0^1 \sum_{p=1}^{m} X_p^l(\xi) [\widehat{\psi_p}, \widehat{\psi_j}](\xi)\, e^{2\pi i k \xi}\, d\xi \\ &= \int_0^1 e_l^* \, G^{-1}(\xi)\, G(\xi)\, e_j \, e^{2\pi i k \xi}\, d\xi = \delta_{l,j}\, \delta_{k,0}.\end{aligned}$$

□

Remark 1. For a finitely generated shift invariant space V, the number

$$\ell(V) = \min\{n \in \mathbb{N}, \exists\, \psi_1, \dots, \psi_n \in V, V = S(\psi_1, \dots, \psi_n)\},$$

is called the length of V, [1], and a set of generators is called a minimal set of generators if it contains exactly $\ell(V)$ elements.

If the Gramian of a finite set $\{\psi_1, \dots, \psi_m\}$ is a.e. positive definite, then

$$\operatorname*{ess\,sup}_{\xi \in \mathbb{T}} \operatorname{rk}(G(\xi)) = m, \tag{4}$$

where $\operatorname{rk}(A)$ denotes the rank of a matrix A. But the left hand side in (4) denotes also the minimal number of generators of the finite shift invariant space $S(\psi_1, \dots, \psi_m)$. It follows that minimality of translates implies that $\{\psi_1, \dots, \psi_m\}$ is a minimal set of generators for the space generated by their translates. Moreover, a minimal set of generators is always linear independent.

Example 3. Let $\psi_1, \psi_2, \psi_3 \in L^2(\mathbb{R})$ be defined by

$$\begin{aligned}\widehat{\psi_1} &= \chi_{[-\frac{1}{2},\frac{1}{2}]} + \chi_{[\frac{1}{2},\frac{3}{2}]} \\ \widehat{\psi_1} &= \chi_{[-\frac{1}{2},\frac{1}{2}]} + 2\chi_{[\frac{3}{2},\frac{5}{2}]} \\ \widehat{\psi_1} &= \chi_{[\frac{1}{2},\frac{3}{2}]} + 3\chi_{[\frac{3}{2},\frac{5}{2}]}\end{aligned}$$

Then the associate Gramian is

$$G(\xi) = \begin{pmatrix} 2 & 1 & 1 \\ 1 & 5 & 6 \\ 1 & 6 & 10 \end{pmatrix}, \quad \text{for } \xi \in \mathbb{T}.$$

Since $G(\xi)$ is constant, and $\det G(t) = 25$, for all $\xi \in \mathbb{T}$, it follows that the system is minimal.

Example 4. Let $\psi_1, \psi_2 \in L^2(\mathbb{R})$ such that

$$\widehat{\psi_1}(\xi) = \sin(2\pi\xi)\, \chi_{[0,1]} + \chi_{[1,2]}, \quad \widehat{\psi_2}(\xi) = \cos(2\pi\xi)\, \chi_{[0,1]}.$$

Then the Gramian is

$$G(\xi) = \begin{pmatrix} \sin^2(2\pi\xi) + 1 & \sin(2\pi\xi)\cos(2\pi\xi) \\ \sin(2\pi\xi)\cos(2\pi\xi) & \cos^2(2\pi\xi) \end{pmatrix}, \quad \text{for } \xi \in \mathbb{T}.$$

The determinant of $G(\xi)$ is $\cos^2(2\pi\xi)$, hence $G(\xi)$ is a.e. positive definite in $\mathbb{T}$. The main diagonal entries of $G^{-1}(\xi)$ are 1 and $\tan^2(2\pi\xi)+\frac{1}{\cos^2(2\pi\xi)}$. It follows that the system of translates of ψ_1,ψ_2 is not minimal.

The following is a natural extension of the result obtained for one function in [14]. We don't have a complete characterization for ℓ^2-linear independence, but it should be noted that, when $m=1$, the additional hypothesis in 2., about the eigenvectors, is always satisfied.

Theorem 12. *Let $G(\xi)$ be the Gramian associated to $\psi_1,\dots,\psi_m \in L^2(\mathbb{R})$. Then*

1. *If $G(\xi)$ is a.e. positive definite then $\mathscr{B}_m$ is ℓ^2-linear independent.*
2. *If $G(\xi)$ is not positive definite in a set of positive measure E, and there exists a measurable eigenvector function corresponding to the zero eigenvalue which agrees with a non-zero trigonometric polynomial a.e. in E, then $\mathscr{B}_m$ is not ℓ^2-linear independent.*

Proof. Assume $G(\xi)$ is a.e. positive definite. Let $c_{i,k}\in\mathbb{C}$, $i=1,\dots,m$, $k\in\mathbb{Z}$, be a nonzero sequence such that $\sum_{i=1}^m\sum_{k\in\mathbb{Z}}|c_{i,k}|^2<+\infty$ and

$$\| \sum_{j=1}^{m}\sum_{|k|\le n} c_{j,k}T_k\psi_j \|^2_{L^2(\mathbb{R})} \underset{n}{\longrightarrow} 0.$$

Now, for j fixed, each $X_j^n(\xi)=\sum_{|k|\le n}c_{j,k}e^{-2\pi ik\xi}$ converges, a.e and in the $L^2(\mathbb{T})$ norm, to a function X_j with Fourier coefficients $(c_{j,-k})_{k\in\mathbb{Z}}\in\ell^2(\mathbb{Z})$. If we consider the vector function $X=(X_1,\dots,X_m)^T$, it follows from (3) that

$$\int_0^1 X^n(\xi)^*G(\xi)X^n(\xi)\,d\xi \underset{n}{\longrightarrow} 0,$$

so that

$$\int_0^1 X(\xi)^*G(\xi)X(\xi)\,d\xi = 0.$$

Note that $X(\xi)\neq 0$ in a set E of positive measure, otherwise all coefficients $c_{j,k}$ would be zero. The above equality then implies that $X(\xi)^*G(\xi)X(\xi)=0$, for all $\xi\in E$, and so G is not a.e. positive definite.

Conversely assume G is not a.e. positive definite. Since $G(\xi)$ is always a.e positive semidefinite, it means that there exists a measurable set E, $|E|>0$, such that the smallest eigenvalue $\lambda_{\min}(\xi)$ of $G(\xi)$ is zero for all $\xi\in E$. Let us denote by $X(\xi)$ the eigenvector corresponding to $\lambda(\xi)$, which agrees with a nonzero trigonometric polynomial a.e in E. Hence

$$X(\xi)^*G(\xi)X(\xi)=0,\quad a.e.\ \xi\in E.$$

Since $X(\xi)$ is a trigonometric polynomial, each entry X_j is a polynomial, say

$$X_j(\xi)=\sum_{|h|\le p_j}a_{j,h}e^{2\pi ih\xi},\quad j=1,\dots,m,\ a.e.\ \xi\in E,$$

and adding zero coefficients, if needed, we can always assume $p_j = p$, for all $j = 1, \dots, m$.

Note that if $|E| = 1$ then $\mathscr{B}_m$ is obviously not ℓ^2-linear independent. So we can assume $|E^c| > 0$.

Consider the characteristic function of E, χ_E. Then $g \in U^\infty$ provided by Theorem 6, corresponding to $\varepsilon = 1/2$, is not zero, since otherwise we get the contradiction $|\{\xi \in \mathbb{T}, \chi_E(\xi) \neq 0\}| = |E| \leq \frac{|E|}{2}$; the support of g is contained in E since $|g| \leq |\chi_E|$; and finally $\|g\|_{U^\infty} \leq \text{const}(1 + \log 2)$ implies that the partial sums of Fourier series of g, $S_n g(\xi)$, are uniformly bounded in the uniform norm.

Moreover $S_n g(\xi) \to g(\xi)$, a.e. and so

$$|S_n g(\xi)|^2 X(\xi)^* G(\xi) X(\xi) \to_n |g(\xi)|^2 X(\xi)^* G(\xi) X(\xi), \quad a.e.\ \xi \in \mathbb{T},$$

but the latter function is identically zero since $g(\xi) = 0$ a.e. in E^c and $X(\xi)^* G(\xi) X(\xi) = 0$ a.e. in E.

Hence by the uniform boundedness of $S_n g(\xi)$ and the dominated convergence theorem it follows that

$$\int_0^1 |S_n g(\xi)|^2 X(\xi)^* G(\xi) X(\xi)\, d\xi \to 0.$$

Note that, for big n, each entry in $S_n g(\xi) X(\xi)$ is

$$\begin{aligned}\sum_{|k|\leq n} \sum_{|h|\leq p} \hat{g}(k)\, a_{j,h}\, e^{2\pi i(k+h)\xi} &= \sum_{l=-n+p}^{n-p} \sum_{|h|\leq p} \hat{g}(l-h)\, a_{j,h}\, e^{2\pi i l\xi} \\ &+ \sum_{l=n-p+1}^{n+p} \sum_{h=l-n}^{p} \hat{g}(l-h)\, a_{j,h}\, e^{2\pi i l\xi} \\ &+ \sum_{l=-n-p}^{-n+p-1} \sum_{h=-p}^{l+n} \hat{g}(l-h)\, a_{j,h}\, e^{2\pi i l\xi}.\end{aligned}$$

If we set

$$c_{j,l} = \sum_{|h|\leq p} \hat{g}(l-h)\, a_{j,h},$$

then obviously $\sum_{j=1}^m \sum_{l\in\mathbb{Z}} |c_{j,l}|^2 < +\infty$, and we get by (3) that the following sum of translates

$$\begin{aligned}\sum_{j=1}^m \sum_{|l|\leq n-p} c_{j,l} T_l \psi_j &+ \sum_{j=1}^m \sum_{l=n-p+1}^{n+p} \sum_{h=l-n}^{p} \hat{g}(l-h)\, a_{j,h}\, T_l \psi_j \\ &+ \sum_{j=1}^m \sum_{l=-n-p}^{-n+p-1} \sum_{h=-p}^{l+n} \hat{g}(l-h)\, a_{j,h}\, T_l \psi_j\end{aligned}$$

goes to zero in the $L^2(\mathbb{R})$ norm, as $n \to +\infty$.

On the other hand, each tail in the above sum goes to zero in the $L^2(\mathbb{R})$ norm, as $n \to +\infty$. For example, after a change of variable,

$$\int_{\mathbb{R}} |\sum_{j=1}^{m} \sum_{l=n-p+1}^{n+p} \sum_{h=l-n}^{p} \hat{g}(l-h)\, a_{j,h}\, \psi_j(x-l)|^2\, dx$$

$$= \int_{\mathbb{R}} |\sum_{j=1}^{m} \sum_{k=-p+1}^{p} \sum_{h=k}^{p} \hat{g}(k+n-h)\, a_{j,h}\, \psi_j(x-k)|^2\, dx \longrightarrow_n 0,$$

for $\hat{g}(k+n-h) \to_n 0$, and the dominate convergence theorem.

Finally we deduce

$$\| \sum_{j=1}^{m} \sum_{|l| \le n-p} c_{j,l} T_l \psi_j \|_{L^2(\mathbb{R})} \to_n 0,$$

and so $\mathscr{B}_m$ is not ℓ^2-linear independent. □

References

1. de Boor C, DeVore RA, RonA. The structure of finitely generated shift-invariant spaces in $L_2(\mathbb{R}^d)$. J Funct Anal. 1994;119:37–78.
2. Christensen O. An introduction to frames and Riesz bases. Boston:Birkhauser; 2003.
3. Daubechies I, Grossmann A, Meyer Y. Painless nonorthogonal expansions. J Math Phys. 1986;27:1271–83.
4. Duffin RJ, Schaeffer AC. A class of nonharmonic Fourier series. Trans AMS. 1952;72:341–66.
5. Hernández E, Šikić H, Weiss G, Wilson E. On the properties of the integer translates of a square integrable function. Harmonic analysis and partial differential equations. Contemp Math. 2010;505:233–49.
6. Jia RQ, Micchelli CA. On linear independence for integer translates of a finite number of functions. Proc Edinb Math Soc. 1993;36:69–85.
7. Kislyakov SV. A sharp correction theorem. Studia Math. 1995;113:177–96.
8. Menchoff D. Sur la convergence uniforme des séries de Fourier. Rec Math [Mat. Sbornik] N S. 1942;53:67–96.
9. Nielsen M, Šikić H. Schauder bases of integer translates. Appl Comput Harmon Anal. 2007;23:259–62.
10. Paluszyński M. A note on integer translates of a square integrable function on $\mathbb{R}$. Colloq Math. 2010;118:593–7.
11. Reid WT. Some elementary properties of proper values and proper vectors of matrix functions. SIAM J Appl Math. 1970;18:259–66.
12. Ron A. A necessary and sufficient condition for the linear independence of the integer translates of a compactly supported distribution. Constr Approx. 1989;5:297–308.
13. Ron A, Shen Z. Frames and stable bases for shift-invariant subspaces of $L_2(\mathbb{R}^d)$. Can J Math. 1995;47:1051–94.
14. Saliani S. ℓ^2-linear independence for the system of integer translates of a square integrable function. Proc Amer Math Soc. 2013;141:937–41.
15. Šikić H, Slami´c I. Linear independence and sets of uniqueness. Glas Mat. 2012;47:415–20.

16. Šikić H, Speegle D. Dyadic PFW's and W_o-bases. In: Muić G, editor. Functional analysis IX. Aarhus University:Aarhus; 2007.
17. Singer I. Bases in Banach spaces I. Berlin:Springer; 1970.
18. Vinogradov SA. A strengthened form of Kolmogorov's theorem on the conjugate function and interpolation properties of uniformly convergent power series. In: Spectral theory of functions and operators II. A translation of Trudy Mat Inst Steklov. 1981;155:7–40. Proc. Steklov Inst. Math. no. 1, Providence:American Mathematical Society; 1983.

Part XII
RADAR and Communications: Design, Theory, and Applications

This part contains four chapters devoted to applications of space, time, and frequency signal modeling. The contributions are written by leading experts from academia and industry.

WIESE together with ROSCA AND CLAUSSEN study the Maximum Likelihood Estimator for Directions of Arrival in the case of an unknown number of moving wideband sources. They study a reversible jump particle filter to implement a Monte Carlo method for computing the posterior distributions associated to multiple models. The authors use a mixed Bayesian-frequentist approach to deal with model selection (unknown number of sources) and a nonstationary environment (moving sources). Their simulations compare the particle filter algorithm performance with an iterative procedure based on subspace methods with known number of sources. Using 2000 particles yield a particle filter and kernel based Maximum A Posteriori estimator with superior performance than of subspace methods.

BRODZIK presents a new design of L polyphase (unit magnitiude) cyclic sequences of length N (with N/L integer) that are simultaneously Golay, zero correlation zone (ZCZ), and have all-zero cross-correlation. The design is performed in finite Zak transform domain and it extends two previous constructions, one by the same author, the other by Mow. Next the author analyzes the computational complexity of this algorithm and shows it is of the form N^2/L^2+2N.

LEVITAS and NUNN survey the state-of-the-art radar scene in waveform design and waveform processing. First the authors describe traditional design requirements: maximizing detection sensitivity (or, equivalently, minimizing the loss), maintaining required range resolution, minimizing eclipsing by a large discrete or by large distributed clutter, decreasing Doppler sensitivity, spectral compliance, and generating multiple uncorrelated codes. Next the authors present examples of traditional codes and analyze their performance. Specifically they look at linear FM codes, Lewis-Kretschmer Palindromic P4 codes, nonlinear FM codes, Barker codes, and pseudorandom noise codes. Then the authors consider more stringent design requirements accounting for increased clutter background levels, higher spectral occupancy, and multiple antenna systems (MIMO waveforms).

The last chapter of this part is by STROHMER and WANG , which presents a compressive sampling based MIMO radar system. The setup assumes multiple radar returns measured by a multiple-transmit-multiple-receive antenna system. The authors formulate a regularized least square estimator with a l^1-norm penalty of multiple target return coefficients. They provide probabilistic guarantees of correct detection. Next the authors analyze the performance of compressive MIMO radars using Kerdock waveforms. Again they provide probabilistic guarantees of correct detection. Then the authors consider off-grid errors and perform a sensitivity analysis. In the last part of the chapter, the authors present numerical examples of the system performance. Specifically they obtain ROC curves by solving the LASSO problem for transmit Kerdock waveforms and addditive white Gaussian noise with variable SNR levels. Their simulations show superior performance of this scheme.

Polyphase Golay Sequences with Semi-Polyphase Fourier Transform and All-Zero Crosscorrelation: Construction B

Andrzej K. Brodzik

Abstract A new design of polyphase cyclic sequence sets of size L and sequence length $N = LM$, where $L, M \in \mathbb{Z}$, that are simultaneously Golay, zero correlation zone (ZCZ), and have all-zero crosscorrelation, is described. The sequences have a semi-polyphase Fourier transform with a constant nonzero magnitude at M points and zero magnitude at $(L-1)M$ points. The design takes place in Zak space. The use of Zak space setting enables selection of a desirable Fourier transform zero placement, reduces computations, and links the Golay sequence design with the design of perfect sequences.

Keywords Bat chirp · Complementary sequences · Correlation · Discrete chirp · Fourier transform · Fourier transform support · Golay sequences with all-zero crosscorrelation · Perfect sequences · Polyphase · Zak transform · Zak transform support · Zero correlation zone

1 Introduction

Polyphase or unimodular sequences with good correlation properties have use in several applications, including, among others, wireless communications, multistatic radar, and cryptography [11, 12, 14, 18]. In general, design of these sequences is a difficult combinatorial-analytic task [19, 20]. This difficulty can be mediated by narrowing focus of investigation to some special cases. These cases include: (1) the sets of sequences with ideal zero out-of-phase autocorrelation and minimum constant crosscorrelation, called *perfect sequence sets* (PSS), (2) the sets of sequences with both auto and crosscorrelation assuming a zero value for delays contained in the *zero correlation zone* (ZCZ), called *ZCZ sequence sets*, and (3) the sets of Golay

A. K. Brodzik (✉)
Independent Scholar, Boston, USA
e-mail: vespertilionoidea@yahoo.com

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_10

or complementary sequences whose individual autocorrelations add up to an ideal autocorrelation. In this work we also focus on a special case, on certain sets of polyphase sequences with all-zero crosscorrelation, and show that they are both Golay and ZCZ. We treat this case as an integral part of a general sequence design framework based on the Zak space (ZS) methods, which reduces computations and links the design with prior constructions [7, 10]. Before describing the results of this investigation in greater detail, we will formally define some basic terms.

Two N-point complex valued sequences $x^{(0)}$ and $x^{(1)}$ are *complementary* or *Golay*, if their aperiodic autocorrelations, $\hat{z}^{(0)}$ and $\hat{z}^{(1)}$, sum up to an ideal aperiodic autocorrelation [13], i.e.,

$$\hat{z}^{(0)}(n)+\hat{z}^{(1)}(n)=0 \text{ for all } n\neq 0, \tag{1}$$

where *aperiodic autocorrelation* is given by

$$\hat{z}^{(s)}(n)=\sum_{m=0}^{N-1-n} x^{(s)}(m)(x^{(s)})^{*}(m+n),\ \ 0\leq n<N,\ \ s=0,1. \tag{2}$$

More precisely, two N-point sequences satisfying condition (1) are called an *aperiodically complementary sequence pair* (ACSP). Two generalizations of ACSP are of interest here. First, the complementarity property can be extended in an obvious way to an *aperiodically complementary sequence set* (ACSS) of size T. Second, ACSPs and ACSSs can be viewed as special cases of *periodically complementary sequence pairs* (PCSPs) and *periodically complementary sequence sets* (PCSSs) [3]. For PCSP a condition analogous to (1) holds, involving *periodic autocorrelations*, defined by

$$z^{(s)}(n)=\frac{1}{N}\sum_{m=0}^{N-1} x^{(s)}(m)(x^{(s)})^{*}(m-n),\ 0\leq n<N,\ s=0,1, \tag{3}$$

where $m-n$ is taken modulo N. ACSSs and PCSSs are linked by the equation

$$z^{(s)}(n)=\hat{z}^{(s)}(n)+\hat{z}^{(s)}(n-N),\ \ 0\leq n<N,\ \ 0\leq s<T. \tag{4}$$

Extending the condition (1) to the doubly general case, we have that every PCSS of L sequences of length N, $\{x^{(0)},x^{(1)},...,x^{(L-1)}\}$, satisfies the Golay condition

$$z^{(0)}(n)+z^{(1)}(n)+...+z^{(L-1)}(n)=\begin{cases} L, & n=0,\\ 0, & \text{else.}\end{cases} \tag{5}$$

The Golay condition (5) has convenient ZS and Fourier space (FS) counterparts, i.e.,

$$\sum_{l=0}^{L-1}|Z_L^{(l)}(j,k)|=\begin{cases} L,\ k=0,\ \ 0\leq j<L,\\ 0,\ \text{else,}\end{cases} \tag{6}$$

and

$$\sum_{l=0}^{L-1} |\mathbf{x}^{(l)}(m)| = \sqrt{LN}, \;\; 0 \le m < N, \tag{7}$$

respectively, where, $N = LM$, $L, M \in \mathbb{Z}$, $\mathbf{x}$ is the discrete Fourier transform (DFT) of x, and X_L is the finite Zak transform (FZT) of x. FZT is a two-dimensional time-frequency representation of x, closely related to DFT. We will formally introduce FZT in the next section. We note that the FS Golay condition (7), in contrast with the time and ZS conditions, (5) and (6), is expressed in terms of sequences rather than their autocorrelations. As the DFT and the FZT of the sequences considered here are especially closely linked, the ZS Golay condition (6) provides convenient means of manipulating the DFT support structure directly by the ZS methods.

Our point of departure in the PCSS design is one of the best known polyphase sequences, the discrete chirp. Several former results relevant to discrete chirps are fundamental and need to be recalled here. First, a discrete chirp is N-periodic if its parameters, the normalized versions of the chirp rate $\bar{a}$ and the carrier frequency $\bar{b}$, conform to a certain modest restriction [1]. Second, the N-periodicity condition is equivalent to the ZS support condition for chirps, which specifies when the FZT of a discrete chirp is semi-polyphase, i.e., with zero or constant nonzero magnitude [5]. Furthermore, a discrete chirp that satisfies the N-periodicity condition has a DFT with either semi-polyphase or polyphase support, depending on co-primality of $\bar{a}$ and N [4].

It follows from this and from the DFT and FZT conditions for Golay sets, (7) and (6), that

- A discrete chirp with a semi-polyphase FZT has a polyphase DFT, when $(\bar{a}, N) = 1$. The discrete chirp is then a perfect sequence called a *bat chirp* [10]. The set of all N-point bat chirps having identical values of L, M, and $\bar{b}$, but distinct $\bar{a}$, the latter satisfying the above co-primality condition, is a PSS.
- Alternatively, a discrete chirp with a semi-polyphase FZT has a semi-polyphase DFT, when $(\bar{a}, N) = c \neq 1$. In this case we call (slightly abusing the language) the discrete chirp a *Golay sequence*, or a *Golay chirp*. The set of all N-point Golay chirps associated with identical values of L and M, and supported on appropriately specified nonoverlapping sets of FZT rows, is a PCSS.

This taxonomy is important because while not all polyphase Golay sequences have a semi-polyphase DFT, those who do have particularly favorable properties.

In [6, 7] a new construction of PCSS, called construction A, was introduced. This construction is associated with the $L \times M$, $M = \sqrt{L}$, FZT lattice, and is given by a collection of $\sqrt{L}$ bat chirps, appropriately zero-padded in ZS. The construction has the combined properties of a ZCZ sequence set, all-zero crosscorrelation, highly sparse semi-polyphase FZT and DFT, and a high degree of set parameterization, permitting further sequence specialization.

Here, we describe another special case of PCSS contained in the above taxonomy, called the construction B. As construction A, this new construction is performed in ZS and, as construction A, it has the properties of ZCZ sequence set, all-zero crosscorrelation, and highly sparse semi-polyphase FZT and DFT. Unlike construction A, individual sequences of construction B are given by FZTs of discrete chirps whose supports are restricted to single rows and whose DFT supports are combs. This special FZT and DFT support structure is achieved by appropriately specializing the ZS support condition. In comparison with the construction A, the construction B offers a wider choice of dimensions of the FZT lattice, achieves the upper bound on the product of the $|ZCZ|$ and the sequence set size, and reduces computational complexity of crosscorrelation approximately by a factor $\frac{\log_2 N}{3}$, when compared with construction A, and by a factor of $\frac{2\log_2 N+1}{2+M/L}$, when compared with direct FS implementation. These advantages are acquired at the cost of reduced number of available PCSSs and larger alphabet size. Constructions A and B are the two principal PCSS constructions for a non-prime N. Several other, special ZS constructions of PCSS are possible, but will be considered elsewhere.

Former work on Golay sequences is quite extensive. For brevity we will mention here only a few key contributions directly related to our work: the original Golay paper [13], an introduction to the periodic codes [3], a similar design obtained by time domain analysis [17], and our prior ZS construction [6–7]. An extensive bibliography on perfect and Golay sequences is given in [14] and [18].

The content of the chapter is as follows. Section 2 provides overview of FZT. Section 3 introduces the finite chirp. Section 4 describes the construction B and its main properties. Section 5 compares construction B with relevant prior designs.

2 FZT

Take x to be any N-periodic sequence in $\mathscr{C}^N$ and set $e_N(n) := e^{2\pi in/N}$, the N-th root of unity. The discrete Fourier transform (DFT) of x is

$$\mathbf{x}(m) = \sum_{n=0}^{N-1} x(n)e_N(nm), \ \ 0 \leq m < N. \tag{8}$$

Suppose for the remainder of this chapter that $N = LM$, where L and M are positive integers, and set $n = k + rM, \ \ m = j + sL, \ \ 0 \leq k, s < M, \ \ 0 \leq r, j < L$. Then the values of $\mathbf{x}(m)$ on a decimated set, say, $m = j + 0L$, are given by

$$\mathbf{x}(j) = \sum_{k=0}^{M-1} e_N(jk) \sum_{r=0}^{L-1} x(k+rM) e_L(rj). \tag{9}$$

The inner sum in (8) is called the *finite Zak transform* (FZT) of x [22],

$$X_L(j,k) = \sum_{r=0}^{L-1} x(k+rM) e_L(rj). \tag{10}$$

It follows that computing X_L requires M L-point DFTs of the sequences

$$x(k), x(k+M), \ldots, x(k+(L-1)M), \;\; 0 \leq k < M. \tag{11}$$

For $L = N$ and $M = 1$ the FZT X_N is identical with the DFT of x. For $L = 1$ and $M = N$ the FZT X_1 is identical with x. The FZT has several applications in mathematics, quantum mechanics and signal analysis. Both the continuous Zak transform and the FZT play a major role in the analysis of time–frequency representations, including ambiguity functions and Weyl-Heisenberg expansions. Here, we will state, without proofs, only a few basic properties of the FZT. For a more extensive review of Zak transform theory and a historical background, the reader is referred to [16].

Like the DFT, the FZT is a one-to-one mapping. A signal x can be recovered from its FZT by the formula

$$x(k+rM) = L^{-1} \sum_{j=0}^{L-1} X_L(j,k) e_L(-rj). \tag{12}$$

Define the inner product on $\mathscr{C}^L \times \mathscr{C}^M$ by

$$\langle X_L, Y_L \rangle = \sum_{j=0}^{L-1} \sum_{k=0}^{M-1} X_L(j,k) Y_L^*(j,k). \tag{13}$$

We have

$$||x||^2 = \frac{1}{L} ||X_L||^2. \tag{14}$$

Up to a scale factor, the FZT is a linear isometry from $\mathscr{C}^N$ onto $\mathscr{C}^L \times \mathscr{C}^M$. The FZT is periodic in the frequency variable and quasiperiodic in the time variable, i.e.,

$$X_L(j+L,k) = X_L(j,k), \tag{15}$$

and

$$X_L(j,k+M) = X_L(j,k) e_L(-j). \tag{16}$$

A related property describes the FZT of time and frequency shifts of x. Set $y(n) = x(n-c)$ and $z(n) = x(n)e_N(dn)$, where $0 \le c < M$ and $0 \le d < L$. Then the FZTs of y and z are given by

$$Y_L(j,k) = X_L(j,k-c) \tag{17}$$

and

$$Z_L(j,k) = X_L(j+d,k)e_N(dk). \tag{18}$$

Consider the *cyclic crosscorrelation* of two N-periodic polyphase sequences, x and y, given by

$$z(n) = (y \odot x)_n := \frac{1}{N}\sum_{m=0}^{N-1} y(m)x^*(m-n), \quad 0 \le n < N, \tag{19}$$

where $m-n$ is taken modulo N. When $y = x$, the cyclic crosscorrelation is called the *cyclic autocorrelation*. Take X_L, Y_L, and Z_L to be the FZTs of x, y, and z in (19), respectively. Then

$$Z_L(j,k)=\frac{1}{N}\sum_{l=0}^{M-1} Y_L(j,l)X_L^*(j,l-k). \tag{20}$$

The result of this operation can be viewed as an assembly of L M-point time domain crosscorrelations performed on frequency slices of the $L \times M$ Zak space (ZS) signals, X_L and Y_L. The ZS correlation formula was previously used to reveal an intimate relationship of PSSs with certain permutation sequences, and resulting in replacement of the analysis of PSSs with the analysis of permutations [10], and in our prior work on Golay sequences [7]. The construction of the Golay sequence set given here can be viewed as an extension of this work.

3 The Finite Chirp

Consider the *discrete chirp*

$$x(n) = e_{L^2}\left(\frac{an^2}{2}\right)e_L(bn), \quad 0 \le n < N, \tag{21}$$

where a is the *discrete chirp rate*, b is the *discrete carrier frequency*, and a, $b \in \mathbb{R}$. To compactify expressions, we will use the following normalized chirp parameters, $\bar{a} = aK$, $\bar{\bar{a}} = aK^2$ and $\bar{b} = bK$, where $K = M/L$.

Take $n = k + rM$, $0 \leq k < M$, $0 \leq r < L$, as in Sec. 2. Then (21) can be expressed as

$$x(k+rM) = e_N\left(\frac{\bar{a}k^2}{2} + \bar{b}Lk + \bar{\bar{a}}Lkr\right) e\left(\frac{\bar{\bar{a}}r^2}{2} + \bar{b}r\right). \tag{22}$$

We impose two conditions on $x(n)$. First, we require that $x(n)$ be *periodic* with period N, i.e.,

$$x(n+N) = x(n). \tag{23}$$

Second, we require that X_L has a *semi-polyphase support*, i.e.,

$$|X_L(j,k)| = \begin{cases} A,\ (j,k) \in \mathrm{supp}(X_L) \subset \mathbb{N}_L \times \mathbb{N}_M, \\ \\ 0,\ \text{else}, \end{cases} \tag{24}$$

where $A = \frac{||X_L||_2}{\sqrt{\mathscr{S}(X_L)}} \in \mathbb{R}$, $\mathscr{S}(X_L)$ is the cardinality of the support of X_L, and "$\subset$" denotes "a proper subset of".

The periodicity and semi-polyphase support conditions facilitate development of highly efficient and flexible algorithms for chirp and chirp-like signal processing [1, 4–10].

Both, periodicity and semi-polyphase support, restricts the values of $\bar{a}$, $\bar{\bar{a}}$, $\bar{b}$, and L.

The periodicity condition

$$\bar{a} \in \mathbb{Z} \text{ and } \frac{\bar{\bar{a}}L^2}{2} + \bar{b}L \in \mathbb{Z}, \tag{25}$$

follows directly from (22). The conditions for semi-polyphase support of X_L are somewhat more difficult to obtain. In [5] several conditions of semi-polyphase support were derived. The most general of these conditions is given by

$$\bar{a} \in \mathbb{Z},\ \bar{\bar{a}} = \frac{n}{d} \in \mathbb{Q} \text{ and } \frac{nL}{2} + \bar{b}L \in \mathbb{Z}. \tag{26}$$

Since $M \in \mathbb{Z}$, $d|L$, and hence it follows that the conditions (25) and (26) are equivalent.

Theorem 1. *A discrete chirp is periodic iff the FZT of a discrete chirp is semi-polyphase.*

4 Construction B

In this section a new Zak space construction of **p**olyphase (in time), **s**emi-polyphase (in FS), **Z**CZ, **a**ll-zero crosscorrelation, cyclic **G**olay sequence sets (PSZAG) is described. We begin with a suitable restriction of the ZSC (26).

4.1 Restriction of ZSC

Consider the sequence set

$$\mathscr{S}_{\gamma,P} = \{x^{(\gamma,p)},\ p \in \mathbb{Z}_L\}, \tag{27}$$

for some $\gamma \in \mathbb{Z}$, $(\gamma, M) = P$, of polyphase sequences (22), with parameters a and b given either by

$$a = \gamma L^2/M,\ \ b = p/M + L/2M,\ \ \gamma,\ L,\ M \in \mathbb{Z}_{odd},\ \ p \in \mathbb{Z}_L, \tag{28}$$

or

$$a = \gamma L^2/M,\ \ b = p/M,\ \ \gamma,\ L,\ M \in \mathbb{Z},\ \ \gamma M \in \mathbb{Z}_{even},\ \ p \in \mathbb{Z}_L. \tag{29}$$

In either case

$$\bar{a} = \gamma L \in \mathbb{Z}\ \text{ and }\ \frac{\bar{a}L^2}{2} + bM = (\gamma L^2/2 + b)M \in \mathbb{Z}, \tag{30}$$

and hence the ZSC (26) is satisfied. As will be seen later, since (28–29) is a restriction of ZSC (26), $x^{(\gamma,p)}$ is a special case of the finite chirp x, equipped with some additional desirable properties, however it is not a bat chirp, as $(\bar{a}, N) = LP \neq 1$.

Suppose, for convenience, that the following extension of the condition (28) holds:

$$a = \gamma L^2/M,\ \ b = p/M + L/2M,\ \ \gamma,\ L,\ M,\ p \in \mathbb{Z}. \tag{31}$$

Then (22) can be rewritten as

$$x^{(\gamma,p)} = e_M\left(\gamma\frac{k^2}{2} + \frac{pk}{L} + \frac{k}{2}\right) e_L(pr). \tag{32}$$

Expression (32) directly accommodates the condition (28), or, with the term $k/2$ removed, the condition (29). We will use (32), with the understanding that the results obtained are easily adaptable to sequences conforming to condition (29). The parity of γ, L, and M will subsequently be considered only when it is necessary.

Taking FZT of (32) yields the following result

Theorem 2.

$$X_L^{(\gamma,p)}(j,k) = \begin{cases} Lx_k^{(\gamma,p)},\ (j+p) \bmod L \equiv 0, \\ \\ 0, \qquad \text{else}, \end{cases} \tag{33}$$

where $x_k^{(\gamma,p)} = e_M(\gamma\frac{k^2}{2} + \frac{pk}{L} + \frac{k}{2})$.

Theorem 2 is the principal tool in the analysis of $\mathscr{S}_{\gamma,P}$. In the next subsection, we will use it to investigate correlation properties of sequences in $\mathscr{S}_{\gamma,P}$.

4.2 Correlation Properties

We investigate auto and crosscorrelation properties of sequences in $\mathscr{S}_{\gamma,P}$. Crosscorrelation of any sequence pair in $\mathscr{S}_{\gamma,P}$ is ideal. Individual autocorrelations of sequences in $\mathscr{S}_{\gamma,P}$ are not ideal, however sequences in $\mathscr{S}_{\gamma,1}$ have M-point $|ZACZ|$, and the sets $\mathscr{S}_{\gamma,1}$ and $\mathscr{S}'_{\gamma,P}$ (an enlargement of $\mathscr{S}_{\gamma,P}$ defined in (41)) are Golay.

Crosscorrelation

The next result follows directly from (33) and the ZS correlation formula (20).

Theorem 3. *Any two distinct sequences in $\mathscr{S}_{\gamma,P}$ have perfect all-zero crosscorrelation.*

Autocorrelation

Using (33) and (20) we can compute the ZS autocorrelation of $x^{(\gamma,p)}$ as follows

$$\begin{aligned} & Z_L^{(\gamma,p)}(j,k) \\ & = \frac{1}{N}\sum_{l=0}^{M-1} X_L^{(\gamma,p)}(j,l) X_L^{(\gamma,p)^*}(j,l-k) \\ & = \begin{cases} \frac{L^2}{N}\sum_{l=0}^{M-1} e_M\left(\gamma\frac{l^2}{2} + \frac{pl}{L} + \frac{l}{2}\right) e_M\left(\gamma\frac{-(l-k)^2}{2} - \frac{p(l-k)}{L} - \frac{l-k}{2}\right),\ j = L-p, \\ \\ 0, \qquad\qquad \text{else}, \end{cases} \\ & = \begin{cases} \frac{L}{M} e_M\left(\gamma\frac{-k^2}{2} + \frac{pk}{L} + \frac{k}{2}\right)\sum_{l=0}^{M-1} e_M(\gamma kl),\ j = L-p, \\ \\ 0, \qquad\qquad \text{else}, \end{cases} \\ & = \begin{cases} Le_M(\frac{-\gamma M^2}{2P^2}t^2 + \frac{pM}{PL}t + \frac{M}{2P}t),\ j = L-p,\ \ k = \frac{tM}{P},\ \ t \in \mathbb{Z}_P, \\ \\ 0, \qquad\qquad \text{else}. \end{cases} \end{aligned} \tag{34}$$

The inverse FZT of $Z_L^{(\gamma,p)}$ yields

$$z^{(\gamma,p)}(k+rM) = \begin{cases} e_M(\frac{-\gamma M^2}{2P^2}t^2 + \frac{pM}{PL}t + \frac{M}{2P}t)e_L(pr), & k = \frac{tM}{P},\ \ t \in \mathbb{Z}_P, \\ 0, & \text{else.} \end{cases} \tag{35}$$

Specializing the chirp rate factor γ in (35) endows construction B with the next property.

Theorem 4. *Take*

$$(\gamma, M) = 1. \tag{36}$$

Then

$$z^{(\gamma,p)}(k+rM) = \begin{cases} e_L(pr), & k = 0, \\ 0, & \text{else,} \end{cases} \tag{37}$$

and hence the set $\mathscr{S}_{\gamma,1}$ *is a ZCZ set with M-point* $|ZACZ|$.

Autocorrelation Sum

Consider again the general case, $(\gamma, M) = P$. From (35)

$$\begin{aligned} &\sum_{p=0}^{L-1} z^{(\gamma,p)}(k+rM) \\ &= \begin{cases} e_{2P}(\frac{-\gamma M}{P}t^2 + t)\sum_{p=0}^{L-1} e_{PL}(p(rP+t)), & k = \frac{tM}{P},\ \ t \in \mathbb{Z}_P, \\ 0, & \text{else,} \end{cases} \\ &= \begin{cases} L, & r = t = 0, \\ e_{2P}(\frac{-\gamma M}{P}t^2 + t)\sum_{p=0}^{L-1} e_{PL}(p(rP+t)), & k = \frac{tM}{P},\ \ t \in \mathbb{Z}_P/\{0\}, \\ 0, & \text{else.} \end{cases} \end{aligned} \tag{38}$$

It is apparent that the sum in (38)

$$\sum_{p=0}^{L-1} e_{PL}(p(rP+t)) = 0 \tag{39}$$

iff the condition (36) is satisfied. It follows then that the sum of cyclic autocorrelations of all sequences in $\mathscr{S}_{\gamma,1}$,

$$\sum_{p=0}^{L-1} z^{(\gamma,p)}(k+rM) = \begin{cases} L, \ r=k=0, \\ \\ 0, \ \text{else}, \end{cases} \tag{40}$$

is an ideal autocorrelation sequence, or, equivalently, by (5),

Theorem 5. *The set $\mathscr{S}_{\gamma,1}$ is cyclic Golay.*

This result can be extended to a larger set of sequences. Suppose $(\gamma,M) = P \neq 1$, and take

$$\mathscr{S}'_{\gamma,P} = \{x^{(\gamma,p)}, \ p \in \mathbb{Z}_{PL}\}. \tag{41}$$

It follows directly from (38) that the sum

$$\sum_{p=0}^{PL-1} z^{(\gamma,p)}(k+rM) = \begin{cases} PL, \ r=k=0, \\ \\ 0, \quad \text{else}, \end{cases} \tag{42}$$

yields an ideal autocorrelation sequence, and hence

Theorem 6. *The set $\mathscr{S}'_{\gamma,P}$ is cyclic Golay.*

The sequence sets $\mathscr{S}_{\gamma,1}$ and $\mathscr{S}'_{\gamma,P}$ are two distinct Golay set constructions; $\mathscr{S}_{\gamma,1}$ is Golay due to restriction on (γ,M), and $\mathscr{S}'_{\gamma,P} \supset \mathscr{S}_{\gamma,P}$ is Golay due to inclusion in the set of certain modulations of $\mathscr{S}_{\gamma,P}$. Moreover, $\mathscr{S}_{\gamma,1}$ is a ZCZ sequence set, while $\mathscr{S}'_{\gamma,P}$ is not.

4.3 DFT Support

Below we will show that $x^{(\gamma,p)}$ is semi-polyphase in the frequency domain. We focus on sequences satisfying the condition (28) and assume $(\gamma,M) = 1$. The general case was considered in [4].

We need first to define the quadratic Gauss sum [2], which will be subsequently used in the derivation. Set

$$G_n(m) = \sum_{k=0}^{n-1} e_n(mk^2) = \begin{cases} \left(\frac{m}{n}\right)\sqrt{n}, & n \equiv 1 \mod 4, \\ \\ 0, & n \equiv 2 \mod 4, \\ \\ \left(\frac{m}{n}\right) i\sqrt{n}, & n \equiv 3 \mod 4, \\ \\ \left(\frac{n}{m}\right)(1+i^m)\sqrt{n}, & n \equiv 0 \mod 4, \end{cases} \tag{43}$$

where $\left(\frac{m}{n}\right)$ is the Jacobi symbol given by the product

$$\left(\frac{m}{n}\right) = \left(\frac{m}{n_1}\right)\left(\frac{m}{n_2}\right)\cdots\left(\frac{m}{n_l}\right), \tag{44}$$

$n = n_1 n_2 \cdots n_l$, n_k are odd positive primes, not necessarily distinct, m is any integer,

$$\left(\frac{m}{n_k}\right) = \begin{cases} 1, & m \text{ is a quadratic residue modulo } n_k, \\ -1, & m \text{ is a quadratic nonresidue modulo } n_k, \\ 0, & n_k \mid m, \end{cases} \tag{45}$$

is the Legendre symbol, and $\left(\frac{m}{1}\right) = 1$.

Now we can proceed with computing the DFT support. Taking the DFT of $x^{(\gamma,p)}$ as an inverse FZT of $X_L^{(\gamma,p)}$ in (33) in k, yields

$$\begin{aligned} &\mathbf{x}^{(\gamma,p)}(j+sL) \\ &\quad= L\sum_{k=0}^{M-1} X_L^{(\gamma,p)}(j,k)e_N(jk)e_M(sk) \\ &\quad= \begin{cases} L\sum_{k=0}^{M-1} e_M\left(\gamma\frac{k^2}{2} + \frac{pk}{L} + \frac{k}{2}\right)e_N(jk)e_M(sk), & j+p \mod L \equiv 0, \\ \\ 0, & \text{else,} \end{cases} \\ &\quad= \begin{cases} L\sum_{k=0}^{M-1} e_M\left(\frac{\gamma k^2+(2s+3)k}{2}\right), & j+p \mod L \equiv 0, \\ \\ 0, & \text{else,} \end{cases} \end{aligned} \tag{46}$$

Since the expression in the sum is M-periodic, we have that for $j+p \mod L \equiv 0$,

$$\begin{aligned} S_1 &:= L\sum_{k=0}^{M-1} e_M\left(\frac{\gamma k^2 + (2s+3)k}{2}\right) \\ &= \frac{L}{2}\sum_{k=0}^{2M-1} e_{2M}\left(\gamma k^2 + (2s+3)k\right). \end{aligned} \tag{47}$$

Since $(\gamma, M) = 1$, there is $\gamma' \in \mathbb{Z}$ such that $\gamma\gamma' \equiv 1 \mod 2M$. Then the Galois conjugate [15] of S_1

$$\begin{aligned}
S_1' &= \frac{L}{2}\sum_{k=0}^{2M-1} e_{2M}\left(\gamma'(\gamma k^2+(2s+3)k)\right) \\
&= \frac{L}{2}\sum_{k=0}^{2M-1} e_{2M}\left((k^2+\gamma'(2s+3)k\right) \\
&= \frac{L}{2}\sum_{k=0}^{2M-1} e_{2M}\left(\left(k+\gamma'\frac{2s+3}{2}\right)^2-\gamma'^2\frac{(2s+3)^2}{4}\right) \\
&= \frac{L}{2}e_{2M}\left(-\gamma'^2\frac{(2s+3)^2}{4}\right)\sum_{k=0}^{2M-1} e_{2M}\left(\left(k+\gamma'\frac{2s+3}{2}\right)^2\right) \\
&= \frac{L}{2}e_{2M}\left(-\gamma'^2\frac{(2s+3)^2}{4}\right)\sum_{k=0}^{2M-1} e_{8M}\left((2k+\gamma'(2s+3))^2\right). \qquad (48)
\end{aligned}$$

Since γ' is odd, $\gamma'(2s+3)$ is odd and hence

$$\begin{aligned}
S_1' &= \frac{L}{2}e_{2M}\left(-\gamma'^2\frac{(2s+3)^2}{4}\right)\sum_{k=0}^{2M-1} e_{8M}\left((2k+1)^2\right) \\
&= \frac{L}{4}e_{2M}\left(-\gamma'^2\frac{(2s+3)^2}{4}\right)\sum_{k=0,\ k\ odd}^{8M-1} e_{8M}\left(k^2\right) \\
&= \frac{L}{4}e_{2M}\left(-\gamma'^2\frac{(2s+3)^2}{4}\right)\left\{\sum_{k=0}^{8M-1} e_{8M}\left(k^2\right)-\sum_{k=0,\ k\ even}^{8M-1} e_{8M}\left(k^2\right)\right\} \\
&= \frac{L}{4}e_{2M}\left(-\gamma'^2\frac{(2s+3)^2}{4}\right)\left\{\sum_{k=0}^{8M-1} e_{8M}\left(k^2\right)-\sum_{k=0}^{4M-1} e_{2M}\left(k^2\right)\right\} \qquad (49)
\end{aligned}$$

which by (43) is equal to

$$S_1' = \frac{L}{4}e_{2M}\left(-\gamma'^2\frac{(2s+3)^2}{4}\right)\sum_{k=0}^{8M-1} e_{8M}\left(k^2\right) \qquad (50)$$

and, by the inverse mapping,

$$\begin{aligned}
S_1 &= \frac{L}{4}e_{2M}\left(-\gamma'\frac{(2s+3)^2}{4}\right)\sum_{k=0}^{8M-1} e_{8M}\left(\gamma k^2\right) \\
&= \frac{L}{4}e_{2M}\left(-\gamma'\frac{(2s+3)^2}{4}\right)\left(\frac{8M}{\gamma}\right)(1+i^{\gamma})\sqrt{8M}, \qquad (51)
\end{aligned}$$

the last equality resulting from evaluation of the quadratic Gauss sum (43).

Inserting (51) into (46) and taking the magnitude of the lhs yields the following result.

Theorem 7. [4]

$$|\mathbf{x}^{(\gamma,p)}(j+sL)| = \begin{cases} L\sqrt{M},\ j=L-p,\ \ s\in\mathbb{Z}_M,\ \ p\in\mathbb{Z}_L, \\ \\ 0, \qquad \text{else}, \end{cases} \tag{52}$$

and hence $\mathbf{x}^{(\gamma,p)}$ is semi-polyphase.

Moreover, since

$$\sum_{p\in\mathbb{Z}_L} |\mathbf{x}^{(\gamma,p)}(j+sL)| = L\sqrt{M},\ \text{ for all } j\in\mathbb{Z}_L \ \text{ and } \ s\in\mathbb{Z}_M, \tag{53}$$

it follows from (7) that the set $\mathscr{S}_{\gamma,1}$ is cyclic Golay, which reproves (40).

Using arguments similar to the ones given in [4], it can be furthermore shown that the semi-polyphase property is shared by all sequences in $\mathscr{S}'_{\gamma,P}$, provided either condition (28) or (29) holds, but for sequences in $\mathscr{S}_{\gamma,P}$ and $\mathscr{S}'_{\gamma,P}$ the DFT support decreases by P wrt sequences in $\mathscr{S}_{\gamma,1}$. It follows that

$$\sum_{p\in\mathbb{Z}_{PL},\ x\in\mathscr{S}'_{\gamma,P}} |\mathbf{x}^{(\gamma,p)}| \tag{54}$$

is polyphase and

$$\sum_{p\in\mathbb{Z}_{L},\ x\in\mathscr{S}_{\gamma,P}} |\mathbf{x}^{(\gamma,p)}| \tag{55}$$

is semi-polyphase, and hence, as shown in previous subsection, the cyclic Golay property holds for $\mathscr{S}_{\gamma,1}$ and $\mathscr{S}'_{\gamma,P}$, but not for $\mathscr{S}_{\gamma,P}$.

Example 1 Take $L=2$, $K=M/L\in\mathbb{Z}$, $\gamma\in\mathbb{Z}_{odd}$, $b=p/2K$, $a=2\gamma/K$, and $(\gamma,K)=1$. The sequences

$$x^{(\gamma,p)} = e_{4K}\left(\gamma k^2+pk\right) e_2(pr),\ \ p=0,1,$$

form a cyclic Golay pair.
Suppose $\gamma=1$. Then the DFT of $x^{(\gamma,p)}$,

$$\mathbf{x}^{(\gamma,p)}(j+sL) = \begin{cases} 2(1+i)\sqrt{K}e_{4K}(-s^2),\ j=L-p, \\ \\ 0, \qquad\qquad\qquad \text{else}. \end{cases}$$

Some Other Examples of Construction B

Example 2 $L\times M=3\times 3,\, a=3\gamma,\ \gamma\in\{1\},\ b=p/3+1/2,\ p\in\mathbb{Z}_3.$

Example 3 $L\times M=7\times 5,\, a=49\gamma/5,\ \gamma\in\{1,3\}),\ b=p/5+7/10,\ p\in\mathbb{Z}_7.$

Example 4 $L\times M=5\times 7,\, a=25\gamma/7,\ \gamma\in\{1,3,5\},\ b=p/7+10/7,\ p\in\mathbb{Z}_5.$

Example 5 $L\times M=9\times 3,\, a=27\gamma,\ \gamma\in\{1\},\ b=p/9+3/2,\ p\in\mathbb{Z}_9.$

Example 6 $L\times M=3\times 9,\, a=\gamma,\ \gamma\in\{1,5,7\},\ b=p/3+1/6,\ p\in\mathbb{Z}_3.$

For b chosen so that $bM\in\mathbb{Z}$, the respective sets of allowed values of γ are $\{2\}$, $\{2,4\}$, $\{2,4,6\}$, $\{2\}$, and $\{2,4,8\}$.

5 Comparison with Some Prior Constructions

Construction B is related to construction A [7] and to the Mow construction [17].

Comparison with Construction A

In comparison with construction A, construction B has three key advantages and two disadvantages. The advantages are:

1. **FZT lattice tesselation**. While construction A is essentially limited to $L\times M$ lattices, where $L,\ M=\sqrt{L}\in\mathbb{Z}$, construction B permits all $L,\ M\in\mathbb{Z}$,
2. **Upper bound on** $|ZACZ|\times|\mathscr{S}_{\gamma,1}|$. Construction B achieves the upper bound, N (compared to N/M for construction A), on the product of $|ZACZ|$ and the number of sequences in the set (M and L, respectively), and
3. **Computational complexity of crosscorrelation**. As described in [7], computation of the crosscorrelation of x and its delayed version, y, includes three stages: FZT of y, ZS crosscorrelation of x and y, and inverse FZT of the result. For construction B complexity of the first and the third stage is reduced due to the special support structure of X_L and Y_L (Theorem 2). The next paragraph shows these computational stages in detail, both in ZS and in direct FS implementation.

Algorithm

- Off-line
  - Compute $L\times M$ FZT / N-point DFT of x, yielding X_L and $\mathbf{x}$, respectively.
- On-line

- Compute $L \times M$ FZT / N-point DFT of y, yielding

$$Y_L(j,k) = \begin{cases} \sum_{r=0}^{L-1} y(k+rM)e_L(rp), & j = p, \\ 0, & \text{else.} \end{cases}$$

where $0 \leq k < M$, and

$$\mathbf{y}(m) = \sum_{n=0}^{N-1} y(n)e_N(nm),$$

where $0 \leq m < N$.

- Perform ZS / FS crosscorrelation z of x and y, yielding

$$Z_L(j,k) = \begin{cases} \frac{1}{N}\sum_{l=0}^{M-1} Y_L(p,l)X_L^*(p,l-k), & j = p, \\ 0, & \text{else.} \end{cases}$$

where $0 \leq k < M$, and

$$\mathbf{z}(m) = \mathbf{y}(m)\mathbf{x}^*(m),$$

where $0 \leq m < N$, respectively.

- Compute the inverse FZT / DFT of the result of previous stage,

$$z(k+rM) = \frac{1}{L} Z_L(p,k)e_L(-rp),$$

where $0 \leq k < M$, $0 \leq r < L$, and

$$z(n) = \frac{1}{N}\sum_{m=0}^{N-1} \mathbf{z}(n)e_N(-nm),$$

where $0 \leq m < N$.

Since FZT of a sequence in $\mathscr{S}_{\gamma,P}$ is supported on a single row of X_L, its computation is reduced to taking the product in (33), $x(k+rM)e_L(-rp)$, which requires N multiplications. The remaining two stages require M^2 and N multiplication, respectively, with the inverse FZT taken as the outer product of $Z_L(p,k)$ and $e_L(-rp)$. Overall, the computational complexity of ZS realization of construction B is

$$\mathscr{O}_B^{ZS} = M^2 + 2N \tag{56}$$

and of direct FS realization of construction B is

$$\mathscr{O}_B^{FS} = N + 2N\log_2 N, \tag{57}$$

which is greater than computational complexity of ZS realization by nearly a factor of

$$\frac{\mathscr{O}_B^{FS}}{\mathscr{O}_B^{ZS}} = \frac{N + 2N\log_2 N}{M^2 + 2N} = \frac{2\log_2 N + 1}{2 + M/L} \approx \log_2 N. \tag{58}$$

To compare ZS realizations of constructions A and B we need to select an appropriate FZT lattice tesselation. For a standard realization of construction A, $L = N^{2/3}$, it's computational complexity is

$$\mathscr{O}_A^{ZS} = N^{2/3} + \frac{2}{3}N\log_2 N, \tag{59}$$

which is greater than computational complexity of construction B,

$$\mathscr{O}_B^{ZS} = N^{2/3} + 2N, \tag{60}$$

by nearly a factor of

$$\frac{\mathscr{O}_A^{ZS}}{\mathscr{O}_B^{ZS}} = \frac{N^{2/3} + \frac{2}{3}N\log_2 N}{N^{2/3} + 2N} \approx \frac{\log_2 N}{3}. \tag{61}$$

The Number of PCSSs and the Alphabet Size

Take $\Phi(n)$ to be the Euler totient function that counts the number of positive integers, less or equal n, that are co-prime with n. One disadvantage of construction B (considering the smaller of the two sets of PCSSs, $\mathscr{S}_{\gamma,1}$) is in the number of available PCSSs: instead of $M!L^M$ distinct sequence sets for each distinct choice of FZT lattice tesselation parameters, only $\Phi(M)L$ sets are available, if trivial operations, such as constant phase multiplications, are ignored. The factor $\Phi(M)$ denotes the available choices of γ and the factor L, the available choices of p. In the latter case, observe that insertion of the factor $e_M(k)$ into (46), that corresponds to taking $p' = p + L$ in (32), permutes the nonzero values of $\mathbf{x}$, but does not change $\mathbf{x}$ support size or structure. Another potential disadvantage of construction B is the number of phases, N, which is M times greater than the number of phases required for construction A.

Construction A+B

The constructions A and B can be mixed to produce a combined set while retaining all PSZAG properties. The first four properties follow from the construction. The cyclic Golay property can be derived either from the DFT support results or directly, from slightly extended autocorrelation computations.

Example 7 Take $L=5, M=4, a=L^2/M, b=1/M$ for a sequence from construction B, so that

$$X_L^{(0)}(j,k)=\begin{cases} Le_M(\frac{25k^2}{4}+\frac{k}{4}),\ j=4, \\ 0, \qquad\qquad\qquad \text{else,}\end{cases}$$

and $X_L^{(1)}$ is an FZT of a sequence from construction A,

$$X_L^{(1)}(j,k)=\begin{cases} L,\ j=k, \\ 0,\ \text{else.}\end{cases}$$

It follows from that

$$\operatorname{supp}(\mathbf{x}^{(0)})=\{4,9,14,19\},$$

$$\operatorname{supp}(\mathbf{x}^{(1)})=\{0,1,2,3,5,6,7,8,10,11,12,13,15,16,17,18\},$$

and

$$z^{(0)}(n)+4\,z^{(1)}(n)=0 \ \text{ for } \ n\neq 0.$$

The factor 4 in the above results from unequal supports and hence unequal magnitudes of $\mathbf{x}^{(0)}$ and $\mathbf{x}^{(1)}$.

Multicolumn Construction

Construction B can be generalized to a multicolumn construction [9], however, in this case the DFT of x, in general, does not preserve the semi-polyphase property.

Relation to the Mow Result

In [17] Mow proposed a time domain approach for the design of polyphase cyclic Golay sequence sets with all-zero crosscorrelation. He conjectured that his approach includes all possible sets of this kind. Both, the construction B given here, and the construction A described in [7], disprove this conjecture, as they permit the ratio of sequence length and set size, M, to contain a square, a case explicitly rejected in [17]. For example, the parameter choices $(\gamma,L,M)=(1,7,8)$, $(\gamma,L,M)=(1,8,24)$, $(\gamma,L,M)=(1,8,4)$, and $(\gamma,L,M)=(1,8,8)$, with a selection of an appropriate b, all yield PSZAG sequence sets. Constructions B has also the additional advantage in that it is performed in ZS, which provides more insight into the structure of sequences and more economy of crosscorrelation computation than the Mow approach.

References

1. An M, Brodzik AK, Tolimieri R. Ideal sequence design in time-frequency space, with applications to radar, sonar and communication systems.Boston: Birkhäuser; 2008.
2. Berndt BC, Evans RJ, Williams KS. Gauss and Jacobi sums. New York: Wiley-Interscience; 1998.
3. Bömer L, Antweiler M. Periodic complementary binary sequences. IEEE Trans Info Theory. 1990;36(6):1487–94 .
4. Brodzik AK. On the Fourier transform of finite chirps. IEEE Signal Process Lett. 2006; 13(9):541–4.
5. Brodzik AK. Characterization of a Zak space support of a discrete chirp. IEEE Trans Info Theory. 2007;53(6):2190–203.
6. Brodzik AK. New polyphase sequence sets with all-zero crosscorrelation. Proc 2012 IEEE ISIT, Cambridge, MA; 2012.
7. Brodzik AK. On certain sets of polyphase sequences with semi-polyphase Zak and Fourier transforms. IEEE Trans Info Theory. 2013; 59(10):6907–16, .
8. Brodzik AK. Construction of sparse representations of perfect polyphase sequences in Zak space with applications to radar and communications. EURASIP Journal on Advances in Signal Processing, Special Issue on Recent Advances in Non-stationary Signal Processing; 2011.
9. Brodzik AK. Taxonomy of Zak space sequence designs and some further properties, in preparation.
10. Brodzik AK, Tolimieri R. Bat chirps with good properties: Zak space construction of perfect polyphase sequences. IEEE Transactions on Info. Theory. 2009;55(4):1804–14,.
11. Chen H-H, Chiu H-W, Guizani M. Orthogonal complementary codes for interference-free CDMA technologies. IEEE Wireless Comm. 2006;68–79.
12. Frost SW, Rigling B. Sidelobe predictions for spectrally-disjoint radar waveforms. Proc. 2012 IEEE National Radar Conference, Atlanta, GA; 2012.
13. Golay M. Complementary series. IEEE Trans Info Theory (2012);7(2):82–7.
14. Helleseth T, Kumar PV. Sequences with low correlatio. In: Pless VS, Huffman WC, Editors. Handbook of coding theory. Elsevier; 1998. p. 1765–854.
15. Hungerford T. Abstract algebra. New York: Springer; 1998.
16. Janssen AJEM. The Zak transform: a signal transform for sampled time-continuous signals. Philips J Res. 1988;43:23–69
17. Mow WH. A new unified construction of perfect root-of-unity sequences. ISSSTA'1996, 955–956; 1996.
18. Parker MG, Paterson KG, Tellambura C. Golay complementary sequences. Wiley Encyclopedia of Telecommunications; 2004.
19. Gilbert J, Rzeszotnik Z. The norm of the Fourier transform on finite abelian groups. Ann Inst Fourier. 2010;60(4):1317–46.
20. Tao T. An uncertainty principle for cyclic groups of prime order. Math Res Lett. (2005);12:121–127 .
21. Welch LR. Lower bounds on the maximum crosscorrelation of signals. IEEE Trans Info Theory. (1974);20(3):397–9.
22. Zak, J. Finite translations in solid state physics. Phys Rev Lett. (1967);19:1385–97.

Reversible Jump Particle Filter (RJPF) for Wideband DOA Tracking

Thomas Wiese, Justinian Rosca and Heiko Claussen

Abstract We extend the maximum likelihood method for wideband direction of arrival (DOA) estimation to the case of an unknown number of moving sources. The extension is nontrivial because closed-form expressions for the conditional signal covariance matrices are no longer available. We propose a reversible jump particle filter (RJPF) based estimation of the source angles, which has been successfully used in narrowband DOA estimation of moving sources. We discuss added difficulties in DOA estimation compared to frequency retrieval problems. These difficulties are addressed by appropriate modifications of the underlying stochastic model. Finally, we show how an estimator of the number of sources and their positions can be constructed from a discrete representation of their posterior probabilities as provided by the particle filter.

Keywords DOA estimation · Wideband · Particle filter · Reversible jump

1 Introduction

In fields like radar, wireless communications, or speech recognition, one targets the position or direction of certain objects, e.g., airplanes, cell phones, or speakers, relative to a reference. This problem is approached by measurements with a sensor

T. Wiese
Fachgebiet Methoden der Signalverarbeitung, Technische Universität München,
80290 Munich, Germany
e-mail: thomas.wiese@tum.de

J. Rosca · H. Claussen
Siemens Corporation, Corporate Technology, Princeton, NJ 08540, USA
e-mail: justinian.rosca@siemens.com

H. Claussen
e-mail: heiko.claussen@siemens.com

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_11

array of the source signals generated by the objects either actively or passively. To estimate the directions of arrival (DOA) of the sources, it is sufficient to record the time differences of arrival of the signals at the sensors of the array. When *narrowband* signals are emitted, these time differences can be approximated by phase shifts, which makes the problem computationally tractable. Most successful algorithms for the DOA estimation problem are variants of ESPRIT [23] or MUSIC [24] that use subspace fitting techniques [28] to compute solutions fast.

Subspace methods for *wideband* signals infer DOA estimates that fulfill the signal and noise subspace orthogonality over possibly multiple frequencies, for each of which the narrowband signal assumption is satisfied. Signal and noise subspaces are computed from the sample covariance matrix of the measured data. In particular, coherent subspace methods (CSSM) [29] compute reference signal and noise subspaces by transforming all data to a reference frequency and deriving optimal DOA estimates with corresponding steering vectors orthogonal to the noise subspace at the reference frequency. Good initial estimates are necessary for the transformation to a reference frequency to work [27]. This issue is alleviated by methods like BI-CSSM or TOPS [13, 32]. A more important drawback of subspace methods, however, is that their performance degrades with increasing signal correlation amongst the sources.

Bayesian DOA methods [14] stand in contrast to the above by grounding estimation in a decision theoretic framework, which is mathematically elegant, more general, and powerful to express the update of beliefs under evidence of new measurements, that is, the posterior probability density function of the DOAs. However, the increased expressibility comes at the expense of increased computational effort to calculate a solution.

In our recent work, we presented a particle filter based algorithm, which we called multiple source tracking (MUST) [31], that can obtain DOA estimates of possibly correlated sources at lower signal to noise ratios (SNR) compared to subspace techniques. Moreover, it can track sources in moving source scenarios that pose difficulty to outstanding subspace methods.

In this chapter, we extend wideband DOA estimation introduced in [31] to the case of an unknown number of moving sources by means of a reversible jump particle filter (RJPF) based estimation of the source angles. Reversible jump Markov chain Monte Carlo (MCMC) methods were introduced in [8] and have since been used to generate efficient approximations in Bayesian inference under model uncertainty for the frequency retrieval problem [3, 19], which is related to the DOA problem, and for narrowband DOA estimation [12].

The estimator of the number of sources and their positions can be constructed from a discrete representation of their posterior probabilities provided by the particle filter. In [19] and [31], the estimator of the positions is preceded by a clustering, or label switching, step. This step increases the efficiency of the estimator as no, or few, particles need to be discarded. However, in this chapter, we argue that this technique can lead to biased estimates if either the signals are too correlated or if the angles between steering vectors for different angles are too small. We present an iterative algorithm that approximates a maximum a posteriori (MAP) estimate of all source angles, simultaneously.

This chapter is arranged in the following order: In the Sect. 2 we define the DOA problem for a known number of sources and present two possible formulations for the likelihood function. In Sect. 3, we explore the likelihood function that is obtained if the nuisance parameters are modeled as random variables. We refer to some fundamental detection limits that motivate the use of nonlinear arrays and wideband methods. Furthermore, our discussion of subspace methods and its limitations provides some intuition regarding the shape of the likelihood function.

In Sect. 4 we introduce the DOA problem for an unknown number of sources as a Bayesian model selection problem. We extend the likelihood function to comprise the unknown number of sources and discuss the influence of the nuisance parameters on the estimator for the number of sources. Furthermore, we show that the modified likelihood function provides a better estimator for the number of sources.

In Sect. 5 we provide a particle filter based algorithm for online calculation of the posterior distributions of the unknown number of sources and their angles of arrival. Subsequently, in Sect. 6, we present an algorithm that calculates point estimates of the number of sources and their angles from the discrete representation of the posterior density function provided by the particle filter. Finally, in Sect. 7, we show the simulation results that highlight the possible performance gains of a particle filter over subspace methods for a known and unknown number of sources. We summarize our results and indicate possible directions of future research in Sect. 8.

2 DOA Estimation for a Known Number of Sources

We consider a linear array of M sensors with distances d_m between the mth and the first sensor, which we take as the reference sensor. These sensors record the superpositions of unknown wavefronts, for example, acoustic signals, from different sources located at different directions. We assume that all sources and microphones are in a plane, hence the problem is two-dimensional. The k sources are located at angles $\theta = (\theta_1, \dots, \theta_k)$ with respect to the sensor array axis and transmit narrowband signals with wavelengths λ_n on $n = 1, \dots, N$ independent frequencies. If during one observation period, the complex amplitudes of the signals transmitted in the nth frequency are given by $s_n = (s_{1,n}, \dots, s_{k,n})^T$, the vector $x_n = (x_{1,n}, \dots, x_{M,n})^T$ of received signals can be modeled as a circularly symmetric complex Gaussian random variable

$$x_n \sim \mathcal{N}_{\mathbb{C}}\left(A_n(\theta)s_n,\, \sigma_n^2 I\right) \tag{1}$$

with receiver noise variance σ_n^2 and the design matrix $A_n(\theta)$ consisting of *steering vectors* $a_n(\theta_\kappa), \kappa = 1, \dots, k$ with

$$a_n(\theta_\kappa) = \left(1 \ \mathrm{e}^{\mathrm{i}\pi \sin(\theta_\kappa) d_2/(\lambda_n/2)} \ \cdots \ \mathrm{e}^{\mathrm{i}\pi \sin(\theta_\kappa) d_M/(\lambda_n/2)}\right)^T . \tag{2}$$

All results and problem formulations in this chapter are in terms of the sines of the angles as the source angles θ_κ appear in the likelihood function only through their sines.

In practice, the N narrowband signals are obtained by transforming real valued time-domain data into the frequency domain, e.g., by using the windowed discrete Fourier transform (DFT), and selecting the N positive frequency bins of interest. Multiple observation periods refer to different positions of the windows. It has been noted that finite window-length effects give rise to model-mismatch and additional statistical dependencies among the channels [14]. These effects are neglected here. Furthermore, when referring to signal correlations, we speak of the statistical dependencies between the coefficients $s_{\kappa,n}$ and $s_{\kappa',n}$ for different sources $\kappa, \kappa' \leq k$ at the *same frequency* over different observation periods. That is, by use of finite windows, we may obtain (approximations of) uncorrelated coherent signals.

The full likelihood function for the observed data $x = (x_1, \ldots, x_N)$ and a known number of sources can be written as

$$p(x|k,\theta,s,\sigma_n^2) = \prod_{n=1}^{N} \frac{1}{\pi^M \sigma_n^{2M}} \exp\left(-\sigma_n^{-2} \|x_n - A_n(\theta)s_n\|^2\right) \tag{3}$$

where the dependence on k is reflected by the dimension of θ and, thus, the number of columns of A_n.

2.1 A Bayesian Approach for the Nuisance Parameters

We briefly present a Bayesian approach for handling the unknown parameters that appear in the likelihood function (3), i.e., the noise variances σ_n^2 and the signals s_n. In Bayesian statistics, priors are introduced for all unknown parameters. These are chosen as a compromise between a minimum of information content and analytic tractability. In the DOA and frequency retrieval problems, it is common to use Jeffrey's uninformative prior

$$p(\sigma_n^2) \propto \frac{1}{\sigma_n^2} \tag{4}$$

for the noise variances and the so-called *g-prior*

$$s_n|k,\theta,\sigma_n^2 \sim \mathcal{N}_{\mathbb{C}}\left(0, \sigma_n^2 \delta_n^2 (A_n(\theta)^H A_n(\theta))^{-1}\right) \tag{5}$$

for the signal amplitudes [3, 12, 14]. The parameter δ_n^2 is interpreted as an expected signal to noise ratio at the nth frequency. In a fully Bayesian approach, δ_n^2 is also modeled as a random variable with its own (hyper-) prior distribution. However, this parameter is of little importance if the number of sources is known. Therefore, we postpone a discussion of δ_n^2 until Sect. 4.

In the following, we calculate the marginal Bayesian likelihood function $p(x|k,\theta)$. Let s and σ^2 denote the collections of the s_n and σ_n^2 for all n. Then, the marginal

Bayesian likelihood function is given as

$$\begin{aligned} p(x|k,\theta) &= \int p(x|k,\theta,s,\sigma^2)p(s,\sigma^2|k,\theta)\,d(s,\sigma^2) \\ &= \int p(x|k,\theta,s,\sigma^2)p(s|k,\theta,\sigma^2)p(\sigma^2)\,d(s,\sigma^2)\,. \end{aligned} \tag{6}$$

We first integrate with respect to the signals s. This is done separately for each frequency n. The *g-prior* is chosen such that

$$\begin{aligned} &\int p(x_n|k,\theta,s_n,\sigma_n^2)p(s_n|k,\theta,\sigma_n^2)\,ds_n \\ &= \frac{\det(A_n^H A_n)}{(\pi\sigma_n^2)^{M+k}\delta_n^{2k}} \int \exp\left(-\frac{1}{\sigma_n^2}\|x_n - A_n s_n\|^2 - \frac{1}{(\sigma_n\delta_n)^2} s_n^H A_n^H A_n s_n\right) ds_n \\ &= \frac{1}{\pi^M \sigma_n^{2M}(1+\delta_n^2)^k} \exp\left(-\sigma_n^{-2} x_n^H \left(I - \frac{\delta_n^2}{1+\delta_n^2} P_n\right) x_n\right) \end{aligned} \tag{7}$$

remains a Gaussian distribution. Here, $P_n = A_n(A_n^H A_n)^{-1}A_n^H$ is the projection onto the range space of A_n. We dropped the argument θ for legibility. In a second step, we integrate over the noise variances σ_n^2.

We distinguish two cases: First, we use independent noise variances σ_n^2 and integrate for each frequency separately. Writing $c_n = x_n^H(I - \delta_n^2/(1+\delta_n^2)P_n)x_n > 0$, we calculate the integral as

$$\begin{aligned} \int p(x_n|k,\theta,\sigma_n^2)p(\sigma_n^2)\,d\sigma_n^2 &\propto \int_0^\infty \frac{1}{\sigma_n^2}\frac{1}{\sigma_n^{2M}} \exp(-\sigma_n^{-2}c_n)\,d\sigma_n^2 \\ &= c_n^{-M}(M-1)! \end{aligned} \tag{8}$$

and, combining over all frequencies, we obtain

$$-\log p(x|k,\theta) = NM \log\left(\sqrt[N]{\prod_{n=1}^N c_n}\right) + k\sum_{n=1}^N \log\left(1+\delta_n^2\right) + C \tag{9}$$

where C is independent of k and θ.

In the second case, we assume that all noise variances σ_n^2 and the parameters δ_n^2 are equal. We write, with abuse of notation, $\sigma_n^2 = \sigma^2$ and $\delta_n^2 = \delta^2$. In this case, the product is under the integral sign and we obtain

$$\begin{aligned} \int \prod_{n=1}^N p(x_n|k,\theta,\sigma^2)p(\sigma^2)\,d\sigma^2 &\propto \int_0^\infty \frac{1}{\sigma^2}\frac{1}{\sigma^{2MN}} \exp\left(-\sigma^{-2}\sum_{n=1}^N c_n\right) d\sigma^2 \\ &= \left(\sum_{n=1}^N c_n\right)^{-NM} (MN-1)! \end{aligned} \tag{10}$$

The negative log-likelihood function now reads

$$-\log p(x|k,\theta) = NM \log\left(\frac{1}{N}\sum_{n=1}^{N} c_n\right) + kN \log\left(1+\delta^2\right) + C \tag{11}$$

where we changed the scaling factor C to emphasize the similarities between both cases: If independent noise variances are used for all channels, then the first term in the likelihood is proportional to the geometric mean of the residual noise energies c_n. If a common noise variance is used, the geometric mean is replaced by the arithmetic mean. In both cases, the first term describes how well the model fits the data, while the second term can be interpreted as a penalty for the model complexity (see Sect. 4 for a discussion).

We now further simplify the likelihood function for a common noise variance and δ^2 value. The average over c_n can then be written as

$$\begin{aligned}\frac{1}{N}\sum_{n=1}^{N} c_n &= \frac{1}{N}\sum_{n=1}^{N} x_n^H \left(I - \frac{\delta^2}{1+\delta^2} P_n\right) x_n \\ &= \frac{1}{1+\delta^2}\left(\frac{1}{N}\sum_{n=1}^{N} x_n^H x_n + \delta^2 \frac{1}{N}\sum_{n=1}^{N} x_n^H \left(I - P_n\right) x_n\right) \\ &= \frac{\delta^2}{1+\delta^2}\left(\frac{E_x}{\delta^2} + E_{\text{noise}}(\theta,k)\right)\end{aligned} \tag{12}$$

where

$$E_x = \frac{1}{N}\sum_{n=1}^{N} x_n^H x_n \quad \text{and} \quad E_{\text{noise}}(\theta,k) = \frac{1}{N}\sum_{n=1}^{N} x_n^H (I - P_n) x_n \tag{13}$$

are the average energy of the received signal and the residual noise, respectively. Hence, the negative log-likelihood (11) can be written as

$$-\log p(x|k,\theta) = NM \log\left(\frac{E_x}{\delta^2} + E_{\text{noise}}(\theta,k)\right) + kN \log(1+\delta^2) + C'\,. \tag{14}$$

In the remainder of this chapter, we will refer to (14) as the *Bayesian likelihood* function.

2.2 Alternative Likelihood—Regularization

In this section, we motivate an alternative to the Bayesian likelihood function (14), where the g-prior for the source signals is replaced by a more natural, at least in our view, circularly symmetric Gaussian prior

$$s_n \sim \mathcal{N}_{\mathbb{C}}(0, \tau_n^{-1} I) \tag{15}$$

with known precisions $\tau_n > 0$. This alternative is also concerned with the model selection problem discussed in Sect 4, where we introduce a penalty parameter that is easier to interpret than the δ^2 parameter appearing in (14). Moreover, the possibility to use a weakly informative prior for the source signals ($\tau_n > 0$) aides in separating closely spaced sources (see Fig. 1).

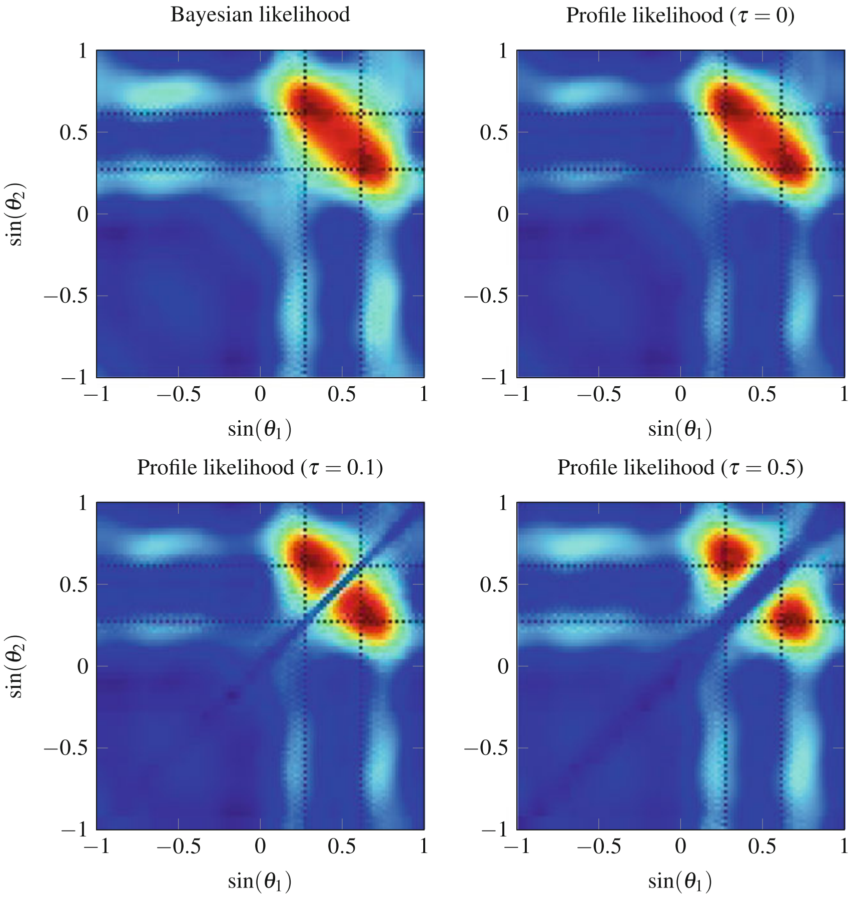


Fig. 1 Likelihoods with five microphones, five snapshots, and a single frequency. The *top left* image shows the Bayesian likelihood function for two sources located at $\theta_1 = 15$ and $\theta_2 = 37$ degrees indicated by the *dashed lines*, while the other images show the profile likelihoods for different values of the regularization parameter τ

A conditional maximum likelihood (ML) estimator [26] for the angles is obtained as

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\theta} p\left(x|k, \theta, \hat{s}_n(\theta)\right) \tag{16}$$

where

$$\hat{s}_n(\theta) = \left(A_n(\theta)^H A_n(\theta) + \tau I\right)^{-1} A_n^H(\theta) x_n \tag{17}$$

is the maximum a posteriori estimator of s_n given the measurements and the source locations. Hence, in ML estimation, only the profile likelihood

$$p(x|k,\theta) \propto \prod_{n=1}^{N} \exp\left(-\sigma^{-2} \|x_n - A_n(\theta)\hat{s}_n(\theta)\|^2\right) \tag{18}$$

is used. Let $\widetilde{P}_n = A_n \left(A_n^H A_n + \tau_n I\right) A_n^H$ denote the regularized projection operator and

$$\widetilde{E}_{\text{noise}} = \frac{1}{N} \sum_{n=1}^{N} x_n^H (I - \widetilde{P}_n) x_n \tag{19}$$

the corresponding average residual noise energy. The negative log-likelihood function reads

$$-\log p(x|k,\theta) = \sigma^{-2} N \widetilde{E}_{\text{noise}} \,. \tag{20}$$

Note that $x_n|k,\theta$ is still a circularly symmetric Gaussian random variable for each n under this likelihood function. We call (20) the (negative logarithm of the) *profile likelihood* function for x. Figure 1 shows that both likelihood functions are similar if a flat (improper) prior, $\tau = 0$, is used for the source signals. However, if nonzero values are used for τ, the two source locations are better separated with the profile likelihood (see Sect. 3.6 for a discussion).

3 Discussion—Known Number of Sources

The DOA problem has been studied extensively in the literature, often in the guise of the frequency retrieval problem. We briefly discuss the relationship between both problems in Sect. 3.1. After that, in Sect. 3.2, we refer to the fundamental limits regarding the maximal number of angles that can be identified with a given array of sensors. In Sect. 3.3, we briefly review the MUSIC algorithm, because a subspace perspective provides valuable insights into the geometry of the likelihood function of the DOA problem. Subsequently, in Sects. 3.4 and 3.5, we highlight the drawbacks of subspace methods in wideband problems or if the source signals are correlated. Finally, in Sect. 3.6, we draw on the subspace view to explain why the regularization occurring in the profile likelihood function helps separating closely spaced sources.

3.1 Relationship of DOA Estimation to Frequency Retrieval

The DOA problem is closely related to the frequency retrieval problem, where M samples of a superposition of pure sine waves of unknown frequencies are recorded at a single sensor [3, 22]. The goal is to estimate the frequencies and, perhaps, the number of sources. This problem can be transformed into a DOA problem by interpreting the time samples as distances between microphones and the unknown frequencies as products of a single frequency with the sine of the source angle. While in frequency retrieval, one typically records a sufficiently large number of samples, the equivalent DOA problem consists of a single snapshot and a number of microphones that equals the number of time samples in the frequency retrieval problem. In DOA estimation, the number of sensors is typically limited to a small number and data are recorded during several observation intervals. In contrast to the former problem, DOA estimation is subject to ambiguities caused by constructive and destructive interference. This is highlighted in Fig. 2, where likelihood functions and their marginals are compared for a scenario with two sources.

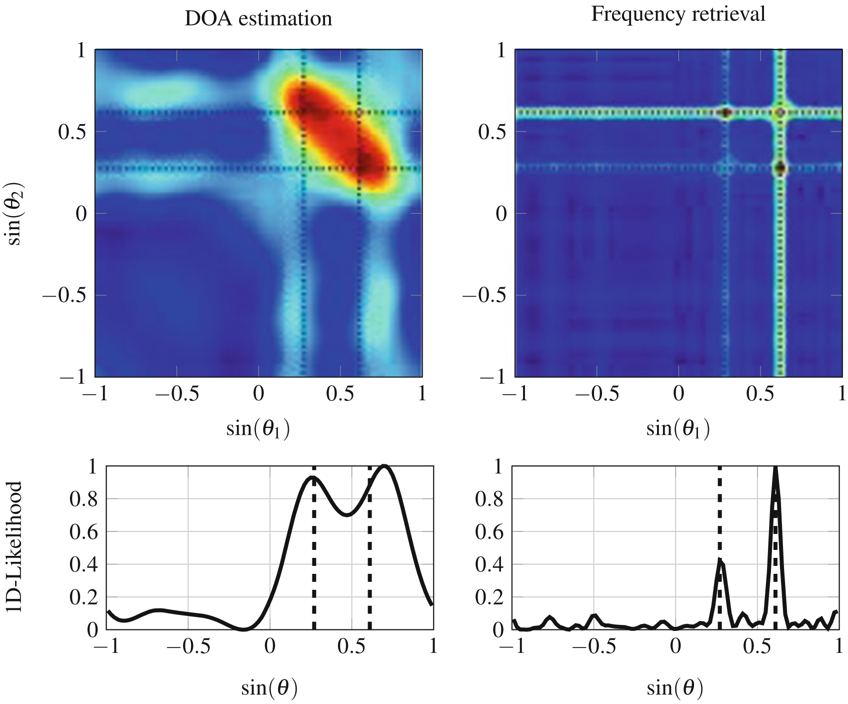


Fig. 2 *Left*: marginal likelihood in DOA estimation with five microphones and five snapshots. *Right*: marginal likelihood in DOA estimation with 25 microphones and a single snapshot—this configuration resembles the frequency retrieval problem. The *dashed lines* indicate the true source locations

3.2 Conditions for Identifiability

A preliminary question that needs to be solved before using an algorithm that estimates k angles of arrivals is whether it is feasible. If a uniform linear array (ULA) of M sensors is used and the sources emit narrowband signals, this question can be answered affirmatively only if k satisfies the *identifiability condition*

$$k \leq 2rM/(2r+1) \tag{21}$$

where r is the rank of the sample covariance matrix of the observed signals x [5]. If the source signal covariance matrix has full rank and if enough samples are collected, (21) simplifies to $k < M-1$.

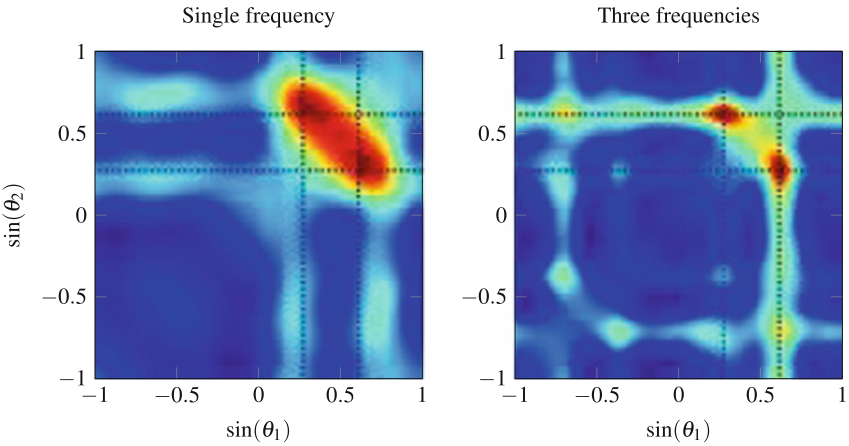
This condition is of little importance in the frequency retrieval problem, where the sensors take the role of time-domain samples and where enough samples are available (M is large). In DOA estimation, however, this constraint is severe as only few sensors are available.

In [30] it is shown that (21) still applies if the sensors are spaced nonuniformly and if the largest distance satisfies $d_M/(\lambda/2) < M-1$. On the contrary, if $d_M/(\lambda/2) > M-1$, there always exist sets of angles θ and $\tilde{\theta}$ and corresponding source signals s and $\tilde{s}$, for which [17]

$$A(\theta)s = A(\tilde{\theta})\tilde{s}. \tag{22}$$

Thus, one cannot distinguish whether a given set of observations is generated by sources located at θ or $\tilde{\theta}$. If, however, additional assumptions regarding the statistics of the source signals are made, it is possible to identify more than M sources [2]. Of particular interest are *fully augmentable* nonuniform arrays with as many integer numbers of pairs of intersensor spacings (measured in half wavelengths) as possible [16]. An example for $M = 4$ is given by the normalized positions $\{0,1,4,6\}$. If $M \leq 4$, it is thus possible to separate up to $M(M-1)/2-1$ source signals using data augmentation techniques [1]. For $M \geq 5$, the maximal number of separable sources is strictly smaller than $M(M-1)/2-1$, but still larger than the ULA limit of $M-1$ sources.

For two sets of wideband sources to be indistinguishable, the ambiguity condition (22) must be satisfied on all frequencies $n = 1,\dots,N$, simultaneously. Hence, ambiguities are less likely. Furthermore, the use of wideband signals improves the accuracy of the estimator (the array has a larger aperture at higher frequencies) while maintaining the identifiability condition $d_M/(\lambda/2) \leq M-1$ at the lowest frequency. This effect is shown in Fig. 3.


Single frequency

Three frequencies

Fig. 3 Likelihood in DOA estimation with five microphones and five snapshots. *Left*: single normalized frequencies $f_1 = 1$. *Right*: three normalized frequencies $f_1 = 1.0, f_2 = 1.5, f_3 = 2.0$

3.3 Subspace Based Methods

Let us assume that two sources at angles θ_1 and θ_2 are transmitting at a single frequency, that $s \sim \mathcal{N}(0, \Sigma_s)$, and that receiver noise is negligible. Then, x is a random vector, distributed according to

$$x \sim \mathcal{N}\left(0, A(\theta)\Sigma_s A(\theta)^H\right) . \tag{23}$$

Thus, the measurements x lie in the subspace spanned by the steering vectors $a(\theta_1)$ and $a(\theta_2)$. This subspace is two-dimensional if the identifiability condition (21) is fulfilled, for example if $M \geq 3$ and if Σ_s has full rank. The subspace can be recovered from the eigenvectors corresponding to nonzero eigenvalues of the sample covariance matrix of the measurements x collected through multiple observation periods. These eigenvectors span the *signal subspace* while the eigenvectors with zero eigenvalues span the *noise subspace*. The identifiability condition ensures that only steering vectors $a(\theta)$ that intersect this subspace correspond to the true angles θ_1 and θ_2. If there was another angle θ_3 for which $a(\theta_3)$ intersects the subspace, $A((\theta_1, \theta_2))$ would generate the same range space as $A((\theta_1, \theta_3))$ and source signals could be found such that (22) is true. Thus, the angles can be recovered by finding those steering vectors that lie in the signal subspace or, equivalently, that are orthogonal to the noise subspace.

Let us denote by W the matrix whose columns are the eigenvectors that are in the noise subspace. Then, the angles θ_1 and θ_2 fulfill

$$a(\theta_\kappa)^H W^H W a(\theta_\kappa) = 0, \quad \kappa = 1, 2 . \tag{24}$$

In practice, noise is present at the receiver and the subspace can only be approximately recovered from the eigenvectors of the sample covariance matrix corresponding to the largest eigenvalues. One then selects those angles θ that correspond to the maxima of

$$\frac{1}{\|a(\theta)^H W^H W a(\theta)\|^2}\,. \tag{25}$$

This last expression is known as the *MUSIC spectrum*. One obtains the incoherent MUSIC (IMUSIC) algorithm by calculating the MUSIC spectrum for each frequency, separately, and then finding the angles that correspond to the peaks of the average MUSIC spectrum.

3.4 Wideband Signals in Subspace Methods

When proceeding to wideband signals, it is no longer possible to apply subspace methods to the sample covariance matrix of the received signals as there is one for each frequency. Incoherent subspace methods such as incoherent MUSIC find source angles that minimize the average of the criterion (25) over all frequencies. As shown in Fig. 4, coherent subspace methods, such as CSSM and WAVES [6, 29] outperform their incoherent counterpart IMUSIC. These methods form a single covariance matrix from information over all channels. This is achieved through the use of focusing matrices [10].

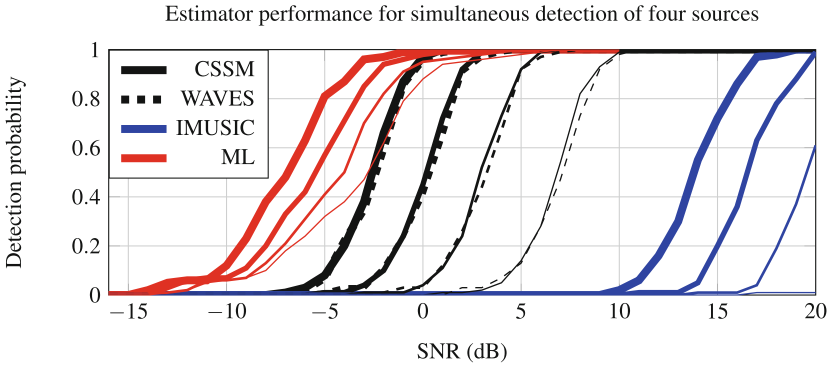


Fig. 4 Probability of simultaneous detection of four sources located at $\theta_1 = 8, \theta_2 = 13, \theta_3 = 33$, and $\theta_4 = 37$ (degrees). The correlation coefficient ρ is increased from the *bold lines* to the *thin lines* as $\rho = 0.0, 0.5, 0.75, 0.9$

The parameters used to calculate the detection probabilities of the estimators shown in Fig. 4 are similar to those used in the DOA literature [6, 29]. Four sources located at $\theta_1 = 8, \theta_2 = 13, \theta_3 = 33$, and $\theta_4 = 37$ (degrees) emit circularly symmetric signals with unit power at each normalized frequency $f_1 = 0.8, f_2 = 0.9, f_3 = 1.0, f_4 = 1.1, f_5 = 1.2$ and during each of 20 observation periods. The amplitudes of the signals are uncorrelated between different observation periods, but within each observation period, the off-diagonal elements of the signal correlation matrix are set to a factor $0 \leq \rho \leq 1$. The signals received at the uniform linear array of ten sensors with half-wavelength spacing with respect to the reference frequency f_3 are corrupted by circularly symmetric white Gaussian noise with power $\sigma^2 = 1/\mathrm{SNR}$ for each sensor and each frequency. For CSSM and WAVES, the focusing matrices are generated using the true source locations. This is an *unrealistic* assumption and significantly improves performance of these methods. It is shown in [27] that CSSM estimates are biased if the focusing angles are not the true directions of arrival. The ML estimate is found by evaluating the likelihood function for all possible combinations of the source locations on a regular grid with 100 nodes in each dimension.

As performance criterion we used the detection probability, which is in line with the cost function one would use to find the MAP estimate of the angles. The sources are detected correctly if $|\sin\theta_\kappa - \sin\hat{\theta}_\kappa| \leq 0.07$ for all $\kappa = 1,\ldots,4$. The curves in Fig. 4 show the fraction of correct detections from a total of 100 Monte Carlo simulations with parameters as described above. Note that the ML method suffers to a much lesser degree from signal correlation than subspace methods. Furthermore, at low SNR levels, the ML estimator provides a significant performance gain even for uncorrelated signals ($\rho = 0$). However, the ML estimation requires solving a nonlinear optimization problem, which is impractical. Therefore, we propose a particle filter method in Sect. 5 that can be seen as a heuristic to solve this problem.

3.5 Why Subspace Methods Suffer from Correlation

A drawback of subspace methods is their limited robustness for increasing signal correlation. This can be seen in the following example. Let us select θ_1 and θ_2 such that $a(\theta_1)$ and $a(\theta_2)$ are orthogonal. Note that for any distinct θ_1 and θ_2 , this is approximately true for large M. Furthermore, let the signal covariance matrix be given by

$$\Sigma_s = \sigma_s^2 \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \tag{26}$$

with $0 \ll \rho < 1$. Then, one can verify that the eigenvalue decomposition of the covariance matrix Σ_x of x is given by

$$U \begin{pmatrix} \sigma_s^2(1+\rho)+\sigma^2 & 0 \\ 0 & \sigma_s^2(1-\rho)+\sigma^2 \end{pmatrix} U^H + \sigma^2 V V^H$$

where U is given by

$$U = (2M)^{-1/2} A(\theta) \begin{pmatrix} -1 & 1 \\ 1 & 1 \end{pmatrix} \tag{27}$$

and where V is such that the matrix $(U\ V)$ is the eigenbasis of Σ_x. As ρ approaches one, the second largest eigenvalue of Σ_x approaches the noise variance σ^2, i.e., the effective SNR decreases with increasing correlation. This drawback of subspace methods is mentioned in [11].

3.6 Regularization in the Profile Likelihood

We give an explanation for the effect of regularization of the profile likelihood function at hand of an example with two closely spaced sources that transmit on a single frequency. If θ_1 and θ_2 are close enough that the parametric curve $a(\theta)$ can be approximated by a straight line between $a(\theta_1)$ and $a(\theta_2)$, then all combinations of $a(\tilde{\theta}_1)$ and $a(\tilde{\theta}_2)$ span approximately the same subspace for $\theta_1 \leq \tilde{\theta}_1, \tilde{\theta}_2 \leq \theta_2$. Let P denote the projector onto this subspace. From the eigenvalue decomposition $A^H(\tilde{\theta})A(\tilde{\theta}) = U \operatorname{diag}(\xi_1, \xi_2)U^H$, we can calculate P and $\tilde{P}$ as

$$P = UU^H, \quad \tilde{P} = U \operatorname{diag}\left(\frac{\xi_1}{\xi_1+\tau}, \frac{\xi_2}{\xi_2+\tau}\right) U^H. \tag{28}$$

Thus, the value of the negative log-likelihood (20),

$$\begin{aligned}\sigma^{-2}x^H(I-\tilde{P})x &= \sigma^{-2}x^H(I-P)x + \sigma^{-2}x^H(P-\tilde{P})x \\ &= \sigma^{-2}x^H(I-P)x + \sigma^{-2}x^H U \operatorname{diag}\left(\frac{\tau}{\xi_1+\tau}, \frac{\tau}{\xi_2+\tau}\right) U^H x\end{aligned} \tag{29}$$

depends on $\tilde{\theta}$ only through the eigenvalues ξ_1 and ξ_2 of $A^H(\tilde{\theta})A(\tilde{\theta})$. One can verify that these are given by

$$\xi_1(\tilde{\theta}) = M^2(1+|\cos\zeta|), \quad \xi_2(\tilde{\theta}) = M^2(1-|\cos\zeta|), \tag{30}$$

where

$$\zeta = \frac{a^H(\tilde{\theta}_1)a(\tilde{\theta}_2)}{\|a(\tilde{\theta}_1)\|\|a(\tilde{\theta}_2)\|} \tag{31}$$

is the angle between both steering vectors. As the SNR is decreased, $U^H x$ tends toward a two-dimensional circularly symmetric Gaussian distribution with variance σ^2 and the expected value of the second term in (29) is given by

$$E\left[\sigma^{-2}x^H U \operatorname{diag}\left(\frac{\tau}{\xi_1+\tau}, \frac{\tau}{\xi_2+\tau}\right) U^H x\right] \approx \frac{\tau}{\xi_1+\tau} + \frac{\tau}{\xi_2+\tau}. \tag{32}$$

One can verify that this term is large if the steering vectors are aligned and small if ζ is large. In conclusion, while the first term of the negative log-likelihood function (29) does not change as $\tilde{\theta}_1$ and $\tilde{\theta}_2$ are varied between θ_1 and θ_2, the second term is smaller, i.e., the likelihood is larger, if the distance between $\tilde{\theta}_1$ and $\tilde{\theta}_2$ is increased. Thus, separation of closely spaced sources is simplified by increased regularization with τ.

4 DOA Estimation for an Unknown Number of Sources

Throughout the previous sections, we described how estimates for the source positions can be obtained if the number of sources is known. However, this information is not always available. In a complete DOA estimation algorithm, this part is referred to as the *tracking* part of the algorithm and must be complemented by a *detection* or *initialization* part [15]. Also, this latter part is referred to as a model selection problem.

In Sect. 4.1, we discuss the Bayesian likelihood function under this new perspective. In Sect. 4.2, we modify the profile likelihood function in a way that permits comparison of models with different numbers of sources and with the Bayesian likelihood function.

4.1 Bayesian Model Selection

We start with the Bayesian likelihood function from (14). If the number of sources k is unknown, one commonly introduces a truncated Poisson prior distribution,

$$k|\Lambda \ \sim \ \mathrm{Poi}_{k_{\max}}(\Lambda) \tag{33}$$

where $\Pr(k > k_{\max}) = 0$. The parameter Λ represents the expected number of sources, which, in a fully Bayesian framework, is also modeled as a random variable with a hyperprior [3, 19].

It is not obvious how an ML estimator for both variables, k and θ, can be obtained from the likelihood function. A possible approach is to find the ML estimate $\hat{\theta}(k)$ for each $k \leq k_{\max}$ and to use k that minimizes the negative log-likelihood function:

$$\hat{k}_{\mathrm{ML}} = \arg\min_{k} NM \log\left(\frac{E_x}{\delta^2} + E_{\mathrm{noise}}(\hat{\theta}(k), k)\right) + kN\log(1+\delta^2)\,. \tag{34}$$

If the Poisson prior is used for the number of sources, one can similarly find a MAP estimate of k from

$$\hat{k}_{\mathrm{MAP}} = \arg\min_{k} NM \log\left(\frac{E_x}{\delta^2} + E_{\mathrm{noise}}(\hat{\theta}(k), k)\right) + kN\log(1+\delta^2) - k\log\Lambda + \log k! \tag{35}$$

That the *data-fit* term $NM\log\left(E_x/\delta^2 + E_{\mathrm{noise}}\right)$ is a decreasing function of k is compensated by the *penalty* term $kN\log(1+\delta^2)$ or $kN\log(1+\delta^2) - k\log\Lambda + \log k!$, respectively. It is known that both terms need to be carefully balanced for an effective model selection [18]. In a Bayesian framework, this is achieved by varying δ^2, which represents the expected SNR. This parameter determines by how much the residual noise energy E_{noise} needs to be reduced if a source is added such that a more complex model is preferred.

The parameter δ^2 has a strong influence on the estimation results [3, 12, 19] as shown in Fig. 5 (top row). The probability $\mathbb{P}(k_{\mathrm{ML}} = k)$ is plotted for different values of k and δ^2 for a scenario with two closely spaced sources. The parameters for this simulation are chosen in Sect. 3.4 except that only two sources at $\theta_1 = 33$ and $\theta_2 = 37$ (degrees) and a single observation period are used. Figure 5 shows how many of 1000 Monte Carlo simulations result in the ML estimator (34) $\hat{k}_{\mathrm{ML}} = k$ for $0 \leq k \leq 4$. The curves also show that the weight between the penalty term and the data-fit term is shifted in a nonmonotonic fashion as δ^2 is increased.

Note that for 0 dB SNR, there exists no δ^2 to balance the data-fit and the penalty terms such that $k = 2$ becomes the most likely number of sources (the curve for $k = 2$ is always below the curve for $k = 1$).

It is known that no suitable noninformative prior exists in the context of model selection problems [19]. Instead a weakly informative conjugate prior (inverse gamma), $\delta_n^2 \sim \mathscr{I}\mathscr{G}(\alpha,\beta)$, has been proposed, where $\alpha = 2$ and β is either chosen ad-hoc or estimated by the data, yielding an empirical Bayes method [3, 19]. Still, the hyper-parameter β has a strong influence on the estimation results [20, 21]. Therefore, we introduce a more intuitive penalty parameter for the profile likelihood function in Sect. 4.2.

4.2 Alternative Profile Likelihood and Model Selection

Recall the profile likelihood function from Sect. 2.2

$$-\log p(x|k,\theta) = \sigma^{-2} N \widetilde{E}_{\mathrm{noise}} . \tag{36}$$

We introduce an additive penalty term to the log-likelihood, in the spirit of the Akaike information criterion (AIC), to compare models of different orders [18]:

$$-\log p(x|k,\theta) = \sigma^{-2} N \widetilde{E}_{\mathrm{noise}} + k . \tag{37}$$

For the profile likelihood, the noise variance σ^2 determines the balance between the data-fit and the penalty terms. From an algorithmic point of view, it is useful to rescale the log-likelihood as

$$-\log p(x|k,\theta) = \gamma(\widetilde{E}_{\mathrm{noise}}/E_x + k\, \varphi/M) \tag{38}$$

where $\gamma > 0$ determines the steepness of the likelihood function, which is useful to control particle depletion for the particle filter to be proposed in Sect. 5, and where φ is a normalized penalty parameter. If no source signals are present, the term $\tilde{E}_{\mathrm{noise}}/E_x$ is reduced by approximately $1/M$ for each added source. This gives a rule of thumb to choose φ. It is interesting to note that this expression can be derived by linearization of (34).

As evident from (38) and shown in Fig. 5 (middle row, see Sect. 4.1, for the simulation parameters and $\gamma = N$), an increase of φ leads to a larger penalty, i.e., the

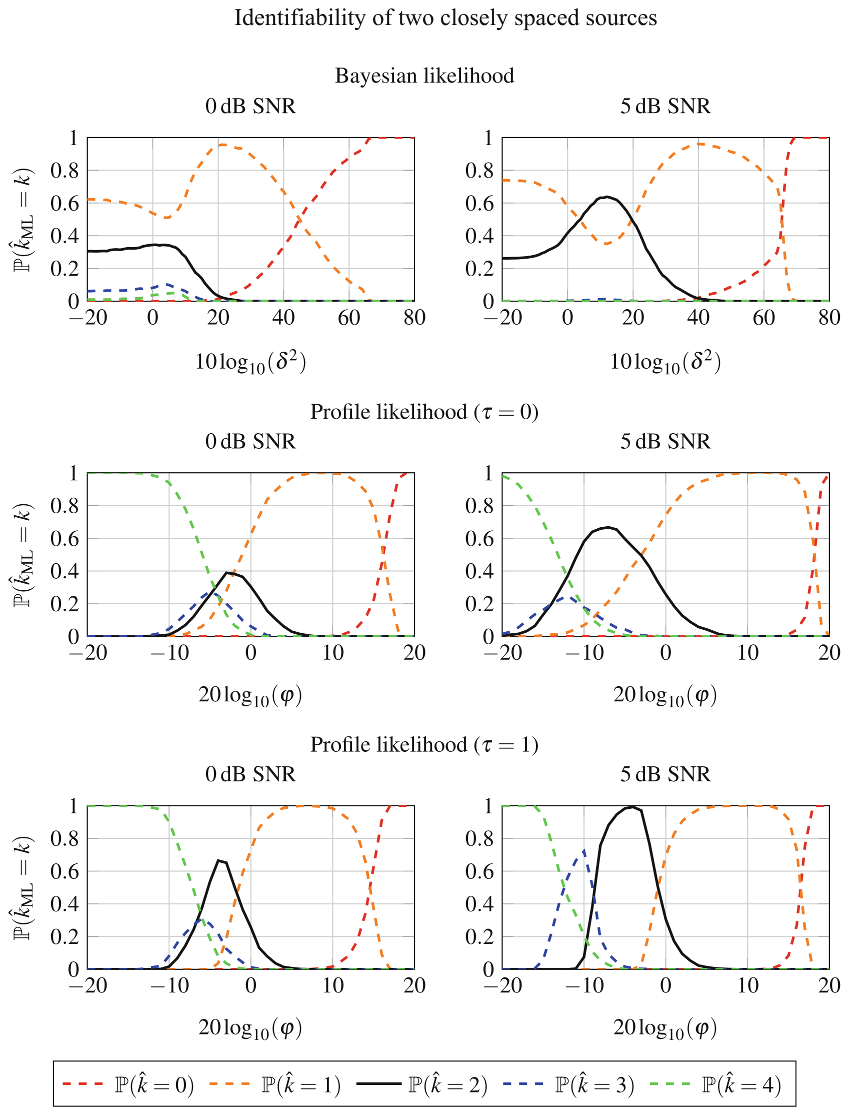


Fig. 5 Probability that the ML estimator for the number of sources equals k (*different lines*). The probability that the ML estimator finds the correct number of sources $k = 2$ is shown by the *black, solid curve*. *Top row*: Bayesian likelihood. *Center row*: profile likelihood with $\tau = 0$. *Bottom row*: profile likelihood with $\tau = 1$

parameter φ has a direct and predictable impact on the model choice. Furthermore, and in contrast to the Bayesian likelihood, there is a range of parameters φ for which $k = 2$ is the most probable outcome of the ML estimator (even for 0 dB SNR).

Finally, if a regularization parameter $\tau = 1$ is used, the probability of detecting the correct number of sources is increased (Fig. 5, bottom row).

5 Moving Sources—A Particle Filter Approximation

In the previous section, we analyzed and constructed a suitable likelihood function for the DOA problem. For static scenarios, it is relatively straightforward to combine the likelihoods from multiple observation periods into an overall likelihood function. The difficulty then is to develop an efficient optimization algorithm that finds the most likely number of sources and angles [25].

However, we are interested in situations where the sources are allowed to move between observation periods. Furthermore, the position estimates shall be updated when new observations are available. In this case, no analytic expressions for the resulting likelihood function can be found. Furthermore, the use of sliding windows would result in location estimates that lag behind. Classically, adaptive filters are used to address this issue. These recursively produce position estimates as convex combinations between a previous location estimate and one that is based on the current observation. A step size parameter determines the weighting between the old and the new estimates.

Stochastic adaptive filters, in contrast, do not update point estimates but rather the complete posterior probability distribution of the source locations. The most widely known stochastic adaptive filter is the Kalman filter, which needs a linear Gaussian relationship between the parameters and observations. For non-Gaussian and/or nonlinear environments as in the DOA problem, particle filters can generate sample-based approximations of the posterior distribution of the source locations. An estimator is needed on top of the filter to obtain point estimates (see Sect. 6).

In Sect. 5.1 we briefly introduce particle filters and show how these are used to solve the DOA problem for a known number of sources. Particle filters have fast convergence properties under the assumption that a complete description of the stochastic model is available. In particular, transition probabilities that describe the motion of the source locations must be available. Particle filters have been successfully applied to the narrowband DOA problem in [14], where a time domain formulation was used and to the wideband DOA problem in [31]. Subsequently, in Sect. 5.2, we introduce reversible jump particle filters [8] that also provide estimates for the number of sources. These have also been used for the narrowband DOA and frequency retrieval problems based on the Bayesian likelihood [3, 12].

5.1 Particle Filters

A crucial step in adaptive filter design is to choose the correct step size. For stochastic adaptive filters, this problem is solved by introducing a Markovian *transition kernel* that describes the stochastic dependence between the source locations of subsequent observation intervals. When the number of sources is known, we propose a transition kernel of the form

$$p(\theta_j(t)|\theta_j(t-1)) = \alpha\, p_{\mathscr{U}}(\theta_j(t)) + (1-\alpha) p_{\mathscr{N}}\left((\theta_j(t)-\theta_j(t-1))/\nu_\theta\right) \tag{39}$$

where $p_{\mathscr{U}}$ is a global distribution and $p_{\mathscr{N}}$ is a local distribution around the previous location with scale parameter ν_θ that adjusts the expected speed of the moving sources. In our simulations, we use a uniform distribution for $\sin\theta_\kappa$ as $p_{\mathscr{U}}$ and a normal distribution for $\sin\theta_\kappa$ as $p_{\mathscr{N}}$ with the endpoints of the interval $[-1,1]$ glued together.

This combination of local and global proposal densities is known as a *small world proposal density* [9] and is an attempt to integrate the initialization/detection part of an overall DOA estimation algorithm with the tracking part. The authors of [9] give a precise rule for the selection of α that requires exact knowledge of the posterior probability density function. However, they argue that $\alpha \in [10^{-4}, 10^{-1}]$ is a good rule of thumb.

Let $I(t)$ denote the collection of all measurements up to the current time $t \in \mathbb{N}$ and let $p(\theta|k, I(t))$ denote the posterior distribution of the source locations given all information $I(t)$ under the assumption of a constant and known number of sources k. At the first observation period, the posterior is proportional to the likelihood. Then, once new measurements become available, Bayes rule provides a recursive update formula

$$p(\theta(t)|k,I(t)) \propto \int p(x(t)|k,\theta(t))\, p(\theta(t)|\theta(t-1))\, p(\theta(t-1)|k,I(t-1))\, d\theta(t-1)\,. \tag{40}$$

Similarly, if $\left\{(\theta^i, w^i) : i = 1,\ldots,Q\right\}$ is a discrete representation of Q weighted samples of $p(\theta(t-1)|k,I(t-1))$,

$$p(\theta(t-1)|k,I(t-1)) = \sum_{i=1}^{Q} w^i \delta_{\theta^i}(\theta(t-1)) \tag{41}$$

where the δ_{θ^i} are Dirac masses at θ^i, then a representation of $p(\theta(t)|k,I(t))$ is obtained as follows [4]: First, for each *particle* (θ^i, w^i) a sample is generated from the transition kernel,

$$\tilde{\theta}^i \sim p(\tilde{\theta}^i|\theta^i)\,. \tag{42}$$

In a second step, the weights are updated with the likelihood and normalized,

$$\tilde{w}^i = \frac{w^i p(x(t)|k,\tilde{\theta}^i)}{\sum_{j=1}^{Q} w^j p(x(t)|k,\tilde{\theta}^j)}\,. \tag{43}$$

The new samples $\{(\tilde{\theta}^i, \tilde{w}^i) : i = 1, \ldots, Q\}$ are a discrete representation of $p(\theta(t)|k, I(t))$. The problem of particle depletion, all weight is concentrated on a single particle after several iterations, is addressed by introducing an optional step in which all indices $i = 1, \ldots, Q$ are resampled from the multinomial distribution described by the new weights $\tilde{w}^i$. This particle filter is known as a sampling importance resampling (SIR) filter [4]. To increase efficiency and reduce dependency among particles, this step is only done if the *effective number* of particles $Q_{\text{eff}} = (\sum_{i=1}^{Q}(w^i)^2)^{-1}$ falls below a threshold $Q_{\min}$.

The ratio between the scale parameter v_θ and the noise variance σ^2 determines the reactivity of the particle filter. That is, if much confidence is put into new measurements (σ^2 is small) or if it is assumed that the sources are fast moving (v_θ is large) the influence of the old approximation is reduced.

In practice, it may be advisable to use a larger noise variance in the particle filter algorithm if the true noise variance is small, as otherwise particle depletion becomes a problem. In any case, exact knowledge of the noise variance σ^2 and the velocity parameter v_θ is seldom available and we propose the following heuristic: First choose the parameter v_θ large enough that the particle filter can track fast moving sources, thereby keeping the trade-off between steady state estimation accuracy and tracking capability in mind. Second, use an online adaptation procedure that reduces, or increases, σ^2 if Q_{eff} decreases, or increases, too quickly. We found that a particle depletion rate of twenty percent per time step is healthy.

5.2 Reversible Jump Particle Filters

If the number of sources k is not known in advance, one extends the particle filter to generate samples from the joint posterior of k and θ, which is defined over the union of the sets $\Theta_k = \{-1 \leq \sin\theta_\kappa \leq 1, 1 \leq \kappa \leq k\}$ for $0 \leq k \leq k_{\max}$. Note that the dimension of Θ_k depends on k. In theory, it is possible to use the same particle filter as before, but one needs to ensure that the proportionality constants of the posterior do not vary across particles of different dimensions before the weights are jointly re-normalized. Resampling is done for each dimension separately. However, such an approach wastes resources as the number of particles per dimension is fixed from the beginning.

RJPFs that is, particle filters that incorporate single reversible jump MCMC moves [8], are more appealing [12]. These allow for transitions between sets Θ_k and $\Theta_{k'}$. Here, we restrict these transitions to moves between sets where $|k - k'| \leq 1$ and we speak of a birth move if the new number of sources k' equals $k + 1$ and of a death move if $k' = k - 1$. The moves can be introduced into the particle filter in the context of resample moves [7]. These are MCMC steps with a Markov kernel that has the posterior $p(\theta(1), k(1), \ldots, \theta(t), k(t)|I(t))$ as its invariant distribution. Such moves theoretically require knowledge of the complete past trajectory $\{k(t'), \theta(t') : t' \leq t\}$ for each particle, which is not practical. Therefore, it is proposed in [7] to approximate the resampling by only considering the most recent past.

A transition kernel that has $p(k(t),\theta(t)|x(t))$ as invariant distribution can be implemented as a mixture of local moves that update θ while leaving k unchanged and trans-dimensional moves that update k and keep the common parameters in θ unchanged. For an in-depth description of the reversible jump MCMC method, refer to [19].

In the following, we propose a resampling step that has been used in [12] and has $p(k(t),\theta(t)|x(t))$ as invariant distribution. In practice, this means that the posterior of $\theta(t)$ is biased toward the likelihood from the current time step, which is why we only update k and not θ. The trans-dimensional MCMC step is implemented by first selecting a *birth step* with probability b_k or a *death step* with probability d_k, where

$$b_k = c \min\left(1, \frac{p(k+1)}{p(k)}\right), \qquad d_k = c \min\left(1, \frac{p(k)}{p(k+1)}\right). \tag{44}$$

The parameter $0 \le c \le 0.5$ adjusts the frequency of birth and death steps. If a birth step is selected, a new source angle is generated according to $\sin\vartheta \sim \mathscr{U}_{[-1,1]}$ and inserted into θ at the right place, that is, $\theta^+ = \mathrm{sort}\{\theta,\vartheta\}$. Then, as in the ordinary Metropolis–Hastings algorithm, the acceptance probability of the proposal $(k+1,\theta^+)$ is calculated from the Metropolis–Hastings ratio (see [19], Proposition 1.11)

$$\begin{aligned} r_{\text{birth}}((k,\theta),(k+1,\theta^+)) &= \frac{d_{k+1}}{b_k}\frac{p(k+1,\theta^+|x(t))}{p(k,\theta|x(t))}\frac{1}{1/2} \\ &= \frac{d_{k+1}}{b_k}\frac{p(x(t)|(k+1,\theta^+))}{p(x(t)|(k,\theta))} \end{aligned} \tag{45}$$

where the factor $1/2$ is due to the uniform distribution on $[-1,1]$ and where the Jacobian appearing in [8] is, in our case, the identity matrix. The birth move is accepted with probability

$$\alpha = \min\left(1, r_{\text{birth}}((k,\theta),(k+1,\theta^+))\right). \tag{46}$$

The death move is the inverse of the birth move: a source $\kappa \in \{1,\dots,k\}$ is randomly selected and removed from θ. The remaining source angles are denoted θ^-. For this move, the Metropolis–Hastings ratio is given by

$$r_{\text{death}}((k,\theta),(k-1,\theta^-)) = r_{\text{birth}}((k-1,\theta^-),(k,\theta))^{-1} \tag{47}$$

and the move is accepted with probability

$$\alpha = \min\left(1, r_{\text{death}}((k,\theta),(k-1,\theta^-))\right). \tag{48}$$

The RJPF algorithm is summarized in Algorithm 3.

Algorithm 3 Reversible Jump Particle Filter

```
1. Initialize $k^i \sim p(k^i)$ and $\sin\theta^i_\kappa \sim \mathscr{U}_{[-1,1]}$ for $\kappa = 1,\dots,k^i$ and $i = 1,\dots,Q$
2. Set weights and calculate the effective number of particles

   $$w^i = \frac{p(x(1)|k^i,\theta^i)}{\sum_{j=1}^{Q} p(x(1)|k^j,\theta^j)} \qquad Q_{\mathrm{eff}}(1) = \left(\sum_{i=1}^{Q}(w^i)^2\right)^{-1} \qquad (49)$$

3. For $t = 2,\dots$, update the particles according to:
   a. Resample $\tilde{\theta}^i \sim p(\tilde{\theta}^i|\theta^i)$ from the transition kernel for each particle $i = 1,\dots,Q$
   b. Update weights and calculate the effective number of particles for each $i = 1,\dots,Q$:

      $$\tilde{w}^i = \frac{w^i p(x(t)|k^i,\tilde{\theta}^i)}{\sum_{j=1}^{Q} w^j p(x(t)|k^j,\tilde{\theta}^j)} \qquad Q_{\mathrm{eff}}(t) = \left(\sum_{i=1}^{Q}(\tilde{w}^i)^2\right)^{-1} \qquad (50)$$

   c. Set $(w^i,\theta^i) = (\tilde{w}^i,\tilde{\theta}^i)$ for $i = 1,\dots,Q$
   d. If the number of effective particles is too small, $Q_{\mathrm{eff}} < Q_{\mathrm{min}}$, resample indices $j_i, i = 1,\dots,Q$ from $p(j_i = \ell) = w^\ell$. Then set $(w^i,\theta^i) = (1/Q,\theta^{j_i})$
   e. Decrease the noise variance $\sigma^2$ by a small amount if the particle depletion ratio $1 - Q_{\mathrm{eff}}(t)/Q_{\mathrm{eff}}(t-1)$ is too large and increase $\sigma^2$ if the ratio is too small
   f. For each particle $i = 1,\dots,Q$, with probability $b_{k^i}$ perform a birth step:
      i. Propose $\vartheta$ according $\sin\vartheta^i \sim \mathscr{U}_{[-1,1]}$, insert $\vartheta^i$ into $\theta^i$ to obtain $\theta^+$, and calculate the Metropolis–Hastings ratio and acceptance probability $\alpha^i$ from (45) and (46)
      ii. With probability $\alpha^i$, set $\theta^i = \theta^+$ and $k^i \leftarrow k^i + 1$
   g. For each particle $i = 1,\dots,Q$, if no birth step was performed, do a death step with probability $d_{k^i}$:
      i. Randomly select an index $\kappa \in \{1,\dots,k^i\}$, remove $\theta^i_\kappa$ from $\theta^i$ to obtain $\theta^-$, and calculate the Metropolis–Hastings ratio and acceptance probability $\alpha^i$ from (47) and (48)
      ii. With probability $\alpha^i$, set $\theta^i = \theta^+$ and $k^i \leftarrow k^i + 1$
```

6 Estimation of the Number of Sources and Their Angles

At each time step, the RJPF provides a discrete representation of the current posterior probability distribution for the number of sources and their angles in the form of Q particles $(k^i,\theta^i,w^i), i = 1,\dots,Q$, of the number of sources, their angles, and weights. The goal of an estimator is to combine the information contained in all particles into a single estimate of the number of sources and their angles.

In Sect. 6.1, we discuss why this is a difficult problem and why we think that, in general, location estimates from particles of different dimensions should not be combined. After that, in Sect. 6.2, we propose a kernel-based method for combined estimation of the number of sources and their angles.

6.1 DOA Estimation for Nested Models

The joint estimation problem of the number of sources and their angles is a nested model choice problem with common parameters. One should, in general, refrain from estimating common parameters by averaging over all models [18].

An example why this is not advisable is that of two closely spaced sources, which we already encountered in Sect. 4.1. As shown in Fig. 6, the most likely source location of the model with a single source corresponds, approximately, to the mean value of the two true source locations. Hence, the information contained in the single-source likelihood function should not be used to improve estimates of either source in the two-source model.

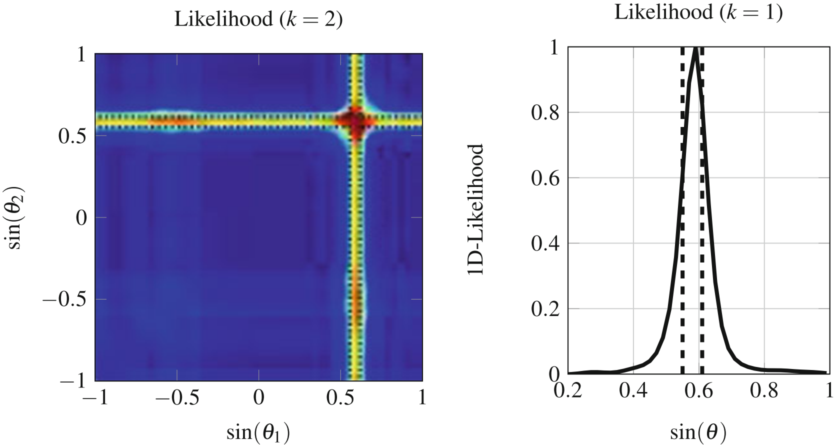


Fig. 6 Likelihoods in DOA estimation with ten microphones and two closely spaced sources. *Left*: likelihood conditioned on the presence of two sources. *Right*: likelihood conditioned on the presence of a single source. The *dashed lines* indicate the true source locations

If a RJPF is used, no particles $i = 1, \dots, Q$ with $k^i \neq k$ should be used to generate an estimator for k source directions. However, this reduces the efficiency of the filter, because the information contained in the discarded particles is unused. Therefore, it has been proposed to use particles from different dimensions regardless of the possible problems [19]. For well-behaved scenarios, e.g., if the number of microphones is large or if many independent frequency channels are available, the steering vectors $a_n(\theta_\kappa)$ are approximately orthogonal and the entries $\theta^i_\kappa, \kappa \leq k^i$ of a particle i of dimension k^i can be associated to the entries of another particle i' with $\theta^{i'}_{\kappa'}, \kappa' \leq k^{i'}$ and $k^{i'} > k^i$. That is, the common parameters also have common values and an estimator could be based on all particles. For such an estimator, however, one needs to associate the dimensions of particles with lesser dimensions to those of particles with higher dimensions, i.e., one needs to solve a *label switching* problem [19].

If, as we propose, one estimates k source locations and uses only particles with $k^i = k$, the label switching problem is reduced, but not removed. For example, if three sources are present at locations $\theta_\kappa, 1 \leq \kappa \leq 3$, then particles with $\theta_1^i \approx \theta_1, \theta_3^i \approx \theta_2$ and θ_2^i in between θ_1 and θ_2 are likely to survive several iterations of the particle filter. Thus, the particle is in fact tracking only two sources and should not be used for estimating three sources. However, these situations are difficult to detect. In the following section, we present a kernel based approach that uses only particles that have all entries close to the true source locations.

6.2 Kernel Based Estimation

In the following, we discuss how the number of sources k and their angles θ can be estimated using the weighted samples $(k^i, \theta^i, w^i), i = 1, \ldots, Q$ of the posterior distribution from the RJPF. One possible way to estimate the number of sources is to use the Bayesian MAP estimator

$$\hat{k}_{\mathrm{MAP}} = \arg\max_k \sum_{i=1}^{Q} w^i \delta_{k^i,k} \tag{51}$$

where $\delta_{k^i,k}$ denotes the Kronecker delta. In a second step, one calculates the MAP estimate $\hat{\theta}_{\mathrm{MAP}}$ using only those particles $i = 1, \ldots, Q$ with $k^i = \hat{k}_{\mathrm{MAP}}$. To calculate the MAP estimate $\hat{\theta}_{\mathrm{MAP}}|\hat{k}_{\mathrm{MAP}}$, we first need a kernel density estimate of the continuous posterior $p(\theta|\hat{k}_{\mathrm{MAP}}, I(t))$. We propose an iterative procedure that approximately solves the following $\hat{k}_{\mathrm{MAP}}$-dimensional nonlinear optimization problem: Let ψ^2 be the variance of a circularly symmetric Gaussian kernel. Then our goal is to solve

$$\hat{\theta}_{\mathrm{MAP},k} = \arg\max_\theta \sum_{i=1}^{Q} \exp\left(-\psi^{-2}\left\|\sin\theta - \sin\theta^i\right\|^2\right) w^i \delta_{k,k^i} \tag{52}$$

for $k = \hat{k}_{\mathrm{MAP}}$.

We propose to solve (52) as follows: At each iteration, marginal MAP estimates are calculated for each source $\kappa = 1, \ldots, k$. But instead of using the weights w^i, these are weighted by the product of the kernel functions from all other dimensions $\kappa' \neq \kappa$:

$$\hat{\theta}_\kappa = \arg\max_{\theta_\kappa} \sum_{i=1}^{Q} \exp\left(-\psi^{-2}\left\|\sin\theta_\kappa - \sin\theta_\kappa^i\right\|^2\right) w^i \delta_{k,k^i} \prod_{\kappa' \neq \kappa} \Phi(\kappa', i)\,. \tag{53}$$

The weighting function Φ, which is initialized with all ones, is then updated using the new marginal location estimate

$$\Phi(\kappa, i) = \exp\left(-\tilde{\psi}^{-2}\left\|\sin\hat{\theta}_\kappa - \sin\theta_\kappa^i\right\|^2\right). \tag{54}$$

This ensures that only those particles are used that have reliable entries θ_κ in all dimensions simultaneously and alleviates the difficulties hinted to in Sect. 6.1. To find a good balance between the number of particles on which the estimator is based and the accuracy of the estimator, we use two different kernel variances ψ^2 and $\tilde{\psi}^2 \gg \psi^2$. The first variance ψ^2 is small so that an MAP estimator is generated for each marginal (as ψ^2 is increased, the estimator resembles the minimum mean squared error (MMSE) estimator). The second variance $\tilde{\psi}^2$ is large to prevent that the estimator discards too many particles. Algorithm 4 provides a summary.

Figure 7 (top) shows the marginal distributions $p(\theta_\kappa)$ for the example from section "Wideband Signals in Subspace Method" where two pairs of closely spaced sources are recorded by an array of ten sensors. The peak for the marginal of the fourth source coincides with that of the marginal for the third source. At the final stage of the algorithm, the marginals are weighted with the product of the values of the kernel density functions of the other dimensions. These weighted marginals are shown in Fig. 7 (center). Those particles for which θ_4^i is close to θ_3 have smaller weights than those particles with θ_4^i close to θ_4, as desired. The final location estimate is then generated by fitting the kernel functions to the marginals, as shown in Fig. 7 (bottom).

Algorithm 4 Iterative kernel-based MAP estimator

```
1. Initialize $\Phi(\kappa,i) = 1$ for $\kappa = 1,\ldots,k$ and $i = 1,\ldots,Q$
2. Until converged, for $\kappa = 1,\ldots,k$ do
    a. Calculate weight factors $\tilde{w}^i = w^i \prod_{\kappa' \neq \kappa} \Phi(\kappa',i)$
    b. Calculate marginal estimate for $\theta_\kappa$ by solving the one-dimensional kernel problem
           $\hat{\theta}_\kappa|k = \arg\max_{\theta} \sum_{i=1}^{Q} \exp\left(-\psi^{-2}\left\|\sin\theta - \sin\theta_\kappa^i\right\|^2\right)\tilde{w}^i\delta_{k,k^i}$    (55)
    c. Update weights
           $\Phi(\kappa,i) = \exp\left(-\tilde{\psi}^{-2}\left\|\sin\hat{\theta}_\kappa - \sin\theta_\kappa^i\right\|^2\right)$    (56)
```

We found that the overall estimator for the number of sources and their directions can be improved if the estimation is performed jointly from

$$\hat{k}_{\mathrm{MAP}} = \arg\max_{k} \sum_{i=1}^{Q} \exp\left(-\psi_k^{-2}\left\|\sin\hat{\theta}_{\mathrm{MAP},k} - \sin\theta^i\right\|^2\right) w^i \delta_{k,k^i} \tag{57}$$

where ψ_k is such that the integrals over the kernels are the same for each dimension, i.e.,

$$\psi_k^{-2} = \sqrt[k]{(2\pi)^{k-1}\psi_1^{-2}}. \tag{58}$$

Of course, this increases the computational load of the algorithm.

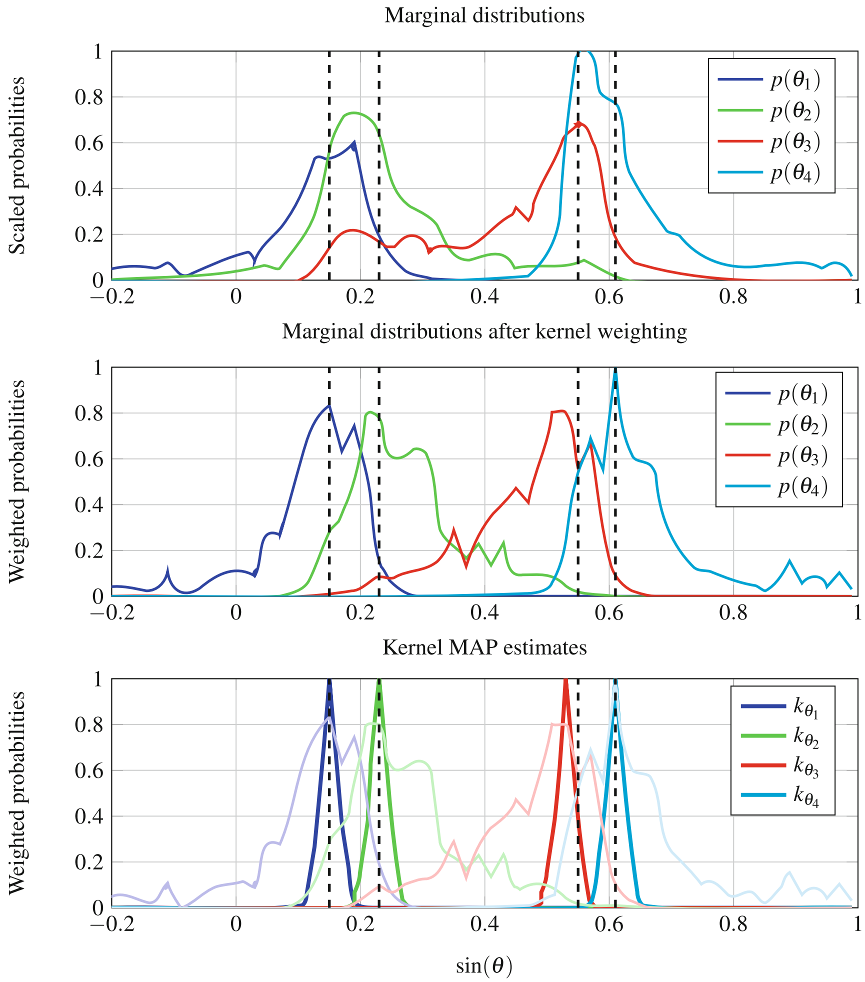


Fig. 7 *Top row*: marginal densities for source angles as generated from the particle filter. *Center row*: weighted marginal densities after Algorithm 4 has been applied to find the source locations. *Bottom row*: fitted Gaussian kernel functions that approximately solve (52)

7 Simulations

We first use the same simulation as in Sect. 3.4 to compare the particle filter algorithm performance with the iterative estimator to subspace methods for a known number of sources (we refer the reader to Sect. 3.4 for the simulation parameters). The particle filter uses a total of $Q = 2000$ particles. The transition kernel is a mixture of a uniform distribution on $[-1,1]$ for $\sin\theta_\kappa$ and a normal distribution centered at $\sin\theta_\kappa(t-1)$ with standard deviation 0.03. The mixing parameter is set to $\alpha = 0.01$. The scaling factor is initially set to $\gamma = N = 5$ and increases by 10 % if $1 - Q_{\text{eff}}(t)/Q_{\text{eff}}(t-1) < 0.25$ or decreases by 10 % if $1 - Q_{\text{eff}}(t)/Q_{\text{eff}}(t-1) > 0.5$.

Resampling is done if $Q_{\mathrm{eff}} < 0.2Q$. The parameters for the kernel based MAP estimator are set to $\psi = 0.02$ and $\tilde{\psi} = 0.1$. Before the first time step, the particles are initialized according to a uniform distribution for $\sin\theta_\kappa^i, i = 1,\ldots,Q, \kappa = 1,\ldots,4$.

Figure 8 shows that the performance of the particle filter and the kernel based MAP estimator is superior to that of subspace methods. Interestingly, at low SNR values, the particle filter exhibits a higher detection rate than the ML estimator. This might be explained by the combination of the uniform initialization, the MAP estimator, and a smoothing effect of the transition kernel, which could lead to a larger region of attraction of the correct peak of the likelihood.

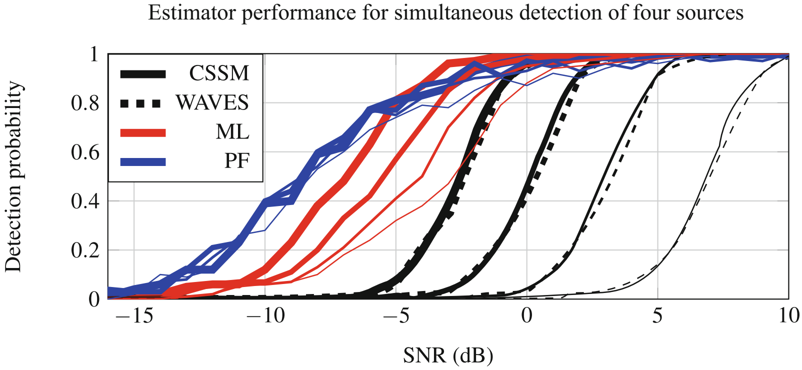


Fig. 8 Probability of simultaneous detection of four sources located at $\theta_1 = 8$, $\theta_2 = 13, \theta_3 = 33$, and $\theta_4 = 37$ (degrees). The correlation coefficient ρ increases from the *thick lines* to the *thin lines* as follows: $\rho = 0.0, 0.5, 0.75, 0.9$

In the following, we highlight the capabilities of the RJPF on a second series of simulations. The simulation parameters are as described in Sect. 3.4 except for the number of sources and their positions. The total number of sources is varied ($k = 1,\ldots,4$) and one of the sources only transmits intermittently. Furthermore, one source moves with constant velocity and crosses the remaining sources. The SNR is set to 0 dB in each experiment. The parameters for the particle filter are as before, except that no mixing is used ($\alpha = 0$). For the RJPF, this role is assumed by birth moves, which occur with probability $b_k = 0.1$ for $1 \leq k \leq 5$ and with $b_0 = 1$ so that $k_{\max} = 6$. For the probability of death moves, we also use $d_k = 0.1$ for $1 \leq k \leq 6$. The penalty parameter for the model complexity is set to $\varphi = 1.2$, that is, slightly larger than one. In particular, we used the same parameters for all simulations.

The RJPF is compared with an empowered implementation of a CSSM based tracking method that is based on the 20 most recent observations. The focusing matrices for the CSSM method are calculated using the true source locations of all active sources. Before the first appearance of the first source, random values are chosen for its angle. For the remaining intervals of inactivity, the final location of the previous activity period is used.

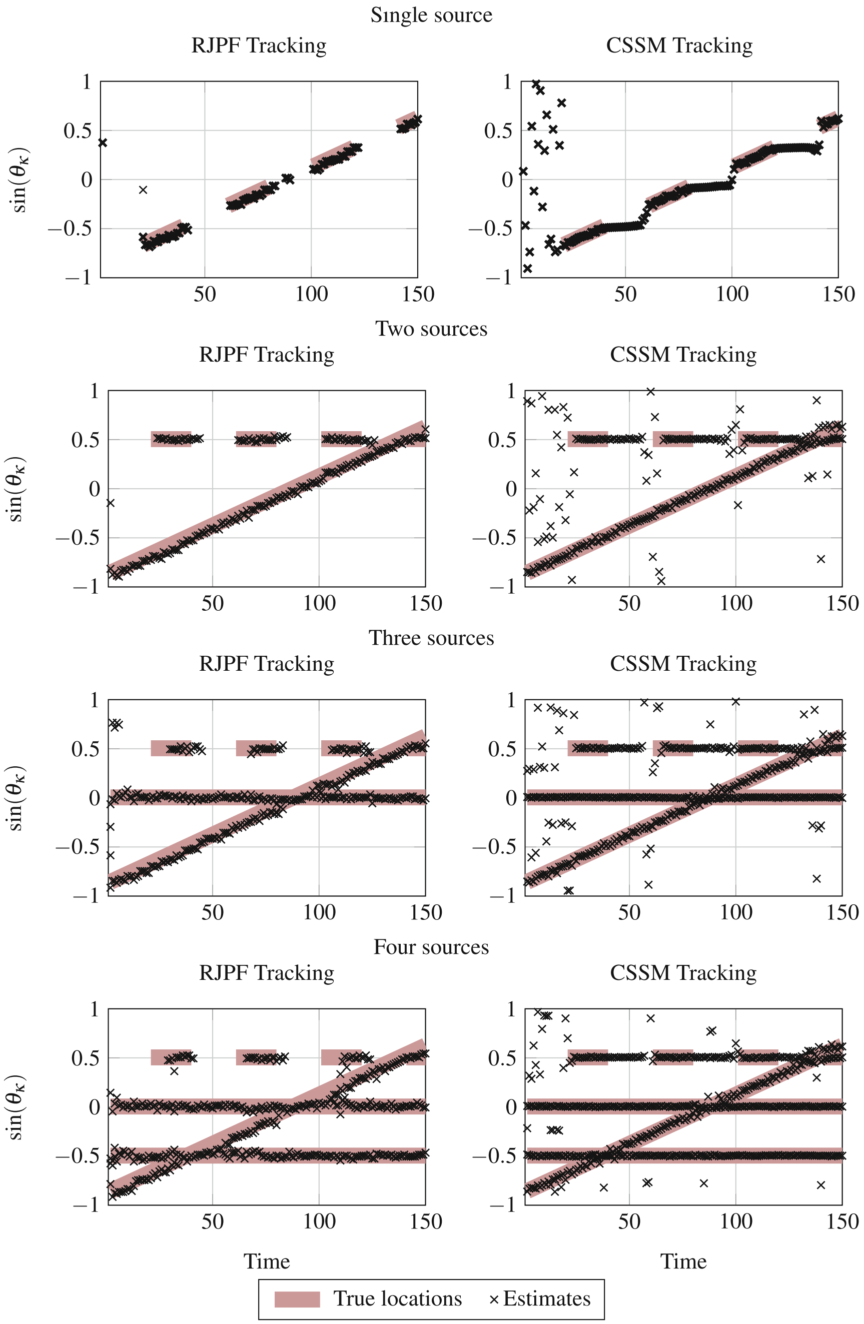


Fig. 9 Results of reversible jump particle filter (RJPF) and CSSM for tracking a variable number of sources

Figure 9 (left column) shows that the RJPF provides reliable estimates of the number of sources and their positions. Obviously, if two sources are at the same position, they are indistinguishable and correctly detected as a single source. In contrast, if CSSM is used with all of the true source angles given as input, erroneous estimates result at trajectory crossings (right column) or during outages. The estimates of the RJPF slightly lag behind the true trajectory. This lag can be reduced by increasing the standard deviation of the local transition kernel. The reason why the estimates for CSSM tracking do not show this lag is that the focusing matrices are generated using the true source locations. Such an assumption is necessary for generating baseline CSSM results or the CSSM method would not work.

8 Conclusion

Bayesian DOA estimation gives the framework for extensions beyond what has been achieved with subspace based methods. This chapter extends a prior work using particle filter concepts to wideband DOA estimation of a variable number of moving sources. We present a detailed analysis of the DOA problem and highlight difficulties that lie with the Bayesian formulation, in particular regarding the choice of the δ^2 parameter. We depart from the rigorous Bayesian framework and introduce the profile likelihood function and corresponding regularization and show that this approach improves the estimation of the correct number of sources. We also argue that the label switching problem does not need to be solved in theory. That is, if enough particles are available one can safely discard all particles that track an incorrect number of sources.

Future research will develop RJPFs that automatically adjust some of the model parameters, most notably the penalty parameter for model complexity. We believe that this can be achieved by allowing interaction between the estimator and particle filter, specifically by using information from the estimator to adapt parameters of the particle filter.

References

1. Abramovich YI, Gray DA, Gorokhov AY, Spencer NK. Positive-definite toeplitz completion in DOA estimation for nonuniform linear antenna arrays. Part I. Fully augmentable arrays. IEEE Trans Signal Process. 1998;46(9):2458–71.
2. Abramovich YI, Spencer NK, Gorokhov AY. Identifiability and manifold ambiguity in DOA estimation for nonuniform linear antenna arrays. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP). 1999;5:2845–8.
3. Andrieu C, Doucet A. Joint Bayesian model selection and estimation of noisy sinusoids via reversible jump MCMC. IEEE Trans Signal Process. 1999;47(10):2667–76.
4. Arulampalam S, Maskell S, Gordon N, Clapp T. A tutorial on particle filters for on-line non-linear/non-Gaussian Bayesian tracking. IEEE Trans Signal Process. 2001;50:174–88.

5. Bresler Y, Macovski A. On the number of signals resolvable by a uniform linear array. IEEE Trans Acoust Speech Signal Process. 1986;34(6):1361–75.
6. Di Claudio ED, Parisi R. WAVES: weighted average of signal subspaces for robust wideband direction finding. IEEE Trans Signal Process. 2001;49(10):2179–91.
7. Doucet A, Johansen AM. A tutorial on particle filtering and smoothing: fifteen years later. In Crisan D, Rozovsky B, editors, Handbook of nonlinear filtering. Cambridge: Cambridge University Press; 2009.
8. Green PJ. Reversible jump Markov chain Monte Carlo computation and Bayesian model determination. Biometrika. 1995;82(4):711–32.
9. Guan Y, Fleißner R, Joyce P, Krone SM. Markov chain Monte Carlo in small worlds. Stat Comput. 2006;16(2):193–202.
10. Hung H, Kaveh M. Focussing matrices for coherent signal-subspace processing. IEEE Trans Acoust Speech Signal Process. 1988;36(8):1272–81.
11. Krim H, Viberg M. Two decades of array signal processing research: the parametric approach. IEEE Signal Process Mag. 1996;13(4):67 –94.
12. Larocque J-R, Reilly JP, Ng W. Particle filters for tracking an unknown number of sources. IEEE Trans Signal Process. 2002;50(12):2926–37.
13. Lee T-S. Efficient wideband source localization using beamforming invariance technique. IEEE Trans Signal Process. 1994;42(6):1376–87.
14. Orton M, Fitzgerald W. A Bayesian approach to tracking multiple targets using sensor arrays and particle filters. IEEE Trans Signal Process. 2002;50(2):216–23.
15. Pertilä P. Online blind speech separation using multiple acoustic speaker tracking and time-frequency masking. Comput Speech Lang. 2013;27(3):683–702.
16. Pillai S, Haber F, Bar-Ness Y. A new approach to array geometry for improved spatial spectrum estimation. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume 10, 1985 p. 1816–19.
17. Proukakis C, Manikas A. Study of ambiguities of linear arrays. In IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP), volume iv, 1994. pp. 549–52.
18. Robert CP. The Bayesian choice. 2nd ed. New York: Springer; 2001.
19. Roodaki A. Signal Decompositions Using Trans-Dimensional Bayesian Methods. PhD thesis, Ecole Supérieure d'Electricité, Gif-sur-Yvette, France, 2012.
20. Roodaki A, Bect J, Fleury G. An empirical Bayes approach for joint Bayesian model selection and estimation of sinusoids via reversible jump MCMC. In European Signal Processing Conference (EUSIPCO), 2010. pp. 1048–52.
21. Roodaki A, Bect J, Fleury G. On the joint Bayesian model selection and estimation of sinusoids via reversible jump MCMC in low snr simulations. In International Conference on Information Sciences, Signal Processing and their Applications (ISSPA), 2010. pp. 5–8.
22. Roodaki A, Bect J, Fleury G. Relabeling and summarizing posterior distributions in signal decomposition problems when the number of components is unknown. arXiv:1301.1650 [stat], January 2013.
23. Roy R, Kailath T. ESPRIT-estimation of signal parameters via rotational invariance techniques. IEEE Trans Acoust Speech Signal Process. 1989;37(7):984-95.
24. Schmidt R. Multiple emitter location and signal parameter estimation. IEEE Trans Antennas Propag. 1986;34(3):276–80.
25. Stoica P, Moses Rl, Friedlander B, Soderstrom T. Maximum likelihood estimation of the parameters of multiple sinusoids from noisy measurements. IEEE Trans Acoust Speech Signal Process. 1989;37(3):378–92.
26. Stoica P, Nehorai A. Performance study of conditional and unconditional direction-of-arrival estimation. IEEE Trans Acoust Speech Signal Process. 1990;38(10):1783–95.
27. Swingler DN, Krolik J. Source location bias in the coherently focused high-resolution broadband beamformer. IEEE Trans Acoust Speech Signal Process. 1989;37(1):143 –45.
28. Viberg M, Ottersten B. Sensor array processing based on subspace fitting. IEEE Trans Signal Process. 1991;39(5):1110 –21.

29. Wang H, Kaveh M. Coherent signal-subspace processing for the detection and estimation of angles of arrival of multiple wide-band sources. IEEE Trans Acoust Speech Signal Process. 1985;33(4):823–31.
30. Wax M, Ziskind I. On unique localization of multiple sources by passive sensor arrays. IEEE Trans Acoust Speech Signal Process. 1989;37(7):996–1000.
31. Wiese T, Claussen H, Rosca J. Particle filter based DOA estimation for multiple source tracking (MUST). In Asilomar Conference on Signals, Systems and Computers, 624–28 November 2011.
32. Yoon Y-S, Kaplan LM, McClellan JH. TOPS: New DOA estimator for wideband signals. IEEE Trans Signal Process. 2006;54(6):1977–89.

Advances in Radar Waveform Development

Menachem Levitas and Carroll Nunn

Abstract This chapter provides a brief overview of the current state-of-the-art in radar waveforms and waveform processing development. It begins by describing the first principles of pulse compression, the traditional performance measures, and some of the waveforms and processing commonly used in existing systems. With this background as reference, it proceeds to introduce additional spectral and temporal requirements imposed on waveforms utilized by state-of-the-art radar systems as results of greater detection sensitivities, more difficult clutter environments, higher anticipated spectral occupancies, and closer relationship between waveforms and antenna patterns. It then describes new technological capabilities, both algorithmic and hardware-related, which make these complex requirements attainable, and provides several examples of waveforms that meet such requirements.

Keywords Waveform · Matched-filter · Mismatched-filter · Pulse-compression · Time-sidelobes · Spectral-compliance · Auto-correlation · Cross-correlation

1 The Time–Bandwidth Product

Due to the Fourier transform relationship of temporal and spectral behavior, the temporal resolution of a waveform is related to its bandwidth as per the inverse relation as described in Eq. 1:

$$\sigma(t) \propto \frac{1}{B} \tag{1}$$

M. Levitas (✉)
900 Arrington Dr., Silver Spring, MD 20901, USA
e-mail: levitasmenachem@gmail.com

C. Nunn
Johns Hopkins University Applied Physics Lab, Laurel, USA
e-mail: Carroll.Nunn@jhuapl.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_12

The corresponding relation between range resolution and bandwidth is given by,

$$\sigma(r) = \frac{c \cdot \sigma(t)}{2} \propto \frac{c}{2 \cdot B} \tag{2}$$

In which the letter c represents the speed of light.

The relation between the duration of a pulse, τ, and its corresponding bandwidth, B, depends on the details of the waveform. When it consists of pure sine wave modulation, the bandwidth is given by,

$$B \approx \frac{1}{\tau} \tag{3}$$

Whereas, the general relation is given by:

$$B \cdot \tau \geq 1 \tag{4}$$

Figure 1 illustrates both cases:

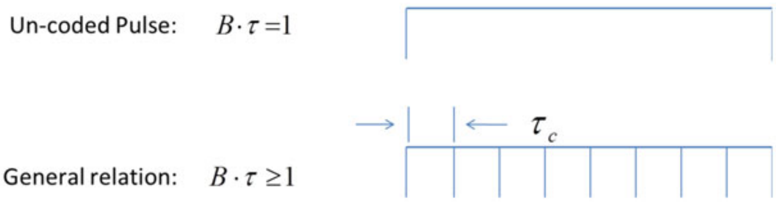

Un-coded Pulse: $B \cdot \tau = 1$
τ_c
General relation: $B \cdot \tau \geq 1$

Fig. 1 Time-bandwidth relationship between coded and un-coded pulses

The product, n, of the pulse duration and the pulse bandwidth is called the time-bandwidth product—see Eq. 5 below. We note that, whereas the pulse duration is on the order of its temporal resolution in the absence of coding, in the presence of coding it is larger.

$$B \approx \frac{n}{\tau} = \frac{1}{\tau_c} \, i.e. B \cdot \tau \approx n \tag{5}$$

Where, $\tau_c = \frac{\tau}{n}$, is referred to as the compressed pulsewidth, and represents the main response width of the processed pulse.

Several considerations influence the time-bandwidth product selection:

1. The bandwidth is determined from the lowest acceptable range resolution $B \geq \frac{c}{2 \cdot \delta(r)_{max}}$.
2. The product of peak power and pulse duration, which is the energy contained in the pulse, is partially determined by required detection sensitivity: i.e., the

pulse energy should be sufficient to support detection sensitivity requirements

$$E = P_T \cdot \tau = n \cdot P_T \cdot \tau_c \geq E_{min}. \tag{6}$$

3. Low peak power is desired for practicality of implementation and for low signal detectability by hostile receivers

$$P_T \leq P_{T,max}. \tag{7}$$

The time–bandwidth product, achieved via coding, is used to facilitate the required trade-off

$$Time\!-\!bandwidth\,product = n \geq \frac{E_{min}}{P_T \cdot \tau_{c,max}} = \frac{E_{min} \cdot c}{P_T \cdot 2 \cdot \delta\,(r)_{max}}. \tag{8}$$

The processing of a coded pulse is accomplished via temporal convolution (or its equivalent processing in the frequency domain) against matched or mismatched filters, and is called ‘Pulse Compression’. Matched filter processing consists of convolving the pulse waveform against its complex conjugate. Mismatched filter processing consists of convolving the pulse waveform against a different i.e., mismatched code, which may also be matched or mismatched in duration i.e., code length: See Fig. 2.

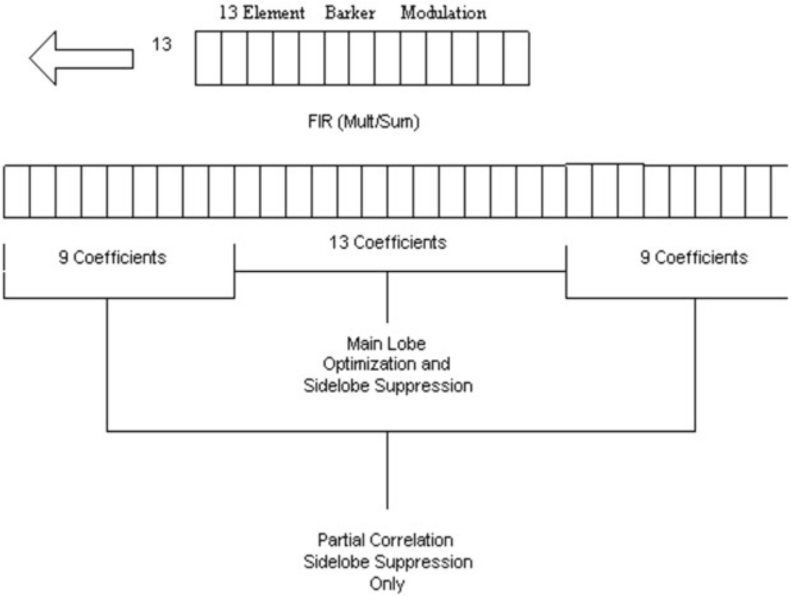

13 Element Barker Modulation
13
FIR (Mult/Sum)
9 Coefficients
13 Coefficients
9 Coefficients
Main Lobe Optimization and Sidelobe Suppression
Partial Correlation Sidelobe Suppression Only

Fig. 2 An example of a code and mismatched length filter

Conventional selection of codes and filters involves several performance considerations, the most common of which includes: minimum peak power loss and main response broadening, and required peak and integrated time-sidelobe levels. They are discussed below together with several other performance measures.

2 Traditional Performance Measures From Codes/Filters

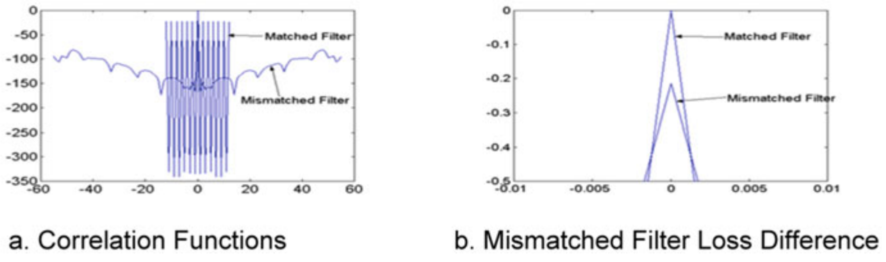
The discussion here will focus on several performance requirements in rough order of importance. It will be noted that the so-called order of importance is very general and can vary from application to application.

2.1 Maximizing Sensitivity

Detection sensitivity is maximized by maintaining constant signal amplitude, and by minimizing the loss. The loss is defined in Eq. 9 as the difference in peak signal-to-noise ratio obtained under matched filter conditions, and under the mismatched filter implemented.

$$Loss_{dB} = SNR_{matchedfilterpeak} - SNR_{mismatchedfilterpeak}. \tag{9}$$

Figure 3 below shows the correlation functions of a certain example code against its matched filter and against a certain mismatched filter. The right hand plot shows their corresponding peak SNR responses. Whereas the time-sidelobes achieved via the mismatched filter are much superior i.e., significantly lower than those achieved via matched filtering, the loss penalty, i.e., lower peak SNR response, is evident. Part of the waveform design problem is to achieve required time sidelobe levels without exceeding the maximum allowed loss.



a. Correlation Functions

b. Mismatched Filter Loss Difference

Fig. 3 Correlation functions of matched and mismatched filters along with mismatched filter loss

2.2 Maintaining Required Range Resolution

Range resolution is maintained by keeping the peak response broadening within required levels. Figure 4 shows several filter broadenings that are by-products of several corresponding mismatched filters when applied to the so-called P4 code, together with the corresponding peak matched filter response. (The P4 code is a digital form of the linear FM code which is discussed further below.) To provide adequate visual comparison of the various broadenings, the peak responses in Fig. 4 are all normalized to unity.

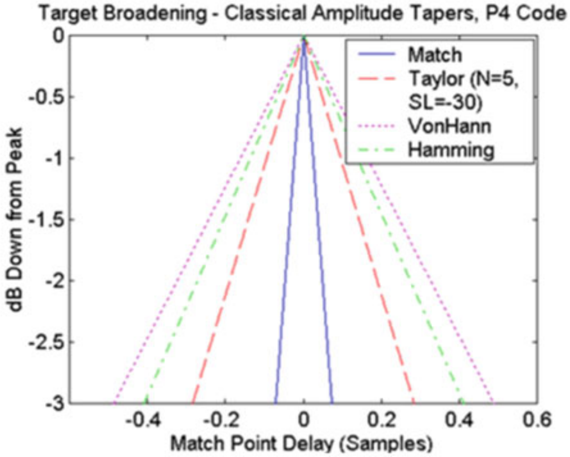


Fig. 4 Peak responses obtained from a matched filtered P4 code, and from several mismatched filters applied to the P4 code

2.3 Peak and Integrated Time-Sidelobes to Minimize Eclipsing by a Large Discrete or by Large Distributed Clutter

Unless the time-sidelobe response level at the position of the scatterer is sufficiently small, a large discrete scatterer placed in the time-sidelobes could eclipse a small target placed at the peak response. This condition places requirements on the peak time-sidelobe. Distributed clutter can be sensed from the entire time-sidelobe region as it adds-up across the sidelobe extent. The need to attenuate such strong distributed clutter places requirements on the so-called integrated time-sidelobes, which is their power sum. Figure 5 [1] shows three levels of time sidelobes achieved from a certain example code using mismatched filters of three different lengths. It is seen that extremely low time-sidelobes (peak and integrated) can be obtained via long mismatched filters. In practical radar systems, a so-called 'noise-floor', generated by imperfect hardware, will prevent the time-sidelobes from going below certain levels.

The waveform designer will, therefore, be advised not to over-design, since going beyond a certain time-sidelobe performance level will result in increasing losses and peak broadenings, without corresponding benefits in terms of lower time-sidelobes

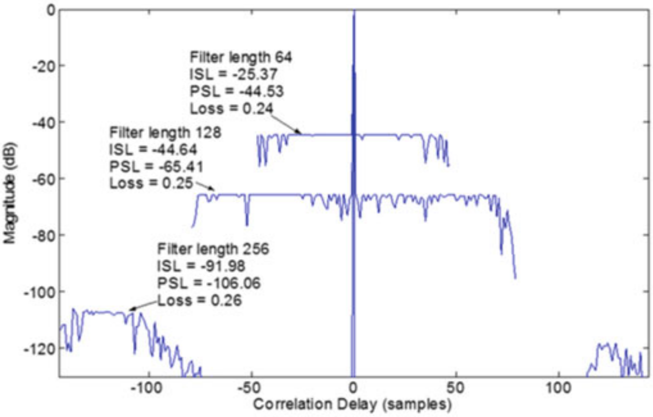


Fig. 5 Code correlation against three mismatched filters of increasing lengths

2.4 Low Doppler Sensitivity

Some radar applications require low Doppler sensitivity i.e., the ability to achieve detection sensitivity independent of target Doppler. The corresponding waveform requirement is that the full pulse compression performance be obtained independent of target Doppler. Some pulse compression codes e.g., the Barker family are very sensitive to target Doppler and may, therefore, not be suitable for such applications. On the other hand, other codes linear frequency modulation (FM) are extremely tolerant to Doppler. Figure 6 (generated by Dr. Lawrence Welch) shows on the left a three dimensional (3D) diagram which represents the so-called Pseudo-noise (PN) code ambiguity diagram . Range response is shown along one horizontal axis and the Doppler response across the other. It is seen that the peak response exists only for very low Doppler values. If a code of this nature is utilized in conjunction with, e.g., very fast closing objects, the signal needs to be channeled across several parallel filters, in each of which, a different Doppler compensation is to be applied prior to compression. These parallel filters require additional processing power. The plot on the right of Fig. 6 (generated by Dr. Lawrence Welch) shows a similar diagram for the P4 code. Clearly, the peak response of this code is nearly independent of the target Doppler.

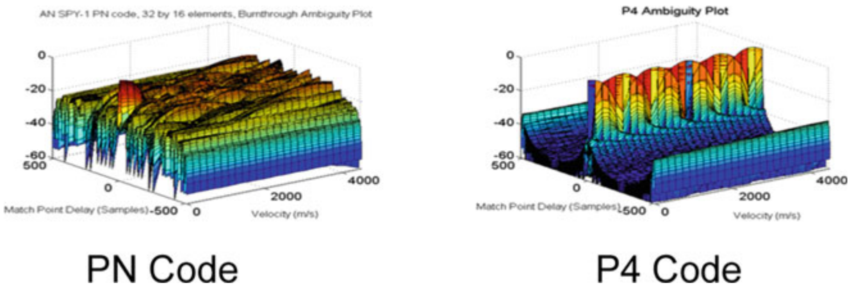


Fig. 6 Range Doppler ambiguity diagrams

2.5 *Spectral Compliance*

Certain levels of spectral compliance have always been required of radar codes. They were specified in terms of far sideband levels and rate of decline across the transition from the in-band to the side-band regions. Due to the trend of increasing spectral occupancy, these requirements are now becoming more severe and far more detailed. We will discuss these separately in later sections of this chapter.

2.6 *Multiple Uncorrelated (Orthogonal) Codes*

Requirements for multiple uncorrelated codes existed in some past applications. (The PN codes occasionally referred to also as PRN (for Pseudo Random Noise) codes fulfilled this requirement. Far more exacting requirements are associated with state of the art and future systems. Their increased challenge is partially due to their combination with other requirements and, partially, due to the occasional need to implement stated levels of partial cross-correlation, resulting in various levels of pseudo-orthogonality. They will be discussed further below.

2.7 *Potential Generation in Real Time*

Some codes were utilized in older systems due to their ease of implementation via analog circuits. Today, with the advent of powerful digital systems and arbitrary waveform generators, these considerations are no longer relevant.

It is seen that, for a waveform to be capable of meeting the array of potentially conflicting requirements described above, many degrees of freedom will be needed. Some of these are supplied by the code itself. A mismatched filter, especially when mismatched in length, will add more. That is the reason why, in stressing future applications, one will generally be talking not in terms of codes alone, but in terms of code/filter pairs.

3 Some Examples of Traditional Codes and Their Respective Performances in Conjunction With Matched and Mismatched Filters

Radar pulse signals can be un-coded, frequency-coded, or phase-coded (see Fig. 7).

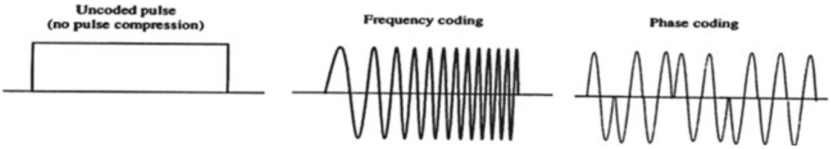


Fig. 7 Un-coded, frequency-coded, and phase-coded pulses

In what follows, we describe a number of traditional codes together with their respective properties. We do not attempt to cover all of the possible traditional codes and their transmission modulations. For that, extensive literature exists. Here we merely wish to provide a glimpse into the world of existing codes and to the fact that each of these codes has fixed characteristics that render it acceptable under a limited set of circumstances, but not necessarily when the requirements become diverse and difficult to meet.

3.1 Linear FM Codes

The linear frequency modulated (linear FM, or LFM) code increases, or decreases the frequency in linear fashion under the pulse envelope, as is shown in Fig. 8. The LFM codes have been the most popular due to their ease of implementation, analog processing, and Doppler tolerance. A popular compression scheme has been traditionally implemented by converting the Radio Frequency (RF) signal to an acoustic signal equivalent using a suitable transducer, and processing the resulting acoustic signal via a dispersive, so-called, surface acoustic wave (SAW) device. The dispersive device provides the required filtering resulting in a compressed signal that is then converted back to electric signal.

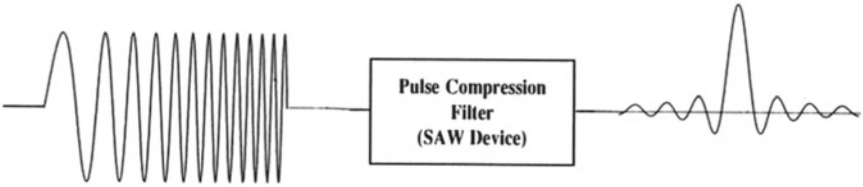


Fig. 8 Linear FM code, compression filtering, and compressed pulse

The principal characteristics of the LFM signal are as follows: (a) Continuous waveform, (b) Good Doppler tolerance, (c) Range–Doppler coupling, (d) Easy to generate and process, (e.) Easy to filter with classical windows, (f.) High matched filter time-sidelobes, resulting in need to utilize mismatched filters, and (g.) A single code: i.e., no diverse code family. The Range–Doppler coupling is due to the fact that the same compressed waveform results from target range displacement as from a certain corresponding target Doppler speed.

3.2 Lewis-Kretschmer Palindromic P4 Code [2]

The P4 code is a digital code in which the phase of the i'th element is given by Eq. 10.

$$\phi_i = \left(\frac{45}{n}\right)(2i-1)^2 - 90\,(2i-1)\,. \tag{10}$$

Where ϕ_i are the discrete phases of the expanded transmit pulse in degrees, n is the transmitted pulse sequence length, and i is the CHIP (change in phase) number (i = 1 to M)

From the quadratic relationship of the phase to the CHIP number, it is evident that the P4 code is a digital form of LFM and that, therefore, it shares in the above LFM properties.

3.3 Nonlinear FM Code [3]: Frequency Varies NonLinearly in Time

The ability to vary the frequency nonlinearly in time, provides the ability to create spectral weighting functions that can be designed to provide very low time sidelobes. The mechanization is illustrated via Fig. 9 below, which shows how the carrier frequency is swept at such rates so as to provide the spectrum a desired weighting function. The principal characteristics include: (a.) Continuous waveform, (b.) Ability to generate low time-sidelobes with short codes, (c.) Doppler tolerance, (d.) Ease of generation, (e.) Very large peak response broadening, i.e., it requires - sometimes significantly—broader bandwidth to achieve stated range resolution.

3.4 The Barker Codes [4]

The Barker code family represents a small group of constant-amplitude, bi-phase codes which have the lowest peak sidelobe level possible. The longest member of this family has the length of 13 which is not long enough for many applications. Its principal characteristics are as follows: a. Codes are bi-phase, with peak compressed value equal to the square of the code's length n, b. Peak sidelobe level is 1 under

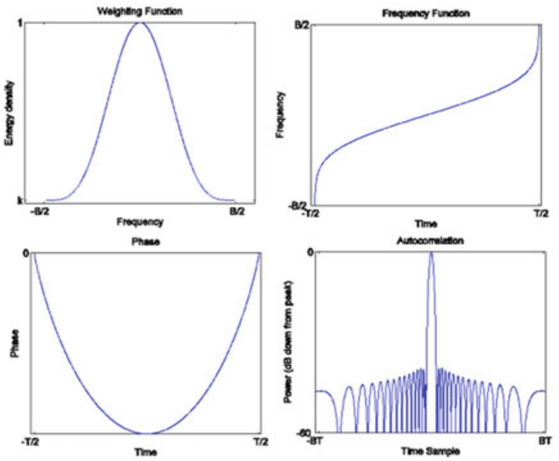


Fig. 9 Nonlinear frequency modulation process. Clockwise from *top left*: frequency weighting function; sweep rates needed to accomplish that; phase vs. time; and compressed pulse

matched filtering (i.e., $1/n^2$ beneath the peak response), c. Longest known code is 13, d. Doppler sensitivity is high;

The code [+1, +1, +1, -1, +1] is an example that features Barker code length of 5. The matched filter is the same as the code. Figure 10 shows the compressed pulse in the absence of Doppler shift.

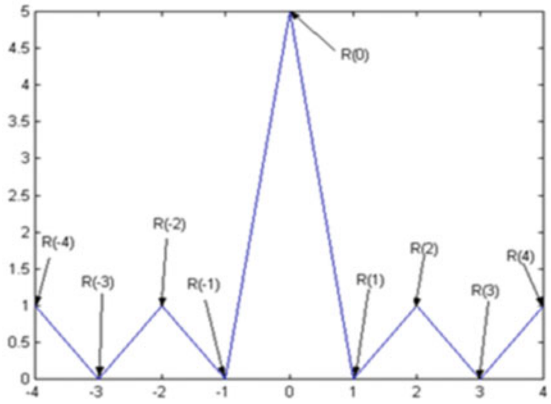


Fig. 10 Compressed form of Barker 5 waveform via matched filter (voltage domain)

Characteristics of this code include:

1. Peak power = 14 dB
2. Peak sidelobe = 0 dB (i.e. 14 dB beneath the peak response)
3. Integrated sidelobes = 8 dB beneath the peak response
4. Root-mean-square (RMS) sidelobes = 17.5 dB beneath the peak response.

3.5 *The Pseudo-Random-Noise (PRN, or PN), or, Maximal Length Codes [5]*

The PN (or PRN) family of codes provides a large number of orthogonal, pseudo-random, bi-phase codes that are easily generated, on-the-fly, via either analog or digital means. Their generation process is described in Fig. 11. Their principal characteristics include the following interesting mathematical properties:

1. Codes are Bi-Phase
2. They exist in odd lengths, i.e., 2N−1
3. Large number of orthogonal codes exist
4. The number of 1's and −1s differ by 1
5. 1/2 the runs are length 1, 1/4 are length 2, 1/8 are length 3, etc.
6. They are generated using a shift register with feedback (see Fig. 11)

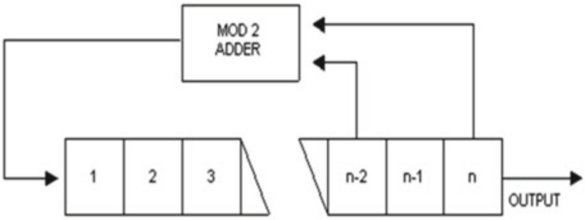


Fig. 11 Generation of the PN code family via shift register and a MOD 2 adder

Practical characteristics include the facts that (a) generating these codes on-the-fly in a live radar eliminates need to store codes in memory, (b) The peak time-sidelobe power is 1/n beneath the peak, and (c) The codes are very Doppler sensitive.

4 New Radar Waveform Requirements, Derived from Increasingly More Challenging Performance Requirements Coupled with More Strenuous Operating Environments

Current, state-of-the-art, and future radar systems are facing ever increasing challenges and their corresponding performance requirements become correspondingly more stringent. These, in turn, reflected in corresponding tightening of requirements on the waveform components of such systems. The following paragraph summarizes the operational challenges and corresponding waveform requirements.

Decreasing target signal and increasing clutter background levels Together with ever increasing detection range requirements, the reflected echoes from targets of interest also become weaker owing to smaller physical sizes of such threats and to stealthier designs. At the same time, competing clutter becomes progressively more severe, with littoral scenarios becoming predominant for Navy operations and with urban clutter gaining in prominence. Medium PRF radar waveforms, which are characterized by both range and Doppler ambiguities, will cause strong near-by clutter to fold together with weak, remote, target echoes, further exacerbating the dynamic-range and signal stability challenges. The corresponding waveform requirements, implemented to avoid the eclipsing of remote small targets by large near-by objects, include ever decreasing peak and integrated time-sidelobe levels. In the presence of very large competing discrete objects the time sidelobes required are so low, that they may only be implementable across narrow notches that must be strategically located relative to the peak response.

Higher spectral occupancy Military RF bands are shrinking due to the ever growing demand for broader bandwidth imposed by commercial systems. Some commercial bands due, e.g., to GPS and Wi-Fi, either extend into military bands, or are, otherwise, located in their immediate proximity. Increasing numbers of military RF sensors share the available spectrum and pose mutual interference threats at the same time that growing detection sensitivity requirements make such military sensors more vulnerable to RF interference. As a result, radar and communications systems are required to operate with greater RF spectral efficiency. To achieve that, waveforms are required that feature lower spectral sideband and, sometimes, spectral notching within the instantaneous band (i.e., the so-called, signal in-band), and/or in the sidebands. Because the spectral environment is dynamically changing, the numbers, widths, and depths of such spectral notches may need to be varied in real-time, or in near-real-time.

Multiple-Input-Multiple-Output (MIMO) waveforms The MIMO technology is growing and is becoming more popular both in radar and in communications. The application of this technology to radar requires antenna systems to radiate different waveforms either on element-to-element, or on sub-array to sub-array basis. Transition to MIMO technology involves a complex cost/benefit trade-off. Part of the perceived benefits includes the ability to trace each component of the received signal to the specific element, or sub-array that generated it. This, in turn allows the receiver to form digitally both transmit and receive beams, making the radar doubly digital. Additional advantages include the increasing ability to trade time vs. energy in spatial sectors of interest, and increased angle measurement accuracy due to the two-way coherent path available. To provide appropriate signal separation, however, the various waveforms need to be orthogonal. In some applications pseudo-orthogonalities of carefully controlled extents are needed. These, in part, are used to limit voltage standing wave ratios, and, in part, to control the angular width of the transmit sector.

The means whereby it becomes possible to meet the above waveform requirements are discussed in the following section. Section 6 presents pertinent waveform examples.

5 New Emerging Technologies That Help Meet New, More Stringent, Requirements

New composite requirements from radar waveforms include detailed temporal and spectral behaviors, subject to stringent loss and peak broadening. Sometimes they also include Doppler sensitivity constraints and, in some applications, such requirements need to be met separately for each member of a large orthogonal, or pseudo-orthogonal, group of codes. Because of this multiplicity of requirements and, because individual requirements tend to be more challenging than could be met by previously specified, current known codes, such as those described in Section 3 above, classical codes are often no longer suitable. Fortunately, new technologies have emerged which make it possible to generate new, custom made, codes of much higher performance levels than previously available, and to apply such codes to practical radar systems.

Very significant capabilities in custom code/filter generation have become available with the advent of practical and powerful optimization techniques and their custom application to the radar waveform synthesis discipline [6, 7]. These techniques make it possible to generate large numbers of compliant phase codes in relatively short computation time using common desktop personal computers. Furthermore, these phase codes are continuous, rather than quantized, as codes traditionally were (i.e., bi-phase, quadri-phase or, general polyphase). New hardware advances—which include powerful, low-cost, multi-bit, floating point, digital computing; deep, fast, and low-cost, memories; and arbitrary waveform generators make both the code synthesis and their applications to common radar system both feasible and practical. Codes can now be computed on-the-fly, or they can be computed off-line and stored in memory. In that second case, they can be stored in very large numbers and recovered and implemented in real time.

The fast and flexible code/filter synthesis is essential since exhaustive search is not possible for anything but short codes of relatively low level of quantization. For example, an octal code of length 100 is one of something on the order of 10^{90} possible combinations. A computer system capable of checking a billion combinations per second would have checked todate less than 10^{27} combinations if it had started when the universe is purported to have been created.

Techniques such as constrained optimization can be used to search for and find acceptable codes. In doing so, the algorithms do not search for the best overall code, but for codes that meet or exceed stated sets of requirements. The multiple dimension space, in which each code or code/filter combination is a point, is dense with good codes.

There normally exist extremely large numbers of codes that meet any reasonable combination of requirements, and they are distributed across the entire multidimensional space. All the same, these potential solutions, referred as ‘Local Minima’ of the optimization cost function, represent an extremely small fraction of the entire space. The optimization techniques seek the best solutions in the neighborhoods of a starting potential solution : hence the name Local Optimization. Similar optimization algorithms have been shown to be very effective in antenna pattern synthesis applications as well.

Each variant of a constrained optimization technique can be formulated as follows:

Minimize f(x), subject to $g_i(x) \leq k_i$ for i=1..L. Where, f(x) is a so-called cost, or objective function, and $g_i(x)$ is the i'th component of an L-dimensions constraints vector. Both the cost function and the constraints vector need to be formulated separately for each optimization problem. The cost function can include, for example, the integrated sidelobe levels and various spectral constraints in the sidebands and in-band. The constraints vector can include the individual sidelobe levels and the loss associated with the filtered code.

Problems which are couched in this framework can be solved using the methods found in many references, for example [8].

To illustrate how a code/filter pair with minimum Integrated Sidelobes (ISL) and with filter losses less than a given constant can be found, we use the following procedure: Let the n chips of a n element constant amplitude pulse compression waveform be represented by c_i for i= 1..n. Also let the m chips of the mismatched filter f be represented by f_i for i=1..m. In general f_i does not have constant amplitude. Next, let $\mathrm{R_i}$(c,f) for i=1..(n+m-1) be the cross correlation of the normalized code c (the normalization is explicitly part of the function) with the filter f, having its correlation peak in the first position then

ISL can be expressed via:

$$ISL(c,f) = \sum_{i \neq 1} \left(\frac{R_i}{R_1} \right)^2 . \tag{11}$$

Similarly, the mismatched filter loss can be expressed as:

$$Loss(c,f) = \frac{\sum\limits_{i=1}^{n} (c_i c_i^*) \sum\limits_{i=1}^{m} (f_i f_i^*)}{R_i^2} . \tag{12}$$

Where c_i and f_i are the respective code and filter components. In the case of simple constraints configuration, the procedure could consist of minimizing the $ISL(c,f)$ subject to the $Loss(c,f) \leq L$, $R_1(c,f) = 1$

6 Examples of Optimal Waveforms Based on Different Sets of Requirements

This section presents specific examples of code/filter pairs that were synthesized via the above optimization techniques.

6.1 Longest Known Barker-Level Code

Figure 12 shows the compressed matched filter response of the longest known Barker level polyphase code [9]. The code is 77 CHIPs long, which represents 37.7 dB of peak sidelobe level beneath the peak response. This was achieved via local optimization technique without additional spectral constraints. It has been subsequently shown elsewhere, by the same author (Carroll Nunn) that significant spectral performance constraints can be met using the local optimization technique, when starting with the code in Fig. 12, which result in minor compromises to the peak and integrated sidelobes. This ability to preserve the good temporal properties of the code in the presence of significant spectral constraints is generally achieved via the additional degrees of freedom obtained via the application of appropriate mismatched filter.

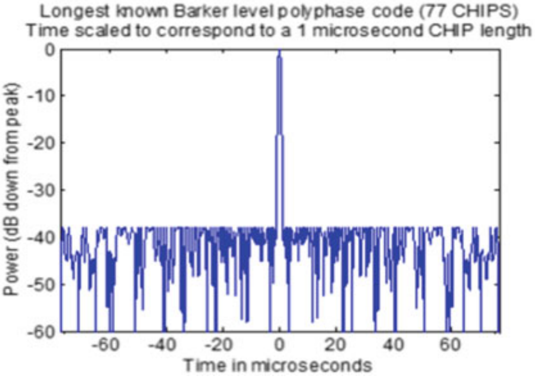


Fig. 12 Autocorrelation function of longest known Barker level code no spectral constraints

6.2 Computation Efficiency of Optimization Algorithm

Figure 13 [7] whose coordinate axes represent peak and integrated sidelobe performance contains three spots or regions in Integrated/Peak Sidelobes (ISL/PSL) space in red, green, and blue. Each region contains 25,000 points, each of which represents a distinct code, or a code/mismatched filter pair, found via the local optimization technique. All codes are 32 CHIPs long. All mismatched filters are 64 CHIPs long. The red region contains the best 25,000 codes optimized for matched filter ISL. The green represents the ISL optimized mismatched filter performance for the codes in the red spot. The blue spot represents the best code/mismatched filter pair performances achieved when the code and filter are optimized together. About a million codes or code/mismatched filter pairs were generated per optimization type. This scatter diagram shows both the performance capability possessed by optimization algorithms, and their processing speed which allows so many independent codes to be computed in reasonable time.

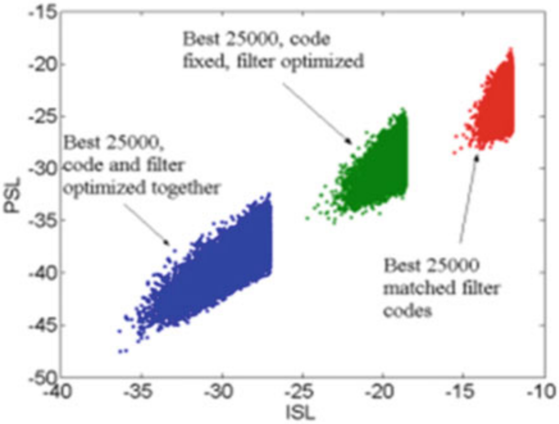


Fig. 13 Scatter diagram for codes demonstrating ability to find large numbers of different codes

6.3 Temporal Waveform Notching

Figure 14 [10] shows a compressed waveform in which a 15 dB time sidelobe notch has been implemented over a desired interval. Given the available number of degrees of freedom in the optimization process, it is generally possible to implement narrow notches that are far deeper than the general time sidelobe levels. The 15 dB notch makes it possible to detect smaller targets located within the notch of a much larger target in its immediate vicinity.

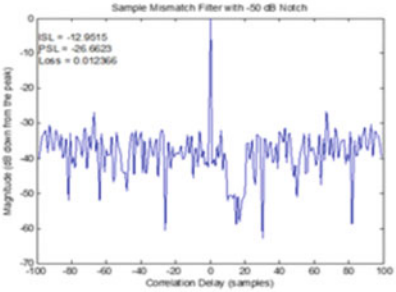


Fig. 14 Compressed code with a 15 dB notch in a selected interval of the time sidelobes

Figure 15 shows the compressed pulse response of a target next to which are three targets that are around 35 dB smaller. Clearly the targets are completely buried in the time sidelobes that are around 35 dB RMS beneath the peak response themselves. When the incoming pulse echo is split into multiple processing paths, processed via a mismatched filter bank, creating time sidelobe notches, such as the one if Fig. 14, that are spread across the entire time sidelobe extent, and when the bottom of all notches are quilted together to generate a much lower synthesized time sidelobe response, the three targets are exposed as in Fig. 16.

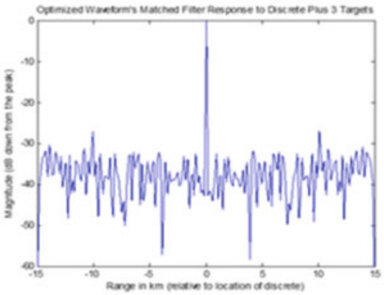


Fig. 15 Pulse compressed via matched filter

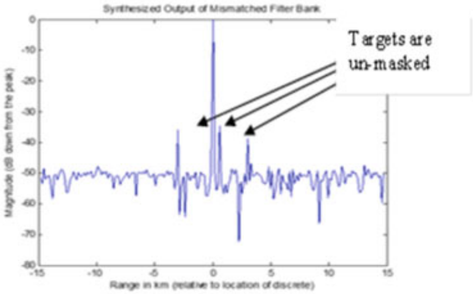


Fig. 16 Pulse compressed via a bank of mismatched filters, resulting in 15 dB lower time-sidelobes

6.4 Spectral Compliance Examples

Here we discuss two examples. The first shows a uniform sideband control. The second shows in-band and sideband spectral notching.

6.4.1 Control of General Sideband Levels

As was mentioned in this paper, ever increasing spectral occupancy forces separate RF systems to operate in the immediate spectral neighborhoods of each other. This imposes much more strenuous spectral requirements on spectral sidebands. Figure 17 shows the spectral response of a code without sideband control (blue) superimposed on an, identical bandwidth, optimized code with careful sideband control (green). It is seen that, in the case shown, the sidebands are brought nearly instantly to a level 50 dB beneath peak, and are maintained there.

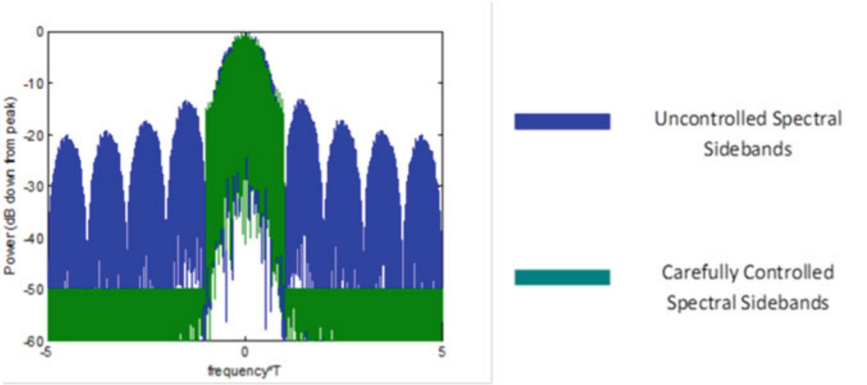
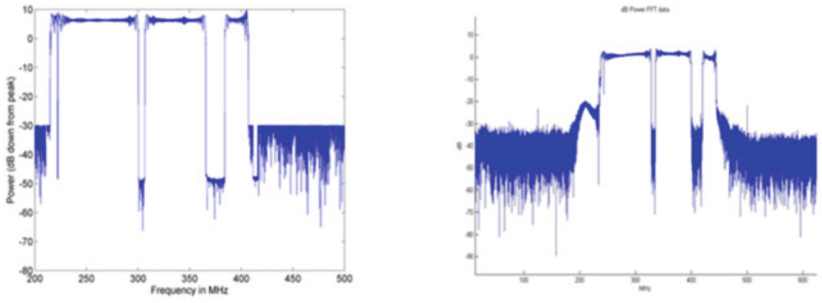


Fig. 17 Respective spectra of codes with controlled and uncontrolled sideband levels

6.4.2 In-Band and Sideband Spectral Notching

Figure 18a and b [11] show a broad, high range resolution, pulse spectrum which required notching in specific in-band and sideband places. Notch locations, depths, and widths were input parameters to the optimization algorithm. Figure 18a shows the ideal spectrum as produced by a desktop PC. It represents the spectrum generated by the arbitrary waveform generator as it is fed into the transmitter. Figure 18b shows the spectrum of the measured transmitter output. It is corrupted to some extent by the transmitter's noise, which generally causes the spectral notches and sideband levels to become less deep. This example shows the performance of an existing transmitter in which no changes whatever had been made to accommodate the input spectrum requirements. A higher quality transmitter, or a transmitter with adaptive equalization loop, could improve the results significantly. It had also been shown that, through the application of a properly optimized mismatched filter, the range performance of the compressed pulses had largely been preserved in the face of this very exacting spectral behavior.



a. Spectrum at waveform generator b. Spectrum measured at HPA

Fig. 18 Notched main and sideband spectrum **a** ideal, and **b** as measured after passage through unmodified physical transmitter

6.5 *Orthogonal Code Families*

Figure 19 shows auto and cross-correlation functions for a family of four orthogonal codes. Note the excellent time sidelobe behavior of both the auto correlations (diagonal figures) and cross correlation functions. Each of the waveforms in questions was several thousand CHIPs long.

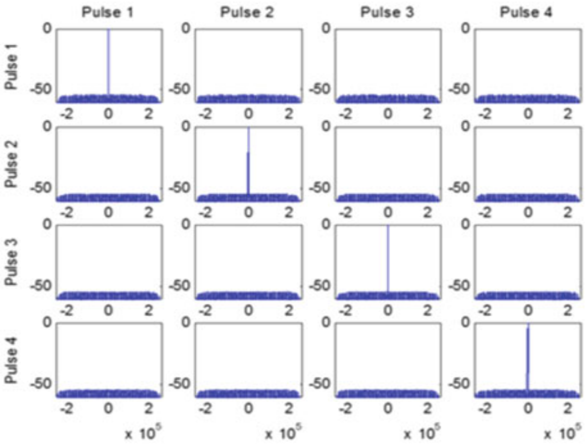


Fig. 19 Matched filter properties of four codes optimized to have −55 dB autocorrelation and cross-correlation sidelobes

6.6 Application of Optimization Techniques to Antenna Pattern Synthesis

Figure 20 [12] shows the pattern of a 15,000 element antenna array in which a 5° horizon notch has been incorporated via optimization. The horizon notch extends 5° in elevation and 180° in azimuth, covering the entire forward hemisphere of the array. The mainbeam is elevated to approximately 5° about the horizon. Using the local optimization technique, such notches have been computed very quickly and efficiently for similarly large arrays in the presence of roll, pitch, and yaw, which rendered the array in asymmetrical positions relative to the Earth's horizon. The notches are useful in elevated mainbeam positions as they help to further attenuate terrain clutter and terrain-based RF interference sources.

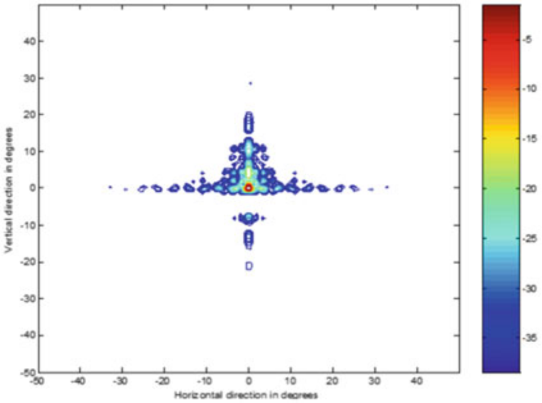


Fig. 20 A five degree notch starting 1.85 3 dB beam widths below the main beam. This notch requires 0.82 dB of insertion loss

6.7 Waveforms and Antenna Patterns Combined: Control of MIMO Radiating Sector Via Code Families of Appropriate Degrees of Quasi-Correlation

Figure 21 shows a typical bi-static MIMO based array architecture. The transmit array contains N elements, each of which transmits a different waveform. When the waveforms are orthogonal the radiation pattern covers the entire forward hemisphere in a manner that is controlled by the embedded element pattern designed for this array. The received array, which contains M elements, is shown on the right. (M is not necessarily equal to N.) The processing of the leftmost element is also shown. Though not explicitly shown, the same processing occurs behind each element. It is seen that the signal is split N ways. Each of these N paths is pulse compressed (i.e., filtered) based on a specific member of the N transmitted waveforms. In so doing, each path extracts the signal emanating from one particular transmit elements. The outputs from the N channels is fed into a beamformer which form the transmit beam relative to the part of space being covered. The outputs of these transmit beamformers are then combined across the M to receive elements to form the receive beam.

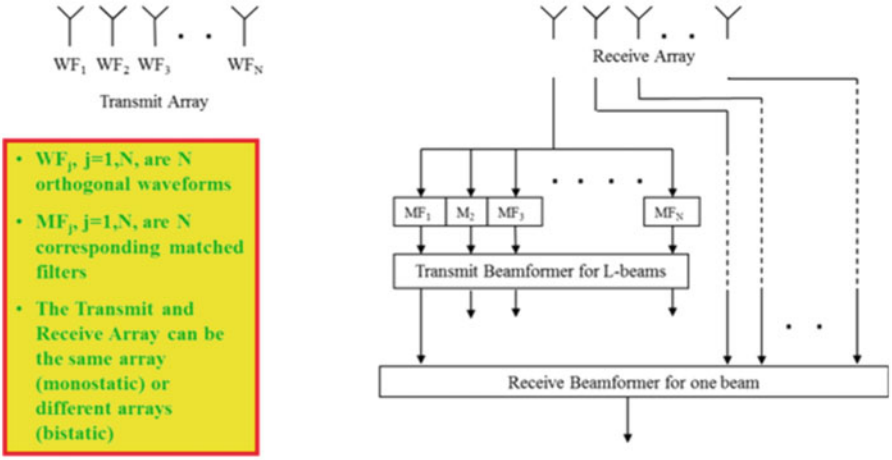

WF_1 WF_2 WF_3 WF_N
Transmit Array
• WF_j, j=1,N, are N orthogonal waveforms
• MF_j, j=1,N, are N corresponding matched filters
• The Transmit and Receive Array can be the same array (monostatic) or different arrays (bistatic)
Receive Array
MF_1
M_2
MF_3
MF_N
Transmit Beamformer for L-beams
Receive Beamformer for one beam

Fig. 21 A typical MIMO scheme

It is seen that this MIMO scheme is doubly digital in that both transmit and receive beams are formed digitally upon receiving, which provides the radar an extraordinary degree of adaptive control. The transmit pattern is governed by the selection of the transmit waveforms, which demonstrates that in this MIMO application both the waveform and array applications become intertwined. By careful selection of the transmit waveforms, transmit patterns can be formed of desired directivities anywhere between the maximum full array directivity (represented by fully correlated waveforms), and minimum, single element, directivity (represented by totally uncorrelated waveforms).

References

1. Nunn CJ, Kretschmer FF. Performance of pulse compression code and filter pairs optimized for loss and integrated sidelobe level. The Record of the IEEE 2007 International Radar Conference, IEEE Aerospace and Electronic Systems Society; 2007.
2. Lewis BL, Kretschmer FF. Aspects of radar signal processing. Norwood: Artech House Inc.; 1966.
3. Felhauer T. Design and analysis of New P(n,k) polyphase pulse compression codes. IEEE Trans Aerosp Electron Syst. 1994;30(3):865–74.
4. Cook CE, Bernfeld M. Radar signals. Norwood: Academic Press Inc.; 1967.
5. Nathanson FE. Radar design principles, 2nd ed. Raleigh: Scitech Publishing Inc.; 1999.

6. Nunn CJ, Welch LR. Multi-parameter local optimization for the design of superior matched filter polyphase pulse compression codes. The record of the IEEE 2000 international radar conference, IEEE Aerospace and Electronic Systems Society; 2000.
7. Nunn CJ. Constrained optimization applied to pulse compression codes, and filters. The record of the IEEE 2005 international radar conference, IEEE Aerospace and Electronic Systems Society; 2005.
8. Leuenberger DG. Linear and nonLinear programming, 2nd ed. Reading: Addison-Wesley; 1984.
9. Nunn CJ, Coxson GE. Polyphase pulse compression codes with optimal peak and integrated side-lobes. IEEE Trans Aerosp Electron Syst. 2009;45(2):775–81.
10. Nunn CJ. Advanced waveforms for spaceborne SAR and GMTI radars. Technical Seminar, Technology Service Corporation; 2011.
11. Nunn CJ, Moyer L. Spectrally-compliant waveforms for wideband radar. IEEE Aerosp Syst Mag. 2012;27(8):11–5.
12. Nunn CJ, Levitas M, Muldoon S, Prenatt M. Real time compatible, phase only pattern notching algorithm for very large arrays. The proceedings of the 2006 Tri-Service Radar Symposium; 2006.

Adventures in Compressive Sensing Based MIMO Radar

Thomas Strohmer and Haichao Wang

Abstract While radar has been around for many decades, novel developments in recent years have led to significant breakthroughs as well as to exciting new mathematical challenges. In this chapter, we consider a multiple-input-multiple-output (MIMO) radar system. Using sparsity as a key ingredient of our approach and tools from compressive sensing, we derive a mathematical framework for the imaging of targets in the azimuth-range-Doppler domain. Our analysis comprises uniformly spaced linear arrays with random waveforms, as well as random sensor arrays with deterministic waveforms. We also derive results that do not require the "on-the-grid" assumption often used in compressive sensing radar. Algorithmic aspects and numerical simulations are presented as well.

Keywords Sparsity · Radar · Compressive sensing · MIMO · Random sensor arrays · Convex optimization · Random matrices · Kerdock code

1 Introduction

In recent years, radar systems employing multiple antennas at the transmitter and the receiver (also referred to as MIMO radar, where MIMO stands for multiple-input multiple-output) have attracted enormous attention in the engineering and signal processing community [11, 23, 24]. MIMO radar is supposed to offer a range of benefits over classical radar. Yet, to be able to rake in those benefits, MIMO radar should be combined with the recently developed framework of compressive sensing [15, 28, 31]. Taking advantage of the sparsity of a radar scene via methods based on

T. Strohmer (✉) · H. Wang
University of California, Davis, 95616 CA, USA
e-mail: strohmer@math.ucdavis.edu

H. Wang
e-mail: hchwang@ucdavis.edu

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3_13

compressive sensing can improve the performance of radar systems under certain conditions and is therefore of considerable practical interest. This article contains a partial overview of the results found in [31] and [33], as well as some new results for compressive SIMO radar "off-the-grid."

Let us first introduce the basic radar setup. A radar system illuminates a region of interest in order to detect the location, velocity, and reflectivity of the objects (targets) in its field of view. We consider the following standard (narrowband) radar model [30]. Suppose a target located at *range* r is traveling with *constant velocity* v and has *reflection coefficient* a. Suppose further just for the moment that we have only one target, one transmitter and one receiver (in which case we cannot detect direction). After transmitting signal $s(t)$, the receiver observes the reflected signal

$$y(t) \;=\; as(t-\tau_r)e^{2\pi i\omega_v t} \tag{1}$$

where $\tau_r = 2r/c$ is the round trip time of flight, c is the speed of light, $\omega_v \approx -2\omega_0 v/c$ is the Doppler shift, and ω_0 is the carrier frequency. The basic idea is that the *range-velocity* information (r,v) of the target can be inferred from the observed *time delay-Doppler shift* (τ_r,ω_v) of s in (1). For only one target this can be done conveniently by correlating the received signal y with time-frequency shifted versions of the transmitted signal. Since, we are dealing with bandlimited signals, it suffices to consider discrete signals sampled at a properly chosen rate Δ_t. It is therefore common practice to compute

$$V(\tau,\omega) := \sum_l y(l\Delta_t)\overline{\mathbf{s}(l\Delta_t-\tau)e^{2\pi i\omega l}} \tag{2}$$

and then locate the largest value of $|V(\tau,\omega)|$ in order to detect the target in the range-Doppler domain.

In the presence of multiple targets more sophisticated methods are necessary. In order to resolve azimuth in addition to range and Doppler, we need to employ an array of antennas. We assume an array of N_T transmit and N_R receiver antennas that are colocated (also known as mono-static radar) as illustrated in Fig. 1. A more detailed description of the setup is postponed to Section 2 The transmit antennas send simultaneously probing signals, which can differ from antenna to antenna and can be chosen to our specifications. It is convenient to divide the region of interest into range-azimuth-Doppler cells corresponding to distance, direction, and velocity, respectively. Let $\mathbf{A}$ be a measurement matrix whose columns correspond to the signal recorded at each receive antenna from a single unit-strength scatterer at a specific range-azimuth-Doppler cell. Let $\mathbf{x}$ denote a vector whose elements represent the complex amplitudes of the scatterers. In many cases the radar scene is sparse in the sense that only a small fraction (often a *very* small fraction) of the cells is occupied by the objects of interest. In this case most of the entries of $\mathbf{x}$ will be zero, but we do not know which ones, otherwise we would have located the targets already. With $\mathbf{w}$ representing a noise vector, we are faced with the linear system of equations

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}, \tag{3}$$

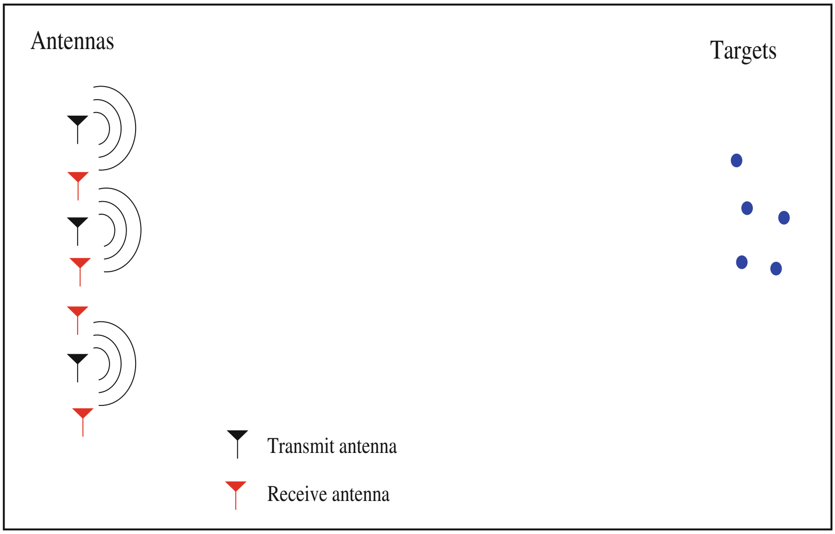


Fig. 1 Schematic illustration of the MIMO random array setup: The distance between antennas and targets is assumed to be large compared to the aperture (i.e., this is the "far-field" scenario). The transmit antennas as well as the receive antennas are colocated

where $\mathbf{y}$ is a vector of measurements collected by the receive antennas over an observation interval. Typically this system will be *underdetermined*, which implies that it will have infinitely many solutions. What comes to our rescue here is the sparsity of $\mathbf{x}$. While conventional radar processing techniques do not take full advantage of sparsity of the radar scene, the recent development of compressive sensing provides us with the possibility to optimally utilize this property [15, 28, 31]. The approach pursued in this chapter to obtain a sparse solution of (3) is based on the lasso [36], which gained tremendous popularity in connection with compressive sensing. The lasso solves

$$\min_{\mathbf{x}} \quad \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2+\lambda\|\mathbf{x}\|_1, \tag{4}$$

where the parameter $\lambda > 0$ trades off goodness of fit with sparsity.

2 Problem Setup

We consider a baseband version of a MIMO radar employing N_T antennas at the transmitter and N_R antennas at the receiver. We assume for convenience that transmitters and receivers are colocated, cf. Fig. 1. Furthermore, we assume a coherent propagation scenario, i.e., the element spacing is sufficiently small so that the radar return from a given scatterer is fully correlated across the array. The arrays and all the scatterers are assumed to be in the same 2-D plane. The extension to the 3-D case is straightforward.

The array manifolds $\mathbf{a}_T(\beta)$, $\mathbf{a}_R(\beta)$ are given by

$$\mathbf{a}_T(\beta) = \left[e^{2\pi i p_1 \beta}, e^{2\pi i p_2 \beta}, \dots, e^{2\pi i p_{N_T} \beta}\right]^T, \tag{5}$$

and

$$\mathbf{a}_R(\beta) = \left[e^{2\pi i q_1 \beta}, e^{2\pi i q_2 \beta}, \dots, e^{2\pi i q_{N_R} \beta}\right]^T, \tag{6}$$

where the p_j's and q_j's are normalized antenna spacings (distance divided by wavelength).

The jth transmit antenna repeatedly transmits the signal $s_j(t)$, which is assumed to be a periodic, continuous-time signal of period-duration T seconds and bandwidth B. We observe the back-scattered signal over a duration T, and since its bandwidth is B, at each receive antennas we sample the observed signal at a rate of $1/\Delta_s$ where $\Delta_s \le \frac{1}{2B}$. For simplicity, we choose $\Delta_s = \frac{1}{2B}$ resulting in $N_s := 2TB$ many samples per period duration[1]. It is convenient to introduce the finite-length vector $\mathbf{s}_j$ associated with s_j, via $\mathbf{s}_j(l) := s_j(l\Delta_s), l = 1, \dots, N_s$.

Let $\mathbf{Z}(t;\beta,\tau,f)$ be the $N_R \times N_s$ noise-free received signal matrix from a unit strength target at direction β, delay τ, and Doppler f (corresponding to its radial velocity with respect to the radar). Then

$$\mathbf{Z}(t;\beta,\tau,f) = \mathbf{a}_R(\beta)\mathbf{a}_T^T(\beta)\mathbf{S}_{\tau,f}^T,$$

where $\mathbf{S}_{\tau,f}$ is a $N_s \times N_T$ matrix whose columns are the circularly delayed and Doppler shifted signals $s_j(t-\tau)e^{2\pi i f t}$.

We let $\mathbf{z}(t;\beta,\tau,f) = \mathrm{vec}\{\mathbf{Z}\}(t;\beta,\tau,f)$ be the noise-free vectorized received signal. We set up a discrete azimuth-range-Doppler grid $\{\beta_l, \tau_j, f_k\}$ for $1 \le l \le N_\beta$, $1 \le j \le N_\tau$, and $1 \le k \le N_f$, where $\Delta_\beta, \Delta_\tau$, and Δ_f denote the corresponding discretization stepsizes. Using vectors $\mathbf{z}(t;\beta_l,\tau_j,f_k)$ for all grid points (β_l,τ_j,f_k) we construct a complete response matrix $\mathbf{A}$ whose columns are $\mathbf{z}(t;\beta_l,\tau_j,f_k)$ for $1 \le l \le N_\beta$ and $1 \le j \le N_\tau$, $1 \le k \le N_f$. In other words, $\mathbf{A}$ is a $N_R N_s \times N_\tau N_\beta N_f$ matrix with columns

$$\mathbf{A}_{\beta,\tau,f} = \mathbf{a}_R(\beta) \otimes \mathbf{S}_{\tau,f}\mathbf{a}_T(\beta). \tag{7}$$

Assume that the radar illuminates a scene consisting of S scatterers located on S points of the (β_l,τ_j,f_k)-grid. Let $\mathbf{x}$ be a sparse vector whose non-zero elements are the complex amplitudes of the scatterers in the scene. The zero elements correspond to grid points which are not occupied by scatterers. We can then define the radar signal $\mathbf{y}$ received from this scene by

$$\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w} \tag{8}$$

[1] Actually the received signal will have a somewhat larger bandwidth $B_1 > B$ due to the Doppler effect. Our results could be easily modified to incorporate this increased bandwidth. Since in practice this increase in bandwidth is small, for convenience we simply assume $B \approx B_1$.

where $\mathbf{y}$ is an $N_R N_s \times 1$ vector, $\mathbf{x}$ is an $N_\tau N_\beta N_f \times 1$ sparse vector, and $\mathbf{w}$ is an $N_R N_s \times 1$ complex Gaussian noise vector. Our goal is to solve for $\mathbf{x}$, i.e., to locate the scatterers (and their reflection coefficients) in the azimuth-delay-Doppler domain.

Remark: The assumption that the targets lie on the grid points, while common in compressive sensing, is certainly quite restrictive. A violation of this assumption will result in a model mismatch, sometimes dubbed *gridding error*, which can potentially be quite severe [8, 16]. Recently some interesting strategies have been proposed to overcome this gridding error [9, 34]. But these methods, at least in their current form, are not directly applicable to our setting. We will answer this question partially in chapter "Recursive Computation of Spherical Harmonic Rotation Coefficients of Large Degree."

As mentioned in the introduction, a standard approach to solve (8) when x is sparse, is

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{A}\mathbf{x}-\mathbf{y}\|_2^2 + \lambda\|\mathbf{x}\|_1, \tag{9}$$

which is also known as lasso [36]. Instead of (9), we will use the *debiased lasso*. That means first we compute an approximation $\tilde{I}$ for the support of $\mathbf{x}$ by solving (9). This is the detection step. Then, in the estimation step, we "debias" the solution by computing the amplitudes of $\mathbf{x}$ via solving the reduced-size least squares problem $\min\|\mathbf{A}_{\tilde{I}}\mathbf{x}_{\tilde{I}} - \mathbf{y}\|_2$, where $\mathbf{A}_{\tilde{I}}$ is the submatrix of $\mathbf{A}$ consisting of the columns corresponding to the index set $\tilde{I}$, and similarly for $\mathbf{x}_{\tilde{I}}$.

We assume that the locations of the targets are random. To be precise, we assume that the S nonzero coefficients of x are selected uniformly at random and the phases of the non-zero entries of x are random and uniformly distributed in $[0, 2\pi)$. We will refer to this model as the generic S-sparse model. We note that by adopting the proof techniques developed in [18], it seems very plausible that one can drop the assumption that the target locations are randomly chosen.

3 Random Waveforms and Deterministic Antenna Arrays

In this section, we assume that the antennas are uniformly spaced in each array, i.e., $p_j = d_T(j-1)$ in (5) and $q_j = d_R(j-1)$ in (6). It is known that the spatial characteristics of a MIMO radar are closely related to that of a virtual array with $N_T N_R$ antennas, whose array manifold is $\mathbf{a}(\beta) = \mathbf{a}_T(\beta) \otimes \mathbf{a}_R(\beta)$. It is known [12] that the following choices for the spacing of the transmit and receive array spacing will yield a uniformly spaced virtual array with half wavelength spacing:

$$d_R = 0.5, d_T = 0.5N_R; \tag{10}$$

$$d_T = 0.5, d_R = 0.5N_T. \tag{11}$$

Both of these choices lead to a virtual array whose aperture is $0.5(N_T N_R - 1)$ wavelengths. This is the largest virtual aperture free of grating lobes. The choices (10) and (11) will also show up again in our theoretical analysis, see Theorem 1.

We assume that $s_i(t)$ is a periodic, continuous-time white Gaussian noise signal of period duration T seconds and bandwidth B. The transmit waveforms are normalized so that the total transmit power is fixed, independent of the number of transmit antennas. Thus, we assume that the entries of $s_i(t)$ have variance $\frac{1}{N_T}$.

Theorem 1. *Consider* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, *where* $\mathbf{A}$ *is as defined in* (7) *and* $\mathbf{w}_i \in \mathscr{CN}(0, \sigma^2)$. *Choose the discretization stepsizes to be* $\Delta_\beta = \frac{2}{N_R N_T}$, $\Delta_\tau = \frac{1}{2B}$ *and* $\Delta_f = \frac{1}{T}$. *Let* $d_T = 1/2, d_R = N_T/2$ *or* $d_T = N_R/2, d_R = 1/2$, *and suppose that*

$$N_t \geq 128 \qquad \textit{and} \qquad \big(\log(N_\tau N_\beta)\big)^3 \leq N_t, \tag{12}$$

If $\mathbf{x}$ *is drawn from the generic K-sparse scatterer model with*

$$K \leq \frac{c_0 N_\tau N_R}{6 \log(N_\tau N_f N_\beta)} \tag{13}$$

for some constant $c_0 > 0$, *and if*

$$\min_{k \in I} |\mathbf{x}_k| > \frac{10\sigma}{\sqrt{N_R N_t}} \sqrt{2 \log N_\tau N_f N_\beta}, \tag{14}$$

then the solution $\tilde{\mathbf{x}}$ *of the debiased lasso computed with* $\lambda = 2\sigma\sqrt{2\log(N_\tau N_f N_\beta)}$ *obeys*

$$\operatorname{supp}(\tilde{\mathbf{x}}) = \operatorname{supp}(\mathbf{x}), \tag{15}$$

with probability at least

$$(1-p_1)(1-p_2)(1-p_3)(1-p_4), \tag{16}$$

and

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\sigma\sqrt{12 N_t N_R}}{\|\mathbf{y}\|_2} \tag{17}$$

with probability at least

$$(1-p_1)(1-p_2)(1-p_3)(1-p_4)(1-p_5), \tag{18}$$

where

$$p_1 = e^{-\frac{(1-\sqrt{1/3})^2 N_t}{2}} + N_T e^{-(\sqrt{3/2}-\sqrt{2})N_t},$$

$$\begin{aligned} p_2 = {} & 2(N_R N_T)^{-1} + 2(N_\tau N_R N_T)^{-1} + 2(N_f N_R N_T)^{-1} \\ & + 6(N_\tau N_f N_R N_T)^{-1} + 2e^{-\frac{N_t(\sqrt{2}-1)^2}{4}}, \end{aligned}$$

$$p_3 = N_R N_T e^{-\frac{(1-\sqrt{1/3})^2 N_t}{2}}, \qquad p_4 = e^{-\frac{N_R N_t}{25}},$$

and

$$p_5 = 2(N_\tau N_\beta)^{-1}(2\pi \log(N_\tau N_\beta) + S(N_\tau N_\beta)^{-1}) + \mathscr{O}((N_\tau N_\beta)^{-2\log 2}).$$

The proof of the above theorem requires several steps. We need two key estimates, one concerns a bound for the operator norm of $\mathbf{A}$, the other one concerns a bound for the coherence of $\mathbf{A}$. We start with deriving a bound for $\|\mathbf{A}\|_{\mathrm{op}}$.

Lemma 1. *Let* $\mathbf{A}$ *be as defined in Theorem 1. Then*

$$\mathbb{P}\Big(\|\mathbf{A}\|_{op}^2 \le 2N_t N_f N_R N_T\Big) \ge 1 - N_T e^{-N_t(\frac{3}{2}-\sqrt{2})}. \tag{19}$$

Proof. There holds $\|\mathbf{A}\|_{\mathrm{op}}^2 = \|\mathbf{A}\mathbf{A}^*\|_{\mathrm{op}}$. It is convenient to consider $\mathbf{A}\mathbf{A}^*$ as block matrix

$$\begin{bmatrix} \mathbf{B}_{1,1} & \mathbf{B}_{1,2} \dots & \mathbf{B}_{1,N_R} \\ \vdots & \ddots & \vdots \\ \mathbf{B}_{N_R,1}^* & & \mathbf{B}_{N_R,N_R} \end{bmatrix},$$

where the blocks $\{\mathbf{B}_{i,i'}\}_{i,i'=1}^{N_R}$ are matrices of size $N_t \times N_t$. We claim that $\mathbf{A}\mathbf{A}^*$ is a block-Toeplitz matrix (i.e., $\mathbf{B}_{i,i'} = \mathbf{B}_{i+1,i'+1}, i = 1,\dots,N_R-1$) and the individual blocks $\mathbf{B}_{i,i'}$ are circulant matrices. To see this, recall the structure of $\mathbf{A}$ and consider the entry $\mathbf{B}_{[i,l;i',l']}$, $i,i' = 1,\dots,N_R; l,l' = 1,\dots,N_t$:

$$\mathbf{B}_{[i,l;i',l']} = (\mathbf{A}\mathbf{A}^*)_{[i,l;i',l']} = \sum_\beta \sum_\tau \sum_f \mathbf{A}_{[i,l;\tau,f,\beta]} \mathbf{A}_{[i',l';\tau,f,\beta]} \tag{20}$$

$$= \sum_\beta e^{j2\pi d_R(i-i')\beta} \sum_{k=1}^{N_T} \sum_{k'=1}^{N_T} e^{j2\pi d_T(k-k')\beta} G_{k,k'}(l,l') \sum_{m=1}^{N_f} e^{j2\pi(l-l')\Delta_t m\Delta_f} \tag{21}$$

$$= \sum_{n=0}^{N_R N_T -1} e^{j2\pi(i-i')\frac{nN_T}{N_R N_T}} \sum_{k=1}^{N_T} \sum_{k'=1}^{N_T} e^{j2\pi(k-k')\frac{n}{N_R N_T}} G_{k,k'}(l,l') N_f \delta_{l-l'} \tag{22}$$

$$= N_T N_R N_f \sum_{k=1}^{N_T} \|\mathbf{s}_k\|^2 \delta_{i-i'} \delta_{l-l'} \tag{23}$$

where we have used in (22) that $N_f = \frac{2B}{\Delta_f} = 2BT$. Noting that

$$\sum_{m=1}^{N_f} e^{j2\pi(l-l')m\Delta_t\Delta_f} = N_f \delta_{l-l'},$$

we obtain

$$\mathbf{A}\mathbf{A}^* = (N_T N_R N_f \sum_{k=1}^{N_T} \|\mathbf{s}_k\|^2)\, \mathbf{I}, \tag{24}$$

i.e., $\mathbf{A}\mathbf{A}^*$ is just a scaled identity matrix. Since $\mathbf{s}_k$ is a Gaussian random vector with $\mathbf{s}_k(j) \sim \mathscr{C}\mathscr{N}(0,1)$, Lemma 7 in the appendix yields

$$\mathbb{P}\Big(\|\mathbf{s}_k\|_2^2 - (\mathbb{E}\|\mathbf{s}_k\|_2)^2 \geq t(t + 2\mathbb{E}\|\mathbf{s}_k\|_2)\Big) \leq e^{-t^2/2}, \tag{25}$$

where we note that $\mathbb{E}\|\mathbf{s}_k\|_2 = \sqrt{\frac{N_t}{N_T}}$. We choose $t = (\sqrt{2}-1)\sqrt{N_t}$, and obtain, after forming the union bound over $k = 1,\dots,N_t - 1$,

$$\mathbb{P}\Big(\sum_{k=1}^{N_T} \|\mathbf{s}_k\|_2^2)^2 \geq 2N_t\Big) \leq N_T e^{-N_t(\frac{3}{2}-\sqrt{2})}. \tag{26}$$

The bound (19) now follows from (24).

Next we establish a coherence bound for $\mathbf{A}$.

Lemma 2. *Let* $\mathbf{A}$ *be as defined in Theorem* 1. *Assume that*

$$N_\tau N_f \geq \sqrt{N_\beta}, \quad \log(N_\tau N_f N_\beta) < \frac{N_t}{30}. \tag{27}$$

Then

$$\max_{(\tau,f,\beta)\neq(\tau',f',\beta')} \big|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'}\rangle\big| \leq 3N_R\sqrt{N_t \log(N_\tau N_f N_\beta)} \tag{28}$$

with probability at least $1-2(N_RN_T)^{-1}-2(N_\tau N_RN_T)^{-1}-2(N_fN_RN_T)^{-1}-6(N_\tau N_f N_R N_T)^{-1}$.

Proof. We have that

$$\mathbf{A}_{\tau,f,\beta} = \mathbf{a}_R(\beta) \otimes (\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)),$$

from which we readily compute

$$\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'}\rangle = \langle \mathbf{a}_R(\beta), \mathbf{a}_R(\beta')\rangle \langle \mathbf{S}_{\tau,f}\mathbf{a}_T(\beta), \mathbf{S}_{\tau',f'}\mathbf{a}_T(\beta')\rangle. \tag{29}$$

We use the discretization $\beta = n\Delta_\beta$, $\beta' = n'\Delta_\beta$, where $\Delta_\beta = \frac{2}{N_RN_T}$, $n,n' = 1,\dots,N_\beta$, with $N_\beta = N_RN_T$, and obtain after a standard calculation

$$\langle \mathbf{a}_R(\beta), \mathbf{a}_R(\beta')\rangle = \begin{cases} N_R & \text{if } n-n' = kN_R \text{ for } k = 0,\dots,N_T-1, \\ 0 & \text{if } n-n' \neq kN_R, \end{cases} \tag{30}$$

and

$$\langle \mathbf{a}_T(\beta), \mathbf{a}_T(\beta')\rangle = \begin{cases} 0 & \text{if } n-n' = kN_R \text{ for } k = 1,\dots,N_T-1, \\ \langle \mathbf{a}_T(\beta), \mathbf{a}_T(\beta)\rangle & \text{if } n-n' = 0. \end{cases} \tag{31}$$

As a consequence of (30), concerning β,β' we only need to focus on the case $n - n' = kN_R$ for $k = 1,\dots,N_T - 1$. Moreover, a standard calculation shows that

$$|\langle \mathbf{S}_{\tau,f}\mathbf{a}_T(\beta), \mathbf{S}_{\tau',f'}\mathbf{a}_T(\beta')\rangle| = |\langle \mathbf{S}_{\tau-\tau',f-f'}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta')\rangle| \tag{32}$$

for $\tau, \tau' = 0, \dots, N_\tau - 1, f, f' = 0, \dots, N_f - 1$, thus we only need to consider $|\langle \mathbf{S}_{\tau,f} \mathbf{a}_T(\beta), \mathbf{S}\mathbf{a}_T(\beta') \rangle|$. We distinguish several cases.

Case (a) $\beta \neq \beta', \tau = 0, f = 0$**:** Based on (30) and (31), to bound $|\langle \mathbf{a}_R(\beta), \mathbf{a}_R(\beta') \rangle \langle \mathbf{S} \mathbf{a}_T(\beta), \mathbf{S}\mathbf{a}_T(\beta') \rangle|$ we only need to consider those n, n' for which $n - n'$ is not a multiple of N_R, in which case $\mathbf{a}_T(\beta)$ and $\mathbf{a}_T(\beta')$ are orthogonal. We have

$$|\langle \mathbf{a}_R(\beta), \mathbf{a}_R(\beta') \rangle \langle \mathbf{S}\mathbf{a}_T(\beta), \mathbf{S}\mathbf{a}_T(\beta') \rangle| \leq N_R \, |\langle \mathbf{S}^* \mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta') \rangle|. \tag{33}$$

By Lemma 8 there holds

$$\mathbb{P}\Big(|\langle \mathbf{S}^* \mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta') \rangle| \geq t N_t \Big) \leq 2 \exp\Big(-N_t \frac{t^2}{C_1 + C_2 t}) \Big) \tag{34}$$

for all $0 < t < 1$, where $C_1 = \frac{4e}{\sqrt{6\pi}}$ and $C_2 = \sqrt{8}e$. We choose $t = 3\sqrt{\frac{1}{N_t} \log(N_\tau N_f N_R N_T)}$ in (34) and get

$$\mathbb{P}\Big(|\langle \mathbf{S}^* \mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta') \rangle| \geq 3\sqrt{N_t \log(N_\tau N_f N_R N_T)} \Big) \tag{35}$$

$$\leq 2 \exp\Big(- \frac{9 \log(N_\tau N_f N_R N_T)}{C_1 + \frac{3C_2}{\sqrt{N_t}} \log(N_\tau N_f N_R N_T)} \Big). \tag{36}$$

We claim that

$$\frac{9 \log(N_\tau N_f N_R N_T)}{C_1 + \frac{3C_2}{\sqrt{N_t}} \log(N_\tau N_f N_R N_T)} \geq 2 \log(N_R N_T). \tag{37}$$

To verify this we first note that (37) is equivalent to

$$9 \log N_\tau N_f \geq \log(N_R N_T) (2C_1 + \frac{6C_2}{\sqrt{N_t}} \sqrt{\log(N_\tau N_f N_\beta)} - 9).$$

Using both assumptions in (27) and the fact that $2C_1 + \frac{6C_2}{\sqrt{30}} - 9 \leq \frac{9}{2}$ we obtain

$$9 \log N_\tau N_f \geq \log N_\beta (2C_1 + \frac{6C_2}{\sqrt{30}} - 9) \geq \log N_\beta (2C_1 + \frac{6C_2}{\sqrt{N_t}} \sqrt{\log N_\tau N_f N_\beta} - 9),$$

which establishes (37). Substituting (37) into (35) gives

$$\mathbb{P}\Big(|\langle \mathbf{S}^* \mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta') \rangle| \geq 3\sqrt{N_t \log(N_\tau N_f N_R N_T)} \Big) \leq 2 \exp\big(-2 \log(N_R N_T) \big). \tag{38}$$

To bound $\max |\langle \mathbf{A}_{\tau,\beta}, \mathbf{A}_{\tau,\beta'} \rangle|$ we only have to take the union bound over $N_R N_T$ different possibilities associated with β, β', as $\tau = 0$ and $f = 0$. Forming now the union bound, and using (33), yields

$$\mathbb{P}\Big(|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau,f,\beta'} \rangle| \leq 3N_R \sqrt{N_t \log(N_\tau N_f N_R N_T)} \Big) \geq 1 - 2(N_R N_T)^{-1}. \tag{39}$$

Case (b) $\beta \neq \beta', \tau \neq 0, f = 0$: We need to consider $|\langle \mathbf{S}_\tau \mathbf{a}_T(\beta), \mathbf{S}\mathbf{a}_T(\beta')\rangle|$ where $\beta = n\Delta_\beta$, $\beta' = n'\Delta_\beta$, with $n - n' = kN_R$ for $k = 1, \dots, N_T - 1$. Since the entries of $\mathbf{S}$ are i.i.d. Gaussian random variables, it follows that the entries of $\mathbf{S}_\tau \mathbf{a}_T(\beta)$ are i.i.d. $\mathscr{CN}(0,1)$-distributed, and similar for $\mathbf{S}\mathbf{a}_T(\beta')$. Moreover, the fact that $\langle \mathbf{a}_T(\beta), \mathbf{a}_T(\beta')\rangle = 0$ implies that $\mathbf{S}_\tau \mathbf{a}_T(\beta)$ and $\mathbf{S}\mathbf{a}_T(\beta')$ are independent. Consequently, the entries of $\sum_{l=0}^{N_t-1} \overline{(\mathbf{S}_\tau \mathbf{a}_T(\beta))_l} (\mathbf{S}\mathbf{a}_T(\beta'))_l$ are jointly independent. Therefore, we can apply Lemma 10 with $t = 3\sqrt{N_t}\sqrt{\log(N_\tau N_R N_T)}$, form the union bound over the $N_\tau N_R N_T$ possibilities associated with τ (we do not take advantage of the fact we actually have only $N_\tau - 1$ and not N_τ possibilities for τ) and β, β' (here, we take again into account property (30)), and eventually obtain

$$\mathbb{P}\Big(|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f,\beta'}\rangle| \leq 3N_R\sqrt{N_t \log(N_\tau N_R N_T)}\Big) \geq 1 - 2(N_\tau N_R N_T)^{-1}. \tag{40}$$

Case (c) $\beta \neq \beta', \tau = 0, f \neq 0$: It is well known that $(\mathbf{T}_\tau \mathbf{x})^\wedge = \mathbf{M}_{-\tau}\hat{\mathbf{x}}$. Hence, by Parseval's theorem, $\langle \mathbf{T}_\tau \mathbf{x}, \mathbf{y}\rangle = \langle \mathbf{M}_{-\tau}\hat{\mathbf{x}}, \hat{\mathbf{y}}\rangle$. Since the normal distribution is invariant under Fourier transform, this case is therefore already covered by Case (b), and we leave the details to the reader. We get

$$\mathbb{P}\Big(|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau,f',\beta'}\rangle| \leq 3N_R\sqrt{N_t \log(N_f N_R N_T)}\Big) \geq 1 - 2(N_f N_R N_T)^{-1}. \tag{41}$$

Case (d) $\beta \neq \beta', \tau \neq 0, f \neq 0$: This is similar to Case (b). The only difference is that we have $N_t N_f N_R N_T$ different possibilities to consider when forming the union bound (the additional factor N_f is of course due to frequency shifts associated with the Doppler effect). Thus in this case the bound reads

$$\mathbb{P}\Big(|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'}\rangle| \leq 3N_R\sqrt{N_t \log(N_\tau N_f N_R N_T)}\Big) \geq 1 - 2(N_\tau N_f N_R N_T)^{-1}. \tag{42}$$

Case (e) $\beta = \beta'$: We need to bound $|\langle \mathbf{M}_f \mathbf{T}_\tau \mathbf{S}\mathbf{a}_T(\beta), \mathbf{S}\mathbf{a}_T(\beta)\rangle|$, where we recall that $\mathbf{S}\mathbf{a}_T(\beta)$ is a Gaussian random vector with variance N_T. (We note that a related case is covered by Theorem 5.1 in [27], which considers $\langle \mathbf{M}_f \mathbf{T}_\tau h, h\rangle$, where h is a Steinhaus sequence.) Since, each of the entries of $\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)$ and of $\mathbf{S}\mathbf{a}_T(\beta)$ is a sum of N_T i.i.d. Gaussian random variables of variance $1/N_T$, we can write

$$|\langle \mathbf{S}_{\tau,f}\mathbf{a}_T(\beta), \mathbf{S}\mathbf{a}_T(\beta)\rangle| = |\sum_{l=0}^{N_t-1} e^{-j2\pi l f/N_t}\bar{g}_{l-\tau} g_l|, \tag{43}$$

where $g_l \sim \mathscr{N}(0,1)$. Note that the terms in this sum are no longer all jointly independent. But similar to the proof of Theorem 5.1 in [27] we observe that for any $\tau \neq 0$ we can split the index set $0, \dots, N_t - 1$ into two subsets $\Lambda_\tau^1, \Lambda_\tau^2 \subset \{0, \dots, N_t - 1\}$, each of size $N_t/2$, such that the $N_t/2$ variables $e^{-j2\pi l f/N_t}\bar{g}(l-\tau)g(l)$ are jointly independent for $l \in \Lambda_\tau^1$, and analogous for Λ_τ^2. (For convenience we assume here that N_t is even, but with a negligible modification the argument also applies for odd N_t). In other words, each of the sums $\sum_{l \in \Lambda_\tau^r} e^{-j2\pi l f/N_t}\bar{g}(l-\tau)g(l), r = 1, 2$, contains only jointly independent terms. Hence, we can apply Lemma 10 and obtain

$$\mathbb{P}\Big(\big|\sum_{l\in\Lambda_\tau^r} e^{-j2\pi l f/N_t}\bar{g}(l-\tau)g(l)\big|>t\Big)\le 2\exp\Big(-\frac{t^2}{N_t/2+2t)}\Big) \tag{44}$$

for all $t>0$. Choosing $t=\frac{3}{2}\sqrt{N_t\log(N_tN_fN_RN_T)}$ gives

$$\begin{aligned}\mathbb{P}\Big(\big|\sum_{l\in\Lambda_\tau^r} e^{-j2\pi l f/N_t}\bar{g}(l-\tau)g(l)\big|>\frac{3}{2}\sqrt{N_t\log(N_tN_fN_RN_T)}\Big)&\\ \le 2\exp\Big(-\frac{\frac{9}{4}N_t\log(N_tN_fN_RN_T)}{\frac{N_t}{2}+3\sqrt{N_t\log(N_tN_RN_T)}}\Big)&\\ \le 2\exp\Big(-\frac{9\log(N_tN_fN_RN_T)}{2+12\sqrt{\frac{\log(N_tN_fN_RN_T)}{N_t}}}\Big).&\end{aligned} \tag{45}$$

Condition (27) implies that $12\sqrt{\frac{\log(N_tN_fN_RN_T)}{N_t}}\le\frac{5}{2}$, hence the estimate in (45) becomes

$$\begin{aligned}\mathbb{P}\Big(\big|\sum_{l\in\Lambda_\tau^r} e^{-j2\pi l f/N_t}\bar{g}(l-\tau)g(l)\big|>\frac{3}{2}\sqrt{\log(N_tN_fN_RN_T)}\sqrt{N_t}\Big)&\\ \le 2\exp\Big(-\frac{9\log(N_tN_fN_RN_T)}{2+\frac{5}{2}}\Big)&\\ =2\exp\big(-2\log(N_tN_fN_RN_T)\big)&\\ =2(N_tN_fN_RN_T)^{-2}.&\end{aligned} \tag{46}$$

Using Eq. (43), inequality (46), and the pigeonhole principle, we obtain

$$\mathbb{P}\Big(|\langle\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta),\mathbf{S}\mathbf{a}_T(\beta)\rangle|>3\sqrt{N_t\log(N_tN_fN_RN_T)}\Big)\le 4(N_tN_fN_RN_T)^{-2}, \tag{47}$$

Combining this estimate with (29) yields

$$\mathbb{P}\Big(|\langle\mathbf{A}_{\tau,f,\beta},\mathbf{A}_{\tau',f',\beta}\rangle|\ge 3N_R\sqrt{N_t\log(N_\tau N_fN_RN_T)}\Big)\le 4(N_tN_fN_RN_T)^{-2}, \tag{48}$$

We apply the union bound over the $\frac{N_t}{2}N_fN_TN_R$ different possibilities and arrive at

$$\begin{aligned}\mathbb{P}\Big(\max|\langle\mathbf{A}_{\tau,f,\beta},\mathbf{A}_{\tau',f',\beta}\rangle|\le 3N_R\sqrt{N_t\log(N_\tau N_fN_RN_T)}\Big)&\\ \ge 1-4(N_tN_fN_RN_T)^{-1},&\end{aligned} \tag{49}$$

where the maximum is taken over all $\tau,\tau',\beta,\beta',f,f'$ with $\tau\ne\tau'$.

To apply Theorem 1.3 in [5], we need to normalize the columns of $\mathbf{A}$. The following result shows the corresponding bounds for $\tilde{\mathbf{A}}$.

Lemma 3. *Let $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{D}^{-1}$, where the entries of the $N_\tau N_f N_\beta \times N_\tau N_f N_\beta$ diagonal matrix are given by $\mathbf{D}_{(\tau,f,\beta),(\tau,f,\beta)} = \|\mathbf{A}_{\tau,f,\beta}\|_2$. Under the conditions of Theorem 1 there holds*

$$\mathbb{P}\Big(\|\tilde{\mathbf{A}}\|_{op}^2 < 6N_T N_f\Big) \geq 1 - p_1, \tag{50}$$

where

$$p_1 = e^{-\frac{(1-\sqrt{1/3})^2 N_t}{2}} + N_T e^{-(\frac{3}{2}-\sqrt{2})N_t},$$

and

$$\mathbb{P}\Big(\mu(\tilde{\mathbf{A}}) \leq 6\sqrt{\frac{1}{N_t}\log(N_\tau N_f N_R N_T)}\Big) \geq 1 - p_2, \tag{51}$$

where

$$p_2 = 2(N_R N_T)^{-1} + 2(N_\tau N_R N_T)^{-1} + 2(N_f N_R N_T)^{-1}$$
$$+6(N_\tau N_f N_R N_T)^{-1} + 2e^{-\frac{N_t(\sqrt{2}-1)^2}{4}}.$$

Proof. We have

$$\|\tilde{\mathbf{A}}\|_{\text{op}}^2 \leq \frac{\|\mathbf{A}\|_{\text{op}}^2}{\max_{\tau,f,\beta}\|\mathbf{A}_{\tau,f,\beta}\|_2^2}. \tag{52}$$

Recall that

$$\mathbf{A}_{\tau,f,\beta} = \mathbf{a}_R(\beta) \otimes (\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)), \tag{53}$$

hence $\|\mathbf{A}_{\tau,f,\beta}\|_2^2 = \|\mathbf{a}_R(\beta)\|_2^2 \|\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)\|_2^2$. Since the entries $(\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta))_k \sim \mathscr{C}\mathscr{N}(0,N_T)$, we have $\mathbb{E}\|\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)\| = \sqrt{N_t}$, and thus by Lemma 7

$$\mathbb{P}\Big(\sqrt{N_t} - \|\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)\|_2 > t\Big) \leq e^{-\frac{t^2}{2}}, \tag{54}$$

for all $t > 0$, hence

$$\mathbb{P}\Big(\frac{1}{\|\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)\|_2^2} < \frac{1}{(\sqrt{N_t}-t)^2}\Big) \geq 1 - e^{-\frac{t^2}{2}}, \tag{55}$$

Choosing $t = (1-\sqrt{1/3})\sqrt{N_t}$ in (55) and forming the union bound only over the $N_R N_T$ different possibilities associated with β (note that $\|\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)\|_2 = \|\mathbf{S}\mathbf{a}_T(\beta)\|_2$ for all τ and f), gives

$$\mathbb{P}\Big(\frac{1}{\max\limits_{\tau,f,\beta}\|\mathbf{A}_{\tau,f,\beta}\|_2^2} < \frac{3}{N_t N_R}\Big) \geq 1 - N_R N_T e^{-\frac{N_t(1-\sqrt{1/3})^2}{2}}. \tag{56}$$

The diligent reader may convince herself that the probability in (56) is indeed close to one under the condition (12). We insert (19) and (56) into (52) and obtain

$$\mathbb{P}\Big(\|\tilde{\mathbf{A}}\|_{\text{op}}^2 < 6N_T N_f\Big) \geq 1 - e^{-\frac{N_t(1-\sqrt{1/3})^2}{2}} - N_T e^{-(\frac{3}{2}-\sqrt{2})N_t}. \tag{57}$$

which proves (50).

To establish (51) we denote $\mathbf{D}^{-1}_{(\tau,f,\beta),(\tau,f,\beta)} := \|\mathbf{A}_{\tau,f,\beta}\|_2^{-1}$ and observe that

$$\mu(\tilde{\mathbf{A}}) \leq \max\left\{\mathbf{D}^{-1}_{(\tau,f,\beta),(\tau,f,\beta)}|(\mathbf{A}^*\mathbf{A})_{(\tau,f,\beta),(\tau',f',\beta')}|\mathbf{D}^{-1}_{(\tau',f',\beta'),(\tau',f',\beta')}\right\}, \tag{58}$$

where the maximum is taken over all $(\tau,f,\beta) \neq (\tau',f',\beta')$, Using Lemma 7 and (53) we compute

$$\mathbb{P}\Big(\|\mathbf{A}_{\tau,f,\beta}\|_2 > \sqrt{N_t N_R} - \sqrt{N_R}t\Big) \geq 1 - e^{-\frac{t^2}{2}}. \tag{59}$$

Therefore

$$\mathbb{P}\Big(\frac{1}{\|\mathbf{A}_{\tau,f,\beta}\|_2} < \frac{1}{\sqrt{N_t N_R} - \sqrt{N_R}t}\Big) \geq 1 - e^{-\frac{t^2}{2}}, \tag{60}$$

and thus

$$\mathbb{P}\Big(|\tilde{\mathbf{A}}^*\tilde{\mathbf{A}})_{(\tau,f,\beta),(\tau',f',\beta')}| \leq \frac{1}{(\sqrt{N_t N_R} - \sqrt{N_R}t)^2}|(\mathbf{A}^*\mathbf{A})_{(\tau,f,\beta),(\tau',f',\beta')}|\Big) \geq 1 - 2e^{-\frac{t^2}{2}}, \tag{61}$$

By choosing $t = (1 - 1/\sqrt{2})\sqrt{N_t}$, we can write (61) as

$$\mathbb{P}\Big(|\tilde{\mathbf{A}}^*\tilde{\mathbf{A}})_{(\tau,f,\beta),(\tau',f',\beta')}| \leq \frac{2}{N_t N_R}|(\mathbf{A}^*\mathbf{A})_{(\tau,f,\beta),(\tau',f',\beta')}|\Big) \geq 1 - 2e^{-\frac{N_t(\sqrt{2}-1)^2}{4}}. \tag{62}$$

Finally, plugging (62) into (58) and using (28) we arrive at

$$\mathbb{P}\Big(\mu(\tilde{\mathbf{A}}) \leq 6\sqrt{\frac{1}{N_t}\log(N_\tau N_f N_R N_T)}\Big) \geq 1 - p_2, \tag{63}$$

where

$$p_2 = 2(N_R N_T)^{-1} + 2(N_\tau N_R N_T)^{-1} + 2(N_f N_R N_T)^{-1} + 6(N_\tau N_f N_R N_T)^{-1} + 2e^{-\frac{N_t(\sqrt{2}-1)^2}{4}}$$

Proof. (of Theorem 1) We first point out that the assumptions of Theorem 1 imply that the conditions of Lemma 1 and Lemma 5 are fulfilled.

Note that the solution $\tilde{\mathbf{x}}$ of (9) and the solution $\tilde{\mathbf{z}}$ of the following lasso problem

$$\min_{\mathbf{z}} \frac{1}{2}\|\mathbf{A}\mathbf{D}^{-1}\mathbf{z} - \mathbf{y}\|_2^2 + \lambda\|\mathbf{z}\|_1, \qquad \text{with } \lambda = 2\sigma\sqrt{2\log(N_\tau N_R N_T)}, \tag{64}$$

satisfy $\text{supp}(\tilde{\mathbf{x}}) = \text{supp}(\mathbf{D}^{-1}\tilde{\mathbf{z}})$.

We will first establish the claims in Theorem 1 for the system $\tilde{\mathbf{A}}\mathbf{z} = \mathbf{y}$ where $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{D}^{-1}$, $\mathbf{z} = \mathbf{D}\mathbf{x}$ and then switch back to $\mathbf{A}\mathbf{x} = \mathbf{y}$.

We verify first condition (133). Property (14) and the fact that $\mathbf{z} = \mathbf{D}\mathbf{x}$ imply that

$$|z_k| \geq \frac{10\|\mathbf{A}_{\tau,\beta}\|_2}{\sqrt{N_R N_t}} \sigma \sqrt{2\log(N_\tau N_f N_\beta)}, \qquad \text{for } k \in S. \tag{65}$$

Using Lemma 7 we get that

$$\mathbb{P}\Big(\|\mathbf{A}_{\tau,\beta}\| \geq \sqrt{N_R N_t} - t\Big) \geq 1 - e^{-\frac{t^2}{2}}. \tag{66}$$

Choosing $t = \frac{2}{10}\sqrt{N_R N_t}$ and combining (66) with (65) gives

$$|z_k| \geq 8\sigma\sqrt{2\log(N_\tau N_f N_\beta)}, \qquad \text{for } k \in S,$$

with probability at least $1 - e^{-\frac{N_R N_t}{25}}$, thus establishing condition (133).

Note that $\tilde{\mathbf{A}}$ has unit-norm columns as required by Theorem 5. It remains to verify condition (131). Using the assumption (12), and the coherence bound (51) we compute

$$\begin{aligned}\mu^2(\tilde{\mathbf{A}}) \leq 36\frac{1}{N_t}\log(N_\tau N_f N_R N_T) &\leq 36\frac{\log(N_\tau N_f N_R N_T)}{\log^3(N_\tau N_f N_R N_T)} \\ &= \frac{36}{\log^2(N_\tau N_f N_R N_T)},\end{aligned}$$

which holds with probability as in (51), and thus the coherence property (131) is fulfilled.

Furthermore, using (50) we see that condition (13) implies

$$K \leq \frac{c_0 N_\tau N_R}{6\log(N_\tau N_f N_R N_T)} \leq \frac{c_0 N_\tau N_f N_R N_T}{\|\tilde{\mathbf{A}}\|_{\mathrm{op}}^2 \log(N_\tau N_f N_R N_T)}$$

with probability as stated in (50). Thus assumption (132) of Theorem 5 is also fulfilled (with high probability) and we obtain that

$$\mathrm{supp}(\tilde{\mathbf{z}}) = \mathrm{supp}(\mathbf{z}). \tag{67}$$

We note that the relation $\mathrm{supp}(\tilde{\mathbf{x}}) = \mathrm{supp}(\mathbf{x})$ holds with the same probability as the relation $\mathrm{supp}(\tilde{\mathbf{z}}) = \mathrm{supp}(\mathbf{z})$ (see Eq. (67)), since $\mathrm{supp}(\mathbf{z}) = \mathrm{supp}(\mathbf{x})$ and multiplication by an invertible diagonal matrix does not change the support of a vector. This establishes (83) with the corresponding probability.

As a consequence of (135) we have the following error bound

$$\frac{\|\tilde{\mathbf{z}} - \mathbf{z}\|_2}{\|\mathbf{z}\|_2} \leq \frac{\sigma\sqrt{3N_R N_t}}{\|\mathbf{y}\|_2} \tag{68}$$

which holds with probability at least

$$\big(1-p_1\big)\big(1-p_2\big)\big(1-p_4\big)\big(1-p_5\big)-\mathscr{O}\big((N_\tau N_\beta)^{-2\log 2}\big),$$

where the probabilities p_1, p_2, p_4, p_5 are as in Theorem 1. Using the fact that $\tilde{\mathbf{z}} = \mathbf{D}\tilde{\mathbf{x}}$, we compute

$$\frac{1}{\kappa(\mathbf{D})}\frac{\|\tilde{\mathbf{x}}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \le \frac{\|\mathbf{D}(\tilde{\mathbf{x}}-\mathbf{x})\|_2}{\|\mathbf{D}\mathbf{x}\|_2} = \frac{\|\tilde{\mathbf{z}}-\mathbf{z}\|_2}{\|\mathbf{z}\|_2},$$

or, equivalently,

$$\frac{\|\tilde{\mathbf{x}}-\mathbf{x}\|_2}{\|\mathbf{x}\|_2} \le \kappa(\mathbf{D})\frac{\|\tilde{\mathbf{z}}-\mathbf{z}\|_2}{\|\mathbf{z}\|_2}. \tag{69}$$

Proceeding along the lines of (54)–(56), we estimate

$$\mathbb{P}\big(\kappa(\mathbf{D}) \le 2\big) \ge 1 - N_R N_T e^{-\frac{N_t(1-\sqrt{1/3})^2}{2}}. \tag{70}$$

The bound (85) follows now from combining (68) with (69) and (70).

4 Deterministic Waveforms and Random Antenna Arrays

In this section, we are considering MIMO radar with random sensor arrays and deterministic waveforms. Specifically, we will investigate the use of Kerdock codes as transmission waveforms.

Random sensor arrays have been around for half a century. The pioneering work [25, 26] by Lo contains a mathematical analysis of important specific characteristics of random arrays, such as sidelope behavior and antenna gain. There is extensive engineering literature that deals with random arrays in connection with phased array radar technology, e.g., see [10]. Recently, Carin made an explicit connection between the areas of random sensor arrays and compressive sensing [6]. He has shown that algorithms developed in these two seemingly different areas are in fact highly interrelated. The setup in [6] is quite different from ours, since the paper is only concerned with angular resolution (thus transmission waveforms do not even explicitly enter into the model), while it is often crucial in practice to be able to estimate range and Doppler as well. Moreover, the theoretical analysis in [6] follows more an engineering style and places less emphasis on mathematical rigor. The paper [7] provides interesting results for the angular estimation of stationary targets. Its setup is similar to that in [6], and quite different from ours, as it does not deal with waveform design nor with moving targets.

Kerdock codes have been proposed for radar in [17], albeit in the setting of a single transmit antenna. Kerdock codes are known to perform rather poorly[2] even in the case of single targets as considered in [17]. Only in the setting of mulitple transmit antennas can Kerdock codes exhibit their enormous potential.

[2] This poor performance is caused by Property (ii) in Theorem 2.

The previous section considers a MIMO radar setting with a very specific (non-random) choice for the antenna locations, but random waveforms, while the current section deals with randomly spaced antennas, but very specific, deterministic waveforms. At first glance, the difference may appear to be mainly semantic. But in practice, the second setting has many advantages. From an engineer's viewpoint random waveforms have several drawback over properly designed deterministic waveforms: they are much harder to implement on a digital device (requiring more complicated hardware, more memory, ...); and they exhibit a larger peak-to-average-power ratio. On the other hand it makes no difference from the viewpoint of physics or hardware, if we place the antennas at random or at deterministic locations. In particular, the current section yields some important insights, which cannot be inferred from the previous section: We obtain a theoretical framework for radar operating with random antenna arrays, a technique which have been around for half a century; we show that Kerdock sequences, which are not useful for SISO or SIMO radar[3], are excellent for MIMO radar; our approach allows for waveforms that satisfy a number of properties which are very desirable in practice, and are not satisfied by random waveforms.

4.1 Kerdock Codes

In this section, we introduce one particularly useful set of transmission waveforms. Due to the setup in Section 2 it suffices that we deal with discrete, finite-length sequences as transmission signals. We briefly review the construction of Kerdock codes and some of their fundamental properties. There is a long list of properties that radar waveforms should satisfy. As we will see in this chapter, Kerdock codes fulfill many of them. Kerdock codes over $\mathbb{Z}_2$ (i.e., binary Kerdock codes) were originally introduced in [20]. In the seminal paper [2] the authors extend Kerdock codes from $\mathbb{Z}_2$ to $\mathbb{Z}_4$. By doing so, they uncover many fascinating properties of Kerdock codes and reveal numerous deep connections between coding theory, discrete geometry and group theory. In the same paper, the authors also extend Kerdock codes to the setting of $\mathbb{Z}_p$, where p is an odd prime.

Kerdock codes are an example of so-called mutually unbiased bases [32, 39]. Kerdock codes have also been proposed for use in communications engineering [14, 19]. In [17] the authors suggest the use of Kerdock codes for radar, based on the peculiar properties of the discrete ambiguity function associated with Kerdock codes. We emphasize however that for the single transmit antenna radar scenario Kerdock codes would actually perform rather badly, as discussed after Theorem 2. It is only in the setting of multiple transmit antennas that Kerdock codes become useful for radar.

For the remainder of this chapter, we will only be concerned with Kerdock codes over $\mathbb{Z}_p$. Some of the Kerdock codes over $\mathbb{Z}_p$, namely those corresponding

[3] SISO stands for single-input-single-output radar, and SIMO for single-input-multiple-output radar (i.e., a radar with one transmit and multiple receive antennas).

to Desarguesian planes in the language of [2], have also been derived earlier in [22] and [21]. A simple way to construct these Kerdock codes is via eigenvectors of time-frequency shift operators. Let p be an odd prime number. For each $k = 0,\dots,p-1$ we compute the eigenvector decomposition of $\mathbf{T}_1\mathbf{M}_k$ (which always exists, since $\mathbf{T}_1\mathbf{M}_k$ is a unitary matrix)

$$\mathbf{U}_{(k)}\boldsymbol{\Sigma}_{(k)}\mathbf{U}^*_{(k)} = \mathbf{T}_1\mathbf{M}_k, \tag{71}$$

where the unitary matrix $\mathbf{U}_{(k)}$ contains the eigenvectors of $\mathbf{T}_1\mathbf{M}_k$ and the diagonal matrix $\boldsymbol{\Sigma}_{(k)}$ the associated eigenvalues[4]. Furthermore, we define $\mathbf{U}_{(p)} := I_p$. Now, let $\mathbf{u}_{k,j}$ be the jth column of $\mathbf{U}_{(k)}$. The set consisting of the p^2+p vectors $\{\mathbf{u}_{k,j}, k = 0,\dots,p; j = 0,\dots,p-1\}$ forms a $\mathbb{Z}_p$-Kerdock code. There are numerous equivalent ways to derive this Kerdock code, but, as pointed out earlier, not *all* Kerdock codes over $\mathbb{Z}_p$ are equivalent (see also the comment following Corollary 11.6 in [2]). But we will be a bit sloppy, and simply refer to the Kerdock code constructed above as *the* Kerdock code.

In the following theorem, we collect those key properties of Kerdock codes that are most relevant for radar. These properties are either explicitly proved in [2, 17] or can be derived easily from properties stated in those papers.

Theorem 2. *Kerdock codes over $\mathbb{Z}_p$, where p is an odd prime, satisfy the following properties:*

(i) Mutually unbiased bases: For all $k = 0,\dots,p$ and all $j = 0,\dots,p-1$, there holds:

$$|\langle \mathbf{u}_{k,j}, \mathbf{u}_{k',j'}\rangle| = \begin{cases} 1 & \text{if } k = k', j = j', \\ 0 & \text{if } k = k', j \neq j', \\ \frac{1}{\sqrt{p}} & \text{if } k \neq k'. \end{cases}$$

(ii)Time-frequency "autocorrelation":

(a) For any fixed $(f,l) \neq (0,0)$ there exists a unique k_0 such that

$$|\langle \mathbf{M}_f\mathbf{T}_l\mathbf{u}_{k_0,j}, \mathbf{u}_{k_0,j}\rangle| = 1 \qquad \text{for } j = 0,\dots,p-1, \tag{72}$$

$$|\langle \mathbf{M}_f\mathbf{T}_l\mathbf{u}_{k,j}, \mathbf{u}_{k,j}\rangle| = 0 \qquad \text{for } k \neq k_0. \tag{73}$$

(b) For any fixed $0 \le k \le p-1$, there exist (f_r, l_r), $r = 1,\dots,p$ such that

$$|\langle \mathbf{M}_{f_r}\mathbf{T}_{l_r}\mathbf{u}_{k,j}, \mathbf{u}_{k,j}\rangle| = 1 \qquad \text{for } j = 0,\dots,p-1, \tag{74}$$

(iii)Time-frequency cross-correlation: For all $k \neq k'$ and all f and l there holds:

$$|\langle \mathbf{M}_f\mathbf{T}_l\mathbf{u}_{k,j}, \mathbf{u}_{k',j}\rangle| \le \frac{1}{\sqrt{p}} \qquad \text{for } j = 0,\dots,p-1. \tag{75}$$

[4] The attentive reader will have noticed that $\mathbf{U}_{(0)}$ is just the $p \times p$ DFT matrix $\mathbf{F}_p$.

(iv) Polyphase property (Roots of unity property) in time and in frequency:
For any $k = 0, \dots, p-1; j = 0, \dots, p-1$, *there holds:*

$$\mathbf{u}_{k,j}(l) = e^{2\pi i r/p} \quad \text{for some } r \in \{0, \dots, p-1\}. \tag{76}$$

For any $k = 1, \dots, p; j = 0, \dots, p-1$, *there holds:*

$$\hat{\mathbf{u}}_{k,j}(l) = e^{2\pi i r/p} \quad \text{for some } r \in \{0, \dots, p-1\}. \tag{77}$$

Proof. Property (i) is proved for instance in Lemma 11.3 in [2]. Properties (ii) and (iii) appear in Theorem 3 of [17]. Statement (76) of property (iv) follows from the comment right after Corollary 11.6 in [2]. Finally, statement (77) of property (iv) follows from (71) together with property (3) and the well-known fundamental relationships

$$\mathbf{F}_p \mathbf{T}_x \mathbf{F}_p^* = \mathbf{M}_{-x}, \qquad \mathbf{F}_p \mathbf{M}_x \mathbf{F}_p^* = \mathbf{T}_x.$$

Kerdock codes have been proposed for adaptive radar in [17]. We emphasize again though that Kerdock codes would not be very effective for a radar system with a single transmit antenna (SISO or SIMO radar). This can be easily seen as follows: Assume we only have one antenna that transmits one waveform $\mathbf{s}$. Due to (74), $\mathbf{s}$ is (up to a constant phase factor) equal to $\mathbf{M}_f \mathbf{T}_l \mathbf{s}$ for some (f, l). In practice this ambiguity prevents us from determining the distance and the velocity of the object, when using Kerdock codes for SISO or SIMO radar.

As a consequence of the aforementioned ambiguity we will not use *all* of the Kerdock codes as transmission signals for our MIMO radar, instead we will choose one code for each index k. The reason is that we need the waveforms to have low time-frequency cross-correlation, while (75) only holds when k and k' are different.

Definition 1 (Kerdock waveforms). Let $\{\mathbf{u}_{k,j}, k = 0, \dots, p, j = 0, \dots, p-1\}$ be a Kerdock code over $\mathbb{Z}_p$. The *Kerdock waveforms* $\mathbf{k}_0, \dots, \mathbf{k}_r$, where $r < p$, are given by $\mathbf{k}_k = \mathbf{u}_{k,j}$ for some arbitrary j. In other words, for each $k = 0, \dots, r-1$ we pick an arbitrary vector from the orthonormal basis $\{\mathbf{u}_{k,j}\}_{j=0}^{p-1}$.

Note that Kerdock waveforms do not include any unit vectors, since only the first r unitary matrices $\mathbf{U}_{(0)}, \dots, \mathbf{U}_{(r-1)}$ are considered and r is strictly less than p (recall that $\mathbf{U}_{(p)} = \mathbf{I}_p$).

4.2 Compressive MIMO Radar and Kerdock Waveforms

Now we are ready to state the main theorem of this section, which illustrates the usefulness of Kerdock waveforms for compressive sensing based MIMO radar.

Theorem 3. *Consider* $\mathbf{y} = \mathbf{A}\mathbf{x} + \mathbf{w}$, *where* $\mathbf{A}$ *is defined as in* (7) *and* $\mathbf{w}_j \in \mathscr{C}N(0, \sigma^2)$. *Assume that the positions of the transmit and receive antennas* p_j*'s and* q_j*'s are chosen i.i.d. uniformly in* $[0, \frac{N_R N_T}{2}]$ *at random. Suppose further that each transmit* *antenna sends a different Kerdock waveform, i.e., the columns of the signal matrix* **S** *are different Kerdock waveforms. Choose the discretization stepsizes to be*

$$\Delta_\beta = \frac{2}{N_R N_T}, \quad \Delta_\tau = \frac{1}{2B}, \quad \Delta_f = \frac{1}{T}. \tag{78}$$

Furthermore, suppose that $\Delta_s = \frac{1}{2B}$ *(i.e.,* $N_s = 2TB$*) and*

$$\max\left(N_R N_T, 32 N_T^3 \log N_\tau N_f N_\beta\right) \leq N_s, \tag{79}$$

and also

$$\log^2 N_\tau N_f N_\beta \leq N_T \leq N_R. \tag{80}$$

If **x** *is drawn from the generic S-sparse scatterer model with*

$$S \leq \frac{c_0 N_\tau}{\log N_\tau N_f N_\beta} \tag{81}$$

for some constant $c_0 > 0$*, and if*

$$\min_{k \in I} |\mathbf{x}_k| > \frac{8\sqrt{3}\sigma}{\sqrt{N_R N_T}} \sqrt{2 \log N_\tau N_f N_\beta}, \tag{82}$$

then the solution $\tilde{\mathbf{x}}$ *of the debiased lasso computed with* $\lambda = 2\sigma\sqrt{2\log N_\tau N_f N_\beta}$ *satisfies*

$$\operatorname{supp}(\tilde{\mathbf{x}}) = \operatorname{supp}(\mathbf{x}), \tag{83}$$

with probability at least

$$1 - p_1, \tag{84}$$

and

$$\frac{\|\tilde{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{5\sigma\sqrt{3 N_R N_s}}{\|\mathbf{y}\|_2} \tag{85}$$

with probability at least

$$(1 - p_1)(1 - p_2), \tag{86}$$

where

$$\begin{aligned} p_1 = 16 N_\tau^{-2} N_R^{-1} + 8 N_\tau^{-2} N_f^{-2} + 4 N_T N_\tau^{-2} N_f^{-2} + 4(N_\tau N_f)^{-1} \\ + 4 N_\tau^{-3} N_f^{-3} N_R^{-2} N_T^{-1} + 8 N_T^{-2} (N_\tau N_f N_R)^{-3}, \end{aligned}$$

and

$$\begin{aligned} p_2 = 2(N_\tau N_f N_\beta)^{-1} (2\pi \log(N_\tau N_f N_\beta) + K(N_\tau N_f N_\beta)^{-1}) \\ + \mathcal{O}((N_\tau N_f N_\beta)^{-2\log 2}). \end{aligned}$$

Remarks and Discussion of the Theorem:

1. Theorem 3 demonstrates several advantages of the proposed sparsity-based approach compared to standard methods such as the matched-filter approach or minimal ℓ_2-norm based methods. First of all, unlike other methods (with the exception of [31]) that aim at target detection and estimation in the azimuth-range-Doppler domain, our approach provides guaranteed performance under explicit conditions. In terms of target detection capability, there is no limit on the dynamical range. A target can be detected as long as its reflection coefficient exceeds the noise level, see (82). Moreover, the resolution limits are actually attainable, and not just for one target, but for quite a large number of targets (as specified in (81)), and furthermore even if a "weak" target is located near a "strong" target.
2. The relations in (78) can be interpreted as guaranteed achievable resolutions in the azimuth-range-Doppler domain. The angular resolution $\Delta_\beta = \frac{2}{N_R N_T}$ means that we have a virtual array whose aperture is $\frac{N_R N_T}{2}$ (and not just $N_R + N_T$). The choices $\Delta_\tau = \frac{1}{2B}, \quad \Delta_f = \frac{1}{T}$ correspond to the resolutions with respect to range and Doppler, respectively.
3. We note that a virtual array of size $\frac{N_R N_T}{2}$ can also be achieved with a specific uniformly interleaved antenna arrangement, see [31]. However, in contrast to [31], Theorem 3 shows that we do not need to resort to random waveforms, but instead can employ deterministic waveforms that satisfy a variety of additional desirable properties.
4. The conditions in (79) and (80) are less straightforward to interpret than condition (78), as they are in part "footprints" of the proof. For instance, we believe that the factor N_T^3 in (79) is somewhat artificial and a more sophisticated approach may be able to eliminate or at least reduce that factor. However, the (in practice rather mild) condition in (79) that the number of samples N_s should at least be as large as $N_T N_R$ seems genuine. Moreover, it is not hard to see that we cannot improve our results by oversampling the received signal (i.e., by increasing N_s via setting $\Delta_s < \frac{1}{2B}$), since oversampling would neither improve the coherence of the matrix, nor provide any new, independent measurements.
5. The conditions in (80) show that we need both N_T and N_R to be larger than 1, i.e., we need to have an actual MIMO radar (and not a SISO or SIMO radar) in order for the theorem to hold. The latter condition $N_T \leq N_R$ in (80) is by no means necessary, but rather makes our computations a little cleaner. We could change it to, say, $N_T \leq 2N_R$, then the theorem would remain true with a slightly different probability of success.
6. It may seem that the conditions in (79) and (80) are a bit restrictive. But, in practice, our method works with a broad range of parameters as the simulations in Section 6 show.
7. The choice of Kerdock waveforms is by no means necessary. Other waveforms can give similar result, see Theorem 6.1 in [33].

Similar to the proof of Theorem 1, the proof of the above theorem requires an estimate of the bound for the operator norm of $\mathbf{A}$ and a bound for the coherence of $\mathbf{A}$.

Lemma 4. *Let* $\mathbf{A}$ *be the matrix in Theorem* 3 *satisfying* (79). *Then*

$$\mathbb{P}\Big(\|\mathbf{A}\|_{op}^2 \leq 2N_f N_R^2 N_T^2\Big) \geq 1 - 8N_\tau^{-2} N_R^{-1}. \tag{87}$$

Proof. Since $\|\mathbf{A}\|_{\mathrm{op}}^2 = \|\mathbf{A}\mathbf{A}^*\|_{\mathrm{op}}$, we express the matrix $\mathbf{B} = \mathbf{A}\mathbf{A}^*$ as block matrix

$$\begin{bmatrix} \mathbf{B}_{1,1} & \mathbf{B}_{1,2} & \dots & \mathbf{B}_{1,N_R} \\ \vdots & & \ddots & \vdots \\ \mathbf{B}_{N_R,1} & \dots & & \mathbf{B}_{N_R,N_R} \end{bmatrix},$$

where the blocks $\{\mathbf{B}_{j,j'}\}_{j,j'=1}^{N_R}$ are matrices of size $N_t \times N_t$.

Via a simple permutation we can turn $\mathbf{B}$ into a matrix $\mathbf{C}$ with blocks $\{\mathbf{C}_{l,l'}\}_{l,l'=1}^{N_s}$ of size $N_R \times N_R$, where the (j,j')th entry of the block $\mathbf{C}_{l,l'}$ is defined as

$$\begin{aligned} \mathbf{C}_{[l,j;l'j']} = \mathbf{B}_{[j,l;j',l']} = (\mathbf{A}\mathbf{A}^*)_{[j,l;j',l']} &= \sum_\beta \sum_\tau \sum_f \mathbf{A}_{[j,l;\tau,f,\beta]} \overline{\mathbf{A}_{[j',l';\tau,f,\beta]}} \\ &= \sum_\beta e^{2\pi i(q_j - q_{j'})\beta} \sum_{k=1}^{N_T} \sum_{k'=1}^{N_T} e^{2\pi i(p_k - p_{k'})\beta} \langle \mathbf{T}_l \mathbf{k}_k, \mathbf{T}_{l'} \mathbf{k}_{k'} \rangle \sum_{m=1}^{N_f} e^{2\pi i(l-l')\Delta_t m \Delta_f} \\ &= \delta_{l,l'} N_f \sum_\beta e^{2\pi i(q_j - q_{j'})\beta} \sum_{k=1}^{N_T} \sum_{k'=1}^{N_T} e^{2\pi i(p_k - p_{k'})\beta} \langle \mathbf{T}_{l-l'} \mathbf{k}_k, \mathbf{k}_{k'} \rangle. \end{aligned} \tag{88}$$

From (88) it is easy to see that $\mathbf{C}$ is block-diagonal and that all the diagonal-blocks are identical. So we only have to bound the first block $\mathbf{C}_{1,1}$.

$$\begin{aligned} \mathbf{C}_{[1,j;1,j']} &= N_f \sum_\beta e^{2\pi i(q_j - q_{j'})\beta} \sum_{k=1}^{N_T} \sum_{k'=1}^{N_T} e^{2\pi i(p_k - p_{k'})\beta} \langle \mathbf{k}_k, \mathbf{k}_{k'} \rangle \\ &= N_f \sum_{n=0}^{N_R N_T - 1} e^{2\pi i(q_j - q_{j'})\frac{n}{N_R N_T}} \sum_{k=1}^{N_T} \sum_{k'=1}^{N_T} e^{2\pi i(p_k - p_{k'})\frac{n}{N_R N_T}} \langle \mathbf{k}_k, \mathbf{k}_{k'} \rangle. \end{aligned}$$

Define $c_n = \sum_{k=1}^{N_T} \sum_{k'=1}^{N_T} e^{2\pi i(p_k - p_k')\frac{n}{N_R N_T}} \langle \mathbf{k}_k, \mathbf{k}_{k'} \rangle$, then

$$\mathbf{C}_{1,1} = N_f \sum_{n=0}^{N_R N_T - 1} c_n X_n,$$

where X_n is the matrix-valued random variable given by

$$(X_n)_{j,j'} = e^{2\pi i(q_j - q_{j'})\frac{n}{N_R N_T}}$$

and therefore $\|X_n\|_{\text{op}} = N_R$.

Note that $\mathbb{E}(e^{2\pi i(p_k - p_{k'})n}) = 0$ and $|\langle \mathbf{k}_k, \mathbf{k}_{k'} \rangle| \le \frac{1}{\sqrt{N_s}}$ for $k \neq k'$. Choosing $t = 2\sqrt{\frac{N_T}{N_s}}\sqrt{\log N_\tau N_R N_T}$ in (124) of Lemma 11, we arrive at

$$\mathbb{P}\Big(|c_n| \le N_T(1 + 4\sqrt{\frac{N_T}{N_s}}\sqrt{\log N_\tau N_R N_T})\Big) \ge 1 - 8N_T(N_\tau N_R N_T)^{-2},$$

then the assumption in (79) implies that $16 N_T \log N_\tau N_R N_T \le N_s$, therefore

$$\mathbb{P}\Big(|c_n| \le 2N_T\Big) \ge 1 - 8N_T(N_\tau N_R N_T)^{-2}.$$

We apply the union bound over the $N_R N_T$ possibilities associated with n and get

$$\mathbb{P}\Big(\max |c_n| \le 2N_T\Big) \ge 1 - 8N_\tau^{-2} N_R^{-1},$$

which implies that

$$\mathbb{P}\Big(\|\mathbf{C}_{1,1}\|_{\text{op}} \le 2N_f N_R^2 N_T^2\Big) \ge 1 - 8N_\tau^{-2} N_R^{-1}.$$

Then the fact that $\|\mathbf{B}\|_{\text{op}} = \|\mathbf{C}\|_{\text{op}} = \|\mathbf{C}_{1,1}\|_{\text{op}}$ will give us the desired conclusion.

Lemma 5. *Let* $\mathbf{A}$ *be the matrix in Theorem 3 satisfying* (79) *and* (80). *Then*

$$\max_{(\tau,f,\beta)\neq(\tau',f',\beta')} \left|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'} \rangle\right| \le 16 N_R \log N_\tau N_f N_R N_T \tag{89}$$

with probability at least

$$1 - 8N_\tau^{-2} N_f^{-2} - 4N_T N_\tau^{-2} N_f^{-2} - 4(N_\tau N_f)^{-1} - 4N_\tau^{-3} N_f^{-3} N_R^{-2} N_T^{-1} \\ - 8N_T^{-2}(N_\tau N_f N_R)^{-3}.$$

Proof. We need to find an upper bound for $\max |\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'} \rangle|$ where the maximum is taken over $(\tau, f, \beta) \neq (\tau', f', \beta')$. Recall that $\mathbf{S}_{\tau,f} = \mathbf{M}_f \mathbf{T}_\tau \mathbf{S}$. It follows from the definition that

$$\mathbf{A}_{\tau,f,\beta} = \mathbf{a}_R(\beta) \otimes (\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)),$$

from which we readily compute

$$|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'} \rangle| = |\langle \mathbf{a}_R(\beta), \mathbf{a}_R(\beta') \rangle| |\langle \mathbf{S}_{\tau,f}\mathbf{a}_T(\beta), \mathbf{S}_{\tau',f'}\mathbf{a}_T(\beta') \rangle|.$$

We use the discretization $\beta = n\Delta_\beta$, $\beta' = n'\Delta_\beta$, where $\Delta_\beta = \frac{2}{N_R N_T}$, $n, n' = 1, \dots, N_\beta$, with $N_\beta = N_R N_T$.

Since

$$|\langle \mathbf{S}_{\tau,f}\mathbf{a}_T(\beta), \mathbf{S}_{\tau',f'}\mathbf{a}_T(\beta') \rangle| = |\langle \mathbf{S}_{\tau-\tau',f-f'}\mathbf{a}_T(\beta), \mathbf{S}\mathbf{a}_T(\beta') \rangle|$$

for $\tau, \tau' = 0, \ldots, N_\tau - 1, f, f' = 0, \ldots, N_f - 1$. We can confine the range of values for τ, τ' to $\tau' = 0, \tau = 0, \ldots, N_\tau - 1$ and f, f' to $f' = 0, f = 0, \ldots, N_f - 1$, then we only need to estimate $|\langle \mathbf{S}_{\tau,f}\mathbf{a}_T(\beta), \mathbf{S}\mathbf{a}_T(\beta')\rangle|$. We now consider three cases.

Case (i) $\beta \neq \beta', \tau = 0, f = 0$**:**

By Theorem 4.5 in [18], for any $t_1 > 0$

$$\mathbb{P}\Big(|\langle \mathbf{a}_R(\beta), \mathbf{a}_R(\beta')\rangle| \geq t_1\Big) \leq 4\exp\Big(-\frac{t_1^2}{4N_R}\Big),$$

choosing $t_1 = 2\sqrt{2}\sqrt{N_R}\sqrt{\log N_\tau N_f N_R N_T}$ will give us that

$$\mathbb{P}\Big(|\langle \mathbf{a}_R(\beta), \mathbf{a}_R(\beta')\rangle| \leq 2\sqrt{2}\sqrt{N_R}\sqrt{\log N_\tau N_f N_R N_T}\Big) \geq 1 - 4(N_\tau N_f N_R N_T)^{-2}. \quad (90)$$

Define $\mathbf{M} := \mathbf{S}^*\mathbf{S}$ with entries m_{jk}, then $|m_{kj}| = |\langle \mathbf{k}_k, \mathbf{k}_j\rangle| \leq \frac{1}{\sqrt{N_s}}$ for $k \neq j$ and m_{jj}=1. We choose s=$2\sqrt{2}\sqrt{N_T}\sqrt{\log N_\tau N_f N_R N_T}$ and t=$2\sqrt{2}\sqrt{\frac{N_T}{N_s}}\sqrt{\log N_\tau N_f N_R N_T}$ in (123) of Lemma 11 and get

$$\mathbb{P}\Big(|\langle \mathbf{S}^*\mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta')\rangle| \leq (2\sqrt{2}\sqrt{N_T} + 2\sqrt{2}N_T\sqrt{\frac{N_T}{N_s}})\sqrt{\log N_\tau N_f N_R N_T}\Big)$$
$$\geq 1 - 4(N_\tau N_f N_R N_T)^{-2} - 4N_T(N_\tau N_f N_R N_T)^{-2},$$

combined with (90),

$$\mathbb{P}\Big(|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau,f',\beta'}\rangle| \leq 8(\sqrt{N_R N_T} + N_T\sqrt{\frac{N_R N_T}{N_s}})\log N_\tau N_f N_R N_T\Big)$$
$$\geq 1 - 8(N_\tau N_f N_R N_T)^{-2} - 4N_T(N_\tau N_f N_R N_T)^{-2}.$$

After taking the union bound over $(N_R N_T)^2$ different possibilities associated with β, β', we will have that

$$\mathbb{P}\Big(\max|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau,f',\beta'}\rangle| \leq 8(\sqrt{N_R N_T} + N_T\sqrt{\frac{N_R N_T}{N_s}})\log N_\tau N_f N_R N_T\Big)$$
$$\geq 1 - 8N_\tau^{-2}N_f^{-2} - 4N_T N_\tau^{-2}N_f^{-2}.$$

A little algebra, using (79) and (80), shows that

$$8(\sqrt{N_R N_T} + N_T\sqrt{\frac{N_R N_T}{N_s}}) \leq 16N_R,$$

therefore

$$\mathbb{P}\Big(\max|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau,f,\beta'}\rangle| \leq 16N_R \log N_\tau N_f N_R N_T\Big)$$
$$\geq 1 - 8N_\tau^{-2}N_f^{-2} - 4N_T N_\tau^{-2}N_f^{-2}. \quad (91)$$

Case (ii) $\beta \neq \beta', (\tau, f) \neq (0,0)$**:**

As in case (i), we have $\beta \neq \beta'$, therefore (90) holds. Define $\mathbf{C} = \mathbf{S}^*_{\tau,f}\mathbf{S}$ with entries c_{jk}. From the properties of $\mathbf{k}_j$, we have that

$$|c_{kj}| = |\langle \mathbf{M}_f \mathbf{T}_\tau \mathbf{k}_k, \mathbf{k}_j \rangle| \leq \frac{1}{\sqrt{N_s}} \text{ for } k \neq j$$

and there exists j_0 such that $|c_{j_0 j_0}| = 1$ and $c_{jj} = 0$ for $j \neq j_0$. Then

$$|\langle \mathbf{S}^*_{\tau,f}\mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta')\rangle| \leq 1 + |\langle \mathbf{C}'\mathbf{a}_T(\beta), \mathbf{a}_T(\beta')\rangle|,$$

where $\mathbf{C}'$ is a zero-diagonal matrix which coincides with $\mathbf{C}$ at off-diagonal entries. Clearyl, $\mathbf{C}'$ satisfies the condition for (121) to hold. Choosing $t = 4\frac{\sqrt{N_T}}{\sqrt{N_s}} \sqrt{\log N_\tau N_f N_R N_T}$ in (121) of Lemma 11 yields

$$\mathbb{P}\Big(|\langle \mathbf{S}^*_{\tau,f}\mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta')\rangle| \leq 1 + 4\frac{N_T\sqrt{N_T}}{\sqrt{N_s}}\sqrt{\log N_\tau N_f N_R N_T|}\Big) \geq 1 - 4N_T(N_\tau N_f N_R N_T)^{-4}, \tag{92}$$

from the assumption that $32N_T^3 \log N_\tau N_f N_R N_T \leq N_s$, together with (90), we will get

$$\mathbb{P}\Big(|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'}\rangle| \leq 4\sqrt{2}\sqrt{N_R}\sqrt{\log N_\tau N_f N_R N_T}\Big) \geq 1 - 4(N_\tau N_f N_R N_T)^{-2} - 4N_T(N_\tau N_f N_R N_T)^{-4}.$$

By (80), we deduce $\log N_\tau N_f N_R N_T \leq N_R$. Therefore

$$\mathbb{P}\Big(|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'}\rangle| \leq 4\sqrt{2}N_R\Big) \geq 1 - 4(N_\tau N_f N_R N_T)^{-2} - 4N_T(N_\tau N_f N_R N_T)^{-4}.$$

We apply the union bound over $N_\tau N_f N_R^2 N_T^2$ possibilities and arrive at

$$\mathbb{P}\Big(\max|\angle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'}\rangle| \leq 4\sqrt{2}N_R\Big) \geq 1 - 4(N_\tau N_f)^{-1} - 4N_\tau^{-3}N_f^{-3}N_R^{-2}N_T^{-1}. \tag{93}$$

Case (iii) $\beta = \beta', (\tau, f) \neq (0,0)$**:**

Note that the matrix $\mathbf{C} = \mathbf{S}^*_{\tau,f}\mathbf{S}$ has exactly the same properties as in Case (ii) above. Following the same argument as we show (92) and applying (122) of Lemma 11 combined with the assumption as in (79) gives us that

$$\mathbb{P}\Big(|\langle \mathbf{S}^*_{\tau,f}\mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta)\rangle| \leq 1 + 4\sqrt{2}\frac{N_T\sqrt{N_T}}{\sqrt{N_s}}\sqrt{\log N_\tau N_f N_R N_T}\Big) \geq 1 - 8N_T(N_\tau N_f N_R N_T)^{-4},$$

which implies that

$$\mathbb{P}\Big(|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta'}\rangle| \leq 2N_R\Big) \geq 1 - 8N_T^{-3}(N_\tau N_f N_R)^{-4},$$

We apply the union bound over the $N_\tau N_f N_R N_T$ possibilities associated with τ, f, and β

$$\mathbb{P}\Big(\max|\langle \mathbf{A}_{\tau,f,\beta}, \mathbf{A}_{\tau',f',\beta}\rangle| \leq 2N_R\Big) \geq 1 - 8N_T^{-2}(N_\tau N_f N_R)^{-3}. \tag{94}$$

(91), (93), and (94) will give the conclusion.

To apply Theorem 1.3 in [5], we need to normalize the columns of $\mathbf{A}$. We first have the following result which shows the lower and upper bounds of the norm of columns of $\mathbf{A}$.

Lemma 6. *Let* $\mathbf{A}$ *be defined as in Theorem 3 satisfying* (79)*, then*

$$\mathbb{P}\Big(\frac{1}{3}N_R N_T \leq \min\|\mathbf{A}_{\tau,f,\beta}\|_2^2 \leq \max\|\mathbf{A}_{\tau,f,\beta}\|_2^2 \leq \frac{5}{3}N_R N_T\Big) \geq 1 - 8N_\tau^{-2}N_R^{-1}. \tag{95}$$

Proof. Recall that

$$\begin{aligned}\|\mathbf{A}_{\tau,f,\beta}\|_2^2 = \|\mathbf{a}_R(\beta)\|_2^2\|\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta)\|_2^2 &= N_R\langle \mathbf{S}_{\tau,f}^*\mathbf{S}_{\tau,f}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta)\rangle \\ &= N_R\langle \mathbf{S}^*\mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta)\rangle.\end{aligned}$$

Setting $t = 2\sqrt{\frac{N_T}{N_s}}\sqrt{\log N_\tau N_R N_T}$ in (124) of Lemma 11 yields

$$\begin{aligned}&\mathbb{P}\Big(N_T(1 - 4\sqrt{\frac{N_T}{N_s}}\sqrt{\log N_\tau N_R N_T}) \leq |\langle \mathbf{S}^*\mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta)\rangle| \leq \\ &\quad N_T(1 + 4\sqrt{\frac{N_T}{N_s}}\sqrt{\log N_\tau N_R N_T})\Big) \geq 1 - 8N_T(N_\tau N_R N_T)^{-2}.\end{aligned}$$

An easy calculation from (79) leads

$$4\sqrt{\frac{N_T}{N_s}}\sqrt{\log N_\tau N_R N_T} \leq \frac{2}{3},$$

which indeed implies

$$\mathbb{P}\Big(\frac{1}{3}N_T \leq |\langle \mathbf{S}^*\mathbf{S}\mathbf{a}_T(\beta), \mathbf{a}_T(\beta)\rangle| \leq \frac{5}{3}N_T\Big) \geq 1 - 8N_T(N_\tau N_R N_T)^{-2}. \tag{96}$$

Since the above probability does not depend on τ or f, we take all $N_R N_T$ possibilities of β and conclude the proof of the lemma.

Corollary 1. *Suppose* $\tilde{\mathbf{A}} = \mathbf{A}\mathbf{D}^{-1}$ *where* $\mathbf{D}$ *is the* $N_\tau N_f N_\beta \times N_\tau N_f N_\beta$ *diagonal matrix defined by* $\mathbf{D}_{(\tau,f,\beta),(\tau,f,\beta)} = \|\mathbf{A}_{\tau,f,\beta}\|_2$*, or in other words* $\tilde{\mathbf{A}}$ *is the matrix with* *unit-norm columns from* $\mathbf{A}$. *Then*

$$\mathbb{P}\Big(\|\tilde{\mathbf{A}}\|_{op}^2 \le 6N_f N_R N_T\Big) \ge 1 - 16 N_\tau^{-2} N_R^{-1}, \tag{97}$$

and

$$\mathbb{P}\Big(\mu(\tilde{\mathbf{A}}) \le 48 \frac{\log N_\tau N_f N_R N_T}{N_T}\Big) \ge 1 - p_3, \tag{98}$$

where

$$\begin{aligned} p_3 = {} & 8N_\tau^{-2}N_R^{-1} + 8N_\tau^{-2}N_f^{-2} + 4N_T N_\tau^{-2} N_f^{-2} + 4(N_\tau N_f)^{-1} \\ & + 4N_\tau^{-3} N_f^{-3} N_R^{-2} N_T^{-1} + 8N_T^{-2}(N_\tau N_f N_R)^{-3}. \end{aligned}$$

Proof. This corollary is a direct consequence of Lemma 4, Lemma 5, and Lemma 6.

Proof. (of Theorem 3) Similar to Theorem 1, the proof follows from Lemma 6, Corollary 1, and Theorem 5.

5 Off-the-Grid Compressive Sensing Radar

In the previous sections, we have assumed that the targets lie on the grid points, while common in compressive sensing, is certainly quite restrictive. A violation of this assumption will result in a model mismatch, sometimes dubbed *gridding error*, which can potentially be quite severe [8, 16]. Recently some interesting strategies have been proposed to overcome this gridding error [9, 34]. But these methods, at least in their current form, are not directly applicable to our setting.

In this section, we are concerned with the scenario where the targets lie in a continuous domain. We consider the Doppler-free case, i.e., we are interested in recovering the angular location and distance of the targets. The approach in this section draws heavily from recent results by Candès and Ferndanez-Granda on super-resolution [4].

Before introducing the main result in this section, we first define the total variation norm (TV-norm) for measures. The TV-norm of a complex measure ν on a measurable set Ω is defined to be

$$|\nu|(\Omega) = \sup \sum_{B\in\pi} |\nu(B)|,$$

where the supremum is taken over all partitions π of the measurable set Ω into a finite number of disjoint measurable subsets. The TV-norm can be interpreted as being the continuous analog to the ℓ_1 norm for discrete signals.

The following total variation minimization has been proposed recently in connection with super-resolution problems [4].

$$\min_{\tilde{\mathbf{x}}} \|\tilde{\mathbf{x}}\|_{TV} \qquad \text{s.t. } \mathbf{A}\tilde{\mathbf{x}} = \mathbf{y} \tag{99}$$

We are going to show that the TV-norm minimization can also be used in grid-free compressive sensing radar, at least for the SIMO case.

5.1 Off-the-Grid Compressive Sensing SIMO Radar

We first consider SIMO radar. Suppose we have single transmit antenna transmitting signal

$$\mathbf{s}(t) = \sum_{|k|\le f_c} d_k e^{j2\pi kt}$$

and N_R receive antennas. If the receive antennas are uniformly spaced with distance $d_R = \frac{1}{2}$(distance divided by wavelength), the array manifold will be

$$\mathbf{a}_R(\beta) = \begin{bmatrix} e^{j2\pi q_1\beta} \\ e^{j2\pi q_2\beta} \\ \vdots \\ e^{j2\pi q_{N_R}\beta} \end{bmatrix} \tag{100}$$

where $q_l = d_R(l-1) = \frac{l-1}{2}$.

In this case the sensing matrix $\mathbf{A}$ has columns

$$\mathbf{A}_{\tau\beta} = \mathbf{a}_R(\beta) \otimes \mathbf{T}_\tau\mathbf{s} = \begin{bmatrix} e^{j2\pi q_1\beta}\mathbf{T}_\tau\mathbf{s} \\ e^{j2\pi q_2\beta}\mathbf{T}_\tau\mathbf{s} \\ \vdots \\ e^{j2\pi q_{N_R}\beta}\mathbf{T}_\tau\mathbf{s} \end{bmatrix}. \tag{101}$$

Consequently the equation

$$\mathbf{A}\mathbf{x} = \mathbf{y}, \tag{102}$$

can be written as

$$\begin{bmatrix} e^{j2\pi q_1\beta}\mathbf{T}_\tau\mathbf{s} \\ e^{j2\pi q_2\beta}\mathbf{T}_\tau\mathbf{s} \\ \vdots \\ e^{j2\pi q_{N_R}\beta}\mathbf{T}_\tau\mathbf{s} \end{bmatrix}_{\tau\beta} \mathbf{x} = \begin{bmatrix} \mathbf{y}_1 \\ \mathbf{y}_2 \\ \vdots \\ \mathbf{y}_{N_R} \end{bmatrix}, \tag{103}$$

or, alternatively, we can write for $1 \le l \le N_R$,

$$\left[e^{j2\pi q_l\beta}\mathbf{T}_\tau\mathbf{s}\right]_{\tau\beta} \mathbf{x} = \mathbf{y}_l, \tag{104}$$

Since **A** is a matrix with infinitely many columns, we have that

$$\sum_{\tau\beta} e^{j2\pi q_l \beta} \mathbf{T}_\tau \mathbf{s} x_{\tau\beta} = \mathbf{y}_l. \tag{105}$$

Theorem 4. *Suppose we have a single transmit antenna, N_R receive antennas, and that the transmit antenna sends a band-limited signal*

$$\mathbf{s}(t) = \sum_{|k|\le f_c} d_k e^{j2\pi kt},$$

such that the entries of the discrete Fourier transform of $\mathbf{d} = (d_k)_{|d_k|\le f_c}$ are all nonzero. Assume that the received antennas are uniformly spaced with distance $\frac{1}{2}$. We take $n = 2f_c + 1$ samples of the received signal from each receive antenna. If the targets $x_k = (\tau_k, \beta_k)$ satisfy the minimum separation condition, i.e.,

$$|\tau_l - \tau_k| \ge \frac{4.76}{n}, \;\; or \;\; |\beta_l - \beta_k| \ge \frac{4.76}{N_R},$$

for $l \ne k$, then the total variation minimization (99) will recover $\mathbf{x}$ exactly.

Remarks:

1. We do not minimize the total variation on the system directly, a Fourier transform is performed before that.
2. This theorem only concerns the noise free case, the efficiency of the algorithm with respect to different SNR is demonstrated in the simulation section. A theorem for the noisy case can be derived by adopting the proof techniques in [3].
3. The factor 4.76 is definitely not optimal. We believe this factor should be 1 as we will see in the simulation section.

Proof. Taking the Fourier transform of (105) gives us that

$$\sum_{\tau\beta} e^{j2\pi q_l \beta} \mathbf{M}_{-\tau} \hat{\mathbf{s}} x_{\tau\beta} = \hat{\mathbf{y}}_l. \tag{106}$$

If we define the matrix **B** with the same dimensions as **A** such that each column of **B** is of the form

$$\begin{bmatrix} e^{j2\pi q_1 \beta} \mathbf{M}_{-\tau} \hat{\mathbf{s}} \\ e^{j2\pi q_2 \beta} \mathbf{M}_{-\tau} \hat{\mathbf{s}} \\ \vdots \\ e^{j2\pi q_{N_R} \beta} \mathbf{M}_{-\tau} \hat{\mathbf{s}} \end{bmatrix}, \tag{107}$$

then (103) is equivalent to

$$\begin{bmatrix} e^{j2\pi q_1\beta}\mathbf{M}_{-\tau}\hat{\mathbf{s}} \\ e^{j2\pi q_2\beta}\mathbf{M}_{-\tau}\hat{\mathbf{s}} \\ \vdots \\ e^{j2\pi q_{N_R}\beta}\mathbf{M}_{-\tau}\hat{\mathbf{s}} \end{bmatrix}_{\tau\beta} \mathbf{x} = \begin{bmatrix} \hat{\mathbf{y}}_1 \\ \hat{\mathbf{y}}_2 \\ \vdots \\ \hat{\mathbf{y}}_{N_R} \end{bmatrix}. \tag{108}$$

Note that $\mathbf{x}$ is the same in both (103) and (108), and we assume $\mathbf{x}$ to satisfy the minimal distance condition in time and azimuth.

We are going to minimize the TV-norm of $\mathbf{x}$ such that (108) holds. Similar as in compressed sensing, in order for the total-variation solution to be exact, it is sufficient to show the existence of certain dual polynomial. Suppose the vector $\mathbf{x}$ that we want to recover has support T. Then for any $\mathscr{W} \in \mathbb{C}^{|T|}$ with $|v_j| = 1$, we want to construct a polynomial $q(\tau,\beta)$ in Img$(\mathbf{B}^*)$ such that

$$\begin{cases} q(\tau_j,\beta_j) = v_j, & \text{if } (\tau_j,\beta_j) \in T \\ |q(\tau,\beta)| < 1, & \text{if } (\tau,\beta) \notin T \end{cases} \tag{109}$$

From the definition of $\mathbf{B}$, the polynomial $q(\tau,\beta)$ is of the form

$$q(\tau,\beta) = \sum_{l=1}^{N_R}\sum_{k=1}^{2f_c+1} c_{lk} e^{j2\pi q_l\beta} e^{-j2\pi k\tau} s_k \tag{110}$$

defined on $[0,1]\times[-1,1]$, where $s_k = \hat{\mathbf{s}}(\frac{k-1}{2f_c+1})$.

After change of variables, we want to construct polynomial

$$q(\tau,\beta) = \sum_{l=0}^{N_R-1}\sum_{k=1}^{2f_c+1} c_{lk} s_k e^{j2\pi l\beta} e^{-j2\pi k\tau} \tag{111}$$

defined on $[0,1]\times[-1/2,1/2]$, where we have used the optimal antenna locations such that $q_l = \frac{l-1}{2}$.

Define the ceiling function $\lceil a \rceil$ to be the smallest integer that is greater than or equal to a. Denote $f_\beta = \lceil \frac{N_R-1}{2} \rceil$.

$$\begin{aligned} q(\tau,\beta)e^{-j2\pi f_\beta\beta}e^{j2\pi(f_c+1)\tau} &= \sum_{l=0}^{N_R-1}\sum_{k=1}^{2f_c+1} c_{lk}s_k e^{j2\pi(l-f_\beta)\beta} e^{-j2\pi(k-f_c-1)\tau} \\ &= \sum_{l=-f_\beta}^{N_R-1-f_\beta}\sum_{k=-f_c}^{f_c} c_{lk}s_k e^{j2\pi l\beta} e^{j2\pi k\tau}, \end{aligned}$$

or

$$q(\tau,\beta)e^{-j2\pi f_\beta \beta}e^{j2\pi(f_c+1)\tau} = \sum_{l=-f_\beta}^{N_R-1-f_\beta}\sum_{k=-f_c}^{f_c} c_{lk}s_k e^{j2\pi l(\beta-\frac{1}{2})}e^{j2\pi k\tau}$$
$$= -\sum_{l=-f_\beta}^{N_R-1-f_\beta}\sum_{k=-f_c}^{f_c} c_{lk}s_k e^{j2\pi l\beta}e^{j2\pi k\tau}$$

defined on $[0,1]\times[0,1]$.

The fact that s_k's are all nonzero and the following proposition will make sure that such a polynomial

$$\tilde{q}(\tau,\beta) = q(\tau,\beta)e^{-j2\pi f_\beta \beta}e^{j2\pi(f_c+1)\tau}$$

exists.

Proposition 1. *Let $T=\{r_1,r_2,\dots\}\in\mathbb{T}^2$, where $r_k=(\tau_k,\beta_k)$ be a family of points obeying the minimum distance condition*

$$|\tau_l-\tau_k|\geq\frac{4.76}{n},\ \ or\ \ |\beta_l-\beta_k|\geq\frac{4.76}{N_R}.$$

Assume $f_\tau, f_\beta\geq 512$. Then for any vector $v\in\mathbb{R}^T$ with $|v_i|=1$, there exists a trigonometric polynomial q with Fourier series coefficients supported on

$$\{-f_\tau,-f_\tau+1,\dots,f_\tau\}\times\{-f_\beta,-f_\beta+1,\dots,f_\beta\}$$

with the property

$$\begin{cases} q(\tau_i,\beta_i)=v_i, & if\ (\tau_i,\beta_i)\in T \\ |q(\tau,\beta)|<1, & if\ (\tau,\beta)\notin T. \end{cases} \tag{112}$$

The proof of Proposition 1 is a general case of Proposition C.1 in [4], where the authors consider $f_\tau=f_\beta$.

In fact we define

$$K^{2D}(r)=K_1(x)K_2(y),$$

where K_1 and K_2 are the squares of Fejer kernels defined as the following

$$K_1(t)=[\frac{\sin((\frac{f_\tau}{2}+1)\pi t)}{(\frac{f_\tau}{2}+1)\sin(\pi t)}]^4$$

and

$$K_2(t)=[\frac{\sin((\frac{f_\beta}{2}+1)\pi t)}{(\frac{f_\beta}{2}+1)\sin(\pi t)}]^4.$$

Then we interpolate the sign pattern using K^{2D} and its derivatives $K^{2D}_{(1,0)}$ and $K^{2D}_{(0,1)}$,

$$q(r) = \sum_{r_i \in T} [\alpha_i K^{2D}(r - r_i) + \beta_{1i} K^{2D}_{(1,0)}(r - r_i) + \beta_{2i} K^{2D}_{(0,1)}(r - r_i)]$$

Following the same technique as for the proof of C.1 in [4] we can show that there exist coefficient vectors α and β such that $q(r)$ interpolates the values at r_i and the magnitude of q reaches a local maximum at these points, i.e., (112) holds.

We still have to address the question how to solve (99) numerically. In [4] an approach via semi-definite programming is proposed. While theoretically appealing, the disadvantange of that approach is the relatively high computational complexity. Instead, one might prefer to simply choose a fine discretization of the parameter space and compute an approximate solution to (99) via standard linear programming, where the accuracy depends on the discretization step. The validity of this simple approach is supported by the analysis in [35]. We adopt this approach also for the corresponding "off-the-grid" numerical simulations in the next section.

An important open problem is the extension of Theorem 4 to the MIMO case and in particular to the Doppler case. The challenge in the latter case is that (commutative) Fourier-analytic methods alone will no longer suffice, as the underlying group is the (non-commutative) Heisenberg group.

6 Numerical Experiments

In this section, we will illustrate our theoretical results and the associated numerical algorithms via computer simulations. We use the Matlab Toolbox TFOCS [1] and choose in TFOCS Auslender and Teboulle's single-projection method to solve (9). The main computational costs per iteration of this method are the operations $\mathbf{Ax}$ and $\mathbf{A}^*\mathbf{y}$. One can of course set up $\mathbf{A}$ explicitly and perform regular matrix multiplication. But due to the special structure of $\mathbf{A}$, we make the certain observation to take advantage of FFT to accelerate the computations.

In each experiment, S scatterers are placed randomly on the range-azimuth-Doppler grid, i.e, the vector $\mathbf{x}$ has S entries at random locations along the vector. White Gaussian noise is added to the composite data vector $\mathbf{Ax}$ with variance σ^2 determined to produce the specified output signal-to-noise ratio (SNR). The lasso solution $\hat{\mathbf{x}}$ is calculated with λ as specified in Theorems 1 and 3. The experiment is repeated 50 times. In each experiment we use an independent noise realization.

The probability of detection P_d and the probability of false alarm P_{fa} are computed as follows. The values of the estimated vector $\hat{\mathbf{x}}$ corresponding to the true scatterer locations are compared to a threshold. Detection is declared whenever a value exceeds the threshold. The probability of detection is defined as the number of detections divided by the total number of scatterers S. Next the values of the estimated vector $\hat{\mathbf{x}}$ corresponding to locations not containing scatterers are compared to the same threshold. A false alarm is declared whenever one of these values exceeds the threshold. The probability of false alarm is defined as the number of false alarms divided by $n-S$, where n is the signal dimension. The probabilities of detection and false alarm are averaged over the 50 repetitions of the experiment.

The probabilities are computed for a range of values of the threshold to produce the so-called receiver operating characteristics (ROC; [14, 28, 25]) the graph of P_d versus P_{fa}. As the threshold decreases, the probability of detection increases and so does the probability of false alarm. In practice the threshold is usually adjusted to achieve a specified probability of false alarm. We note that the probability of detection increases as the SNR increases and decreases as S, the number of scatterers increases. We carry out two sets of simulations with different parameters to show the performance of the algorithms.

The first set of simulations is done to compare the Kerdock waveforms and the white Gaussian noise waveforms at different SNR levels. When Kerdock waveforms are used, the locations of the transmit and receive antennas are chosen i.i.d. randomly in $[0, \frac{N_R N_T}{2}]$ (as in Theorem 3). When white Gaussian noise waveforms are used, we choose $d_R = 0.5$ and $d_T = 0.5N_R$ as the parameters regarding locations of the transmit and receive antennas (as one of the cases in Theorem 1). The following parameters are used: $N_T = 6, N_R = 6, N_s = 37, N_f = 37$; $S_{\max} = 20$ and the actual number of targets is $S = S_{\max}/2, S_{\max}, 2S_{\max}$. Furthermore, the bandwidth B is chosen to be 5 MHz, thus the sampling rate $\Delta_s = 10^{-7}$s and $T = 37 \times 10^{-7}$s. Yet, we emphasize that in practice one would choose a much larger value for T. Moreover, here we consider a single-pulse experiment (as this is also the setting of our theoretical framework), whereas in practice one would integrate the received waveforms over several periods T, thus leading to a considerably better Doppler resolution. However, the main purpose of the simulations is to validate the theories developed in this chapter. A truly practical simulation is beyond the scope of this chapter, as it also would have to incorporate pulse repetition rates and other variables and restrictions, which however might in part conceal the essence of our theoretical results.

Figures 2–3 show comparisons between the detection capability of the proposed methods. Depicted are the probability of detection versus probability of false alarm for the aforementioned setup.

Figure 4 illustrates the estimation capability of the proposed methods. We depict the relative ℓ_2 error $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}$ averaged over 50 trials for different SNR values (ranging from 0 to 25 dB).

In the second set of simulations, we consider the offgrid SIMO radar. Theorem 4 concerns the noise free case. In this set of simulations, we are going to demonstrate the efficiency of the algorithm with respect to different SNR. We test our algorithm in a discrete setting with $\rho = 3$ times finer stepsize in both range and angle, that is $\Delta_{\tau'} = \frac{\Delta_\tau}{\rho}$ and $\Delta_{\beta'} = \frac{\Delta_\beta}{\rho}$. In order to do that, we choose $N_T = 1, N_R = 10, N_s = 11, S = 6, 9, 12$. We choose the factor in the minimum separation condition to be 1, that is for any two targets, their ranges are separated by Δ_τ or their angles are separated by Δ_β. Figures 5 and 6 show the detection capability of the proposed method. Depicted are

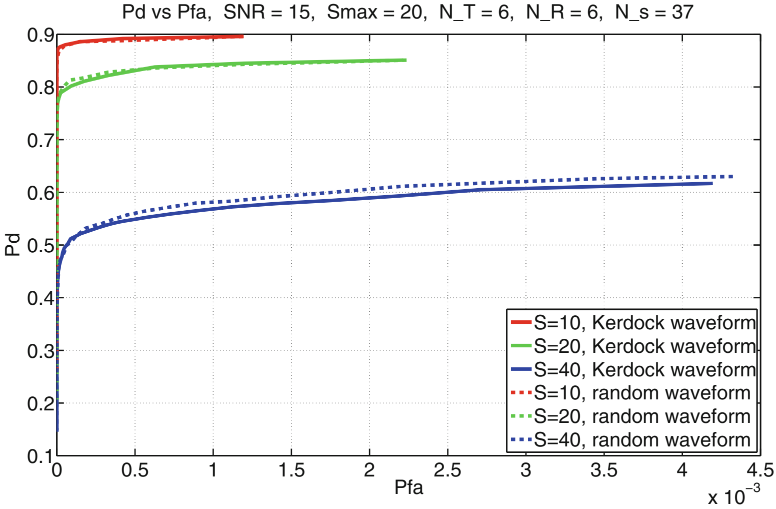


Fig. 2 ROC comparisons between Kerdock waveforms and random waveforms for different sparsity. The SNR level is 15 dB

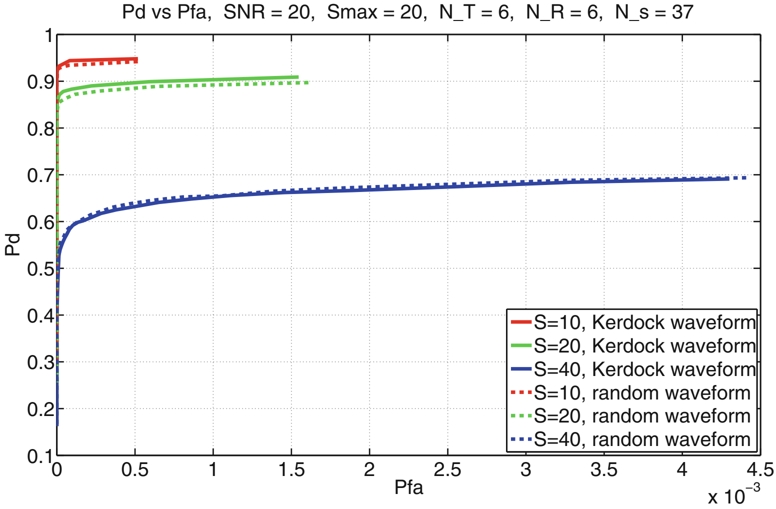


Fig. 3 ROC comparisons between Kerdock waveforms and random waveforms for different sparsity. The SNR level is 20 dB

the probability of detection versus probability of false alarm for the aforementioned setup. Figure 7 illustrates the relative ℓ_2 error $\frac{\|\mathbf{x}-\hat{\mathbf{x}}\|_2}{\|\mathbf{x}\|_2}$ averaged over 50 trials for different SNR values (ranging from 15 to 30 dB).

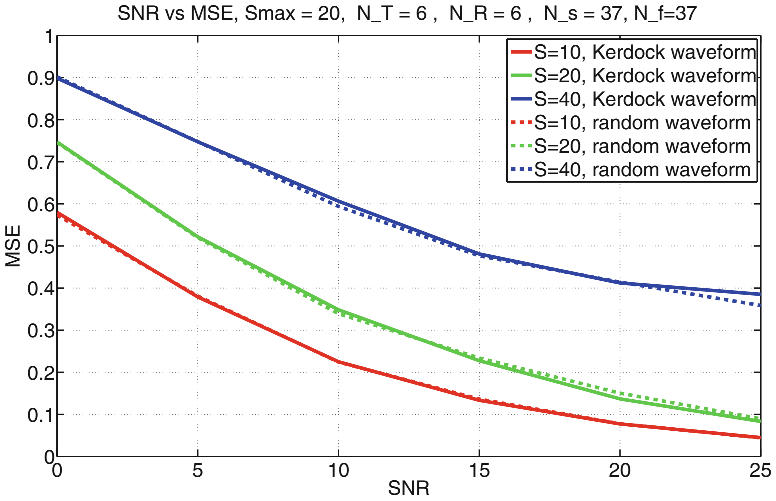


Fig. 4 Relative mean square error versus SNR for MIMO radar employing Kerdock waveforms and random waveforms

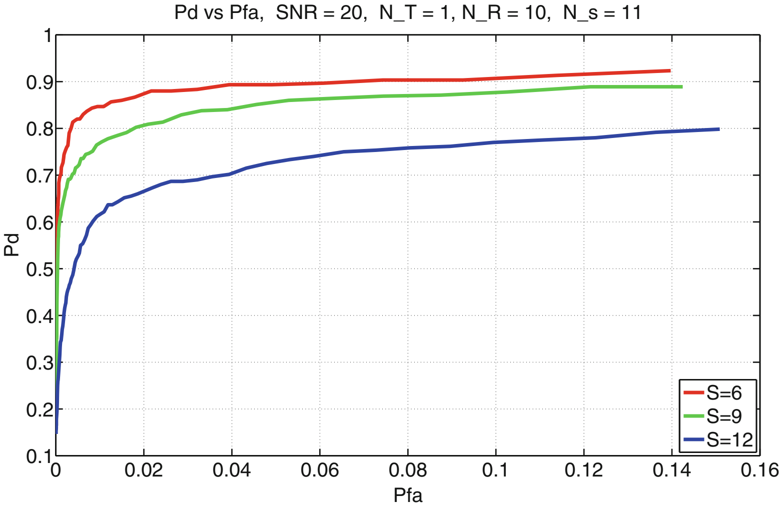


Fig. 5 Probability of detection versus probability of false alarm for SIMO radar with $\rho = 3$ times finer stepsize using random waveform for different sparsity. The SNR level is 20 dB

Acknowledgements T. Strohmer and H. Wang acknowledge support from the NSF via grant DTRA-DMS 1042939, and from DARPA via grant N66001-11-1-4090.

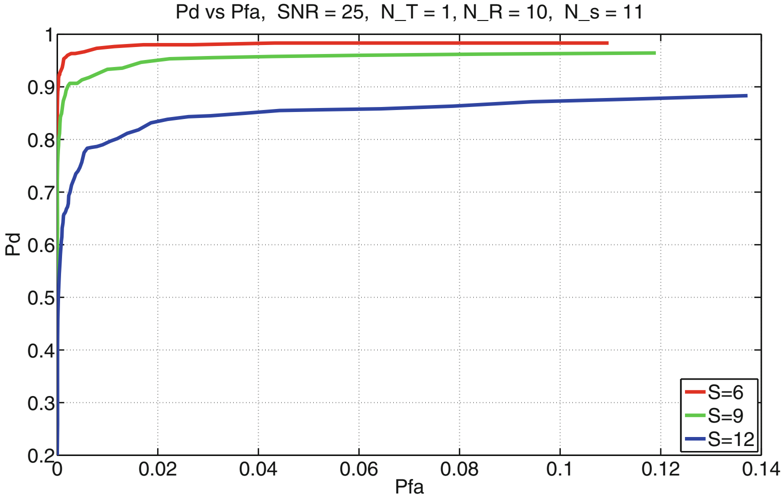


Fig. 6 Probability of detection versus probability of false alarm for SIMO radar with $\rho = 3$ times finer stepsize using random waveform for different sparsity. The SNR level is 25 dB

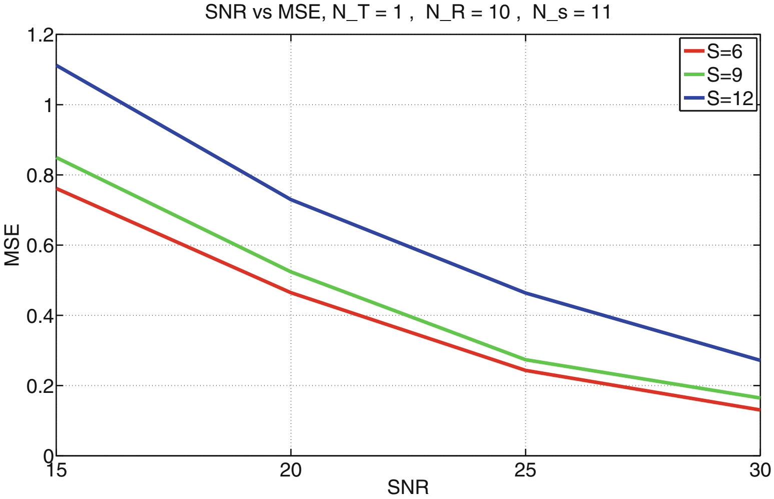


Fig. 7 Relative mean square error versus SNR for SIMO radar with $\rho = 3$ times finer stepsize using random waveform for different sparsity

Appendix A

In this appendix we collect some auxiliary results.

Lemma 7. *[38, Proposition 34] Let $\mathbf{x} \in \mathbb{C}^n$ be a vector with $x_k \sim \mathscr{CN}(0,\sigma^2)$, then for every $t > 0$ one has*

$$\mathbb{P}\Big(\|\mathbf{x}\|_2 - \mathbb{E}\|\mathbf{x}\|_2 > t\Big) \le e^{-\frac{t^2}{2\sigma^2}}. \tag{113}$$

The following lemma is a rescaled version of Lemma 3.1 in [29].

Lemma 8. *Let $\mathbf{A} \in \mathbb{C}^{n\times m}$ be a Gaussian random matrix with $A_{i,j} \sim \mathscr{CN}(0,\sigma^2)$. Then for all $\mathbf{x},\mathbf{y} \in \mathbb{C}^m$ with $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = \sqrt{m}$ and all $t > 0$*

$$\mathbb{P}\Big\{|\frac{\sigma^2}{n}\langle \mathbf{Ax},\mathbf{Ay}\rangle - \langle \mathbf{x},\mathbf{y}\rangle| > tm\Big\} \le 2\exp\Big(-n\frac{t^2}{C_1 + C_2 t}\Big), \tag{114}$$

with $C_1 = \frac{4e}{\sqrt{6\pi}}$ and $C_2 = \sqrt{8}e$.

For convenience we state the following version of Bernstein's inequality, which will be used in the proof of Lemma 10.

Lemma 9 (See e.g. [37]). *Let $X_1,\dots,X_n$ be independent random varibles with zero mean such that*

$$\mathbb{E}|X_i|^p \le \frac{1}{2}p!K^{p-2}v_i, \qquad \text{for all } i = 1,\dots,n; p \in \mathbb{N}, p \ge 2, \tag{115}$$

for some constants $K > 0$ and $v_i > 0, i = 1,\dots,n$. Then, for all $t > 0$

$$\mathbb{P}\Big(|\sum_{i=1}^n X_i| \ge t\Big) \le 2\exp\Big(-\frac{t^2}{2v + Kt}\Big), \tag{116}$$

where $v := \sum_{i=1}^n v_i$.

We also need the following deviation inequality for unbounded random variables. It is a complex-valued and slightly sharpened version of Lemma 6 in [13]. Our proof strategy differs at certain steps from that of Lemma 6 in [13] (and our proof is a bit shorter).

Lemma 10. *Let x_i and y_i, $i = 1,\dots,n$, be sequences of i.i.d. complex Gaussian random variables with variance σ. Then,*

$$\mathbb{P}\Big(|\sum_{i=1}^n \bar{x}_i y_i| > t\Big) \le 2\exp\big(-\frac{t^2}{\sigma^2(n\sigma^2 + 2t)}\big). \tag{117}$$

Proof. In order to apply Bernstein's inequality, we need to compute the moments $\mathbb{E}|x_i y_i|^p$. Since x_i and y_i are independent, there holds

$$\mathbb{E}(|x_i y_i|^p) = \mathbb{E}(|x_i|^p)\mathbb{E}(|y_i|^p) = (\mathbb{E}(|x_i|^p))^2. \tag{118}$$

The moments of x_i are well-known:

$$\mathbb{E}|x_i|^{2p} = p!\,\sigma^{2p}, \tag{119}$$

hence

$$(\mathbb{E}|x_i|^{2p})^2 = (2p!)^2(\sigma^{2p})^2 \le \frac{1}{4}(2p)!(\sigma^2)^{2p} \le \frac{1}{2}(2p)!(\sigma^2)^{2p-2}\frac{(\sigma^2)^2}{2}. \tag{120}$$

We apply Bernstein's inequality (116) with $K = \sigma^2$ and $v_i = \frac{(\sigma^2)^2}{2}, i = 1, \ldots, n$ and obtain (117).

Lemma 11. *Suppose M is an $m \times m$ matrix, α and β are two joint independent random vectors in $\mathbb{C}^m$ with zero means and $|\alpha_k| = |\beta_k| = 1$ for $k = 1, \ldots, m$. If n is a positive constant, then for any $t > 0$ and $s > 0$,*

1. if $|m_{kj}| \le \frac{1}{\sqrt{n}}$ for all k, j, then

$$\mathbb{P}\Big(|\langle M\alpha, \beta\rangle| \le mt\Big) \ge 1 - 4m\exp\Big(-\frac{t^2}{4\frac{m}{n}}\Big). \tag{121}$$

and

$$\mathbb{P}\Big(|\langle M\alpha, \alpha\rangle| \le 2mt\Big) \ge 1 - 8m\exp\Big(-\frac{t^2}{2\frac{m}{n}}\Big), \tag{122}$$

2. if $|m_{kj}| \le \frac{1}{\sqrt{n}}$ for $k \ne j$ and $m_{jj} = 1$, then

$$\mathbb{P}\Big(|\langle M\alpha, \beta\rangle| \le s + mt\Big) \ge 1 - 4\exp\Big(-\frac{s^2}{4m}\Big) - 4m\exp\Big(-\frac{t^2}{4\frac{m}{n}}\Big), \tag{123}$$

and

$$\mathbb{P}\Big(m(1-2t) \le |\langle M\alpha, \alpha\rangle| \le m(1+2t)\Big) \ge 1 - 8m\exp\Big(-\frac{t^2}{2\frac{m}{n}}\Big). \tag{124}$$

Proof. Note that

$$\begin{aligned}\langle M\alpha, \beta\rangle &= \sum_{k,j=1}^{m} m_{kj}\alpha_j\bar{\beta}_k \\ &= \sum_{l=1}^{m}\sum_{j=1}^{m} m_{j\oplus l, j}\alpha_j\bar{\beta}_{j\oplus l},\end{aligned}$$

where $\oplus$ denotes addition modulo m.

Let us first assume that $|m_{kj}| \le \frac{1}{\sqrt{n}}$.

Since α and β are jointly independent, then for any l, the entries in $\sum_{j=1}^{m} m_{j\oplus l,j}$ $\alpha_j \bar{\beta}_{j\oplus l}$ are all jointly independent and it is easy to check that $\mathbb{E}(m_{j\oplus l,j}\alpha_j\bar{\beta}_{j\oplus l}) = 0$ and $|m_{j\oplus l,j}\alpha_j\bar{\beta}_{j\oplus l}| = |m_{j\oplus l,j}|$, then Theorem 4.5 in [18] will give

$$\mathbb{P}\Big(|\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\beta}_{j\oplus l}| \le t\Big) \ge 1 - 4\exp\Big(-\frac{t^2}{4\sum_j |m_{j\oplus l,j}|^2}\Big)$$
$$\ge 1 - 4\exp\Big(-\frac{t^2}{4\frac{m}{n}}\Big). \tag{125}$$

We take all m different choices of l, then

$$\mathbb{P}\Big(|\sum_{l=1}^{m}\sum_{j=1}^{m} m_{i\oplus l,j}\alpha_j\bar{\beta}_{j\oplus l}| \le mt\Big) \ge 1 - 4m\exp\Big(-\frac{t^2}{4\frac{m}{n}}\Big), \tag{126}$$

which proves (121).

Now consider

$$\langle M\alpha, \alpha\rangle = \sum_{l=1}^{m}\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\alpha}_{j\oplus l},$$

but different from above, the entries in $\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\alpha}_{j\oplus l}$ are no longer all jointly independent. But similar to the proof of Theorem 5.1 in [27] and Lemma 3 in [31], we observe that for any l we can split the index set $1,\dots,m$ into two subsets $T_l^1, T_l^2 \subset \{1,\dots,m\}$, each of size $m/2$, such that the $m/2$ variables $\alpha_j\bar{\alpha}_{j\oplus l}$ are jointly independent for $j \in T_l^1$, and analogous for T_l^2. (For convenience we assume here that m is even, but with a negligible modification the argument also applies for odd m.) In other words, each of the sums $\sum_{j\in T_l^r} m_{j\oplus l,j}\alpha_j\bar{\alpha}_{j\oplus l}, r = 1,2$, contains only jointly independent terms.

So for each l,

$$\mathbb{P}\Big(|\sum_{j\in T_l^r} m_{j\oplus l,j}\alpha_j\bar{\alpha}_{j\oplus l}| \le t\Big) \ge 1 - 4\exp\Big(-\frac{t^2}{2\frac{m}{n}}\Big), \tag{127}$$

which implies that

$$\mathbb{P}\Big(|\sum_{j} m_{j\oplus l,j}\alpha_j\bar{\alpha}_{j\oplus l}| \le 2t\Big) \ge 1 - 8\exp\Big(-\frac{t^2}{2\frac{m}{n}}\Big), \tag{128}$$

Again, we take all m different choices of l, then

$$\mathbb{P}\Big(|\sum_{l=1}^{m}\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\alpha}_{j\oplus l}| \le 2mt\Big) \ge 1 - 8m\exp\Big(-\frac{t^2}{2\frac{m}{n}}\Big), \tag{129}$$

which proves (122).

Now let us assume that $|m_{kj}| \leq \frac{1}{\sqrt{n}}$ for $k \neq j$ and $m_{jj} = 1$.

$$\begin{aligned}\langle M\alpha, \beta\rangle &= \sum_{j=1}^{m} m_{jj}\alpha_j\bar{\beta}_j + \sum_{l=1}^{m-1}\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\beta}_{j\oplus l} \\ &= \sum_{j=1}^{m} \alpha_j\bar{\beta}_j + \sum_{l=1}^{m-1}\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\beta}_{j\oplus l}.\end{aligned}$$

Since α and β are joint independent and $|\alpha_j\bar{\beta}_j| = 1$,

$$\mathbb{P}\Big(|\sum_{j=1}^{m}\alpha_j\bar{\beta}_j| \leq s\Big) \geq 1 - 4\exp\Big(-\frac{s^2}{4m}\Big). \tag{130}$$

Similar to the proof of (126) above, we have that

$$\mathbb{P}\Big(|\sum_{l=1}^{m-1}\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\beta}_{j\oplus l}| \leq (m-1)t\Big) \geq 1 - 4(m-1)\exp\Big(-\frac{t^2}{4\frac{m}{n}}\Big),$$

together with (130), it follows

$$\mathbb{P}\Big(|\langle M\alpha, \beta\rangle| \leq s + (m-1)t\Big) \geq 1 - 4\exp\Big(-\frac{s^2}{4m}\Big) - 4(m-1)\exp\Big(-\frac{t^2}{4\frac{m}{n}}\Big),$$

which proves (123).

Finally,

$$\langle M\alpha, \alpha\rangle = \sum_{j=1}^{m} m_{jj} + \sum_{l=1}^{m-1}\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\alpha}_{j\oplus l} = m + \sum_{l=1}^{m-1}\sum_{j=1}^{m} m_{j\oplus l,j}\alpha_j\bar{\alpha}_{j\oplus l},$$

then (124) results from similar proof as for (122) and the triangle inequality.

Appendix B

We consider a general linear system of equations $\Psi\mathbf{x} = \mathbf{y}$, where $\Psi \in \mathbb{C}^{n\times m}$, $\mathbf{x} \in \mathbb{C}^m$ and $n \leq m$. We introduce the following generic K-sparse model:

- The support $I \subset \{1, \ldots, m\}$ of the K nonzero coefficients of $\mathbf{x}$ is selected uniformly at random.
- The non-zero entries of $\mathrm{sgn}(\mathbf{x})$ form a Steinhaus sequence, i.e., $\mathrm{sgn}(\mathbf{x}_k) := \mathbf{x}_k/|\mathbf{x}_k|, k \in I$, is a complex random variable that is uniformly distributed on the unit circle.

The following theorem is a slightly extended version of Theorem 1.3 in [5], see [31] for its proof.

Theorem 5. *Given* $\mathbf{y} = \Psi\mathbf{x} + \mathbf{w}$, *where* Ψ *has all unit-*ℓ_2*-norm columns,* $\mathbf{x}$ *is drawn from the generic K-sparse model and* $\mathbf{w}_i \sim \mathscr{C}\mathscr{N}(0, \sigma^2)$. *Assume that*

$$\mu(\Psi) \leq \frac{C_0}{\log m}, \tag{131}$$

where $C_0 > 0$ *is a constant independent of* n, m. *Furthermore, suppose*

$$K \leq \frac{c_0 m}{\|\Psi\|_{op}^2 \log m} \tag{132}$$

for some constant $c_0 > 0$ *and that*

$$\min_{k \in I} |\mathbf{x}_k| > 8\sigma\sqrt{2\log m}. \tag{133}$$

Then the solution $\hat{\mathbf{x}}$ *to the debiased lasso computed with* $\lambda = 2\sigma\sqrt{2\log m}$ *obeys*

$$\operatorname{supp}(\hat{\mathbf{x}}) = \operatorname{supp}(\mathbf{x}), \tag{134}$$

and

$$\frac{\|\hat{\mathbf{x}} - \mathbf{x}\|_2}{\|\mathbf{x}\|_2} \leq \frac{\sigma\sqrt{3n}}{\|\mathbf{y}\|_2} \tag{135}$$

with probability at least

$$1 - 2m^{-1}(2\pi \log m + Km^{-1}) - \mathscr{O}(m^{-2\log 2}). \tag{136}$$

References

1. Becker S, Candes E, Grant M. Templates for convex cone problems with applications to sparse signal recovery. Math Program Comput. 2011;3(3):165–218.
2. Calderbank AR, Cameron PJ, Kantor WM, Seidel JJ. Z_4-Kerdock codes, orthogonal spreads, and extremal Euclidean line-sets. Proc London Math Soc., (3). 1997;75(2):436–80.
3. Candès E, Fernandez-Granda C. Super-resolution from noisy data. J Fourier Anal Appl. 2013;19(6):1229–54.
4. Candès E, Fernandez-Granda C. Towards a mathematical theory of super-resolution. Commun Pure Appl Math. (to appear).
5. Candès EJ, Plan Y. Near-ideal model selection by ℓ_1 minimization. Ann Stat. 2009;37(5A):2145–77.
6. Carin L. On the relationship between compressive sensing and random sensor arrays. IEEE Antennas Propag Mag. 2009;51(5):72–81.
7. Cevher V, Boufounos PT, Baraniuk RG, Gilbert AC, Strauss MJ. Strauss. Near-optimal bayesian localization via incoherence and sparsity. Proceedings of the 2009 International Conference on Information Processing in Sensor Networks (IPSN), 13–16 April; 2009. p. 205–16.
8. Chi Y, Scharf LL, Pezeshki A, Calderbank AR. Sensitivity to basis mismatch in compressed sensing. IEEE Trans Signal Process. 2011;59(5):2182–95.

9. Fannjiang A, Liao W. Coherence pattern—guided compressive sensing with unresolved grids. SIAM J Imaging Sci. 2012;5:179–202.
10. Fenn AJ, Temme DH, Delaney WP, Courtney WE. The development of phased-array radar technology. Lincoln Lab J. 2000;12(2):321–40.
11. Friedlander B. Adaptive signal design for MIMO radar. In Li J, Stoica P, editors. MIMO radar signal processing, chapter 5. Wiley; 2009.
12. Friedlander B. On the relationship between MIMO and SIMO radars. IEEE Trans Signal Process. 2009;57(1):394–8.
13. Haupt J, Bajwa W, Raz G, Nowak R. Toeplitz compressed sensing matrices with applications to sparse channel estimation. IEEE Trans Inform Theory. 2010;56(11):5862–75.
14. Heath R, Strohmer T, Paulraj A. On quasi-orthogonal signatures for CDMA systems. IEEE Trans Info Theory. 2006;52(3):1217-26.
15. Herman M, Strohmer T. High-resolution radar via compressed sensing. IEEE Trans Signal Process. 2009;57(6):2275–84.
16. Herman M, Strohmer T. General deviants: an analysis of perturbations in compressed sensing. IEEE J Sel Top Signal Process Special Issue Compress Sens. 2010;4(2):342–49.
17. Howard SD, Calderbank AR, Moran W. The finite Heisenberg-Weyl groups in radar and communications. EURASIP J Appl Signal Process. 2006;1–12:2006.
18. Hügel M., Rauhut H, Strohmer T. Remote sensing via ℓ_1-minimization. Found Comput Math. (to appear).
19. Inoue T, Heath RW. Kerdock codes for limited feedback mimo systems. Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing, 2008. p. 3113–6.
20. Kerdock A. Studies of low-rate binary codes (ph.d. thesis abstr.). IEEE Trans Inf Theory. 1972;18(2):316.
21. König H. Isometric embeddings of euclidean spaces into finite-dimensional ℓ_p-spaces. Banach Cent Publ. 1995;34:79–87.
22. Levenstein VI. Bounds on the maximal cardinality of a code with bounded modulus of the inner product. Soviet Math Dokl. 1982;25:526–31.
23. Li J., Stoica P. MIMO radar with colocated antennas: review of some recent work. IEEE Signal Process Mag. 2007;24(5):106–14.
24. Li J, Stoica P, editors. MIMO radar signal processing. Wiley; 2009.
25. Lo Y. A mathematical theory of antenna arrays with randomly spaced element. IEEE Trans Antennas Propag. 1964;12(3):257–68.
26. Lo Y. A probalistic approach to the problem of large antenna arrays. J Res Nat Bur Stand. 1964;68D(5):1011–9.
27. Pfander GE, Rauhut H, Tanner J. Identification of matrices having a sparse representation. IEEE Trans Signal Process. 2008;56(11):5376–88.
28. Potter LC, Ertin E, Parker JT, Cetin M. Sparsity and compressed sensing in radar imaging. Proc IEEE. 2010;98(6):1006–20.
29. Rauhut H, Schnass K, Vandergheynst P. Compressed sensing and redundant dictionaries. IEEE Trans Inf Theory. 2008;54(5):2210–9.
30. Rihaczek AW. High-resolution radar. Boston: Artech House; 1996. (originally published: McGraw-Hill, NY, 1969).
31. Strohmer T, Friedlander B. Analysis of sparse MIMO radar. Appl Comput Harmon Anal. 2014;37:361–88.
32. Strohmer T, Heath R. Grassmannian rames with applications to coding and communications. Appl Comput Harmon Anal. 2003;14(3):257–75.
33. Strohmer T, Wang H. Accurate imaging of moving targets via random sensor arrays and Kerdock codes. Inverse Prob. 2013;29(2013):085001.
34. Tang G, Bhaskar BN, Shah P, Recht B. Compressed sensing off the grid. Preprint, [arvix:1207.6053]; 2012.
35. Tang G, Bhaskar BN, Recht B. Sparse recovery over continuous dictionaries: Just discretize. Asilomar Conference Signals, Systems, Computers, Asilomar; 2013.

36. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Statist Soc Ser B. 1996;58(1):267–88.
37. van der Vaart AW, Wellner JA. Weak convergence and empirical processes. Springer Series in Statistics. New York: Springer-Verlag; 1996. (With applications to statistics).
38. Vershynin R. Introduction to the non-asymptotic analysis of random matrices. In Eldar CY, Kutyniok G, editors, Compressed sensing: theory and applications. Cambridge University Press; 2012.
39. Wootters WK, Fields BD. Optimal state-determination by mutually unbiased measurements. Ann Phys. 191(2):363–81, 1989.

Applied and Numerical Harmonic Analysis Consists of 67 Titles

A. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences* (ISBN 978-0-8176-3924-2)

C.E. D'Attellis and E.M. Fernandez-Berdaguer: *Wavelet Theory and Harmonic Analysis in Applied Sciences* (ISBN 978-0-8176-3953-2)

H.G. Feichtinger and T. Strohmer: *Gabor Analysis and Algorithms* (ISBN 978-0-8176-3959-4)

R. Tolimieri and M. An: *Time-Frequency Representations* (ISBN 978-0-8176-3918-1)

T.M. Peters and J.C. Williams: *The Fourier Transform in Biomedical Engineering* (ISBN 978-0-8176-3941-9)

G.T. Herman: *Geometry of Digital Spaces* (ISBN 978-0-8176-3897-9)

A. Teolis: *Computational Signal Processing with Wavelets* (ISBN 978-0-8176-3909-9)

J. Ramanathan: *Methods of Applied Fourier Analysis* (ISBN 978-0-8176-3963-1)

J.M. Cooper: *Introduction to Partial Differential Equations with MATLAB* (ISBN 978-0-8176-3967-9)

A. Procházka, N.G. Kingsbury, P.J. Payner, and J. Uhlir: *Signal Analysis and Prediction* (ISBN 978-0-8176-4042-2)

W. Bray and C. Stanojevic: *Analysis of Divergence* (ISBN 978-1-4612-7467-4)

G.T. Herman and A. Kuba: *Discrete Tomography* (ISBN 978-0-8176-4101-6)

K. Gröchenig: *Foundations of Time-Frequency Analysis* (ISBN 978-0-8176-4022-4)

L. Debnath: *Wavelet Transforms and Time-Frequency Signal Analysis* (ISBN 978-0-8176-4104-7)

J.J. Benedetto and P.J.S.G. Ferreira: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)

D.F. Walnut: *An Introduction to Wavelet Analysis* (ISBN 978-0-8176-3962-4)

A. Abbate, C. DeCusatis, and P.K. Das: *Wavelets and Subbands* (ISBN 978-0-8176-4136-8)

O. Bratteli, P. Jorgensen, and B. Treadway: *Wavelets Through a Looking Glass* (ISBN 978-0-8176-4280-8)

H.G. Feichtinger and T. Strohmer: *Advances in Gabor Analysis* (ISBN 978-0-8176-4239-6)

O. Christensen: *An Introduction to Frames and Riesz Bases* (ISBN 978-0-8176-4295-2)

L. Debnath: *Wavelets and Signal Processing* (ISBN 978-0-8176-4235-8)

G. Bi and Y. Zeng: *Transforms and Fast Algorithms for Signal Analysis and Representations* (ISBN 978-0-8176-4279-2)

J.H. Davis: *Methods of Applied Mathematics with a MATLAB Overview* (ISBN 978-0-8176-4331-7)

J.J. Benedetto and A.I. Zayed: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)

E. Prestini: *The Evolution of Applied Harmonic Analysis* (ISBN 978-0-8176-4125-2)

L. Brandolini, L. Colzani, A. Iosevich, and G. Travaglini: *Fourier Analysis and Convexity* (ISBN 978-0-8176-3263-2)

W. Freeden and V. Michel: *Multiscale Potential Theory* (ISBN 978-0-8176-4105-4)

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3

O. Christensen and K.L. Christensen: *Approximation Theory* (ISBN 978-0-8176-3600-5)
O. Calin and D.-C. Chang: *Geometric Mechanics on Riemannian Manifolds* (ISBN 978-0-8176-4354-6)
J.A. Hogan: *TimeFrequency and TimeScale Methods* (ISBN 978-0-8176-4276-1)
C. Heil: *Harmonic Analysis and Applications* (ISBN 978-0-8176-3778-1)
K. Borre, D.M. Akos, N. Bertelsen, P. Rinder, and S.H. Jensen: *A Software-Defined GPS and Galileo Receiver* (ISBN 978-0-8176-4390-4)
T. Qian, M.I. Vai, and Y. Xu: *Wavelet Analysis and Applications* (ISBN 978-3-7643-7777-9)
G.T. Herman and A. Kuba: *Advances in Discrete Tomography and Its Applications* (ISBN 978-0-8176-3614-2)
M.C. Fu, R.A. Jarrow, J.-Y. Yen, and R.J. Elliott: *Advances in Mathematical Finance* (ISBN 978-0-8176-4544-1)
O. Christensen: *Frames and Bases* (ISBN 978-0-8176-4677-6)
P.E.T. Jorgensen, J.D. Merrill, and J.A. Packer: *Representations, Wavelets, and Frames* (ISBN 978-0-8176-4682-0)
M. An, A.K. Brodzik, and R. Tolimieri: *Ideal Sequence Design in Time-Frequency Space* (ISBN 978-0-8176-4737-7)
S.G. Krantz: *Explorations in Harmonic Analysis* (ISBN 978-0-8176-4668-4)
B. Luong: *Fourier Analysis on Finite Abelian Groups* (ISBN 978-0-8176-4915-9)
G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 1* (ISBN 978-0-8176-4802-2)
C. Cabrelli and J.L. Torrea: *Recent Developments in Real and Harmonic Analysis* (ISBN 978-0-8176-4531-1)
M.V. Wickerhauser: *Mathematics for Multimedia* (ISBN 978-0-8176-4879-4)
B. Forster, P. Massopust, O. Christensen, K. Gröchenig, D. Labate, P. Vandergheynst, G. Weiss, and Y. Wiaux: *Four Short Courses on Harmonic Analysis* (ISBN 978-0-8176-4890-9)
O. Christensen: *Functions, Spaces, and Expansions* (ISBN 978-0-8176-4979-1)
J. Barral and S. Seuret: *Recent Developments in Fractals and Related Fields* (ISBN 978-0-8176-4887-9)
O. Calin, D.-C. Chang, and K. Furutani, and C. Iwasaki: *Heat Kernels for Elliptic and Sub-elliptic Operators* (ISBN 978-0-8176-4994-4)
C. Heil: *A Basis Theory Primer* (ISBN 978-0-8176-4686-8)
J.R. Klauder: *A Modern Approach to Functional Integration* (ISBN 978-0-8176-4790-2)
J. Cohen and A.I. Zayed: *Wavelets and Multiscale Analysis* (ISBN 978-0-8176-8094-7)
D. Joyner and J.-L. Kim: *Selected Unsolved Problems in Coding Theory* (ISBN 978-0-8176-8255-2)
G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 2* (ISBN 978-0-8176-4943-2)
J.A. Hogan and J.D. Lakey: *Duration and Bandwidth Limiting* (ISBN 978-0-8176-8306-1)
G. Kutyniok and D. Labate: *Shearlets* (ISBN 978-0-8176-8315-3)
P.G. Casazza and P. Kutyniok: *Finite Frames* (ISBN 978-0-8176-8372-6)
V. Michel: *Lectures on Constructive Approximation* (ISBN 978-0-8176-8402-0)
D. Mitrea, I. Mitrea, M. Mitrea, and S. Monniaux: *Groupoid Metrization Theory* (ISBN 978-0-8176-8396-2)
T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 1* (ISBN 978-0-8176-8375-7)
T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 2* (ISBN 978-0-8176-8378-8)
D.V. Cruz-Uribe and A. Fiorenza: *Variable Lebesgue Spaces* (ISBN 978-3-0348-0547-6)
W. Freeden and M. Gutting: *Special Functions of Mathematical (Geo-)Physics* (ISBN 978-3-0348-0562-9)
A. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences, Volume 2: Linear and Nonlinear Dynamics of Continuous Media* (ISBN 978-0-8176-3942-6)

S. Foucart and H. Rauhut: *A Mathematical Introduction to Compressive Sensing* (ISBN 978-0-8176-4947-0)

G. Herman and J. Frank: *Computational Methods for Three-Dimensional Microscopy Reconstruction* (ISBN 978-1-4614-9520-8)

A. Paprotny and M. Thess: *Realtime Data Mining: Self-Learning Techniques for Recommendation Engines* (ISBN 978-3-319-01320-6)

A. Zayed and G. Schmeisser: *New Perspectives on Approximation and Sampling Theory* (ISBN 978-3-319-08800-6)

R. Balan, M.J. Begué, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 3* (ISBN 978-3-319-13229-7)

For an up-to-date list of titles in ANHA, please visit http://www.springer.com/series/4968

Index

R. Balan et al. (eds.), *Excursions in Harmonic Analysis, Volume 3*,
Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-13230-3