Applied and Numerical Harmonic Analysis

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

Peter G. Casazza • Gitta Kutyniok
Editors

Finite Frames

Theory and Applications

Birkhäuser

Editors
Peter G. Casazza
Department of Mathematics
University of Missouri
Columbia, MO, USA

Gitta Kutyniok
Institut für Mathematik
Technische Universität Berlin
Berlin, Germany

ISBN 978-0-8176-8372-6 ISBN 978-0-8176-8373-3 (eBook)
DOI 10.1007/978-0-8176-8373-3
Springer New York Heidelberg Dordrecht London

Library of Congress Control Number: 2012948835

Mathematics Subject Classification (2010): 41A63, 42C15, 47A05, 94A12, 94A20

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.birkhauser-science.com)

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 the 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 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; and second, to determine which phenomena can be constructed from given classes of harmonics, as done, e.g., 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), 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 deal-

ing 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

Frame theory is nowadays a fundamental research area in mathematics, computer science, and engineering with many exciting applications in a variety of different fields. Introduced in 1952 by Duffin and Schaeffer, its significance for signal processing has been revealed in the pioneering work by Daubechies, Grossman, and Meyer in 1986. Since then frame theory has quickly become the key approach whenever redundant, yet stable, representations of data are required. Frames in finite-dimensional spaces, i.e., finite frames, are a very important class of frames due to their significant relevance in applications. This book is the first comprehensive introduction to both the theory and applications of finite frames, with various chapters outlining diverse directions of this intriguing research area.

Today, frame theory provides an extensive framework for the analysis and decomposition of signals in a stable and redundant way, accompanied by various reconstruction procedures. Its main methodological ingredients are the representation systems which form a frame. In fact, a frame can be regarded as the most natural generalization of the notion of an orthonormal basis. To be more specific, let $(\varphi_i)_{i=1}^M$ be a family of vectors in $\mathbb{R}^N$ or $\mathbb{C}^N$. Then these vectors form a frame if there exist constants $0 < A \leq B < \infty$ such that $A\|x\|_2 \leq \|(\langle x, \varphi_i\rangle)_{i=1}^M\|_{\ell_2} \leq B\|x\|_2$ holds for all x in the underlying space. The constants A and B determine the condition of a frame, which is optimal for $A = B = 1$, leading to the class of Parseval frames. It is evident that the notion of a frame allows the inclusion of redundant systems in the sense of overcomplete systems. This is key to the resilience of frames to disturbances (such as, e.g., noise, erasures, and quantization) of the frame coefficients $(\langle x, \varphi_i\rangle)_{i=1}^M$ associated with a signal x. These frame coefficients can be utilized, for instance, for edge detection in an image, for the transmission of a speech signal, or for recovery of missing data. Although the analysis operator $x \mapsto (\langle x, \varphi_i\rangle)_{i=1}^M$ maps a signal into a higher-dimensional space, frame theory also provides efficient methods for reconstructing the signal.

New theoretical insights and novel applications are continually arising, because the underlying principles of frame theory are basic ideas which are fundamental to a wide canon of areas of research. In this sense, frame theory might be regarded as partly belonging to applied harmonic analysis, functional analysis, and operator

theory, as well as numerical linear algebra and matrix theory. Some of its countless applications are in biology, geophysics, imaging sciences, quantum computing, speech recognition, and wireless communication, to name a few.

In this book we depict the current state of the research in finite frame theory and cover the progress which has been made in the area over the last twenty years. It is suitable for both a researcher who is interested in the latest developments in finite frame theory, and also for a graduate student who seeks an introduction into this exciting research area.

This book comprises (in addition to the introduction) twelve chapters, which cover a variety of topics in the theory and applications of finite frames written by well-known leading experts in the subject. The necessary background for the subsequent chapters is provided in the comprehensive introductory chapter on finite frame theory. The twelve chapters can be divided into four topical groups: Frame properties (Chaps. 2–4), special classes of frames (Chaps. 5 and 6), applications of frames (Chaps. 7–11), and extensions of the concept of a frame (Chaps. 12 and 13). Every chapter contains the current state of its respective field and can be read independently of the others. We now provide a brief summary of the content of each chapter.

Chapter 1 provides a comprehensive introduction to the basics of finite frame theory. After answering the question *why frames?*, background material from Hilbert space theory and operator theory is presented. The authors then introduce the basics of finite frame theory and the operators connected with a frame. After this preparation the reader is equipped with an overview of well-known results on the reconstruction of signals, the construction of special frames, frame properties, and applications.

Chapter 2 deals with constructing frames with prescribed lengths of the frame vectors and a prescribed spectrum of the frame operator. Several years of research have now led to a complete solution of this problem, which is presented in this chapter. The authors show in great detail how methods stemming from the Spectral Tetris algorithm can be utilized to achieve an algorithmic solution to the problem.

Chapter 3 is devoted to the problem of partitioning a frame into a minimal number of linearly independent or a maximal number of spanning subsets. A direct application of the Rado-Horn theorem would solve the first problem, but it is much too inefficient and does not make use of frame properties. However, the authors improve the Rado-Horn theorem and derive various results solving the problem in special cases using frame properties.

Chapter 4 accommodates the fact that (besides analytic and algebraic properties) frames can also be analyzed from a geometric standpoint. Accompanied by several examples, it is shown how methods from algebraic geometry can be successfully exploited to obtain local coordinate systems on the algebraic variety of frames with prescribed frame operator and frame vector lengths. After that, angles and metrics on the Grassmannian variety are defined. They are then used to prove that the generic Parseval frames are dense in the class of Parseval frames. The chapter ends with a survey of results on signal reconstruction without phase from an algebraic geometry viewpoint.

Chapter 5 establishes a connection between finite group theory and finite frame theory. The frames of investigation are called group frames. These are frames which are induced by unitary group actions on the underlying Hilbert space; harmonic frames are a special class of group frames. One of the highlights of the chapter is the utilization of group frames to construct equiangular frames, which are most desirable in applications due to their resilience to erasures.

Chapter 6 provides a basic self-contained introduction to Gabor frames on finite Abelian groups. In the first half of the chapter the main ideas of Gabor analysis in signal processing are illuminated, and fundamental results for Gabor frames are proved. The second half deals with geometric properties such as linear independence, coherence, and the restricted isometry property for Gabor synthesis matrices, which then gives rise to the utilization of Gabor frames in compressed sensing.

Chapter 7 studies the suitability of frames for signal recovery from encoded, noisy, or erased data with controllable accuracy. After providing a survey of results on the resilience of frames with respect to noisy measurements, the author analyzes the effects of erasures and error correction. One main result states that equiangular and random Parseval frames are optimally robust against such disturbances.

Chapter 8 considers frame quantization, which is essential for the process of digitizing analog signals. The authors introduce the reader to the ideas and principles of memoryless scalar quantization as well as to first order and higher order sigma-delta quantization algorithms, and discuss their performance in terms of the reconstruction error. In particular, it is shown that an appropriate choice of quantization scheme and encoding operator leads to an error decaying exponentially with the oversampling rate.

Chapter 9 surveys recent work on sparse signal processing which has become a novel paradigm in the last year. The authors address problems such as exact or lossy recovery, estimation, regression, and support detection of sparse signals in both the deterministic and probabilistic regimes. The significance of frames for this methodological approach is, for instance, shown by revealing the special role of equal norm tight frames for detecting the presence of a sparse signal in noise.

Chapter 10 considers the connection of finite frames and filter banks. After the introduction of basic related operations, such as convolution, downsampling, the discrete Fourier transform, and the Z-transform, the polyphase representation of filter banks is proved to hold, and its properties and advantages are discussed. Thereafter, the authors show how various desiderata for the frame connected with a filter bank can be realized.

Chapter 11 is split into two parts. The first part presents a variety of conjectures stemming from diverse areas of research in pure and applied mathematics as well as engineering. Intriguingly, all these conjectures are equivalent to the famous 1959 Kadison-Singer problem. The second part is devoted to the Paulsen problem, which is formulated in pure frame theoretical terms and is also still unsolved.

Chapter 12 presents one generalization of frames, called probabilistic frames. The collection of these frames is a set of probability measures which contains the usual finite frames as point measures. The authors present the basic properties of probabilistic frames and survey a range of areas such as directional statistics, in which this concept implicitly appears.

Chapter 13 introduces fusion frames, which are a generalization of frames designed for and perfectly suited to model distributed processing. They analyze signals by projecting them onto multidimensional subspaces, in contrast to frames which consider only one-dimensional projections. Various results are reviewed, including fusion frame constructions, sparse recovery from fusion frame measurements, and specific applications of fusion frames.

The first editor thanks Janet Tremain for her unending support and help during the preparation of this book.

Columbia, MO, USA — Peter G. Casazza
Berlin, Germany — Gitta Kutyniok

Contents

Contributors

Waheed U. Bajwa Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, Piscataway, NJ, USA

Bernhard G. Bodmann Mathematics Department, University of Houston, Houston, TX, USA

Jameson Cahill Department of Mathematics, University of Missouri, Columbia, MO, USA

Peter G. Casazza Department of Mathematics, University of Missouri, Columbia, MO, USA

Martin Ehler Institute of Biomathematics and Biometry, Helmholtz Zentrum München, Neuherberg, Germany

Matthew Fickus Department of Mathematics, Air Force Institute of Technology, Wright-Patterson AFB, OH, USA

Gitta Kutyniok Institut für Mathematik, Technische Universität Berlin, Berlin, Germany

Melody L. Massar Department of Mathematics, Air Force Institute of Technology, Wright-Patterson AFB, OH, USA

Dustin G. Mixon Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ, USA

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

Ali Pezeshki Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO, USA

Götz E. Pfander School of Engineering and Science, Jacobs University, Bremen, Germany

Friedrich Philipp Institut für Mathematik, Technische Universität Berlin, Berlin, Germany

Miriam J. Poteet Department of Mathematics, Air Force Institute of Technology, Wright-Patterson AFB, OH, USA

Alexander M. Powell Department of Mathematics, Vanderbilt University, Nashville, TN, USA

Rayan Saab Department of Mathematics, Duke University, Durham, NC, USA

Darrin Speegle Department of Mathematics and Computer Science, Saint Louis University, St. Louis, MO, USA

Nate Strawn Department of Mathematics, Duke University, Durham, NC, USA

Shayne Waldron Department of Mathematics, University of Auckland, Auckland, New Zealand

Özgür Yılmaz Department of Mathematics, University of British Columbia, Vancouver, BC, Canada

Chapter 1
Introduction to Finite Frame Theory

Peter G. Casazza, Gitta Kutyniok, and Friedrich Philipp

Abstract To date, frames have established themselves as a standard notion in applied mathematics, computer science, and engineering as a means to derive redundant, yet stable decompositions of a signal for analysis or transmission, while also promoting sparse expansions. The reconstruction procedure is then based on one of the associated dual frames, which—in the case of a Parseval frame—can be chosen to be the frame itself. In this chapter, we provide a comprehensive review of the basics of finite frame theory upon which the subsequent chapters are based. After recalling some background information on Hilbert space theory and operator theory, we introduce the notion of a frame along with some crucial properties and construction procedures. Then we discuss algorithmic aspects such as basic reconstruction algorithms and present brief introductions to diverse applications and extensions of frames. The subsequent chapters of this book will then extend key topics in many intriguing directions.

Keywords Applications of finite frames · Construction of frames · Dual frames · Frames · Frame operator · Grammian operator · Hilbert space theory · Operator theory · Reconstruction algorithms · Redundancy · Tight frames

1.1 Why Frames?

The Fourier transform has been a major tool in analysis for over 100 years. However, it solely provides frequency information, and hides (in its phases) information concerning the moment of emission and duration of a signal. D. Gabor resolved this

P.G. Casazza
Mathematics Department, University of Missouri, Columbia, MO 65211, USA
e-mail: casazzap@missouri.edu

G. Kutyniok (✉) · F. Philipp
Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany
e-mail: kutyniok@math.tu-berlin.de

F. Philipp
e-mail: philipp@math.tu-berlin.de

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_1,

problem in 1946 [92] by introducing a fundamental new approach to signal decomposition. Gabor's approach quickly became the paradigm for this area, because it provided resilience to additive noise, quantization, and transmission losses as well as an ability to capture important signal characteristics. Unbeknownst to Gabor, he had discovered the fundamental properties of a frame without any of the formalism. In 1952, Duffin and Schaeffer [79] were studying some deep problems in nonharmonic Fourier series for which they required a formal structure for working with highly overcomplete families of exponential functions in $L^2[0, 1]$. For this, they introduced the notion of a *Hilbert space frame*, in which Gabor's approach is now a special case, falling into the area of *time-frequency analysis* [97]. Much later—in the late 1980s—the fundamental concept of frames was revived by Daubechies, Grossman and Mayer [76] (see also [75]), who showed its importance for data processing.

Traditionally, frames were used in signal and image processing, nonharmonic Fourier series, data compression, and sampling theory. But today, frame theory has ever-increasing applications to problems in both pure and applied mathematics, physics, engineering, and computer science, to name a few. Several of these applications will be investigated in this book. Since applications mainly require frames in finite-dimensional spaces, this will be our focus. In this situation, a frame is a spanning set of vectors—which are generally *redundant* (*overcomplete*), requiring control of its condition numbers. Thus a typical frame possesses more frame vectors than the dimension of the space, and each vector in the space will have infinitely many representations with respect to the frame. It is this *redundancy of frames* which is key to their significance for applications.

The role of redundancy varies depending on the requirements of the applications at hand. First, redundancy gives greater design *flexibility*, which allows frames to be constructed to fit a particular problem in a manner not possible by a set of linearly independent vectors. For instance, in areas such as quantum tomography, classes of orthonormal bases with the property that the modulus of the inner products of vectors from different bases are a constant are required. A second example comes from speech recognition, when a vector needs to be determined by the absolute value of the frame coefficients (up to a phase factor). A second major advantage of redundancy is *robustness*. By spreading the information over a wider range of vectors, resilience against losses (*erasures*) can be achieved. Erasures are, for instance, a severe problem in wireless sensor networks when transmission losses occur or when sensors are intermittently fading out, or in modeling the brain where memory cells are dying out. A further advantage of spreading information over a wider range of vectors is to mitigate the effects of noise in the signal.

These examples represent a tiny fraction of the theory and applications of frame theory that you will encounter in this book. New theoretical insights and novel applications are continually arising due to the fact that the underlying principles of frame theory are basic ideas which are fundamental to a wide canon of areas of research. In this sense, frame theory might be regarded as partly belonging to applied harmonic analysis, functional analysis, operator theory, numerical linear algebra, and matrix theory.

1.1.1 The Role of Decompositions and Expansions

Focusing on the finite-dimensional situation, let x be given data which we assume to belong to some real or complex N-dimensional Hilbert space $\mathscr{H}^N$. Further, let $(\varphi_i)_{i=1}^M$ be a representation system (i.e., a spanning set) in $\mathscr{H}^N$, which might be chosen from an existing catalog, designed depending on the type of data we are facing, or learned from sample sets of the data.

One common approach to data processing consists in the *decomposition* of the data x according to the system $(\varphi_i)_{i=1}^M$ by considering the map

$$x \mapsto \big(\langle x, \varphi_i \rangle\big)_{i=1}^M.$$

As we will see, the generated sequence $(\langle x, \varphi_i \rangle)_{i=1}^M$ belonging to $\ell_2(\{1, \dots, M\})$ can then be used, for instance, for transmission of x. Also, a careful choice of the representation system enables us to solve a variety of analysis tasks. As an example, under certain conditions the positions and orientations of edges of an image x are determined by those indices $i \in \{1, \dots, M\}$ belonging to the largest coefficients in magnitude $|\langle x, \varphi_i \rangle|$, i.e., by hard thresholding, in the case that $(\varphi_i)_{i=1}^M$ is a shearlet system (see [115]). Finally, the sequence $(\langle x, \varphi_i \rangle)_{i=1}^M$ allows compression of x, which is in fact the heart of the new JPEG2000 compression standard when choosing $(\varphi_i)_{i=1}^M$ to be a wavelet system [140].

An accompanying approach is the *expansion* of the data x by considering sequences $(c_i)_{i=1}^M$ satisfying

$$x = \sum_{i=1}^{M} c_i \varphi_i.$$

It is well known that suitably chosen representation systems allow sparse sequences $(c_i)_{i=1}^M$ in the sense that $\|c\|_0 = \#\{i : c_i \neq 0\}$ is small. For example, certain wavelet systems typically sparsify natural images in this sense (see, for example, [77, 122, 133] and the references therein). This observation is key to allowing the application of the abundance of existing sparsity methodologies such as compressed sensing [86] to x. In contrast to this viewpoint which assumes x as explicitly given, the approach of expanding the data is also highly beneficial in the case where x is only implicitly given, which is, for instance, the problem all partial differential equation (PDE) solvers face. Hence, using $(\varphi_i)_{i=1}^M$ as a generating system for the trial space, the PDE solver's task reduces to computing $(c_i)_{i=1}^M$, which is advantageous for deriving efficient solvers provided that—as before—a sparse sequence does exist (see, e.g., [73, 106]).

1.1.2 Beyond Orthonormal Bases

To choosing the representation system $(\varphi_i)_{i=1}^N$ to form an orthonormal basis for $\mathscr{H}^N$ is the standard choice. However, the linear independence of such a system causes a variety of problems for the aforementioned applications.

Starting with the *decomposition* viewpoint, using $(\langle x, \varphi_i \rangle)_{i=1}^N$ for transmission is far from being robust to erasures, since the erasure of only a single coefficient causes a true information loss. Also, for analysis tasks orthonormal bases can be unfavorable, since they do not allow any flexibility in design, which is needed, for instance, in the design of directional representation systems. In fact, it is conceivable that no orthonormal basis with paralleling properties such as curvelets or shearlets does exist.

Also, from an *expansion* point of view, the utilization of orthonormal bases is not advisable. A particular problem affecting sparsity methodologies as well as the utilization for PDE solvers is the uniqueness of the sequence $(c_i)_{i=1}^M$. This non-flexibility prohibits the search for a sparse coefficient sequence.

It is evident that these problems can be tackled by allowing the system $(\varphi_i)_{i=1}^M$ to be redundant. Certainly, numerical stability issues in the typical processing of data

$$x \mapsto \big(\langle x, \varphi_i \rangle\big)_{i=1}^M \mapsto \sum_{i=1}^M \langle x, \varphi_i \rangle \tilde{\varphi}_i \approx x$$

with an adapted system $(\tilde{\varphi}_i)_{i=1}^M$ must be taken into account. This leads naturally to the notion of a (*Hilbert space*) *frame*. The main idea is to have a controlled norm equivalence between the data x and the sequence of coefficients $(\langle x, \varphi_i \rangle)_{i=1}^M$.

The area of frame theory is very closely related to other research fields in both pure and applied mathematics. General (Hilbert space) frame theory—in particular, including the infinite-dimensional situation—intersects functional analysis and operator theory. It also bears close relations to the area of applied harmonic analysis, in which the design of representation systems, typically by a careful partitioning of the Fourier domain, is one major objective. Some researchers even consider frame theory as belonging to this area. Restricting to the finite-dimensional situation—in which customarily the term *finite frame theory* is used—the classical areas of matrix theory and numerical linear algebra have close intersections, but also, for instance, the novel area of compressed sensing, as already pointed out.

Nowadays, frames have established themselves as a standard notion in applied mathematics, computer science, and engineering. Finite frame theory deserves special attention due to its importance for applications, and might be even considered a research area of its own. This is also the reason why this book specifically focuses on finite frame theory. The subsequent chapters will show the diversity of this rich and vivid research area to date, ranging from the development of frameworks to analyzing specific properties of frames, the design of different classes of frames, various applications of frames, and extensions of the notion of a frame.

1.1.3 Outline

In Sect. 1.2 we first provide some background information on Hilbert space theory and operator theory to make this book self-contained. Frames are then subsequently introduced in Sect. 1.3, followed by a discussion of the four main operators associated with a frame, namely, the analysis, synthesis, frame, and Gramian operators (see Sect. 1.4). Reconstruction results and algorithms naturally including the notion of a dual frame are the focus of Sect. 1.5. This is followed by the presentation of different constructions of tight as well as non-tight frames (Sect. 1.6), and a discussion of some crucial properties of frames, in particular, their spanning properties, the redundancy of a frame, and equivalence relations among frames in Sect. 1.7. This chapter is concluded with brief introductions to diverse applications and extensions of frames (Sects. 1.8 and 1.9).

1.2 Background Material

Let us start by recalling some basic definitions and results from Hilbert space theory and operator theory, which will be required for all subsequent chapters. We do not include the proofs of the presented results; instead, we refer to the standard literature such as, for instance, [152] for Hilbert space theory and [70, 104, 129] for operator theory. We emphasize that all following results are solely stated in the finite-dimensional setting, which is the focus of this book.

1.2.1 Review of Basics from Hilbert Space Theory

Letting N be a positive integer, we denote by $\mathscr{H}^N$ a real or complex N-dimensional Hilbert space. This will be the space considered throughout this book. Sometimes, if it is convenient, we identify $\mathscr{H}^N$ with $\mathbb{R}^N$ or $\mathbb{C}^N$. By $\langle \cdot, \cdot \rangle$ and $\| \cdot \|$ we denote the inner product on $\mathscr{H}^N$ and its corresponding norm, respectively.

Let us now start with the origin of frame theory, which is the notion of an orthonormal basis. Alongside, we recall the basic definitions we will also require in the sequel.

Definition 1.1 A vector $x \in \mathscr{H}^N$ is called *normalized* if $\|x\| = 1$. Two vectors $x, y \in \mathscr{H}^N$ are called *orthogonal* if $\langle x, y \rangle = 0$. A system $(e_i)_{i=1}^k$ of vectors in $\mathscr{H}^N$ is called

(a) *complete* (or a *spanning set*) if $\operatorname{span}\{e_i\}_{i=1}^k = \mathscr{H}^N$.
(b) *orthogonal* if for all $i \neq j$, the vectors e_i and e_j are orthogonal.
(c) *orthonormal* if it is orthogonal and each e_i is normalized.
(e) *an orthonormal basis* for $\mathscr{H}^N$ if it is complete and orthonormal.

A fundamental result in Hilbert space theory is *Parseval's identity*.

Proposition 1.1 (Parseval's Identity) *If $(e_i)_{i=1}^N$ is an orthonormal basis for $\mathscr{H}^N$, then, for every $x \in \mathscr{H}^N$, we have*

$$\|x\|^2 = \sum_{i=1}^{N} \big|\langle x, e_i\rangle\big|^2.$$

Interpreting this identity from a signal processing point of view, it implies that the energy of the signal is preserved under the map $x \mapsto (\langle x, e_i\rangle)_{i=1}^N$, which we will later refer to as the analysis map. We also mention at this point that this identity is not only satisfied by orthonormal bases. In fact, redundant systems ("non-bases") such as $(e_1, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \ldots, \frac{1}{\sqrt{N}}e_N, \ldots, \frac{1}{\sqrt{N}}e_N)$ also satisfy this equality, and will later be coined *Parseval frames*.

Parseval's identity has the following implication, which shows that a vector x can be recovered from the coefficients $(\langle x, e_i\rangle)_{i=1}^N$ by a simple procedure. Thus, from an application point of view, this result can also be interpreted as a reconstruction formula.

Corollary 1.1 *If $(e_i)_{i=1}^N$ is an orthonormal basis for $\mathscr{H}^N$, then, for every $x \in \mathscr{H}^N$, we have*

$$x = \sum_{i=1}^{N} \langle x, e_i\rangle e_i.$$

Next, we present a series of basic identities and inequalities, which are exploited in various proofs.

Proposition 1.2 *Let $x, \tilde{x} \in \mathscr{H}^N$.*

(i) Cauchy-Schwarz inequality. *We have*

$$\big|\langle x, \tilde{x}\rangle\big| \le \|x\|\|\tilde{x}\|,$$

with equality if and only if $x = c\tilde{x}$ for some constant c.

(ii) Triangle inequality. *We have*

$$\|x + \tilde{x}\| \le \|x\| + \|\tilde{x}\|.$$

(iii) Polarization identity (real form). *If $\mathscr{H}^N$ is real, then*

$$\langle x, \tilde{x}\rangle = \frac{1}{4}\big[\|x + \tilde{x}\|^2 - \|x - \tilde{x}\|^2\big].$$

(iv) Polarization identity (complex form). *If $\mathscr{H}^N$ is complex, then*

$$\langle x, \tilde{x}\rangle = \frac{1}{4}\big[\|x + \tilde{x}\|^2 - \|x - \tilde{x}\|^2\big] + \frac{i}{4}\big[\|x + i\tilde{x}\|^2 - \|x - i\tilde{x}\|^2\big].$$

(v) Pythagorean theorem. *Given pairwise orthogonal vectors* $(x_i)_{i=1}^M \in \mathscr{H}^N$, *we have*

$$\left\| \sum_{i=1}^{M} x_i \right\|^2 = \sum_{i=1}^{M} \|x_i\|^2.$$

We next turn to considering subspaces in $\mathscr{H}^N$, again starting with the basic notation and definitions.

Definition 1.2 Let $\mathscr{W}, \mathscr{V}$ be subspaces of $\mathscr{H}^N$.

(a) A vector $x \in \mathscr{H}^N$ is called *orthogonal to* $\mathscr{W}$ (denoted by $x \perp \mathscr{W}$), if

$$\langle x, \tilde{x} \rangle = 0 \quad \text{for all } \tilde{x} \in \mathscr{W}.$$

The *orthogonal complement* of $\mathscr{W}$ is then defined by

$$\mathscr{W}^\perp = \left\{ x \in \mathscr{H}^N : x \perp \mathscr{W} \right\}.$$

(b) The subspaces $\mathscr{W}$ and $\mathscr{V}$ are called *orthogonal subspaces* (denoted by $\mathscr{W} \perp \mathscr{V}$), if $\mathscr{W} \subset \mathscr{V}^\perp$ (or, equivalently, $\mathscr{V} \subset \mathscr{W}^\perp$).

The notion of *orthogonal direct sums*, which will play an essential role in Chap. 13, can be regarded as a generalization of Parseval's identity (Proposition 1.1).

Definition 1.3 Let $(\mathscr{W}_i)_{i=1}^M$ be a family of subspaces of $\mathscr{H}^N$. Then their *orthogonal direct sum* is defined as the space

$$\left(\sum_{i=1}^{M} \oplus \mathscr{W}_i \right)_{\ell^2} := \mathscr{W}_1 \times \cdots \times \mathscr{W}_M$$

with inner product defined by

$$\langle x, \tilde{x} \rangle = \sum_{i=1}^{M} \langle x_i, \tilde{x}_i \rangle \quad \text{for all } x = (x_i)_{i=1}^M,\ \tilde{x} = (\tilde{x}_i)_{i=1}^M \in \left(\sum_{i=1}^{M} \oplus \mathscr{W}_i \right)_{\ell^2}.$$

The extension of Parseval's identity can be seen when choosing $\tilde{x} = x$ yielding $\|x\|^2 = \sum_{i=1}^M \|x_i\|^2$.

1.2.2 Review of Basics from Operator Theory

We next introduce the basic results from operator theory used throughout this book. We first recall that each linear operator has an associated matrix representation.

Definition 1.4 Let $T : \mathscr{H}^N \to \mathscr{H}^K$ be a linear operator, let $(e_i)_{i=1}^N$ be an orthonormal basis for $\mathscr{H}^N$, and let $(g_i)_{i=1}^K$ be an orthonormal basis for $\mathscr{H}^K$. Then the *matrix representation of* T (*with respect to the orthonormal bases* $(e_i)_{i=1}^N$ *and* $(g_i)_{i=1}^K$) is a matrix of size $K \times N$ and is given by $A = (a_{ij})_{i=1,j=1}^{K,N}$, where

$$a_{ij} = \langle T e_j, g_i \rangle.$$

For all $x \in \mathscr{H}^N$ with $c = (\langle x, e_i \rangle)_{i=1}^N$ we have

$$Tx = Ac.$$

1.2.2.1 Invertibility

We start with the following definition.

Definition 1.5 Let $T : \mathscr{H}^N \to \mathscr{H}^K$ be a linear operator.

(a) The *kernel* of T is defined by $\ker T := \{x \in \mathscr{H}^N : Tx = 0\}$. Its *range* is $\operatorname{ran} T := \{Tx : x \in \mathscr{H}^N\}$, sometimes also called the *image* and denoted by $\operatorname{im} T$. The *rank of* T, $\operatorname{rank} T$, is the dimension of the range of T.
(b) The operator T is called *injective* (or *one-to-one*), if $\ker T = \{0\}$, and *surjective* (or *onto*), if $\operatorname{ran} T = \mathscr{H}^K$. It is called *bijective (or invertible)*, if T is both injective and surjective.
(c) The *adjoint operator* $T^* : \mathscr{H}^K \to \mathscr{H}^N$ is defined by

$$\langle Tx, \tilde{x} \rangle = \langle x, T^* \tilde{x} \rangle \quad \text{for all } x \in \mathscr{H}^N \text{ and } \tilde{x} \in \mathscr{H}^K.$$

(d) The *norm* of T is defined by

$$\|T\| := \sup\{\|Tx\| : \|x\| = 1\}.$$

The next result states several relations between these notions.

Proposition 1.3

(i) *Let* $T : \mathscr{H}^N \to \mathscr{H}^K$ *be a linear operator. Then*

$$\dim \mathscr{H}^N = N = \dim \ker T + \operatorname{rank} T.$$

Moreover, if T *is injective, then* T^*T *is also injective.*
(ii) *Let* $T : \mathscr{H}^N \to \mathscr{H}^N$ *be a linear operator. Then* T *is injective if and only if it is surjective. Moreover,* $\ker T = (\operatorname{ran} T^*)^\perp$, *and hence*

$$\mathscr{H}^N = \ker T \oplus \operatorname{ran} T^*.$$

If $T : \mathscr{H}^N \to \mathscr{H}^N$ is an injective operator, then T is obviously invertible. If an operator $T : \mathscr{H}^N \to \mathscr{H}^K$ is not injective, we can make T injective by restricting it to $(\ker T)^\perp$. However, $T|_{(\ker T)^\perp}$ might still not be invertible, since it does not need to be surjective. This can be ensured by considering the operator $T : (\ker T)^\perp \to \operatorname{ran} T$, which is now invertible.

The Moore-Penrose inverse of an injective operator provides a one-sided inverse for the operator.

Definition 1.6 Let $T : \mathscr{H}^N \to \mathscr{H}^K$ be an injective, linear operator. The *Moore-Penrose inverse* of T, $T^\dagger$, is defined by

$$T^\dagger = \left(T^* T\right)^{-1} T^*.$$

It is immediate to prove invertibility from the left as stated in the following result.

Proposition 1.4 *If $T : \mathscr{H}^N \to \mathscr{H}^K$ is an injective, linear operator, then $T^\dagger T = Id$.*

Thus, $T^\dagger$ plays the role of the inverse on $\operatorname{ran} T$—not on all of $\mathscr{H}^K$. It projects a vector from $\mathscr{H}^K$ onto $\operatorname{ran} T$ and then inverts the operator on this subspace.

A more general notion of this inverse is called the *pseudoinverse*, which can be applied to a non-injective operator. In fact, it adds one more step to the action of $T^\dagger$ by first restricting to $(\ker T)^\perp$ to enforce injectivity of the operator followed by application of the Moore-Penrose inverse of this new operator. This pseudoinverse can be derived from the singular value decomposition. Recalling that by fixing orthonormal bases of the domain and range of a linear operator we derive an associated unique matrix representation; we begin by stating this decomposition in terms of a matrix.

Theorem 1.1 *Let A be an $M \times N$ matrix. Then there exist an $M \times M$ matrix U with $U^* U = Id$, and $N \times N$ matrix V with $V^* V = Id$, and an $M \times N$ diagonal matrix Σ with nonnegative, decreasing real entries on the diagonal such that*

$$A = U \Sigma V^*.$$

Hereby, an $M \times N$ diagonal matrix with $M \neq N$ is an $M \times N$ matrix $(a_{ij})_{i=1,\, j=1}^{M,N}$ with $a_{ij} = 0$ for $i \neq j$.

Definition 1.7 Let A be an $M \times N$ matrix, and let U, Σ, and V be chosen as in Theorem 1.1. Then $A = U \Sigma V^*$ is called the *singular value decomposition (SVD)* of A. The column vectors of U are called the *left singular vectors*, and the column vectors of V are referred to as the *right singular vectors* of A.

The pseudoinverse A^+ of A can be deduced from the SVD in the following way.

Theorem 1.2 *Let A be an $M \times N$ matrix, and let $A = U\Sigma V^*$ be its singular value decomposition. Then*

$$A^+ = V\Sigma^+ U^*,$$

where Σ^+ is the $N \times M$ diagonal matrix arising from Σ^* by inverting the nonzero (diagonal) entries.

1.2.2.2 Riesz bases

In the previous subsection, we recalled the notion of an orthonormal basis. However, sometimes the requirement of orthonormality is too strong, but uniqueness of a decomposition as well as stability are to be retained. The notion of a Riesz basis, which we next introduce, satisfies these desiderata.

Definition 1.8 A family of vectors $(\varphi_i)_{i=1}^N$ in a Hilbert space $\mathscr{H}^N$ is a *Riesz basis* with *lower* (respectively, *upper*) *Riesz bounds* A (resp. B), if, for all scalars $(a_i)_{i=1}^N$, we have

$$A\sum_{i=1}^N |a_i|^2 \leq \left\| \sum_{i=1}^N a_i\varphi_i \right\|^2 \leq B\sum_{i=1}^N |a_i|^2.$$

The following result is immediate from the definition.

Proposition 1.5 *Let $(\varphi_i)_{i=1}^N$ be a family of vectors. Then the following conditions are equivalent.*

(i) *$(\varphi_i)_{i=1}^N$ is a Riesz basis for $\mathscr{H}^N$ with Riesz bounds A and B.*
(ii) *For any orthonormal basis $(e_i)_{i=1}^N$ for $\mathscr{H}^N$, the operator T on $\mathscr{H}^N$ given by $Te_i = \varphi_i$ for all $i = 1, 2, \ldots, N$ is an invertible operator with $\|T\|^2 \leq B$ and $\|T^{-1}\|^{-2} \geq A$.*

1.2.2.3 Diagonalization

Next, we continue our list of important properties of linear operators.

Definition 1.9 A linear operator $T : \mathscr{H}^N \to \mathscr{H}^K$ is called

(a) *self-adjoint*, if $\mathscr{H}^N = \mathscr{H}^K$ and $T = T^*$.
(b) *normal*, if $\mathscr{H}^N = \mathscr{H}^K$ and $T^*T = TT^*$.
(c) *an isometry*, if $\|Tx\| = \|x\|$ for all $x \in \mathscr{H}^N$.
(d) *positive*, if $\mathscr{H}^N = \mathscr{H}^K$, T is self-adjoint, and $\langle Tx, x\rangle \geq 0$ for all $x \in \mathscr{H}^N$.
(e) *unitary*, if it is a surjective isometry.

From the variety of basic relations and results of those notions, the next proposition presents a selection of those which will be required in the sequel.

Proposition 1.6 *Let $T : \mathscr{H}^N \to \mathscr{H}^K$ be a linear operator.*

(i) *We have $\|T^*T\| = \|T\|^2$, and T^*T and TT^* are self-adjoint.*
(ii) *If $\mathscr{H}^N = \mathscr{H}^K$, the following conditions are equivalent.*
  (1) *T is self-adjoint.*
  (2) *$\langle Tx, \tilde{x}\rangle = \langle x, T\tilde{x}\rangle$ for all $x, \tilde{x} \in \mathscr{H}^N$.*
  (3) *If $\mathscr{H}^N$ is complex, $\langle Tx, x\rangle \in \mathbb{R}$ for all $x \in \mathscr{H}^N$.*
(iii) *The following conditions are equivalent.*
  (1) *T is an isometry.*
  (2) *$T^*T = Id$.*
  (3) *$\langle Tx, T\tilde{x}\rangle = \langle x, \tilde{x}\rangle$ for all $x, \tilde{x} \in \mathscr{H}^N$.*
(iv) *The following conditions are equivalent.*
  (1) *T is unitary.*
  (2) *T and T^* are isometric.*
  (3) *$TT^* = Id$ and $T^*T = Id$.*
(v) *If U is a unitary operator, then $\|UT\| = \|T\| = \|TU\|$.*

Diagonalizations of operators are frequently utilized to derive an understanding of the action of an operator. The following definitions lay the groundwork for this theory.

Definition 1.10 Let $T : \mathscr{H}^N \to \mathscr{H}^N$ be a linear operator. A nonzero vector $x \in \mathscr{H}^N$ is an *eigenvector* of T with *eigenvalue* λ, if $Tx = \lambda x$. The operator T is called *orthogonally diagonalizable*, if there exists an orthonormal basis $(e_i)_{i=1}^N$ of $\mathscr{H}^N$ consisting of eigenvectors of T.

We start with an easy observation.

Proposition 1.7 *For any linear operator $T : \mathscr{H}^N \to \mathscr{H}^K$, the nonzero eigenvalues of T^*T and TT^* are the same.*

If the operator is unitary, self-adjoint, or positive, we have more information on the eigenvalues stated in the next result, which follows immediately from Proposition 1.6.

Corollary 1.2 *Let $T : \mathscr{H}^N \to \mathscr{H}^N$ be a linear operator.*

(i) *If T is unitary, then its eigenvalues have modulus one.*
(ii) *If T is self-adjoint, then its eigenvalues are real.*
(iii) *If T is positive, then its eigenvalues are nonnegative.*

This fact allows us to introduce a condition number associated with each invertible positive operator.

Definition 1.11 Let $T : \mathscr{H}^N \to \mathscr{H}^N$ be an invertible positive operator with eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Then its *condition number* is defined by $\frac{\lambda_1}{\lambda_N}$.

We next state a fundamental result in operator theory which has its analogue in the infinite-dimensional setting called the *spectral theorem*.

Theorem 1.3 *Let $\mathscr{H}^N$ be complex and let $T : \mathscr{H}^N \to \mathscr{H}^N$ be a linear operator. Then the following conditions are equivalent.*

(i) *T is normal.*
(ii) *T is orthogonally diagonalizable.*
(iii) *There exists a diagonal matrix representation of T.*
(iv) *There exist an orthonormal basis $(e_i)_{i=1}^N$ of $\mathscr{H}^N$ and values $\lambda_1, \ldots, \lambda_N$ such that*

$$Tx = \sum_{i=1}^{N} \lambda_i \langle x, e_i \rangle e_i \quad \text{for all } x \in \mathscr{H}^N.$$

In this case,

$$\|T\| = \max_{1 \leq i \leq N} |\lambda_i|.$$

Since every self-adjoint operator is normal, we obtain the following corollary (which is independent of whether $\mathscr{H}^N$ is real or complex).

Corollary 1.3 *A self-adjoint operator is orthogonally diagonalizable.*

Another consequence of Theorem 1.3 is the following result, which in particular allows the definition of the n-th root of a positive operator.

Corollary 1.4 *Let $T : \mathscr{H}^N \to \mathscr{H}^N$ be an invertible positive operator with normalized eigenvectors $(e_i)_{i=1}^N$ and respective eigenvalues $(\lambda_i)_{i=1}^N$, let $a \in \mathbb{R}$, and define an operator $T^a : \mathscr{H}^N \to \mathscr{H}^N$ by*

$$T^a x = \sum_{i=1}^{N} \lambda_i^a \langle x, e_i \rangle e_i \quad \text{for all } x \in \mathscr{H}^N.$$

Then T^a is a positive operator and $T^a T^b = T^{a+b}$ for $a, b \in \mathbb{R}$. In particular, T^{-1} and $T^{-1/2}$ are positive operators.

Finally, we define the trace of an operator, which, by using Theorem 1.3, can be expressed in terms of eigenvalues.

Definition 1.12 Let $T : \mathscr{H}^N \to \mathscr{H}^N$ be an operator. Then, the *trace* of T is defined by

$$\operatorname{Tr} T = \sum_{i=1}^{N} \langle T e_i, e_i \rangle, \tag{1.1}$$

where $(e_i)_{i=1}^N$ is an arbitrary orthonormal basis for $\mathscr{H}^N$.

The trace is well defined since the sum in Eq. (1.1) is independent of the choice of the orthonormal basis.

Corollary 1.5 *Let $T : \mathscr{H}^N \to \mathscr{H}^N$ be an orthogonally diagonalizable operator, and let $(\lambda_i)_{i=1}^N$ be its eigenvalues. Then*

$$\operatorname{Tr} T = \sum_{i=1}^{N} \lambda_i .$$

1.2.2.4 Projection operators

Subspaces are closely intertwined with associated projection operators which map vectors onto the subspace either orthogonally or not. Although orthogonal projections are more often used, in Chap. 13 we will require the more general notion.

Definition 1.13 Let $P : \mathscr{H}^N \to \mathscr{H}^N$ be a linear operator. Then P is called a *projection*, if $P^2 = P$. This projection is called *orthogonal*, if P is in addition self-adjoint.

For brevity, *orthogonal projections* are often simply referred to as *projections* provided there is no danger of misinterpretation.

Relating to our previous comment, for any subspace $\mathscr{W}$ of $\mathscr{H}^N$, there exists a unique orthogonal projection P of $\mathscr{H}^N$ having $\mathscr{W}$ as its range. This projection can be constructed as follows: Let m denote the dimension of $\mathscr{W}$, and choose an orthonormal basis $(e_i)_{i=1}^m$ of $\mathscr{W}$. Then, for any $x \in \mathscr{H}^N$, we set

$$Px = \sum_{i=1}^{m} \langle x, e_i \rangle e_i .$$

It is important to notice that also $Id - P$ is an orthogonal projection of $\mathscr{H}^N$, this time onto the subspace $\mathscr{W}^\perp$.

An orthogonal projection P has the crucial property that each given vector of $\mathscr{H}^N$ is mapped to the closest vector in the range of P.

Lemma 1.1 *Let $\mathscr{W}$ be a subspace of $\mathscr{H}^N$, let P be the orthogonal projection onto $\mathscr{W}$, and let $x \in \mathscr{H}^N$. Then*

$$\|x - Px\| \leq \|x - \tilde{x}\| \quad \text{for all } \tilde{x} \in \mathscr{W}.$$

Moreover, if $\|x - Px\| = \|x - \tilde{x}\|$ for some $\tilde{x} \in \mathscr{W}$, then $\tilde{x} = Px$.

The next result gives the relationship between trace and rank for projections. This follows from the definition of an orthogonal projection and Corollaries 1.3 and 1.5.

Proposition 1.8 *Let P be the orthogonal projection onto a subspace $\mathscr{W}$ of $\mathscr{H}^N$, and let $m = \dim \mathscr{W}$. Then P is orthogonally diagonalizable with eigenvalue 1 of multiplicity m and eigenvalue 0 of multiplicity $N - m$. In particular, we have that* $\operatorname{Tr} P = m$.

1.3 Basics of Finite Frame Theory

We start by presenting the basics of finite frame theory. For illustration purposes, we then present some exemplary frame classes. At this point, we also refer to the monographs and books [34, 35, 99, 100, 111] as well as to [65, 66] for infinite-dimensional frame theory.

1.3.1 Definition of a Frame

The definition of a (Hilbert space) frame originates from early work by Duffin and Schaeffer [79] on nonharmonic Fourier series. The main idea, as discussed in Sect. 1.1, is to weaken Parseval's identity and yet still retain norm equivalence between a signal and its frame coefficients.

Definition 1.14 A family of vectors $(\varphi_i)_{i=1}^M$ in $\mathscr{H}^N$ is called a *frame for $\mathscr{H}^N$*, if there exist constants $0 < A \leq B < \infty$ such that

$$A\|x\|^2 \leq \sum_{i=1}^{M} \left|\langle x, \varphi_i \rangle\right|^2 \leq B\|x\|^2 \quad \text{for all } x \in \mathscr{H}^N. \tag{1.2}$$

The following notions are related to a frame $(\varphi_i)_{i=1}^M$.

(a) The constants A and B as in (1.2) are called the *lower and upper frame bound* for the frame, respectively. The largest lower frame bound and the smallest upper frame bound are denoted by A_{op}, B_{op} and are called the *optimal frame bounds*.

(b) Any family $(\varphi_i)_{i=1}^M$ satisfying the right-hand side inequality in (1.2) is called a *B-Bessel sequence*.
(c) If $A = B$ is possible in (1.2), then $(\varphi_i)_{i=1}^M$ is called an *A-tight frame*.
(d) If $A = B = 1$ is possible in (1.2), i.e., Parseval's identity holds, then $(\varphi_i)_{i=1}^M$ is called a *Parseval frame*.
(e) If there exists a constant c such that $\|\varphi_i\| = c$ for all $i = 1, 2, \ldots, M$, then $(\varphi_i)_{i=1}^M$ is an *equal norm frame*. If $c = 1$, $(\varphi_i)_{i=1}^M$ is a *unit norm frame*.
(f) If there exists a constant c such that $|\langle \varphi_i, \varphi_j \rangle| = c$ for all $i \neq j$, then $(\varphi_i)_{i=1}^M$ is called an *equiangular* frame.
(g) The values $(\langle x, \varphi_i \rangle)_{i=1}^M$ are called the *frame coefficients* of the vector x with respect to the frame $(\varphi_i)_{i=1}^M$.
(h) The frame $(\varphi_i)_{i=1}^M$ is called *exact*, if $(\varphi_i)_{i \in I}$ ceases to be a frame for $\mathscr{H}^N$ for every $I = \{1, \ldots, M\} \setminus \{i_0\}$, $i_0 \in \{1, \ldots, M\}$.

We can immediately make the following useful observations.

Lemma 1.2 *Let $(\varphi_i)_{i=1}^M$ be a family of vectors in $\mathscr{H}^N$.*

(i) *If $(\varphi_i)_{i=1}^M$ is an orthonormal basis, then $(\varphi_i)_{i=1}^M$ is a Parseval frame. The converse is not true in general.*
(ii) *$(\varphi_i)_{i=1}^M$ is a frame for $\mathscr{H}^N$ if and only if it is a spanning set for $\mathscr{H}^N$.*
(iii) *$(\varphi_i)_{i=1}^M$ is a unit norm Parseval frame if and only if it is an orthonormal basis.*
(iv) *If $(\varphi_i)_{i=1}^M$ is an exact frame for $\mathscr{H}^N$, then it is a basis of $\mathscr{H}^N$, i.e., a linearly independent spanning set.*

Proof (i) The first part is an immediate consequence of Proposition 1.1. For the second part, let $(e_i)_{i=1}^N$ and $(g_i)_{i=1}^N$ be orthonormal bases for $\mathscr{H}^N$. Then $(e_i/\sqrt{2})_{i=1}^N \cup (g_i/\sqrt{2})_{i=1}^N$ is a Parseval frame for $\mathscr{H}^N$, but not an orthonormal basis.

(ii) If $(\varphi_i)_{i=1}^M$ is not a spanning set for $\mathscr{H}^N$, then there exists $x \neq 0$ such that $\langle x, \varphi_i \rangle = 0$ for all $i = 1, \ldots, M$. Hence, $(\varphi_i)_{i=1}^M$ cannot be a frame. Conversely, assume that $(\varphi_i)_{i=1}^M$ is not a frame. Then there exists a sequence $(x_n)_{n=1}^\infty$ of normalized vectors in $\mathscr{H}^N$ such that $\sum_{i=1}^M |\langle x_n, \varphi_i \rangle|^2 < 1/n$ for all $n \in \mathbb{N}$. Hence, the limit x of a convergent subsequence of $(x_n)_{n=1}^\infty$ satisfies $\langle x, \varphi_i \rangle = 0$ for all $i = 1, \ldots, M$. Since $\|x\| = 1$, it follows that $(\varphi_i)_{i=1}^M$ is not a spanning set.

(iii) By the Parseval property, for each $i_0 \in \{1, \ldots, M\}$, we have

$$\|\varphi_{i_0}\|_2^2 = \sum_{i=1}^M \big|\langle \varphi_{i_0}, \varphi_i \rangle\big|^2 = \|\varphi_{i_0}\|_2^4 + \sum_{i=1,\, i \neq i_0}^M \big|\langle \varphi_{i_0}, \varphi_i \rangle\big|^2.$$

Since the frame vectors are normalized, we conclude that

$$\sum_{i=1,\, i \neq i_0}^M \big|\langle \varphi_{i_0}, \varphi_i \rangle\big|^2 = 0 \quad \text{for all } i_0 \in \{1, \ldots, M\}.$$

Hence $\langle \varphi_i, \varphi_j \rangle = 0$ for all $i \neq j$. Thus, $(\varphi_i)_{i=1}^M$ is an orthonormal system which is complete by (ii), and (iii) is proved.

(iv) If $(\varphi_i)_{i=1}^M$ is a frame, by (ii), it is also a spanning set for $\mathscr{H}^N$. Towards a contradiction, assume that $(\varphi_i)_{i=1}^M$ is linearly dependent. Then there exist some $i_0 \in \{1, \ldots, M\}$ and values λ_i, $i \in I := \{1, \ldots, M\} \setminus \{i_0\}$ such that

$$\varphi_{i_0} = \sum_{i \in I} \lambda_i \varphi_i.$$

This implies that $(\varphi_i)_{i \in I}$ is also a frame, thus contradicting exactness of the frame. □

Before presenting some insightful basic results in frame theory, we first discuss some examples of frames to develop an intuitive understanding.

1.3.2 Examples

By Lemma 1.2 (iii), orthonormal bases are unit norm Parseval frames (and vice versa). However, applications typically require *redundant* Parseval frames. One basic way to approach this construction problem is to build redundant Parseval frames using orthonormal bases, and we will present several examples in the sequel. Since the associated proofs are straightforward, we leave them to the interested reader.

Example 1.1 Let $(e_i)_{i=1}^N$ be an orthonormal basis for $\mathscr{H}^N$.

(1) The system

$$(e_1, 0, e_2, 0, \ldots, e_N, 0)$$

is a Parseval frame for $\mathscr{H}^N$. This example indicates that a Parseval frame can indeed contain zero vectors.

(2) The system

$$\left(e_1, \frac{e_2}{\sqrt{2}}, \frac{e_2}{\sqrt{2}}, \frac{e_3}{\sqrt{3}}, \frac{e_3}{\sqrt{3}}, \frac{e_3}{\sqrt{3}}, \ldots, \frac{e_N}{\sqrt{N}}, \ldots, \frac{e_N}{\sqrt{N}}\right)$$

is a Parseval frame for $\mathscr{H}^N$. This example indicates two important issues. First, a Parseval frame can have multiple copies of a single vector. Second, the norms of vectors of an (infinite) Parseval frame can converge to zero.

We next consider a series of examples of non-Parseval frames.

Example 1.2 Let $(e_i)_{i=1}^N$ be an orthonormal basis for $\mathscr{H}^N$.

(1) The system

$$(e_1, e_1, \ldots, e_1, e_2, e_3, \ldots, e_N)$$

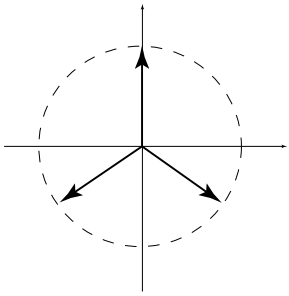


Fig. 1.1 Mercedes-Benz frame

with the vector e_1 appearing $N+1$ times, is a frame for $\mathscr{H}^N$ with frame bounds 1 and $N+1$.

(2) The system

$$(e_1, e_1, e_2, e_2, e_3, e_3, \ldots, e_N)$$

is a 2-tight frame for $\mathscr{H}^N$.

(3) The union of L orthonormal bases of $\mathscr{H}^N$ is a unit norm L-tight frame for $\mathscr{H}^N$, generalizing (2).

A particularly interesting example is the smallest truly redundant Parseval frame for $\mathbb{R}^2$, which is typically coined the *Mercedes-Benz frame*. The reason for this naming becomes evident in Fig. 1.1.

Example 1.3 The *Mercedes-Benz frame* for $\mathbb{R}^2$ is the equal norm tight frame for $\mathbb{R}^2$ given by:

$$\left(\sqrt{\tfrac{2}{3}}\begin{pmatrix}0\\1\end{pmatrix}, \sqrt{\tfrac{2}{3}}\begin{pmatrix}\frac{\sqrt{3}}{2}\\-\frac{1}{2}\end{pmatrix}, \sqrt{\tfrac{2}{3}}\begin{pmatrix}-\frac{\sqrt{3}}{2}\\-\frac{1}{2}\end{pmatrix}\right)$$

Note that this frame is also equiangular.

For more information on the theoretical aspects of equiangular frames we refer to [60, 91, 120, 139]. A selection of their applications is reconstruction without phase [5, 6], erasure-resilient transmission [15, 102], and coding [136]. We also refer to the Chaps. 4, 5 in this book for more details on equiangular frames.

Another standard class of examples can be derived from the *discrete Fourier transform* (*DFT*) *matrix*.

Example 1.4 Given $M \in \mathbb{N}$, we let $\omega = \exp(\frac{2\pi i}{M})$. Then the DFT matrix in $\mathbb{C}^{M\times M}$ is defined by

$$D_M = \frac{1}{\sqrt{M}}\left(\omega^{jk}\right)_{j,k=0}^{M-1}.$$

This matrix is a unitary operator on $\mathbb{C}^M$. Later (see Corollary 1.11) it will be seen that the selection of any N rows from D_M yields a Parseval frame for $\mathbb{C}^N$ by taking the associated M column vectors.

There also exist particularly interesting classes of frames such as Gabor frames utilized primarily for audio processing. Among the results on various aspects of Gabor frames are uncertainty considerations [113], linear independence [119], group-related properties [89], optimality analysis [127], and applications [67, 74, 75, 87, 88]. Chapter 6 provides a survey on this class of frames. Another example is the class of group frames, for which various constructions [24, 101, 147], classifications [64], and intriguing symmetry properties [146, 148] have been studied. A comprehensive presentation can be found in Chap. 5.

1.4 Frames and Operators

For the rest of this introduction we set $\ell_2^M := \ell_2(\{1, \dots, M\})$. Note that this space in fact coincides with $\mathbb{R}^M$ or $\mathbb{C}^M$, endowed with the standard inner product and the associated Euclidean norm.

The analysis, synthesis, and frame operators determine the operation of a frame when analyzing and reconstructing a signal. The Gramian operator is perhaps not that well known, yet it crucially illuminates the behavior of a frame $(\varphi_i)_{i=1}^M$ embedded as an N-dimensional subspace in the high-dimensional space ℓ_2^M.

1.4.1 Analysis and Synthesis Operators

Two of the main operators associated with a frame are the analysis and synthesis operators. The analysis operator—as the name suggests—analyzes a signal in terms of the frame by computing its frame coefficients. We start by formalizing this notion.

Definition 1.15 Let $(\varphi_i)_{i=1}^M$ be a family of vectors in $\mathscr{H}^N$. Then the associated *analysis operator* $T : \mathscr{H}^N \to \ell_2^M$ is defined by

$$Tx := \big(\langle x, \varphi_i \rangle\big)_{i=1}^M, \quad x \in \mathscr{H}^N.$$

In the following lemma we derive two basic properties of the analysis operator.

Lemma 1.3 *Let $(\varphi_i)_{i=1}^M$ be a sequence of vectors in $\mathscr{H}^N$ with associated analysis operator T.*

(i) *We have*

$$\|Tx\|^2 = \sum_{i=1}^M \big|\langle x, \varphi_i \rangle\big|^2 \quad \text{for all } x \in \mathscr{H}^N.$$

Hence, $(\varphi_i)_{i=1}^M$ is a frame for $\mathscr{H}^N$ if and only if T is injective.

(ii) *The adjoint operator* $T^* : \ell_2^M \to \mathscr{H}^N$ *of* T *is given by*

$$T^*(a_i)_{i=1}^M = \sum_{i=1}^M a_i \varphi_i.$$

Proof (i) This is an immediate consequence of the definition of T and the frame property (1.2).

(ii) For $x = (a_i)_{i=1}^M$ and $y \in \mathscr{H}^N$, we have

$$\left\langle T^* x, y \right\rangle = \langle x, Ty \rangle = \left\langle (a_i)_{i=1}^M, \left(\langle y, \varphi_i \rangle\right)_{i=1}^M \right\rangle = \sum_{i=1}^M a_i \overline{\langle y, \varphi_i \rangle} = \left\langle \sum_{i=1}^M a_i \varphi_i, y \right\rangle.$$

Thus, T^* is as claimed. □

The second main operator associated to a frame, the synthesis operator, is now defined as the adjoint operator to the analysis operator given in Lemma 1.3(ii).

Definition 1.16 Let $(\varphi_i)_{i=1}^M$ be a sequence of vectors in $\mathscr{H}^N$ with associated analysis operator T. Then the associated *synthesis operator* is defined to be the adjoint operator T^*.

The next result summarizes some basic, yet useful, properties of the synthesis operator.

Lemma 1.4 *Let* $(\varphi_i)_{i=1}^M$ *be a sequence of vectors in* $\mathscr{H}^N$ *with associated analysis operator* T.

(i) *Let* $(e_i)_{i=1}^M$ *denote the standard basis of* ℓ_2^M. *Then for all* $i = 1, 2, \ldots, M$, *we have* $T^* e_i = T^* P e_i = \varphi_i$, *where* $P : \ell_2^M \to \ell_2^M$ *denotes the orthogonal projection onto* $\operatorname{ran} T$.

(ii) $(\varphi_i)_{i=1}^M$ *is a frame if and only if* T^* *is surjective.*

Proof The first claim follows immediately from Lemma 1.3 and the fact that $\ker T^* = (\operatorname{ran} T)^\perp$. The second claim is a consequence of $\operatorname{ran} T^* = (\ker T)^\perp$ and Lemma 1.3(i). □

Often frames are modified by the application of an invertible operator. The next result shows not only the impact on the associated analysis operator, but also the fact that the new sequence again forms a frame.

Proposition 1.9 *Let* $\Phi = (\varphi_i)_{i=1}^M$ *be a sequence of vectors in* $\mathscr{H}^N$ *with associated analysis operator* T_Φ *and let* $F : \mathscr{H}^N \to \mathscr{H}^N$ *be a linear operator. Then the analysis operator of the sequence* $F\Phi = (F\varphi_i)_{i=1}^M$ *is given by*

$$T_{F\Phi} = T_\Phi F^*.$$

Moreover, if Φ is a frame for $\mathscr{H}^N$ and F is invertible, then $F\Phi$ is also a frame for $\mathscr{H}^N$.

Proof For $x \in \mathscr{H}^N$ we have

$$T_{F\Phi}x = \big(\langle x, F\varphi_i\rangle\big)_{i=1}^M = \big(\langle F^*x, \varphi_i\rangle\big)_{i=1}^M = T_\Phi F^*x.$$

This proves $T_{F\Phi} = T_\Phi F^*$. The *moreover* part follows from Lemma 1.4(ii). □

Next, we analyze the structure of the matrix representation of the synthesis operator. This matrix is of fundamental importance, since this is what most frame constructions in fact focus on; see also Sect. 1.6.

The first result provides the form of this matrix along with stability properties.

Lemma 1.5 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T. Then a matrix representation of the synthesis operator T^* is the $N \times M$ matrix given by*

$$\begin{bmatrix} | & | & \cdots & | \\ \varphi_1 & \varphi_2 & \cdots & \varphi_M \\ | & | & \cdots & | \end{bmatrix}.$$

Moreover, the Riesz bounds of the row vectors of this matrix equal the frame bounds of the column vectors.

Proof The form of the matrix representation is obvious. To prove the *moreover* part, let $(e_j)_{j=1}^N$ be the corresponding orthonormal basis of $\mathscr{H}^N$ and for $j = 1, 2, \ldots, N$ let

$$\psi_j = \big[\langle\varphi_1, e_j\rangle, \langle\varphi_2, e_j\rangle, \ldots, \langle\varphi_M, e_j\rangle\big]$$

be the row vectors of the matrix. Then for $x = \sum_{j=1}^N a_j e_j$ we obtain

$$\begin{aligned}\sum_{i=1}^M \big|\langle x, \varphi_i\rangle\big|^2 &= \sum_{i=1}^M \left|\sum_{j=1}^N a_j\langle e_j, \varphi_i\rangle\right|^2 = \sum_{j,k=1}^N a_j\overline{a_k}\sum_{i=1}^M \langle e_j, \varphi_i\rangle\langle\varphi_i, e_k\rangle \\ &= \sum_{j,k=1}^N a_j\overline{a_k}\langle\psi_k, \psi_j\rangle = \left\|\sum_{j=1}^N \overline{a_j}\psi_j\right\|^2.\end{aligned}$$

The claim follows from here. □

A much stronger result (Proposition 1.12) can be proven for the case in which the matrix representation is derived using a specifically chosen orthonormal basis. However, the choice of this orthonormal basis requires the introduction of the frame operator in the following Sect. 1.4.2.

1.4.2 The Frame Operator

The frame operator might be considered the most important operator associated with a frame. Although it is "merely" the concatenation of the analysis and synthesis operators, it encodes crucial properties of the frame, as we will see in the sequel. Moreover, it is also fundamental for the reconstruction of signals from frame coefficients (see Theorem 1.8).

1.4.2.1 Fundamental properties

The precise definition of the frame operator associated with a frame is as follows.

Definition 1.17 Let $(\varphi_i)_{i=1}^M$ be a sequence of vectors in $\mathcal{H}^N$ with associated analysis operator T. Then the associated *frame operator* $S : \mathcal{H}^N \to \mathcal{H}^N$ is defined by

$$Sx := T^*Tx = \sum_{i=1}^{M} \langle x, \varphi_i \rangle \varphi_i, \quad x \in \mathcal{H}^N.$$

A first observation concerning the close relation of the frame operator to frame properties is the following lemma.

Lemma 1.6 *Let $(\varphi_i)_{i=1}^M$ be a sequence of vectors in $\mathcal{H}^N$ with associated frame operator S. Then, for all $x \in \mathcal{H}^N$,*

$$\langle Sx, x \rangle = \sum_{i=1}^{M} \big| \langle x, \varphi_i \rangle \big|^2.$$

Proof The proof follows directly from $\langle Sx, x \rangle = \langle T^*Tx, x \rangle = \|Tx\|^2$ and Lemma 1.3(i). □

Clearly, the frame operator $S = T^*T$ is self-adjoint and positive. The most fundamental property of the frame operator—if the underlying sequence of vectors forms a frame—is its invertibility, which is crucial for the reconstruction formula.

Theorem 1.4 *The frame operator S of a frame $(\varphi_i)_{i=1}^M$ for $\mathcal{H}^N$ with frame bounds A and B is a positive, self-adjoint invertible operator satisfying*

$$A \cdot Id \leq S \leq B \cdot Id.$$

Proof By Lemma 1.6, we have

$$\langle Ax, x \rangle = A\|x\|^2 \leq \sum_{i=1}^{M} \big| \langle x, \varphi_i \rangle \big|^2 = \langle Sx, x \rangle \leq B\|x\|^2 = \langle Bx, x \rangle \quad \text{for all } x \in \mathcal{H}^N.$$

This implies the claimed inequality. □

The following proposition follows directly from Proposition 1.9.

Proposition 1.10 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S, and let F be an invertible operator on $\mathscr{H}^N$. Then $(F\varphi_i)_{i=1}^M$ is a frame with frame operator FSF^*.*

1.4.2.2 The special case of tight frames

Tight frames can be characterized as those frames whose frame operator equals a positive multiple of the identity. The next result provides a variety of similarly frame-operator-inspired classifications.

Proposition 1.11 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T and frame operator S. Then the following conditions are equivalent.*

(i) *$(\varphi_i)_{i=1}^M$ is an A-tight frame for $\mathscr{H}^N$.*
(ii) *$S = A \cdot Id$.*
(iii) *For every $x \in \mathscr{H}^N$,*

$$x = A^{-1} \cdot \sum_{i=1}^{M} \langle x, \varphi_i \rangle \varphi_i .$$

(iv) *For every $x \in \mathscr{H}^N$,*

$$A\|x\|^2 = \sum_{i=1}^{M} \big|\langle x, \varphi_i \rangle\big|^2 .$$

(v) *$T/\sqrt{A}$ is an isometry.*

Proof (i) $\Leftrightarrow$ (ii) $\Leftrightarrow$ (iii) $\Leftrightarrow$ (iv) These are immediate from the definition of the frame operator and from Theorem 1.4.

(ii) $\Leftrightarrow$ (v) This follows from the fact that $T/\sqrt{A}$ is an isometry if and only if $T^*T = A \cdot Id$. □

A similar result for the special case of a Parseval frame can be easily deduced from Proposition 1.11 by setting $A = 1$.

1.4.2.3 Eigenvalues of the frame operator

Tight frames have the property that the eigenvalues of the associated frame operator all coincide. We next consider the general situation, i.e., frame operators with arbitrary eigenvalues.

The first and maybe even most important result shows that the largest and smallest eigenvalues of the frame operator are the optimal frame bounds of the frame. Optimality refers to the smallest upper frame bound and the largest lower frame bound.

Theorem 1.5 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S having eigenvalues $\lambda_1 \geq \cdots \geq \lambda_N$. Then λ_1 coincides with the optimal upper frame bound and λ_N is the optimal lower frame bound.*

Proof Let $(e_i)_{i=1}^N$ denote the normalized eigenvectors of the frame operator S with respective eigenvalues $(\lambda_j)_{j=1}^N$ written in decreasing order. Let $x \in \mathscr{H}^N$. Since $x = \sum_{j=1}^M \langle x, e_j\rangle e_j$, we obtain

$$Sx = \sum_{j=1}^N \lambda_j \langle x, e_j\rangle e_j.$$

By Lemma 1.6, this implies

$$\begin{aligned}\sum_{i=1}^M \big|\langle x, \varphi_i\rangle\big|^2 = \langle Sx, x\rangle &= \left\langle \sum_{j=1}^N \lambda_j \langle x, e_j\rangle e_j, \sum_{j=1}^N \langle x, e_j\rangle e_j \right\rangle \\ &= \sum_{j=1}^N \lambda_j \big|\langle x, e_j\rangle\big|^2 \leq \lambda_1 \sum_{j=1}^N \big|\langle x, e_j\rangle\big|^2 = \lambda_1 \|x\|^2.\end{aligned}$$

Thus $B_{\text{op}} \leq \lambda_1$, where B_{op} denotes the optimal upper frame bound of the frame $(\varphi_i)_{i=1}^M$. The claim $B_{\text{op}} = \lambda_1$ then follows from

$$\sum_{i=1}^M \big|\langle e_1, \varphi_i\rangle\big|^2 = \langle Se_1, e_1\rangle = \langle \lambda_1 e_1, e_1\rangle = \lambda_1.$$

The claim concerning the lower frame bound can be proven similarly. □

From this result, we can now draw the following immediate conclusion about the Riesz bounds.

Corollary 1.6 *Let $(\varphi_i)_{i=1}^N$ be a frame for $\mathscr{H}^N$. Then the following statements hold.*

(i) *The optimal upper Riesz bound and the optimal upper frame bound of $(\varphi_i)_{i=1}^N$ coincide.*
(ii) *The optimal lower Riesz bound and the optimal lower frame bound of $(\varphi_i)_{i=1}^N$ coincide.*

Proof Let T denote the analysis operator of $(\varphi_i)_{i=1}^N$ and S the associated frame operator having eigenvalues $(\lambda_i)_{i=1}^N$ written in decreasing order. We have

$$\lambda_1 = \|S\| = \|T^*T\| = \|T\|^2 = \|T^*\|^2$$

and

$$\lambda_N = \|S^{-1}\|^{-1} = \|(T^*T)^{-1}\|^{-1} = \|(T^*)^{-1}\|^{-2}.$$

Now, both claims follow from Theorem 1.5, Lemma 1.4, and Proposition 1.5. □

The next theorem reveals a relation between the frame vectors and the eigenvalues and eigenvectors of the associated frame operator.

Theorem 1.6 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S having normalized eigenvectors $(e_j)_{j=1}^N$ and respective eigenvalues $(\lambda_j)_{j=1}^N$. Then for all $j = 1, 2, \ldots, N$ we have*

$$\lambda_j = \sum_{i=1}^M |\langle e_j, \varphi_i \rangle|^2.$$

In particular,

$$\operatorname{Tr} S = \sum_{j=1}^N \lambda_j = \sum_{i=1}^M \|\varphi_i\|^2.$$

Proof This follows from $\lambda_j = \langle Se_j, e_j \rangle$ for all $j = 1, \ldots, N$ and Lemma 1.6. □

1.4.2.4 Structure of the synthesis matrix

As already promised in Sect. 1.4.1, we now apply the previously derived results to obtain a complete characterization of the synthesis matrix of a frame in terms of the frame operator.

Proposition 1.12 *Let $T : \mathscr{H}^N \to \ell_2^M$ be a linear operator, let $(e_j)_{j=1}^N$ be an orthonormal basis of $\mathscr{H}^N$, and let $(\lambda_j)_{j=1}^N$ be a sequence of positive numbers. By A denote the $N \times M$ matrix representation of T^* with respect to $(e_j)_{j=1}^N$ (and the standard basis $(\hat{e}_i)_{i=1}^M$ of ℓ_2^M). Then the following conditions are equivalent.*

(i) *$(T^*\hat{e}_i)_{i=1}^M$ forms a frame for $\mathscr{H}^N$ whose frame operator has eigenvectors $(e_j)_{j=1}^N$ and associated eigenvalues $(\lambda_j)_{j=1}^N$.*
(ii) *The rows of A are orthogonal, and the j-th row square sums to λ_j.*
(iii) *The columns of A form a frame for ℓ_2^N, and $AA^* = \operatorname{diag}(\lambda_1, \ldots, \lambda_N)$.*

Proof Let $(f_j)_{j=1}^N$ be the standard basis of ℓ_2^N and denote by $U : \ell_2^N \to \mathscr{H}^N$ the unitary operator which maps f_j to e_j. Then $T^* = UA$.

(i)⇒(ii) For $j, k \in \{1, \dots, N\}$ we have

$$\langle A^* f_j, A^* f_k \rangle = \langle TUf_j, TUf_k \rangle = \langle T^* T e_j, e_k \rangle = \lambda_j \delta_{jk},$$

which is equivalent to (ii).

(ii)⇒(iii) Since the rows of A are orthogonal, we have $\operatorname{rank} A = N$, which implies that the columns of A form a frame for ℓ_2^N. The rest follows from $\langle AA^* f_j, f_k \rangle = \langle A^* f_j, A^* f_k \rangle = \lambda_j \delta_{jk}$ for $j, k = 1, \dots, N$.

(iii)⇒(i) Since $(A\hat{e}_i)_{i=1}^M$ is a spanning set for ℓ_2^N and $T^* = UA$, it follows that $(T^*\hat{e}_i)_{i=1}^M$ forms a frame for $\mathscr{H}^N$. Its analysis operator is given by T, since for all $x \in \mathscr{H}^N$,

$$\big(\langle x, T^*\hat{e}_i \rangle\big)_{i=1}^M = \big(\langle Tx, \hat{e}_i \rangle\big)_{i=1}^M = Tx.$$

Moreover,

$$T^*Te_j = UAA^*U^*e_j = U \operatorname{diag}(\lambda_1, \dots, \lambda_N) f_j = \lambda_j U f_j = \lambda_j e_j,$$

which completes the proof. □

1.4.3 Gramian Operator

Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T. The previous subsection was concerned with properties of the frame operator defined by $S = T^*T : \mathscr{H}^N \to \mathscr{H}^N$. Of particular interest is also the operator generated by first applying the synthesis and then the analysis operator. Let us first state the precise definition before discussing its importance.

Definition 1.18 Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T. Then the operator $G : \ell_2^M \to \ell_2^M$ defined by

$$G(a_i)_{i=1}^M = TT^*(a_i)_{i=1}^M = \left(\sum_{i=1}^M a_i \langle \varphi_i, \varphi_k \rangle \right)_{k=1}^M = \sum_{i=1}^M a_i \big(\langle \varphi_i, \varphi_k \rangle \big)_{k=1}^M$$

is called the *Gramian* (*operator*) of the frame $(\varphi_i)_{i=1}^M$.

Note that the (canonical) matrix representation of the Gramian of a frame $(\varphi_i)_{i=1}^M$ for $\mathscr{H}^N$ (which will also be called the *Gramian matrix*) is given by

$$\begin{bmatrix} \|\varphi_1\|^2 & \langle \varphi_2, \varphi_1 \rangle & \cdots & \langle \varphi_M, \varphi_1 \rangle \\ \langle \varphi_1, \varphi_2 \rangle & \|\varphi_2\|^2 & \cdots & \langle \varphi_M, \varphi_2 \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle \varphi_1, \varphi_M \rangle & \langle \varphi_2, \varphi_M \rangle & \cdots & \|\varphi_M\|^2 \end{bmatrix}.$$

One property of the Gramian is immediate. In fact, if the frame is unit norm, then the entries of the Gramian matrix are exactly the cosines of the angles between the frame vectors. Hence, for instance, if a frame is equiangular, then all off-diagonal entries of the Gramian matrix have the same modulus.

The fundamental properties of the Gramian operator are collected in the following result.

Theorem 1.7 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T, frame operator S, and Gramian operator G. Then the following statements hold.*

(i) *An operator U on $\mathscr{H}^N$ is unitary if and only if the Gramian of $(U\varphi_i)_{i=1}^M$ coincides with G.*
(ii) *The nonzero eigenvalues of G and S coincide.*
(iii) *$(\varphi_i)_{i=1}^M$ is a Parseval frame if and only if G is an orthogonal projection of rank N (namely onto the range of T).*
(iv) *G is invertible if and only if $M = N$.*

Proof (i) This follows immediately from the fact that the entries of the Gramian matrix for $(U\varphi_i)_{i=1}^M$ are of the form $\langle U\varphi_i, U\varphi_j\rangle$.

(ii) Since TT^* and T^*T have the same nonzero eigenvalues (see Proposition 1.7), the same is true for G and S.

(iii) It is immediate to prove that G is self-adjoint and has rank N. Since T is injective, T^* is surjective, and

$$G^2 = \big(TT^*\big)\big(TT^*\big) = T\big(T^*T\big)T^*,$$

it follows that G is an orthogonal projection if and only if $T^*T = Id$, which is equivalent to the frame being Parseval.

(iv) This is immediate by (ii). □

1.5 Reconstruction from Frame Coefficients

The analysis of a signal is typically performed by merely considering its frame coefficients. However, if the task is transmission of a signal, the ability to reconstruct the signal from its frame coefficients and also to do so efficiently becomes crucial. Reconstruction from coefficients with respect to an orthonormal basis was discussed in Corollary 1.1. However, reconstruction from coefficients with respect to a redundant system is much more delicate and requires the utilization of another frame, called the dual frame. If computing such a dual frame is computationally too complex, a circumvention of this problem is the frame algorithm.

1.5.1 Exact Reconstruction

We start by stating an exact reconstruction formula.

Theorem 1.8 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S. Then, for every $x \in \mathscr{H}^N$, we have*

$$x = \sum_{i=1}^{M} \langle x, \varphi_i \rangle S^{-1}\varphi_i = \sum_{i=1}^{M} \langle x, S^{-1}\varphi_i \rangle \varphi_i .$$

Proof The proof follows directly from the definition of the frame operator in Definition 1.17 by writing $x = S^{-1}Sx$ and $x = SS^{-1}x$. □

Notice that the first formula can be interpreted as a reconstruction strategy, whereas the second formula has the flavor of a decomposition. We further observe that the sequence $(S^{-1}\varphi_i)_{i=1}^M$ plays a crucial role in the formulas in Theorem 1.8. The next result shows that this sequence indeed also constitutes a frame.

Proposition 1.13 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame bounds A and B and with frame operator S. Then the sequence $(S^{-1}\varphi_i)_{i=1}^M$ is a frame for $\mathscr{H}^N$ with frame bounds B^{-1} and A^{-1} and with frame operator S^{-1}.*

Proof By Proposition 1.10, the sequence $(S^{-1}\varphi_i)_{i=1}^M$ forms a frame for $\mathscr{H}^N$ with associated frame operator $S^{-1}S(S^{-1})^* = S^{-1}$. This in turn yields the frame bounds B^{-1} and A^{-1}. □

This new frame is called the *canonical dual frame*. In the sequel, we will discuss that other dual frames may also be utilized for reconstruction.

Definition 1.19 Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator denoted by S. Then $(S^{-1}\varphi_i)_{i=1}^M$ is called the *canonical dual frame* for $(\varphi_i)_{i=1}^M$.

The canonical dual frame of a Parseval frame is now easily determined by Proposition 1.13.

Corollary 1.7 *Let $(\varphi_i)_{i=1}^M$ be a Parseval frame for $\mathscr{H}^N$. Then its canonical dual frame is the frame $(\varphi_i)_{i=1}^M$ itself, and the reconstruction formula in Theorem 1.8 reads*

$$x = \sum_{i=1}^{M} \langle x, \varphi_i \rangle \varphi_i , \quad x \in \mathscr{H}^N .$$

As an application of the above reconstruction formula for Parseval frames, we prove the following proposition which again shows the close relation between Parseval frames and orthonormal bases already indicated in Lemma 1.2.

Proposition 1.14 (Trace Formula for Parseval Frames) *Let $(\varphi_i)_{i=1}^M$ be a Parseval frame for $\mathscr{H}^N$, and let F be a linear operator on $\mathscr{H}^N$. Then*

$$\mathrm{Tr}(F) = \sum_{i=1}^{M} \langle F\varphi_i, \varphi_i \rangle.$$

Proof Let $(e_j)_{j=1}^N$ be an orthonormal basis for $\mathscr{H}^N$. Then, by definition,

$$\mathrm{Tr}(F) = \sum_{j=1}^{N} \langle Fe_j, e_j \rangle.$$

This implies

$$\begin{aligned}\mathrm{Tr}(F) &= \sum_{j=1}^{N} \left\langle \sum_{i=1}^{M} \langle Fe_j, \varphi_i \rangle \varphi_i, e_j \right\rangle = \sum_{j=1}^{N} \sum_{i=1}^{M} \left\langle e_j, F^*\varphi_i \right\rangle \langle \varphi_i, e_j \rangle \\ &= \sum_{i=1}^{M} \left\langle \sum_{j=1}^{N} \langle \varphi_i, e_j \rangle e_j, F^*\varphi_i \right\rangle = \sum_{i=1}^{M} \left\langle \varphi_i, F^*\varphi_i \right\rangle = \sum_{i=1}^{M} \langle F\varphi_i, \varphi_i \rangle.\end{aligned}$$

□

As already announced, many other dual frames for reconstruction exist. We next provide a precise definition.

Definition 1.20 Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$. Then a frame $(\psi_i)_{i=1}^M$ is called a *dual frame* for $(\varphi_i)_{i=1}^M$, if

$$x = \sum_{i=1}^{M} \langle x, \varphi_i \rangle \psi_i \quad \text{for all } x \in \mathscr{H}^N.$$

Dual frames, which do not coincide with the canonical dual frame, are often coined *alternate dual frames*.

Similar to the different forms of the reconstruction formula in Theorem 1.8, dual frames can also achieve reconstruction in different ways.

Proposition 1.15 *Let $(\varphi_i)_{i=1}^M$ and $(\psi_i)_{i=1}^M$ be frames for $\mathscr{H}^N$ and let T and $\tilde{T}$ be the analysis operators of $(\varphi_i)_{i=1}^M$ and $(\psi_i)_{i=1}^M$, respectively. Then the following conditions are equivalent.*

(i) *We have $x = \sum_{i=1}^M \langle x, \psi_i \rangle \varphi_i$ for all $x \in \mathscr{H}^N$.*
(ii) *We have $x = \sum_{i=1}^M \langle x, \varphi_i \rangle \psi_i$ for all $x \in \mathscr{H}^N$.*
(iii) *We have $\langle x, y \rangle = \sum_{i=1}^M \langle x, \varphi_i \rangle \langle \psi_i, y \rangle$ for all $x, y \in \mathscr{H}^N$.*
(iv) *$T^*\tilde{T} = Id$ and $\tilde{T}^*T = Id$.*

Proof Clearly (i) is equivalent to $T^*\tilde{T} = Id$, which holds if and only if $\tilde{T}^*T = Id$. The equivalence of (iii) can be derived in a similar way. □

One might ask what distinguishes the canonical dual frame from the alternate dual frames besides its explicit formula in terms of the initial frame. Another seemingly different question is which properties of the coefficient sequence in the decomposition of some signal x in terms of the frame (see Theorem 1.8),

$$x = \sum_{i=1}^{M} \langle x, S^{-1}\varphi_i \rangle \varphi_i,$$

uniquely distinguishes it from other coefficient sequences; redundancy allows infinitely many coefficient sequences. Interestingly, the next result answers both questions simultaneously by stating that this coefficient sequence has minimal ℓ_2-norm among all sequences—in particular those, with respect to alternate dual frames—representing x.

Proposition 1.16 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S, and let $x \in \mathscr{H}^N$. If $(a_i)_{i=1}^M$ are scalars such that $x = \sum_{i=1}^M a_i \varphi_i$, then*

$$\sum_{i=1}^{M} |a_i|^2 = \sum_{i=1}^{M} \left|\langle x, S^{-1}\varphi_i \rangle\right|^2 + \sum_{i=1}^{M} \left|a_i - \langle x, S^{-1}\varphi_i \rangle\right|^2.$$

Proof Letting T denote the analysis operator of $(\varphi_i)_{i=1}^M$, we obtain

$$\left(\langle x, S^{-1}\varphi_i \rangle\right)_{i=1}^M = \left(\langle S^{-1}x, \varphi_i \rangle\right)_{i=1}^M \in \operatorname{ran} T.$$

Since $x = \sum_{i=1}^M a_i \varphi_i$, it follows that

$$\left(a_i - \langle x, S^{-1}\varphi_i \rangle\right)_{i=1}^M \in \ker T^* = (\operatorname{ran} T)^{\perp}.$$

Considering the decomposition

$$(a_i)_{i=1}^M = \left(\langle x, S^{-1}\varphi_i \rangle\right)_{i=1}^M + \left(a_i - \langle x, S^{-1}\varphi_i \rangle\right)_{i=1}^M,$$

the claim is immediate. □

Corollary 1.8 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$, and let $(\psi_i)_{i=1}^M$ be an associated alternate dual frame. Then, for all $x \in \mathscr{H}^N$,*

$$\left\| \left(\langle x, S^{-1}\varphi_i \rangle\right)_{i=1}^M \right\|_2 \leq \left\| \left(\langle x, \psi_i \rangle\right)_{i=1}^M \right\|_2.$$

We wish to mention that sequences which are minimal in the ℓ_1-norm also play a crucial role to date due to the fact that the ℓ_1-norm promotes sparsity. The interested reader is referred to Chap. 9 for further details.

1.5.2 Properties of Dual Frames

While we focused on properties of the canonical dual frame in the last subsection, we next discuss properties shared by all dual frames. The first question arising is: How do you characterize all dual frames? A comprehensive answer is provided by the following result.

Proposition 1.17 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T and frame operator S. Then the following conditions are equivalent.*

(i) $(\psi_i)_{i=1}^M$ *is a dual frame for* $(\varphi_i)_{i=1}^M$.
(ii) *The analysis operator T_1 of the sequence $(\psi_i - S^{-1}\varphi_i)_{i=1}^M$ satisfies*

$$\operatorname{ran} T \perp \operatorname{ran} T_1.$$

Proof We set $\tilde{\varphi}_i := \psi_i - S^{-1}\varphi_i$ for $i = 1, \ldots, M$ and note that

$$\sum_{i=1}^M \langle x, \psi_i\rangle \varphi_i = \sum_{i=1}^M \big\langle x, \tilde{\varphi}_i + S^{-1}\varphi_i\big\rangle \varphi_i = x + \sum_{i=1}^M \langle x, \tilde{\varphi}_i\rangle \varphi_i = x + T^*T_1 x$$

holds for all $x \in \mathscr{H}^N$. Hence, $(\psi_i)_{i=1}^M$ is a dual frame for $(\varphi_i)_{i=1}^M$ if and only if $T^*T_1 = 0$. But this is equivalent to (ii). □

From this result, we have the following corollary which provides a general formula for all dual frames.

Corollary 1.9 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T and frame operator S with associated normalized eigenvectors $(e_j)_{j=1}^N$ and respective eigenvalues $(\lambda_j)_{j=1}^N$. Then every dual frame $\{\psi_i\}_{i=1}^M$ for $(\varphi_i)_{i=1}^M$ is of the form*

$$\psi_i = \sum_{j=1}^N \left(\frac{1}{\lambda_j} \langle \varphi_i, e_j\rangle + \overline{h_{ij}} \right) e_j, \quad i = 1, \ldots, M,$$

where each $(h_{ij})_{i=1}^M$, $j = 1, \ldots, N$, is an element of $(\operatorname{ran} T)^\perp$.

Proof If ψ_i, $i = 1, \ldots, M$, is of the given form with sequences $(h_{ij})_{i=1}^M \in \ell_2^M$, $j = 1, \ldots, N$, then $\psi_i = S^{-1}\varphi_i + \tilde{\varphi}_i$, where $\tilde{\varphi}_i := \sum_{j=1}^N \overline{h_{ij}} e_j$, $i = 1, \ldots, M$. The analysis operator $\tilde{T}$ of $(\tilde{\varphi}_i)_{i=1}^M$ satisfies $\tilde{T} e_j = (h_{ij})_{i=1}^M$. The claim follows from this observation. □

As a second corollary, we derive a characterization of all frames which have a uniquely determined dual frame. It is evident that this unique dual frame coincides with the canonical dual frame.

Corollary 1.10 *A frame $(\varphi_i)_{i=1}^M$ for $\mathscr{H}^N$ has a unique dual frame if and only if $M = N$.*

1.5.3 Frame Algorithms

Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S, and assume we are given the image of a signal $x \in \mathscr{H}^N$ under the analysis operator, i.e., the sequence $(\langle x, \varphi_i\rangle)_{i=1}^M$ in ℓ_2^M. Theorem 1.8 has already provided us with the reconstruction formula

$$x = \sum_{i=1}^{M} \langle x, \varphi_i\rangle S^{-1}\varphi_i$$

by using the canonical dual frame. Since inversion is typically not only computationally expensive, but also numerically instable, this formula might not be utilizable in practice.

To resolve this problem, we will next discuss three iterative methods to derive a converging sequence of approximations of x from the knowledge of $(\langle x, \varphi_i\rangle)_{i=1}^M$. The first on our list is called the *frame algorithm*.

Proposition 1.18 (Frame Algorithm) *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame bounds A, B and frame operator S. Given a signal $x \in \mathscr{H}^N$, define a sequence $(y_j)_{j=0}^\infty$ in $\mathscr{H}^N$ by*

$$y_0 = 0, \qquad y_j = y_{j-1} + \frac{2}{A+B} S(x - y_{j-1}) \quad \text{for all } j \geq 1.$$

Then $(y_j)_{j=0}^\infty$ converges to x in $\mathscr{H}^N$, and the rate of convergence is

$$\|x - y_j\| \leq \left(\frac{B-A}{B+A}\right)^j \|x\|, \quad j \geq 0.$$

Proof First, for all $x \in \mathscr{H}^N$, we have

$$\left\langle \left(Id - \frac{2}{A+B}S\right)x, x\right\rangle = \|x\|^2 - \frac{2}{A+B}\sum_{i=1}^{M} \left|\langle x, \varphi_i\rangle\right|^2 \leq \|x\|^2 - \frac{2A}{A+B}\|x\|^2$$
$$= \frac{B-A}{A+B}\|x\|^2.$$

Similarly, we obtain

$$-\frac{B-A}{B+A}\|x\|^2 \leq \left\langle \left(Id - \frac{2}{A+B}S\right)x, x\right\rangle,$$

which yields

$$\left\| Id - \frac{2}{A+B} S \right\| \leq \frac{B-A}{A+B}. \tag{1.3}$$

By the definition of y_j, for any $j \geq 0$,

$$x - y_j = x - y_{j-1} - \frac{2}{A+B} S(x - y_{j-1}) = \left(Id - \frac{2}{A+B} S \right)(x - y_{j-1}).$$

Iterating this calculation, we derive

$$x - y_j = \left(Id - \frac{2}{A+B} S \right)^j (x - y_0), \quad \text{for all } j \geq 0.$$

Thus, by (1.3),

$$\begin{aligned} \|x - y_j\| &= \left\| \left(Id - \frac{2}{A+B} S \right)^j (x - y_0) \right\| \\ &\leq \left\| Id - \frac{2}{A+B} S \right\|^j \|x - y_0\| \\ &\leq \left(\frac{B-A}{A+B} \right)^j \|x\|. \end{aligned}$$

The result is proved. □

Note that, although the iteration formula in the frame algorithm contains x, the algorithm does not depend on the knowledge of x but only on the frame coefficients $(\langle x, \varphi_i \rangle)_{i=1}^M$, since $y_j = y_{j-1} + \frac{2}{A+B}(\sum_i \langle x, \varphi_i \rangle \varphi_i - S y_{j-1})$.

One drawback of the frame algorithm is the fact that not only does the convergence rate depend on the ratio of the frame bounds, i.e., the condition number of the frame, but it depends on it in a highly sensitive way. This causes the problem that a large ratio of the frame bounds leads to very slow convergence.

To tackle this problem, in [96], the *Chebyshev method* and the *conjugate gradient methods* were introduced, which are significantly better adapted to frame theory and lead to faster convergence than the frame algorithm. These two algorithms will next be discussed. We start with the *Chebyshev algorithm*.

Proposition 1.19 (Chebyshev Algorithm, [96]) *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame bounds A, B and frame operator S, and set*

$$\rho := \frac{B-A}{B+A} \quad \textit{and} \quad \sigma := \frac{\sqrt{B} - \sqrt{A}}{\sqrt{B} + \sqrt{A}}.$$

Given a signal $x \in \mathscr{H}^N$, *define a sequence* $(y_j)_{j=0}^{\infty}$ *in* $\mathscr{H}^N$ *and corresponding scalars* $(\lambda_j)_{j=1}^{\infty}$ *by*

$$y_0 = 0, \qquad y_1 = \frac{2}{B+A} Sx, \quad \text{and} \quad \lambda_1 = 2,$$

and for $j \geq 2$, *set*

$$\lambda_j = \frac{1}{1 - \frac{\rho^2}{4}\lambda_{j-1}} \quad \text{and} \quad y_j = \lambda_j \left(y_{j-1} - y_{j-2} + \frac{2}{B+A} S(x - y_{j-1}) \right) + y_{j-2}.$$

Then $(y_j)_{j=0}^{\infty}$ *converges to* x *in* $\mathscr{H}^N$, *and the rate of convergence is*

$$\|x - y_j\| \leq \frac{2\sigma^j}{1 + \sigma^{2j}} \|x\|.$$

The advantage of the *conjugate gradient method*, which we will present next, is the fact that it does not require knowledge of the frame bounds. However, as before, the rate of convergence certainly does depend on them.

Proposition 1.20 (Conjugate Gradient Method, [96]) *Let* $(\varphi_i)_{i=1}^{M}$ *be a frame for* $\mathscr{H}^N$ *with frame operator* S. *Given a signal* $x \in \mathscr{H}^N$, *define three sequences* $(y_j)_{j=0}^{\infty}$, $(r_j)_{j=0}^{\infty}$, *and* $(p_j)_{j=-1}^{\infty}$ *in* $\mathscr{H}^N$ *and corresponding scalars* $(\lambda_j)_{j=-1}^{\infty}$ *by*

$$y_0 = 0, \qquad r_0 = p_0 = Sx, \quad \text{and} \quad p_{-1} = 0,$$

and for $j \geq 0$, *set*

$$\lambda_j = \frac{\langle r_j, p_j \rangle}{\langle p_j, Sp_j \rangle}, \qquad y_{j+1} = y_j + \lambda_j p_j, \qquad r_{j+1} = r_j - \lambda_j Sp_j,$$

and

$$p_{j+1} = Sp_j - \frac{\langle Sp_j, Sp_j \rangle}{\langle p_j, Sp_j \rangle} p_j - \frac{\langle Sp_j, Sp_{j-1} \rangle}{\langle p_{j-1}, Sp_{j-1} \rangle} p_{j-1}.$$

Then $(y_j)_{j=0}^{\infty}$ *converges to* x *in* $\mathscr{H}^N$, *and the rate of convergence is*

$$|||x - y_j||| \leq \frac{2\sigma^j}{1 + \sigma^{2j}} |||x||| \quad \text{with } \sigma = \frac{\sqrt{B} - \sqrt{A}}{\sqrt{B} + \sqrt{A}},$$

and $|||\cdot|||$ *is the norm on* $\mathscr{H}^N$ *given by* $|||x||| = \langle x, Sx \rangle^{1/2} = \|S^{1/2}x\|$, $x \in \mathscr{H}^N$.

1.6 Construction of Frames

Applications often require the construction of frames with certain desired properties. As a result of the large diversity of these desiderata, there exists a large number of

construction methods [36, 58]. In this section, we will present a prominent selection of these. For further details and results, for example, the construction of frames through Spectral Tetris [30, 43, 46] and through eigensteps [29], we refer to Chap. 2.

1.6.1 Tight and Parseval Frames

Tight frames are particularly desirable due to the fact that the reconstruction of a signal from tight frame coefficients is numerically optimally stable, as discussed in Sect. 1.5. Most of the constructions we will present modify a given frame so that the result is a tight frame.

We start with the most basic result for generating a Parseval frame, which is the application of $S^{-1/2}$, S being the frame operator.

Lemma 1.7 *If $(\varphi_i)_{i=1}^M$ is a frame for $\mathscr{H}^N$ with frame operator S, then $(S^{-1/2}\varphi_i)_{i=1}^M$ is a Parseval frame.*

Proof By Proposition 1.10, the frame operator for $(S^{-1/2}\varphi_i)_{i=1}^M$ is $S^{-1/2}SS^{-1/2} = Id$. □

Although this result is impressive in its simplicity, from a practical point of view it has various problems, the most significant being that this procedure requires inversion of the frame operator.

However, Lemma 1.7 can certainly be applied if all eigenvalues and respective eigenvectors of the frame operator are given. If only information on the eigenspace corresponding to the largest eigenvalue is missing, then there exists a simple practical method to generate a tight frame by adding a provably minimal number of vectors.

Proposition 1.21 *Let $(\varphi_i)_{i=1}^M$ be any family of vectors in $\mathscr{H}^N$ with frame operator S having eigenvectors $(e_j)_{j=1}^N$ and respective eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$. Let $1 \leq k \leq N$ be such that $\lambda_1 = \lambda_2 = \cdots = \lambda_k > \lambda_{k+1}$. Then*

$$(\varphi_i)_{i=1}^M \cup \left((\lambda_1 - \lambda_j)^{1/2} e_j\right)_{j=k+1}^N \tag{1.4}$$

forms a λ_1-tight frame for $\mathscr{H}^N$.

Moreover, $N-k$ is the least number of vectors which can be added to $(\varphi_i)_{i=1}^M$ to obtain a tight frame.

Proof A straightforward calculation shows that the sequence in (1.4) is indeed a λ_1-tight frame for $\mathscr{H}^N$.

For the *moreover* part, assume that there exist vectors $(\psi_j)_{j\in J}$ with frame operator S_1 satisfying that $(\varphi_i)_{i=1}^M \cup (\psi_j)_{j\in J}$ is an A-tight frame. This implies $A \geq \lambda_1$.

Now define S_2 to be the operator on $\mathscr{H}^N$ given by

$$S_2 e_j = \begin{cases} 0: & 1 \le j \le k, \\ (\lambda_1 - \lambda_j) e_j: & k+1 \le j \le N. \end{cases}$$

It follows that $A \cdot Id = S + S_1$ and

$$S_1 = A \cdot Id - S \ge \lambda_1 Id - S = S_2.$$

Since S_2 has $N-k$ nonzero eigenvalues, S_1 also has at least $N-k$ nonzero eigenvalues. Hence $|J| \ge N-k$, showing that indeed $N-k$ added vectors is minimal. □

Before we delve into further explicit constructions, we need to first state some fundamental results on tight, and, in particular, Parseval frames.

The most basic invariance property a frame could have is invariance under orthogonal projections. The next result shows that this operation indeed maintains and may even improve the frame bounds. In particular, the orthogonal projection of a Parseval frame remains a Parseval frame.

Proposition 1.22 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame bounds A, B, and let P be an orthogonal projection of $\mathscr{H}^N$ onto a subspace $\mathscr{W}$. Then $(P\varphi_i)_{i=1}^M$ is a frame for $\mathscr{W}$ with frame bounds A, B.*

In particular, if $(\varphi_i)_{i=1}^M$ is a Parseval frame for $\mathscr{H}^N$ and P is an orthogonal projection on $\mathscr{H}^N$ onto $\mathscr{W}$, then $(P\varphi_i)_{i=1}^M$ is a Parseval frame for $\mathscr{W}$.

Proof For any $x \in \mathscr{W}$,

$$A\|x\|^2 = A\|Px\|^2 \le \sum_{i=1}^M \big|\langle Px, \varphi_i\rangle\big|^2 = \sum_{i=1}^M \big|\langle x, P\varphi_i\rangle\big|^2 \le B\|Px\|^2 = B\|x\|^2.$$

This proves the claim. The *in particular* part follows immediately. □

Proposition 1.22 immediately yields the following corollary.

Corollary 1.11 *Let $(e_i)_{i=1}^N$ be an orthonormal basis for $\mathscr{H}^N$, and let P be an orthogonal projection of $\mathscr{H}^N$ onto a subspace $\mathscr{W}$. Then $(Pe_i)_{i=1}^N$ is a Parseval frame for $\mathscr{W}$.*

Corollary 1.11 can be interpreted in the following way: Given an $M \times M$ unitary matrix, if we select any N rows from the matrix, then the column vectors from these rows form a Parseval frame for $\mathscr{H}^N$. The next theorem, known as *Naimark's theorem*, shows that indeed every Parseval frame can be obtained as the result of this kind of operation.

Theorem 1.9 (Naimark's Theorem) *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T, let $(e_i)_{i=1}^M$ be the standard basis of ℓ_2^M, and let $P : \ell_2^M \to \ell_2^M$ be the orthogonal projection onto* $\operatorname{ran} T$. *Then the following conditions are equivalent.*

(i) $(\varphi_i)_{i=1}^M$ *is a Parseval frame for* $\mathscr{H}^N$.
(ii) *For all* $i = 1, \ldots, M$, *we have* $Pe_i = T\varphi_i$.
(iii) *There exist* $\psi_1, \ldots, \psi_M \in \mathscr{H}^{M-N}$ *such that* $(\varphi_i \oplus \psi_i)_{i=1}^M$ *is an orthonormal basis of* $\mathscr{H}^M$.

Moreover, if (iii) *holds, then* $(\psi_i)_{i=1}^M$ *is a Parseval frame for* $\mathscr{H}^{M-N}$. *If* $(\psi_i')_{i=1}^M$ *is another Parseval frame as in* (iii), *then there exists a unique linear operator* L *on* $\mathscr{H}^{M-N}$ *such that* $L\psi_i = \psi_i'$, $i = 1, \ldots, M$, *and* L *is unitary.*

Proof (i)⇔(ii) By Theorem 1.7(iii) $(\varphi_i)_{i=1}^M$ is a Parseval frame if and only if $TT^* = P$. Therefore, (i) and (ii) are equivalent due to $T^*e_i = \varphi_i$ for all $i = 1, \ldots, M$.

(i)⇒(iii) We set $c_i := e_i - T\varphi_i$, $i = 1, \ldots, M$. Then, by (ii), $c_i \in (\operatorname{ran} T)^\perp$ for all i. Let $\Phi : (\operatorname{ran} T)^\perp \to \mathscr{H}^{M-N}$ be unitary and put $\psi_i := \Phi c_i$, $i = 1, \ldots, M$. Then, since T is isometric,

$$\langle \varphi_i \oplus \psi_i, \varphi_k \oplus \psi_k \rangle = \langle \varphi_i, \varphi_k \rangle + \langle \psi_i, \psi_k \rangle = \langle T\varphi_i, T\varphi_k \rangle + \langle c_i, c_k \rangle = \delta_{ik},$$

which proves (iii).

(iii)⇒(i) This follows directly from Corollary 1.11.

Concerning the *moreover* part, it follows from Corollary 1.11 that $(\psi_i)_{i=1}^M$ is a Parseval frame for $\mathscr{H}^{M-N}$. Let $(\psi_i')_{i=1}^M$ be another Parseval frame as in (iii) and denote the analysis operators of $(\psi_i)_{i=1}^M$ and $(\psi_i')_{i=1}^M$ by F and F', respectively. We make use of the decomposition $\mathscr{H}^M = \mathscr{H}^N \oplus \mathscr{H}^{M-N}$. Note that both $U := (T, F)$ and $U' := (T, F')$ are unitary operators from $\mathscr{H}^M$ onto ℓ_2^M. By P_{M-N} denote the projection of $\mathscr{H}^M$ onto $\mathscr{H}^{M-N}$ and set

$$L := P_{M-N}U'^*U|_{\mathscr{H}^{M-N}} = P_{M-N}U'^*F.$$

Let $y \in \mathscr{H}^N$. Then, since $U|_{\mathscr{H}^N} = U'|_{\mathscr{H}^N} = T$, we have $P_{M-N}U'^*Uy = P_{M-N}y = 0$. Hence,

$$L\psi_i = P_{M-N}U'^*U(\varphi_i \oplus \psi_i) = P_{M-N}U'^*e_i = P_{M-N}\big(\varphi_i \oplus \psi_i'\big) = \psi_i'.$$

The uniqueness of L follows from the fact that both $(\psi_i)_{i=1}^M$ and $(\psi_i')_{i=1}^M$ are spanning sets for $\mathscr{H}^{M-N}$.

To show that L is unitary, we observe that, by Proposition 1.10, the frame operator of $(L\psi_i)_{i=1}^M$ is given by LL^*. The claim $LL^* = Id$ now follows from the fact that the frame operator of $(\psi_i')_{i=1}^M$ is also the identity. □

The simplest way to construct a frame from a given one is just to scale the frame vectors. Therefore, it seems desirable to have a characterization of the class of

frames which can be scaled to a Parseval frame or a tight frame (which is equivalent). We term such frames scalable.

Definition 1.21 A frame $(\varphi_i)_{i=1}^M$ for $\mathscr{H}^N$ is called (*strictly*) *scalable*, if there exist nonnegative (respectively, positive) numbers $a_1, \dots, a_M$ such that $(a_i\varphi_i)_{i=1}^M$ is a Parseval frame.

The next result is closely related to Naimark's theorem.

Theorem 1.10 [116] *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with analysis operator T. Then the following statements are equivalent.*

(i) *$(\varphi_i)_{i=1}^M$ is strictly scalable.*
(ii) *There exists a linear operator $L : \mathscr{H}^{M-N} \to \ell_2^M$ such that $TT^* + LL^*$ is a positive definite diagonal matrix.*
(iii) *There exists a sequence $(\psi_i)_{i=1}^M$ of vectors in $\mathscr{H}^{M-N}$ such that $(\varphi_i \oplus \psi_i)_{i=1}^M$ forms a complete orthogonal system in $\mathscr{H}^M$.*

If $\mathscr{H}^N$ is real, then the following result applies, which can be utilized to derive a geometric interpretation of scalability. For this we once more refer to [116].

Theorem 1.11 [116] *Let $\mathscr{H}^N$ be real and let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ without zero vectors. Then the following statements are equivalent.*

(i) *$(\varphi_i)_{i=1}^M$ is not scalable.*
(ii) *There exists a self-adjoint operator Y on $\mathscr{H}^N$ with $\operatorname{Tr}(Y) < 0$ and $\langle Y\varphi_i, \varphi_i\rangle \geq 0$ for all $i = 1, \dots, M$.*
(iii) *There exists a self-adjoint operator Y on $\mathscr{H}^N$ with $\operatorname{Tr}(Y) = 0$ and $\langle Y\varphi_i, \varphi_i\rangle > 0$ for all $i = 1, \dots, M$.*

We finish this subsection with an existence result of tight frames with prescribed norms of the frame vectors. Its proof in [44] heavily relies on a deep understanding of the frame potential and is a pure existence proof. However, in special cases constructive methods are presented in [56].

Theorem 1.12 [44] *Let $N \leq M$, and let $a_1 \geq a_2 \geq \cdots \geq a_M$ be positive real numbers. Then the following conditions are equivalent.*

(i) *There exists a tight frame $(\varphi_i)_{i=1}^M$ for $\mathscr{H}^N$ satisfying $\|\varphi_i\| = a_i$ for all $i = 1, 2, \dots, M$.*
(ii) *For all $1 \leq j < N$,*

$$a_j^2 \leq \frac{\sum_{i=j+1}^M a_i^2}{N-j}.$$

(iii) *We have*

$$\sum_{i=1}^{M} a_i^2 \geq N a_1^2.$$

Equal norm tight frames are even more desirable, but are difficult to construct. A powerful method, called *Spectral Tetris*, for such constructions was recently derived in [46], see Chap. 2. This methodology even generates sparse frames [49], which reduce the computational complexity and also ensure high compressibility of the synthesis matrix—which then is a sparse matrix. However, we caution the reader that Spectral Tetris has the drawback that it often generates multiple copies of the same frame vector. For practical applications, this is typically avoided, since the frame coefficients associated with a repeated frame vector do not provide any new information about the incoming signal.

1.6.2 Frames with Given Frame Operator

It is often desirable not only to construct tight frames, but more generally to construct frames with a prescribed frame operator. Typically in such a case the eigenvalues of the frame operator are given assuming that the eigenvectors are the standard basis. Applications include, for instance, noise reduction if colored noise is present.

The first comprehensive results containing necessary and sufficient conditions for the existence and the construction of tight frames with frame vectors of a prescribed norm were derived in [44] and [56]; see also Theorem 1.12. The result in [44] was then extended in [57] to the following theorem, which now also includes prescribing the eigenvalues of the frame operator.

Theorem 1.13 [57] *Let S be a positive self-adjoint operator on $\mathscr{H}^N$, and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N > 0$ be the eigenvalues of S. Further, let $M \geq N$, and let $c_1 \geq c_2 \geq \cdots \geq c_M$ be positive real numbers. Then the following conditions are equivalent.*

(i) *There exists a frame $(\varphi_i)_{i=1}^M$ for $\mathscr{H}^N$ with frame operator S satisfying $\|\varphi_i\| = c_i$ for all $i = 1, 2, \ldots, M$.*
(ii) *For every $1 \leq k \leq N$, we have*

$$\sum_{j=1}^{k} c_j^2 \leq \sum_{j=1}^{k} \lambda_j \quad \text{and} \quad \sum_{i=1}^{M} c_i^2 = \sum_{j=1}^{N} \lambda_j.$$

However, it is often preferable to utilize equal norm frames, since then, roughly speaking, each vector provides the same coverage for the space. In [57], it was shown that there always exists an equal norm frame with a prescribed frame operator. This is the content of the next result.

Theorem 1.14 [57] *For every $M \geq N$ and every invertible positive self-adjoint operator S on $\mathscr{H}^N$ there exists an equal norm frame for $\mathscr{H}^N$ with M elements and frame operator S. In particular, there exist equal norm Parseval frames with M elements in $\mathscr{H}^N$ for every $N \leq M$.*

Proof We define the norm of the to-be-constructed frame to be c, where

$$c^2 = \frac{1}{M} \sum_{j=1}^{N} \lambda_j.$$

It is sufficient to prove that the conditions in Theorem 1.13(ii) are satisfied for $c_i = c$ for all $i = 1, 2, \ldots, M$. The definition of c immediately implies the second condition.

For the first condition, we observe that

$$c_1^2 = c^2 = \frac{1}{M} \sum_{j=1}^{N} \lambda_j \leq \lambda_1.$$

Hence this condition holds for $j = 1$. Now, toward a contradiction, assume that there exists some $k \in \{2, \ldots, N\}$ for which this condition fails for the first time by counting from 1 upward, i.e.,

$$\sum_{j=1}^{k-1} c_j^2 = (k-1)c^2 \leq \sum_{j=1}^{k-1} \lambda_j, \quad \text{but} \quad \sum_{j=1}^{k} c_j^2 = kc^2 > \sum_{j=1}^{k} \lambda_j.$$

This implies

$$c^2 \geq \lambda_k \quad \text{and thus} \quad c^2 \geq \lambda_j \quad \text{for all } k+1 \leq j \leq N.$$

Hence,

$$Mc^2 \geq kc^2 + (N-k)c^2 > \sum_{j=1}^{k} \lambda_j + \sum_{j=k+1}^{N} c_j^2 \geq \sum_{j=1}^{N} \lambda_j + \sum_{j=k+1}^{N} \lambda_j = \sum_{j=1}^{N} \lambda_j,$$

which is a contradiction. The proof is completed. □

By an extension of the aforementioned algorithm *Spectral Tetris* [30, 43, 47, 49] to non-tight frames, Theorem 1.14 can be constructively realized. The interested reader is referred to Chap. 2. We also mention that an extension of Spectral Tetris to construct fusion frames (cf. Sect. 1.9) exists. Further details on this topic are contained in Chap. 13.

1.6.3 Full Spark Frames

Generic frames are those optimally resilient against erasures. The precise definition is as follows.

Definition 1.22 A frame $(\varphi_i)_{i=1}^M$ for $\mathscr{H}^N$ is called a *full spark frame*, if the erasure of any $M-N$ vectors leaves a frame; i.e., for any $I \subset \{1,\dots,M\}$, $|I| = M-N$, the sequence $(\varphi_i)_{i=1, i\notin I}^M$ is still a frame for $\mathscr{H}^N$.

It is evident that such frames are of significant importance for applications. A first study was undertaken in [126]. Recently, using methods from algebraic geometry, equivalence classes of full spark frames were extensively studied [26, 80, 135]. It was shown, for instance, that equivalence classes of full spark frames are dense in the Grassmannian variety. For the readers to be able to appreciate these results, Chap. 4 provides an introduction to algebraic geometry followed by a survey about this and related results.

1.7 Frame Properties

As already discussed, crucial properties of frames such as erasure robustness, resilience against noise, or sparse approximation properties originate from spanning and independence properties of frames [13], which are typically based on the Rado-Horn theorem [103, 128] and its redundant version [54]. These, in turn, are only possible because of their redundancy [12]. This section will shed some light on these issues.

1.7.1 Spanning and Independence

As is intuitively clear, the frame bounds imply certain spanning properties which are detailed in the following result. This theorem should be compared to Lemma 1.2, which presented some first statements about spanning sets in frames.

Theorem 1.15 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame bounds A and B. Then the following holds.*

(i) $\|\varphi_i\|^2 \le B_{\mathrm{op}}$ *for all* $i = 1, 2, \dots, M$.
(ii) *If, for some* $i_0 \in \{1,\dots,M\}$, *we have* $\|\varphi_{i_0}\|^2 = B_{\mathrm{op}}$, *then* $\varphi_{i_0} \perp \operatorname{span}\{\varphi_i\}_{i=1, i\neq i_0}^M$.
(iii) *If, for some* $i_0 \in \{1,\dots,M\}$, *we have* $\|\varphi_{i_0}\|^2 < A_{\mathrm{op}}$, *then* $\varphi_{i_0} \in \operatorname{span}\{\varphi_i\}_{i=1, i\neq i_0}^M$.

In particular, if $(\varphi_i)_{i=1}^M$ is a Parseval frame, then either $\varphi_{i_0} \perp \operatorname{span}\{\varphi_i\}_{i=1, i\neq i_0}^M$ (and in this case $\|\varphi_i\| = 1$) or $\|\varphi_{i_0}\| < 1$.

Proof For any $i_0 \in \{1,\dots,M\}$ we have

$$\|\varphi_{i_0}\|^4 \le \|\varphi_{i_0}\|^4 + \sum_{i\neq i_0} \big|\langle \varphi_{i_0}, \varphi_i\rangle\big|^2 = \sum_{i=1}^M \big|\langle \varphi_{i_0}, \varphi_i\rangle\big|^2 \le B_{\mathrm{op}}\|\varphi_{i_0}\|^2. \tag{1.5}$$

The claims (i) and (ii) now directly follow from (1.5).

(iii) Let P denote the orthogonal projection of $\mathscr{H}^N$ onto $(\text{span}\{\varphi_i\}_{i=1,i\neq i_0}^M)^\perp$. Then

$$A_{\text{op}}\|P\varphi_{i_0}\|^2 \leq \|P\varphi_{i_0}\|^4 + \sum_{i=1,i\neq i_0}^M \left|\langle P\varphi_{i_0}, \varphi_i\rangle\right|^2 = \|P\varphi_{i_0}\|^4.$$

Hence, either $P\varphi_{i_0} = 0$ (and thus $\varphi_{i_0} \in \text{span}\{\varphi_i\}_{i=1,i\neq i_0}^M$) or $A_{\text{op}} \leq \|P\varphi_{i_0}\|^2 \leq \|\varphi_{i_0}\|^2$. This proves (iii). □

Ideally, we are interested in having an exact description of a frame in terms of its spanning and independence properties. The following questions could be answered by such a measure: How many disjoint linearly independent spanning sets does the frame contain? After removing these, how many disjoint linearly independent sets which span hyperplanes does it contain? And many more.

One of the main results in this direction is the following from [13].

Theorem 1.16 [13] *Every unit norm tight frame $(\varphi_i)_{i=1}^M$ for $\mathscr{H}^N$ with $M = kN + j$ elements, $0 \leq j < N$, can be partitioned into k linearly independent spanning sets plus a linearly independent set of j elements.*

For its proof and further related results we refer to Chap. 3.

1.7.2 Redundancy

As we have discussed and will be seen throughout this book, redundancy is the key property of frames. This fact makes it even more surprising that, until recently, not much attention has been paid to introduce meaningful quantitative measures of redundancy. The classical measure of the *redundancy* of a frame $(\varphi_i)_{i=1}^M$ for $\mathscr{H}^N$ is the quotient of the number of frame vectors and the dimension of the ambient space, i.e., $\frac{M}{N}$. However, this measure has serious problems in distinguishing, for instance, the two frames in Example 1.2 (1) and (2) by assigning the same redundancy measure $\frac{2N}{N} = 2$ to both of them. From a frame perspective these two frames are very different, since, for instance, one contains two spanning sets whereas the other just contains one.

Recently, in [12] a new notion of redundancy was proposed which seems to better capture the spirit of what redundancy should represent. To present this notion, let $\mathbb{S} = \{x \in \mathscr{H}^N : \|x\| = 1\}$ denote the unit sphere in $\mathscr{H}^N$, and let $P_{\text{span}\{x\}}$ denote the orthogonal projection onto the subspace $\text{span}\{x\}$ for some $x \in \mathscr{H}^N$.

Definition 1.23 Let $\Phi = (\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$. For each $x \in \mathbb{S}$, the *redundancy function* $\mathscr{R}_\Phi : \mathbb{S} \to \mathbb{R}^+$ is defined by

$$\mathscr{R}_\Phi(x) = \sum_{i=1}^M \|P_{\text{span}\{\varphi_i\}}x\|^2.$$

Then the *upper redundancy* of Φ is defined by

$$\mathscr{R}_{\Phi}^{+} = \max_{x \in \mathbb{S}} \mathscr{R}_{\Phi}(x),$$

and the *lower redundancy* of Φ is defined by

$$\mathscr{R}_{\Phi}^{-} = \min_{x \in \mathbb{S}} \mathbb{R}_{\Phi}(x).$$

Moreover, Φ has *uniform redundancy*, if

$$\mathscr{R}_{\Phi}^{-} = \mathscr{R}_{\Phi}^{+}.$$

One might hope that this new notion of redundancy provides information about spanning and independence properties of the frame, since these are closely related to questions such as, say, whether a frame is resilient with respect to deletion of a particular number of frame vectors. Indeed, such a link exists and is detailed in the next result.

Theorem 1.17 [12] *Let $\Phi = (\varphi_i)_{i=1}^{M}$ be a frame for $\mathscr{H}^N$ without zero vectors. Then the following conditions hold.*

(i) *Φ contains $\lfloor \mathscr{R}_{\Phi}^{-} \rfloor$ disjoint spanning sets.*
(ii) *Φ can be partitioned into $\lceil \mathscr{R}_{\Phi}^{+} \rceil$ linearly independent sets.*

Various other properties of this notion of redundancy are known, such as additivity or its range, and we refer to [12] and Chap. 3 for more details.

At this point, we point out that this notion of upper and lower redundancy coincides with the optimal frame bounds of the normalized frame $(\frac{\varphi_i}{\|\varphi_i\|})_{i=1}^{M}$, after deletion of zero vectors. The crucial point is that with this viewpoint Theorem 1.17 combines analytic and algebraic properties of Φ.

1.7.3 Equivalence of Frames

We now consider equivalence classes of frames. As in other research areas, the idea is that frames in the same equivalence class share certain properties.

1.7.3.1 Isomorphic frames

The following definition states one equivalence relation for frames.

Definition 1.24 Two frames $(\varphi_i)_{i=1}^{M}$ and $(\psi_i)_{i=1}^{M}$ for $\mathscr{H}^N$ are called *isomorphic*, if there exists an operator $F : \mathscr{H}^N \to \mathscr{H}^N$ satisfying $F\varphi_i = \psi_i$ for all $i = 1, 2, \ldots, M$.

We remark that—due to the spanning property of frames—an operator F as in the above definition is both invertible and unique. Moreover, note that in [4] the isomorphy of frames with an operator F as above was termed F-equivalence.

The next theorem characterizes the isomorphy of two frames in terms of their analysis and synthesis operators.

Theorem 1.18 *Let $(\varphi_i)_{i=1}^M$ and $(\psi_i)_{i=1}^M$ be frames for $\mathscr{H}^N$ with analysis operators T_1 and T_2, respectively. Then the following conditions are equivalent.*

(i) $(\varphi_i)_{i=1}^M$ *is isomorphic to* $(\psi_i)_{i=1}^M$.
(ii) $\operatorname{ran} T_1 = \operatorname{ran} T_2$.
(iii) $\ker T_1^* = \ker T_2^*$.

If one of (i)–(iii) *holds, then the operator $F : \mathscr{H}^N \to \mathscr{H}^N$ with $F\varphi_i = \psi_i$ for all $i = 1, \dots, N$ is given by $F = T_2^*(T_1^*|_{\operatorname{ran} T_1})^{-1}$.*

Proof The equivalence of (ii) and (iii) follows by orthogonal complementation. In the following let $(e_i)_{i=1}^M$ denote the standard unit vector basis of ℓ_2^M.

(i)⇒(iii) Let F be an invertible operator on $\mathscr{H}^N$ such that $F\varphi_i = \psi_i$ for all $i = 1, \dots, M$. Then Proposition 1.9 implies $T_2 = T_1 F^*$ and hence $FT_1^* = T_2^*$. Since F is invertible, (iii) follows.

(ii)⇒(i) Let P be the orthogonal projection onto $\mathscr{W} := \operatorname{ran} T_1 = \operatorname{ran} T_2$. Then $\varphi_i = T_1^* e_i = T_1^* P e_i$ and $\psi_i = T_2^* e_i = T_2^* P e_i$. The operators T_1^* and T_2^* both map $\mathscr{W}$ bijectively onto $\mathscr{H}^N$. Therefore, the operator $F := T_2^*(T_1^*|_{\mathscr{W}})^{-1}$ maps $\mathscr{H}^N$ bijectively onto itself. Consequently, for each $i \in \{1, \dots, M\}$ we have

$$F\varphi_i = T_2^*\big(T_1^*|_{\mathscr{W}}\big)^{-1} T_1^* P e_i = T_2^* P e_i = \psi_i,$$

which proves (i) as well as the additional statement on the operator F. □

An obvious, though interesting, result in the context of frame isomorphy is that the Parseval frame in Lemma 1.7 is in fact isomorphic to the original frame.

Lemma 1.8 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S. Then the Parseval frame $(S^{-1/2}\varphi_i)_{i=1}^M$ is isomorphic to $(\varphi_i)_{i=1}^M$.*

Similarly, a given frame is also isomorphic to its canonical dual frame.

Lemma 1.9 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S. Then the canonical dual frame $(S^{-1}\varphi_i)_{i=1}^M$ is isomorphic to $(\varphi_i)_{i=1}^M$.*

Intriguingly, it turns out—and will be proven in the following result—that the canonical dual frame is the only dual frame which is isomorphic to a given frame.

Proposition 1.23 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a frame for $\mathscr{H}^N$ with frame operator S, and let $(\psi_i)_{i=1}^M$ and $(\tilde{\psi}_i)_{i=1}^M$ be two different dual frames for Φ. Then $(\psi_i)_{i=1}^M$ and $(\tilde{\psi}_i)_{i=1}^M$ are not isomorphic.*

In particular, $(S^{-1}\varphi_i)_{i=1}^M$ *is the only dual frame for* Φ *which is isomorphic to* Φ.

Proof Let $(\psi_i)_{i=1}^M$ and $(\tilde{\psi}_i)_{i=1}^M$ be different dual frames for Φ. Toward a contradiction, we assume that $(\psi_i)_{i=1}^M$ and $(\tilde{\psi}_i)_{i=1}^M$ are isomorphic, and let F denote the invertible operator satisfying $\psi_i = F\tilde{\psi}_i$, $i = 1, 2, \ldots, M$. Then, for each $x \in \mathscr{H}^N$ we have

$$F^*x = \sum_{i=1}^M \langle F^*x, \tilde{\psi}_i\rangle\varphi_i = \sum_{i=1}^M \langle x, F\tilde{\psi}_i\rangle\varphi_i = \sum_{i=1}^M \langle x, \psi_i\rangle\varphi_i = x.$$

Thus, $F^* = Id$ which implies $F = Id$, a contradiction. □

1.7.3.2 Unitarily isomorphic frames

A stronger version of equivalence is given by the notion of unitarily isomorphic frames.

Definition 1.25 Two frames $(\varphi_i)_{i=1}^M$ and $(\psi_i)_{i=1}^M$ for $\mathscr{H}^N$ are *unitarily isomorphic*, if there exists a unitary operator $U : \mathscr{H}^N \to \mathscr{H}^N$ satisfying $U\varphi_i = \psi_i$ for all $i = 1, 2, \ldots, M$.

In the situation of Parseval frames, though, the notions of isomorphy and unitary isomorphy coincide.

Lemma 1.10 *Let* $(\varphi_i)_{i=1}^M$ *and* $(\psi_i)_{i=1}^M$ *be isomorphic Parseval frames for* $\mathscr{H}^N$. *Then they are even unitarily isomorphic.*

Proof Let F be an invertible operator on $\mathscr{H}^N$ with $F\varphi_i = \psi_i$ for all $i = 1, 2, \ldots, M$. By Proposition 1.10, the frame operator of $(F\varphi_i)_{i=1}^M$ is $FIdF^* = FF^*$. On the other hand, the frame operator of $(\psi_i)_{i=1}^M$ is the identity. Hence, $FF^* = Id$. □

We end this section with a necessary and sufficient condition for two frames to be unitarily isomorphic.

Proposition 1.24 *For two frames* $(\varphi_i)_{i=1}^M$ *and* $(\psi_i)_{i=1}^M$ *for* $\mathscr{H}^N$ *with analysis operators* T_1 *and* T_2, *respectively, the following conditions are equivalent.*

(i) $(\varphi_i)_{i=1}^M$ *and* $(\psi_i)_{i=1}^M$ *are unitarily isomorphic.*
(ii) $\|T_1^*c\| = \|T_2^*c\|$ *for all* $c \in \ell_2^M$.
(iii) $T_1T_1^* = T_2T_2^*$.

Proof (i)⇒(iii) Let U be a unitary operator on $\mathscr{H}^N$ with $U\varphi_i = \psi_i$ for all $i = 1, \ldots, M$. Then, since by Proposition 1.9 we have $T_2 = T_1U^*$, we obtain $T_2T_2^* = T_1U^*UT_1^* = T_1T_1^*$ and thus (iii).

(iii)⇒(ii) This is immediate.

(ii)⇒(i) Since (ii) implies $\ker T_1^* = \ker T_2^*$, it follows from Theorem 1.18 that $U\varphi_i = \psi_i$ for all $i = 1, \ldots, M$, where $U = T_2^*(T_1^*|_{\operatorname{ran} T_1})^{-1}$. But this operator is unitary since (ii) also implies

$$\left\| T_2^* \left(T_1^*|_{\operatorname{ran} T_1}\right)^{-1} x \right\| = \left\| T_1^* \left(T_1^*|_{\operatorname{ran} T_1}\right)^{-1} x \right\| = \|x\|$$

for all $x \in \mathscr{H}^N$. □

1.8 Applications of Finite Frames

Finite frames are a versatile methodology for any application which requires redundant, yet stable, decompositions, e.g., for analysis or transmission of signals, but surprisingly also for more theoretically oriented questions. We state some of these applications in this section, which also coincide with the chapters of this book.

1.8.1 Noise and Erasure Reduction

Noise and erasures are one of the most common problems signal transmissions have to face [130–132]. The redundancy of frames is particularly suitable to reduce and compensate for such disturbances. Pioneering studies can be found in [50, 93–95], followed by the fundamental papers [10, 15, 102, 136, 149]. In addition one is always faced with the problem of suppressing errors introduced through quantization, both pulse code modulation (PCM) [20, 151] and sigma-delta quantization [7, 8, 16, 17]. Theoretical error considerations range from worst to average case scenarios. Different strategies for reconstruction exist depending on whether the receiver is aware or unaware of noise and erasures. Some more recent work also takes into account special types of erasures [18] or the selection of dual frames for reconstruction [121, 123]. Chapter 7 provides a comprehensive survey of these considerations and related results.

1.8.2 Resilience Against Perturbations

Perturbations of a signal are an additional problem faced by signal processing applications. Various results on the ability of frames to be resilient against perturbations are known. One class focuses on generally applicable frame perturbation results [3, 37, 59, 68], some even in the Banach space setting [39, 68]. Yet another topic is that of perturbations of specific frames such as Gabor frames [40], frames containing a Riesz basis [38], or frames for shift-invariant spaces [153]. Finally, extensions such as fusion frames are studied with respect to their behavior under perturbations [52].

1.8.3 Quantization Robustness

Each signal processing application contains an analog-to-digital conversion step, which is called quantization. Quantization is typically applied to the transform coefficients, which in our case are (redundant) frame coefficients; see [94, 95]. Interestingly, the redundancy of the frame can be successfully explored in the quantization step by using sigma-delta algorithms and a particular noncanonical dual frame reconstruction. In most regimes, the performance is significantly better than that obtained by rounding each coefficient separately (PCM). This was first observed in [7, 8]. Within a short amount of time, the error bounds were improved [16, 114], refined quantization schemes were studied [14, 17], specific dual frame constructions for reconstruction were developed [9, 98, 118], and PCM was revisited [105, 151]. The interested reader is referred to Chap. 8, which provides an introduction to the quantization of finite frames.

1.8.4 Compressed Sensing

Since high-dimensional signals are typically concentrated on lower dimensional subspaces, it is a natural assumption that the collected data can be represented by a sparse linear combination of an appropriately chosen frame. The novel methodology of compressed sensing, initially developed in [32, 33, 78], utilizes this observation to show that such signals can be reconstructed from very few nonadaptive linear measurements by linear programming techniques. For an introduction, we refer to the books [84, 86] and the survey [25]. Finite frames thus play an essential role, both as sparsifying systems and in designing the measurement matrix. For a selection of studies focusing in particular on the connection to frames, we refer to [1, 2, 31, 69, 141, 142]; for the connection to structured frames such as fusion frames, see [22, 85]. Chapter 9 provides an introduction to compressed sensing and the connection to finite frame theory.

There exists yet another intriguing connection of finite frames to sparsity methodologies, namely, aiming for sparse frame vectors to ensure low computational complexity. For this, we refer to the two papers [30, 49] and to Chap. 13.

1.8.5 Filter Banks

Filter banks are the basis for most signal processing applications. We exemplarily mention the general books [125, 145] and those with a particular focus on wavelets [75, 134, 150], as well as the beautiful survey articles [109, 110]. Usually, several filters are applied in parallel to an input signal, followed by downsampling. This processing method is closely related to the decomposition with respect to finite frames provided that the frame consists of equally spaced translates of a fixed set of vectors,

first observed in [19, 21, 71, 72] and later refined and extended in [62, 63, 90, 112]. This viewpoint has the benefit of providing a deeper understanding of filtering procedures, while retaining the potential of extensions of classical filter bank theory. We refer to Chap. 10, which provides an introduction into filter banks and their connections with finite frame theory.

1.8.6 Stable Partitions

The Feichtinger conjecture in frame theory conjectures the existence of certain partitions of frames into sequences with "good" frame bounds; see [41]. Its relevance becomes evident when modeling distributed processing, and stable frames are required for the local processing units (see also Sect. 1.9 on fusion frames). The fundamental papers [48, 55, 61] then linked this conjecture to a variety of open conjectures in what is customarily called pure mathematics such as the Kadison-Singer problem in C^*-algebras [107]. Chapter 11 provides an introduction into these connections and their significance. A particular focus of this chapter is also on the Paulsen problem [11, 27, 45], which provides error estimates on the ability of a frame to be simultaneously (almost) equal norm and (almost) tight.

1.9 Extensions

Typically motivated by applications, various extensions of finite frame theory have been developed over the last years. In this book, Chaps. 12 and 13 are devoted to the main two generalizations, whose key ideas we will now briefly describe.

- *Probabilistic Frames.* This theory is based on the observation that finite frames can be regarded as mass points distributed in $\mathscr{H}^N$. As an extension, probabilistic frames, which were introduced and studied in [81–83], constitute a class of general probability measures, again with appropriate stability constraints. Applications include, for instance, directional statistics in which probabilistic frames can be utilized to measure inconsistencies of certain statistical tests [108, 143, 144]. For more details on the theory and applications of probabilistic frames, we refer to Chap. 12.
- *Fusion Frames.* Signal processing by finite frames can be regarded as projections onto one-dimensional subspaces. In contrast to this, fusion frames, introduced in [51, 53], analyze and process a signal by (orthogonal) projections onto multidimensional subspaces, which again have to satisfy some stability conditions. They also allow for a local processing in the different subspaces. This theory is in fact a perfect fit to applications requiring distributed processing; we refer to the series of papers [22, 23, 28, 30, 42, 43, 46, 63, 117, 124]. We also mention that a closely related generalization called G-frames exists, which however does not admit any additional (local) structure and which is unrelated to applications (see, for instance, [137, 138]). A detailed introduction to fusion frame theory can be found in Chap. 13.

Acknowledgements The authors are grateful to Andreas Heinecke and Emily King for their extensive proofreading and various useful comments, which have significantly improved the presentation. P.G.C. acknowledges support by NSF Grants DMS 1008183 and ATD 1042701 as well as by AFOSR Grant FA9550-11-1-0245. G.K. acknowledges support by the Einstein Foundation Berlin, by Deutsche Forschungsgemeinschaft (DFG) Grant SPP-1324 KU 1446/13 and DFG Grant KU 1446/14, and by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin. F.P. was supported by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin.

References

1. Bajwa, W.U., Calderbank, R., Jafarpour, S.: Why Gabor frames? Two fundamental measures of coherence and their role in model selection. J. Commun. Netw. **12**, 289–307 (2010)
2. Bajwa, W.U., Calderbank, R., Mixon, D.G.: Two are better than one: fundamental parameters of frame coherence. Appl. Comput. Harmon. Anal. **33**, 58–78 (2012)
3. Balan, R.: Stability theorems for Fourier frames and wavelet Riesz bases. J. Fourier Anal. Appl. **3**, 499–504 (1997)
4. Balan, R.: Equivalence relations and distances between Hilbert frames. Proc. Am. Math. Soc. **127**, 2353–2366 (1999)
5. Balan, R., Bodmann, B.G., Casazza, P.G., Edidin, D.: Painless reconstruction from magnitudes of frame coefficients. J. Fourier Anal. Appl. **15**, 488–501 (2009)
6. Balan, R., Casazza, P.G., Edidin, D.: On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**, 345–356 (2006)
7. Benedetto, J.J., Powell, A.M., Yilmaz, Ö.: Sigma-delta ($\Sigma\Delta$) quantization and finite frames. IEEE Trans. Inf. Theory **52**, 1990–2005 (2006)
8. Benedetto, J.J., Powell, A.M., Yilmaz, Ö.: Second order sigma-delta quantization of finite frame expansions. Appl. Comput. Harmon. Anal. **20**, 126–148 (2006)
9. Blum, J., Lammers, M., Powell, A.M., Yilmaz, Ö.: Sobolev duals in frame theory and sigma-delta quantization. J. Fourier Anal. Appl. **16**, 365–381 (2010)
10. Bodmann, B.G.: Optimal linear transmission by loss-insensitive packet encoding. Appl. Comput. Harmon. Anal. **22**, 274–285 (2007)
11. Bodmann, B., Casazza, P.G.: The road to equal-norm Parseval frames. J. Funct. Anal. **258**, 397–420 (2010)
12. Bodmann, B.G., Casazza, P.G., Kutyniok, G.: A quantitative notion of redundancy for finite frames. Appl. Comput. Harmon. Anal. **30**, 348–362 (2011)
13. Bodmann, B.G., Casazza, P.G., Paulsen, V.I., Speegle, D.: Spanning and independence properties of frame partitions. Proc. Am. Math. Soc. **140**, 2193–2207 (2012)
14. Bodmann, B., Lipshitz, S.: Randomly dithered quantization and sigma-delta noise shaping for finite frames. Appl. Comput. Harmon. Anal. **25**, 367–380 (2008)
15. Bodmann, B.G., Paulsen, V.I.: Frames, graphs and erasures. Linear Algebra Appl. **404**, 118–146 (2005)
16. Bodmann, B., Paulsen, V.: Frame paths and error bounds for sigma-delta quantization. Appl. Comput. Harmon. Anal. **22**, 176–197 (2007)
17. Bodmann, B., Paulsen, V., Abdulbaki, S.: Smooth frame-path termination for higher order sigma-delta quantization. J. Fourier Anal. Appl. **13**, 285–307 (2007)
18. Bodmann, B.G., Singh, P.K.: Burst erasures and the mean-square error for cyclic Parseval frames. IEEE Trans. Inf. Theory **57**, 4622–4635 (2011)
19. Bölcskei, H., Hlawatsch, F.: Oversampled cosine modulated filter banks with perfect reconstruction. IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process. **45**, 1057–1071 (1998)
20. Bölcskei, H., Hlawatsch, F.: Noise reduction in oversampled filter banks using predictive quantization. IEEE Trans. Inf. Theory **47**, 155–172 (2001)

21. Bölcskei, H., Hlawatsch, F., Feichtinger, H.G.: Frame-theoretic analysis of oversampled filter banks. IEEE Trans. Signal Process. **46**, 3256–3269 (1998)
22. Boufounos, B., Kutyniok, G., Rauhut, H.: Sparse recovery from combined fusion frame measurements. IEEE Trans. Inf. Theory **57**, 3864–3876 (2011)
23. Bownik, M., Luoto, K., Richmond, E.: A combinatorial characterization of tight fusion frames, preprint
24. Broome, H., Waldron, S.: On the construction of highly symmetric tight frames and complex polytopes, preprint
25. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**, 34–81 (2009)
26. Cahill, J.: Flags, frames, and Bergman spaces. Master's Thesis, San Francisco State University (2009)
27. Cahill, J., Casazza, P.G.: The Paulsen problem in operator theory. Oper. Matrices (to appear)
28. Cahill, J., Casazza, P.G., Li, S.: Non-orthogonal fusion frames and the sparsity of fusion frame operators. J. Fourier Anal. Appl. **18**, 287–308 (2012)
29. Cahill, J., Fickus, M., Mixon, D.G., Poteet, M.J., Strawn, N.K.: Constructing finite frames of a given spectrum and set of lengths. Appl. Comput. Harmon. Anal. (to appear)
30. Calderbank, R., Casazza, P.G., Heinecke, A., Kutyniok, G., Pezeshki, A.: Sparse fusion frames: existence and construction. Adv. Comput. Math. **35**, 1–31 (2011)
31. Candès, E.J., Eldar, Y., Needell, D., Randall, P.: Compressed sensing with coherent and redundant dictionaries. Appl. Comput. Harmon. Anal. **31**, 59–73 (2011)
32. Candès, E.J., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**, 1207–1223 (2006)
33. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**, 489–509 (2006)
34. Casazza, P.G.: Modern tools for Weyl-Heisenberg (Gabor) frame theory. Adv. Imaging Electron Phys. **115**, 1–127 (2000)
35. Casazza, P.G.: The art of frame theory. Taiwan. J. Math. **4**, 129–201 (2000)
36. Casazza, P.G.: Custom building finite frames. In: Wavelets, Frames and Operator Theory. Papers from the Focused Research Group Workshop, University of Maryland, College Park, MD, USA, 15–21 January 2003. Contemp. Math., vol. 345, pp. 15–21. Am. Math. Soc., Providence (2003)
37. Casazza, P.G., Christensen, O.: Perturbation of operators and applications to frame theory. J. Fourier Anal. Appl. **3**, 543–557 (1997)
38. Casazza, P.G., Christensen, O.: Frames containing a Riesz basis and preservation of this property under perturbations. SIAM J. Math. Anal. **29**, 266–278 (1998)
39. Casazza, P.G., Christensen, O.: The reconstruction property in Banach spaces and a perturbation theorem. Can. Math. Bull. **51**, 348–358 (2008)
40. Casazza, P.G., Christensen, O., Lammers, M.C.: Perturbations of Weyl-Heisenberg frames. Hokkaido Math. J. **31**, 539–553 (2002)
41. Casazza, P.G., Christensen, O., Lindner, A., Vershynin, R.: Frames and the Feichtinger conjecture. Proc. Am. Math. Soc. **133**, 1025–1033 (2005)
42. Casazza, P.G., Fickus, M.: Minimizing fusion frame potential. Acta Appl. Math. **107**, 7–24 (2009)
43. Casazza, P.G., Fickus, M., Heinecke, A., Wang, Y., Zhou, Z.: Spectral tetris fusion frame constructions. J. Fourier Anal. Appl. Published online, April 2012
44. Casazza, P.G., Fickus, M., Kovačević, J., Leon, M., Tremain, J.C.: A physical interpretation for finite tight frames. In: Heil, C. (ed.) Harmonic Analysis and Applications (in Honor of John Benedetto), pp. 51–76. Birkhaüser, Basel (2006)
45. Casazza, P.G., Fickus, M., Mixon, D.: Auto-tuning unit norm frames. Appl. Comput. Harmon. Anal. **32**, 1–15 (2012)
46. Casazza, P.G., Fickus, M., Mixon, D., Wang, Y., Zhou, Z.: Constructing tight fusion frames. Appl. Comput. Harmon. Anal. **30**, 175–187 (2011)

47. Casazza, P.G., Fickus, M., Mixon, D., Wang, Y., Zhou, Z.: Constructing tight fusion frames. Appl. Comput. Harmon. Anal. **30**, 175–187 (2011)
48. Casazza, P.G., Fickus, M., Tremain, J.C., Weber, E.: The Kadison-Singer problem in mathematics and engineering—a detailed account. In: Operator Theory, Operator Algebras and Applications. Proceedings of the 25th Great Plains Operator Theory Symposium, University of Central Florida, FL, USA, 7–12 June 2005. Contemp. Math., vol. 414, pp. 297–356. Am. Math. Soc., Providence (2006)
49. Casazza, P.G., Heinecke, A., Krahmer, F., Kutyniok, G.: Optimally Sparse frames. IEEE Trans. Inf. Theory **57**, 7279–7287 (2011)
50. Casazza, P.G., Kovačević, J.: Equal-norm tight frames with erasures. Adv. Comput. Math. **18**, 387–430 (2003)
51. Casazza, P.G., Kutyniok, G.: Frames of subspaces. In: Wavelets, Frames and Operator Theory. Papers from the Focused Research Group Workshop, University of Maryland, College Park, MD, USA, 15–21 January 2003. Contemp. Math., vol. 345, pp. 15–21. Am. Math. Soc., Providence (2003)
52. Casazza, P.G., Kutyniok, G.: Robustness of fusion frames under erasures of subspaces and of local frame vectors. In: Radon Transforms, Geometry, and Wavelets. Contemp. Math., vol. 464, pp. 149–160. Am. Math. Soc., Providence (2008)
53. Casazza, P.G., Kutyniok, G., Li, S.: Fusion frames and distributed processing. Appl. Comput. Harmon. Anal. **25**, 114–132 (2008)
54. Casazza, P.G., Kutyniok, G., Speegle, D.: A redundant version of the Rado-Horn theorem. Linear Algebra Appl. **418**, 1–10 (2006)
55. Casazza, P.G., Kutyniok, G., Speegle, D.: A decomposition theorem for frames and the Feichtinger conjecture. Proc. Am. Math. Soc. **136**, 2043–2053 (2008)
56. Casazza, P.G., Leon, M.: Existence and construction of finite tight frames. J. Concr. Appl. Math. **4**, 277–289 (2006)
57. Casazza, P.G., Leon, M.: Existence and construction of finite frames with a given frame operator. Int. J. Pure Appl. Math. **63**, 149–158 (2010)
58. Casazza, P.G., Leonhard, N.: Classes of finite equal norm Parseval frames. In: Frames and Operator Theory in Analysis and Signal Processing. AMS-SIAM Special Session, San Antonio, TX, USA, 12–15 January 2006. Contemp. Math., vol. 451, pp. 11–31. Am. Math. Soc., Providence (2008)
59. Casazza, P.G., Liu, G., Zhao, C., Zhao, P.: Perturbations and irregular sampling theorems for frames. IEEE Trans. Inf. Theory **52**, 4643–4648 (2006)
60. Casazza, P.G., Redmond, D., Tremain, J.C.: Real equiangular frames. In: 42nd Annual Conference on Information Sciences and Systems. CISS 2008, pp. 715–720 (2008)
61. Casazza, P.G., Tremain, J.C.: The Kadison-Singer problem in mathematics and engineering. Proc. Natl. Acad. Sci. **103**, 2032–2039 (2006)
62. Chai, L., Zhang, J., Zhang, C., Mosca, E.: Frame-theory-based analysis and design of oversampled filter banks: direct computational method. IEEE Trans. Signal Process. **55**, 507–519 (2007)
63. Chebira, A., Fickus, M., Mixon, D.G.: Filter bank fusion frames. IEEE Trans. Signal Process. **59**, 953–963 (2011)
64. Chien, T., Waldron, S.: A classification of the harmonic frames up to unitary equivalence. Appl. Comput. Harmon. Anal. **30**, 307–318 (2011)
65. Christensen, O.: An Introduction to Frames and Riesz Bases. Birkhäuser Boston, Boston (2003)
66. Christensen, O.: Frames and Bases: An Introductory Course. Birkhäuser, Boston (2008)
67. Christensen, O., Feichtinger, H.G., Paukner, S.: Gabor analysis for imaging. In: Handbook of Mathematical Methods in Imaging, pp. 1271–1307. Springer, Berlin (2011)
68. Christensen, O., Heil, C.: Perturbations of Banach frames and atomic decompositions. Math. Nachr. **185**, 33–47 (1997)
69. Cohen, A., Dahmen, W., DeVore, R.A.: Compressed sensing and best k-term approximation. J. Am. Math. Soc. **22**, 211–231 (2009)

70. Conway, J.B.: A Course in Functional Analysis, 2nd edn. Springer, Berlin (2010)
71. Cvetković, Z., Vetterli, M.: Oversampled filter banks. IEEE Trans. Signal Process. **46**, 1245–1255 (1998)
72. Cvetković, Z., Vetterli, M.: Tight Weyl-Heisenberg frames in $\ell_2(\mathbb{Z})$. IEEE Trans. Signal Process. **46**, 1256–1259 (1998)
73. Dahmen, W., Huang, C., Schwab, C., Welper, G.: Adaptive Petrov-Galerkin methods for first order transport equation. SIAM J. Numer. Anal. (to appear)
74. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theory **36**, 961–1005 (1990)
75. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
76. Daubechies, I., Grossman, A., Meyer, Y.: Painless nonorthogonal expansions. J. Math. Phys. **27**, 1271–1283 (1985)
77. Dong, B., Shen, Z.: MRA-Based Wavelet Frames and Applications. IAS/Park City Math. Ser., vol. 19 (2010)
78. Donoho, D.L.: Compressed sensing. IEEE Trans. Inf. Theory **52**, 1289–1306 (2006)
79. Duffin, R., Schaeffer, A.: A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)
80. Dykema, K., Strawn, N.: Manifold structure of spaces of spherical tight frames. Int. J. Pure Appl. Math. **28**, 217–256 (2006)
81. Ehler, M.: Random tight frames. J. Fourier Anal. Appl. **18**, 1–20 (2012)
82. Ehler, M., Galanis, J.: Frame theory in directional statistics. Stat. Probab. Lett. **81**, 1046–1051 (2011)
83. Ehler, M., Okoudjou, K.A.: Minimization of the probabilistic p-frame potential. J. Stat. Plan. Inference **142**, 645–659 (2012)
84. Elad, M.: Sparse and Redundant Representations: From Theory to Applications in Signal and Image Processing. Springer, Berlin (2010)
85. Eldar, Y.C., Kuppinger, P., Bölcskei, H.: Block-sparse signals: uncertainty relations and efficient recovery. IEEE Trans. Signal Process. **58**, 3042–3054 (2010)
86. Eldar, Y., Kutyniok, G. (eds.): Compressed Sensing: Theory and Applications. Cambridge University Press, Cambridge (2012)
87. Feichtinger, H.G., Gröchenig, K.: Gabor frames and time-frequency analysis of distributions. J. Funct. Anal. **146**, 464–495 (1996)
88. Feichtinger, H.G., Strohmer, T. (eds.): Gabor Analysis and Algorithms: Theory and Applications. Birkhäuser, Boston (1998)
89. Feichtinger, H.G., Strohmer, T., Christensen, O.: A group-theoretical approach to Gabor analysis. Opt. Eng. **34**, 1697–1704 (1995)
90. Fickus, M., Johnson, B.D., Kornelson, K., Okoudjou, K.: Convolutional frames and the frame potential. Appl. Comput. Harmon. Anal. **19**, 77–91 (2005)
91. Fickus, M., Mixon, D.G., Tremain, J.C.: Steiner equiangular tight frames. Linear Algebra Appl. **436**, 1014–1027 (2012)
92. Gabor, D.: Theory of communication. J. Inst. Electr. Eng. **93**, 429–457 (1946)
93. Goyal, V.K., Kelner, J.A., Kovačević, J.: Multiple description vector quantization with a coarse lattice. IEEE Trans. Inf. Theory **48**, 781–788 (2002)
94. Goyal, V.K., Kovačević, J., Kelner, J.A.: Quantized frame expansions with erasures. Appl. Comput. Harmon. Anal. **10**, 203–233 (2001)
95. Goyal, V., Vetterli, M., Thao, N.T.: Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory **44**, 16–31 (1998)
96. Gröchenig, K.: Acceleration of the frame algorithm. IEEE Trans. Signal Process. **41**, 3331–3340 (1993)
97. Gröchenig, K.: Foundations of Time-Frequency Analysis. Birkhäuser, Boston (2000)
98. Güntürk, C.S., Lammers, M., Powell, A.M., Saab, R., Yilmaz, Ö.: Sobolev duals for random frames and sigma-delta quantization of compressed sensing measurements, preprint
99. Han, D., Kornelson, K., Larson, D.R., Weber, E.: Frames for Undergraduates. American Mathematical Society, Student Mathematical Library, vol. 40 (2007)

100. Han, D., Larson, D.R.: Frames, bases and group representations. Mem. Am. Math. Soc. **147**, 1–103 (2000)
101. Hay, N., Waldron, S.: On computing all harmonic frames of n vectors in $\mathbb{C}^d$. Appl. Comput. Harmon. Anal. **21**, 168–181 (2006)
102. Holmes, R.B., Paulsen, V.I.: Optimal frames for erasures. Linear Algebra Appl. **377**, 31–51 (2004)
103. Horn, A.: A characterization of unions of linearly independent sets. J. Lond. Math. Soc. **30**, 494–496 (1955)
104. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
105. Jimenez, D., Wang, L., Wang, Y.: White noise hypothesis for uniform quantization errors. SIAM J. Appl. Math. **28**, 2042–2056 (2007)
106. Jokar, S., Mehrmann, V., Pfetsch, M., Yserentant, H.: Sparse approximate solution of partial differential equations. Appl. Numer. Math. **60**, 452–472 (2010)
107. Kadison, R., Singer, I.: Extensions of pure states. Am. J. Math. **81**, 383–400 (1959)
108. Kent, J.T., Tyler, D.E.: Maximum likelihood estimation for the wrapped Cauchy distribution. J. Appl. Stat. **15**, 247–254 (1988)
109. Kovačević, J., Chebira, A.: Life beyond bases: the advent of frames (Part I). IEEE Signal Process. Mag. **24**, 86–104 (2007)
110. Kovačević, J., Chebira, A.: Life beyond bases: the advent of frames (Part II). IEEE Signal Process. Mag. **24**, 115–125 (2007)
111. Kovačević, J., Chebira, A.: An introduction to frames. Found. Trends Signal Process. **2**, 1–100 (2008)
112. Kovačević, J., Dragotti, P.L., Goyal, V.K.: Filter bank frame expansions with erasures. IEEE Trans. Inf. Theory **48**, 1439–1450 (2002)
113. Krahmer, F., Pfander, G.E., Rashkov, P.: Uncertainty in time-frequency representations on finite abelian groups and applications. Appl. Comput. Harmon. Anal. **25**, 209–225 (2008)
114. Krahmer, F., Saab, R., Ward, R.: Root-exponential accuracy for coarse quantization of finite frame expansions. IEEE J. Int. Theory **58**, 1069–1079 (2012)
115. Kutyniok, G., Labate, D. (eds.): Shearlets: Multiscale Analysis for Multivariate Data. Birkhäuser, Boston (2012)
116. Kutyniok, G., Okoudjou, K.A., Philipp, F., Tuley, E.K.: Scalable frames, preprint
117. Kutyniok, G., Pezeshki, A., Calderbank, A.R., Liu, T.: Robust dimension reduction, fusion frames, and Grassmannian packings. Appl. Comput. Harmon. Anal. **26**, 64–76 (2009)
118. Lammers, M., Powell, A.M., Yilmaz, Ö.: Alternative dual frames for digital-to-analog conversion in sigma-delta quantization. Adv. Comput. Math. **32**, 73–102 (2010)
119. Lawrence, J., Pfander, G.E., Walnut, D.F.: Linear independence of Gabor systems in finite dimensional vector spaces. J. Fourier Anal. Appl. **11**, 715–726 (2005)
120. Lemmens, P., Seidel, J.: Equiangular lines. J. Algebra **24**, 494–512 (1973)
121. Lopez, J., Han, D.: Optimal dual frames for erasures. Linear Algebra Appl. **432**, 471–482 (2010)
122. Mallat, S.: A Wavelet Tour of Signal Processing: The Sparse Way. Academic Press, San Diego (2009)
123. Massey, P.: Optimal reconstruction systems for erasures and for the q-potential. Linear Algebra Appl. **431**, 1302–1316 (2009)
124. Massey, P.G., Ruiz, M.A., Stojanoff, D.: The structure of minimizers of the frame potential on fusion frames. J. Fourier Anal. Appl. **16**, 514–543 (2010)
125. Oppenheim, A.V., Schafer, R.W.: Digital Signal Processing. Prentice Hall, New York (1975)
126. Püschel, M., Kovačević, J.: Real tight frames with maximal robustness to erasures. In: Proc. Data Compr. Conf., pp. 63–72 (2005)
127. Qiu, S., Feichtinger, H.: Discrete Gabor structure and optimal representation. IEEE Trans. Signal Process. **43**, 2258–2268 (1995)
128. Rado, R.: A combinatorial theorem on vector spaces. J. Lond. Math. Soc. **37**, 351–353 (1962)
129. Rudin, W.: Functional Analysis, 2nd edn. McGraw-Hill, New York (1991)

130. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423, 623–656 (1948)
131. Shannon, C.E.: Communication in the presence of noise. Proc. I.R.E. **37**, 10–21 (1949)
132. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. The University of Illinois Press, Urbana (1949)
133. Shen, Z.: Wavelet frames and image restorations. In: Proceedings of the International Congress of Mathematicians (ICM 2010), Hyderabad, India, August 1927. Invited lectures, vol. IV, pp. 2834–2863. World Scientific/Hindustan Book Agency, Hackensack/New Delhi (2011)
134. Strang, G., Nguyen, T.: Wavelets and Filter Banks. Wellesley-Cambridge Press, Cambridge (1996)
135. Strawn, N.: Finite frame varieties: nonsingular points, tangent spaces, and explicit local parameterizations. J. Fourier Anal. Appl. **17**, 821–853 (2011)
136. Strohmer, T., Heath, R.W. Jr.: Grassmannian frames with applications to coding and communication. Appl. Comput. Harmon. Anal. **14**, 257–275 (2003)
137. Sun, W.: G-frames and G-Riesz bases. J. Math. Anal. Appl. **322**, 437–452 (2006)
138. Sun, W.: Stability of G-frames. J. Math. Anal. Appl. **326**, 858–868 (2007)
139. Sustik, M.A., Tropp, J.A., Dhillon, I.S., Heath, R.W. Jr.: On the existence of equiangular tight frames. Linear Algebra Appl. **426**, 619–635 (2007)
140. Taubman, D.S., Marcellin, M.: JPEG2000: Image Compression Fundamentals, Standards and Practice. Kluwer International Series in Engineering & Computer Science (2001)
141. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**, 2231–2242 (2004)
142. Tropp, J.A.: Just relax: convex programming methods for identifying sparse signals in noise. IEEE Trans. Inf. Theory **52**, 1030–1051 (2006)
143. Tyler, D.E.: A distribution-free M-estimate of multivariate scatter. Ann. Stat. **15**, 234–251 (1987)
144. Tyler, D.E.: Statistical analysis for the angular central Gaussian distribution. Biometrika **74**, 579–590 (1987)
145. Vaidyanathan, P.P.: Multirate Systems and Filter Banks. Prentice Hall, Englewood Cliffs (1992)
146. Vale, R., Waldron, S.: Tight frames and their symmetries. Constr. Approx. **21**, 83–112 (2005)
147. Vale, R., Waldron, S.: Tight frames generated by finite nonabelian groups. Numer. Algorithms **48**, 11–27 (2008)
148. Vale, R., Waldron, S.: The symmetry group of a finite frame. Linear Algebra Appl. **433**, 248–262 (2010)
149. Vershynin, R.: Frame expansions with erasures: an approach through the noncommutative operator theory. Appl. Comput. Harmon. Anal. **18**, 167–176 (2005)
150. Vetterli, M., Kovačević, J., Goyal, V.K.: Fourier and Wavelet Signal Processing (2011). http://www.fourierandwavelets.org
151. Wang, Y., Xu, Z.: The performance of PCM quantization under tight frame representations. SIAM J. Math. Anal. (to appear)
152. Young, N.: An Introduction to Hilbert Space. Cambridge University Press, Cambridge (1988)
153. Zhao, P., Zhao, C., Casazza, P.G.: Perturbation of regular sampling in shift-invariant spaces for frames. IEEE Trans. Inf. Theory **52**, 4643–4648 (2006)

Chapter 2
Constructing Finite Frames with a Given Spectrum

Matthew Fickus, Dustin G. Mixon, and Miriam J. Poteet

Abstract Broadly speaking, frame theory is the study of how to produce well-conditioned frame operators, often subject to nonlinear application-motivated restrictions on the frame vectors themselves. In this chapter, we focus on one particularly well-studied type of restriction: having frame vectors of prescribed lengths. We discuss two methods for iteratively constructing such frames. The first method, called Spectral Tetris, produces special examples of such frames, and only works in certain cases. The second method combines the idea behind Spectral Tetris with the classical theory of majorization; this method can build any such frame in terms of a sequence of interlacing spectra, called eigensteps.

Keywords Tight frames · Schur-Horn · Majorization · Interlacing

2.1 Introduction

Although we work over the complex field for the sake of generality, the theory presented here carries over verbatim to the real-variable setting. The *synthesis operator* of a sequence of vectors $\Phi = \{\varphi_m\}_{m=1}^M$ in $\mathbb{C}^N$ is $\Phi : \mathbb{C}^M \to \mathbb{C}^N$, $\Phi y := \sum_{m=1}^M y(m)\varphi_m$. That is, Φ is the $N \times M$ matrix whose columns are the φ_m's. Note that we make no notational distinction between the vectors themselves and the synthesis operator they induce. Φ is said to be a *frame* for $\mathbb{C}^N$ if there exist *frame bounds* $0 < A \leq B < \infty$ such that $A\|x\|^2 \leq \|\Phi^* x\|^2 \leq B\|x\|^2$ for all $x \in \mathbb{C}^N$. The optimal frame bounds A and B of Φ are the least and greatest eigenvalues of the

M. Fickus (✉) · M.J. Poteet
Department of Mathematics, Air Force Institute of Technology, Wright-Patterson AFB, OH 45433, USA
e-mail: matthew.fickus@afit.edu

D.G. Mixon
Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_2,

frame operator

$$\Phi\Phi^* = \sum_{m=1}^{M} \varphi_m \varphi_m^*, \tag{2.1}$$

respectively. Here, φ_m^* is the linear functional $\varphi_m^* : \mathbb{C}^N \to \mathbb{C}$, $\varphi_m^* x := \langle x, \varphi_m \rangle$. In particular, Φ is a frame if and only if the φ_m's span $\mathbb{C}^N$, which necessitates $N \leq M$.

Frames provide numerically stable methods for finding overcomplete decompositions of vectors, and as such are useful tools in various signal processing applications [26, 27]. Indeed, if Φ is a frame, then any $x \in \mathbb{C}^N$ can be decomposed as

$$x = \Phi\tilde{\Phi}^* x = \sum_{m=1}^{M} \langle x, \tilde{\varphi}_m \rangle \varphi_m, \tag{2.2}$$

where $\tilde{\Phi} = \{\tilde{\varphi}_m\}_{m=1}^M$ is a *dual frame* of Φ, meaning it satisfies $\Phi\tilde{\Phi}^* = Id$. The most often-used dual is the *canonical* dual, namely the pseudoinverse $\tilde{\Phi} = (\Phi\Phi^*)^{-1}\Phi$. Computing a canonical dual involves inverting the frame operator. As such, when designing a frame for a given application, it is important to control over the spectrum $\{\lambda_n\}_{n=1}^N$ of $\Phi\Phi^*$. Here and throughout, such spectra are arranged in nonincreasing order, with the optimal frame bounds A and B being λ_N and λ_1, respectively.

Of particular interest are *tight frames*, namely frames for which $A = B$. Note that this occurs precisely when $\lambda_n = A$ for all n, meaning $\Phi\Phi^* = AId$. In this case, the canonical dual is given by $\tilde{\varphi}_m = \frac{1}{A}\varphi_m$, and (2.2) becomes an overcomplete generalization of an orthonormal basis decomposition. Tight frames are not hard to construct; we simply need the rows of Φ to be orthogonal and have constant squared norm A. However, this problem becomes significantly more difficult if we further require the φ_m's—the columns of Φ—to have prescribed lengths.

In particular, much attention has been paid to the problem of constructing *unit norm tight frames* (UNTFs): tight frames for which $\|\varphi_m\| = 1$ for all m. Here, since $NA = \mathrm{Tr}(\Phi\Phi^*) = \mathrm{Tr}(\Phi^*\Phi) = M$, we see that A is necessarily $\frac{M}{N}$. For any $N \leq M$, there always exists at least one corresponding UNTF, namely the *harmonic frame* obtained by letting Φ be an $N \times M$ submatrix of an $M \times M$ discrete Fourier transform [20]. UNTFs are known to be optimally robust with respect to additive noise [21] and erasures [12, 23], and are a generalization of code division multiple access (CDMA) encoders [31, 33]. Moreover, all unit norm sequences Φ satisfy the zeroth-order *Welch bound* $\mathrm{Tr}[(\Phi\Phi^*)^2] \geq \frac{M^2}{N}$, which is achieved precisely when Φ is a UNTF [34, 35]; a physics-inspired interpretation of this fact leading to an optimization-based proof of the existence of UNTFs is given in [3]. We further know that many such frames exist: when $M > N + 1$, the manifold of all $N \times M$ real UNTFs, modulo rotations, is known to have nontrivial dimension [17]. Local parametrizations of this manifold are given in [30]. Much of the recent work on UNTFs has focused on the *Paulsen problem* [4, 9], a type of Procrustes problem [22] concerning how a given frame should be perturbed in order to make it more like a UNTF.

In this chapter, we discuss the main results of [5, 10, 19], which show how to construct *every* UNTF and moreover solve the following more general problem.

Problem 2.1 Given any nonnegative nonincreasing sequences $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$, construct all $\Phi = \{\varphi_m\}_{m=1}^M$ whose frame operator $\Phi\Phi^*$ has spectrum $\{\lambda_n\}_{n=1}^N$ and for which $\|\varphi_m\|^2 = \mu_m$ for all m.

To solve this problem, we build on the existing theory of majorization. To be precise, given two nonnegative nonincreasing sequences $\{\lambda_m\}_{m=1}^M$ and $\{\mu_m\}_{m=1}^M$, we say that $\{\lambda_m\}_{m=1}^M$ *majorizes* $\{\mu_m\}_{m=1}^M$, denoted $\{\lambda_m\}_{m=1}^M \succeq \{\mu_m\}_{m=1}^M$, if:

$$\sum_{m'=1}^{m} \lambda_{m'} \geq \sum_{m'=1}^{m} \mu_{m'}, \quad \forall m = 1, \ldots, M-1, \tag{2.3}$$

$$\sum_{m'=1}^{M} \lambda_{m'} = \sum_{m'=1}^{M} \mu_{m'}. \tag{2.4}$$

A classical result of Schur [29] states that the spectrum of a self-adjoint positive semidefinite matrix necessarily majorizes its diagonal entries. A few decades later, Horn gave a nonconstructive proof of a converse result [24], showing that if $\{\lambda_m\}_{m=1}^M \succeq \{\mu_m\}_{m=1}^M$, then there exists a self-adjoint matrix that has $\{\lambda_m\}_{m=1}^M$ as its spectrum and $\{\mu_m\}_{m=1}^M$ as its diagonal. These two results are collectively known as the Schur-Horn theorem.

Schur-Horn Theorem *There exists a positive semidefinite matrix with spectrum* $\{\lambda_m\}_{m=1}^M$ *and diagonal entries* $\{\mu_m\}_{m=1}^M$ *if and only if* $\{\lambda_m\}_{m=1}^M \succeq \{\mu_m\}_{m=1}^M$.

Over the years, several methods for explicitly constructing Horn's matrices have been found; see [15] for a nice overview. Many current methods rely on Givens rotations [13, 15, 33], while others involve optimization [14]. Regarding frame theory, the significance of the Schur-Horn theorem is that it completely characterizes whether or not there exists a frame whose frame operator has a given spectrum and whose vectors have given lengths. This follows from applying it to the *Gram matrix* $\Phi^*\Phi$, whose diagonal entries are the values $\{\|\varphi_m\|^2\}_{m=1}^M$ and whose spectrum $\{\lambda_m\}_{m=1}^M$ is a zero-padded version of the spectrum $\{\lambda_n\}_{n=1}^N$ of the frame operator $\Phi\Phi^*$. Indeed, majorization inequalities arose during the search for tight frames with given lengths [8, 16], and the explicit connection between frames and the Schur-Horn theorem was noted in [1, 32]. This connection was then exploited to solve various frame theory problems, such as frame completion [28].

Certainly, any solution to Problem 2.1 must account for the fact that frames exist precisely when the Schur-Horn majorization condition is satisfied. In this paper, we solve Problem 2.1 by iteratively selecting frame elements in a way that guarantees majorization holds in the end. We start in Sect. 2.2 by reviewing the UNTF construction method of [10] called Spectral Tetris, which selects one or two frame elements

at a time in a way that preserves the frame operator's eigenbasis. This permits a simple analysis of how the frame operator's spectrum changes with each iteration, but it lacks the generality needed to solve Problem 2.1. Section 2.3 tackles the generality: it discusses a two-step process from [5] which constructs every frame of a given spectrum and set of lengths. The first step, Step A, finds every possible way in which a frame's spectrum evolves when defining one frame element at a time. Step B then finds every possible choice of frame elements that corresponds to each evolution of spectra. Finally, Sects. 2.4 and 2.5 complete this solution to Problem 2.1 by providing explicit algorithms [5, 19] that accomplish Steps A and B, respectively.

2.2 Spectral Tetris

In this section, we discuss the Spectral Tetris method of constructing UNTFs. This method first appeared in [10], and has since been further studied and generalized [6, 7, 11]; this section presents the original version from [10]. Our goal is to construct $N \times M$ synthesis matrices $\Phi = \{\varphi_m\}_{m=1}^M$ which have:

(i) columns of unit norm,
(ii) orthogonal rows, meaning the frame operator $\Phi\Phi^*$ is diagonal,
(iii) rows of equal norm, meaning $\Phi\Phi^*$ is a multiple of the identity matrix.

Spectral Tetris builds such Φ's iteratively; the name stems from the fact that it builds a flat spectrum out of blocks of fixed area. In short, Spectral Tetris ensures that, with each iteration, our matrices leading to Φ will exactly satisfy (i) and (ii), and get closer to satisfying (iii). Here, an illustrative example is helpful.

Example 2.1 Let's play Spectral Tetris to build a UNTF of 11 elements in $\mathbb{C}^4$: a 4×11 matrix whose columns have norm one and whose rows are orthogonal and square sum to $\frac{11}{4}$. We begin with an arbitrary 4×11 matrix, and let the first two frame elements be copies of the first standard basis element δ_1:

$$\Phi = \begin{bmatrix} 1 & 1 & ? & ? & ? & ? & ? & ? & ? & ? & ? \\ 0 & 0 & ? & ? & ? & ? & ? & ? & ? & ? & ? \\ 0 & 0 & ? & ? & ? & ? & ? & ? & ? & ? & ? \\ 0 & 0 & ? & ? & ? & ? & ? & ? & ? & ? & ? \end{bmatrix}. \tag{2.5}$$

If the remaining unknown entries are chosen so that Φ has orthogonal rows, then $\Phi\Phi^*$ will be a diagonal matrix. Currently, the diagonal entries of $\Phi\Phi^*$ are mostly unknown, having the form $\{2+?, ?, ?, ?\}$. Also note that if the remainder of the first row of Φ is set to zero, then the first diagonal entry of $\Phi\Phi^*$ would be $2 < \frac{11}{4}$. Thus, we need to add more weight to this row. However, making the third column of Φ another copy of δ_1 would add too much weight, as $3 > \frac{11}{4}$. Therefore, we need a way to give $\frac{11}{4} - 2 = \frac{3}{4}$ more weight in the first row without compromising either the orthogonality of the rows of Φ or the normality of its columns. The key idea is to realize that, for any $0 \leq x \leq 2$, there exists a 2×2 matrix $T(x)$ with

orthogonal rows and unit-length columns such that $T(x)T^*(x)$ is a diagonal matrix with diagonal entries $\{x, 2-x\}$. Specifically, we have:

$$T(x) := \frac{1}{\sqrt{2}}\begin{bmatrix} \sqrt{x} & \sqrt{x} \\ \sqrt{2-x} & -\sqrt{2-x} \end{bmatrix}, \qquad T(x)T^*(x) = \begin{bmatrix} x & 0 \\ 0 & 2-x \end{bmatrix}.$$

We define the third and fourth columns of Φ in terms of $T(x)$, where $x = \frac{11}{4} - 2 = \frac{3}{4}$:

$$\Phi = \begin{bmatrix} 1 & 1 & \frac{\sqrt{3}}{\sqrt{8}} & \frac{\sqrt{3}}{\sqrt{8}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{5}}{\sqrt{8}} & -\frac{\sqrt{5}}{\sqrt{8}} & ? & ? & ? & ? & ? & ? & ? \\ 0 & 0 & 0 & 0 & ? & ? & ? & ? & ? & ? & ? \\ 0 & 0 & 0 & 0 & ? & ? & ? & ? & ? & ? & ? \end{bmatrix}. \tag{2.6}$$

The diagonal entries of $\Phi\Phi^*$ are now $\{\frac{11}{4}, \frac{5}{4} + ?, ?, ?\}$. The first row now has sufficient weight, and so its remaining entries are set to zero. The second entry is currently falling short by $\frac{11}{4} - \frac{5}{4} = \frac{6}{4} = 1 + \frac{2}{4}$, and as such, we make the fifth column δ_2, while the sixth and seventh arise from $T(\frac{2}{4})$:

$$\Phi = \begin{bmatrix} 1 & 1 & \frac{\sqrt{3}}{\sqrt{8}} & \frac{\sqrt{3}}{\sqrt{8}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{5}}{\sqrt{8}} & -\frac{\sqrt{5}}{\sqrt{8}} & 1 & \frac{\sqrt{2}}{\sqrt{8}} & \frac{\sqrt{2}}{\sqrt{8}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\sqrt{6}}{\sqrt{8}} & -\frac{\sqrt{6}}{\sqrt{8}} & ? & ? & ? & ? \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & ? & ? & ? & ? \end{bmatrix}. \tag{2.7}$$

The diagonal entries of $\Phi\Phi^*$ are now $\{\frac{11}{4}, \frac{11}{4}, \frac{6}{4} + ?, ?\}$, where the third diagonal entry is falling short by $\frac{11}{4} - \frac{6}{4} = \frac{5}{4} = 1 + \frac{1}{4}$. We therefore take the eighth column of Φ as δ_3, let the ninth and tenth columns arise from $T(\frac{1}{4})$, and make the final column be δ_4, yielding the desired UNTF:

$$\Phi = \begin{bmatrix} 1 & 1 & \frac{\sqrt{3}}{\sqrt{8}} & \frac{\sqrt{3}}{\sqrt{8}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \frac{\sqrt{5}}{\sqrt{8}} & -\frac{\sqrt{5}}{\sqrt{8}} & 1 & \frac{\sqrt{2}}{\sqrt{8}} & \frac{\sqrt{2}}{\sqrt{8}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \frac{\sqrt{6}}{\sqrt{8}} & -\frac{\sqrt{6}}{\sqrt{8}} & 1 & \frac{1}{\sqrt{8}} & \frac{1}{\sqrt{8}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \frac{\sqrt{7}}{\sqrt{8}} & -\frac{\sqrt{7}}{\sqrt{8}} & 1 \end{bmatrix}. \tag{2.8}$$

In this construction, column vectors are either introduced one at a time, such as $\{\varphi_1\}$, $\{\varphi_2\}$, $\{\varphi_5\}$, $\{\varphi_8\}$, or $\{\varphi_{11}\}$, or in pairs, such as $\{\varphi_3, \varphi_4\}$, $\{\varphi_6, \varphi_7\}$, or $\{\varphi_9, \varphi_{10}\}$. Each singleton contributes a value of 1 to a particular diagonal entry of $\Phi\Phi^*$, while each pair spreads two units of weight over two entries. Overall, we have formed a flat spectrum, $\{\frac{11}{4}, \frac{11}{4}, \frac{11}{4}, \frac{11}{4}\}$, from blocks of area 1 or 2. This construction is reminiscent of the game Tetris, as illustrated in Fig. 2.1.

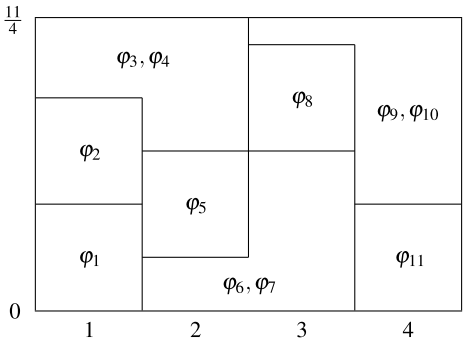


Fig. 2.1 The Spectral Tetris construction of a UNTF of 11 elements for $\mathbb{C}^4$, as detailed in Example 2.1. Each of the four columns corresponds to a diagonal entry of the frame operator $\Phi\Phi^*$, and each block represents the contribution made to these entries by the corresponding frame elements. For example, the single frame element $\{\varphi_2\}$ contributes $\{1, 0, 0, 0\}$ to the diagonal, while the pair $\{\varphi_6, \varphi_7\}$ contributes $\{0, \frac{2}{4}, \frac{6}{4}, 0\}$. The area of the blocks is determined by the number of frame elements that generate them: blocks that arise from a single element have unit area, while blocks that arise from two elements have an area of 2. In order for $\{\varphi_m\}_{m=1}^{11}$ to be a UNTF for $\mathbb{C}^4$, these blocks need to stack to a uniform height of $\frac{11}{4}$. By building a rectangle from blocks of given areas, we are essentially playing Tetris with the spectrum of $\Phi\Phi^*$

We conclude this example by pointing out some useful consequences of this Spectral Tetris construction. First, note that the frame vectors in (2.8) are extremely sparse. In fact, Spectral Tetris constructs optimally sparse UNTFs [11]. Also note that many pairs of frame vectors in this example have mutually disjoint support. In particular, we have that φ_m and $\varphi_{m'}$ are orthogonal whenever $m - m' \geq 5$. This feature of Spectral Tetris frames is exploited in [10] to construct tight fusion frames.

In order to formalize the Spectral Tetris argument used in the previous example, we introduce the following notion.

Definition 2.1 We say that a sequence $\{\varphi_m\}_{m=1}^M$ is an (m_0, n_0)-*proto unit norm tight frame* (PUNTF) for $\mathbb{C}^N$ if:

(i) $\sum_{n=1}^N |\varphi_m(n)|^2 = \begin{cases} 1, & m \leq m_0, \\ 0, & m > m_0, \end{cases}$

(ii) $\sum_{m=1}^M \varphi_m(n)[\varphi_m(n')]^* = 0$ for all $n, n' = 1, \ldots, N, n \neq n'$,

(iii) $\sum_{m=1}^M |\varphi_m(n)|^2 = \begin{cases} \frac{M}{N}, & n < n_0, \\ 0, & n > n_0, \end{cases}$

(iv) $1 \leq \sum_{m=1}^M |\varphi_m(n_0)|^2 \leq \frac{M}{N}$.

Here and throughout, z^* denotes the complex conjugate of a complex scalar z, as it corresponds to the conjugate transpose of a 1×1 matrix. That is, $\{\varphi_m\}_{m=1}^M$ is an (m_0, n_0)-PUNTF for $\mathbb{C}^N$ precisely when its $N \times M$ synthesis matrix Φ vanishes off its upper left $n_0 \times m_0$ submatrix, its nonzero columns have unit norm, and its frame

operator $\Phi\Phi^*$ is diagonal, with the first $n_0 - 1$ diagonal entries being $\frac{M}{N}$, the n_0th entry lying in $[1, \frac{M}{N}]$, and the remaining entries being zero. In particular, setting "?" entries to zero in (2.5), (2.6), (2.7), and (2.8) results in $(2, 1)$-, $(4, 2)$-, $(7, 3)$-, and $(11, 4)$-PUNTFs, respectively. As seen in Example 2.1, the goal of Spectral Tetris is to iteratively create larger PUNTFs from existing ones, continuing until $(m_0, n_0) = (M, N)$, at which point the PUNTF is a UNTF. We now give the precise rules for enlarging a given PUNTF; here, as in the preceding example, $\{\delta_n\}_{n=1}^N$ is the standard basis of $\mathbb{C}^N$.

Theorem 2.1 *Let* $2N \leq M$, $\{\varphi_m\}_{m=1}^M$ *be an* (m_0, n_0)*-PUNTF, and* $\lambda := \sum_{m=1}^M |\varphi_m(n_0)|^2$.

(i) *If* $\lambda \leq \frac{M}{N} - 1$, *then* $m_0 < M$ *and* $\{g_m\}_{m=1}^M$ *is an* $(m_0 + 1, n_0)$*-PUNTF, where*

$$g_m := \begin{cases} \varphi_m, & m \leq m_0, \\ \delta_{n_0}, & m = m_0 + 1, \\ 0, & m > m_0 + 1. \end{cases}$$

(ii) *If* $\frac{M}{N} - 1 < \lambda < \frac{M}{N}$, *then* $m_0 < M - 2$, $n_0 < N$, *and* $\{g_m\}_{m=1}^M$, *with*

$$g_m := \begin{cases} \varphi_m, & m \leq m_0, \\ \sqrt{\frac{1}{2}(\frac{M}{N} - \lambda)}\delta_{n_0} + \sqrt{1 - \frac{1}{2}(\frac{M}{N} - \lambda)}\delta_{n_0+1}, & m = m_0 + 1, \\ \sqrt{\frac{1}{2}(\frac{M}{N} - \lambda)}\delta_{n_0} - \sqrt{1 - \frac{1}{2}(\frac{M}{N} - \lambda)}\delta_{n_0+1}, & m = m_0 + 2, \\ 0, & m > m_0 + 2, \end{cases}$$

is an $(m_0 + 2, n_0 + 1)$*-PUNTF.*

(iii) *If* $\lambda = \frac{M}{N}$ *and* $n_0 < N$, *then* $m_0 < M$ *and* $\{g_m\}_{m=1}^M$, *with*

$$g_m := \begin{cases} \varphi_m, & m \leq m_0, \\ \delta_{n_0+1}, & m = m_0 + 1, \\ 0, & m > m_0 + 1, \end{cases}$$

is an $(m_0 + 1, n_0 + 1)$*-PUNTF for* $\mathbb{C}^N$.

(iv) *If* $\lambda = \frac{M}{N}$ *and* $n_0 = N$, *then* $\{\varphi_m\}_{m=1}^M$ *is a UNTF.*

The proof of Theorem 2.1 can be found in [10]. For this proof, the assumption $2N \leq M$ is crucial; in the case where λ is slightly smaller than $\frac{M}{N}$, the $(n_0 + 1)$th diagonal entry of $\Phi\Phi^*$ must accept nearly two spectral units of weight, which is only possible when the desired Spectral Tetris height $\frac{M}{N}$ is at least 2. At the same time, we note that playing Spectral Tetris can also result in matrices of lesser redundancy, provided larger blocks are used. Indeed, UNTFs of redundancy $\frac{M}{N} \geq \frac{3}{2}$ can be constructed using 3×3 Spectral Tetris submatrices, as we now have two diagonal entries over which to spread at most three units of spectral weight; the blocks themselves are obtained by scaling the rows of a 3×3 discrete Fourier transform matrix.

More generally, UNTFs with redundancy greater than $\frac{J}{J-1}$ can be constructed using $J \times J$ submatrices. Note that these lower levels of redundancy are only bought at the expense of a loss in sparsity, and in particular, a loss of orthogonality relations between the frame elements themselves. These ideas are further explored in [7].

Also note that although this section's results were proved in complex Euclidean space for the sake of consistency, the frames obtained by playing Spectral Tetris with 1×1 and 2×2 submatrices are, in fact, real-valued. The simplicity of this construction rivals that of real harmonic frames, which consist of samples of sines and cosines. In particular, Spectral Tetris provides a very simple proof of the existence of real UNTFs for any $M \geq N$: when $2N \leq M$, the construction is direct; Naimark complements [10] then give real UNTFs with redundancy less than two. Spectral Tetris can also be used to construct nontight frames [6] provided the spectrum is bounded below by 2. Unfortunately, these techniques are insufficient to solve Problem 2.1. The next section details a process for solving that problem.

2.3 The Necessity and Sufficiency of Eigensteps

In the previous section, we presented the Spectral Tetris algorithm, which systematically builds UNTFs one or two vectors at a time. There, the main idea was to iteratively construct frame elements in a manner that changes the frame operator's spectrum in a predictable way while at the same time preserving its eigenbasis. However, Spectral Tetris itself cannot solve Problem 2.1 in generality: it only works with unit vectors and with spectra in which each eigenvalue is at least two in value. Moreover, even in that case, it only seems to produce a narrow class of all possible such frames.

In this section, we present the method of [5], which generalizes the Spectral Tetris idea in a way that provides a complete solution to Problem 2.1. Like Spectral Tetris, this method constructs $\Phi = \{\varphi_m\}_{m=1}^M$ in a manner so that at any given $m = 1, \ldots, M$, we know the spectrum of the frame operator

$$\Phi_m \Phi_m^* = \sum_{m'=1}^{m} \varphi_{m'} \varphi_{m'}^* \tag{2.9}$$

of the partial sequence $\Phi_m := \{\varphi_{m'}\}_{m'=1}^m$. However, unlike Spectral Tetris, this method will not require the eigenbasis of (2.9) to be the standard basis for all m. Indeed, the opposite is true: this method requires this eigenbasis to evolve with m.

The key idea is to realize from (2.9) that $\Phi_{m+1}^* \Phi_{m+1} = \Phi_m^* \Phi_m + \varphi_{m+1}^* \varphi_{m+1}$. From this perspective, Problem 2.1 comes down to understanding how the spectrum of a given positive semidefinite operator $\Phi_m^* \Phi_m$ is affected by the addition of a scaled rank-one projection operator $\varphi_{m+1}^* \varphi_{m+1}$ of trace μ_{m+1}. Such problems have been studied classically, and involve a concept called eigenvalue interlacing.

To be precise, a nonnegative nonincreasing sequence $\{\gamma_n\}_{n=1}^N$ *interlaces* on another such sequence $\{\beta_n\}_{n=1}^N$, denoted $\{\beta_n\}_{n=1}^N \sqsubseteq \{\gamma_n\}_{n=1}^N$, provided that

$$\beta_N \leq \gamma_N \leq \beta_{N-1} \leq \gamma_{N-1} \leq \cdots \leq \beta_2 \leq \gamma_2 \leq \beta_1 \leq \gamma_1. \tag{2.10}$$

The classical theory of eigenvalue interlacing [25] tells us that letting $\{\lambda_{m;n}\}_{n=1}^{N}$ denote the spectrum of (2.9), we necessarily have that $\{\lambda_{m;n}\}_{n=1}^{N} \sqsubseteq \{\lambda_{m+1;n}\}_{n=1}^{N}$. Moreover, if $\|\varphi_m\|^2 = \mu_m$ for all $m = 1, \ldots, M$, then for any such m,

$$\sum_{n=1}^{N} \lambda_{m;n} = \operatorname{Tr}\big(\Phi_m \Phi_m^*\big) = \operatorname{Tr}\big(\Phi_m^* \Phi_m\big) = \sum_{m'=1}^{m} \|\varphi_{m'}\|^2 = \sum_{m'=1}^{m} \mu_{m'}. \tag{2.11}$$

In [19], interlacing spectra that satisfy (2.11) are called a sequence of *outer eigensteps*.

Definition 2.2 Let $\{\lambda_n\}_{n=1}^{N}$ and $\{\mu_m\}_{m=1}^{M}$ be nonnegative nonincreasing sequences. A corresponding sequence of *outer eigensteps* is a sequence $\{\{\lambda_{m;n}\}_{n=1}^{N}\}_{m=0}^{M}$ which satisfies the following four properties:

(i) $\lambda_{0;n} = 0$ for every $n = 1, \ldots, N$,
(ii) $\lambda_{M;n} = \lambda_n$ for every $n = 1, \ldots, N$,
(iii) $\{\lambda_{m-1;n}\}_{n=1}^{N} \sqsubseteq \{\lambda_{m;n}\}_{n=1}^{N}$ for every $m = 1, \ldots, M$,
(iv) $\sum_{n=1}^{N} \lambda_{m;n} = \sum_{n=1}^{m} \mu_n$ for every $m = 1, \ldots, M$.

As we have just discussed, every sequence of vectors whose frame operator has the spectrum $\{\lambda_n\}_{n=1}^{N}$ and whose vectors have squared lengths $\{\mu_m\}_{m=1}^{M}$ generates a sequence of outer eigensteps. By the following theorem, the converse is also true. Specifically, Theorem 2.2 characterizes and proves the existence of sequences of vectors that generate a given sequence of outer eigensteps. We will see that once the outer eigensteps have been chosen, there is little freedom in picking the frame vectors themselves. That is, modulo rotations, the outer eigensteps are the free parameters when designing a frame whose frame operator has a given spectrum and whose vectors have given lengths.

Theorem 2.2 *For any nonnegative nonincreasing sequences $\{\lambda_n\}_{n=1}^{N}$ and $\{\mu_m\}_{m=1}^{M}$, every sequence of vectors $\Phi = \{\varphi_m\}_{m=1}^{M}$ in $\mathbb{C}^N$ whose frame operator $\Phi\Phi^*$ has spectrum $\{\lambda_n\}_{n=1}^{N}$ and which satisfies $\|\varphi_m\|^2 = \mu_m$ for all m can be constructed by the following process*:

Step A. *Pick outer eigensteps $\{\{\lambda_{m;n}\}_{n=1}^{N}\}_{m=0}^{M}$ as in Definition 2.2.*
Step B. *For each $m = 1, \ldots, M$, consider the polynomial*

$$p_m(x) := \prod_{n=1}^{N} (x - \lambda_{m;n}). \tag{2.12}$$

Take any $\varphi_1 \in \mathbb{C}^N$ such that $\|\varphi_1\|^2 = \mu_1$. For each $m = 1, \ldots, M-1$, choose any φ_{m+1} such that

$$\|P_{m;\lambda}\varphi_{m+1}\|^2 = -\lim_{x \to \lambda} (x - \lambda) \frac{p_{m+1}(x)}{p_m(x)} \tag{2.13}$$

for all $\lambda \in \{\lambda_{m;n}\}_{n=1}^N$, where $P_{m;\lambda}$ denotes the orthogonal projection operator onto the eigenspace $\mathrm{N}(\lambda Id - \Phi_m\Phi_m^*)$ of the frame operator $\Phi_m\Phi_m^*$ of the partial sequence $\Phi_m = \{\varphi_{m'}\}_{m'=1}^m$. The limit in (2.13) exists and is nonpositive.

Conversely, any Φ constructed by this process has $\{\lambda_n\}_{n=1}^N$ as the spectrum of $\Phi\Phi^*$ and $\|\varphi_m\|^2 = \mu_m$ for all m. Moreover, for any Φ constructed in this manner, the spectrum of $\Phi_m\Phi_m^*$ is $\{\lambda_{m;n}\}_{n=1}^N$ for all $m = 1, \ldots, M$.

In order to prove Theorem 2.2, we first obtain some supporting results. In particular, the next result gives conditions that a vector must satisfy in order for it to perturb the spectrum of a given frame operator in a desired way, and was inspired by the proof of the Matrix Determinant Lemma and its application in [2].

Theorem 2.3 *Let $\Phi_m = \{\varphi_{m'}\}_{m'=1}^m$ be an arbitrary sequence of vectors in $\mathbb{C}^N$ and let $\{\lambda_{m;n}\}_{n=1}^N$ denote the eigenvalues of the corresponding frame operator $\Phi_m\Phi_m^*$. For any choice of φ_{m+1} in $\mathbb{C}^N$, let $\Phi_{m+1} = \{\varphi_{m'}\}_{m'=1}^{m+1}$. Then for any $\lambda \in \{\lambda_{m;n}\}_{n=1}^N$, the norm of the projection of φ_{m+1} onto the eigenspace $\mathrm{N}(\lambda Id - \Phi_m\Phi_m^*)$ is given by*

$$\|P_{m;\lambda}\varphi_{m+1}\|^2 = -\lim_{x\to\lambda}(x-\lambda)\frac{p_{m+1}(x)}{p_m(x)},$$

where $p_m(x)$ and $p_{m+1}(x)$ denote the characteristic polynomials of $\Phi_m\Phi_m^*$ and $\Phi_{m+1}\Phi_{m+1}^*$, respectively.

Proof For notational simplicity, we let $\Phi := \Phi_m$, $\varphi := \varphi_{m+1}$ and so $\Phi_{m+1}\Phi_{m+1}^* = \Phi\Phi^* + \varphi\varphi^*$. Suppose x is not an eigenvalue of $\Phi_{m+1}\Phi_{m+1}^*$. Then:

$$\begin{aligned} p_{m+1}(x) &= \det\left(xId - \Phi\Phi^* - \varphi\varphi^*\right) \\ &= \det\left(xId - \Phi\Phi^*\right)\det\left(Id - \left(xId - \Phi\Phi^*\right)^{-1}\varphi\varphi^*\right) \\ &= p_m(x)\det\left(Id - \left(xId - \Phi\Phi^*\right)^{-1}\varphi\varphi^*\right). \end{aligned} \tag{2.14}$$

We can simplify the determinant of $Id - (xId - \Phi\Phi^*)^{-1}\varphi\varphi^*$ by multiplying by certain matrices with unit determinant:

$$\begin{aligned} &\det\left(Id - \left(xId - \Phi\Phi^*\right)^{-1}\varphi\varphi^*\right) \\ &\quad= \det\left(\begin{bmatrix} Id & 0 \\ \varphi^* & 1 \end{bmatrix}\begin{bmatrix} Id - (xId - \Phi\Phi^*)^{-1}\varphi\varphi^* & -(xId - \Phi\Phi^*)^{-1}\varphi \\ 0 & 1 \end{bmatrix}\right. \\ &\qquad\left.\times\begin{bmatrix} Id & 0 \\ -\varphi^* & 1 \end{bmatrix}\right) \\ &\quad= \det\left(\begin{bmatrix} Id & 0 \\ \varphi^* & 1 \end{bmatrix}\begin{bmatrix} Id & -(xId - \Phi\Phi^*)^{-1}\varphi \\ -\varphi^* & 1 \end{bmatrix}\right) \end{aligned}$$

$$= \det\left(\begin{bmatrix} Id & -(xId - \Phi\Phi^*)^{-1}\varphi \\ 0 & 1 - \varphi^*(xId - \Phi\Phi^*)^{-1}\varphi \end{bmatrix}\right)$$

$$= 1 - \varphi^*\big(xId - \Phi\Phi^*\big)^{-1}\varphi. \tag{2.15}$$

We now use (2.14) and (2.15) with the spectral decomposition $\Phi\Phi^* = \sum_{n=1}^{N} \lambda_{m;n} u_n u_n^*$:

$$p_{m+1}(x) = p_m(x)\big(1 - \varphi^*\big(xId - \Phi\Phi^*\big)^{-1}\varphi\big) = p_m(x)\left(1 - \sum_{n=1}^{N} \frac{|\langle \varphi, u_n\rangle|^2}{x - \lambda_{m;n}}\right). \tag{2.16}$$

Rearranging (2.16) and grouping the eigenvalues $\Lambda = \{\lambda_{m;n}\}_{n=1}^{N}$ according to multiplicity then gives

$$\frac{p_{m+1}(x)}{p_m(x)} = 1 - \sum_{n=1}^{N} \frac{|\langle \varphi, u_n\rangle|^2}{x - \lambda_{m;n}} = 1 - \sum_{\lambda' \in \Lambda} \frac{\|P_{m;\lambda'}\varphi\|^2}{x - \lambda'}, \quad \forall x \notin \Lambda.$$

As such, for any $\lambda \in \Lambda$,

$$\begin{aligned} \lim_{x\to\lambda} (x - \lambda)\frac{p_{m+1}(x)}{p_m(x)} &= \lim_{x\to\lambda} (x - \lambda)\left(1 - \sum_{\lambda' \in \Lambda} \frac{\|P_{m;\lambda'}\varphi\|^2}{x - \lambda'}\right) \\ &= \lim_{x\to\lambda}\left[(x - \lambda) - \|P_{m;\lambda}\varphi\|^2 - \sum_{\lambda' \neq \lambda} \|P_{m;\lambda'}\varphi\|^2 \frac{x - \lambda}{x - \lambda'}\right] \\ &= -\|P_{m;\lambda}\varphi\|^2, \end{aligned}$$

yielding our claim. □

Though technical, the proofs of the next three lemmas are nonetheless elementary; the interested reader can find them in [5].

Lemma 2.1 *Let $\{\lambda_n\}_{n=1}^{N}$ and $\{\mu_m\}_{m=1}^{M}$ be nonnegative and nonincreasing, and let $\{\{\lambda_{m;n}\}_{n=1}^{N}\}_{m=0}^{M}$ be any corresponding sequence of outer eigensteps as in Definition 2.2. If a sequence of vectors $\Phi = \{\varphi_m\}_{m=1}^{M}$ has the property that the spectrum of the frame operator $\Phi_m\Phi_m^*$ of $\Phi_m = \{\varphi_{m'}\}_{m'=1}^{m}$ is $\{\lambda_{m;n}\}_{n=1}^{N}$ for all $m = 1, \ldots, M$, then the spectrum of $\Phi\Phi^*$ is $\{\lambda_n\}_{n=1}^{N}$ and $\|\varphi_m\|^2 = \mu_m$ for all $m = 1, \ldots, M$.*

Lemma 2.2 *If $\{\beta_n\}_{n=1}^{N}$ and $\{\gamma_n\}_{n=1}^{N}$ are nonincreasing, then $\{\beta_n\}_{n=1}^{N} \sqsubseteq \{\gamma_n\}_{n=1}^{N}$ if and only if*

$$\lim_{x\to\beta_n} (x - \beta_n)\frac{q(x)}{p(x)} \leq 0, \quad \forall n = 1, \ldots, N,$$

where $p(x) = \prod_{n=1}^{N}(x - \beta_n)$ and $q(x) = \prod_{n=1}^{N}(x - \gamma_n)$.

Lemma 2.3 *If* $\{\beta_n\}_{n=1}^N$, $\{\gamma_n\}_{n=1}^N$, *and* $\{\delta_n\}_{n=1}^N$ *are nonincreasing and*

$$\lim_{x\to\beta_n}(x-\beta_n)\frac{q(x)}{p(x)}=\lim_{x\to\beta_n}(x-\beta_n)\frac{r(x)}{p(x)},\quad \forall n=1,\ldots,N,$$

where $p(x)=\prod_{n=1}^N(x-\beta_n)$, $q(x)=\prod_{n=1}^N(x-\gamma_n)$, *and* $r(x)=\prod_{n=1}^N(x-\delta_n)$, *then* $q(x)=r(x)$.

The preceding results in hand, we turn to the main result of this section.

Proof of Theorem 2.2 ($\Rightarrow$) Let $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$ be arbitrary nonnegative nonincreasing sequences, and let $\Phi=\{\varphi_m\}_{m=1}^M$ be any sequence of vectors such that the spectrum of $\Phi\Phi^*$ is $\{\lambda_n\}_{n=1}^N$ and $\|\varphi_m\|^2=\mu_m$ for all $m=1,\ldots,M$. We claim that this particular Φ can be constructed by following Steps A and B.

In particular, consider the sequence $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$ defined by letting $\{\lambda_{m;n}\}_{n=1}^N$ be the spectrum of the frame operator $\Phi_m\Phi_m^*$ of the sequence $\Phi_m=\{\varphi_{m'}\}_{m'=1}^m$ for all $m=1,\ldots,M$ and letting $\lambda_{0;n}=0$ for all n. We claim that $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$ satisfies Definition 2.2 and therefore is a valid sequence of eigensteps. Note that conditions (i) and (ii) of Definition 2.2 are immediately satisfied. To see that $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$ satisfies (iii), consider the polynomials $p_m(x)$ defined by (2.12) for all $m=1,\ldots,M$. In the special case where $m=1$, the desired property (iii) that $\{0\}_{n=1}^N\sqsubseteq\{\lambda_{1;n}\}_{n=1}^N$ follows from the fact that the spectrum $\{\lambda_{1;n}\}_{n=1}^N$ of the scaled rank-one projection $\Phi_1\Phi_1^*=\varphi_1\varphi_1^*$ is the value $\|\varphi_1\|^2=\mu_1$ along with $N-1$ repetitions of 0, the eigenspaces being the span of φ_1 and its orthogonal complement, respectively. Meanwhile, if $m=2,\ldots,M$, Theorem 2.3 gives that

$$\lim_{x\to\lambda_{m-1;n}}(x-\lambda_{m-1;n})\frac{p_m(x)}{p_{m-1}(x)}=-\|P_{m-1;\lambda_{m-1;n}}\varphi_m\|^2\leq 0,\quad \forall n=1,\ldots,N,$$

implying by Lemma 2.2 that $\{\lambda_{m-1;n}\}_{n=1}^N\sqsubseteq\{\lambda_{m;n}\}_{n=1}^N$ as claimed. Finally, (iv) holds, since for any $m=1,\ldots,M$ we have

$$\sum_{n=1}^N\lambda_{m;n}=\mathrm{Tr}\big(\Phi_m\Phi_m^*\big)=\mathrm{Tr}\big(\Phi_m^*\Phi_m\big)=\sum_{m'=1}^m\|\varphi_{m'}\|^2=\sum_{m'=1}^m\mu_{m'}.$$

Having shown that these particular values of $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$ can indeed be chosen in Step A, we next show that our particular Φ can be constructed according to Step B. As the method of Step B is iterative, we use induction to prove that it can yield Φ. Indeed, the only restriction that Step B places on φ_1 is that $\|\varphi_1\|^2=\mu_1$, something our particular φ_1 satisfies by assumption. Now assume that for any $m=1,\ldots,M-1$ we have already correctly produced $\{\varphi_{m'}\}_{m'=1}^m$ by following the method of Step B; we show that we can produce the correct φ_{m+1} by continuing to follow Step B. To be clear, each iteration of Step B does not produce a unique vector, but rather presents a family of φ_{m+1}'s to choose from, and we show that our particular choice of φ_{m+1} lies in this family. Specifically, our choice of

φ_{m+1} must satisfy (2.13) for any choice of $\lambda \in \{\lambda_{m;n}\}_{n=1}^{N}$; the fact that it indeed does so follows immediately from Theorem 2.3. To summarize, we have shown that, by making appropriate choices, we can indeed produce our particular Φ by following Steps A and B, concluding this direction of the proof.

($\Leftarrow$) Now assume that a sequence of vectors $\Phi = \{\varphi_m\}_{m=1}^{M}$ has been produced according to Steps A and B. To be precise, letting $\{\{\lambda_{m;n}\}_{n=1}^{N}\}_{m=0}^{M}$ be the sequence of eigensteps chosen in Step A, we claim that any $\Phi = \{\varphi_m\}_{m=1}^{M}$ constructed according to Step B has the property that the spectrum of the frame operator $\Phi_m\Phi_m^*$ of $\Phi_m = \{\varphi_{m'}\}_{m'=1}^{m}$ is $\{\lambda_{m;n}\}_{n=1}^{N}$ for all $m = 1, \ldots, M$. Note that by Lemma 2.1, proving this claim will yield our stated result that the spectrum of $\Phi\Phi^*$ is $\{\lambda_n\}_{n=1}^{N}$ and that $\|\varphi_m\|^2 = \mu_m$ for all $m = 1, \ldots, M$. As the method of Step B is iterative, we prove this claim by induction. Step B begins by taking any φ_1 such that $\|\varphi_1\|^2 = \mu_1$. As noted above in the proof of the other direction, the spectrum of $\Phi_1\Phi_1^* = \varphi_1\varphi_1^*$ is the value μ_1 along with $N-1$ repetitions of 0. As claimed, these values match those of $\{\lambda_{1;n}\}_{n=1}^{N}$; to see this, note that Definition 2.2(i) and (iii) give $\{0\}_{n=1}^{N} = \{\lambda_{0;n}\}_{n=1}^{N} \sqsubseteq \{\lambda_{1;n}\}_{n=1}^{N}$ and so $\lambda_{1;n} = 0$ for all $n = 2, \ldots, N$, at which point Definition 2.2(iv) implies $\lambda_{1,1} = \mu_1$.

Now assume that for any $m = 1, \ldots, M-1$, the Step B process has already produced $\Phi_m = \{\varphi_{m'}\}_{m'=1}^{m}$ such that the spectrum of $\Phi_m\Phi_m^*$ is $\{\lambda_{m;n}\}_{n=1}^{N}$. We show that following Step B yields a φ_{m+1} such that $\Phi_{m+1} = \{\varphi_{m'}\}_{m'=1}^{m+1}$ has the property that $\{\lambda_{m+1;n}\}_{n=1}^{N}$ is the spectrum of $\Phi_{m+1}\Phi_{m+1}^*$. To do this, consider the polynomials $p_m(x)$ and $p_{m+1}(x)$ defined by (2.12) and pick any φ_{m+1} that satisfies (2.13), namely,

$$\lim_{x\to\lambda_{m;n}} (x - \lambda_{m;n})\frac{p_{m+1}(x)}{p_m(x)} = -\|P_{m;\lambda_{m;n}}\varphi_{m+1}\|^2, \quad \forall n = 1, \ldots, N. \tag{2.17}$$

Letting $\{\hat{\lambda}_{m+1;n}\}_{n=1}^{N}$ denote the spectrum of $\Phi_{m+1}\Phi_{m+1}^*$, our goal is to show that $\{\hat{\lambda}_{m+1;n}\}_{n=1}^{N} = \{\lambda_{m+1;n}\}_{n=1}^{N}$. Equivalently, our goal is to show that $p_{m+1}(x) = \hat{p}_{m+1}(x)$, where $\hat{p}_{m+1}(x)$ is the polynomial

$$\hat{p}_{m+1}(x) := \prod_{n=1}^{N}(x - \hat{\lambda}_{m+1;n}).$$

Since $p_m(x)$ and $\hat{p}_{m+1}(x)$ are the characteristic polynomials of $\Phi_m\Phi_m^*$ and $\Phi_{m+1}\Phi_{m+1}^*$, respectively, Theorem 2.3 gives

$$\lim_{x\to\lambda_{m;n}} (x - \lambda_{m;n})\frac{\hat{p}_{m+1}(x)}{p_m(x)} = -\|P_{m;\lambda_{m;n}}\varphi_{m+1}\|^2, \quad \forall n = 1, \ldots, N. \tag{2.18}$$

Comparing (2.17) and (2.18) gives

$$\lim_{x\to\lambda_{m;n}} (x - \lambda_{m;n})\frac{p_{m+1}(x)}{p_m(x)} = \lim_{x\to\lambda_{m;n}} (x - \lambda_{m;n})\frac{\hat{p}_{m+1}(x)}{p_m(x)}, \quad \forall n = 1, \ldots, N,$$

implying by Lemma 2.3 that $p_{m+1}(x) = \hat{p}_{m+1}(x)$, as desired. □

2.4 Parametrizing Eigensteps

In light of Theorem 2.2, solving Problem 2.1 comes down to finding every valid sequence of outer eigensteps $\{\{\lambda_{m;n}\}_{n=1}^{N}\}_{m=0}^{M}$, see Definition 2.2, for any given nonnegative nonincreasing sequences $\{\lambda_n\}_{n=1}^{N}$ and $\{\mu_m\}_{m=1}^{M}$. In this section, we detail the main results of [19], which give a systematic procedure for finding these eigensteps. We begin with an example from [5].

Example 2.2 We wish to parametrize all eigensteps for a particular case: UNTFs consisting of 5 vectors in $\mathbb{C}^3$. Here, $\lambda_1 = \lambda_2 = \lambda_3 = \frac{5}{3}$ and $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 1$. In light of Step A of Theorem 2.2, we seek outer eigensteps consistent with Definition 2.2; that is, we want to find all sequences $\{\{\lambda_{m;n}\}_{n=1}^{3}\}_{m=1}^{4}$ which satisfy the interlacing conditions

$$\{0\}_{n=1}^{3} \sqsubseteq \{\lambda_{1;n}\}_{n=1}^{3} \sqsubseteq \{\lambda_{2;n}\}_{n=1}^{3} \sqsubseteq \{\lambda_{3;n}\}_{n=1}^{3} \sqsubseteq \{\lambda_{4;n}\}_{n=1}^{3} \sqsubseteq \left\{\frac{5}{3}\right\}_{n=1}^{3}, \tag{2.19}$$

as well as the trace conditions

$$\sum_{n=1}^{3} \lambda_{1;n} = 1, \qquad \sum_{n=1}^{3} \lambda_{2;n} = 2, \qquad \sum_{n=1}^{3} \lambda_{3;n} = 3, \qquad \sum_{n=1}^{3} \lambda_{4;n} = 4. \tag{2.20}$$

Let us write these desired spectra in a table:

m	0	1	2	3	4	5
$\lambda_{m;3}$	0	?	?	?	?	$\frac{5}{3}$
$\lambda_{m;2}$	0	?	?	?	?	$\frac{5}{3}$
$\lambda_{m;1}$	0	?	?	?	?	$\frac{5}{3}$

In this table, the trace condition (2.20) means that the sum of the values in the mth column is $\sum_{n=1}^{m} \mu_n = m$, while the interlacing condition (2.19) means that any value $\lambda_{m;n}$ is at least the neighbor to the upper right $\lambda_{m+1;n+1}$ and no more than its neighbor to the right $\lambda_{m+1;n}$. In particular, for $m = 1$, we have $0 = \lambda_{0;2} \leq \lambda_{1;2} \leq \lambda_{0;1} = 0$ and $0 = \lambda_{0;3} \leq \lambda_{1;3} \leq \lambda_{0;2} = 0$, implying $\lambda_{1;2} = \lambda_{1;3} = 0$. Similarly, for $m = 4$, interlacing requires that $\frac{5}{3} = \lambda_{5;2} \leq \lambda_{4;1} \leq \lambda_{5;1} = \frac{5}{3}$ and $\frac{5}{3} = \lambda_{5;3} \leq \lambda_{4;2} \leq \lambda_{5;2} = \frac{5}{3}$, implying $\lambda_{4;1} = \lambda_{4;2} = \frac{5}{3}$. Applying this same idea again for $m = 2$ and $m = 3$ gives $\lambda_{2;3} = 0$ and $\lambda_{3;1} = \frac{5}{3}$. That is, we necessarily have

m	0	1	2	3	4	5
$\lambda_{m;3}$	0	0	0	?	?	$\frac{5}{3}$
$\lambda_{m;2}$	0	0	?	?	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	0	?	?	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

Moreover, the trace condition (2.20) at $m = 1$ gives $1 = \lambda_{1;1} + \lambda_{1;2} + \lambda_{1;3} = \lambda_{1;1} + 0 + 0$, and so $\lambda_{1;1} = 1$. Similarly, at $m = 4$ we have $4 = \lambda_{4;1} + \lambda_{4;2} + \lambda_{4;3} =$

$\frac{5}{3}+\frac{5}{3}+\lambda_{4;3}$, and so $\lambda_{4;3}=\frac{2}{3}$:

m	0	1	2	3	4	5
$\lambda_{m;3}$	0	0	0	?	$\frac{2}{3}$	$\frac{5}{3}$
$\lambda_{m;2}$	0	0	?	?	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	0	1	?	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

The remaining entries are not fixed. In particular, we let $\lambda_{3;3}$ be some variable x and note that by the trace condition, $3=\lambda_{3;1}+\lambda_{3;2}+\lambda_{3;3}=x+\lambda_{3;2}+\frac{5}{3}$ and so $\lambda_{3;2}=\frac{4}{3}-x$. Similarly, letting $\lambda_{2;2}=y$ gives $\lambda_{2;1}=2-y$:

m	0	1	2	3	4	5
$\lambda_{m;3}$	0	0	0	x	$\frac{2}{3}$	$\frac{5}{3}$
$\lambda_{m;2}$	0	0	y	$\frac{4}{3}-x$	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	0	1	$2-y$	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

(2.21)

We take care to note that x and y in (2.21) are not arbitrary, but instead must be chosen so that the requisite interlacing relations are satisfied:

$$\begin{aligned}
\{\lambda_{3;n}\}_{n=1}^{3}\sqsubseteq\{\lambda_{4;n}\}_{n=1}^{3} &\iff x\le\frac{2}{3}\le\frac{4}{3}-x\le\frac{5}{3},\\
\{\lambda_{2;n}\}_{n=1}^{3}\sqsubseteq\{\lambda_{3;n}\}_{n=1}^{3} &\iff 0\le x\le y\le\frac{4}{3}-x\le 2-y\le\frac{5}{3},\\
\{\lambda_{1;n}\}_{n=1}^{3}\sqsubseteq\{\lambda_{2;n}\}_{n=1}^{3} &\iff 0\le y\le 1\le 2-y.
\end{aligned}\tag{2.22}$$

By plotting each of the 11 inequalities of (2.22) as a half-plane (Fig. 2.2(a)), we obtain a convex pentagon (Fig. 2.2(b)) of all (x,y) such that (2.21) is a valid sequence of eigensteps. This example highlights the key obstacle in using Theorem 2.2 to solve Problem 2.1: finding all valid sequences of eigensteps (2.21) often requires reducing a large system of linear inequalities (2.22). We now consider a result which provides a method for finding all solutions to these systems.

Theorem 2.4 *Let $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$ be nonnegative and nonincreasing where $N\le M$. There exists a sequence of vectors $\Phi=\{\varphi_m\}_{m=1}^M$ in $\mathbb{C}^N$ whose frame operator $\Phi\Phi^*$ has spectrum $\{\lambda_n\}_{n=1}^N$ and for which $\|\varphi_m\|^2=\mu_m$ for all m if and only if $\{\lambda_n\}_{n=1}^N\cup\{0\}_{n=N+1}^M\succeq\{\mu_m\}_{m=1}^M$. Moreover, if $\{\lambda_n\}_{n=1}^N\cup\{0\}_{n=N+1}^M\succeq\{\mu_m\}_{m=1}^M$, then every such Φ can be constructed by the following process*:

Step A: *Let* $\{\lambda_{M;n}\}_{n=1}^N:=\{\lambda_n\}_{n=1}^N$.

For $m=M,\dots,2$, *construct* $\{\lambda_{m-1;n}\}_{n=1}^N$ *in terms of* $\{\lambda_{m;n}\}_{n=1}^N$ *as follows*:

For each $k=N,\dots,1$, *if* $k>m-1$, *take* $\lambda_{m-1;k}:=0$.

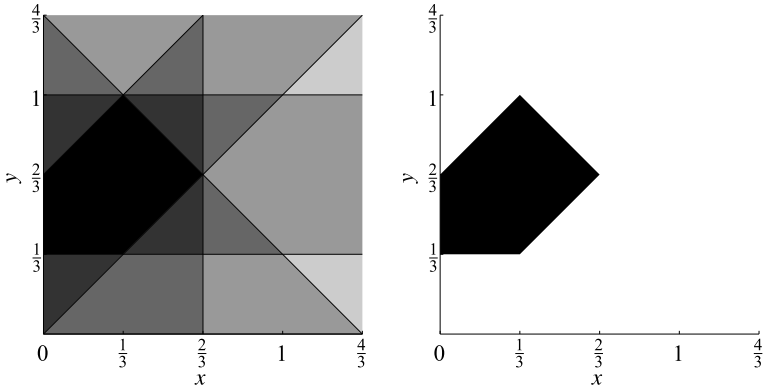


Fig. 2.2 Pairs of parameters (x, y) that generate a valid sequence of eigensteps when substituted into (2.21). To be precise, in order to satisfy the interlacing requirements of Definition 2.2, x and y must be chosen so as to satisfy the 11 pairwise inequalities summarized in (2.22). Each of these inequalities corresponds to a half-plane (a), and the set of pairs (x, y) that satisfy all of them is given by their intersection (b). By Theorem 2.2, any corresponding sequence of eigensteps (2.21) generates a 3×5 UNTF and, conversely, every 3×5 UNTF is generated in this way. As such, x and y may be viewed as the two essential parameters in the set of all such frames

Otherwise, pick any $\lambda_{m-1;k} \in [A_{m-1;k}, B_{m-1;k}]$, *where*:

$$A_{m-1;k} := \max\left\{\lambda_{m;k+1}, \sum_{n=k}^{N} \lambda_{m;n} - \sum_{n=k+1}^{N} \lambda_{m-1;n} - \mu_m\right\},$$

$$B_{m-1;k} := \min\left\{\lambda_{m;k}, \min_{l=1,\ldots,k}\left\{\sum_{n=l}^{m-1} \mu_n - \sum_{n=l+1}^{k} \lambda_{m;n} - \sum_{n=k+1}^{N} \lambda_{m-1;n}\right\}\right\}.$$

Here, by convention, $\lambda_{m;N+1} := 0$ *and sums over empty sets of indices are zero.*
Step B: *Follow Step B of Theorem 2.2.*

Conversely, any Φ *constructed by this process has* $\{\lambda_n\}_{n=1}^N$ *as the spectrum of* $\Phi\Phi^*$ *and* $\|\varphi_m\|^2 = \mu_m$ *for all* m, *and moreover,* $\Phi_m\Phi_m^*$ *has spectrum* $\{\lambda_{m;n}\}_{n=1}^N$.

It turns out that the method of Theorem 2.4 is more easily understood in terms of an alternative but equivalent notion of eigensteps. To be clear, for any given sequence of outer eigensteps $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$, recall from Theorem 2.2 that for any $m = 1, \ldots, M$, the sequence $\{\lambda_{m;n}\}_{n=1}^N$ is the spectrum of the $N \times N$ frame operator $\Phi_m\Phi_m^*$ of the mth partial sequence $\Phi_m = \{\varphi_{m'}\}_{m'=1}^m$. In the following theory, it is more convenient to instead work with the spectrum $\{\lambda_{m;m'}\}_{m'=1}^m$ of the corresponding $m \times m$ Gram matrix $\Phi_m^*\Phi_m$; we use the same notation for both spectra since $\{\lambda_{m;m'}\}_{m'=1}^m$ is a zero-padded version of $\{\lambda_{m;n}\}_{n=1}^N$ or vice versa, depending on whether $m > N$ or $m \leq N$. We refer to the values $\{\{\lambda_{m;m'}\}_{m'=1}^m\}_{m=1}^M$ as a sequence

of *inner eigensteps* since they arise from matrices of inner products of the φ_m's, whereas the outer eigensteps $\{\{\lambda_{m;n}\}_{n=1}^{N}\}_{m=0}^{M}$ arise from sums of outer products of the φ_m's; see Theorem 2.5 below. The following definition makes this precise.

Definition 2.3 Let $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$ be nonnegative nonincreasing sequences. A corresponding sequence of *inner eigensteps* is a sequence $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$ which satisfies the following three properties:

(i) $\lambda_{M;m'} = \lambda_{m'}$ for every $m' = 1, \ldots, M$,
(ii) $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \sqsubseteq \{\lambda_{m;m'}\}_{m'=1}^{m}$ for every $m = 2, \ldots, M$,
(iii) $\sum_{m'=1}^{m} \lambda_{m;m'} = \sum_{m'=1}^{m} \mu_{m'}$ for every $m = 1, \ldots, M$.

To clarify, unlike the outer eigensteps of Definition 2.2, the interlacing relation (ii) here involves two sequences of different length; we write $\{\alpha_{m'}\}_{m'=1}^{m-1} \sqsubseteq \{\beta_{m'}\}_{m'=1}^{m}$ if $\beta_{m'+1} \leq \alpha_{m'} \leq \beta_{m'}$ for all $m' = 1, \ldots, m-1$. As the next example illustrates, inner and outer eigensteps can be put into correspondence with each other.

Example 2.3 We revisit Example 2.2. Here, we pad $\{\lambda_n\}_{n=1}^{3}$ with two zeros so as to match the length of $\{\mu_m\}_{m=1}^{5}$. That is, $\lambda_1 = \lambda_2 = \lambda_3 = \frac{5}{3}$, $\lambda_4 = \lambda_5 = 0$, and $\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5 = 1$. We find every sequence of inner eigensteps $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{5}$, namely every table of the following form:

m	1	2	3	4	5
$\lambda_{m;5}$					0
$\lambda_{m;4}$				?	0
$\lambda_{m;3}$			?	?	$\frac{5}{3}$
$\lambda_{m;2}$		?	?	?	$\frac{5}{3}$
$\lambda_{m;1}$	?	?	?	?	$\frac{5}{3}$

(2.23)

that satisfies the interlacing properties (ii) and trace conditions (iii) of Definition 2.3. To be precise, (ii) gives us $0 = \lambda_{5;5} \leq \lambda_{4;4} \leq \lambda_{5;4} = 0$ and so $\lambda_{4;4} = 0$. Similarly, $\frac{5}{3} \leq \lambda_{5;3} \leq \lambda_{4;2} \leq \lambda_{3;1} \leq \lambda_{4;1} \leq \lambda_{5;1} = \frac{5}{3}$ and so $\lambda_{4;2} = \lambda_{3;1} = \lambda_{4;1} = \frac{5}{3}$, yielding

m	1	2	3	4	5
$\lambda_{m;5}$					0
$\lambda_{m;4}$				0	0
$\lambda_{m;3}$			?	?	$\frac{5}{3}$
$\lambda_{m;2}$		?	?	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	?	?	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

(2.24)

Meanwhile, since $\mu_{m'} = 1$ for all m', the trace conditions (iii) give that the values in the mth column of (2.24) sum to m. Thus, $\lambda_{1;1} = 1$ and $\lambda_{4;3} = \frac{2}{3}$:

m	1	2	3	4	5
$\lambda_{m;5}$					0
$\lambda_{m;4}$				0	0
$\lambda_{m;3}$			?	$\frac{2}{3}$	$\frac{5}{3}$
$\lambda_{m;2}$		?	?	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	1	?	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

Labeling $\lambda_{3;3}$ as x and $\lambda_{2;2}$ as y, (iii) uniquely determines $\lambda_{3;2}$ and $\lambda_{2;1}$:

m	1	2	3	4	5
$\lambda_{m;5}$					0
$\lambda_{m;4}$				0	0
$\lambda_{m;3}$			x	$\frac{2}{3}$	$\frac{5}{3}$
$\lambda_{m;2}$		y	$\frac{4}{3} - x$	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	1	$2 - y$	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

(2.25)

For our particular choice of $\{\lambda_m\}_{m=1}^5$ and $\{\mu_m\}_{m=1}^5$, the preceding argument shows that every corresponding sequence of inner eigensteps is of the form (2.25). Conversely, one may immediately verify that any $\{\{\lambda_{m;m'}\}_{m'=1}^m\}_{m=1}^5$ of this form satisfies (i) and (iii) of Definition 2.3 and moreover satisfies (ii) when $m = 5$. However, in order to satisfy (ii) for $m = 2, 3, 4$, x and y must be chosen to satisfy the ten inequalities:

$$\begin{aligned}
\{\lambda_{3;m'}\}_{m'=1}^3 \sqsubseteq \{\lambda_{4;m'}\}_{m'=1}^4 &\iff 0 \le x \le \frac{2}{3} \le \frac{4}{3} - x \le \frac{5}{3},\\
\{\lambda_{2;m'}\}_{m'=1}^2 \sqsubseteq \{\lambda_{3;m'}\}_{m'=1}^3 &\iff x \le y \le \frac{4}{3} - x \le 2 - y \le \frac{5}{3},\\
\{\lambda_{1;m'}\}_{m'=1}^1 \sqsubseteq \{\lambda_{2;m'}\}_{m'=1}^2 &\iff y \le 1 \le 2 - y.
\end{aligned} \tag{2.26}$$

A quick inspection reveals the system (2.26) to be equivalent to the one derived in the outer eigenstep formulation (2.22) presented in Example 2.2, which is reducible to $0 \le x \le \frac{2}{3}$, $\max\{\frac{1}{3}, x\} \le y \le \min\{\frac{2}{3} + x, \frac{4}{3} - x\}$. Moreover, we see that the outer eigensteps (2.21) that arise from $\{\lambda_1, \lambda_2, \lambda_3\} = \{\frac{5}{3}, \frac{5}{3}, \frac{5}{3}\}$ and the inner eigensteps (2.25) that arise from $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\} = \{\frac{5}{3}, \frac{5}{3}, \frac{5}{3}, 0, 0\}$ are but zero-padded versions of each other. The next result, proven in [18], gives that such a result holds in general.

Theorem 2.5 *Let $\{\lambda_m\}_{m=1}^M$ and $\{\mu_m\}_{m=1}^M$ be nonnegative and nonincreasing, and choose any $N \le M$ such that $\lambda_m = 0$ for every $m > N$. Then every choice of outer eigensteps (Definition 2.2) corresponds to a unique choice of inner eigensteps (Definition 2.3) and vice versa, the two being zero-padded versions of each other.*

Specifically, a sequence of outer eigensteps $\{\{\lambda_{m;n}\}_{n=1}^{N}\}_{m=0}^{M}$ *gives rise to a sequence of inner eigensteps* $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$, *where* $\lambda_{m;m'} := 0$ *whenever* $m' > N$. *Conversely, a sequence of inner eigensteps* $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$ *gives rise to a sequence of outer eigensteps* $\{\{\lambda_{m;n}\}_{n=1}^{N}\}_{m=0}^{M}$, *where* $\lambda_{m;n} := 0$ *whenever* $n > m$.

Moreover, $\{\lambda_{m;n}\}_{n=1}^{N}$ *is the spectrum of the frame operator* $\Phi_m\Phi_m^*$ *if and only if* $\{\lambda_{m;m'}\}_{m'=1}^{m}$ *is the spectrum of the Gram matrix* $\Phi_m^*\Phi_m$.

2.4.1 Top Kill and the Existence of Eigensteps

As discussed earlier in this section, Theorem 2.2 reduces Problem 2.1 to a problem of constructing every possible sequence of outer eigensteps (Definition 2.2). Moreover, by Theorem 2.5, every sequence of outer eigensteps corresponds to a unique sequence of inner eigensteps (Definition 2.3). We now note that if a sequence of inner eigensteps $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$ exists, then $\{\lambda_m\}_{m=1}^{M}$ necessarily majorizes $\{\mu_m\}_{m=1}^{M}$. Indeed, letting $m = M$ in the trace property (iii) of Definition 2.3 immediately gives one of the majorization conditions (2.4); to obtain the remaining condition (2.3) at a given $m = 1, \ldots, M-1$, note that the interlacing property (ii) gives $\lambda_{m;m'} \le \lambda_{M;m'} = \lambda_{m'}$ for all $m' = 1, \ldots, m$, at which point (iii) implies that

$$\sum_{m'=1}^{m} \mu_{m'} = \sum_{m'=1}^{m} \lambda_{m;m'} \le \sum_{m'=1}^{m} \lambda_{m'}.$$

In this section, we prove the converse result, namely that if $\{\lambda_m\}_{m=1}^{M} \succeq \{\mu_m\}_{m=1}^{M}$, then a corresponding sequence of inner eigensteps $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$ exists. The key idea is an algorithm, dubbed *Top Kill*, for transforming any sequence $\{\lambda_{m;m'}\}_{m'=1}^{m}$ that majorizes $\{\mu_{m'}\}_{m'=1}^{m}$ into a new, shorter sequence $\{\lambda_{m;m'}\}_{m'=1}^{m-1}$ that majorizes $\{\mu_{m'}\}_{m'=1}^{m-1}$ and also interlaces with $\{\lambda_{m;m'}\}_{m'=1}^{m}$. In the next section, these new proof techniques lead to a result which shows how to systematically construct every valid sequence of inner eigensteps for a given $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$. We now motivate Top Kill with an example.

Example 2.4 Let $M = 3$, $\{\lambda_1, \lambda_2, \lambda_3\} = \{\frac{7}{4}, \frac{3}{4}, \frac{1}{2}\}$, and $\{\mu_1, \mu_2, \mu_3\} = \{1, 1, 1\}$. Since this spectrum majorizes these lengths, we claim that there exists a corresponding sequence of inner eigensteps $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{3}$. That is, recalling Definition 2.3, we claim that it is possible to find values $\{\lambda_{1;1}\}$ and $\{\lambda_{2;1}, \lambda_{2,2}\}$ which satisfy the interlacing requirements (ii) that $\{\lambda_{1;1}\} \sqsubseteq \{\lambda_{2;1}, \lambda_{2,2}\} \sqsubseteq \{\frac{7}{4}, \frac{3}{4}, \frac{1}{2}\}$ as well as the trace requirements (iii) that $\lambda_{1;1} = 1$ and $\lambda_{2;1} + \lambda_{2;2} = 2$. Indeed, every such sequence of eigensteps is given by the following table:

m	1	2	3
$\lambda_{m;3}$			$\frac{1}{2}$
$\lambda_{m;2}$		x	$\frac{3}{4}$
$\lambda_{m;1}$	1	$2-x$	$\frac{7}{4}$

(2.27)

where x is required to satisfy

$$\frac{1}{2} \leq x \leq \frac{3}{4} \leq 2 - x \leq \frac{7}{4}, \qquad x \leq 1 \leq 2 - x. \tag{2.28}$$

Clearly, any $x \in [\frac{1}{2}, \frac{3}{4}]$ will do. However, when M is large, the table analogous to (2.27) will contain many more variables, leading to a system of inequalities which is much larger and more complicated than (2.28). In such settings, it is not obvious how to construct even a single valid sequence of eigensteps. As such, we consider this same simple example from a different perspective, one that leads to an eigenstep construction algorithm which is easily implementable regardless of the size of M.

The key idea is to view the task of constructing eigensteps as iteratively building a staircase in which the mth level is λ_m units long. For this example in particular, our goal is to build a three-step staircase where the bottom level has length $\frac{7}{4}$, the second level has length $\frac{3}{4}$, and the top level has length $\frac{1}{2}$; the profile of such a staircase is outlined in black in each of the six subfigures of Fig. 2.3. The benefit of visualizing eigensteps in this way is that the interlacing and trace conditions become intuitive staircase-building rules. Specifically, up until the mth step, we will have built a staircase whose levels are of length $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$. To build on top of this staircase, we use m blocks of height 1 whose areas sum to μ_m. Each of these m new blocks is added to its corresponding level of the current staircase, and is required to rest entirely on top of what has been previously built. This requirement corresponds to the interlacing condition (ii) of Definition 2.3, while the trace condition (iii) corresponds to the fact that the areas of these blocks sum to μ_m.

This intuition in mind, we now try to build such a staircase from the ground up. In the first step (Fig. 2.3(a)), we are required to place a single block of area $\mu_1 = 1$ on the first level. The length of this first level is $\lambda_{1;1} = \mu_1$. In the second step, we build up and out from this initial block, placing two new blocks—one on the first level and another on the second—whose total area is $\mu_2 = 1$. The lengths $\lambda_{2;1}$ and $\lambda_{2;2}$ of the new first and second levels depend on how these two blocks are chosen. In particular, choosing first and second level blocks of area $\frac{3}{4}$ and $\frac{1}{4}$, respectively, results in $\{\lambda_{2;1}, \lambda_{2;2}\} = \{\frac{7}{4}, \frac{1}{4}\}$ (Fig. 2.3(b)), which corresponds to a greedy pursuit of the final desired spectrum $\{\frac{7}{4}, \frac{3}{4}, \frac{1}{2}\}$; we fully complete the first level before turning our attention to the second. The problem with this greedy approach is that it doesn't always work, as this example illustrates. Indeed, in the third and final step, we build up and out from the staircase of Fig. 2.3(b) by adding three new blocks—one each for the first, second, and third levels—whose total area is $\mu_3 = 1$. However, in order to maintain interlacing, the new top block must rest entirely on the existing second level, meaning that its length $\lambda_{3;3} \leq \lambda_{2;2} = \frac{1}{4}$ cannot equal the desired value of $\frac{1}{2}$. That is, because of our poor choice in the second step, the "best" we can now do is $\{\lambda_{3;1}, \lambda_{3;2}, \lambda_{3;3}\} = \{\frac{7}{4}, 1, \frac{1}{4}\}$ (Fig. 2.3(c)):

m	1	2	3
$\lambda_{m;3}$			$\frac{1}{4}$
$\lambda_{m;2}$		$\frac{1}{4}$	1
$\lambda_{m;1}$	1	$\frac{7}{4}$	$\frac{7}{4}$

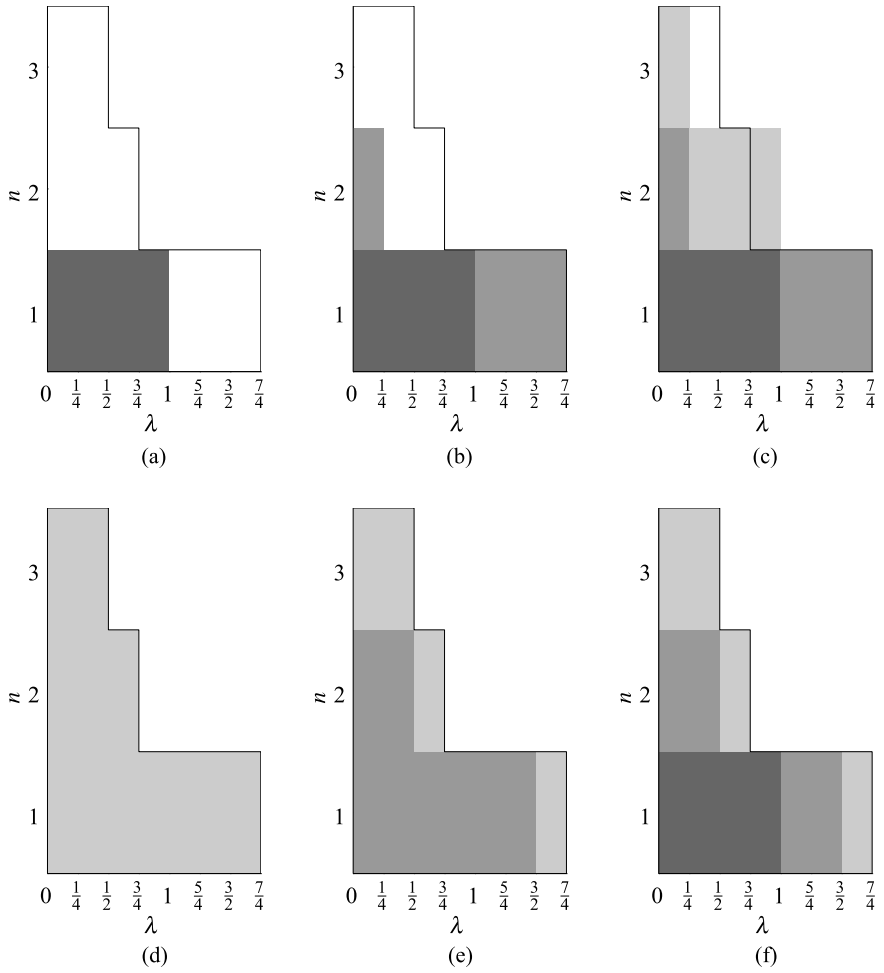


Fig. 2.3 Two attempts at iteratively building a sequence of inner eigensteps for $\{\lambda_1, \lambda_2, \lambda_3\} = \{\frac{7}{4}, \frac{3}{4}, \frac{1}{2}\}$ and $\{\mu_1, \mu_2, \mu_3\} = \{1, 1, 1\}$. As detailed in Example 2.4, the first row represents a failed attempt in which we greedily complete the first level before focusing on those above it. The failure arises from a lack of foresight: the second step does not build a sufficient foundation for the third. The second row represents a second attempt, one that is successful. There, we begin with the final desired staircase and work backward. That is, we chip away at the three-level staircase (**d**) to produce a two-level one (**e**), and then chip away at it to produce a one-level one (**f**). In each step, we remove as much as possible from the top level before turning our attention to the lower levels, subject to the interlacing constraints. We refer to this algorithm for iteratively producing $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$ from $\{\lambda_{m;m'}\}_{m'=1}^{m}$ as Top Kill. Theorem 2.6 shows that Top Kill will always produce a valid sequence of eigensteps from any desired spectrum $\{\lambda_m\}_{m=1}^{M}$ that majorizes a given desired sequence of lengths $\{\mu_m\}_{m=1}^{M}$

This greedy approach fails because it doesn't plan ahead. Indeed, it treats the bottom levels of the staircase as the priority when, in fact, the opposite is true: the top levels are the priority, since they require the most foresight. In particular, for $\lambda_{3;3}$ to achieve its desired value of $\frac{1}{2}$ in the third step, one must lay a suitable foundation in which $\lambda_{2;2} \geq \frac{1}{2}$ in the second step.

In light of this realization, we make another attempt at building our staircase. This time we begin with the final desired spectrum $\{\lambda_{3;1}, \lambda_{3;2}, \lambda_{3;3}\} = \{\frac{7}{4}, \frac{3}{4}, \frac{1}{2}\}$ (Fig. 2.3(d)) and work backward. From this perspective, our task is now to remove three blocks—the entirety of the top level, and portions of the first and second levels—whose total area is $\mu_3 = 1$. Here, the interlacing requirement translates to only being permitted to remove portions of the staircase that were already exposed to the surface at the end of the previous step. After lopping off the top level, which has area $\lambda_{3;3} = \frac{1}{2}$, we need to decide how to chip away $\mu_1 - \lambda_{3;3} = 1 - \frac{1}{2} = \frac{1}{2}$ units of area from the first and second levels, subject to this constraint. At this point, we observe that, in the step that follows, our first task will be to remove the remaining portion of the second level. As such, it is to our advantage to remove as much of the second level as possible in the current step, and only then to turn our attention to the lower levels. That is, we follow Thomas Jefferson's adage, "Never put off until tomorrow what you can do today." We call this approach Top Kill, since it "kills" off as much as possible from the top portions of the staircase. For this example in particular, interlacing implies that we can at most remove a block of area $\frac{1}{4}$ from the second level, leaving $\frac{1}{4}$ units of area to be removed from the first; the resulting two-level staircase—the darker shade in Fig. 2.3(e)—has levels of lengths $\{\lambda_{2;1}, \lambda_{2;2}\} = \{\frac{3}{2}, \frac{1}{2}\}$. In the second step, we then apply this same philosophy, removing the entire second level and a block of area $\mu_2 - \lambda_{2;2} = 1 - \frac{1}{2} = \frac{1}{2}$ from the first, resulting in the one-level staircase (Fig. 2.3(f)) in which $\{\lambda_{1;1}\} = 1$. That is, by working backward we have produced a valid sequence of eigensteps:

m	1	2	3
$\lambda_{m;3}$			$\frac{1}{4}$
$\lambda_{m;2}$		$\frac{1}{2}$	1
$\lambda_{m;1}$	1	$\frac{3}{2}$	$\frac{7}{4}$

The preceding example illustrated a systematic "Top Kill" approach for building eigensteps; we now express these ideas more rigorously. As can be seen in the bottom row of Fig. 2.3, Top Kill generally picks $\lambda_{m-1;m'} := \lambda_{m;m'+1}$ for the larger m''s. Top Kill also picks $\lambda_{m-1;m'} := \lambda_{m;m'}$ for the smaller m''s. The level that separates the larger m''s from the smaller m''s is the lowest level from which a nontrivial area is removed. For this level, say level k, we have $\lambda_{m;k+1} < \mu_m \leq \lambda_{m;k}$. In the levels above k, we have already removed a total of $\lambda_{m;k+1}$ units of area, leaving $\mu_m - \lambda_{m;k+1}$ to be chipped away from $\lambda_{m;k}$, yielding $\lambda_{m-1;k} := \lambda_{m;k} - (\mu_m - \lambda_{m;k+1})$. The next result confirms that Top Kill always produces eigensteps whenever it is possible to do so.

Theorem 2.6 *Suppose* $\{\lambda_{m;m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$, *and define* $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$ *according to Top Kill*; *that is*, *pick any* k *such that* $\lambda_{m;k+1} \le \mu_m \le \lambda_{m;k}$, *and for each* $m' = 1, \ldots, m-1$, *define*:

$$\lambda_{m-1;m'} := \begin{cases} \lambda_{m;m'}, & 1 \le m' \le k-1, \\ \lambda_{m;k} + \lambda_{m;k+1} - \mu_m, & m' = k, \\ \lambda_{m;m'+1}, & k+1 \le m' \le m-1. \end{cases} \tag{2.29}$$

Then $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \sqsubseteq \{\lambda_{m;m'}\}_{m'=1}^{m}$ *and* $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$.

Furthermore, *given nonnegative nonincreasing sequences* $\{\lambda_m\}_{m=1}^{M}$ *and* $\{\mu_m\}_{m=1}^{M}$ *such that* $\{\lambda_m\}_{m=1}^{M} \succeq \{\mu_m\}_{m=1}^{M}$, *define* $\lambda_{M;m'} := \lambda_{m'}$ *for every* $m' = 1, \ldots, M$, *and for each* $m = M, \ldots, 2$, *consecutively define* $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$ *according to Top Kill*. *Then* $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$ *is a valid sequence of inner eigensteps.*

Proof For notational simplicity, we denote $\{\alpha_{m'}\}_{m'=1}^{m-1} := \{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$ and $\{\beta_{m'}\}_{m'=1}^{m} := \{\lambda_{m;m'}\}_{m'=1}^{m}$. Since $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$, we necessarily have that $\beta_m \le \mu_m \le \mu_1 \le \beta_1$, and so there exists $k = 1, \ldots, m-1$ such that $\beta_{k+1} \le \mu_m \le \beta_k$. Though this k may not be unique when subsequent $\beta_{m'}$'s are equal, a quick inspection reveals that any appropriate choice of k will yield the same $\alpha_{m'}$'s, and so Top Kill is well defined. To prove $\{\alpha_{m'}\}_{m'=1}^{m-1} \sqsubseteq \{\beta_{m'}\}_{m'=1}^{m}$, we need to show that

$$\beta_{m'+1} \le \alpha_{m'} \le \beta_{m'} \tag{2.30}$$

for every $m' = 1, \ldots, m-1$. If $1 \le m' \le k-1$, then $\alpha_{m'} := \beta_{m'}$, and so the right-hand inequality of (2.30) holds with equality, at which point the left-hand inequality is immediate. Similarly, if $k+1 \le m' \le m-1$, then $\alpha_{m'} := \beta_{m'+1}$, and so (2.30) holds with equality on the left-hand side. Lastly if $m' = k$, then $\alpha_k := \beta_k + \beta_{k+1} - \mu_m$, and our assumption that $\beta_{k+1} \le \mu_m \le \beta_k$ gives (2.30) in this case:

$$\beta_{k+1} \le \beta_k + \beta_{k+1} - \mu_m \le \beta_k.$$

Thus, $\{\alpha_{m'}\}_{m'=1}^{m-1} \sqsubseteq \{\beta_{m'}\}_{m'=1}^{m}$, as claimed. We next show that $\{\alpha_{m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$. If $j \le k-1$, then since $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$, we have

$$\sum_{m'=1}^{j} \alpha_{m'} = \sum_{m'=1}^{j} \beta_{m'} \ge \sum_{m'=1}^{j} \mu_{m'},$$

as needed. On the other hand, if $j \ge k$, we have

$$\sum_{m'=1}^{j} \alpha_{m'} = \sum_{m'=1}^{k-1} \beta_{m'} + (\beta_k + \beta_{k+1} - \mu_m) + \sum_{m'=k+1}^{j} \beta_{m'+1} = \sum_{m'=1}^{j+1} \beta_{m'} - \mu_m, \tag{2.31}$$

with the understanding that a sum over an empty set of indices is zero. We continue (2.31) by using the facts that $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$ and $\mu_{j+1} \geq \mu_m$:

$$\sum_{m'=1}^{j} \alpha_{m'} = \sum_{m'=1}^{j+1} \beta_{m'} - \mu_m \geq \sum_{m'=1}^{j+1} \mu_{m'} - \mu_m \geq \sum_{m'=1}^{j} \mu_{m'}. \tag{2.32}$$

Note that when $j = m$, the inequalities in (2.32) become equalities, giving the final trace condition.

For the final conclusion, note that one application of Top Kill transforms a sequence $\{\lambda_{m;m'}\}_{m'=1}^{m}$ that majorizes $\{\mu_{m'}\}_{m'=1}^{m}$ into a shorter sequence $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$ that interlaces with $\{\lambda_{m;m'}\}_{m'=1}^{m}$ and majorizes $\{\mu_{m'}\}_{m'=1}^{m-1}$. As such, one may indeed start with $\lambda_{M;m'} := \lambda_{m'}$ and apply Top Kill $M-1$ times to produce a sequence $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$ that immediately satisfies Definition 2.3. □

2.4.2 *Parametrizing Inner Eigensteps*

In the previous subsection, we discussed Top Kill, an algorithm designed to construct a sequence of inner eigensteps from given nonnegative nonincreasing sequences $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$. In this subsection, we use the intuition underlying Top Kill to find a systematic method for producing all such eigensteps. To be precise, treating the values $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M-1}$ as independent variables, it is not difficult to show that the set of all inner eigensteps for a given $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$ form a convex polytope in $\mathbb{R}^{M(M-1)/2}$. Our goal is to find a useful, implementable parametrization of this polytope.

We begin by noting that this polytope is nonempty precisely when $\{\lambda_m\}_{m=1}^{M}$ majorizes $\{\mu_m\}_{m=1}^{M}$. Indeed, as noted at the beginning of the previous section, if such a sequence of eigensteps exists, then we necessarily have that $\{\lambda_m\}_{m=1}^{M} \succeq \{\mu_m\}_{m=1}^{M}$. Conversely, if $\{\lambda_m\}_{m=1}^{M} \succeq \{\mu_m\}_{m=1}^{M}$, then Theorem 2.6 states that Top Kill will produce a valid sequence of eigensteps from $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$. Note that this implies that, for a given $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$, if any given strategy for building eigensteps is successful, then Top Kill will also succeed. In this sense, Top Kill is an optimal strategy. However, Top Kill alone will not suffice to parametrize our polytope, since for a given feasible $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$, it only produces a single sequence of eigensteps when, in fact, there are in general infinitely many such sequences. In the work that follows, we view these non-Top-Kill-produced eigensteps as the result of applying suboptimal generalizations of Top Kill to $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$.

For example, if $\{\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5\} = \{\frac{5}{3}, \frac{5}{3}, \frac{5}{3}, 0, 0\}$ and $\mu_m = 1$ for all $m = 1, \ldots, 5$, every sequence of inner eigensteps corresponds to a choice of the unknown values in (2.23) which satisfies the interlacing and trace conditions (ii) and (iii) of Definition 2.3. There are 10 unknowns in (2.23), and the set of all such eigensteps is

a convex polytope in $\mathbb{R}^{10}$. Although this dimension can be reduced by exploiting the interlacing and trace conditions—the 10 unknowns in (2.23) can be reduced to the two unknowns in (2.25)—this approach to constructing all eigensteps nevertheless requires one to simplify large systems of coupled inequalities, such as (2.26).

We suggest a different method for parametrizing this polytope: to systematically pick the values $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{4}$ one at a time. Top Kill is one way to do this: working from the top levels down, we chip away $\mu_5 = 1$ units of area from $\{\lambda_{5;m'}\}_{m'=1}^{5}$ to successively produce $\lambda_{4;4} = 0$, $\lambda_{4;3} = \frac{2}{3}$, $\lambda_{4;2} = \frac{5}{3}$, and $\lambda_{4;1} = \frac{5}{3}$. We then repeat this process to transform $\{\lambda_{4;m'}\}_{m'=1}^{4}$ into $\{\lambda_{3;m'}\}_{m'=1}^{3}$, and so on; the specific values can be obtained by letting $(x, y) = (0, \frac{1}{3})$ in (2.25). We seek to generalize Top Kill to find *all* ways of picking the $\lambda_{m;m'}$'s one at a time. As in Top Kill, we work backward: we first find all possibilities for $\lambda_{4;4}$, then the possibilities for $\lambda_{4;3}$ in terms of our choice of $\lambda_{4;4}$, then the possibilities for $\lambda_{4;2}$ in terms of our choices of $\lambda_{4;4}$ and $\lambda_{4;3}$, and so on. That is, we iteratively parametrize our polytope in the following order:

$$\lambda_{4;4}, \quad \lambda_{4;3}, \quad \lambda_{4;2}, \quad \lambda_{4;1}, \quad \lambda_{3;3}, \quad \lambda_{3;2}, \quad \lambda_{3;1}, \quad \lambda_{2;2}, \quad \lambda_{2;1}, \quad \lambda_{1;1}.$$

More generally, for any $\{\lambda_m\}_{m=1}^{M}$ and $\{\mu_m\}_{m=1}^{M}$ such that $\{\lambda_m\}_{m=1}^{M} \succeq \{\mu_m\}_{m=1}^{M}$ we construct every possible sequence of eigensteps $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$ by finding all possibilities for any given $\lambda_{m-1;k}$ in terms of $\lambda_{m'';m'}$, where either $m'' > m-1$ or $m'' = m-1$ and $m' > k$. Certainly, any permissible choice for $\lambda_{m-1;k}$ must satisfy the interlacing criteria (ii) of Definition 2.3, and so we have bounds $\lambda_{m;k+1} \leq \lambda_{m-1;k} \leq \lambda_{m;k}$. Other necessary bounds arise from the majorization conditions. Indeed, in order to have both $\{\lambda_{m;m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$ and $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$ we need

$$\mu_m = \sum_{m'=1}^{m} \mu_{m'} - \sum_{m'=1}^{m-1} \mu_{m'} = \sum_{m'=1}^{m} \lambda_{m;m'} - \sum_{m'=1}^{m-1} \lambda_{m-1;m'}, \tag{2.33}$$

and so we may view μ_m as the total change between the eigenstep spectra. Having already selected $\lambda_{m-1;n-1}, \ldots, \lambda_{m-1;k+1}$, we've already imposed a certain amount of change between the spectra, and so we are limited in how much we can change the kth eigenvalue. Continuing (2.33), this fact can be expressed as

$$\mu_m = \lambda_{m;m} + \sum_{m'=1}^{m-1} (\lambda_{m;m'} - \lambda_{m-1;m'}) \geq \lambda_{m;m} + \sum_{m'=k}^{m-1} (\lambda_{m;m'} - \lambda_{m-1;m'}), \tag{2.34}$$

where the inequality follows from the fact that the summands $\lambda_{m;m'} - \lambda_{m-1;m'}$ are nonnegative if $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$ is to be chosen so that $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \sqsubseteq \{\lambda_{m;m'}\}_{m'=1}^{m}$. Rearranging (2.34) then gives a second lower bound on $\lambda_{m-1;k}$ to go along with our

previously mentioned requirement that $\lambda_{m-1;k} \geq \lambda_{m;k+1}$:

$$\lambda_{m-1;k} \geq \sum_{m'=k}^{m} \lambda_{m;m'} - \sum_{m'=k+1}^{m-1} \lambda_{m-1;m'} - \mu_m. \tag{2.35}$$

We next apply the intuition behind Top Kill to obtain other upper bounds on $\lambda_{m-1;k}$ to go along with our previously mentioned requirement that $\lambda_{m-1;k} \leq \lambda_{m;k}$. We caution that what follows is not a rigorous argument for the remaining upper bound on $\lambda_{m-1;k}$, but rather an informal derivation of this bound's expression; the legitimacy of this derivation is formally confirmed in the proof of Theorem 2.7. Recall that, at this point in the narrative, we have already selected $\{\lambda_{m-1;m'}\}_{m'=k+1}^{m-1}$ and are attempting to find all possible choices $\lambda_{m-1;k}$ that will allow the remaining values $\{\lambda_{m-1;m'}\}_{m'=1}^{k-1}$ to be chosen in such a way that:

$$\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \sqsubseteq \{\lambda_{m;m'}\}_{m'=1}^{m}, \qquad \{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}. \tag{2.36}$$

To do this, we recall our staircase-building intuition from the previous section: if it is possible to build a given staircase, then one way to do this is to assign maximal priority to the highest levels, as these are the most difficult to build. As such, for a given choice of $\lambda_{m-1;k}$, if it is possible to choose $\{\lambda_{m-1;m'}\}_{m'=1}^{k-1}$ in such a way that (2.36) holds, then it is reasonable to expect that one way of doing this is to pick $\lambda_{m-1;k-1}$ by chipping away as much as possible from $\lambda_{m;k-1}$, then pick $\lambda_{m-1;k-2}$ by chipping away as much as possible from $\lambda_{m;k-2}$, and so on. That is, we pick some arbitrary value $\lambda_{m-1;k}$, and to test its legitimacy, we apply the Top Kill algorithm to construct the remaining undetermined values $\{\lambda_{m-1;m'}\}_{m'=1}^{k-1}$; we then check whether or not $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$.

To be precise, note that prior to applying Top Kill, the remaining spectrum is $\{\lambda_{m;m'}\}_{m'=1}^{k-1}$, and that the total amount we will chip away from this spectrum is

$$\mu_m - \left(\lambda_{m;n} + \sum_{m'=k}^{m-1} (\lambda_{m;m'} - \lambda_{m-1;m'}) \right). \tag{2.37}$$

To ensure that our choice of $\lambda_{m-1;k-1}$ satisfies $\lambda_{m-1;k-1} \geq \lambda_{m;k}$, we artificially reintroduce $\lambda_{m;k}$ to both (2.37) and the remaining spectrum $\{\lambda_{m;m'}\}_{m'=1}^{k-1}$ before applying Top Kill. That is, we apply Top Kill to $\{\beta_{m'}\}_{m'=1}^{m} := \{\lambda_{m;m'}\}_{m'=1}^{k} \cup \{0\}_{m'=k+1}^{m}$. Specifically in light of Theorem 2.6, in order to optimally subtract

$$\begin{aligned} \mu &:= \mu_m - \left(\lambda_{m;n} + \sum_{m'=k}^{m-1} (\lambda_{m;m'} - \lambda_{m-1;m'}) \right) + \lambda_{m;k} \\ &= \mu_m - \sum_{m'=k+1}^{m} \lambda_{m;m'} + \sum_{m'=k}^{m-1} \lambda_{m-1;m'} \end{aligned}$$

units of area from $\{\beta_{m'}\}_{m'=1}^{m}$, we first pick j such that $\beta_{j+1} \le \mu \le \beta_j$. We then use (2.29) to produce a zero-padded version of the remaining new spectrum $\{\lambda_{m-1;m'}\}_{m'=1}^{k-1} \cup \{0\}_{m'=k}^{m}$:

$$\lambda_{m-1;m'} = \begin{cases} \lambda_{m;m'}, & 1 \le m' \le j-1, \\ \lambda_{m;j} + \lambda_{m;j+1} - \mu_m + \sum_{m''=k+1}^{m} \lambda_{m;m''} - \sum_{m''=k}^{m-1} \lambda_{m-1;m''}, & m' = j \\ \lambda_{m;m'+1}, & j+1 \le m' \le k-1. \end{cases}$$

Picking l such that $j+1 \le l \le k$, we now sum the above values of $\lambda_{m-1;m'}$ to obtain:

$$\begin{aligned} \sum_{m'=1}^{l-1} \lambda_{m-1;m'} &= \sum_{m'=1}^{j-1} \lambda_{m-1;m'} + \lambda_{m-1;j} + \sum_{m'=j+1}^{l-1} \lambda_{m-1;m'} \\ &= \sum_{m'=1}^{l} \lambda_{m;m'} - \mu_m + \sum_{m'=k+1}^{m} \lambda_{m;m'} - \sum_{m'=k}^{m-1} \lambda_{m-1;m'}. \end{aligned} \tag{2.38}$$

Adding $\sum_{m'=1}^{m} \mu_{m'} - \sum_{m'=1}^{m} \lambda_{m;m'} = 0$ to the right-hand side of (2.38) then yields:

$$\begin{aligned} \sum_{m'=1}^{l-1} \lambda_{m-1;m'} &= \sum_{m'=1}^{l} \lambda_{m;m'} - \mu_m + \sum_{m'=k+1}^{m} \lambda_{m;m'} - \sum_{m'=k}^{m-1} \lambda_{m-1;m'} + \sum_{m'=1}^{m} \mu_{m'} \\ &\quad - \sum_{m'=1}^{m} \lambda_{m;m'} = \sum_{m'=1}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{k} \lambda_{m;m'} - \sum_{m'=k}^{m-1} \lambda_{m-1;m'}. \end{aligned} \tag{2.39}$$

Now, in order for $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$ as desired, (2.39) must satisfy:

$$\sum_{m'=1}^{l-1} \mu_{m'} \le \sum_{m'=1}^{l-1} \lambda_{m-1;m'} = \sum_{m'=1}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{k} \lambda_{m;m'} - \sum_{m'=k}^{m-1} \lambda_{m-1;m'}. \tag{2.40}$$

Solving for $\lambda_{m-1;k}$ in (2.40) then gives:

$$\lambda_{m-1;k} \le \sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{k} \lambda_{m;m'} - \sum_{m'=k+1}^{m-1} \lambda_{m-1;m'}. \tag{2.41}$$

Note that, according to how we derived it, (2.41) is valid when $j+1 \le l \le k$. As established in the following theorem, this bound actually holds when $l = 1, \ldots, k$. Overall, the interlacing conditions, (2.35), and (2.41) are precisely the bounds that we verify in the following result.

Theorem 2.7 *Suppose* $\{\lambda_{m;m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$. *Then* $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$ *and* $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1} \sqsubseteq \{\lambda_{m;m'}\}_{m'=1}^{m}$ *if and only if* $\lambda_{m-1;k} \in [A_{m-1;k}, B_{m-1;k}]$ *for ev-*

ery $k = 1, \ldots, m-1$, *where*:

$$A_{m-1;k} := \max\left\{\lambda_{m;k+1}, \sum_{m'=k}^{m} \lambda_{m;m'} - \sum_{m'=k+1}^{m-1} \lambda_{m-1;m'} - \mu_m\right\}, \tag{2.42}$$

$$B_{m-1;k} := \min\left\{\lambda_{m;k}, \min_{l=1,\ldots,k}\left\{\sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{k} \lambda_{m;m'} - \sum_{m'=k+1}^{m-1} \lambda_{m-1;m'}\right\}\right\}. \tag{2.43}$$

Here, we use the convention that sums over empty sets of indices are zero. Moreover, suppose that $\lambda_{m-1;m-1}, \ldots, \lambda_{m-1;k+1}$ *are consecutively chosen to satisfy these bounds. Then* $A_{m-1;k} \leq B_{m-1;k}$, *and so* $\lambda_{m-1;k}$ *can also be chosen from such an interval.*

Proof For notational simplicity, we let $\{\alpha_{m'}\}_{m'=1}^{m-1} := \{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$, $\{\beta_{m'}\}_{m'=1}^{m} := \{\lambda_{m;m'}\}_{m'=1}^{m}$, $A_k := A_{m-1;k}$, and $B_k := B_{m-1;k}$.

($\Rightarrow$) Suppose $\{\alpha_{m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$ and $\{\alpha_{m'}\}_{m'=1}^{m-1} \sqsubseteq \{\beta_{m'}\}_{m'=1}^{m}$. Fix any particular $k = 1, \ldots, m-1$. Note that interlacing gives $\beta_{k+1} \leq \alpha_k \leq \beta_k$, which accounts for the first entries in (2.42) and (2.43). We first show that $\alpha_k \geq A_k$. Since $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$ and $\{\alpha_{m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$, then

$$\mu_m = \sum_{m'=1}^{m} \mu_{m'} - \sum_{m'=1}^{m-1} \mu_{m'} = \sum_{m'=1}^{m} \beta_{m'} - \sum_{m'=1}^{m-1} \alpha_{m'} = \beta_m + \sum_{m'=1}^{m-1} (\beta_{m'} - \alpha_{m'}). \tag{2.44}$$

Since $\{\alpha_{m'}\}_{m'=1}^{m-1} \sqsubseteq \{\beta_{m'}\}_{m'=1}^{m}$, the summands in (2.44) are nonnegative, and so

$$\mu_m \geq \beta_m + \sum_{m'=k}^{m-1} (\beta_{m'} - \alpha_{m'}) = \sum_{m'=k}^{m} \beta_{m'} - \sum_{m'=k+1}^{m-1} \alpha_{m'} - \alpha_k. \tag{2.45}$$

Isolating α_k in (2.45) and combining with the fact that $\alpha_k \geq \beta_{k+1}$ gives $\alpha_k \geq A_k$. We next show that $\alpha_k \leq B_k$. Fix $l = 1, \ldots, k$. Then $\{\alpha_{m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$ implies $\sum_{m'=1}^{l-1} \alpha_{m'} \geq \sum_{m'=1}^{l-1} \mu_{m'}$ and $\sum_{m'=1}^{m-1} \alpha_{m'} = \sum_{m'=1}^{m-1} \mu_{m'}$, and so subtracting gives

$$\sum_{m'=l}^{m-1} \mu_{m'} \geq \sum_{m'=l}^{m-1} \alpha_{m'} = \sum_{m'=k}^{m-1} \alpha_{m'} + \sum_{m'=l}^{k-1} \alpha_{m'} \geq \sum_{m'=k}^{m-1} \alpha_{m'} + \sum_{m'=l}^{k-1} \beta_{m'+1}, \tag{2.46}$$

where the second inequality follows from $\{\alpha_{m'}\}_{m'=1}^{m-1} \sqsubseteq \{\beta_{m'}\}_{m'=1}^{m}$. Since our choice for $l = 1, \ldots, k$ was arbitrary, isolating α_k in (2.46) and combining with the fact that $\alpha_k \leq \beta_k$ gives $\alpha_k \leq B_k$.

($\Leftarrow$) Now suppose $A_k \leq \alpha_k \leq B_k$ for every $k = 1, \ldots, m-1$. Then the first entries in (2.42) and (2.43) give $\beta_{k+1} \leq \alpha_k \leq \beta_k$ for every $k = 1, \ldots, m-1$, that is,

$\{\alpha_{m'}\}_{m'=1}^{m-1} \sqsubseteq \{\beta_{m'}\}_{m'=1}^{m}$. It remains to be shown that $\{\alpha_{m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$. Since $\alpha_k \le B_k$ for every $k = 1, \dots, m-1$, then

$$\alpha_k \le \sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{k} \beta_{m'} - \sum_{m'=k+1}^{m-1} \alpha_{m'}, \quad \forall k = 1, \dots, m-1,\ l = 1, \dots, k. \tag{2.47}$$

Rearranging (2.47) in the case where $l = k$ gives

$$\sum_{m'=k}^{m-1} \alpha_{m'} \le \sum_{m'=k}^{m-1} \mu_{m'}, \quad \forall k = 1, \dots, m-1. \tag{2.48}$$

Moreover, $\alpha_1 \ge A_1$ implies $\alpha_1 \ge \sum_{m'=1}^{m} \beta_{m'} - \sum_{m'=2}^{m-1} \alpha_{m'} - \mu_m$. Rearranging this inequality and applying $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$ then gives

$$\sum_{m'=1}^{m-1} \alpha_{m'} \ge \sum_{m'=1}^{m} \beta_{m'} - \mu_m = \sum_{m'=1}^{m-1} \mu_{m'}. \tag{2.49}$$

Combining (2.49) with (2.48) in the case where $k = 1$ gives

$$\sum_{m'=1}^{m-1} \alpha_{m'} = \sum_{m'=1}^{m-1} \mu_{m'}. \tag{2.50}$$

Subtracting (2.48) from (2.50) completes the proof that $\{\alpha_{m'}\}_{m'=1}^{m-1} \succeq \{\mu_{m'}\}_{m'=1}^{m-1}$.

For the final claim, we first show that the claim holds for $k = m-1$, namely that $A_{m-1} \le B_{m-1}$. Explicitly, we need to show that

$$\max\{\beta_m, \beta_{m-1} + \beta_m - \mu_m\} \le \min\left\{\beta_{m-1}, \min_{l=1,\dots,m-1}\left\{\sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{m-1} \beta_{m'}\right\}\right\}. \tag{2.51}$$

Note that (2.51) is equivalent to the following inequalities holding simultaneously:

(i) $\beta_m \le \beta_{m-1}$,
(ii) $\beta_{m-1} + \beta_m - \mu_m \le \beta_{m-1}$,
(iii) $\beta_m \le \sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{m-1} \beta_{m'},\ \forall l = 1, \dots, m-1$,
(iv) $\beta_{m-1} + \beta_m - \mu_m \le \sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{m-1} \beta_{m'}, \forall l = 1, \dots, m-1$.

First, (i) follows immediately from the fact that $\{\beta_{m'}\}_{m'=1}^{m}$ is nonincreasing. Next, rearranging (ii) gives $\beta_m \le \mu_m$, which follows from $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$. For (iii), the facts that $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$ and $\{\mu_{m'}\}_{m'=1}^{m}$ is nonincreasing imply that

$$\sum_{m'=l+1}^{m} \beta_{m'} \le \sum_{m'=l+1}^{m} \mu_{m'} \le \sum_{m'=l}^{m-1} \mu_{m'}, \quad \forall l = 1, \dots, m-1,$$

which in turn implies (iii). Also for (iv), the facts that $\{\beta_{m'}\}_{m'=1}^{m}$ is nonincreasing and $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$ imply that

$$\beta_{m-1} + \sum_{m'=l+1}^{m} \beta_{m'} \leq \sum_{m'=l}^{m} \beta_{m'} \leq \sum_{m'=l}^{m} \mu_{m'}, \quad \forall l = 1, \ldots, m-1,$$

which in turn implies (iv). We now proceed by induction. Assume α_{k+1} satisfies $A_{k+1} \leq \alpha_{k+1} \leq B_{k+1}$. Given this assumption, we need to show that $A_k \leq B_k$. Considering the definitions (2.42) and (2.43) of A_k and B_k, this is equivalent to the following inequalities holding simultaneously:

(i) $\beta_{k+1} \leq \beta_k$,
(ii) $\sum_{m'=k}^{m} \beta_{m'} - \sum_{m'=k+1}^{m-1} \alpha_{m'} - \mu_m \leq \beta_k$,
(iii) $\beta_{k+1} \leq \sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{k} \beta_{m'} - \sum_{m'=k+1}^{m-1} \alpha_{m'}$, $\forall l = 1, \ldots, k$,
(iv) $\sum_{m'=k}^{m} \beta_{m'} - \sum_{m'=k+1}^{m-1} \alpha_{m'} - \mu_m \leq \sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{k} \beta_{m'} - \sum_{m'=k+1}^{m-1} \alpha_{m'}$, $\forall l = 1, \ldots, k$.

Again, the fact that $\{\beta_{m'}\}_{m'=1}^{m}$ is nonincreasing implies (i). Next, $\alpha_{k+1} \geq A_{k+1}$ gives

$$\alpha_{k+1} \geq \sum_{m'=k+1}^{m} \beta_{m'} - \sum_{m'=k+2}^{m-1} \alpha_{m'} - \mu_m,$$

which is a rearrangement of (ii). Similarly, $\alpha_{k+1} \leq B_{k+1}$ gives

$$\alpha_{k+1} \leq \sum_{m'=l}^{m-1} \mu_{m'} - \sum_{m'=l+1}^{k+1} \beta_{m'} - \sum_{m'=k+2}^{m-1} \alpha_{m'}, \quad \forall l = 1, \ldots, k+1,$$

which is a rearrangement of (iii). Note that we don't use the fact that (iii) holds when $l = k+1$. Finally, (iv) follows from the facts that $\{\beta_{m'}\}_{m'=1}^{m}$ is nonincreasing and $\{\beta_{m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$, since they imply that

$$\beta_k + \sum_{m'=l+1}^{m} \beta_{m'} \leq \sum_{m'=l}^{m} \beta_{m'} \leq \sum_{m'=l}^{m} \mu_{m'}, \quad \forall l = 1, \ldots, k,$$

which is a rearrangement of (iv). □

We now note that, by starting with a sequence $\{\lambda_{M;m'}\}_{m'=1}^{M} = \{\lambda_{m'}\}_{m'=1}^{M}$ that majorizes a given $\{\mu_m\}_{m=1}^{M}$, repeatedly applying Theorem 2.7 to construct $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$ from $\{\lambda_{m;m'}\}_{m'=1}^{m}$ results in a sequence of inner eigensteps (Definition 2.3). Conversely, if $\{\{\lambda_{m;m'}\}_{m'=1}^{m}\}_{m=1}^{M}$ is a valid sequence of inner eigensteps, then for every m, (ii) gives $\{\lambda_{m;m'}\}_{m'=1}^{m-1} \sqsubseteq \{\lambda_{m;m'}\}_{m'=1}^{m}$, while (ii) and (iii) together imply that $\{\lambda_{m;m'}\}_{m'=1}^{m} \succeq \{\mu_{m'}\}_{m'=1}^{m}$ à la the discussion at the beginning of Sect. 2.3. As such, any sequence of inner eigensteps can be constructed by repeatedly applying Theorem 2.7. We now summarize these facts.

Corollary 2.1 *Let* $\{\lambda_m\}_{m=1}^M$ *and* $\{\mu_m\}_{m=1}^M$ *be nonnegative, nonincreasing sequences where* $\{\lambda_m\}_{m=1}^M \succeq \{\mu_m\}_{m=1}^M$. *Then, every corresponding sequence of inner eigensteps* $\{\{\lambda_{m;m'}\}_{m'=1}^m\}_{m=1}^M$ *can be constructed by the following algorithm: let* $\lambda_{M;m'} = \lambda_{m'}$ *for all* $m' = 1, \ldots, M$; *for any* $m = M, \ldots, 2$ *construct* $\{\lambda_{m-1;m'}\}_{m'=1}^{m-1}$ *from* $\{\lambda_{m;m'}\}_{m'=1}^m$ *by picking* $\lambda_{m-1;k} \in [A_{m-1;k}, B_{m-1;k}]$ *for all* $k = m-1, \ldots, 1$, *where* $A_{m-1;k}$ *and* $B_{m-1;k}$ *are* (2.42) *and* (2.43), *respectively. Moreover, any sequence constructed by this algorithm is indeed a corresponding sequence of inner eigensteps.*

We now redo Example 2.3 to illustrate that Corollary 2.1 indeed gives a more systematic way of parametrizing the eigensteps.

Example 2.5 We wish to parametrize the eigensteps corresponding to the UNTFs of 5 vectors in $\mathbb{C}^3$. In the end, we will get the same parametrization of eigensteps as in Example 2.3:

m	1	2	3	4	5
$\lambda_{m;5}$					0
$\lambda_{m;4}$				0	0
$\lambda_{m;3}$			x	$\frac{2}{3}$	$\frac{5}{3}$
$\lambda_{m;2}$		y	$\frac{4}{3} - x$	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	1	$2 - y$	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

(2.52)

where $0 \le x \le \frac{2}{3}$, $\max\{\frac{1}{3}, x\} \le y \le \min\{\frac{2}{3} + x, \frac{4}{3} - x\}$. In what follows, we rederive the above table one column at a time, in order from right to left, and fill in each column from top to bottom. First, the desired spectrum of the final Gram matrix gives us that $\lambda_{5;5} = \lambda_{5;4} = 0$ and $\lambda_{5;3} = \lambda_{5;2} = \lambda_{5;1} = \frac{5}{3}$. Next, we wish to find all $\{\lambda_{4;m'}\}_{m'=1}^4$ such that $\{\lambda_{4;m'}\}_{m'=1}^4 \sqsubseteq \{\lambda_{5;m'}\}_{m'=1}^5$ and $\{\lambda_{4;m'}\}_{m'=1}^4 \succeq \{\mu_{m'}\}_{m'=1}^4$. To this end, taking $m = 5$ and $k = 4$, Theorem 2.7 gives:

$$\begin{aligned}
&\max\{\lambda_{5;5}, \lambda_{5;4} + \lambda_{5;5} - \mu_5\} \le \lambda_{4;4} \\
&\quad \le \min\left\{\lambda_{5;4}, \min_{l=1,\ldots,4}\left\{\sum_{m'=l}^4 \mu_{m'} - \sum_{m'=l+1}^4 \lambda_{5;m'}\right\}\right\}, \\
&0 = \max\{0, -1\} \le \lambda_{4;4} \le \min\left\{0, \frac{2}{3}, \frac{4}{3}, 2, 1\right\} = 0,
\end{aligned}$$

and so $\lambda_{4;4} = 0$. For each $k = 3, 2, 1$, the same approach gives $\lambda_{4;3} = \frac{2}{3}$, $\lambda_{4;2} = \frac{5}{3}$, and $\lambda_{4;1} = \frac{5}{3}$. For the next column, we take $m = 4$. Starting with $k = 3$, we have:

$$\max\{\lambda_{4;4}, \lambda_{4;3} + \lambda_{4;4} - \mu_4\} \leq \lambda_{3;3}$$
$$\leq \min\left\{\lambda_{4;3}, \min_{l=1,\ldots,3}\left\{\sum_{m'=l}^{3} \mu_{m'} - \sum_{m'=l+1}^{3} \lambda_{4;m'}\right\}\right\},$$
$$0 = \max\left\{0, -\frac{1}{3}\right\} \leq \lambda_{3;3} \leq \min\left\{\frac{2}{3}, \frac{2}{3}, \frac{4}{3}, 1\right\} = \frac{2}{3}.$$

Notice that the lower and upper bounds on $\lambda_{3;3}$ are not equal. Since $\lambda_{3;3}$ is our first free variable, we parametrize it: $\lambda_{3;3} = x$ for some $x \in [0, \frac{2}{3}]$. Next, $k = 2$ gives

$$\frac{4}{3} - x = \max\left\{\frac{2}{3}, \frac{4}{3} - x\right\} \leq \lambda_{3;2} \leq \min\left\{\frac{5}{3}, \frac{4}{3} - x, 2 - x\right\} = \frac{4}{3} - x,$$

and so $\lambda_{3;2} = \frac{4}{3} - x$. Similarly, $\lambda_{3;1} = \frac{5}{3}$. Next, we take $m = 3$ and $k = 2$:

$$\max\left\{x, \frac{1}{3}\right\} \leq \lambda_{2;2} \leq \min\left\{\frac{4}{3} - x, \frac{2}{3} + x, 1\right\}.$$

Note that $\lambda_{2;2}$ is a free variable; we parametrize it as $\lambda_{2;2} = y$ such that:

$$y \in \left[\frac{1}{3}, \frac{2}{3} + x\right] \quad \text{if } x \in \left[0, \frac{1}{3}\right], \qquad y \in \left[x, \frac{4}{3} - x\right] \quad \text{if } x \in \left[\frac{1}{3}, \frac{2}{3}\right].$$

Finally, $\lambda_{2,1} = 2 - y$ and $\lambda_{1,1} = 1$.

We conclude by giving a complete constructive solution to Problem 2.1, that is, the problem of constructing every frame of a given spectrum and set of lengths. Recall from the introduction that it suffices to prove Theorem 2.4.

Proof of Theorem 2.4: We first show that such a Φ exists if and only if we have $\{\lambda_m\}_{m=1}^N \cup \{0\}_{m=N+1}^M \succeq \{\mu_m\}_{m=1}^M$. In particular, if such a Φ exists, then Theorem 2.2 implies that there exists a sequence of outer eigensteps corresponding to $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$; by Theorem 2.5, this implies that there exists a sequence of inner eigensteps corresponding to $\{\lambda_m\}_{m=1}^N \cup \{0\}_{m=N+1}^M$ and $\{\mu_m\}_{m=1}^M$; by the discussion at the beginning of Sect. 2.4.1, we necessarily have $\{\lambda_m\}_{m=1}^N \cup \{0\}_{m=N+1}^M \succeq \{\mu_m\}_{m=1}^M$. Conversely, if $\{\lambda_m\}_{m=1}^N \cup \{0\}_{m=N+1}^M \succeq \{\mu_m\}_{m=1}^M$, then Top Kill (Theorem 2.6) constructs a corresponding sequence of inner eigensteps, and so Theorem 2.5 implies that there exists a sequence of outer eigensteps corresponding to $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$, at which point Theorem 2.2 implies that such a Φ exists.

For the remaining conclusions, note that, in light of Theorem 2.2, it suffices to show that every valid sequence of outer eigensteps (Definition 2.2) satisfies the bounds of Step A of Theorem 2.4, and conversely, that every sequence constructed by Step A is a valid sequence of outer eigensteps. Both of these facts follow from the same two results. The first is Theorem 2.5, which establishes a correspondence between every valid sequence of outer eigensteps for $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$ with

a valid sequence of inner eigensteps for $\{\lambda_m\}_{m=1}^N \cup \{0\}_{m=N+1}^M$ and $\{\mu_m\}_{m=1}^M$ and vice versa, the two being zero-padded versions of each other. The second relevant result is Corollary 2.1, which characterizes all such inner eigensteps in terms of the bounds (2.42) and (2.43) of Theorem 2.7. In short, the algorithm of Step A is the outer eigenstep version of the application of Corollary 2.1 to $\{\lambda_m\}_{m=1}^N \cup \{0\}_{m=N+1}^M$; one may easily verify that all discrepancies between the statement of Theorem 2.4 and Corollary 2.1 are the result of the zero padding that occurs in the transition from inner to outer eigensteps. □

2.5 Constructing Frame Elements from Eigensteps

As discussed in Sect. 2.3, Theorem 2.2 provides a two-step process for constructing any and all sequences of vectors $\Phi = \{\varphi_m\}_{m=1}^M$ in $\mathbb{C}^N$ whose frame operator possesses a given spectrum $\{\lambda_n\}_{n=1}^N$ and whose vectors have given lengths $\{\mu_m\}_{m=1}^M$. In Step A, we choose a sequence of outer eigensteps $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$; this process is systematized in Theorem 2.4 of the previous section. In the end, the mth sequence $\{\lambda_{m;n}\}_{n=1}^N$ will become the spectrum of $\Phi_m\Phi_m^*$, where $\Phi_m = \{\varphi_{m'}\}_{m'=1}^m$.

Next, the purpose of Step B is to explicitly construct any and all sequences of vectors whose partial-frame-operator spectra match the outer eigensteps chosen in Step A. The problem with Step B of Theorem 2.2 is that it is not very explicit. Indeed for every $m = 1, \ldots, M-1$, in order to construct φ_{m+1}, we must first compute an orthonormal eigenbasis for $\Phi_m\Phi_m^*$. This problem is readily doable, since the eigenvalues $\{\lambda_{m;n}\}_{n=1}^N$ of $\Phi_m\Phi_m^*$ are already known. It is nevertheless a tedious and inelegant process to do by hand, requiring us to, for example, compute QR-factorizations of $\lambda_{m;n}Id - \Phi_m\Phi_m^*$ for each $n = 1, \ldots, N$. This section is devoted to the following result, which is a version of Theorem 2.2 equipped with a more explicit Step B; though technical, this new and improved Step B is still simple enough to be performed by hand. This material was first presented in [5].

Theorem 2.8 *For any nonnegative nonincreasing sequences $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$, every sequence of vectors $\Phi = \{\varphi_m\}_{m=1}^M$ in $\mathbb{C}^N$ whose frame operator $\Phi\Phi^*$ has spectrum $\{\lambda_n\}_{n=1}^N$ and which satisfies $\|\varphi_m\|^2 = \mu_m$ for all m can be constructed by the following algorithm*:

Step A: *Pick outer eigensteps as in Theorem 2.4.*

Step B: *Let U_1 be any unitary matrix with columns $\{u_{1;n}\}_{n=1}^N$. Let $\varphi_1 = \sqrt{\mu_1}u_{1;1}$. For each $m = 1, \ldots, M-1$:*

- B.1 *Let V_m be an $N \times N$ block-diagonal unitary matrix whose blocks correspond to the distinct values of $\{\lambda_{m;n}\}_{n=1}^N$ with the size of each block being the multiplicity of the corresponding eigenvalue.*
- B.2 *Identify those terms which are common to both $\{\lambda_{m;n}\}_{n=1}^N$ and $\{\lambda_{m+1;n}\}_{n=1}^N$. Specifically*:

- *Let* $\mathscr{I}_m \subseteq \{1,\ldots,N\}$ *consist of those indices* n *such that* $\lambda_{m;n} < \lambda_{m;n'}$ *for all* $n' < n$ *and such that the multiplicity of* $\lambda_{m;n}$ *as a value in* $\{\lambda_{m;n'}\}_{n'=1}^N$ *exceeds its multiplicity as a value in* $\{\lambda_{m+1;n'}\}_{n'=1}^N$.
- *Let* $\mathscr{J}_m \subseteq \{1,\ldots,N\}$ *consist of those indices* n *such that* $\lambda_{m+1;n} < \lambda_{m+1;n'}$ *for all* $n' < n$ *and also such that the multiplicity of* $\lambda_{m;n}$ *in* $\{\lambda_{m+1;n'}\}_{n'=1}^N$ *exceeds its multiplicity as a value in* $\{\lambda_{m;n'}\}_{n'=1}^N$.

The sets $\mathscr{I}_m$ *and* $\mathscr{J}_m$ *have equal cardinality, which we denote* R_m. *Next*:

- *Let* $\pi_{\mathscr{I}_m}$ *be the unique permutation on* $\{1,\ldots,N\}$ *that is increasing on both* $\mathscr{I}_m$ *and* $\mathscr{I}_m^{\mathrm{c}}$ *and such that* $\pi_{\mathscr{I}_m}(n) \in \{1,\ldots,R_m\}$ *for all* $n \in \mathscr{I}_m$. *Let* $\Pi_{\mathscr{I}_m}$ *be the associated permutation matrix* $\Pi_{\mathscr{I}_m}\delta_n = \delta_{\pi_{\mathscr{I}_m}(n)}$.
- *Let* $\pi_{\mathscr{J}_m}$ *be the unique permutation on* $\{1,\ldots,N\}$ *that is increasing on both* $\mathscr{J}_m$ *and* $\mathscr{J}_m^{\mathrm{c}}$ *and such that* $\pi_{\mathscr{J}_m}(n) \in \{1,\ldots,R_m\}$ *for all* $n \in \mathscr{J}_m$. *Let* $\Pi_{\mathscr{J}_m}$ *be the associated permutation matrix* $\Pi_{\mathscr{J}_m}\delta_n = \delta_{\pi_{\mathscr{J}_m}(n)}$.

B.3 *Let* v_m, w_m *be the* $R_m \times 1$ *vectors whose entries are*:

$$v_m\big(\pi_{\mathscr{I}_m}(n)\big) = \left[-\frac{\prod_{n''\in\mathscr{J}_m}(\lambda_{m;n}-\lambda_{m+1;n''})}{\prod_{\substack{n''\in\mathscr{I}_m\\ n''\neq n}}(\lambda_{m;n}-\lambda_{m;n''})}\right]^{\frac{1}{2}}, \quad \forall n \in \mathscr{I}_m,$$

$$w_m\big(\pi_{\mathscr{J}_m}(n')\big) = \left[\frac{\prod_{n''\in\mathscr{I}_m}(\lambda_{m+1;n'}-\lambda_{m;n''})}{\prod_{\substack{n''\in\mathscr{J}_m\\ n''\neq n'}}(\lambda_{m+1;n'}-\lambda_{m+1;n''})}\right]^{\frac{1}{2}}, \quad \forall n' \in \mathscr{J}_m.$$

B.4 $\varphi_{m+1} = U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \left[\begin{smallmatrix} v_m \\ 0 \end{smallmatrix}\right]$, *where the* $N \times 1$ *vector* $\left[\begin{smallmatrix} v_m \\ 0 \end{smallmatrix}\right]$ *is* v_m *with* $N - R_m$ *zeros*.

B.5 $U_{m+1} = U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \left[\begin{smallmatrix} W_m & 0 \\ 0 & Id \end{smallmatrix}\right] \Pi_{\mathscr{J}_m}$ *where* W_m *is the* $R_m \times R_m$ *matrix with entries*:

$$W_m\big(\pi_{\mathscr{I}_m}(n), \pi_{\mathscr{J}_m}(n')\big) = \frac{1}{\lambda_{m+1;n'} - \lambda_{m;n}} v_m\big(\pi_{\mathscr{I}_m}(n)\big) w_m\big(\pi_{\mathscr{J}_m}(n')\big).$$

Conversely, any Φ *constructed by this process has* $\{\lambda_n\}_{n=1}^N$ *as the spectrum of* $\Phi\Phi^*$ *and* $\|\varphi_m\|^2 = \mu_m$ *for all* m. *Moreover, for any* Φ *constructed in this manner and any* $m = 1,\ldots,M$, *the spectrum of the frame operator* $\Phi_m\Phi_m^*$ *arising from the partial sequence* $\Phi_m = \{\varphi_{m'}\}_{m'=1}^m$ *is* $\{\lambda_{m;n}\}_{n=1}^N$, *and the columns of* U_m *form a corresponding orthonormal eigenbasis for* $\Phi_m\Phi_m^*$.

Before proving Theorem 2.8, we give an example of its implementation, with the hope of conveying the simplicity of the underlying idea, and better explaining the heavy notation used in the statement of the result.

Example 2.6 Recall from Example 2.2 that the valid outer eigensteps (2.21) corresponding to 3×5 UNTFs are given by

m	0	1	2	3	4	5
$\lambda_{m;3}$	0	0	0	x	$\frac{2}{3}$	$\frac{5}{3}$
$\lambda_{m;2}$	0	0	y	$\frac{4}{3}-x$	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	0	1	$2-y$	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

where $x \in [0, \frac{2}{3}]$ and $y \in [\max\{\frac{1}{3}, x\}, \min\{\frac{2}{3}+x, \frac{4}{3}-x\}]$. To complete Step A of Theorem 2.8, we pick any valid (x, y). For example, for $(x, y) = (0, \frac{1}{3})$, (2.21) becomes

m	0	1	2	3	4	5
$\lambda_{m;3}$	0	0	0	0	$\frac{2}{3}$	$\frac{5}{3}$
$\lambda_{m;2}$	0	0	$\frac{1}{3}$	$\frac{4}{3}$	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	0	1	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$	$\frac{5}{3}$

(2.53)

Note that this particular choice corresponds to Top Kill. We now perform Step B of Theorem 2.8 for this particular choice of eigensteps. First, we must choose a unitary matrix U_1. Considering the equation for U_{m+1} along with the fact that the columns of U_M will form an eigenbasis for $\Phi\Phi^*$, we see that our choice for U_1 merely rotates this eigenbasis, and hence the entire frame Φ, to our liking. We choose $U_1 = Id$ for simplicity. Thus,

$$\varphi_1 = \sqrt{\mu_1}u_{1;1} = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}.$$

We now iterate, performing Steps B.1 through B.5 for $m = 1$ to find φ_2 and U_2, then performing Steps B.1 through B.5 for $m = 2$ to find φ_3 and U_3, and so on. Throughout this process, the only remaining choices to be made appear in Step B.1. In particular, for $m = 1$ Step B.1 asks us to pick a block-diagonal unitary matrix V_1 whose blocks are sized according to the multiplicities of the eigenvalues $\{\lambda_{1;1}, \lambda_{1;2}, \lambda_{1;3}\} = \{1, 0, 0\}$. That is, V_1 consists of a 1×1 unitary block—a unimodular scalar—and a 2×2 unitary block. There are an infinite number of such V_1's, each leading to a distinct frame. For simplicity, we choose $V_1 = Id$. Having completed Step B.1 for $m = 1$, we turn to Step B.2, which requires us to consider the columns of (2.53) that correspond to $m = 1$ and $m = 2$:

m	1	2
$\lambda_{m;3}$	0	0
$\lambda_{m;2}$	0	$\frac{1}{3}$
$\lambda_{m;1}$	1	$\frac{5}{3}$

(2.54)

In particular, we compute a set of indices $\mathscr{I}_1 \subseteq \{1, 2, 3\}$ that contains the indices n of $\{\lambda_{1;1}, \lambda_{1;2}, \lambda_{1;3}\} = \{1, 0, 0\}$ for which (i) the multiplicity of $\lambda_{1;n}$ as a value of

$\{1,0,0\}$ exceeds its multiplicity as a value of $\{\lambda_{2;1},\lambda_{2;2},\lambda_{2;3}\} = \{\frac{5}{3},\frac{1}{3},0\}$ and (ii) n corresponds to the first occurrence of $\lambda_{1;n}$ as a value of $\{1,0,0\}$; by these criteria, we find $\mathscr{I}_1 = \{1,2\}$. Similarly, $n \in \mathscr{J}_1$ if and only if n indicates the first occurrence of a value $\lambda_{2;n}$ whose multiplicity as a value of $\{\frac{5}{3},\frac{1}{3},0\}$ exceeds its multiplicity as a value of $\{1,0,0\}$, and so $\mathscr{J}_1 = \{1,2\}$. Equivalently, $\mathscr{I}_1$ and $\mathscr{J}_1$ can be obtained by canceling common terms from (2.54), working top to bottom. An explicit algorithm for doing so is given in Table 2.2.

Continuing with Step B.2 for $m = 1$, we now find the unique permutation $\pi_{\mathscr{I}_1} : \{1,2,3\} \to \{1,2,3\}$ that is increasing on both $\mathscr{I}_1 = \{1,2\}$ and its complement $\mathscr{I}_1^{\mathrm{c}} = \{3\}$ and takes $\mathscr{I}_1$ to the first $R_1 = |\mathscr{I}_1| = 2$ elements of $\{1,2,3\}$. In this particular instance, $\pi_{\mathscr{I}_1}$ happens to be the identity permutation, and so $\Pi_{\mathscr{I}_1} = Id$. Since $\mathscr{J}_1 = \{1,2\} = \mathscr{I}_1$, we similarly have that $\pi_{\mathscr{J}_1}$ and $\Pi_{\mathscr{J}_1}$ are the identity permutation and matrix, respectively. For the remaining steps, it is useful to isolate the terms in (2.54) that correspond to $\mathscr{I}_1$ and $\mathscr{J}_1$:

$$\begin{aligned} \beta_2 &= \lambda_{1;2} = 0, & \gamma_2 &= \lambda_{2;2} = \frac{1}{3}, \\ \beta_1 &= \lambda_{1;1} = 1, & \gamma_1 &= \lambda_{2;1} = \frac{5}{3}. \end{aligned} \tag{2.55}$$

In particular, in Step B.3, we find the $R_1 \times 1 = 2 \times 1$ vector v_1 by computing quotients of products of differences of the values in (2.55):

$$\big[v_1(1)\big]^2 = -\frac{(\beta_1-\gamma_1)(\beta_1-\gamma_2)}{(\beta_1-\beta_2)} = -\frac{(1-\frac{5}{3})(1-\frac{1}{3})}{(1-0)} = \frac{4}{9}, \tag{2.56}$$

$$\big[v_1(2)\big]^2 = -\frac{(\beta_2-\gamma_1)(\beta_2-\gamma_2)}{(\beta_2-\beta_1)} = -\frac{(0-\frac{5}{3})(0-\frac{1}{3})}{(0-1)} = \frac{5}{9}, \tag{2.57}$$

yielding $v_1 = \left[\begin{smallmatrix} \frac{2}{3} \\ \frac{\sqrt{5}}{3} \end{smallmatrix}\right]$. Similarly, we compute $w_1 = \left[\begin{smallmatrix} \frac{\sqrt{5}}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} \end{smallmatrix}\right]$ according to the following formulas:

$$\big[w_1(1)\big]^2 = \frac{(\gamma_1-\beta_1)(\gamma_1-\beta_2)}{(\gamma_1-\gamma_2)} = \frac{(\frac{5}{3}-1)(\frac{5}{3}-0)}{(\frac{5}{3}-\frac{1}{3})} = \frac{5}{6}, \tag{2.58}$$

$$\big[w_1(2)\big]^2 = \frac{(\gamma_2-\beta_1)(\gamma_2-\beta_2)}{(\gamma_2-\gamma_1)} = \frac{(\frac{1}{3}-1)(\frac{1}{3}-0)}{(\frac{1}{3}-\frac{5}{3})} = \frac{1}{6}. \tag{2.59}$$

Next, in Step B.4, we form our second frame element $\varphi_2 = U_1 V_1 \Pi_{\mathscr{J}_1}^{\mathrm{T}} \left[\begin{smallmatrix} v_1 \\ 0 \end{smallmatrix}\right]$:

$$\varphi_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{2}{3} \\ \frac{\sqrt{5}}{3} \\ 0 \end{bmatrix} = \begin{bmatrix} \frac{2}{3} \\ \frac{\sqrt{5}}{3} \\ 0 \end{bmatrix}.$$

As justified in the proof of Theorem 2.8, the resulting partial sequence of vectors:

$$\Phi_2 = \begin{bmatrix} \varphi_1 & \varphi_2 \end{bmatrix} = \begin{bmatrix} 1 & \frac{2}{3} \\ 0 & \frac{\sqrt{5}}{3} \\ 0 & 0 \end{bmatrix}$$

has a frame operator $\Phi_2\Phi_2^*$ whose spectrum is $\{\lambda_{2;1}, \lambda_{2;2}, \lambda_{2;3}\} = \{\frac{5}{3}, \frac{1}{3}, 0\}$. Moreover, a corresponding orthonormal eigenbasis for $\Phi_2\Phi_2^*$ is computed in Step B.5; here the first step is to compute the $R_1 \times R_1 = 2 \times 2$ matrix W_1 by computing a pointwise product of a certain 2×2 matrix with the outer product of v_1 with w_1:

$$W_1 = \begin{bmatrix} \frac{1}{\gamma_1 - \beta_1} & \frac{1}{\gamma_2 - \beta_1} \\ \frac{1}{\gamma_1 - \beta_2} & \frac{1}{\gamma_2 - \beta_2} \end{bmatrix} \odot \begin{bmatrix} v_1(1) \\ v_1(2) \end{bmatrix} \begin{bmatrix} w_1(1) \\ w_1(2) \end{bmatrix}^{\mathrm{T}} = \begin{bmatrix} \frac{3}{2} & -\frac{3}{2} \\ \frac{3}{5} & 3 \end{bmatrix} \odot \begin{bmatrix} \frac{2\sqrt{5}}{3\sqrt{6}} & \frac{2}{3\sqrt{6}} \\ \frac{5}{3\sqrt{6}} & \frac{\sqrt{5}}{3\sqrt{6}} \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} \end{bmatrix}.$$

Note that W_1 is a real orthogonal matrix whose diagonal and subdiagonal entries are strictly positive and whose superdiagonal entries are strictly negative; one can easily verify that every W_m has this form. More significantly, the proof of Theorem 2.8 guarantees that the columns of:

$$U_2 = U_1 V_1 \Pi_{\mathscr{J}_1}^{\mathrm{T}} \begin{bmatrix} W_1 & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{J}_1}$$

$$= \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

form an orthonormal eigenbasis of $\Phi_2\Phi_2^*$. This completes the $m = 1$ iteration of Step B; we now repeat this process for $m = 2, 3, 4$. For $m = 2$, in Step B.1 we arbitrarily pick some 3×3 diagonal unitary matrix V_2. Note that if we want a real frame, there are only $2^3 = 8$ such choices of V_2. For simplicity, we choose $V_2 = Id$ in this example. Continuing, Step B.2 involves canceling the common terms in

m	2	3
$\lambda_{m;3}$	0	0
$\lambda_{m;2}$	$\frac{1}{3}$	$\frac{4}{3}$
$\lambda_{m;1}$	$\frac{5}{3}$	$\frac{5}{3}$

to find $\mathscr{I}_2 = \mathscr{J}_2 = \{2\}$, and so

$$\Pi_{\mathscr{I}_2} = \Pi_{\mathscr{J}_2} = \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

In Step B.3, we find that $v_2 = w_2 = [1]$. Steps B.4 and B.5 then give that $\Phi_3 = \begin{bmatrix} \varphi_1 & \varphi_2 & \varphi_3 \end{bmatrix}$ and U_3 are:

$$\Phi_3 = \begin{bmatrix} 1 & \frac{2}{3} & -\frac{1}{\sqrt{6}} \\ 0 & \frac{\sqrt{5}}{3} & \frac{\sqrt{5}}{\sqrt{6}} \\ 0 & 0 & 0 \end{bmatrix}, \qquad U_3 = \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

The columns of U_3 form an orthonormal eigenbasis for the partial frame operator $\Phi_3\Phi_3^*$ with corresponding eigenvalues $\{\lambda_{3;1}, \lambda_{3;2}, \lambda_{3;3}\} = \{\frac{5}{3}, \frac{4}{3}, 0\}$. For the $m = 3$ iteration, we pick $V_3 = Id$ and cancel the common terms in

m	3	4
$\lambda_{m;3}$	0	$\frac{2}{3}$
$\lambda_{m;2}$	$\frac{4}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	$\frac{5}{3}$	$\frac{5}{3}$

to obtain $\mathscr{I}_3 = \{2, 3\}$ and $\mathscr{J}_3 = \{1, 3\}$, implying:

$$\Pi_{\mathscr{I}_3} = \begin{bmatrix} 0 & 1 & 0 \\ 0 & 0 & 1 \\ 1 & 0 & 0 \end{bmatrix}, \qquad \Pi_{\mathscr{J}_3} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix},$$

$$\beta_2 = \lambda_{3;3} = 0, \qquad \gamma_2 = \lambda_{4;3} = \frac{2}{3},$$

$$\beta_1 = \lambda_{3;2} = \frac{4}{3}, \qquad \gamma_1 = \lambda_{4;1} = \frac{5}{3}.$$

In Step B.3, we then compute the $R_3 \times 1 = 2 \times 1$ vectors v_3 and w_3 in a manner analogous to (2.56), (2.57), (2.58) and (2.59):

$$v_3 = \begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{\sqrt{5}}{\sqrt{6}} \end{bmatrix}, \qquad w_3 = \begin{bmatrix} \frac{\sqrt{5}}{3} \\ \frac{2}{3} \end{bmatrix}.$$

Note that in Step B.4, the role of permutation matrix $\Pi_{\mathscr{J}_3}^{\mathrm{T}}$ is that it maps the entries of v_3 onto the $\mathscr{I}_3$ indices, meaning that v_4 lies in the span of the corresponding

eigenvectors $\{u_{3;n}\}_{n\in\mathscr{I}_3}$:

$$\varphi_4 = \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} \\ \frac{\sqrt{5}}{\sqrt{6}} \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ \frac{1}{\sqrt{6}} \\ \frac{\sqrt{5}}{\sqrt{6}} \end{bmatrix} = \begin{bmatrix} -\frac{1}{6} \\ \frac{\sqrt{5}}{6} \\ \frac{\sqrt{5}}{\sqrt{6}} \end{bmatrix}.$$

In a similar fashion, the purpose of the permutation matrices in Step B.5 is to embed the entries of the 2×2 matrix W_3 into the $\mathscr{I}_3 = \{2, 3\}$ rows and $\mathscr{J}_3 = \{1, 3\}$ columns of a 3×3 matrix:

$$U_4 = \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\sqrt{5}}{\sqrt{6}} & -\frac{1}{\sqrt{6}} & 0 \\ \frac{1}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 \\ \frac{\sqrt{5}}{\sqrt{6}} & 0 & -\frac{1}{\sqrt{6}} \\ \frac{1}{\sqrt{6}} & 0 & \frac{\sqrt{5}}{\sqrt{6}} \end{bmatrix}$$

$$= \begin{bmatrix} -\frac{\sqrt{5}}{6} & \frac{\sqrt{5}}{\sqrt{6}} & \frac{1}{6} \\ \frac{5}{6} & \frac{1}{\sqrt{6}} & -\frac{\sqrt{5}}{6} \\ \frac{1}{\sqrt{6}} & 0 & \frac{\sqrt{5}}{\sqrt{6}} \end{bmatrix}.$$

For the last iteration $m = 4$, we again choose $V_4 = Id$ in Step B.1. For Step B.2, note that since

m	4	5
$\lambda_{m;3}$	$\frac{2}{3}$	$\frac{5}{3}$
$\lambda_{m;2}$	$\frac{5}{3}$	$\frac{5}{3}$
$\lambda_{m;1}$	$\frac{5}{3}$	$\frac{5}{3}$

we have $\mathscr{I}_4 = \{3\}$ and $\mathscr{J}_4 = \{1\}$, implying:

$$\Pi_{\mathscr{I}_4} = \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}, \qquad \Pi_{\mathscr{J}_4} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Working through Steps B.3, B.4, and B.5 yields the UNTF:

$$\Phi = \Phi_5 = \begin{bmatrix} 1 & \frac{2}{3} & -\frac{1}{\sqrt{6}} & -\frac{1}{6} & \frac{1}{6} \\ 0 & \frac{\sqrt{5}}{3} & \frac{\sqrt{5}}{\sqrt{6}} & \frac{\sqrt{5}}{6} & -\frac{\sqrt{5}}{6} \\ 0 & 0 & 0 & \frac{\sqrt{5}}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} \end{bmatrix}, \qquad U_5 = \begin{bmatrix} \frac{1}{6} & -\frac{\sqrt{5}}{6} & \frac{\sqrt{5}}{\sqrt{6}} \\ -\frac{\sqrt{5}}{6} & \frac{5}{6} & \frac{1}{\sqrt{6}} \\ \frac{\sqrt{5}}{\sqrt{6}} & \frac{1}{\sqrt{6}} & 0 \end{bmatrix}. \tag{2.60}$$

We emphasize that the UNTF Φ given in (2.60) was based on the particular choice of eigensteps given in (2.53), which arose by choosing $(x, y) = (0, \frac{1}{3})$ in (2.21). Choosing other pairs (x, y) from the parameter set depicted in Fig. 2.2(b) yields other UNTFs. Indeed, since the eigensteps of a given Φ are equal to those of $U\Phi$ for any unitary operator U, we have that each distinct (x, y) yields a UNTF which is not unitarily equivalent to any of the others. For example, by following the algorithm of Theorem 2.8 and choosing $U_1 = Id$ and $V_m = Id$ in each iteration, we obtain the following four additional UNTFs, each corresponding to a distinct corner point of the parameter set:

$$\Phi = \begin{bmatrix} 1 & \frac{2}{3} & 0 & -\frac{1}{3} & -\frac{1}{3} \\ 0 & \frac{\sqrt{5}}{3} & 0 & \frac{\sqrt{5}}{3} & \frac{\sqrt{5}}{3} \\ 0 & 0 & 1 & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \end{bmatrix} \quad \text{for } (x, y) = \left(\frac{1}{3}, \frac{1}{3}\right),$$

$$\Phi = \begin{bmatrix} 1 & \frac{1}{3} & \frac{1}{3} & -\frac{1}{3} & -\frac{1}{\sqrt{3}} \\ 0 & \frac{\sqrt{8}}{3} & \frac{1}{3\sqrt{2}} & -\frac{1}{3\sqrt{2}} & \frac{\sqrt{2}}{\sqrt{3}} \\ 0 & 0 & \frac{\sqrt{5}}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} & 0 \end{bmatrix} \quad \text{for} (x, y) = \left(\frac{2}{3}, \frac{2}{3}\right),$$

$$\Phi = \begin{bmatrix} 1 & 0 & 0 & \frac{1}{\sqrt{3}} & -\frac{1}{\sqrt{3}} \\ 0 & 1 & \frac{2}{3} & -\frac{1}{3} & -\frac{1}{3} \\ 0 & 0 & \frac{\sqrt{5}}{3} & \frac{\sqrt{5}}{3} & \frac{\sqrt{5}}{3} \end{bmatrix} \quad \text{for } (x, y) = \left(\frac{1}{3}, 1\right),$$

$$\Phi = \begin{bmatrix} 1 & \frac{1}{3} & -\frac{1}{\sqrt{3}} & \frac{1}{3} & -\frac{1}{3} \\ 0 & \frac{\sqrt{8}}{3} & \frac{\sqrt{2}}{\sqrt{3}} & \frac{1}{3\sqrt{2}} & -\frac{1}{3\sqrt{2}} \\ 0 & 0 & 0 & \frac{\sqrt{5}}{\sqrt{6}} & \frac{\sqrt{5}}{\sqrt{6}} \end{bmatrix} \quad \text{for } (x, y) = \left(0, \frac{2}{3}\right).$$

Notice that, of the four UNTFs above, the second and fourth are actually the same up to a permutation of the frame elements. This is an artifact of our method of construction, namely, that our choices for eigensteps, U_1, and $\{V_m\}_{m=1}^{M-1}$ determine the *sequence* of frame elements. As such, we can recover all permutations of a given frame by modifying these choices.

We emphasize that these four UNTFs along with that of (2.60) are but five examples from the continuum of all such frames. Indeed, keeping x and y as variables in (2.21) and applying the algorithm of Theorem 2.8—again choosing $U_1 = Id$

and $V_m = Id$ in each iteration for simplicity—yields the frame elements given in Table 2.1. Here, we restrict (x, y) so as to not lie on the boundary of the parameter set of Fig. 2.2(b). This restriction simplifies the analysis, as it prevents all unnecessary repetitions of values in neighboring columns in (2.21). Table 2.1 gives an explicit parametrization for a two-dimensional manifold that lies within the set of all UNTFs consisting of five elements in three-dimensional space. By Theorem 2.8, this can be generalized so as to yield all such frames, provided we both (i) further consider (x, y) that lie on each of the five line segments that constitute the boundary of the parameter set and (ii) throughout generalize V_m to an arbitrary block-diagonal unitary matrix, where the sizes of the blocks are chosen in accordance with Step B.1.

Having discussed the utility of Theorem 2.8, we turn to its proof.

Proof of Theorem 2.8 ($\Leftarrow$) Let $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$ be arbitrary nonnegative nonincreasing sequences and take an arbitrary sequence of outer eigensteps $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$ in accordance with Definition 2.2. Note that here we do not assume that such a sequence of eigensteps actually exists for this particular choice of $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$; if one does not, then this direction of the result is vacuously true.

We claim that any $\Phi = \{\varphi_m\}_{m=1}^M$ constructed according to Step B has the property that for all $m = 1, \ldots, M$, the spectrum of the frame operator $\Phi_m \Phi_m^*$ of $\Phi_m = \{\varphi_{m'}\}_{m'=1}^m$ is $\{\lambda_{m;n}\}_{n=1}^N$, and that the columns of U_m form an orthonormal eigenbasis for $\Phi_m \Phi_m^*$. Note that, by Lemma 2.1, proving this claim will yield our stated result that the spectrum of $\Phi\Phi^*$ is $\{\lambda_n\}_{n=1}^N$ and that $\|\varphi_m\|^2 = \mu_m$ for all $m = 1, \ldots, M$. Since Step B is an iterative algorithm, we prove this claim by induction on m. To be precise, Step B begins by letting $U_1 = \{u_{1;n}\}_{n=1}^N$ and $\varphi_1 = \sqrt{\mu_1} u_{1;1}$. The columns of U_1 form an orthonormal eigenbasis for $\Phi_1 \Phi_1^*$ since U_1 is unitary by assumption and

$$\Phi_1 \Phi_1^* u_{1;n} = \langle u_{1;n}, \varphi_1 \rangle \varphi_1 = \mu_1 \langle u_{1;n}, u_{1;1} \rangle u_{1;1} = \begin{cases} \mu_1 u_{1;1} & n = 1, \\ 0 & n \neq 1, \end{cases}$$

for all $n = 1, \ldots, N$. As such, the spectrum of $\Phi_1 \Phi_1^*$ consists of μ_1 and $N - 1$ repetitions of 0. To see that this spectrum matches the values of $\{\lambda_{1;n}\}_{n=1}^N$, note that, by Definition 2.2, we know $\{\lambda_{1;n}\}_{n=1}^N$ interlaces on the trivial sequence $\{\lambda_{0;n}\}_{n=1}^N = \{0\}_{n=1}^N$ in the sense of (2.10), implying $\lambda_{1;n} = 0$ for all $n \geq 2$; this in hand, this definition also gives that $\lambda_{1;1} = \sum_{n=1}^N \lambda_{1;n} = \mu_1$. Thus, our claim indeed holds for $m = 1$.

We now proceed by induction, assuming that for any given $m = 1, \ldots, M - 1$ the process of Step B has produced $\Phi_m = \{\varphi_{m'}\}_{m'=1}^m$ such that the spectrum of $\Phi_m \Phi_m^*$ is $\{\lambda_{m;n}\}_{n=1}^N$ and that the columns of U_m form an orthonormal eigenbasis for $\Phi_m \Phi_m^*$. In particular, we have $\Phi_m \Phi_m^* U_m = U_m D_m$ where D_m is the diagonal matrix whose diagonal entries are $\{\lambda_{m;n}\}_{n=1}^N$. Defining D_{m+1} analogously from $\{\lambda_{m+1;n}\}_{n=1}^N$, we show that constructing φ_{m+1} and U_{m+1} according to Step B implies $\Phi_{m+1} \Phi_{m+1}^* U_{m+1} = U_{m+1} D_{m+1}$ where U_{m+1} is unitary; doing so proves our claim.

Table 2.1 A continuum of UNTFs. To be precise, for each choice of (x, y) that lies in the interior of the parameter set depicted in Fig. 2.2(b), these five elements form a UNTF for $\mathbb{C}^3$, meaning that its 3×5 synthesis matrix Φ has both unit norm columns and orthogonal rows of constant squared norm $\frac{5}{3}$. These frames were produced by applying the algorithm of Theorem 2.8 to the sequence of eigensteps given in (2.21), choosing $U_1 = Id$ and $V_m = Id$ for all m. These formulas give an explicit parametrization for a two-dimensional manifold that lies within the set of all 3×5 UNTFs

$$\varphi_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}$$

$$\varphi_2 = \begin{bmatrix} 1-y \\ \sqrt{y(2-y)} \\ 0 \end{bmatrix}$$

$$\varphi_3 = \begin{bmatrix} \frac{\sqrt{(3y-1)(2+3x-3y)(2-x-y)}}{6\sqrt{1-y}} & - & \frac{\sqrt{(5-3y)(4-3x-3y)(y-x)}}{6\sqrt{1-y}} \\ \frac{\sqrt{y(3y-1)(2+3x-3y)(2-x-y)}}{6\sqrt{(1-y)(2-y)}} & + & \frac{\sqrt{(5-3y)(2-y)(4-3x-3y)(y-x)}}{6\sqrt{y(1-y)}} \\ \frac{\sqrt{5x(4-3x)}}{3\sqrt{y(2-y)}} & & \end{bmatrix}$$

$$\varphi_4 = \begin{bmatrix} -\frac{\sqrt{(4-3x)(3y-1)(2-x-y)(4-3x-3y)}}{12\sqrt{(2-3x)(1-y)}} & - & \frac{\sqrt{(4-3x)(5-3y)(y-x)(2+3x-3y)}}{12\sqrt{(2-3x)(1-y)}} & - & \frac{\sqrt{x(3y-1)(y-x)(2+3x-3y)}}{4\sqrt{3(2-3x)(1-y)}} & + & \frac{\sqrt{x(5-3y)(2-x-y)(4-3x-3y)}}{4\sqrt{3(2-3x)(1-y)}} \\ -\frac{\sqrt{(4-3x)y(3y-1)(2-x-y)(4-3x-3y)}}{12\sqrt{(2-3x)(1-y)(2-y)}} & + & \frac{\sqrt{(4-3x)(2-y)(5-3y)(y-x)(2+3x-3y)}}{12\sqrt{(2-3x)y(1-y)}} & - & \frac{\sqrt{xy(3y-1)(y-x)(2+3x-3y)}}{4\sqrt{3(2-3x)(1-y)(2-y)}} & - & \frac{\sqrt{x(2-y)(5-3y)(2-x-y)(4-3x-3y)}}{4\sqrt{3(2-3x)y(1-y)}} \\ \frac{\sqrt{5x(2+3x-3y)(4-3x-3y)}}{6\sqrt{(2-3x)y(2-y)}} & + & \frac{\sqrt{5(4-3x)(y-x)(2-x-y)}}{2\sqrt{3(2-3x)y(2-y)}} & & & & \end{bmatrix}$$

$$\varphi_5 = \begin{bmatrix} \frac{\sqrt{(4-3x)(3y-1)(2-x-y)(4-3x-3y)}}{12\sqrt{(2-3x)(1-y)}} & + & \frac{\sqrt{(4-3x)(5-3y)(y-x)(2+3x-3y)}}{12\sqrt{(2-3x)(1-y)}} & - & \frac{\sqrt{x(3y-1)(y-x)(2+3x-3y)}}{4\sqrt{3(2-3x)(1-y)}} & + & \frac{\sqrt{x(5-3y)(2-x-y)(4-3x-3y)}}{4\sqrt{3(2-3x)(1-y)}} \\ \frac{\sqrt{(4-3x)y(3y-1)(2-x-y)(4-3x-3y)}}{12\sqrt{(2-3x)(1-y)(2-y)}} & - & \frac{\sqrt{(4-3x)(2-y)(5-3y)(y-x)(2+3x-3y)}}{12\sqrt{(2-3x)y(1-y)}} & - & \frac{\sqrt{xy(3y-1)(y-x)(2+3x-3y)}}{4\sqrt{3(2-3x)(1-y)(2-y)}} & - & \frac{\sqrt{x(2-y)(5-3y)(2-x-y)(4-3x-3y)}}{4\sqrt{3(2-3x)y(1-y)}} \\ -\frac{\sqrt{5x(2+3x-3y)(4-3x-3y)}}{6\sqrt{(2-3x)y(2-y)}} & + & \frac{\sqrt{5(4-3x)(y-x)(2-x-y)}}{2\sqrt{3(2-3x)y(2-y)}} & & & & \end{bmatrix}$$

Table 2.2 An explicit algorithm for computing the index sets $\mathscr{I}_m$ and $\mathscr{J}_m$ in Step B.2 of Theorem 2.8

01	$\mathscr{I}_m^{(N)} := \{1, \ldots, N\}$
02	$\mathscr{J}_m^{(N)} := \{1, \ldots, N\}$
03	for $n = N, \ldots, 1$
04	if $\lambda_{m;n} \in \{\lambda_{m+1;n'}\}_{n' \in \mathscr{J}_m^{(n)}}$
05	$\mathscr{I}_m^{(n-1)} := \mathscr{I}_m^{(n)} \setminus \{n\}$
06	$\mathscr{J}_m^{(n-1)} := \mathscr{J}_m^{(n)} \setminus \{n'\}$ where $n' = \max\{n'' \in \mathscr{J}_m^{(n)} : \lambda_{m+1;n''} = \lambda_{m;n}\}$
07	else
08	$\mathscr{I}_m^{(n-1)} := \mathscr{I}_m^{(n)}$
09	$\mathscr{J}_m^{(n-1)} := \mathscr{J}_m^{(n)}$
10	end if
11	end for
12	$\mathscr{I}_m := \mathscr{I}_m^{(1)}$
13	$\mathscr{J}_m := \mathscr{J}_m^{(1)}$

To do so, pick any unitary matrix V_m according to Step B.1. To be precise, let K_m denote the number of distinct values in $\{\lambda_{m;n}\}_{n=1}^N$, and for any $k = 1, \ldots, K_m$, let $L_{m;k}$ denote the multiplicity of the kth value. We write the index n as an increasing function of k and l; that is, we write $\{\lambda_{m;n}\}_{n=1}^N$ as $\{\lambda_{m;n(k,l)}\}_{k=1}^{K_m}{}_{l=1}^{L_{m;k}}$ where $n(k,l) < n(k',l')$ if $k < k'$ or if $k = k'$ and $l < l'$. We let V_m be an $N \times N$ block-diagonal unitary matrix consisting of K diagonal blocks, where for any $k = 1, \ldots, K$, the kth block is an $L_{m;k} \times L_{m;k}$ unitary matrix. In the extreme case where all the values of $\{\lambda_{m;n}\}_{n=1}^N$ are distinct, we have that V_m is a diagonal unitary matrix, meaning it is a diagonal matrix whose diagonal entries are unimodular. Even in this case, there is some freedom in how to choose V_m; this is the only freedom that the Step B process provides when determining φ_{m+1}. In any case, the crucial fact about V_m is that its blocks match those corresponding to distinct multiples of the identity that appear along the diagonal of D_m, implying $D_m V_m = V_m D_m$.

Having chosen V_m, we proceed to Step B.2. Here, we produce subsets $\mathscr{I}_m$ and $\mathscr{J}_m$ of $\{1, \ldots, N\}$ that are the remnants of the indices of $\{\lambda_{m;n}\}_{n=1}^N$ and $\{\lambda_{m+1;n}\}_{n=1}^N$, respectively, obtained by canceling the values that are common to both sequences, working backward from index N to index 1. An explicit algorithm for doing so is given in Table 2.2. Note that, for each $n = N, \ldots, 1$ (Line 03), we either remove a single element from both $\mathscr{I}_m^{(n)}$ and $\mathscr{J}_m^{(n)}$ (Lines 04–06) or remove nothing from both (Lines 07–09), meaning that $\mathscr{I}_m := \mathscr{I}_m^{(1)}$ and $\mathscr{J}_m := \mathscr{J}_m^{(1)}$ have the same cardinality, which we denote R_m. Moreover, since $\{\lambda_{m+1;n}\}_{n=1}^N$ interlaces on $\{\lambda_{m;n}\}_{n=1}^N$, then for any real scalar λ whose multiplicity as a value of $\{\lambda_{m;n}\}_{n=1}^N$ is L, we have that its multiplicity as a value of $\{\lambda_{m+1;n}\}_{n=1}^N$ is either $L-1$, L or $L+1$. When these two multiplicities are equal, this algorithm completely removes the corresponding indices from both $\mathscr{I}_m$ and $\mathscr{J}_m$. On the other hand, if the new multiplicity is $L-1$ or $L+1$, then the least such index in $\mathscr{I}_m$ or $\mathscr{J}_m$ is left behind,

respectively, leading to the definitions of $\mathscr{I}_m$ or $\mathscr{J}_m$ given in Step B.2. Having these sets, it is trivial to find the corresponding permutations $\pi_{\mathscr{I}_m}$ and $\pi_{\mathscr{J}_m}$ on $\{1,\ldots,N\}$ and to construct the associated projection matrices $\Pi_{\mathscr{I}_m}$ and $\Pi_{\mathscr{J}_m}$.

We now proceed to Step B.3. For notational simplicity, let $\{\beta_r\}_{r=1}^{R_m}$ and $\{\gamma_r\}_{r=1}^{R_m}$ denote the values of $\{\lambda_{m;n}\}_{n\in\mathscr{I}_m}$ and $\{\lambda_{m+1;n}\}_{n\in\mathscr{J}_m}$, respectively. That is, let $\beta_{\pi_{\mathscr{I}_m}(n)} = \lambda_{m;n}$ for all $n \in \mathscr{I}_m$ and $\gamma_{\pi_{\mathscr{J}_m}(n)} = \lambda_{m+1;n}$ for all $n \in \mathscr{J}_m$. Note that due to the way in which $\mathscr{I}_m$ and $\mathscr{J}_m$ were defined, we have that the values of $\{\beta_r\}_{r=1}^{R_m}$ and $\{\gamma_r\}_{r=1}^{R_m}$ are all distinct, both within each sequence and across the two sequences. Moreover, since $\{\lambda_{m;n}\}_{n\in\mathscr{I}_m}$ and $\{\lambda_{m+1;n}\}_{n\in\mathscr{J}_m}$ are nonincreasing while $\pi_{\mathscr{I}_m}$ and $\pi_{\mathscr{J}_m}$ are increasing on $\mathscr{I}_m$ and $\mathscr{J}_m$ respectively, then the values $\{\beta_r\}_{r=1}^{R_m}$ and $\{\gamma_r\}_{r=1}^{R_m}$ are strictly decreasing. We further claim that $\{\gamma_r\}_{r=1}^{R_m}$ interlaces on $\{\beta_r\}_{r=1}^{R_m}$. To see this, consider the four polynomials:

$$\begin{aligned} p_m(x) &= \prod_{n=1}^{N}(x-\lambda_{m;n}), \qquad & p_{m+1}(x) &= \prod_{n=1}^{N}(x-\lambda_{m+1;n}),\\ b(x) &= \prod_{r=1}^{R_m}(x-\beta_r), & c(x) &= \prod_{r=1}^{R_m}(x-\gamma_r). \end{aligned} \tag{2.61}$$

Since $\{\beta_r\}_{r=1}^{R_m}$ and $\{\gamma_r\}_{r=1}^{R_m}$ were obtained by canceling the common terms from $\{\lambda_{m;n}\}_{n=1}^{N}$ and $\{\lambda_{m+1;n}\}_{n=1}^{N}$, we have that $p_{m+1}(x)/p_m(x) = c(x)/b(x)$ for all $x \notin \{\lambda_{m;n}\}_{n=1}^{N}$. Writing any $r = 1,\ldots,R_m$ as $r = \pi_{\mathscr{I}_m}(n)$ for some $n \in \mathscr{I}_m$, we have that since $\{\lambda_{m;n}\}_{n=1}^{N} \sqsubseteq \{\lambda_{m+1;n}\}_{n=1}^{N}$, applying the "only if" direction of Lemma 2.2 with "$p(x)$" and "$q(x)$" being $p_m(x)$ and $p_{m+1}(x)$ gives

$$\lim_{x\to\beta_r}(x-\beta_r)\frac{c(x)}{b(x)} = \lim_{x\to\lambda_{m;n}}(x-\lambda_{m;n})\frac{p_{m+1}(x)}{p_m(x)} \leq 0. \tag{2.62}$$

Since (2.62) holds for all $r = 1,\ldots,R_m$, applying the "if" direction of Lemma 2.2 with "$p(x)$" and "$q(x)$" being $b(x)$ and $c(x)$ gives that $\{\gamma_r\}_{r=1}^{R_m}$ interlaces on $\{\beta_r\}_{r=1}^{R_m}$.

Taken together, the facts that $\{\beta_r\}_{r=1}^{R_m}$ and $\{\gamma_r\}_{r=1}^{R_m}$ are distinct, strictly decreasing, and interlacing sequences implies that the $R_m \times 1$ vectors v_m and w_m are well defined. To be precise, Step B.3 may be rewritten as finding $v_m(r), w_m(r') \geq 0$ for all $r, r' = 1\ldots, R_m$ such that:

$$\big[v_m(r)\big]^2 = -\frac{\prod_{r''=1}^{R_m}(\beta_r-\gamma_{r''})}{\prod_{\substack{r''=1\\ r''\neq r}}^{R}(\beta_r-\beta_{r''})}, \qquad \big[w_m(r')\big]^2 = \frac{\prod_{r''=1}^{R_m}(\gamma_{r'}-\beta_{r''})}{\prod_{\substack{r''=1\\ r''\neq r'}}^{R}(\gamma_{r'}-\gamma_{r''})}. \tag{2.63}$$

Note that the fact that the β_r's and γ_r's are distinct implies that the denominators in (2.63) are nonzero, and moreover that the quotients themselves are nonzero. In fact, since $\{\beta_r\}_{r=1}^{R_m}$ is strictly decreasing, then for any fixed r, the values $\{\beta_r - \beta_{r''}\}_{r''\neq r}$

can be decomposed into $r-1$ negative values $\{\beta_r - \beta_{r''}\}_{r''=1}^{r-1}$ and $R_m - r$ positive values $\{\beta_r - \beta_{r''}\}_{r''=r+1}^{R_m}$. Moreover, since $\{\beta_r\}_{r=1}^{R_m} \sqsubseteq \{\gamma_r\}_{r=1}^{R_m}$, then for any such r, the values $\{\beta_r - \gamma_{r''}\}_{r''=1}^{R_m}$ can be broken into r negative values $\{\beta_r - \gamma_{r''}\}_{r''=1}^{r}$ and $R_m - r$ positive values $\{\beta_r - \gamma_{r''}\}_{r''=r+1}^{R_m}$. With the inclusion of an additional negative sign, we see that the quantity defining $[v_m(r)]^2$ in (2.63) is indeed positive. Meanwhile, the quantity defining $[w_m(r')]^2$ has exactly $r'-1$ negative values in both the numerator and denominator, namely $\{\gamma_{r'} - \beta_{r''}\}_{r''=1}^{r'-1}$ and $\{\gamma_{r'} - \gamma_{r''}\}_{r''=1}^{r'-1}$, respectively.

Having shown that the v_m and w_m of Step B.3 are well defined, we now take φ_{m+1} and U_{m+1} as defined in Steps B.4 and B.5. Recall that what remains to be shown in this direction of the proof is that U_{m+1} is a unitary matrix and that $\Phi_{m+1} = \{\varphi_{m'}\}_{m'=1}^{m+1}$ satisfies $\Phi_{m+1}\Phi_{m+1}^* U_{m+1} = U_{m+1} D_{m+1}$. To do so, consider the definition of U_{m+1} and recall that U_m is unitary by the inductive hypothesis, V_m is unitary by construction, and that the permutation matrices $\Pi_{\mathscr{I}_m}$ and $\Pi_{\mathscr{J}_m}$ are orthogonal, that is, unitary and real. As such, to show that U_{m+1} is unitary, it suffices to show that the $R_m \times R_m$ real matrix W_m is orthogonal. To do this, recall that eigenvectors corresponding to distinct eigenvalues of self-adjoint operators are necessarily orthogonal. As such, to show that W_m is orthogonal, it suffices to show that the columns of W_m are eigenvectors of a real symmetric operator. To this end, we claim:

$$\big(D_{m;\mathscr{I}_m} + v_m v_m^{\mathrm{T}}\big) W_m = W_m D_{m+1;\mathscr{J}_m}, \qquad W_m^{\mathrm{T}} W_m(r,r) = 1, \quad \forall r = 1, \dots, R_m, \tag{2.64}$$

where $D_{m;\mathscr{I}_m}$ and $D_{m+1;\mathscr{J}_m}$ are the $R_m \times R_m$ diagonal matrices whose rth diagonal entries are given by $\beta_r = \lambda_{m;\pi_{\mathscr{I}_m}^{-1}(r)}$ and $\gamma_r = \lambda_{m+1;\pi_{\mathscr{J}_m}^{-1}(r)}$, respectively. To prove (2.64), note that for any $r, r' = 1, \dots, R_m$,

$$\begin{aligned}\big[\big(D_{m;\mathscr{I}_m} + v_m v_m^{\mathrm{T}}\big) W_m\big](r,r') &= (D_{m;\mathscr{I}_m} W_m)(r,r') + \big(v_m v_m^{\mathrm{T}} W_m\big)(r,r') \\ &= \beta_r W_m(r,r') + v_m(r) \sum_{r''=1}^{R_m} v_m(r'') W_m(r'',r').\end{aligned} \tag{2.65}$$

Rewriting the definition of W_m from Step B.5 in terms of $\{\beta_r\}_{r=1}^{R_m}$ and $\{\gamma_r\}_{r=1}^{R_m}$ gives

$$W_m(r,r') = \frac{v_m(r) w_m(r')}{\gamma_{r'} - \beta_r}. \tag{2.66}$$

Substituting (2.66) into (2.65) gives

$$\begin{aligned}&\big[\big(D_{m;\mathscr{I}_m} + v_m v_m^{\mathrm{T}}\big) W_m\big](r,r') \\ &\quad = \beta_r \frac{v_m(r) w_m(r')}{\gamma_{r'} - \beta_r} + v_m(r) \sum_{r''=1}^{R_m} v_m(r'') \frac{v_m(r'') w_m(r')}{\gamma_{r'} - \beta_{r''}}\end{aligned}$$

$$= v_m(r) w_m(r') \left(\frac{\beta_r}{\gamma_{r'} - \beta_r} + \sum_{r''=1}^{R_m} \frac{[v_m(r'')]^2}{\gamma_{r'} - \beta_{r''}} \right). \tag{2.67}$$

Simplifying (2.67) requires a polynomial identity. To be precise, note that the difference $\prod_{r''=1}^{R_m}(x - \gamma_{r''}) - \prod_{r''=1}^{R_m}(x - \beta_{r''})$ of two monic polynomials is itself a polynomial of degree at most $R_m - 1$, and as such it can be written as the Lagrange interpolating polynomial determined by the R_m distinct points $\{\beta_r\}_{r=1}^{R_m}$:

$$\prod_{r''=1}^{R_m}(x - \gamma_{r''}) - \prod_{r''=1}^{R_m}(x - \beta_{r''}) = \sum_{r''=1}^{R_m} \left(\prod_{r=1}^{R_m}(\beta_{r''} - \gamma_r) - 0 \right) \prod_{\substack{r=1 \\ r \neq r''}}^{R_m} \frac{(x - \beta_r)}{(\beta_{r''} - \beta_r)}$$

$$= \sum_{r''=1}^{R_m} \frac{\prod_{r=1}^{R_m}(\beta_{r''} - \gamma_r)}{\prod_{\substack{r=1 \\ r \neq r''}}^{R_m}(\beta_{r''} - \beta_r)} \prod_{\substack{r=1 \\ r \neq r''}}^{R_m}(x - \beta_r). \tag{2.68}$$

Recalling the expression for $[v_m(r)]^2$ given in (2.63), (2.68) can be rewritten as

$$\prod_{r''=1}^{R_m}(x - \beta_{r''}) - \prod_{r''=1}^{R_m}(x - \gamma_{r''}) = \sum_{r''=1}^{R_m} [v_m(r'')]^2 \prod_{\substack{r=1 \\ r \neq r''}}^{R_m}(x - \beta_r). \tag{2.69}$$

Dividing both sides of (2.69) by $\prod_{r''=1}^{R_m}(x - \beta_{r''})$ gives

$$1 - \prod_{r''=1}^{R_m} \frac{(x - \gamma_{r''})}{(x - \beta_{r''})} = \sum_{r''=1}^{R_m} \frac{[v_m(r'')]^2}{(x - \beta_{r''})}, \quad \forall x \notin \{\beta_r\}_{r=1}^{R_m}. \tag{2.70}$$

For any $r' = 1, \ldots, R_m$, letting $x = \gamma_{r'}$ in (2.70) makes the left-hand product vanish, yielding the identity

$$1 = \sum_{r''=1}^{R_m} \frac{[v_m(r'')]^2}{(\gamma_{r'} - \beta_{r''})}, \quad \forall r' = 1, \ldots, R_m. \tag{2.71}$$

Substituting (2.71) into (2.67) and then recalling (2.66) gives

$$\begin{aligned} &\left[\left(D_{m;\mathscr{I}_m} + v_m v_m^{\mathrm{T}} \right) W_m \right](r, r') \\ &\quad = v_m(r) w_m(r') \left(\frac{\beta_r}{\gamma_{r'} - \beta_r} + 1 \right) \\ &\quad = \gamma_{r'} \frac{v_m(r) w_m(r')}{\gamma_{r'} - \beta_r} = \gamma_{r'} W_m(r, r') = (W_m D_{m+1;\mathscr{J}_m})(r, r'). \end{aligned} \tag{2.72}$$

As (2.72) holds for all $r, r' = 1, \ldots, R_m$ we have the first half of our claim (2.64). In particular, we know that the columns of W_m are eigenvectors of the real symmetric

operator $D_{m;\mathscr{I}_m} + v_m v_m^{\mathrm{T}}$ which correspond to the distinct eigenvalues $\{\gamma_r\}_{r=1}^{R_m}$. As such, the columns of W_m are orthogonal. To show that W_m is an orthogonal matrix, we must further show that the columns of W_m have unit norm, namely the second half of (2.64). To prove this, at any $x \notin \{\beta_r\}_{r=1}^{R_m}$ we differentiate both sides of (2.70) with respect to x to obtain

$$\sum_{r''=1}^{R_m}\left[\prod_{\substack{r=1\\ r\neq r''}}^{R_m}\frac{(x-\gamma_r)}{(x-\beta_r)}\right]\frac{\gamma_{r''}-\beta_{r''}}{(x-\beta_{r''})^2} = \sum_{r''=1}^{R_m}\frac{[v_m(r'')]^2}{(x-\beta_{r''})^2}, \quad \forall x \notin \{\beta_r\}_{r=1}^{R_m}. \tag{2.73}$$

For any $r' = 1, \ldots, R_m$, letting $x = \gamma_{r'}$ in (2.73) makes the left-hand summands where $r'' \neq r'$ vanish; by (2.63), the remaining summand where $r'' = r'$ can be written as

$$\begin{aligned}\frac{1}{[w_m(r')]^2} &= \frac{\prod_{r=1}^{R}(\gamma_{r'}-\gamma_r)}{\prod_{r=1}^{R_m}(\gamma_{r'}-\beta_r)} = \left[\prod_{\substack{r=1\\ r\neq r'}}^{R_m}\frac{(\gamma_{r'}-\gamma_r)}{(\gamma_{r'}-\beta_r)}\right]\frac{\gamma_{r'}-\beta_{r'}}{(\gamma_{r'}-\beta_{r'})^2}\\ &= \sum_{r''=1}^{R_m}\frac{[v_m(r'')]^2}{(\gamma_{r'}-\beta_{r''})^2}.\end{aligned} \tag{2.74}$$

We now use this identity to show that the columns of W_m have unit norm; for any $r' = 1, \ldots, R_m$, (2.66) and (2.74) give:

$$\begin{aligned}\big(W_m^{\mathrm{T}}W_m\big)(r',r') &= \sum_{r''=1}^{R_m}\big[W_m(r'',r')\big]^2 = \sum_{r''=1}^{R_m}\left(\frac{v_m(r'')w_m(r')}{\gamma_{r'}-\beta_{r''}}\right)^2\\ &= \big[w_m(r')\big]^2\sum_{r''=1}^{R_m}\frac{[v_m(r'')]^2}{(\gamma_{r'}-\beta_{r''})^2} = \big[w_m(r')\big]^2\frac{1}{[w_m(r')]^2} = 1.\end{aligned}$$

Having shown that W_m is orthogonal, we have that U_{m+1} is unitary. For this direction of the proof, all that remains to be shown is that $\Phi_{m+1}\Phi_{m+1}^* U_{m+1} = U_{m+1}D_{m+1}$. To do this, write $\Phi_{m+1}\Phi_{m+1}^* = \Phi_m\Phi_m^* + \varphi_{m+1}\varphi_{m+1}^*$ and recall the definition of U_{m+1}:

$$\begin{aligned}\Phi_{m+1}\Phi_{m+1}^*U_{m+1} &= \big(\Phi_m\Phi_m^* + \varphi_{m+1}\varphi_{m+1}^*\big)U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}}\begin{bmatrix}W_m & 0\\ 0 & Id\end{bmatrix}\Pi_{\mathscr{J}_m}\\ &= \Phi_m\Phi_m^*U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}}\begin{bmatrix}W_m & 0\\ 0 & Id\end{bmatrix}\Pi_{\mathscr{J}_m}\\ &\quad + \varphi_{m+1}\varphi_{m+1}^*U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}}\begin{bmatrix}W_m & 0\\ 0 & Id\end{bmatrix}\Pi_{\mathscr{J}_m}.\end{aligned} \tag{2.75}$$

To simplify the first term in (2.75), recall that the inductive hypothesis gives us that $\Phi_m\Phi_m^*U_m = U_m D_m$ and that V_m was constructed to satisfy $D_m V_m = V_m D_m$,

implying:

$$\begin{aligned}
&\Phi_m \Phi_m^* U_m V_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \begin{bmatrix} W_m & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{J}_m} \\
&\quad = U_m V_m D_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \begin{bmatrix} W_m & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{J}_m} \\
&\quad = U_m V_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \big(\Pi_{\mathscr{J}_m} D_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \big) \begin{bmatrix} W_m & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{J}_m}. \qquad (2.76)
\end{aligned}$$

To continue simplifying (2.76), note that $\Pi_{\mathscr{J}_m} D_m \Pi_{\mathscr{J}_m}^{\mathrm{T}}$ is itself a diagonal matrix: for any $n, n' = 1, \ldots, N$, the definition of $\Pi_{\mathscr{J}_m}$ given in Step B.2 gives

$$\big(\Pi_{\mathscr{J}_m} D_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \big)(n, n') = \langle D_m \delta_{\pi_{\mathscr{J}_m}^{-1}(n')}, \delta_{\pi_{\mathscr{J}_m}^{-1}(n)} \rangle = \begin{cases} \lambda_{m;\pi_{\mathscr{J}_m}^{-1}(n)}, & n = n', \\ 0, & n \neq n'. \end{cases}$$

That is, $\Pi_{\mathscr{J}_m} D_m \Pi_{\mathscr{J}_m}^{\mathrm{T}}$ is the diagonal matrix whose first R_m diagonal entries, namely $\{\beta_r\}_{r=1}^{R_m} = \{\lambda_{m;\pi_{\mathscr{J}_m}^{-1}(r)}\}_{r=1}^{R_m}$, match those of the aforementioned $R_m \times R_m$ diagonal matrix $D_{m;\mathscr{J}_m}$ and whose remaining $N - R_m$ diagonal entries $\{\lambda_{m;\pi_{\mathscr{J}_m}^{-1}(n)}\}_{n=R_m+1}^{N}$ form the diagonal of an $(N - R_m) \times (N - R_m)$ diagonal matrix $D_{m;\mathscr{J}_m^{\mathrm{c}}}$:

$$\Pi_{\mathscr{J}_m} D_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} = \begin{bmatrix} D_{m;\mathscr{J}_m} & 0 \\ 0 & D_{m;\mathscr{J}_m^{\mathrm{c}}} \end{bmatrix}. \qquad (2.77)$$

Substituting (2.77) into (2.76) gives:

$$\begin{aligned}
&\Phi_m \Phi_m^* U_m V_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \begin{bmatrix} W_m & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{J}_m} \\
&\quad = U_m V_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \begin{bmatrix} D_{m;\mathscr{J}_m} & 0 \\ 0 & D_{m;\mathscr{J}_m^{\mathrm{c}}} \end{bmatrix} \begin{bmatrix} W_m & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{J}_m} \\
&\quad = U_m V_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \begin{bmatrix} D_{m;\mathscr{J}_m} W_m & 0 \\ 0 & D_{m;\mathscr{J}_m^{\mathrm{c}}} \end{bmatrix} \Pi_{\mathscr{J}_m}. \qquad (2.78)
\end{aligned}$$

Meanwhile, to simplify the second term in (2.75), we recall the definition of φ_{m+1} from Step B.4:

$$\begin{aligned}
&\varphi_{m+1} \varphi_{m+1}^* U_m V_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \begin{bmatrix} W_m & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{J}_m} \\
&\quad = U_m V_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \begin{bmatrix} v_m \\ 0 \end{bmatrix} \begin{bmatrix} v_m^{\mathrm{T}} & 0 \end{bmatrix} \begin{bmatrix} W_m & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{J}_m} \\
&\quad = U_m V_m \Pi_{\mathscr{J}_m}^{\mathrm{T}} \begin{bmatrix} v_m v_m^{\mathrm{T}} W_m & 0 \\ 0 & 0 \end{bmatrix} \Pi_{\mathscr{J}_m}. \qquad (2.79)
\end{aligned}$$

Substituting (2.78) and (2.79) into (2.75), simplifying the result, and recalling (2.64) gives

$$\begin{aligned}\Phi_{m+1}\Phi_{m+1}^* U_{m+1} &= U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \begin{bmatrix} (D_{m;\mathscr{I}_m} + v_m v_m^{\mathrm{T}}) W_m & 0 \\ 0 & D_{m;\mathscr{I}_m^{\mathrm{c}}} \end{bmatrix} \Pi_{\mathscr{I}_m} \\ &= U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \begin{bmatrix} W_m D_{m+1;\mathscr{I}_m} & 0 \\ 0 & D_{m;\mathscr{I}_m^{\mathrm{c}}} \end{bmatrix} \Pi_{\mathscr{I}_m}.\end{aligned}$$

By introducing an extra permutation matrix and its inverse and recalling the definition of U_{m+1}, this simplifies to

$$\begin{aligned}&\Phi_{m+1}\Phi_{m+1}^* U_{m+1} \\ &\quad = U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \begin{bmatrix} W_m & 0 \\ 0 & Id \end{bmatrix} \Pi_{\mathscr{I}_m} \Pi_{\mathscr{I}_m}^{\mathrm{T}} \begin{bmatrix} D_{m+1;\mathscr{I}_m} & 0 \\ 0 & D_{m;\mathscr{I}_m^{\mathrm{c}}} \end{bmatrix} \Pi_{\mathscr{I}_m} \\ &\quad = U_{m+1} \Pi_{\mathscr{I}_m}^{\mathrm{T}} \begin{bmatrix} D_{m+1;\mathscr{I}_m} & 0 \\ 0 & D_{m;\mathscr{I}_m^{\mathrm{c}}} \end{bmatrix} \Pi_{\mathscr{I}_m}. \end{aligned} \tag{2.80}$$

We now partition the $\{\lambda_{m+1;n}\}_{n=1}^N$ of D_{m+1} into $\mathscr{I}_m$ and $\mathscr{I}_m^{\mathrm{c}}$ and mimic the derivation of (2.77), writing D_{m+1} in terms of $D_{m+1;\mathscr{I}_m}$ and $D_{m+1;\mathscr{I}_m^{\mathrm{c}}}$. Note here that by the manner in which $\mathscr{I}_m$ and $\mathscr{I}_m$ were constructed, the values of $\{\lambda_{m;n}\}_{n\in\mathscr{I}_m^{\mathrm{c}}}$ are equal to those of $\{\lambda_{m+1;n}\}_{\mathscr{I}_m^{\mathrm{c}}}$, as the two sets represent exactly those values which are common to both $\{\lambda_{m;n}\}_{n=1}^N$ and $\{\lambda_{m+1;n}\}_{n=1}^N$. As these two sequences are also both in nonincreasing order, we have $D_{m;\mathscr{I}_m^{\mathrm{c}}} = D_{m+1;\mathscr{I}_m^{\mathrm{c}}}$ and so

$$\Pi_{\mathscr{I}_m} D_{m+1} \Pi_{\mathscr{I}_m}^{\mathrm{T}} = \begin{bmatrix} D_{m+1;\mathscr{I}_m} & 0 \\ 0 & D_{m+1;\mathscr{I}_m^{\mathrm{c}}} \end{bmatrix} = \begin{bmatrix} D_{m+1;\mathscr{I}_m} & 0 \\ 0 & D_{m;\mathscr{I}_m^{\mathrm{c}}} \end{bmatrix}. \tag{2.81}$$

Substituting (2.81) into (2.80) yields $\Phi_{m+1}\Phi_{m+1}^* U_{m+1} = U_{m+1} D_{m+1}$, completing this direction of the proof.

($\Rightarrow$) Let $\{\lambda_n\}_{n=1}^N$ and $\{\mu_m\}_{m=1}^M$ be any nonnegative nonincreasing sequences, and let $\Phi = \{\varphi_m\}_{m=1}^M$ be any sequence of vectors whose frame operator $\Phi\Phi^*$ has $\{\lambda_n\}_{n=1}^N$ as its spectrum and has $\|\varphi_m\|^2 = \mu_m$ for all $m = 1, \ldots, M$. We will show that this Φ can be constructed by following Step A and Step B of this result. To see this, for any $m = 1, \ldots, M$, let $\Phi_m = \{\varphi_{m'}\}_{m'=1}^m$ and let $\{\lambda_{m;n}\}_{n=1}^N$ be the spectrum of the corresponding frame operator $\Phi_m\Phi_m^*$. Letting $\lambda_{0;n} := 0$ for all n, the proof of Theorem 2.2 demonstrated that the sequence of spectra $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$ necessarily forms a sequence of outer eigensteps as specified by Definition 2.2. This particular set of eigensteps is the one we choose in Step A.

All that remains to be shown is that we can produce our specific Φ by using Step B. Here, we must carefully exploit our freedom to pick U_1 and the V_m's; the proper choice of these unitary matrices will result in Φ, while other choices will produce other sequences of vectors that are only related to Φ through a potentially complicated series of rotations. Indeed, note that since $\{\{\lambda_{m;n}\}_{n=1}^N\}_{m=0}^M$ is a valid

sequence of eigensteps, then the other direction of this proof, as given earlier, implies that any choice of U_1 and V_m's will result in a sequence of vectors whose eigensteps match those of Φ. Moreover, quantities that we considered in the other direction of the proof that only depended on the choice of eigensteps, such as $\mathscr{I}_m$, $\mathscr{J}_m$, $\{\beta_r\}_{r=1}^{R_m}$, $\{\gamma_r\}_{r=1}^{R_m}$, etc., are thus also well defined in this direction; in the following arguments, we recall several such quantities and make further use of their previously derived properties.

To be precise, let U_1 be any one of the infinite number of unitary matrices whose first column $u_{1;1}$ satisfies $\varphi_1 = \sqrt{\mu_1}u_{1;1}$. We now proceed by induction, assuming that for any given $m = 1, \ldots, M-1$, we have followed Step B and have made appropriate choices for $\{V_{m'}\}_{m'=1}^{m-1}$ so as to correctly produce $\Phi_m = \{\varphi_{m'}\}_{m'=1}^{m}$; we show how the appropriate choice of V_m will correctly produce φ_{m+1}. To do so, we again write the mth spectrum $\{\lambda_{m;n}\}_{n=1}^{N}$ in terms of its multiplicities as $\{\lambda_{m;n(k,l)}\}_{k=1\ l=1}^{K_m\ L_{m;k}}$. For any $k = 1, \ldots, K_m$, Step B of Theorem 2.2 gives that the norm of the projection of φ_{m+1} onto the kth eigenspace of $\Phi_m\Phi_m^*$ is necessarily given by

$$\|P_{m;\lambda_{m;n(k,1)}}\varphi_{m+1}\|^2 = -\lim_{x\to\lambda_{m;n(k,1)}}(x-\lambda_{m;n(k,1)})\frac{p_{m+1}(x)}{p_m(x)}, \tag{2.82}$$

where $p_m(x)$ and $p_{m+1}(x)$ are defined by (2.61). Note that by picking $l = 1$, $\lambda_{m;n(k,1)}$ represents the first appearance of that particular value in $\{\lambda_{m;n}\}_{n=1}^{N}$. As such, these indices are the only ones that are eligible to be members of the set $\mathscr{I}_m$ found in Step B.2. That is, $\mathscr{I}_m \subseteq \{n(k,1) : k = 1, \ldots, K_m\}$. However, these two sets of indices are not necessarily equal, since $\mathscr{I}_m$ only contains n's of the form $n(k,1)$ that satisfy the additional property that the multiplicity of $\lambda_{m;n}$ as a value in $\{\lambda_{m;n'}\}_{n'=1}^{N}$ exceeds its multiplicity as a value in $\{\lambda_{m+1;n}\}_{n=1}^{N}$. To be precise, for any given $k = 1, \ldots, K_m$, if $n(k,1) \in \mathscr{I}_m^{\mathrm{c}}$, then $\lambda_{m;n(k,1)}$ appears as a root of $p_{m+1}(x)$ at least as many times as it appears as a root of $p_m(x)$, meaning in this case that the limit in (2.82) is necessarily zero. If, on the other hand, $n(k,1) \in \mathscr{I}_m$, then writing $\pi_{\mathscr{I}_m}(n(k,1))$ as some $r \in \{1, \ldots, R_m\}$ and recalling the definitions of $b(x)$ and $c(x)$ in (2.61) and $v(r)$ in (2.63), we can rewrite (2.82) as

$$\begin{aligned}\|P_{m;\beta_r}\varphi_{m+1}\|^2 &= -\lim_{x\to\beta_r}(x-\beta_r)\frac{p_{m+1}(x)}{p_m(x)}\\ &= -\lim_{x\to\beta_r}(x-\beta_r)\frac{c(x)}{b(x)} = -\frac{\prod_{r''=1}^{R_m}(\beta_r-\gamma_{r''})}{\prod_{\substack{r''=1\\ r''\neq r}}^{R}(\beta_r-\beta_{r''})} = \big[v_m(r)\big]^2.\end{aligned} \tag{2.83}$$

As such, we can write φ_{m+1} as

$$\begin{aligned}\varphi_{m+1} &= \sum_{k=1}^{K_m} P_{m;\lambda_{m;n(k,1)}}\varphi_{m+1} = \sum_{r=1}^{R_m} P_{m;\beta_r}\varphi_{m+1} = \sum_{r=1}^{R_m} v_m(r)\frac{1}{v_m(r)}P_{m;\beta_r}\varphi_{m+1}\\ &= \sum_{n\in\mathscr{I}_m} v_m\big(\pi_{\mathscr{I}_m}(n)\big)\frac{1}{v_m(\pi_{\mathscr{I}_m}(n))}P_{m;\beta_{\pi_{\mathscr{I}_m}(n)}}\varphi_{m+1}\end{aligned} \tag{2.84}$$

where each $\frac{1}{v_m(\pi_{\mathscr{I}_m}(n))} P_{m;\beta_{\pi_{\mathscr{I}_m}(n)}} \varphi_{m+1}$ has unit norm by (2.83). We now pick a new orthonormal eigenbasis $\hat{U}_m := \{\hat{u}_{m;n}\}_{n=1}^{N}$ for $\Phi_m \Phi_m^*$ that has the property that for any $k = 1, \ldots, K_m$, both $\{u_{m;n(k,l)}\}_{l=1}^{L_{m;k}}$ and $\{\hat{u}_{m;n(k,l)}\}_{l=1}^{L_{m;k}}$ span the same eigenspace and, for every $n(k,1) \in \mathscr{I}_m$, has the additional property that

$$\hat{u}_{m;n(k,1)} = \frac{1}{v_m(\pi_{\mathscr{I}_m}(n(k,1)))} P_{m;\beta_{\pi_{\mathscr{I}_m}(n(k,1))}} \varphi_{m+1}.$$

As such, (2.84) becomes

$$\begin{aligned} \varphi_{m+1} &= \sum_{n \in \mathscr{I}_m} v_m\big(\pi_{\mathscr{I}_m}(n)\big)\hat{u}_{m;n} = \hat{U}_m \sum_{n \in \mathscr{I}_m} v_m\big(\pi_{\mathscr{I}_m}(n)\big)\delta_n \\ &= \hat{U}_m \sum_{r=1}^{R_m} v_m(r)\delta_{\pi_{\mathscr{I}_m}^{-1}(r)} = \hat{U}_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \sum_{r=1}^{R_m} v_m(r)\delta_r = \hat{U}_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \begin{bmatrix} v_m \\ 0 \end{bmatrix}. \end{aligned} \tag{2.85}$$

Letting V_m be the unitary matrix $V_m = U_m^* \hat{U}_m$, the eigenspace spanning condition gives that V_m is block diagonal whose kth diagonal block is of size $L_{m;k} \times L_{m;k}$. Moreover, with this choice of V_m, (2.85) becomes

$$\varphi_{m+1} = U_m U_m^* \hat{U}_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \begin{bmatrix} v_m \\ 0 \end{bmatrix} = U_m V_m \Pi_{\mathscr{I}_m}^{\mathrm{T}} \begin{bmatrix} v_m \\ 0 \end{bmatrix}$$

meaning that φ_{m+1} can indeed be constructed by following Step B. □

Acknowledgements This work was supported by NSF DMS 1042701, NSF CCF 1017278, AFOSR F1ATA01103J001, AFOSR F1ATA00183G003, and the A.B. Krongard Fellowship. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, the U.S. Government, or Thomas Jefferson.

References

1. Antezana, J., Massey, P., Ruiz, M., Stojanoff, D.: The Schur-Horn theorem for operators and frames with prescribed norms and frame operator. Ill. J. Math. **51**, 537–560 (2007)
2. Batson, J., Spielman, D.A., Srivastava, N.: Twice-Ramanujan sparsifiers. In: Proc. STOC'09, pp. 255–262 (2009)
3. Benedetto, J.J., Fickus, M.: Finite normalized tight frames. Adv. Comput. Math. **18**, 357–385 (2003)
4. Bodmann, B.G., Casazza, P.G.: The road to equal-norm Parseval frames. J. Funct. Anal. **258**, 397–420 (2010)
5. Cahill, J., Fickus, M., Mixon, D.G., Poteet, M.J., Strawn, N.: Constructing finite frames of a given spectrum and set of lengths. Appl. Comput. Harmon. Anal. (submitted). arXiv: 1106.0921
6. Calderbank, R., Casazza, P.G., Heinecke, A., Kutyniok, G., Pezeshki, A.: Sparse fusion frames: existence and construction. Adv. Comput. Math. **35**, 1–31 (2011)

7. Casazza, P.G., Fickus, M., Heinecke, A., Wang, Y., Zhou, Z.: Spectral Tetris fusion frame constructions. J. Fourier Anal. Appl.
8. Casazza, P.G., Fickus, M., Kovačević, J., Leon, M.T., Tremain, J.C.: A physical interpretation of tight frames. In: Heil, C. (ed.) Harmonic Analysis and Applications: In Honor of John J. Benedetto, pp. 51–76. Birkhäuser, Boston (2006)
9. Casazza, P.G., Fickus, M., Mixon, D.G.: Auto-tuning unit norm tight frames. Appl. Comput. Harmon. Anal. **32**, 1–15 (2012)
10. Casazza, P.G., Fickus, M., Mixon, D.G., Wang, Y., Zhou, Z.: Constructing tight fusion frames. Appl. Comput. Harmon. Anal. **30**, 175–187 (2011)
11. Casazza, P.G., Heinecke, A., Krahmer, F., Kutyniok, G.: Optimally sparse frames. IEEE Trans. Inf. Theory **57**, 7279–7287 (2011)
12. Casazza, P.G., Kovačević, J.: Equal-norm tight frames with erasures. Adv. Comput. Math. **18**, 387–430 (2003)
13. Casazza, P.G., Leon, M.T.: Existence and construction of finite tight frames. J. Comput. Appl. Math. **4**, 277–289 (2006)
14. Chu, M.T.: Constructing a Hermitian matrix from its diagonal entries and eigenvalues. SIAM J. Matrix Anal. Appl. **16**, 207–217 (1995)
15. Dhillon, I.S., Heath, R.W., Sustik, M.A., Tropp, J.A.: Generalized finite algorithms for constructing Hermitian matrices with prescribed diagonal and spectrum. SIAM J. Matrix Anal. Appl. **27**, 61–71 (2005)
16. Dykema, K., Freeman, D., Kornelson, K., Larson, D., Ordower, M., Weber, E.: Ellipsoidal tight frames and projection decomposition of operators. Ill. J. Math. **48**, 477–489 (2004)
17. Dykema, K., Strawn, N.: Manifold structure of spaces of spherical tight frames. Int. J. Pure Appl. Math. **28**, 217–256 (2006)
18. Fickus, M., Mixon, D.G., Poteet, M.J.: Frame completions for optimally robust reconstruction. Proc. SPIE **8138**, 81380Q/1-8 (2011)
19. Fickus, M., Mixon, D.G., Poteet, M.J., Strawn, N.: Constructing all self-adjoint matrices with prescribed spectrum and diagonal (submitted). arXiv:1107.2173
20. Goyal, V.K., Kovačević, J., Kelner, J.A.: Quantized frame expansions with erasures. Appl. Comput. Harmon. Anal. **10**, 203–233 (2001)
21. Goyal, V.K., Vetterli, M., Thao, N.T.: Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory **44**, 16–31 (1998)
22. Higham, N.J.: Matrix nearness problems and applications. In: Gover, M.J.C., Barnett, S. (eds.) Applications of Matrix Theory, pp. 1–27. Oxford University Press, Oxford (1989)
23. Holmes, R.B., Paulsen, V.I.: Optimal frames for erasures. Linear Algebra Appl. **377**, 31–51 (2004)
24. Horn, A.: Doubly stochastic matrices and the diagonal of a rotation matrix. Am. J. Math. **76**, 620–630 (1954)
25. Horn, R.A., Johnson, C.R.: Matrix Analysis. Cambridge University Press, Cambridge (1985)
26. Kovačević, J., Chebira, A.: Life beyond bases: the advent of frames (Part I). IEEE Signal Process. Mag. **24**, 86–104 (2007)
27. Kovačević, J., Chebira, A.: Life beyond bases: the advent of frames (Part II). IEEE Signal Process. Mag. **24**, 115–125 (2007)
28. Massey, P., Ruiz, M.: Tight frame completions with prescribed norms. Sampl. Theory Signal. Image Process. **7**, 1–13 (2008)
29. Schur, I.: Über eine Klasse von Mittelbildungen mit Anwendungen auf die Determinantentheorie. Sitzungsber. Berl. Math. Ges. **22**, 9–20 (1923)
30. Strawn, N.: Finite frame varieties: nonsingular points, tangent spaces, and explicit local parameterizations. J. Fourier Anal. Appl. **17**, 821–853 (2011)
31. Tropp, J.A., Dhillon, I.S., Heath, R.W.: Finite-step algorithms for constructing optimal CDMA signature sequences. IEEE Trans. Inf. Theory **50**, 2916–2921 (2004)
32. Tropp, J.A., Dhillon, I.S., Heath, R.W., Strohmer, T.: Designing structured tight frames via an alternating projection method. IEEE Trans. Inf. Theory **51**, 188–209 (2005)

33. Viswanath, P., Anantharam, V.: Optimal sequences and sum capacity of synchronous CDMA systems. IEEE Trans. Inf. Theory **45**, 1984–1991 (1999)
34. Waldron, S.: Generalized Welch bound equality sequences are tight frames. IEEE Trans. Inf. Theory **49**, 2307–2309 (2003)
35. Welch, L.: Lower bounds on the maximum cross correlation of signals. IEEE Trans. Inf. Theory **20**, 397–399 (1974)

Chapter 3
Spanning and Independence Properties of Finite Frames

Peter G. Casazza and Darrin Speegle

Abstract The fundamental notion of frame theory is *redundancy*. It is this property which makes frames invaluable in so many diverse areas of research in mathematics, computer science, and engineering, because it allows accurate reconstruction after transmission losses, quantization, the introduction of additive noise, and a host of other problems. This issue also arises in a number of famous problems in pure mathematics such as the Bourgain-Tzafriri conjecture and its many equivalent formulations. As such, one of the most important problems in frame theory is to understand the spanning and independence properties of subsets of a frame. In particular, how many spanning sets does our frame contain? What is the smallest number of linearly independent subsets into which we can partition the frame? What is the least number of Riesz basic sequences that the frame contains with universal lower Riesz bounds? Can we partition a frame into subsets which are nearly tight? This last question is equivalent to the infamous Kadison–Singer problem. In this section we will present the state of the art on partitioning frames into linearly independent and spanning sets. A fundamental tool here is the famous Rado-Horn theorem. We will give a new recent proof of this result along with some nontrivial generalizations of the theorem.

Keywords Spanning sets · Independent sets · Redundancy · Riesz sequence · Rado-Horn theorem · Spark · Maximally robust · Matroid · K-ordering of dimensions

P.G. Casazza (✉)
Department of Mathematics, University of Missouri, Columbia, MO 65211, USA
e-mail: casazzap@missouri.edu

D. Speegle
Department of Mathematics and Computer Science, Saint Louis University, 221 N. Grand Blvd., St. Louis, MO 63103, USA
e-mail: speegled@slu.edu

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_3,

3.1 Introduction

The primary focus of this chapter is the independence and spanning properties of finite frames. More specifically, we will be looking at partitioning frames into sets $\{A_k\}_{k=1}^K$ which are linearly independent, spanning, or both. Since increasing the number of sets in the partition makes it easier for each set to be independent, and harder to span, we will be looking for the smallest K needed to be able to choose independent sets, and the largest K allowed so that we still have each set of vectors spanning. In order to fix notation, let $\Phi = (\varphi_i)_{i=1}^M$ be a set of vectors in $\mathcal{H}^N$, not necessarily a frame. It is clear from dimension counting that if A_i is linearly independent for each $1 \leq i \leq K$, then $K \geq \lceil M/N \rceil$. It is also clear from dimension counting that if A_i spans $\mathcal{H}^N$ for each $1 \leq i \leq K$, then $K \leq \lfloor M/N \rfloor$. So, in terms of linear independence and spanning properties, Φ is most "spread out" if it can be partitioned into $K = \lceil M/N \rceil$ linearly independent sets, $\lfloor M/N \rfloor$ of which are also spanning sets.

This important topic of spanning and independence properties of frames was not developed in frame theory until recently. In [9] we see the first results on decompositions of frames into linearly independent sets. Recently, a detailed study of spanning and independence properties of frames was made in [4]. Also, in [5] we see a new notion of *redundancy* for frames which connects the number of linearly independent and spanning sets of a frame of nonzero vectors $(\varphi_i)_{i=1}^M$ to the largest and smallest eigenvalues of the frame operator of the normalized frame $(\frac{\varphi_i}{\|\varphi_i\|})_{i=1}^M$. In this chapter we will discuss the state of the art on this topic and will also point out the remaining deep, important, open problems on this subject.

Spanning and independence properties of frames are related to several important themes in frame theory. First, a fundamental open problem in frame theory is the Kadison–Singer problem in the context of frame theory, which was originally called the Feichtinger conjecture [9, 10, 14, 16]. The Kadison–Singer problem asks whether for every frame $\Phi = (\varphi_i)_{i \in I}$, not necessarily finite, that is norm bounded below, there exists a finite partition $\{A_j : j = 1, \ldots, J\}$ such that for each $1 \leq j \leq J$, $(\varphi_i)_{i \in A_j}$ is a Riesz sequence. Since every Riesz sequence is, in particular, a linearly independent set, it is natural to study partitions of frames into linearly independent sets in order to better understand the Kadison–Singer problem in frame theory.

A second notion related to the spanning and independence properties of frames is that of redundancy. Frames are sometimes described as "redundant" bases, and a theme throughout frame theory is to make the notion of redundancy precise. Two properties that have been singled out as desirable properties of redundancy are: redundancy should measure the maximal number of disjoint spanning sets, and redundancy should measure the minimal number of disjoint linearly independent sets [5]. Of course, these two numbers are not usually the same, but nonetheless, describing in an efficient way the maximal number of spanning sets and the minimal number of linearly independent sets is a useful goal in quantifying the redundancy of a frame.

A third place where the spanning and independence properties of frames are vital, concerns *erasures*. During transmission, it is possible that frame coefficients are lost (erasures) or corrupted; then we have to try to do accurate reconstruction after losses

of frame coefficients. This can be done if the remaining frame vectors still span the space. So, for example, if a frame contains at least two spanning sets, then we can still do perfect reconstruction after the loss of one frame vector.

A fundamental tool for working with spanning and independence properties of frames is the celebrated *Rado-Horn theorem* [19, 22]. This theorem gives a necessary and sufficient condition for a frame to be partitioned into K disjoint linearly independent sets. The terminology *Rado-Horn theorem* was introduced in the paper [6]. The Rado-Horn theorem is a problem for frame theory in that it is impractical in applications. In particular, it requires doing a computation on every subset of the frame. What we want, is to be able to identify the minimal number of linearly independent sets into which we can partition a frame by using *properties of the frame* such as the eigenvalues of the frame operator, the norms of the frame vectors, etc. To do this, we will develop a sequence of deep refinements of the Rado-Horn theorem [5, 13] which are able to determine the number of linearly independent and spanning sets of a frame in terms of the properties of the frame. There are at least four proofs of the Rado-Horn theorem today [4, 13, 17, 19, 22]. The original proof is delicate, and the recent refinements [4, 13] are even more so. So we will develop these refinements slowly throughout various sections of this chapter to make this understandable.

Finally, let us recall that any frame $\Phi = (\varphi_i)_{i=1}^M$ with frame operator S is isomorphic to a Parseval frame $S^{-1/2}\Phi = (S^{-1/2}\varphi_i)_{i=1}^M$ and these two frames have the same linearly independent and spanning sets. So in our work we will mostly be working with Parseval frames.

3.1.1 Full Spark Frames

There is one class of frames for which the answers to our questions concerning the partition of the frame into independent and spanning sets are obvious. These are the *full spark frames.*

Definition 3.1 The spark of a frame $(\varphi_i)_{i=1}^M$ in $\mathcal{H}_N$ is the cardinality of the smallest linearly dependent subset of the frame. We say the frame is full spark if every N-element subset of the frame is linearly independent.

Full spark frames have appeared in the literature under the name *generic frames* [7] and *maximally robust to erasures* [11], since these frames have the property that the loss (erasure) of any $M - N$ of the frame elements still leaves a frame. For a full spark frame $(\varphi_i)_{i=1}^M$, any partition $\{A_j\}_{j=1}^K$ of $[1, M]$ into $K = \lceil \frac{M}{N} \rceil$ sets with $|A_j| = N$ for $j = 1, 2, \ldots, K-1$ and A_K the remaining elements has the property that $(\varphi_i)_{i\in A_k}$ is a linearly independent spanning set for all $1 \le k \le K$ and $(\varphi_i)_{i\in A_K}$ is linearly independent (and also spanning if $M = KN$).

It appears that full spark frames are quite specialized and perhaps do not occur very often. But, it is known that every frame is arbitrarily close to a full spark frame.

In [7] it is shown that this result holds even for Parseval frames. That is, the full spark frames are dense in the class of frames and the full spark Parseval frames are dense in the class of Parseval frames.

To prove these results, we do some preliminary work. For a frame $\Phi = (\varphi_i)_{i=1}^M$ with frame operator S, it is known that $(S^{-1/2}\varphi_i)_{i=1}^M$ is the closest Parseval frame to Φ [2, 3, 8, 12, 20]. Recall (see the Chap. 11) that a frame Φ for $\mathcal{H}^N$ is ϵ-nearly Parseval if the eigenvalues of the frame operator of the frame $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N$ satisfy $1 - \epsilon \leq \lambda_N \leq \lambda_1 \leq 1 + \epsilon$.

Proposition 3.1 *Let $(\varphi_i)_{i=1}^M$ be an ϵ-nearly Parseval frame for $\mathcal{H}^N$, with frame operator S. Then $(S^{-1/2}\varphi_i)_{i=1}^M$ is the closest Parseval frame to $(\varphi_i)_{i=1}^M$ and*

$$\sum_{i=1}^M \left\| S^{-1/2}\varphi_i - \varphi_i \right\|^2 \leq N\left(2 - \epsilon - 2\sqrt{1-\epsilon}\right) \leq N\frac{\epsilon^2}{4}.$$

Proof See the section on The Kadison–Singer and Paulsen Problems for a proof. □

We also need to check that a frame which is close to a Parseval frame is itself close to being Parseval.

Proposition 3.2 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a Parseval frame for $\mathcal{H}^N$ and let $\Psi = (\psi_i)_{i=1}^M$ be a frame for $\mathcal{H}^N$ satisfying*

$$\sum_{i=1}^M \|\varphi_i - \psi_i\|^2 < \epsilon < \frac{1}{9}.$$

Then Ψ is a $3\sqrt{\epsilon}$ nearly Parseval frame.

Proof Given $x \in \mathcal{H}^N$ we compute

$$\begin{aligned}\left(\sum_{i=1}^M \left|\langle x, \psi_i\rangle\right|^2\right)^{1/2} &\leq \left(\sum_{i=1}^M \left|\langle x, \varphi_i - \psi_i\rangle\right|^2\right)^{1/2} + \left(\sum_{i=1}^M \left|\langle x, \varphi_i\rangle\right|^2\right)^{1/2} \\ &\leq \|x\|\left(\sum_{i=1}^M \|\varphi_i - \psi_i\|^2\right)^{1/2} + \|x\| \\ &\leq \|x\|\left(1 + \sqrt{\epsilon}\right).\end{aligned}$$

The lower frame bound is similar. □

The final result needed is that if a Parseval frame Φ is close to a frame Ψ with frame operator S, then Φ is close to $S^{-1/2}\Psi$.

Proposition 3.3 *If $\Phi = (\varphi_i)_{i=1}^M$ is a Parseval frame for $\mathcal{H}^N$ and $\Psi = (\psi_i)_{i=1}^M$ is a frame with frame operator S satisfying*

$$\sum_{i=1}^{M} \|\varphi_i - \psi_i\|^2 < \epsilon < \frac{1}{9},$$

then

$$\sum_{i=1}^{M} \left\|\varphi_i - S^{-1/2}\psi_i\right\|^2 < 2\epsilon\left[1 + \frac{9}{4}N\right].$$

Proof We compute

$$\begin{aligned}\sum_{i=1}^{M} \left\|\varphi_i - S^{-1/2}\psi_i\right\|^2 &\le 2\left[\sum_{i=1}^{M} \|\varphi_i - \psi_i\|^2 + \sum_{i=1}^{M} \left\|\psi_i - S^{-1/2}\psi_i\right\|^2\right] \\ &\le 2\left[\epsilon + N\frac{(3\sqrt{\epsilon})^2}{4}\right] \\ &= 2\epsilon\left[1 + \frac{9}{4}N\right],\end{aligned}$$

where in the second inequality we applied Proposition 3.1 to the frame $(\psi_i)_{i=1}^M$ which is $3\sqrt{\epsilon}$ nearly Parseval by Proposition 3.2. □

Now we are ready for the main theorem. We will give a new elementary proof of this result.

Theorem 3.1 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a frame for $\mathcal{H}^N$ and let $\epsilon > 0$. Then there is a full spark frame $\Psi = (\psi_i)_{i=1}^M$ so that*

$$\|\varphi_i - \psi_i\| < \epsilon, \quad \text{for all } i = 1, 2, \ldots, M.$$

Moreover, if Φ is a Parseval frame, then Ψ may be chosen to be a Parseval frame.

Proof Since Φ must contain a linearly independent spanning set, we may assume that $(\varphi_i)_{i=1}^N$ is such a set. We let $\psi_i = \varphi_i$ for $i = 1, 2, \ldots, N$. The complement of the union of all hyperplanes spanned by subsets of $(\varphi_i)_{i=1}^N$ is open and dense in $\mathcal{H}^N$, and so there is a vector ψ_{N+1} in this open set with $\|\varphi_{N+1} - \psi_{N+1}\| < \epsilon$. By definition, $(\psi_i)_{i=1}^{N+1}$ is full spark. Now we continue this argument. The complement of the union of all hyperplanes spanned by subsets of $(\psi_i)_{i=1}^{N+1}$ is an open dense set in $\mathcal{H}^N$, and so we can choose a vector ψ_{N+2} from this set with $\|\varphi_{N+2} - \psi_{N+2}\| < \epsilon$. Again, by construction, $(\psi_i)_{i=1}^{N+2}$ is full spark. Iterating this argument we construct $(\psi_i)_{i=1}^M$.

For the *moreover* part, we choose $\delta > 0$ so that $\delta < \frac{1}{9}$ and

$$2\delta\left[1 + \frac{9}{4}N\right] < \epsilon^2.$$

By the first part of the theorem, we can choose a full spark frame $(\psi_i)_{i=1}^M$ so that

$$\sum_{i=1}^{M} \|\varphi_i - \psi_i\|^2 < \delta.$$

Letting S be the frame operator for $(\psi_i)_{i=1}^M$, we have that $(S^{-1/2}\psi_i)_{i=1}^M$ is a full spark frame, and by Proposition 3.3 we have that

$$\sum_{i=1}^{M} \left\|\varphi_i - S^{-1/2}\psi_i\right\|^2 < 2\delta\left[1 + \frac{9}{4}N\right] < \epsilon^2. \qquad \square$$

We end this section with an open problem.

Problem 3.1 If $(\varphi_i)_{i=1}^M$ is an equal norm Parseval frame for $\mathcal{H}^N$ and $\epsilon > 0$, is there a full spark equal norm Parseval frame $\Psi = (\psi_i)_{i=1}^M$ so that

$$\|\psi_i - \varphi_i\| < \epsilon, \quad \text{for all } i = 1, 2, \ldots, M?$$

We refer the reader to [1] for a discussion of this problem and its relationship to algebraic geometry.

3.2 Spanning and Independence Properties of Finite Frames

The main goal of this section is to show that equal norm Parseval frames of M vectors in $\mathcal{H}^N$ can be partitioned into $\lfloor M/N \rfloor$ bases and one additional set which is linearly independent. In particular, equal norm Parseval frames will contain $\lfloor M/N \rfloor$ spanning sets and $\lceil M/N \rceil$ linearly independent sets.

We begin by relating the algebraic properties of spanning and linear independence to the analytical properties of frames and Riesz sequences.

Proposition 3.4 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$. Then, Φ is a frame for $\mathcal{H}^N$ if and only if* span $\Phi = \mathcal{H}^N$.

Proof If Φ is a frame for $\mathcal{H}^N$ with frame operator S, then $A \cdot Id \leq S$ for some $0 < A$. So Φ must span $\mathcal{H}^N$.

The converse is a standard compactness argument. If Φ is not a frame, then there are vectors $x_n \in \mathcal{H}^N$ with $\|x_n\| = 1$ and satisfying

$$\sum_{i=1}^{M} \left|\langle x_n, \varphi_i\rangle\right|^2 \leq \frac{1}{n}, \quad \text{for all } n = 1, 2, \ldots.$$

Since we are in a finite-dimensional space, by switching to a subsequence of $\{x_n\}_{n=1}^\infty$ if necessary we may assume that $\lim_{n\to\infty} x_n = x \in \mathcal{H}^N$. Now we have

$$\sum_{i=1}^{M} \left|\langle x, \varphi_i\rangle\right|^2 \le 2\left[\sum_{i=1}^{M} \left|\langle x_n, \varphi_i\rangle\right|^2 + \sum_{i=1}^{M} \left|\langle x - x_n, \varphi_i\rangle\right|^2\right]$$
$$\le 2\left[\frac{1}{n} + \sum_{i=1}^{M} \|x - x_n\|^2 \|\varphi_i\|^2\right]$$
$$= 2\left[\frac{1}{n} + \|x - x_n\|^2 \sum_{i=1}^{M} \|\varphi_i\|^2\right].$$

As $n \to \infty$, the right-hand side of the above inequality goes to zero. Hence,

$$\sum_{i=1}^{M} \left|\langle x, \varphi_i\rangle\right|^2 = 0,$$

and so $x \perp \varphi_i$ for all $i = 1, 2, \ldots, M$. That is, Φ does not span $\mathcal{H}^N$. □

Proposition 3.5 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$. Then, Φ is linearly independent if and only if Φ is a Riesz sequence.*

Proof If Φ is a Riesz sequence, then there is a constant $0 < A$ so that for all scalars $\{a_i\}_{i=1}^M$ we have

$$A \sum_{i=1}^{M} |a_i|^2 \le \left\| \sum_{i=1}^{M} a_i \varphi_i \right\|^2.$$

Hence, if $\sum_{i=1}^M a_i \varphi_i = 0$, then $a_i = 0$ for all $i = 1, 2, \ldots, M$.

Conversely, if Φ is linearly independent, then (see the Introduction) the lower Riesz bound of Φ equals the lower frame bound and so Φ is a Riesz sequence. □

Notice that in the two propositions above, we do not say anything about the frame bounds or the Riesz bounds of the sets Φ. The following examples show that the lower frame bounds and Riesz bounds can be close to zero.

Example 3.1 Given $\epsilon > 0$, $N \in \mathbb{N}$, there is a linearly independent set containing N norm one vectors in $\mathcal{H}^N$ with lower frame bound (and hence lower Riesz bound) less than ϵ. To see this, let $(e_i)_{i=1}^N$ be an orthonormal basis for $\mathcal{H}^N$ and define a unit norm linearly independent set

$$\Phi = (\varphi_i)_{i=1}^N = \left(e_1, \frac{e_1 + \sqrt{\epsilon} e_2}{\sqrt{1+\epsilon}}, e_3, \ldots, e_N\right).$$

Now,

$$\sum_{i=1}^{N} \left|\langle e_2, \varphi_i\rangle\right|^2 = \frac{\epsilon}{1+\epsilon} < \epsilon.$$

3.2.1 Applications of the Rado-Horn Theorem I

Returning to the main theme of this chapter, we ask: When is it possible to partition a frame of M vectors for $\mathcal{H}^N$ into K linearly independent [resp., spanning] sets? The main combinatorial tool that we have to study this question is the Rado-Horn theorem.

Theorem 3.2 (Rado-Horn Theorem I) *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$ and $K \in \mathbb{N}$. There exists a partition $\{A_1, \ldots, A_K\}$ of $[1, M]$ such that for each $1 \leq k \leq K$, the set $(\varphi_i : i \in A_k)$ is linearly independent if and only if for every nonempty $J \subset [1, M]$,*

$$\frac{|J|}{\dim \operatorname{span}\{\varphi_i : i \in J\}} \leq K.$$

This theorem was proven in more general algebraic settings in [18, 19, 22], as well as later rediscovered in [17]. We delay the discussion of the proof of this theorem to Sect. 3.3. We content ourselves now with noting that the forward direction of the Rado-Horn Theorem I is essentially obvious. It says that in order to partition Φ into K linearly independent sets, there can not exist a subspace S which contains more than $K \dim(S)$ vectors. The reverse direction indicates that there are no obstructions to partitioning sets of vectors into linearly independent sets other than dimension counting obstructions.

We wish to use the Rado-Horn Theorem I to partition frames into linearly independent sets. Proposition 3.4 tells us that every spanning set is a frame, so it is clear that in order to get strong results we are going to need to make some assumptions about the frame. A natural extra condition is that of an *equal norm Parseval frame*. Intuitively, equal norm Parseval frames have no preferred directions, so it seems likely that one should be able to partition them into a small number of linearly independent sets. We will be able to do better than that; we will relate the minimum norm of the vectors in the Parseval frame to the number of linearly independent sets into which the frame can be partitioned.

Proposition 3.6 *Let $0 < C < 1$ and let Φ be a Parseval frame with M vectors for $\mathcal{H}^N$ such that $\|\varphi\|^2 \geq C$ for all $\varphi \in \Phi$. Then, Φ can be partitioned into $\lceil \frac{1}{C} \rceil$ linearly independent sets.*

Proof We show that the hypotheses of the Rado-Horn theorem are satisfied. Let $J \subset [1, M]$. Let $S = \operatorname{span}\{\varphi_j : j \in J\}$, and let P denote the orthogonal projection of $\mathcal{H}^N$ onto S. Since the orthogonal projection of a Parseval frame is again a Parseval frame and the sum of the norms squared of the vectors of the Parseval frame is the dimension of the space, we have

$$\begin{aligned} \dim S &= \sum_{j=1}^{M} \|P_S \varphi_j\|^2 \geq \sum_{j \in J} \|P_S \varphi_j\|^2 \\ &= \sum_{j \in J} \|\varphi_j\|^2 \geq |J| C. \end{aligned}$$

Therefore,

$$\frac{|J|}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq \frac{1}{C},$$

and Φ can be partitioned into $\lceil \frac{1}{C} \rceil$ linearly independent sets by the Rado-Horn theorem. □

We now present a trivial way of constructing an equal norm Parseval frame of M vectors for $\mathcal{H}^N$ when N divides M. Let $(e_i)_{i=1}^N$ be an orthonormal basis for $\mathcal{H}^N$ and let $\Phi = (Ce_1, \ldots, Ce_1, Ce_2, \ldots, Ce_2, \ldots, Ce_N, \ldots, Ce_N)$ be the orthonormal basis repeated M/N times, where $C = \sqrt{N/M}$. Then, it is easy to check that Φ is a Parseval frame. Another, slightly less trivial example is to union M/N orthonormal bases with no common elements and to normalize the vectors of the resulting set. In each of these cases, the Parseval frame can be trivially decomposed into M/N bases for $\mathcal{H}^N$. The following corollary can be seen as a partial converse.

Corollary 3.1 *If Φ is an equal norm Parseval frame of M vectors for $\mathcal{H}^N$, then Φ can be partitioned into $\lceil M/N \rceil$ linearly independent sets. In particular, if $M = kN$, then Φ can be partitioned into k Riesz bases.*

Proof This follows immediately from Proposition 3.6 and the fact that

$$\sum_{i=1}^{M} \|\varphi_i\|^2 = N,$$

which tells us that $\|\varphi_i\|^2 = M/N$ for all $i = 1, \ldots, M$. □

The argument above does not give any information about the lower Riesz bounds of the k Riesz bases we get in Corollary 3.1. Understanding these bounds is an exceptionally difficult problem and is equivalent to solving the Kadison–Singer problem (see the Chap. 11).

3.2.2 Applications of the Rado-Horn Theorem II

The Rado-Horn Theorem I has been generalized in several ways. In this section, we present the generalization to matroids and two applications of this generalization to partitioning into spanning and independent sets. We refer the reader to [21] for an introduction to matroid theory.

A *matroid* is a finite set X together with a collection $\mathcal{I}$ of subsets of X, which satisfies three properties:

1. $\emptyset \in \mathcal{I}$
2. if $I_1 \in \mathcal{I}$ and $I_2 \subset I_1$, then $I_2 \in \mathcal{I}$, and
3. if $I_1, I_2 \in \mathcal{I}$ and $|I_1| < |I_2|$, then there exists $x \in I_2 \setminus I_1$ such that $I_1 \cup \{x\} \in \mathcal{I}$.

Traditionally, the sets $I \in \mathcal{I}$ are called *independent* sets, which can lead to some confusion. For this chapter, we will use *linearly independent* to denote linear independence in the vector space sense, and *independent* to denote independence in the matroid sense. The rank of a set $E \subset X$ is defined to be the cardinality of a maximal independent (in the matroid sense) set contained in E.

There are many examples of matroids, but perhaps the most natural one comes from considering linear independence. Given a frame (or other finite collection of vectors) Φ in $\mathcal{H}^N$, define

$$\mathcal{I} = \{I \subset \Phi : I \text{ is linearly independent}\}.$$

It is easy to see that $(\Phi, \mathcal{I})$ is a matroid.

Another, slightly more involved example is to let X be a finite set which spans $\mathcal{H}^N$, and

$$\mathcal{I} = \left\{I \subset X : \text{span}(X \setminus I) = \mathcal{H}^N\right\}.$$

Then, in the definition of matroid, properties (1) and (2) are immediate. To see property (3), let I_1, I_2 be as in (3). We have that $\text{span}(X \setminus I_1) = \text{span}(X \setminus I_2) = \mathcal{H}^N$. Let $E_1 = X \setminus I_1$ and $E_2 = X \setminus I_2$; then, we have $|E_1| > |E_2|$. Find a basis G_1 for $\mathcal{H}^N$ by first taking a maximal linearly independent subset F of $E_1 \cap E_2$, and adding elements from E_1 to form a basis. Then find another basis G_2 for $\mathcal{H}^N$ by taking F and adding elements from E_2. Since $|E_1| > |E_2|$, there must be an element $x \in E_1 \setminus E_2$ which was not chosen to be in G_1. Note that $x \in I_2 \setminus I_1$, and $I_1 \cup \{x\} \in \mathcal{I}$, since $X \setminus (I_1 \cup \{x\})$ contains G_1, which is a basis. Another important source of examples is graph theory.

There is a natural generalization of the Rado-Horn theorem to the matroid setting.

Theorem 3.3 (Rado-Horn Theorem II) [18] *Let $(X, \mathcal{I})$ be a matroid, and let K be a positive integer. A set $J \subset X$ can be partitioned into K independent sets if and only if for every subset $E \subset J$,*

$$\frac{|E|}{\text{rank}(E)} \leq K. \tag{3.1}$$

We will be applying the matroid version of the Rado-Horn theorem to frames in Theorem 3.5 below, but first let us illustrate a more intuitive use. Consider the case of a collection Φ of M vectors where we wish to partition Φ into K linearly independent sets after discarding up to L vectors from Φ. It is natural to guess, based on our experience with the Rado-Horn theorem, that this is possible if and only if for every nonempty $J \subset [1, M]$

$$\frac{|J| - L}{\dim \text{span}\{\varphi_j : j \in J\}} \leq K.$$

However, it is not immediately obvious how to prove this from the statement of the Rado-Horn theorem. In the following theorem, we prove that, in some instances, the above conjecture is correct. Unfortunately, the general case will have to wait until we prove a different extension of the Rado-Horn theorem in Theorem 3.6.

Proposition 3.7 *Let Φ be a collection of M vectors in $\mathcal{H}^N$ and $K, L \in \mathbb{N}$. If there exists a set H with $|H| \leq L$ such that the set $\Phi \setminus H$ can be partitioned into K linearly independent sets, then for every nonempty $J \subset [1, M]$*

$$\frac{|J| - L}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K.$$

Proof If $J \subset [1, M] \setminus H$, then the Rado-Horn Theorem I implies

$$\frac{|J|}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K.$$

For general J with $|J| \geq L + 1$, notice that

$$\frac{|J| - L}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq \frac{|J \setminus H|}{\dim \operatorname{span}\{\varphi_j : j \in J \setminus H\}} \leq K,$$

as desired. □

Proposition 3.8 *Let Φ be a collection of M vectors in $\mathcal{H}^N$ indexed by $[1, M]$ and let $L \in \mathbb{N}$. Let $\mathcal{I} = \{I \subset [1, M]$: there exists a set $H \subset I$ with $|H| \leq L$ such that $I \setminus H$ is linearly independent$\}$. Then $(\Phi, \mathcal{I})$ is a matroid.*

Proof As usual, the first two properties of matroids are immediate. For the third property, let $I_1, I_2 \in \mathcal{I}$ with $|I_1| < |I_2|$. There exist H_1 and H_2 such that $I_j \setminus H_j$ is linearly independent and $|H_j| \leq L$ for $j = 1, 2$. If $|H_1|$ can be chosen so that $|H_1| < L$, then we can add any vector to I_1 and still have the new set linearly independent. If $|H_1|$ must be chosen to have cardinality L, then $|I_1 \setminus H_1| < |I_2 \setminus H_2|$ and both sets are linearly independent, so there is a vector $x \in (I_2 \setminus H_2) \setminus (I_1 \setminus H_1)$ so that $(I_1 \setminus H_1) \cup \{x\}$ is linearly independent. By the assumption that H_1 must be chosen to have cardinality L, $x \notin H_1$. Therefore, $x \notin I_1$ and $I_1 \cup \{x\} \in \mathcal{I}$, as desired. □

Theorem 3.4 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a collection of M vectors in $\mathcal{H}^N$. Let $K, L \in \mathbb{N}$. There exists a set H with $|H| \leq LK$ such that the set $\Phi \setminus H$ can be partitioned into K linearly independent sets if and only if, for every nonempty $J \subset [1, M]$,*

$$\frac{|J| - LK}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K.$$

Proof The forward direction is a special case of Proposition 3.7. For the reverse direction, define the matroid $(\Phi, \mathcal{I})$ as in Proposition 3.8. By the matroid version of the Rado-Horn theorem, we can partition Φ into K independent sets if and only if, for every nonempty $J \subset [1, M]$,

$$\frac{|J|}{\operatorname{rank}(\{\varphi_j : j \in J\})} \leq K.$$

We now show that this follows if, for every nonempty $J \subset [1, M]$,

$$\frac{|J| - LK}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K.$$

Suppose we have for every nonempty $J \subset [1, M]$,

$$\frac{|J| - LK}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K.$$

Let $J \subset [1, M]$. Note that if we can remove fewer than L vectors from $(\varphi_j)_{j\in J}$ to form a linearly independent set, then $\operatorname{rank}(\{\varphi_j : j \in J\}) = |J|$, so

$$\frac{|J|}{\operatorname{rank}(\{\varphi_j : j \in J\})} = 1 \leq K.$$

On the other hand, if we need to remove at least L vectors from $(\varphi_j)_{j\in J}$ to form a linearly independent set, then $\operatorname{rank}(\{\varphi_j : j \in J\}) = \dim \operatorname{span}\{\varphi_j : j \in J\} + L$, so

$$\begin{aligned} |J| &\leq K \dim \operatorname{span}\{\varphi_j : j \in J\} + LK \\ &= K \operatorname{rank}\big(\{\varphi_j : j \in J\}\big), \end{aligned}$$

as desired. Therefore, if for every $J \subset [1, M]$,

$$\frac{|J| - LK}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K,$$

then there is a partition $\{A_i\}_{i=1}^K$ of $[1, M]$ such that $(\varphi_j : j \in A_i) \in \mathcal{I}$ for each $1 \leq i \leq K$. By the definition of our matroid, for each $1 \leq i \leq K$, there exists $H_i \subset A_i$ with $|H_i| \leq L$ such that $(\varphi_j : j \in A_i \setminus H_i)$ is linearly independent. Let $H = \bigcup_{i=1}^K H_i$ and note that $|H| \leq LK$ and $J \setminus H$ can be partitioned into K linearly independent sets. □

The matroid version of the Rado-Horn Theorem will be applied to finite frames in the following theorem.

Theorem 3.5 *Let $\delta > 0$. Suppose that $\Phi = (\varphi_i)_{i=1}^M$ is a Parseval frame of M vectors for $\mathcal{H}^N$ with $\|\varphi_i\|^2 \leq 1 - \delta$ for all $\varphi \in \Phi$. Let $R \in \mathbb{N}$ such that $R \geq \frac{1}{\delta}$. Then, it is possible to partition $[1, M]$ into R sets $\{A_1, \ldots, A_R\}$ such that, for each $1 \leq r \leq R$, the family $(\varphi_j : j \notin A_r)$ spans $\mathcal{H}^N$.*

Proof Let $\mathcal{I} = \{E \subset [1, M] : \operatorname{span}\{\varphi_j : j \notin E\} = \mathcal{H}^N\}$. Since any frame is a spanning set, we have that $([1, M], \mathcal{I})$ is a matroid. By the Rado-Horn Theorem II, it suffices to show (3.1) for each subset of $[1, M]$. Let $E \subset [1, M]$. Define $S = \operatorname{span}\{\varphi_j : j \notin E\}$, and let P be the orthogonal projection onto $S^\perp$. Since the orthogonal projection of a Parseval frame is again a Parseval frame, we have that $(P\varphi : \varphi \in \Phi)$ is a Parseval frame for $S^\perp$. Moreover, we have

$$\begin{aligned} \dim S^\perp &= \sum_{j=1}^M \|P\varphi_j\|^2 = \sum_{j\in E} \|P\varphi_j\|^2 \\ &\leq |E|(1 - \delta). \end{aligned}$$

Let M be the largest integer smaller than or equal to $|E|(1-\delta)$. Since $\dim S^\perp \leq M$, we have that there exists a set $E_1 \subset E$ such that $|E_1| = M$ and $\operatorname{span}\{P\varphi_j : j \in E_1\} =$

$S^\perp$. Let $E_2 = E \setminus E_1$. We show that E_2 is independent. For this, write $h \in \mathcal{H}^N$ as $h = h_1 + h_2$, where $h_1 \in S$ and $h_2 \in S^\perp$. We have that $h_2 = \sum_{j\in E_1} \alpha_j Pf\varphi_j$ for some choice of $\{\alpha_j : j \in E_1\}$. Write $\sum_{j\in E_1} \alpha_j\varphi_j = g_1 + h_2$, where $g_1 \in S$. Then, there exist $\{\alpha_j : j \notin E\}$ such that $\sum_{j\notin E} \alpha_j\varphi_j = h_1 - g_1$. So,

$$\sum_{j\notin E_2} \alpha_j\varphi_j = h,$$

and thus E_2 is independent.

Now, since E contains an independent set of cardinality $|E| - M$, it follows that $\text{rank}(E) \geq |E| - M \geq |E| - |E|(1-\delta) = \delta|E|$. Therefore,

$$\frac{|E|}{\text{rank}(E)} \leq \frac{1}{\delta} \leq R,$$

as desired. □

3.2.3 Applications of the Rado-Horn Theorem III

Up to this point, we have mostly focused on linear independence properties of frames. We now turn to spanning properties. We present a more general form of the Rado-Horn theorem, which describes what happens when the vectors cannot be partitioned into linearly independent sets.

The worst possible blockage that can occur preventing us from partitioning a frame $(\varphi_i)_{i=1}^M$ into K linearly independent sets would be the case where there are disjoint subsets (not necessarily a partition) $\{A_k\}_{k=1}^K$ of $[1, M]$ with the property

$$\text{span}(\varphi_i)_{i\in A_j} = \text{span}(\varphi_i)_{i\in A_k}, \quad \text{for all } 1 \leq j, k \leq K.$$

The following improvement of the Rado-Horn theorem shows the surprising fact that this is really the only blockage that can occur.

Theorem 3.6 (Rado-Horn Theorem III) *Let $\Phi = (\varphi_i)_{i=1}^M$ be a collection of vectors in $\mathcal{H}^N$ and $K \in \mathbb{N}$. Then the following conditions are equivalent.*

(1) *There exists a partition $\{A_k : k = 1, \ldots, K\}$ of $[1, M]$ such that for each $1 \leq k \leq K$ the set $\{\varphi_j : j \in A_k\}$ is linearly independent.*
(2) *For all $J \subset I$,*

$$\frac{|J|}{\dim \text{span}\{\varphi_j : j \in J\}} \leq K. \tag{3.2}$$

Moreover, in the case that either of the conditions above fails, there exists a partition $\{A_k : k = 1, \ldots, K\}$ of $[1, M]$ and a subspace S of $\mathcal{H}^N$ such that the following three conditions hold.

(a) *For all $1 \leq k \leq K$, $S = \text{span}\{\varphi_j : j \in A_k \text{ and } \varphi_j \in S\}$.*
(b) *For $J = \{i \in I : \varphi_i \in S\}$, $\frac{|J|}{\dim \text{span}(\{\varphi_i : i\in J\})} > K$.*

(c) *For each* $1 \leq k \leq K$, $\{P_{S^\perp}\varphi_i : i \in A_k, \varphi_i \notin S\}$ *is linearly independent, where* $P_{S^\perp}$ *is the orthogonal projection onto* $S^\perp$.

For the purposes of this chapter, we are restricting to $\mathcal{H}^N$, but the result also holds with a slightly different statement for general vector spaces; see [13] for details.

The statement of Theorem 3.6 is somewhat involved, and the proof even more so, so we delay the proof until Sect. 3.4. For now, we show how Theorem 3.6 can be applied in two different cases. For our first application, we will provide a proof of Theorem 3.4 in the general case.

Theorem 3.7 *Let* $\Phi = (\varphi_i)_{i=1}^M$ *be a collection of* M *vectors in* $\mathcal{H}^N$. *Let* $K, L \in \mathbb{N}$. *There exists a set* H *with* $|H| \leq L$ *such that the set* $\Phi \setminus H$ *can be partitioned into* K *linearly independent sets if and only if, for every nonempty* $J \subset [1, M]$,

$$\frac{|J| - L}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K.$$

Proof The forward direction is Proposition 3.7. For the reverse direction, if Φ can be partitioned into K linearly independent sets, then we are done. Otherwise, we can apply the alternative in Theorem 3.6 to obtain a partition $\{A_k : 1 \leq k \leq K\}$ and a subspace S satisfying the properties listed.

For $1 \leq k \leq K$, let $A_k^1 = \{j \in A_k : \varphi_j \in S\}$, and $A_k^2 = A_k \setminus A_k^1 = \{j \in A_k : \varphi_j \notin S\}$. For each $1 \leq k \leq K$, let $B_k \subset A_k^1$ be defined such that $(\varphi_j : j \in B_k)$ is a basis for S, which is possible by property (a) in Theorem 3.6. Letting $J = \bigcup_{k=1}^K A_k^1$ and applying

$$\frac{|J| - L}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K$$

yields that there are at most L vectors in J which are not in one of the B_k's. Let $H = J \setminus \bigcup_{k=1}^K B_k$. Since $|H| \leq L$, it suffices to show that letting $C_k = B_k \cup A_k^2$ partitions $[1, M] \setminus H$ into linearly independent sets.

Indeed, fix k and assume that $\sum_{j \in C_k} a_k \varphi_k = 0$. Then

$$\begin{aligned} 0 &= \sum_{j \in C_k} a_k P_{S^\perp} \varphi_j \\ &= \sum_{j \in A_k^2} a_k P_{S^\perp} \varphi_j. \end{aligned}$$

So $a_k = 0$ for all $k \in A_k^2$ by property (c) in Theorem 3.6. This implies that

$$\begin{aligned} 0 &= \sum_{j \in C_k} a_k \varphi_j \\ &= \sum_{j \in B_k} a_k \varphi_j, \end{aligned}$$

and so $a_k = 0$ for all $k \in B_k$. Therefore, $\{C_k\}$ is a partition of $[1, M] \setminus H$ such that for each $1 \leq k \leq K$, the set $(\varphi_j : j \in C_k)$ is linearly independent. □

We now present an application that is more directly related to frame theory. This theorem will be combined with Theorem 3.10 to prove Lemma 3.2.

Theorem 3.8 *Let $\Phi = (\varphi_i)_{i=1}^M$ be an equal norm Parseval frame for $\mathcal{H}^N$. Let $K = \lfloor M/N \rfloor$. Then there exists a partition $\{A_k\}_{k=1}^K$ of $[1, M]$ so that*

$$\text{span}\,\{\varphi_i : i \in A_j\} = \mathcal{H}^N, \quad \text{for all } j = 1, 2, \ldots, K.$$

Our method of proof of Theorem 3.8 involves induction on the dimension N. In order to apply the induction step, we will project onto a subspace, which, while it preserves the Parseval frame property, does not preserve equal norm of the vectors. For this reason, we state a more general theorem that is more amenable to an induction proof.

Theorem 3.9 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a frame for $\mathcal{H}^N$ with lower frame bound $A \geq 1$, let $\|\varphi_i\|^2 \leq 1$ for all $i \in [1, M]$, and set $K = \lfloor A \rfloor$. Then there exists a partition $\{A_k\}_{k=1}^K$ of $[1, M]$ so that*

$$\text{span}\,\{\varphi_i : i \in A_k\} = \mathcal{H}^N, \quad \text{for all } k = 1, 2, \ldots, K.$$

In particular, the number of frame vectors in a unit norm frame with lower frame bound A is greater than or equal to $\lfloor A \rfloor N$.

We will need the following lemma, which we state without proof.

Lemma 3.1 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a collection of vectors in $\mathcal{H}^N$ and let $I_k \subset [1, M]$, $k = 1, 2, \ldots, K$ be a partition of Φ into linearly independent sets. Assume that there is a partition of $[1, M]$ into $\{A_k\}_{k=1}^K$ so that*

$$\text{span}\,(\varphi_i)_{i \in A_k} = \mathcal{H}^N, \quad \text{for all } k = 1, 2, \ldots, K.$$

Then,

$$\text{span}\,\{\varphi_i\}_{i \in I_k} = \mathcal{H}^N, \quad \text{for all } k = 1, 2, \ldots, K.$$

Proof of Theorem 3.9 We replace $(\varphi_i)_{i=1}^M$ by $(\frac{1}{\sqrt{K}}\varphi_i)_{i=1}^M$ so that our frame has lower frame bound greater than or equal to 1 and $\|\varphi_i\|^2 \leq \frac{1}{K}$, for all $i \in [1, M]$. Assume the frame operator for $(\varphi_i)_{i=1}^M$ has eigenvectors $(e_j)_{j=1}^N$ with respective eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 1$. We proceed by induction on N.

We first consider $N = 1$: Since

$$\sum_{i=1}^M \|\varphi_i\|^2 \geq 1, \quad \text{and} \quad \|\varphi_i\|^2 \leq \frac{1}{K}, \tag{3.3}$$

it follows that $|\{i \in I : \varphi_i \neq 0\}| \geq K$ and so we have a partition of the frame into K spanning sets.

Next, we assume the induction hypothesis holds for any Hilbert space of dimension N and let $\mathcal{H}^{N+1}$ be a Hilbert space of dimension $N + 1$. We check two cases.

Case I: Suppose there exists a partition $\{A_k\}_{k=1}^K$ of $[1, M]$ so that $(\varphi_i)_{i\in A_k}$ is linearly independent for all $k = 1, 2, \ldots, K$. In this case,

$$N + 1 \leq (N + 1)\lambda_N \leq \sum_{j=1}^{N+1} \lambda_j = \sum_{i=1}^{M} \|\varphi_i\|^2 \leq M\frac{1}{K},$$

and hence,

$$M \geq K(N + 1).$$

However, by linear independence, we have

$$M = \sum_{k=1}^{K} |A_k| \leq K(N + 1).$$

Thus, $|A_k| = N + 1$ for every $k = 1, 2, \ldots, K$ and so $(\varphi_i)_{i\in A_k}$ is spanning for $1 \leq k \leq K$.

Case II: Suppose $(\varphi_i)_{i=1}^M$ cannot be partitioned into K linearly independent sets. In this case, let $\{A_k\}_{k=1}^K$ and a subspace $\emptyset \neq S \subset \mathcal{H}^{N+1}$ be given by Theorem 3.6. If $S = \mathcal{H}^{N+1}$, we are done. Otherwise, let P be the orthogonal projection onto the subspace S. Let

$$A_k' = \{i \in A_k : \varphi_i \notin S\}, \qquad B = \bigcup_{k=1}^{K} A_k'.$$

By Theorem 3.6(c), $((Id - P)\varphi_i)_{i\in A_k'}$ is linearly independent for all $k = 1, 2, \ldots, K$.

Now, $((Id - P)\varphi_i)_{i\in B}$ has lower frame bound 1 in $(Id - P)(\mathcal{H}^{N+1})$, $\dim(Id - P)(\mathcal{H}^{N+1}) \leq N$ and

$$\left\|(Id - P)\varphi_i\right\|^2 \leq \|\varphi_i\|^2 \leq \frac{1}{K}$$

for all $i \in B$. Applying the induction hypothesis, we can find a partition $\{B_k\}_{k=1}^K$ of B with $\text{span}((Id - P)\varphi_i)_{i\in B_k} = (Id - P)(\mathcal{H}^{N+1})$ for all $k = 1, 2, \ldots, K$. Now, we can apply Lemma 3.1 together with the partition $\{B_k\}_{k=1}^K$ to conclude span $((Id - P)\varphi_i)_{i\in A_k'} = (Id - P)(\mathcal{H}^{N+1})$, and hence

$$\text{span}(\varphi_i)_{i\in A_k} = \text{span}\left\{S, \left((Id - P)\varphi_i\right)_{i\in A_k'}\right\} = \mathcal{H}^{N+1}. \qquad \square$$

Up to this point, we have seen that an equal norm Parseval frame with M vectors in $\mathcal{H}^N$ can be partitioned into $\lfloor M/N \rfloor$ spanning sets and $\lceil M/N \rceil$ linearly independent sets. We now show that there is a single partition which accomplishes both the spanning and linear independence properties.

Theorem 3.10 *Let $\Phi = (\varphi_i)_{i=1}^M$ be an equal norm Parseval frame for $\mathcal{H}^N$ and let $K = \lceil M/N \rceil$. There exists a partition $\{A_k\}_{k=1}^K$ of $[1, M]$ such that*

1. *$(\varphi_i : i \in A_k)$ is linearly independent for $1 \leq k \leq K$, and*
2. *$(\varphi_i : i \in A_k)$ spans $\mathcal{H}^N$ for $1 \leq k \leq K - 1$.*

The proof of Theorem 3.10 is immediate from Corollary 3.1, Theorem 3.8, and Lemma 3.2 below.

Lemma 3.2 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a finite collection of vectors in $\mathcal{H}^N$ and let $K \in \mathbb{N}$. Assume*

1. *Φ can be partitioned into $K+1$-linearly independent sets, and*
2. *Φ can be partitioned into a set and K spanning sets.*

Then there is a partition $\{A_k\}_{k=1}^{K+1}$ so that $(\varphi_j)_{j\in A_k}$ is a linearly independent spanning set for all $k = 2, 3, \ldots, K+1$ and $(\varphi_i)_{i\in A_1}$ is a linearly independent set.

The proof of Lemma 3.2 requires yet another extension of the Rado-Horn theorem, which we have not yet discussed and will be proven at the end of Sect. 3.4.

3.3 The Rado-Horn Theorem I and Its Proof

In this and the following sections, we discuss the proofs of the Rado-Horn Theorems I and III. Although the forward direction is essentially obvious, the reverse direction of the Rado-Horn Theorem I, while elementary, is not simple to prove. Our present goal is a proof of the case $K = 2$, which contains many of the essential ideas of the general proof without some of the bookkeeping difficulties in the general case. The proof of the general case of the Rado-Horn Theorem III will be presented below, and it contains a proof of the Rado-Horn Theorem I. The main idea for the reverse direction is to take as a candidate partition one that maximizes the sum of the dimensions associated with the partition. Then, if that does not partition the set into linearly independent subsets, one can construct a set of interconnected linearly dependent vectors which directly contradicts the hypotheses of the Rado-Horn Theorem I.

As mentioned above, the forward direction of the Rado-Horn Theorem I is essentially obvious, but we provide a formal proof in the following lemma.

Lemma 3.3 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$ and $K \in \mathbb{N}$. If there exists a partition $\{A_1, \ldots, A_K\}$ of $[1, M]$ such that, for each $1 \le k \le K$, $(\varphi_i : i \in A_k)$ is linearly independent, then for every nonempty $J \subset [1, M]$,*

$$\frac{|J|}{\dim \operatorname{span}\{\varphi_i : i \in J\}} \le K.$$

Proof Let $\{A_1, \ldots, A_K\}$ partition Φ into linearly independent sets. Let J be a nonempty subset of $[1, M]$. For each $1 \le k \le K$, let $J_k = J \cap A_k$. Then,

$$|J| = \sum_{k=1}^{K} |J_k| = \sum_{k=1}^{K} \dim \operatorname{span}\big(\{\varphi_i : i \in J_k\}\big) \le K \dim \operatorname{span}\big(\{\varphi_i : i \in J\}\big),$$

as desired. □

The Rado-Horn Theorem I tells us that if we want to partition vectors into K linearly independent subsets, there are no nontrivial obstructions. The only obstruction is that there cannot be a subspace S which contains more than $K\dim(S)$ of the vectors that we wish to partition.

The first obstacle to proving the Rado-Horn Theorem I is coming up with a candidate partition which should be linearly independent. There are several ways to do this. The most common, used in [17–19, 22], is to build the partition while proving the theorem. In [13], it was noticed that any partition which maximizes the sums of dimensions (as explained below) must partition Φ into linearly independent sets, provided any partition can do so. Given a set $\Phi \subset \mathcal{H}^N$ indexed by $[1, M]$ and a natural number K, we say that a partition $\{A_1, \ldots, A_K\}$ of $[1, M]$ *maximizes the K-sum of dimensions* of Φ if, for any partition $\{B_1, \ldots, B_K\}$ of $[1, M]$,

$$\sum_{k=1}^{K} \dim \operatorname{span}\{\varphi_j : j \in A_k\} \geq \sum_{k=1}^{K} \dim \operatorname{span}\{\varphi_j : j \in B_k\}.$$

There are two things to notice about a partition $\{A_1, \ldots, A_K\}$ which maximizes the K-sum of dimensions. First, such a partition will always exist since we are dealing with finite sets. Second, such a partition will partition Φ into K linearly independent sets if it is possible for any partition to do so. That is the content of the next two propositions.

Proposition 3.9 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$, $K \in \mathbb{N}$, and $\{A_k\}_{k=1}^K$ be a partition of $[1, M]$. The following conditions are equivalent.*

(1) *For every $k \in \{1, \ldots, K\}$, $(\varphi_j : j \in A_k)$ is linearly independent.*
(2) $\sum_{k=1}^{K} \dim \operatorname{span}\{\varphi_j : j \in A_k\} = M$.

Proof (1) $\Rightarrow$ (2) Clearly,

$$\sum_{k=1}^{K} \dim \operatorname{span}\{\varphi_j : j \in A_k\} = \sum_{k=1}^{K} |A_k| = M.$$

(2) $\Rightarrow$ (1) Note that

$$M = \sum_{k=1}^{K} \dim \operatorname{span}\{\varphi_j : j \in A_k\} \leq \sum_{k=1}^{K} |A_k| = M.$$

Therefore, $\dim \operatorname{span}\{\varphi_j : j \in A_k\} = |A_k|$ for each $1 \leq k \leq K$ and $(\varphi_j : j \in A_k)$ is linearly independent. □

Proposition 3.10 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$ and $K \in \mathbb{N}$. If $\{A_k\}_{k=1}^K$ maximizes the K-sum of dimensions of Φ and there exists a partition $\{B_k\}_{k=1}^K$ such that for each $1 \leq k \leq K$, $(\varphi_j : j \in B_k)$ is linearly independent, then $(\varphi_j : j \in A_k)$ is linearly independent for each $1 \leq k \leq K$.*

Proof We have

$$\begin{aligned} M &= \sum_{k=1}^{K} \dim \operatorname{span}\{\varphi_j : j \in B_k\} \\ &\leq \sum_{k=1}^{K} \dim \operatorname{span}\{\varphi_j : j \in A_k\} \leq M. \end{aligned}$$

Therefore, $(\varphi_j : j \in A_k)$ is linearly independent for each $1 \leq j \leq M$ by Proposition 3.9. □

A third way of partitioning Φ to prove the Rado-Horn Theorem I was given in [4], though not explicitly. Given Φ as above and $K \in \mathbb{N}$, we say a partition $\{A_k\}_{k=1}^K$ *maximizes the K-ordering of dimensions* if the following holds. Given any partition $\{B_k\}_{k=1}^K$ of $[1, M]$, if for every $1 \leq k \leq K$, $\dim \operatorname{span}\{\varphi_j : j \in A_k\} \leq \dim \operatorname{span}\{\varphi_j : j \in B_k\}$, then

$$\dim \operatorname{span}\{\varphi_j : j \in A_k\} = \dim \operatorname{span}\{\varphi_j : j \in B_k\}, \quad \text{for every } 1 \leq k \leq K.$$

It is easy to see that any partition which maximizes the K-sum of dimensions also maximizes the K-ordering of dimensions. The next proposition shows that the converse holds, at least in the case that one can partition into linearly independent sets. Therefore, when proving the Rado-Horn theorem, it makes sense to begin with a partition which maximizes the K-ordering of dimensions. We do not present a proof of this proposition, but mention that it follows from Theorem 3.12.

Proposition 3.11 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$ and $K \in \mathbb{N}$. If $\{A_k\}_{k=1}^K$ maximizes the K-ordering of dimensions of Φ and there exists a partition $\{B_k\}_{k=1}^K$ such that for each $1 \leq k \leq K$, the set $(\varphi_j : j \in B_k)$ is linearly independent, then for each $1 \leq k \leq K$, the set $(\varphi_j : j \in A_k)$ is linearly independent.*

A second obstacle to proving the Rado-Horn Theorem I is proving that a candidate partition into linearly independent sets really does partition into linearly independent sets. Our strategy will be to suppose that it does not partition into linearly independent sets, and directly construct a set $J \subset [1, M]$ which violates the hypotheses of the Rado-Horn Theorem I. In order to construct J, we will imagine moving the linearly dependent vectors from one element of the partition to another element of the partition. The first observation is that if a partition maximizes the K-ordering of dimensions, and there is a linearly dependent vector in one of the elements of the partition, then that vector is in the span of each element of the partition.

Proposition 3.12 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$, $K \in \mathbb{N}$, and let $\{A_k\}_{k=1}^K$ be a partition of $[1, M]$ which maximizes the K-ordering of dimensions of Φ. Fix $1 \leq m \leq K$. Suppose that there exist scalars $\{a_j\}_{j \in A_m}$, not all of which are zero, such that $\sum_{j \in A_m} a_j \varphi_j = 0$. Let $j_0 \in A_m$ be such that $a_{j_0} \neq 0$. Then for each $1 \leq n \leq K$,*

$$\varphi_{j_0} \in \operatorname{span}\{\varphi_j : j \in A_n\}.$$

Proof Since removing φ_{j_0} from A_m will not decrease the dimension of the span, adding φ_{j_0} to any of the other A_n's will not increase the dimension of their spans. □

A simple, but useful, observation is that if we start with a partition $\{A_k\}_{k=1}^K$ which maximizes the K-ordering of dimensions of Φ, then a new partition obtained by moving one linearly dependent vector out of some A_k into another $A_{k'}$ will also maximize the K-ordering of dimensions.

Proposition 3.13 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$, $K \in \mathbb{N}$, and let $\{A_k\}_{k=1}^K$ be a partition of $[1, M]$ which maximizes the K-ordering of dimensions of Φ. Fix $1 \leq m \leq K$. Suppose that there exist scalars $\{a_j\}_{j \in A_m}$, not all of which are zero, such that $\sum_{j \in A_m} a_j \varphi_j = 0$. Let $j_0 \in A_m$ be such that $a_{j_0} \neq 0$. For every $1 \leq n \leq K$, the partition $\{B_k\}_{k=1}^K$ given by*

$$B_k = \begin{cases} A_k & k \neq m, n, \\ A_m \setminus \{j_0\} & k = m, \\ A_n \cup \{j_0\} & k = n, \end{cases}$$

also maximizes the K-ordering of dimensions of Φ.

Proof By Proposition 3.12, the new partition has exactly the same dimension of spans as the old partition. □

The idea for constructing the set J which will contradict the hypotheses of the Rado-Horn Theorem I is to suppose that a partition which maximizes the K-ordering of dimensions does not partition into linearly independent sets. We will take a vector which is linearly dependent, and then see that it is in the span of each of the other elements of the partition. We create new partitions, which again maximize the K-ordering of dimensions, by moving the linearly dependent vector into other sets of the partition. The partition element to which we moved the vector will also be linearly dependent. We then repeat and take the index of all vectors which can be reached in such a way as our set J. It is easy to imagine that the bookkeeping aspect of this proof will become involved relatively quickly. For that reason, we will restrict to the case $K = 2$ and prove the Rado-Horn Theorem I in that case, using the same idea that will work in the general case. The bookkeeping in this case is somewhat easier, yet all of the ideas are already there.

A key concept in our proof of the Rado-Horn Theorem I is that of a *chain of dependencies* of length P. Given two collections of vectors $(\varphi_j : j \in A_1)$ and $(\varphi_j : j \in A_2)$, where $A_1 \cap A_2 = \emptyset$, we define a chain of dependencies of length P to be a finite sequence of distinct indices $\{i_1, i_2, \ldots, i_P\} \subset A_1 \cup A_2$ with the following properties:

1. i_k will be an element of A_1 for odd indices k, and an element of A_2 for even indices k,
2. $\varphi_{i_1} \in \text{span}\{\varphi_j : j \in A_1 \setminus \{i_1\}\}$, and $\varphi_{i_1} \in \text{span}\{\varphi_j : j \in A_2\}$,
3. for odd k, $1 < k \leq P$, $\varphi_{i_k} \in \text{span}\{\varphi_j : j \in (A_1 \cup \{i_2, i_4, \ldots, i_{k-1}\}) \setminus \{i_1, i_3, \ldots, i_{k-2}\}\}$ and $\varphi_{i_k} \in \text{span}\{\varphi_j : j \in (A_2 \cup \{i_1, i_3, \ldots, i_{k-2}\}) \setminus \{i_2, i_4, \ldots, i_{k-1}\}\}$,

4. for even k, $1 < k \leq P$, $\varphi_{i_k} \in \text{span}\{\varphi_j : j \in (A_2 \cup \{i_1, i_3, \ldots, i_{k-1}\}) \setminus \{i_2, i_4, \ldots, i_k\}\}$, and $\varphi_{i_k} \in \text{span}\{\varphi_j : j \in (A_1 \cup \{i_2, i_4, \ldots, i_{k-2}\}) \setminus \{i_1, i_3, \ldots, i_{k-1}\}\}$.

A chain of dependencies is constructed as follows. Start with a linearly dependent vector. Moving that vector to another set in the partition cannot increase the sum of the dimensions of the spans, so that vector is also in the span of the vectors in the set to which it has been moved. Now, that makes the new set linearly dependent, so take a second vector, which is linearly dependent in the second set, and move it to a third set. Again, the second vector is in the span of the vectors in the third set. Continuing in this fashion gives a chain of dependencies.

With this new definition, it is easier to describe the technique of the proof of the Rado-Horn Theorem I. Suppose that a partition which maximizes the 2-ordering of dimensions does not partition into linearly independent sets. Let J be the union of all of the chains of dependencies. We will show that J satisfies

$$\frac{|J|}{\dim \text{span}\{\varphi_i : i \in J\}} > 2.$$

Example 3.2 We give an example of chains of dependencies in $\mathcal{H}^3$. Let $\varphi_1 = \varphi_5 = (1, 0, 0)^T$, $\varphi_2 = \varphi_6 = (0, 1, 0)^T$, $\varphi_3 = \varphi_7 = (0, 0, 1)^T$, and $\varphi_4 = (1, 1, 1)^T$. Suppose also that $A_1 = \{1, 2, 3, 4\}$ and $A_2 = \{5, 6, 7\}$. Then, the set $\{4, 5, 1, 6, 2, 7, 3\}$ is a chain of dependencies of length 7. Note also that $\{4, 5, 1\}$ is a chain of dependencies of length 3.

Note that if we let J be the union of all of the sets of dependencies based on the partition $\{A_1, A_2\}$, then

$$\frac{|J|}{\dim \text{span}\{\varphi_i : i \in J\}} = \frac{7}{3} > 2.$$

The following example illustrates what can happen if we do not start with a partition which maximizes the K-ordering of dimensions.

Example 3.3 Let $\varphi_1 = (1, 0, 0)^T, \varphi_2 = (0, 1, 0)^T, \varphi_3 = (1, 1, 0)^T, \varphi_4 = (1, 0, 0)^T$, $\varphi_5 = (0, 0, 1)^T$, and $\varphi_6 = (0, 1, 1)^T$. Imagine starting with our partition consisting of $A_1 = \{1, 2, 3\}$ and $A_2 = \{4, 5, 6\}$. We can make a chain of dependencies $\{3, 6\}$, but notice that $\{\varphi_6, \varphi_1, \varphi_2\}$ is linearly independent. This indicates that we have removed one linear dependence, and in fact, the new partition $B_1 = \{1, 2, 6\}$, $B_2 = \{3, 4, 5\}$ is linearly independent.

Note that the new partition does maximize the K-ordering of dimensions.

A slight generalization of Proposition 3.13 is given below.

Lemma 3.4 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$, and suppose that Φ cannot be partitioned into two linearly independent sets. Let $\{A_1, A_2\}$ be a partition of $[1, M]$ which maximizes the 2-ordering of dimensions. Let $\{i_1, \ldots, i_P\}$ be a chain of dependencies of length*

P based on the partition $\{A_1, A_2\}$. For each $1 \leq k \leq P$, the partition $\{B_1(k), B_2(k)\}$ given by

$$B_1(k) = \left(A_1 \cup \bigcup_{1 \leq j \leq k/2} \{i_{2j}\}\right) \Big\backslash \bigcup_{1 \leq j \leq (k+1)/2} \{i_{2j-1}\},$$

$$B_2(k) = \left(A_2 \cup \bigcup_{1 \leq j \leq (k+1)/2} \{i_{2j-1}\}\right) \Big\backslash \bigcup_{1 \leq j \leq k/2} \{i_{2j}\}$$

also maximizes the 2-ordering of dimensions.

We introduce one notational convenience at this point. Given a set $A \subset [1, M]$, a finite sequence of elements $\{i_1, \ldots, i_P\}$, and disjoint sets $Q, R \subset [1, P]$, we define

$$A(Q; R) = \left(A \cup \bigcup_{j \in Q} \{i_j\}\right) \Big\backslash \bigcup_{j \in R} \{i_j\}.$$

Lemma 3.5 *Let $\Phi = \Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$, and suppose that Φ cannot be partitioned into two linearly independent sets. Let $\{A_1, A_2\}$ be a partition of $[1, M]$ which maximizes the 2-ordering of dimensions. Let J be the union of all chains of dependencies of Φ based on the partition $\{A_1, A_2\}$. Let $J_1 = J \cap A_1$ and $J_2 = J \cap A_2$, and $S = \text{span}\{\varphi_i : i \in J\}$. Then,*

$$S = \text{span}\{\varphi_i : i \in J_k\}$$

for $k = 1, 2$.

Proof We will prove the lemma in the case $k = 1$, the other case being similar. It suffices to show that for every chain of dependencies $\{i_1, \ldots, i_P\}$, all of the even indexed vectors φ_k are in the span of J_1, which we will do by induction.

Note that $\varphi_{i_2} \in \text{span}\{\varphi_i : i \in A_1 \setminus \{i_1\}\}$. Therefore, there exist scalars $\{a_i : i \in A_1 \setminus \{i_1\}\}$ such that

$$\varphi_{i_2} = \sum_{i \in A_1 \setminus \{i_1\}} a_i \varphi_i.$$

Let $i \in A_1 \setminus \{i_1\}$ be such that $a_i \neq 0$. We show that $\{i_1, i_2, i\}$ is a chain of dependencies of length 3. First, note that $\varphi_i \in \text{span}\{\varphi_j : j \in A_1(\{2\}; \{1\})\}$. By Lemma 3.4, the partition $\{A_1(\{2\}; \{1\}), A_2(\{1\}; \{2\})\}$ maximizes the 2-ordering of dimensions. Since φ_i is a dependent vector in $(\varphi_j : j \in A_1(\{2\}; \{1\}))$, the partition $\{A_1(\{2\}; \{1, i\}), A_2(\{1, i\}; \{2\})\}$ has the same dimensions as the partition $\{A_1(\{2\}; \{1\}), A_2(\{1\}; \{2\})\}$. In particular, $\varphi_i \in \text{span}\{\varphi_j : j \in A_2(\{1\}; \{2\})\}$. Therefore $\{i_1, i_2, i\}$ is a chain of dependencies of length 3, and $\varphi_{i_2} \in \text{span}\{\varphi_j : j \in J_1\}$.

Now, suppose that $\varphi_{i_2}, \ldots, \varphi_{i_{2m-2}} \in \text{span}\{\varphi_j : j \in J_1\}$. We show that $\varphi_{i_{2m}} \in \text{span}\{\varphi_j : j \in J_1\}$. Note that $\varphi_{i_{2m}} \in \text{span}\{\varphi_j : j \in A_1(\{2, 4, \ldots, 2m-2\}; \{1, 3, \ldots, 2m-1\})\}$. Therefore, there exist scalars $\{a_i : i \in A_1(\{2, 4, \ldots, 2m-2\}; \{1, 3, \ldots, 2m-1\})\}$ such that

$$\varphi_{i_{2m}} = \sum_{i \in A_1(\{2,4,\ldots,2m-2\}; \{1,3,\ldots,2m-1\})} a_i \varphi_i. \tag{3.4}$$

By the induction hypothesis, for the even indices $j < 2m$, $\varphi_j \in \operatorname{span}\{\varphi_i : i \in J_1\}$, so it suffices to show that for all $i \in A_1(\emptyset; \{1, 3, \ldots, 2m-1\})$ such that $a_i \neq 0$, the set $\{i_1, \ldots, i_{2m}, i\}$ is a chain of dependencies. (Note that there may not be any i in this set.) By (3.4), $\varphi_i \in \operatorname{span}\{\varphi_j : j \in A_1(\{2, 4, \ldots, 2m\}; \{1, 3 \ldots, 2m-1\})\}$. By Lemma 3.4, the partition $\{A_1(\{2, 4, \ldots, 2m\}; \{1, 3, \ldots, 2m-1\}), A_2(\{1, 3 \ldots, 2m-1\}; \{2, 4, \ldots, 2m\})\}$ maximizes the 2-ordering of dimensions. Therefore, since φ_i is a dependent vector in $(\varphi_j : j \in A_1(\{2, 4, \ldots, 2m\}; \{1, 3, \ldots, 2m-1\}))$, moving i into the second partition by forming the new partition $\{A_1(\{2, 4, \ldots, 2m\}; \{1, 3, \ldots, 2m-1, i\}), A_2(\{1, 3, \ldots, 2m-1, i\}; \{2, 4, \ldots, 2m\})\}$ does not change the dimensions. In particular,

$$\varphi_i \in \operatorname{span}\big\{\varphi_j : j \in A_2\{1, 3, \ldots, 2m-1\}; \{2, 4, \ldots, 2m\}\big\}.$$

Therefore $\{i_1, i_2, \ldots, i_{2m}, i\}$ is a chain of dependencies of length $2m+1$, and $\varphi_{i_{2m}} \in \operatorname{span}\{\varphi_j : j \in J_1\}$. □

Theorem 3.11 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$. If for every nonempty $J \subset [1, M]$,*

$$\frac{|J|}{\dim \operatorname{span}\{\varphi_i : i \in J\}} \leq 2,$$

then Φ can be partitioned into two linearly independent sets.

Proof Suppose that Φ cannot be partitioned into two linearly independent sets. We will construct a set J such that

$$\frac{|J|}{\dim \operatorname{span}\{\varphi_i : i \in J\}} > 2.$$

Let $\{A_1, A_2\}$ be a partition of $[1, M]$ which maximizes the 2-ordering of dimensions. By hypothesis, this partition of $[1, M]$ does not partition Φ into linearly independent sets, so at least one of the collections $(\varphi_j : j \in A_k), k = 1, 2$ must be linearly dependent. Without loss of generality, we assume that $(\varphi_j : j \in A_1)$ is linearly dependent.

Let J be the union of all chains of dependencies based on the partition $\{A_1, A_2\}$. We claim that J satisfies

$$\frac{|J|}{\dim \operatorname{span}\{\varphi_i : i \in J\}} > 2.$$

Indeed, let $J_1 = J \cap A_1$ and $J_2 = J \cap A_2$. By Lemma 3.5, $(\varphi_j : j \in J_k), k = 1, 2$ span the same subspace $S = (\varphi_j : j \in J)$. Since $(\varphi_j : j \in J_1)$ is not linearly independent, $|J_1| > \dim S$. Therefore,

$$\begin{aligned} |J| &= |J_1| + |J_2| \\ &> \dim S + \dim S = 2 \dim\{\varphi_j : j \in J\}, \end{aligned}$$

and the theorem is proved.

A careful reading of the proof of Theorem 3.11 yields that we have proven more than what has been advertised. In fact, we have essentially proven the more general Theorem 3.12 in the special case of partitioning into two sets. □

3.4 The Rado-Horn Theorem III and Its Proof

The final section of this chapter is devoted to the proof of the Rado-Horn Theorem III, which we recall below (see Theorem 3.6). We did not include all elements of the theorem, as a discussion of partitions maximizing the K-ordering of dimensions would have taken us too far astray at that time, and we only needed the full version of the theorem in the proof of Lemma 3.2, whose proof we have delayed until the end of this section.

Theorem 3.12 (Rado-Horn Theorem III) *Let $\Phi = (\varphi_i)_{i=1}^M$ be a collection of vectors in $\mathcal{H}^N$ and $K \in \mathbb{N}$. Then the following conditions are equivalent.*

(1) *There exists a partition $\{A_k : k = 1, \dots, K\}$ of $[1, M]$ such that for each $1 \leq k \leq K$ the set $(\varphi_j : j \in A_k)$ is linearly independent.*
(2) *For all $J \subset I$,*

$$\frac{|J|}{\dim \operatorname{span}\{\varphi_j : j \in J\}} \leq K. \tag{3.5}$$

Moreover, in the case that both of the conditions above are true, any partition which maximizes the K-ordering of dimensions will partition the vectors into linearly independent sets. In the case that either of the conditions above fails, there exists a partition $\{A_k : k = 1, \dots, K\}$ of $[1, M]$ and a subspace S of $\mathcal{H}^N$ such that the following three conditions hold.

(a) *For all $1 \leq k \leq K$, $S = \operatorname{span}\{\varphi_j : j \in A_k \text{ and } \varphi_j \in S\}$.*
(b) *For $J = \{i \in I : \varphi_i \in S\}$, $\frac{|J|}{\dim \operatorname{span}\{\varphi_i : i \in J\}} > K$.*
(c) *For each $1 \leq k \leq K$, $(P_{S^\perp}\varphi_i : i \in A_k, \varphi_i \notin S)$ is linearly independent, where $P_{S^\perp}$ is the orthogonal projection onto $S^\perp$.*

We saw in the previous section how to prove the more elementary version of the Rado-Horn theorem in the case of partitioning into two subsets. The details of the proof in the general setting are similar, and where the proofs follow the same outline we will omit them. The interested reader can refer to [4, 13] for full details.

As before, our general plan is to start with a partition which maximizes the K-ordering of dimensions. We will show that if that partition does not partition into linearly independent sets, then we can construct a set J which directly contradicts the hypotheses of the Rado-Horn theorem. The set J constructed will span the subspace S in the conclusion of the theorem.

Let $\{A_1, \dots, A_K\}$ be a partition of $[1, M]$ and let $\{i_1, \dots, i_P\} \subset [1, M]$. We say that $\{a_1, \dots, a_P\}$ are the associated partition indices if for all $1 \leq p \leq P$, $i_p \in A_{a_p}$. We define the chain of partitions $\{\mathcal{A}^j\}_{j=1}^P$ associated with $\mathcal{A} = \{A_1, \dots, A_K\}$ and $\{i_1, \dots, i_P\}$ as follows. Let $\mathcal{A}^1 = \mathcal{A}$, and given that the partitions $\mathcal{A}^j = \{A_k^j\}_{k=1}^K$

have been defined for $1 \le j \le p$ and $p \le P$, we define $\mathcal{A}^{p+1} = \{A_1^{p+1}, \dots, A_K^{p+1}\}$ by

$$A_k^{p+1} = \begin{cases} A_k^p & k \neq a_p, a_{p+1}, \\ A_{a_p}^p \setminus \{i_p\} & k = a_p, \\ A_{a_{p+1}}^p \cup \{i_p\} & k = a_{p+1}. \end{cases}$$

A chain of dependencies of length P based on the partition $\{A_1, \dots, A_K\}$ is a set of distinct indices $\{i_1, \dots, i_P\} \subset [1, M]$ with associated partition indices $\{a_1, \dots, a_P\}$ and the $P+1$ associated partitions $\{A_k^p\}_{k=1}^K$, $1 \le p \le P+1$ such that the following conditions are met.

1. $a_p \neq a_{p+1}$ for all $1 \le p < P$.
2. $a_1 = 1$.
3. $\varphi_{i_1} \in \text{span}\{\varphi_j : j \in A_1^2\}$, and $\varphi_{i_1} \in \text{span}\{\varphi_j : j \in A_{a_2}^1\}$.
4. $\varphi_{i_p} \in \text{span}\{\varphi_j : j \in A_{a_p}^p \setminus \{i_p\}\}$ for all $1 < p \le P$.
5. $\varphi_{i_p} \in \text{span}\{\varphi_j : j \in A_{a_{p+1}}^p\}$ for all $1 < p < P$.

Lemma 3.6 *With the notation above, for each $1 \le p \le P+1$, the partition $\{A_k^p\}_{k=1}^K$ maximizes the K-ordering of dimensions.*

Proof As in Lemma 3.4, when we are constructing the pth partition, we are taking a vector that is dependent in the $(p-1)$st partition, and moving it to a new partition element. Since removing the dependent vector does not reduce the dimension, all of the dimensions in the pth partition must remain the same. Hence, it maximizes the K-ordering of dimensions. □

Lemma 3.7 *Let $\Phi = (\varphi_i)_{i=1}^M \subset \mathcal{H}^N$, and suppose that Φ cannot be partitioned into K linearly independent sets. Let $\{A_1, \dots, A_K\}$ be a partition of $[1, M]$ which maximizes the K-ordering of dimensions. Let J be the union of all chains of dependencies of Φ based on the partition $\{A_1, \dots, A_K\}$. For $1 \le k \le K$, let $J_k = J \cap A_k$, and let $S = \text{span}\{\varphi_i : i \in J\}$. Then,*

$$S = \text{span}\{\varphi_i : i \in J_k\}$$

for $k = 1, \dots, K$.

Proof We sketch the proof for $k = 1$. The details are similar to Lemma 3.5. Clearly, it suffices to show that if $\{i_1, \dots, i_P\}$ is a chain of dependencies based on $\{A_1, \dots, A_K\}$, then each $\varphi_{i_p} \in \text{span}\{\varphi_i : i \in J_1\}$ for each $1 \le p \le P$. For $p = 1$, this is true since $a_1 = 1$. (For $k \neq 1$, it is true since moving a dependent vector from A_1 to A_k cannot increase the dimension of $(\varphi_i : i \in A_k)$.)

Proceeding by induction on p, assume that $\varphi_{i_1}, \dots, \varphi_{i_{p-1}} \in \text{span}\{\varphi_i : i \in J_1\}$. Let $\{a_1, \dots, a_P\}$ be the associated partition indices and $\mathcal{A}^p = \{A_k^p\}_{k=1}^K$ the associated partitions. If $a_p = 1$, then we are done. Otherwise, we know that $\varphi_{i_p} \in \text{span}\{\varphi_j : j \in$

$A_{a_p}^{p+1}\}$. Note that $i_p \in A_{a_p}^p$ and $i_p \notin A_{a_p}^{p+1}$. Therefore, removing i_p from $A_{a_p}^p$ does not change the span of the vectors indexed by $A_{a_p}^p$, and by Lemma 3.6,

$$\varphi_{i_p} \in \operatorname{span}\{\varphi_j : j \in A_1^p\}.$$

Write

$$\varphi_{i_p} = \sum_{j \in A_1^p} \alpha_j \varphi_j$$

for some scalars α_j. We claim that for each j such that $\alpha_j \neq 0$, $\varphi_j \in \operatorname{span}\{\varphi_i : i \in J_1\}$. Since $A_1^p \subset A_1 \cup \{i_1, \ldots, i_{p-1}\}$, by the induction hypothesis it suffices to show that whenever $j_0 \in A_1^p \setminus \{i_1, \ldots, i_p\}$, $\varphi_{j_0} \in \operatorname{span}\{\varphi_i : i \in J_1\}$. To do so, we claim that $\{i_1, \ldots, i_p, j_0\}$ is a chain with associated indices $\{a_1, \ldots, a_p, 1\}$. Indeed, noting that $A_1^{p+1} = (A_1^p \cup \{i_p\})$, property 4 of a chain of dependencies ensures that

$$\varphi_{j_0} \in \operatorname{span}\{\varphi_i : i \in (A_1^p \cup \{i_p\}) \setminus \{j_0\}\}. \qquad \square$$

Proof of Theorem 3.12 Suppose that Φ cannot be partitioned into K linearly independent sets. Let $\mathcal{A}$ be a partition of $[1, M]$ which maximizes the K-ordering of subspaces. By hypothesis, this partition does not partition Φ into linearly independent sets, so without loss of generality, we assume that $(\varphi_i : i \in A_1)$ is linearly dependent.

Let J be the union of all chains of dependencies based on the partition $\mathcal{A}$ and $S = \operatorname{span}\{\varphi_i : i \in J\}$. By Lemma 3.7, J satisfies

$$J = \{i \in [1, M] : \varphi_i \in S\}.$$

We show that J and S satisfy the conclusions of Theorem 3.12.

First, let $J_k = A_k \cap J$ for $1 \leq k \leq K$. We have that $\operatorname{span}\{\varphi_i : i \in J_k\} = S$ for $1 \leq k \leq K$ by Lemma 3.7, and $|J_1| > \dim S$ by the assumption that $\mathcal{A}$ does not partition into linearly independent sets. Therefore,

$$|J| = \sum_{k=1}^{K} |J_k| > K \dim S = K \dim \operatorname{span}\{\varphi_i : i \in J\}.$$

In particular, if it were possible to partition into linearly independent sets, $\mathcal{A}$ would do it.

To see (a) in the list of conclusions in Theorem 3.12, note that $S \supset \operatorname{span}\{\varphi_i : i \in A_k, \varphi_i \in S\}$ is obvious, and $S \subset \operatorname{span}\{\varphi_i : i \in A_k, \varphi_i \in S\}$ follows from Lemma 3.7. Part (b) follows from Lemma 3.7 and the computations above.

It remains to prove (c). Suppose there exist $\{\alpha_j\}_{j \in A_k \setminus J}$ not all zero such that $\sum_{j \in A_k \setminus J} \alpha_j \varphi_j \in S$. Since J is the union of the set of all chains of dependencies, $\sum_{j \in A_k \setminus J} \alpha_j \varphi_j \neq 0$. Let $\{\beta_j\}_{j \in J_k}$ be scalars such that

$$\sum_{j \in A_k \setminus J} \alpha_j \varphi_j = \sum_{j \in J_k} \beta_j \varphi_j. \tag{3.6}$$

Choose j_0 and a chain of dependencies $\{i_1, \ldots, i_{P-1}, j_0\}$ such that $\beta_{j_0} \neq 0$ and such that P is the minimum length of all chains of dependencies whose final element is

in $\{\beta_j : j \neq 0\}$. Let $m \in A_k \setminus J$ such that $\alpha_m \neq 0$. We claim that $\{i_1, \ldots, i_{P-1}, m\}$ is a chain of dependencies, which contradicts $m \notin J$ and finishes the proof.

The key observation to proving the claim is to observe that the minimality of the length of the chain $\{i_1, \ldots, i_{P-1}, j_0\}$ forces

$$\{j : \beta_j \neq 0\} \cup \{j : \alpha_j \neq 0\} \subset A_{a_P}^P. \tag{3.7}$$

To verify property 5 of a chain of dependencies, since $\varphi_{i_{P-1}} \in \operatorname{span}\{\varphi_j : j \in A_{a_P}^P \setminus \{j_0\}\}$, (3.6) and (3.7) imply that $\varphi_{i_{P-1}} \in \operatorname{span}\{\varphi_j : j \in A_{a_P}^P \setminus \{m\}\}$. To see property 4 of a chain of dependencies, write

$$\varphi_{j_0} = \sum_{j \in A_{a_P}^P \setminus \{j_0\}} \gamma_j \varphi_j.$$

If $\gamma_m \neq 0$, then $\varphi_m \in \operatorname{span}\{\varphi_i : i \in A_{a_P}^P \setminus \{m\}\}$ directly from the above equation. If $\gamma_m = 0$, then replacing φ_{j_0} in (3.6) with $\sum_{j \in A_{a_P}^P \setminus \{j_0\}} \gamma_j \varphi_j$ shows that $\varphi_m \in \operatorname{span}\{\varphi_i : i \in A_{a_P}^P \setminus \{m\}\}$. □

We end with a proof of Lemma 3.2, which we restate now for the reader's convenience.

Theorem 3.13 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a finite collection of vectors in $\mathcal{H}^N$, and let $K \in \mathbb{N}$. Assume*

1. *Φ can be partitioned into $K+1$-linearly independent sets, and*
2. *Φ can be partitioned into a set and K spanning sets.*

Then there is a partition $\{A_k\}_{k=1}^{K+1}$ so that $(\varphi_j)_{j \in A_k}$ is a linearly independent spanning set for all $k = 2, 3, \ldots, K+1$ and $(\varphi_i)_{i \in A_1}$ is a linearly independent set.

Proof We choose a partition $\{A_k\}_{k=1}^{K+1}$ of $[1, M]$ that maximizes $\dim \operatorname{span}\{\varphi_j\}_{j \in A_1}$ taken over all partitions so that the last K sets span $\mathcal{H}^N$. If $\{B_k\}_{k=1}^{K+1}$ is a partition of $[1, M]$ such that for all $1 \leq k \leq K+1$,

$$\dim \operatorname{span}\{\varphi_j\}_{j \in B_i} \geq \dim \operatorname{span}\{\varphi_j\}_{j \in A_i},$$

then

$$\dim \operatorname{span}\{\varphi_j\}_{j \in A_i} = \dim \operatorname{span}\{\varphi_j\}_{j \in B_i}$$

for all $i = 2, \ldots, K+1$ since $\dim \operatorname{span}\{\varphi_j\}_{j \in A_i} = N$, and $\dim \operatorname{span}\{\varphi_j\}_{j \in A_1} \geq \dim \operatorname{span}\{\varphi_j\}_{j \in B_1}$ by construction. This means that the partition $\{A_k\}_{k=1}^{K+1}$ maximizes the $(K+1)$-ordering of dimensions. By Theorem 3.12, since there is a partition of Φ into $K+1$ linearly independent sets, $\{A_k\}_{k=1}^{K+1}$ partitions Φ into linearly independent sets, as desired. □

3.5 The Maximal Number of Spanning Sets in a Frame

In this section, we determine the maximal number of spanning sets contained in a frame. Partitioning into spanning sets has not been studied as much as partitioning into linearly independent sets, and several of the results in this section are, as far as we know, new.

In one sense, the difficulties associated with choosing spanning sets contained in a frame is very similar to the difficulties associated with choosing linearly independent sets. Namely, choosing spanning sets at random will not necessarily provide the maximum number of spanning sets. A trivial example is given in $\mathbb{R}^2$ by the frame (e_1, e_1, e_2, e_1+e_2) where $e_1 = (1,0)^T$, $e_2 = (0,1)^T$. If we choose (e_2, e_1+e_2), then we can only get one spanning set, while if we choose (e_1, e_2), (e_1, e_1+e_2) we get two spanning sets. Recently [15], the problem of determining the maximal number of spanning sets was resolved. We begin with some preliminary results.

Theorem 3.14 *Let P be a projection on $\mathcal{H}^M$ and let $(e_i)_{i=1}^M$ be an orthonormal basis for $\mathcal{H}^M$. If $I \subset \{1, 2, \ldots, M\}$, the following are equivalent*:

(1) $(Pe_i)_{i\in I}$ *spans* $P(\mathcal{H}^M)$.
(2) $((Id - P)e_i)_{i\in I^c}$ *is linearly independent.*

Proof (1) $\Rightarrow$ (2) Assume that $((Id - P)e_i)_{i\in I^c}$ is not linearly independent. Then there exist scalars $\{b_i\}_{i\in I^c}$, not all zero, so that

$$\sum_{i\in I^c} b_i(Id - P)e_i = 0.$$

It follows that

$$x = \sum_{i\in I^c} b_i e_i = \sum_{i\in I^c} b_i Pe_i \in P\big(\mathcal{H}^M\big).$$

Thus,

$$\langle x, Pe_j\rangle = \langle Px, e_j\rangle = \sum_{i\in I^c} b_i\langle e_i, e_j\rangle = 0, \quad \text{if } j \in I.$$

So $x \perp \operatorname{span}\{Pe_i\}_{i\in I}$ and hence this family is not spanning for $P(\mathcal{H}^M)$.

(2) $\Rightarrow$ (1) We assume that $\operatorname{span}\{Pe_i\}_{i\in I} \neq P(\mathcal{H}^M)$. That is, there is a $0 \neq x \in P(\mathcal{H}^M)$ so that $x \perp \operatorname{span}\{Pe_i\}_{i\in I}$. Also, $x = \sum_{i=1}^M \langle x, e_i\rangle Pe_i$. Then

$$\langle x, Pe_i\rangle = \langle Px, e_i\rangle = \langle x, e_i\rangle = 0, \quad \text{for all } i \in I.$$

Hence, $x = \sum_{i\in I^c}\langle x, e_i\rangle e_i$. That is,

$$\sum_{i\in I^c}\langle x, e_i\rangle e_i = x = Px = \sum_{i\in I^c}\langle x, e_i\rangle Pe_i.$$

That is,

$$\sum_{i\in I^c}\langle x, e_i\rangle(I - P)e_i = 0,$$

i.e., $((Id - P)e_i)_{i\in I^c}$ is not linearly independent. □

We state an immediate consequence.

Corollary 3.2 *Let P be a projection on $\mathcal{H}^M$. The following are equivalent*:

(1) *There is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ so that $(Pe_i)_{i\in A_j}$ spans $P(\mathcal{H}^M)$ for all $j = 1, 2, \ldots, r$.*
(2) *There is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ so that $((Id-P)e_i)_{i\in A_j^c}$ is linearly independent for every $j = 1, 2, \ldots, r$.*

Now we can prove the main result, which gives the maximal number of spanning sets contained in a frame. Recall that this problem is independent of applying an invertible operator to the frame and hence we only need to prove the result for Parseval frames.

Theorem 3.15 [15] *Let $(\varphi_i)_{i=1}^M$ be a Parseval frame for $\mathcal{H}^N$, let P be a projection on $\mathcal{H}^M$ with $(\varphi_i)_{i=1}^M = (Pe_i)_{i=1}^M$ where $(e_i)_{i=1}^M$ is an orthonormal basis for $\mathcal{H}^M$, and let $(\psi_i)_{i=1}^{(r-1)M}$ be the multiset*

$$\begin{aligned}&\big\{(Id-P)e_1, \ldots, (Id-P)e_1, (Id-P)e_2, \ldots, (Id-P)_2, \ldots,\\ &(Id-P)e_M, \ldots, (Id-P)e_M\big\}.\end{aligned} \tag{3.8}$$

The following are equivalent:

(1) *$(\varphi_i)_{i=1}^M$ can be partitioned into r spanning sets.*
(2) *$(\psi_i)_{i=1}^{(r-1)M}$ can be partitioned into r linearly independent sets.*
(3) *For all $I \subset \{1, 2, \ldots, (r-1)M\}$,*

$$\frac{|I|}{\dim \operatorname{span}\{\psi_i\}_{i\in I}} \leq r. \tag{3.9}$$

Proof (1) $\Rightarrow$ (2) Let $\{A_j\}_{j=1}^r$ be a partition of $\{1, 2, \ldots, M\}$ so that $(Pe_i)_{i\in A_j}$ is spanning for every $j = 1, 2, \ldots, r$. Then $((Id-P)e_i)_{i\in A_j^c}$ is linearly independent for every $j = 1, 2, \ldots, r$. Since $\{A_j\}_{j=1}^r$ is a partition, each $(Id-P)e_i$ appears in exactly $r-1$ of the collections $((Id-P)e_i)_{i\in A_j^c}$. So the multiset $(\psi_i)_{i=1}^{(r-1)M}$ has a partition into r linearly independent sets.

(2) $\Rightarrow$ (1) Let $\{A_j\}_{j=1}^r$ be a partition of $\{1, 2, \ldots, (r-1)M\}$ so that $((Id-P)e_i)_{i\in A_j}$ is linearly independent for all $j = 1, 2, \ldots, r$. Since the collection $((Id-P)e_i)_{i\in A_j}$ is linearly independent, it contains at most one of the r copies of $(Id-P)e_i$ for each $i = 1, 2, \ldots, M$. Hence, each $(Id-P)e_i$ is in exactly $r-1$ of the collections $((Id-P)e_i)_{i\in A_j}$. That is, each i is in all but one of these sets A_j. For each $j = 1, 2, \ldots, r$, let B_j be the complement of A_j in $\{1, 2, \ldots, M\}$. Since $((Id-P)e_i)_{i\in A_j}$ is linearly independent, $(Pe_i)_{i\in B_j}$ is spanning. Also, for all $i, j = 1, \ldots, r$ with $i \neq j$, we have $B_i \cap B_j = \emptyset$, since if $k \in B_i \cap B_j$ then $k \notin A_i$, and $k \notin A_j$, which is a contradiction.

(2) $\Leftrightarrow$ (3) This is the Rado-Horn Theorem I. $\square$

3.6 Problems

We end with the problems which are still left open in this theory. The Rado-Horn theorem and its variants tell us the minimal number of linearly independent sets into which we can partition a frame. But this is unusable in practice, since it requires doing a calculation for every subset of the frame. What we have done in this chapter is to try to use the Rado-Horn theorem to identify, *in terms of frame properties*, the minimal number of linearly independent sets into which we can partition a frame. We have shown that there are many cases where we can do this, but the general problem is still open.

Problem 3.2 Identify, in terms of frame properties, the minimal number of linearly independent sets into which we can partition a frame.

By frame properties we mean using the eigenvalues of the frame operator of a frame $(\varphi_i)_{i=1}^M$, the norms of the frame vectors, or the norms of the vectors of the associated Parseval frame or perhaps the norms of the frame vectors of the canonical Parseval frame associated to $\{\frac{\varphi_i}{\|\varphi_i\|}\}_{i=1}^M$.

The main problem concerning spanning and independence properties of frames is the following.

Problem 3.3 Given a frame Φ for $\mathcal{H}^N$, find integers $r_0, r_1, \ldots, r_{N-1}$ so that Φ can be partitioned into r_0 sets of codimension 0 (i.e., r_0 spanning sets), r_1 sets of codimension 1, and in general, r_i sets of codimension i for $i = 0, 1, 2, \ldots, N-1$. Moreover, do this in a maximal way in the sense that r_0 is the maximal number of spanning sets, and whenever we take r_0 spanning sets out of the frame, r_1 is the maximal number of hyperplanes we can obtain from the remaining vectors, and whenever r_0, r_1 are known, r_2 is the maximal number of subsets of codimension 2 which can be obtained from the remaining vectors, etc.

Finally, we need to know how to answer the above problems in practice.

Problem 3.4 Find real-time algorithms for answering the problems above.

Problem 3.4 is particularly difficult, since it requires finding an algorithm for proving the Rado-Horn theorem just to get started.

Acknowledgements The first author is supported by NSF DMS 1008183, NSF ATD 1042701, and AFOSR FA9550-11-1-0245.

References

1. Alexeev, B., Cahill, J., Mixon, D.G.: Full spark frames, preprint
2. Balan, R.: Equivalence relations and distances between Hilbert frames. Proc. Am. Math. Soc. **127**(8), 2353–2366 (1999)

3. Bodmann, B.G., Casazza, P.G.: The road to equal-norm Parseval frames. J. Funct. Anal. **258**(2), 397–420 (2010)
4. Bodmann, B.G., Casazza, P.G., Paulsen, V.I., Speegle, D.: Spanning and independence properties of frame partitions. Proc. Am. Math. Soc. **40**(7), 2193–2207 (2012)
5. Bodmann, B.G., Casazza, P.G., Kutyniok, G.: A quantitative notion of redundancy for finite frames. Appl. Comput. Harmon. Anal. **30**, 348–362 (2011)
6. Bourgain, J.: Λ_p-Sets in Analysis: Results, Problems and Related Aspects. Handbook of the Geometry of Banach Spaces, vol. I, pp. 195–232. North-Holland, Amsterdam (2001)
7. Cahill, J.: Flags, frames, and Bergman spaces. M.S. Thesis, San Francisco State University (2010)
8. Casazza, P.G.: Custom building finite frames. In: Wavelets, Frames and Operator Theory, College Park, MD, 2003. Contemp. Math., vol. 345, pp. 61–86. Am. Math. Soc., Providence (2004)
9. Casazza, P., Christensen, O., Lindner, A., Vershynin, R.: Frames and the Feichtinger conjecture. Proc. Am. Math. Soc. **133**(4), 1025–1033 (2005)
10. Casazza, P.G., Fickus, M., Weber, E., Tremain, J.C.: The Kadison–Singer problem in mathematics and engineering—a detailed account. In: Han, D., Jorgensen, P.E.T., Larson, D.R. (eds.) Operator Theory, Operator Algebras and Applications. Contemp. Math., vol. 414, pp. 297–356 (2006)
11. Casazza, P.G., Kovačević, J.: Equal-norm tight frames with erasures. Adv. Comput. Math. **18**(2–4), 387–430 (2003)
12. Casazza, P., Kutyniok, G.: A generalization of Gram-Schmidt orthogonalization generating all Parseval frames. Adv. Comput. Math. **18**, 65–78 (2007)
13. Casazza, P.G., Kutyniok, G., Speegle, D.: A redundant version of the Rado-Horn theorem. Linear Algebra Appl. **418**, 1–10 (2006)
14. Casazza, P.G., Kutyniok, G., Speegle, D., Tremain, J.C.: A decomposition theorem for frames and the Feichtinger conjecture. Proc. Am. Math. Soc. **136**, 2043–2053 (2008)
15. Casazza, P.G., Peterson, J., Speegle, D.: Private communication
16. Casazza, P.G., Tremain, J.: The Kadison–Singer problem in mathematics and engineering. Proc. Natl. Acad. Sci. **103**(7), 2032–2039 (2006)
17. Christensen, O., Lindner, A.: Decompositions of Riesz frames and wavelets into a finite union of linearly independent sets. Linear Algebra Appl. **355**, 147–159 (2002)
18. Edmonds, J., Fulkerson, D.R.: Transversals and matroid partition. J. Res. Natl. Bur. Stand. Sect. B **69B**, 147–153 (1965)
19. Horn, A.: A characterization of unions of linearly independent sets. J. Lond. Math. Soc. **30**, 494–496 (1955)
20. Janssen, A.J.E.M.: Zak transforms with few zeroes and the tie. In: Feichtinger, H.G., Strohmer, T. (eds.) Advances in Gabor Analysis, pp. 31–70. Birkhäuser, Boston (2002)
21. Oxley, J.: Matroid Theory. Oxford University Press, New York (2006)
22. Rado, R.: A combinatorial theorem on vector spaces. J. Lond. Math. Soc. **37**, 351–353 (1962)

Chapter 4
Algebraic Geometry and Finite Frames

Jameson Cahill and Nate Strawn

Abstract Interesting families of finite frames often admit characterizations in terms of algebraic constraints, and thus it is not entirely surprising that powerful results in finite frame theory can be obtained by utilizing tools from algebraic geometry. In this chapter, our goal is to demonstrate the power of these techniques. First, we demonstrate that algebro-geometric ideas can be used to explicitly construct local coordinate systems that reflect intuitive degrees of freedom within spaces of finite unit norm tight frames (and more general spaces), and that optimal frames can be characterized by useful algebraic conditions. In particular, we construct locally well-defined real-analytic coordinate systems on spaces of finite unit norm tight frames, and we demonstrate that many types of optimal Parseval frames are dense and that further optimality can be discovered through embeddings that naturally arise in algebraic geometry.

Keywords Algebraic geometry · Elimination theory · Plücker embedding · Finite frames

4.1 Introduction

Our goal in this chapter is to demonstrate that ideas from algebraic geometry can be used to obtain striking results in finite frame theory. Traditionally, the frame theory community has focused on tools from harmonic and functional analysis. By contrast, algebro-geometric techniques have only been exploited in the past few years because of the relatively recent interest in the theory of finite frames.

There are two central reasons why the frame theory community has begun to develop an extensive theory of finite frames. First, there is a hope within the community that a deeper understanding of finite frames may help resolve longstanding

J. Cahill
Department of Mathematics, University of Missouri, Columbia, MO, USA
e-mail: jccbd@mail.missouri.edu

N. Strawn (✉)
Department of Mathematics, Duke University, Durham, NC, USA
e-mail: nstrawn@math.duke.edu

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_4,

problems in infinite-dimensional frame theory (such as the Kadison-Singer problem [8, 22]). Second, computer implementations of frames are necessarily finite, and we must have a theory of finite frames to demonstrate that these implementations are accurate and robust (one manifestation of this is in the Paulsen problem [6]). It turns out that interesting families of finite frames can be identified with algebraic varieties. That is, they are solutions to systems of algebraic equations, or they live in equivalence classes of solutions to algebraic systems. For example, real Parseval frames satisfy the algebraic system of equations arising from the entries of $\Phi\Phi^T = Id$. In what follows, we shall apply ideas from algebraic geometry to study finite unit norm tight frames and Parseval frames.

Finite unit norm tight frames obey length constraints and a frame operator constraint. Maintaining the frame operator constraint is the most complex obstruction to parameterizing these spaces. A rather fruitful perspective on this constraint is obtained by considering the frame operator as a sum of dyadic products:

$$S = \sum_{i=1}^{M} \phi_i \phi_i^T .$$

Supposing that $\Lambda \subset [M]$ contains the indices of a basis inside of the frame Φ, we then have

$$\sum_{i \in \Lambda} \phi_i \phi_i^T = S - \sum_{i \in [M] \setminus \Lambda} \phi_i \phi_i^T .$$

By continuity, we should be able to locally articulate the ϕ_i's with indices in $[M] \setminus \Lambda$ while ensuring that the left-hand side of this equation remains a viable frame operator for the basis. As the free vectors move, the basis reacts elastically to maintain the overall frame operator. Additionally, the basis contributes extra degrees of freedom. It turns out that this intuition can be formalized, and tools from elimination theory can be used to explicitly compute the resulting coordinate systems on spaces of finite unit norm tight frames (more generally, frames with fixed vector lengths and a fixed frame operator). It should be noted that the chapter "Constructing Finite Frames with a Given Spectrum" also contained in this volume has coordinate systems where the free parameters directly control a system of eigensteps. In contrast, the coordinates derived in our chapter have free parameters that directly control the spatial location of frame vectors. We provide a technical justification for these coordinate systems by characterizing the tangent spaces (Theorem 4.3) on the space of finite unit norm tight frames (and more general frames), and then applying the real-analytic inverse function theorem (Theorem 4.4). An extensive example is provided to convey the central ideas behind these results.

Parseval frames which are equivalent up to an invertible transform can be identified with the Grassmannian variety, which allows us to define a concrete notion of distance between these equivalence classes. Using this distance, we can demonstrate that equivalence classes of generic frames (robust to $M - N$ arbitrary erasures [20]) are dense in the Grassmannian variety. Moreover, the Plücker embedding allows us to construct algebraic equations which characterize generic frames that are

also numerically maximally robust to erasures. Finally, we demonstrate that sufficient redundancy implies that the frames that can be used to solve the phaseless reconstruction problem form a dense subset.

4.1.1 Preliminaries

We shall now discuss the necessary preliminary concepts and notation. The Zariski topology is a fundamental idea in algebraic geometry. The zero sets of multivariate polynomials form a basis for the closed sets of the Zariski topology on $\mathscr{H}^n$. Thus, the closed sets are given by

$$\mathscr{C} = \left\{ C \subset \mathscr{H}^n : C = \bigcap_{i=1}^{k} p_i^{-1}\big(\{0\}\big) \text{ for some polynomials } \{p_i\}_{i=1}^{k} \right\}. \tag{4.1}$$

It is not difficult to deduce that this induces a topology [13]. An important property of this topology is that the nontrivial open sets are dense in the Euclidean topology.

We shall often use $[a]$ to denote the a-set $\{1, \ldots, a\}$, and $[a, b] = \{a, a + 1, \ldots, b\}$. For sets $P \subset [M]$ and $Q \subset [N]$ and any M by N matrix X, we let X_Q denote the matrix obtained by deleting the columns with indices outside of Q, and we let $X_{P \times Q}$ denote the matrix obtained by deleting entries with indices outside of $P \times Q$. For any submanifold $\mathscr{M}$ embedded in the space of M by N matrices, we set

$$T_X \mathscr{M} = \left\{ Y : Y = \frac{d}{dt}\gamma(t)\Big|_{t=0} \text{ for a smooth path } \gamma \text{ in } \mathscr{M} \text{ with } \gamma(0) = X \right\}.$$

4.2 Elimination Theory for Frame Constraints

Elimination theory consists of techniques for solving multivariate polynomial systems. Generally, one successively "eliminates" variables by combining equations. Variables are eliminated until a univariate polynomial is obtained, and then those solutions are used to "backsolve" and acquire all of the solutions to the multivariate system. Gaussian elimination is perhaps the most well-known application of elimination-theoretic techniques. Given a consistent system of linear constraints, Gaussian elimination can be carried out to produce a parameterization of the solution space. For higher-order polynomial systems, generalizing this kind of elimination can be quite tricky, but it simplifies in a few notable cases. For example, square roots allow us to construct locally well-defined coordinate systems for the space of solutions to a single spherical constraint:

$$\sum_{i=1}^{N} x_i^2 = 1 \quad \Longrightarrow \quad x_1 = \pm \sqrt{1 - \sum_{i=2}^{N} x_i^2}.$$

This example demonstrates that we can parameterize the top or bottom cap of a hypersphere in terms of the variables x_i for $i = 2, \ldots, N$. Note that these parameterizations are both valid as long as $\sum_{i=2}^{N} x_i^2 \leq 1$, and that they are also analytic inside of this region.

The finite unit norm tight frames of M vectors in $\mathbb{R}^N$ are completely characterized by the algebraic constraints

$$\phi_i^T \phi_i = \sum_{j=1}^{N} \phi_{ji}^2 = 1 \quad \text{for } i = 1, \ldots, N \quad \text{and} \quad \Phi \Phi^T = \frac{M}{N} Id_N,$$

and hence the space of finite unit norm tight frames is an algebraic variety. Moreover, these constraints are all quadratic constraints, so the space of finite unit norm tight frames is also a quadratic variety. Computing solutions for a general quadratic variety is NP-hard [10], but we shall soon see that spaces of finite unit norm tight frames often admit tractable local solutions.

Finite unit norm tight frames for $\mathbb{R}^2$ admit simple parameterizations because they can be identified with closed planar chains (see [3]).

Proposition 4.1 *For any frame Φ of M vectors in $\mathbb{R}^2$, identify $(\phi_i)_{i=1}^M$ with the sequence of complex variables $\{z_i\}_{i=1}^N$ with $\mathrm{Re}(z_i) = \phi_{1i}$ and $\mathrm{Im}(z_i) = \phi_{2i}$. Then Φ is a finite unit norm tight frame if and only if $|z_i|^2 = 1$ for $i = 1, \ldots, N$ and $\sum_{i=1}^{N} z_i^2 = 0$.*

To induce a parameterization on finite unit norm tight frames with M vectors in $\mathbb{R}^2$, we may place $M - 2$ links in a planar chain starting at the origin, and to close the chain with two links of length one, there are only finitely many viable solutions. This parameterization betrays the fact that the local parameterizations arise from locally arbitrary perturbation of $M - 2$ vectors. This intuition extends to finite unit norm tight frames of $\mathbb{R}^N$, but the reacting basis for $N > 2$ has nontrivial degrees of freedom.

More generally, for a list of squared vector lengths $\mu \in \mathbb{R}_+^M$ and a target frame operator S (a symmetric, positive definite N by N matrix), we may extend this intuition to the algebraic variety of frames with squared vector lengths indexed by μ and with frame operator S. We shall call these frames the (μ, S)-frames, and we let $\mathscr{F}_{\mu,S}$ denote the space of all such frames. The following majorization condition (introduced to the frame community in [7]) characterizes the μ and S such that $\mathscr{F}_{\mu,S}$ is nonempty, and we shall implicitly assume that μ and S satisfy this condition for the remainder of this section.

Theorem 4.1 *Let $\mu \in \mathbb{R}_+^M$ and let S denote an N by N symmetric positive definite operator. The space $\mathscr{F}_{\mu,S}$ is not empty if and only if*

$$\max_{\{A \subset [M] : |A| = k\}} \sum_{i \in A} \mu_i \leq \sum_{i=1}^{k} \lambda_i(S) \quad \textit{for all } k \in [N],$$

and $\sum_{i=1}^{M} \mu_i = \sum_{i=1}^{N} \lambda_i(S)$. *Here,* $\{\lambda_i(S)\}_{i=1}^{N}$ *are the eigenvalues of* S *listed in nonincreasing order.*

In this section, we shall rigorously validate the intuition that coordinates on $\mathscr{F}_{\mu,S}$ essentially arise from free articulation of $M - N$ vectors on a sphere, and restricted articulation of a basis. First, we shall present a simple example that depicts how formal local coordinates may be constructed on a space of frames with fixed vector lengths and a fixed frame operator. In order to validate the formal local coordinates, we first characterize the tangent spaces on these frame varieties and proceed to demonstrate injectivity of the tangent spaces onto candidate parameter spaces. This allows us to invoke the real-analytic inverse function theorem to ensure that locally well-defined real-analytic coordinate patches do exist. Finally, we use this existence result to validate explicit expressions for the formal coordinates. While all of these results are also true in the complex case, we shall only consider real frames, because the notation is less cumbersome, and the arguments are very similar.

4.2.1 A Motivating Example

In this example, we demonstrate how coordinates can be obtained for a space of bases in $\mathbb{R}^3$ with fixed lengths and a fixed frame operator. This is the simplest nontrivial case, but our approach requires a decent amount of effort. The benefit is that the approach of this example works in general, with minor modifications.

We consider the case $M = N = 3$,

$$\mu = \begin{bmatrix} 1 \\ 1 \\ 1 \end{bmatrix}, \quad \text{and} \quad \Phi = \begin{bmatrix} 1 & \sqrt{2}/2 & 0 \\ 0 & \sqrt{2}/2 & \sqrt{2}/2 \\ 0 & 0 & \sqrt{2}/2 \end{bmatrix},$$

$$S = \Phi\Phi^T = \begin{bmatrix} 3/2 & 1/2 & 0 \\ 1/2 & 1 & 1/2 \\ 0 & 1/2 & 1/2 \end{bmatrix}.$$

Let us count the constraints on $\mathscr{F}_{\mu,S}$ to determine its dimension as a manifold. Each of the three length conditions imposes a constraint. Because of symmetry, the frame operator condition imposes $3 + 2 + 1 = 6$ constraints. Since $\mathscr{F}_{\mu,S} \subset \mathbb{R}^{3\times 3}$, it would seem that this algebraic variety is zero dimensional. However, we have counted one of the constraints twice because $\text{trace}(S) = \sum \mu_i$. Thus, $\mathscr{F}_{\mu,S}$ is a one-dimensional algebraic variety. Consequently, we look for parameterizations of the form

$$\Phi(t) = \begin{bmatrix} \phi_{11}(t) & \phi_{12}(t) & \phi_{13}(t) \\ \phi_{21}(t) & \phi_{22}(t) & \phi_{23}(t) \\ t & \phi_{32}(t) & \phi_{33}(t) \end{bmatrix}, \qquad \Phi(0) = \Phi.$$

The constraints are $\operatorname{diag}(\Phi^T(t)\Phi(t)) = [111]^T$ and

$$\Phi(t)\Phi(t)^T = S \quad \Longleftrightarrow \quad \Phi(t)^T S^{-1}\Phi(t) = \Phi(t)^T \begin{bmatrix} 1 & -1 & 1 \\ -1 & 3 & -3 \\ 1 & -3 & 5 \end{bmatrix} \Phi(t) = Id_3.$$

We proceed inductively through the columns of $\Phi(t)$. The constraints that only involve the first column are the normality condition and the condition imposed by $S_{11} = 1$,

$$\begin{aligned} \phi_{11}^2 + \phi_{21}^2 + t^2 &= 1, \\ \phi_{11}^2 + 3\phi_{21}^2 + 5t^2 - 2\phi_{11}\phi_{21} + 2\phi_{11}t - 6\phi_{21}t &= 1. \end{aligned}$$

Viewing these two multinomials as polynomials in ϕ_{21} with coefficients in ϕ_{11} and t, we have

$$\begin{aligned} \phi_{21}^2 + \big(\phi_{11}^2 + t^2 - 1\big) &= 0, \\ 3\phi_{21}^2 + (-2\phi_{11} - 6t)\phi_{21} + \big(\phi_{11}^2 + 5t^2 + 2\phi_{11}t - 1\big) &= 0. \end{aligned}$$

To perform elimination on this system, we need to invoke the following proposition (which is a simple exercise using Gaussian elimination).

Proposition 4.2 *Suppose $\alpha_i, \beta_i \in \mathbb{R}$ for $i = 0, 1, 2$ and $\alpha_2, \beta_2 \neq 0$. The quadratics $p = \alpha_2\xi^2 + \alpha_1\xi + \alpha_0$ and $q = \beta_2\xi^2 + \beta_1\xi + \beta_0$ have a mutual zero if and only if the Bézout determinant satisfies*

$$Bz(p, q) := (\alpha_2\beta_1 - \alpha_1\beta_2)(\alpha_1\beta_0 - \alpha_0\beta_1) - (\alpha_2\beta_0 - \alpha_0\beta_2)^2 = 0. \tag{4.2}$$

Applying this proposition to the last two quadratics, we can eliminate ϕ_{21} to obtain

$$\begin{aligned} 0 &= \big[(1)(-2\phi_{11} - 6t) - (0)(3)\big]\big[(0)\big(\phi_{11}^2 + 5t^2 + 2\phi_{11}t - 1\big) \\ &\quad - \big(\phi_{11}^2 + t^2 - 1\big)(-2\phi_{11} - 6t)\big] - \big[(1)\big(\phi_{11}^2 + 5t^2 + 2\phi_{11}t - 1\big) \\ &\quad - (3)\big(\phi_{11}^2 + t^2 - 1\big)\big]^2 \\ &= 8\phi_{11}^4 + 16t\phi_{11}^3 + \big(36t^2 - 12\big)\phi_{11}^2 + \big(32t^3 - 16t\big)\phi_{11} + \big(40t^4 - 28t^2 + 4\big). \end{aligned}$$

Solving for ϕ_{11} in terms of t, we obtain the four possible solutions:

$$\phi_{11}(t) = \pm\sqrt{1 - 2t^2}, -t \pm \frac{1}{2}\sqrt{-6t^2 + 2}.$$

The condition $\phi_{11}(0) = 1$ leaves us with just one possible solution:

$$\phi_{11}(t) = \sqrt{1 - 2t^2},$$

and we readily verify that this implies $\phi_{21}(t) = t$. Having solved for the first column, we consider the constraints that have not been satisfied, but which only depend on the first and second columns:

$$\begin{aligned} \phi_{12}^2 + \phi_{22}^2 + \phi_{32}^2 &= 1, \\ \phi_{12}^2 + 3\phi_{22}^2 + 5\phi_{32}^2 - 2\phi_{12}\phi_{22} + 2\phi_{12}\phi_{32} - 6\phi_{22}\phi_{32} &= 1, \\ x^T S^{-1} y = \sqrt{1-2t^2}\phi_{12} - \sqrt{1-2t^2}\phi_{22} + \big(\sqrt{1-2t^2} + 2t\big)\phi_{32} &= 0. \end{aligned} \tag{4.3}$$

By continuity, we know that $\phi_{11}(t) \neq 0$ near $t = 0$, so we may solve the third equation for ϕ_{12} to obtain

$$\phi_{12} = \phi_{22} - \big(1 + 2t/\sqrt{1-2t^2}\big)\phi_{32}.$$

This allows us to eliminate ϕ_{12} from the first two equations, and we may view these new equations as quadratics in ϕ_{22} with coefficients in ϕ_{32} and t:

$$\begin{aligned} &2\phi_{22}^2 + \big[\big(-2 - 4t/\sqrt{1-2t^2}\big)\phi_{32}\big]\phi_{22} \\ &\quad + \big[\big(2 + 4t/\sqrt{1-2t^2} + 4t^2/\big(1-2t^2\big)\big)\phi_{32}^2 - 1\big] = 0, \\ &2\phi_{22}^2 + [-4\phi_{32}]\phi_{22} + \big[\big(4 + 4t^2/\big(1-2t^2\big)\big)\phi_{32}^2 - 1\big] = 0. \end{aligned}$$

We now solve for ϕ_{32} in terms of t so that the Bézout determinant of this system vanishes, and we obtain only three solutions,

$$\phi_{32}(t) = 0, \pm\frac{1}{2}\sqrt{2-4t^2}.$$

Since $\phi_{32}(0) = 0$, we are left with the solution $\phi_{32}(t) = 0$. Substitution into (4.3) immediately implies that $\phi_{12}(t) = \phi_{22}(t)$ for all t, so we must conclude that $\phi_{12}(t) = \phi_{22}(t) = \sqrt{2}/2$ for all t.

We now solve for the final column, ϕ_3. Noting that conditions on ϕ_2 are also imposed upon ϕ_3, we see that

$$\phi_{33}(t) = 0, \pm\frac{1}{2}\sqrt{2-4t^2}.$$

However, $\phi_{33}(0) = \sqrt{2}/2$, so we have that $\phi_{33}(t) = \frac{1}{2}\sqrt{2-4t^2}$. A similar line of reasoning reveals that

$$\phi_{23}(t) = \pm\sqrt{2}/2, \pm\frac{1}{2}\sqrt{2-4t^2}.$$

Invoking the orthogonality condition,

$$\phi_2^T S^{-1}\phi_3 = \sqrt{2}\phi_{23} - \sqrt{2}\phi_{33} = 0,$$

we may eliminate the constant solutions, and $\phi_{23}(0) = \sqrt{2}/2$ leaves us with $\phi_{23}(t) = \frac{1}{2}\sqrt{2-4t^2}$. Using the spherical condition $\phi_{13}^2 + \phi_{23}^2 + \phi_{33}^2 = 1$ and the orthogonality condition $x^T S^{-1} z = 0$, we obtain $\phi_{13}(t) = -\sqrt{2}t$. Thus, the final solution is

$$\Phi(t) = \begin{bmatrix} \sqrt{1-2t^2} & \sqrt{2}/2 & -\sqrt{2}t \\ t & \sqrt{2}/2 & \frac{1}{2}\sqrt{2-4t^2} \\ t & 0 & \frac{1}{2}\sqrt{2-4t^2} \end{bmatrix}.$$

This parameterization is relatively simple because the first and third columns form an orthonormal basis of $\text{span}\{\phi_1(0), \phi_3(0)\}$ for all t. If we had observed this at the beginning of the example, the parameterizations would follow very quickly. However, a generic frame does not contain an orthonormal basis and the approach of this example is generically effective.

We may immediately exploit the idea behind this example to construct formal coordinate systems around arbitrary frames in $\mathscr{F}_{\mu,S}$, However, it is not immediately clear that any of these formal coordinate systems are locally well defined. Our first challenge is to demonstrate that there are unique, valid coordinate systems. We shall then endeavor to identify these with the formal solutions that can be constructed in a manner echoing this example.

4.2.2 Tangent Spaces on $\mathscr{F}_{\mu,S}$

We first turn our attention to the problem of characterizing the tangent spaces of $\mathscr{F}_{\mu,S}$. The reason for this is twofold. First, if the tangent is not well defined, then we are not guaranteed that the algebraic variety is locally diffeomorphic to an open subset of Euclidean space. That is, smooth coordinate charts may not be available. The second reason is that we have a procedure for constructing formal coordinate systems (as illustrated by the preceding example), but we would like to know that these formal coordinate systems are actually valid in some open neighborhood. To obtain this validation, we want to demonstrate injectivity of a Jacobian in order to invoke a form of the inverse function theorem. Demonstrating the injectivity ensures that our coordinate map does not collapse or exhibit a pinched point, and we have to characterize the tangent spaces of $\mathscr{F}_{\mu,S}$ to carry out the demonstration.

For $\mu \in \mathbb{R}_+^M$ and N by N symmetric positive definite S, let

$$\mathbb{T}_{\mu,N} = \left\{\Phi = (\phi_i)_{i=1}^M \subset \mathbb{R}^N : \|\phi_i\|^2 = \mu_i \text{ for all } i = 1, \ldots, M\right\}$$

and

$$\text{St}_{S,M} = \left\{\Phi = (\phi_i)_{i=1}^M \subset \mathbb{R}^N : \Phi\Phi^T = S\right\}$$

denote the generalized torus and generalized Stiefel manifold respectively. For brevity, we shall simply call these the torus and the Stiefel manifold. Clearly, we

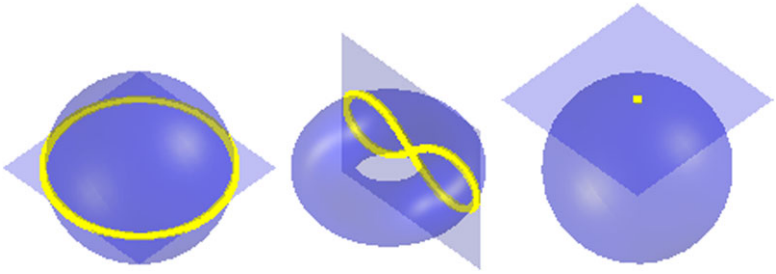


Fig. 4.1 Full transversality of an intersection (*left*) ensures that the intersection forms a manifold with a formula for the dimension. The *central* figure demonstrates that local failure of transversality results in crossings inside the intersection (the lemniscate). On the *right*, we see that degeneracy occurs when transversality fails completely

have that

$$\mathscr{F}_{\mu,S} = \mathbb{T}_{\mu,N} \cap \mathrm{St}_{S,M}.$$

Suppose that $\mathscr{F}_{\mu,S}$ is nonempty, set $c = \sum_{i=1}^{M} \mu_i$, and define the Frobenius sphere of square radius c by

$$\mathscr{S}_{M,N,c} = \left\{ \Phi = (\phi_i)_{i=1}^{M} \subset \mathbb{R}^N : \sum_{i=1}^{N} \|\phi_i\|^2 = c \right\}.$$

Then, we have the following inclusion diagram:

$$\begin{array}{ccccc} & & \mathscr{S}_{M,N,c} & & \\ & \nearrow & & \nwarrow & \\ \mathrm{St}_{S,M} & & & & \mathbb{T}_{\mu,N}. \\ & \nwarrow & & \nearrow & \\ & & \mathscr{F}_{\mu,S} & & \end{array}$$

In order to demonstrate that formal coordinates are valid using the implicit function theorem, we shall require an explicit characterization of the tangent space $T_\Phi \mathscr{F}_{\mu,S}$ for a given $\Phi \in \mathscr{F}_{\mu,S}$. Given the inclusion diagram, it is natural to ask when

$$T_\Phi \mathscr{F}_{\mu,S} = T_\Phi \mathbb{T}_{\mu,N} \cap T_\Phi \mathrm{St}_{S,M}.$$

That is, when is the tangent space of the intersection equal to the intersection of the tangent spaces? The notion of transversal intersection is our starting point for approaching this question (see [12]).

Definition 4.1 Suppose that $\mathscr{M}$ and $\mathscr{N}$ are smooth submanifolds of the smooth manifold $\mathscr{K}$, and let $x \in \mathscr{M} \cap \mathscr{N}$. We say that $\mathscr{M}$ and $\mathscr{N}$ intersect transversally at x in $\mathscr{N}$ if $T_x\mathscr{K} = T_x\mathscr{M} + T_x\mathscr{N}$. Here, $+$ is the Minkowski sum.

Theorem 4.2 *Suppose that $\mathscr{M}$ and $\mathscr{N}$ are smooth submanifolds of the smooth manifold $\mathscr{K}$, and let $x \in \mathscr{M} \cap \mathscr{N}$. If $\mathscr{N}$ and $\mathscr{M}$ intersect transversally at x in $\mathscr{K}$, then $T_x(\mathscr{M} \cap \mathscr{N})$ is well defined and*

$$T_x(\mathscr{M} \cap \mathscr{N}) = T_x\mathscr{M} \cap T_x\mathscr{N}.$$

That is, the tangent space of the intersection is the intersection of the tangent spaces.

Figure 4.1 provides a visualization of this theorem. To exploit this theorem, we must first determine $T_\Phi \mathscr{S}_{M,N,c}$, $T_\Phi \mathbb{T}_{\mu,N}$, and $T_\Phi \mathrm{St}_{S,M}$. The tangent space for the sphere at Φ is simply the set of matrices "orthogonal" to Φ, or

$$T_\Phi \mathscr{S}_{M,N,c} = \left\{ X = (x_i)_{i=1}^M \subset \mathbb{R}^N : \sum_{i=1}^N \langle x_i, \phi_i \rangle = 0 \right\}.$$

Since the tangent space of a product of manifolds is the product of the tangent spaces, we also have that

$$T_\Phi \mathbb{T}_{\mu,N} = \left\{ X = (x_i)_{i=1}^M \subset \mathbb{R}^N : \langle x_i, \phi_i \rangle = 0 \text{ for all } i = 1, \ldots, N \right\}.$$

The most convenient characterization of $T_\Phi \mathrm{St}_{S,M}$ is obtained by noting that the special orthogonal group $SO(N)$ acts on $\mathrm{St}_{S,M}$ on the right: $(U, \Phi) \mapsto \Phi U$. Since the Lie algebra of $SO(N)$ is the skew-symmetric matrices, it is not difficult to show that

$$T_\Phi \mathrm{St}_{S,M} = \left\{ X = (x_i)_{i=1}^M \subset \mathbb{R}^N : X = \Phi Z, \text{ where } Z = -Z^T \right\}.$$

Having characterized these tangent spaces, we now turn to the problem of characterizing the $\Phi \in \mathscr{F}_{\mu,S}$ at which $\mathbb{T}_{\mu,N}$ and $\mathrm{St}_{S,M}$ intersect transversally in $\mathscr{S}_{M,N,c}$. It turns out that the "bad" Φ are exactly the orthodecomposable frames (see [9]).

Definition 4.2 A frame Φ is said to be *orthodecomposable* if it can be split into two nontrivial subcollections, Φ_1 and Φ_2 satisfying $\Phi_1^* \Phi_2 = 0$. That is, $\operatorname{span} \Phi_1$ and $\operatorname{span} \Phi_2$ are nontrivial orthogonal subspaces.

Clearly, orthodecomposability is intimately related to the correlation structure of the frame's members. In order to demonstrate this equivalence, we shall require the notion of a frame's correlation network.

Definition 4.3 The *correlation network* of a frame $\Phi = (\phi_i)_{i=1}^M$ is the undirected graph $\gamma(\Phi) = (V, E)$, where $V = [M]$ and $(i, j) \in E$ if and only if $\langle \phi_i, \phi_j \rangle$ is nonzero.

Example 4.1 For the Φ defined in the example,

$$\left[\langle \phi_i, \phi_j \rangle \right]_{(i,j) \in [3]^2} = \Phi^T \Phi = \begin{bmatrix} 1 & \sqrt{2}/2 & 0 \\ \sqrt{2}/2 & 1 & 1/2 \\ 0 & 1/2 & 1 \end{bmatrix}.$$

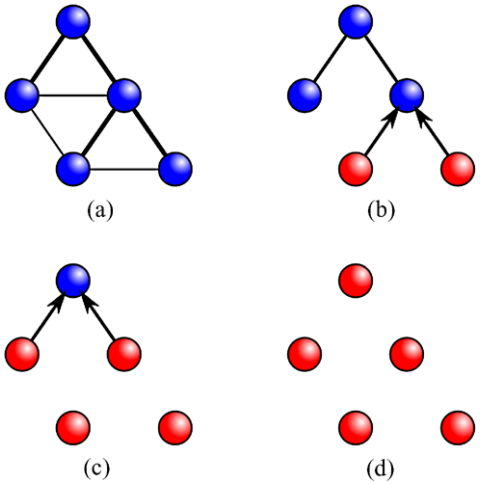


Fig. 4.2 (**a**) A rooted spanning tree is extracted from $\gamma(\Phi)$. Set $z_{ij} = 0$ if (i, j) is not in the spanning tree. (**b**) At nodes whose only children are leaves, fix entries of Z so that (4.4) holds for all these children. Effectively remove these children from the tree. (**c**) Inductively apply (**b**) until only the root remains. (**d**) The conditions on Y ensure that the final equation holds. At this point, all entries of Z have been defined

We conclude that $\gamma(\Phi) = (\{1, 2, 3\}, \{(1, 2), (2, 3)\})$, since ϕ_1, ϕ_3 is the only orthogonal (uncorrelated) pair.

We can now state the main theorem which relates the transversality of the intersection at Φ, the connectivity of the correlation network $\gamma(\Phi)$, and the orthodecomposability of Φ. This result is due to Strawn [21].

Theorem 4.3 *Suppose $\Phi \in \mathscr{F}_{\mu,S}$. Then the following are equivalent*:

(i) $T_\Phi \mathscr{S}_{M,N,c} = T_\Phi \mathbb{T}_{\mu,N} + T_\Phi \mathrm{St}_{S,M}$;
(ii) *For all $Y \in T_\Phi \mathscr{S}_{M,N,c}$, there is a skew-symmetric $Z = [z_{ij}]$ which is a solution to the system*

$$\langle y_i, \phi_i \rangle = \sum_{j \in [M]} z_{ji} \langle \phi_i, \phi_j \rangle \quad \textit{for all } i \in [M]; \tag{4.4}$$

(iii) *Φ is not orthodecomposable*;
(iv) *$\gamma(\Phi)$ is connected.*

The proof of this theorem is fairly straightforward, but its technical details obfuscate the simple intuition. The centerpiece of the argument involves an algorithm for constructing a solution to (4.4) given that $\gamma(\Phi)$ is connected. Because Z is skew-symmetric, this procedure can be interpreted as an algorithm for distributing specified amounts of a resource at the nodes of the correlation network. We illustrate this algorithm in Fig. 4.2.

From this theorem, we immediately obtain a characterization of the tangent spaces of $\mathscr{F}_{\mu,S}$ at non-orthodecomposable frames.

Corollary 4.1 *Assuming that $\Phi \in \mathscr{F}_{\mu,S}$ is not orthodecomposable, we have*

$$\begin{aligned} T_\Phi \mathscr{F}_{\mu,S} &= T_\Phi \mathbb{T}_{\mu,N} \cap T_\Phi \mathrm{St}_{S,M} \\ &= \left\{ X = (x_i)_{i=1}^M \subset \mathbb{R}^N : X = \Phi Z, Z = -Z^T, \mathrm{diag}\left(\Phi^* X\right) = 0 \right\}. \end{aligned} \tag{4.5}$$

4.2.3 Existence of Locally Well-Defined Parameterizations on $\mathscr{F}_{\mu,S}$

Now that we have characterized the tangent spaces on $\mathscr{F}_{\mu,S}$, we proceed to construct a linear map π and a linear parameter space $\Omega \oplus \Delta$ for each non-orthodecomposable $F \in \mathscr{F}_{\mu,S}$ so that $\pi : T_F \mathscr{F}_{\mu,S} \to \Omega \oplus \Delta$ (the Jacobian of $\pi : \mathscr{F}_{\mu,S} \to \Omega \oplus \Delta$) is injective and hence the map $\pi : \mathscr{F}_{\mu,S} \to \Omega \oplus \Delta$ has a locally well-defined inverse by the inverse function theorem [16]. This allows us to conclude that our formal procedure produces valid coordinate systems.

We begin by noting that, by counting the governing constraints, the dimension of a generic nonempty $\mathscr{F}_{\mu,S}$ is

$$\begin{aligned} \dim(\mathscr{F}_{\mu,S}) &= \dim \mathbb{T}_{\mu,N} + \dim \mathrm{St}_{S,M} - \dim \mathscr{S}_{M,N,c} \\ &= (N-1)M + \sum_{i=1}^{M} (M-i) - (MN-1) \\ &= (N-1)(M-N) + \sum_{i=1}^{M-2} i. \end{aligned}$$

Based on our initial example, this calculation, and a little intuition, we expect that it may be possible to obtain a parameterization of the form $\Phi(\Theta, L) = [\Gamma(\Theta) B(\Theta, L)]$, where

$$\begin{aligned} L \in \Delta_N &= \left\{ \delta = (\delta_i)_{i=1}^N \subset \mathbb{R}^N : \delta_{ij} = 0 \text{ if } i \leq j+1 \right\}, \\ \Theta \in \Omega_{M,N} &= \left\{ \omega = (\omega_i)_{i=1}^{M-N} \subset \mathbb{R}^N : \omega_{1i} = 0 \text{ for all } i = 1, \ldots, M-N \right\}, \\ \Gamma(\Theta) &= \begin{bmatrix} \phi_{11}(\theta_1) & \phi_{12}(\theta_2) & \cdots & \phi_{1,M-N}(\theta_{M-N}) \\ \theta_{21} & \theta_{22} & \cdots & \theta_{2,M-N} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_{N1} & \theta_{N2} & \cdots & \theta_{N,M-N} \end{bmatrix}, \end{aligned}$$

and where $B(\Theta, L)$ has the form

$$\begin{bmatrix} \phi_{1,M-N+1} & \phi_{1,M-N+2} & \cdots & \phi_{1,M-3} & \phi_{1,M-2} & \phi_{1,M-1} & \phi_{1M} \\ \phi_{2,M-N+1} & \phi_{2,M-N+2} & \cdots & \phi_{2,M-3} & \phi_{2,M-2} & \phi_{2,M-1} & \phi_{2M} \\ l_{31} & \phi_{3,M-N+2} & \cdots & \phi_{3,M-3} & \phi_{3,M-2} & \phi_{3,M-1} & \phi_{3M} \\ l_{41} & l_{42} & \cdots & \phi_{4,M-3} & \phi_{4,M-2} & \phi_{4,M-1} & \phi_{4M} \\ \vdots & \vdots & \ddots & \vdots & \vdots & \vdots & \vdots \\ l_{N-1,1} & l_{N-1,2} & \cdots & l_{N-1,N-3} & \phi_{N-1,M-2} & \phi_{N-1,M-1} & \phi_{N-1,M} \\ l_{N1} & l_{N2} & \cdots & l_{N,N-3} & l_{N,N-2} & \phi_{N,M-1} & \phi_{NM} \end{bmatrix}.$$

Here, Γ represents the vectors that may be freely perturbed within their sphere, and B parameterizes the basis. Note that $\Gamma(\Theta)$ and $B(\Theta, L)$ are N by $M - N$ and N by N arrays respectively.

In order to exploit this parameter space, we must rotate all of the vectors of Φ so that the resulting tangent space is sufficiently aligned with $\Omega_{M,N} \oplus \Delta_N$. Otherwise, we shall fail to acquire a parameterization with the form we have just described. Notationally, this system of rotations is represented as an array of orthogonal matrices:

$$\mathbf{Q} = (Q_i)_{i=1}^{M} \subset O^M(N).$$

The alignment of the frame Φ using the system of rotations $\mathbf{Q}$ is denoted

$$\mathbf{Q} \star \Phi = (Q_i \phi_i)_{i=1}^{M},$$

and we set $\mathbf{Q}^T = (Q_i^T)_{i=1}^{M}$.

This next theorem (also due to Strawn [21]) sets the stage for applying the real-analytic inverse function theorem [16] by demonstrating injectivity of the Jacobian. In particular, it allows us to know how and when we may use the parameter space $\Omega_{M,N} \oplus \Delta_N$ to obtain coordinates on $\mathscr{F}_{\mu,S}$.

Theorem 4.4 *Suppose $\Phi \in \mathscr{F}_{\mu,S}$ is not orthodecomposable. Then there is a system of rotations $\mathbf{Q} \in O^M(N)$ and an M by M permutation matrix P such that the orthogonal projection*

$$\pi : \mathbf{Q}^T \star T_{\Phi P^T} \mathscr{F}_{P\mu,S} \to \Omega_{M,N} \oplus \Delta_N$$

is injective.

By the real-analytic inverse function theorem, we obtain the following corollary, which ensures that our procedure for constructing formal coordinates (as in the first example) might actually produce well-defined coordinate systems.

Corollary 4.2 *If the conditions of Theorem 4.4 are satisfied, then there is a unique, locally well-defined, real-analytic inverse of π, $\Phi' : \Omega_{M,N} \oplus \Delta_N \to \mathbf{Q}^T \star \mathscr{F}_{P\mu,S}$.*

Remark 4.1 If Φ' is as in the above corollary, then $(\mathbf{Q} \star \Phi'(\Theta, L))P$ is a parameterization around $\Phi \in \mathscr{F}_{\mu,S}$.

The proof of Theorem 4.4 is rather technical, but a simple example should illustrate the nuances. Consider the frame

$$\Phi = \begin{bmatrix} 1 & 1 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & 0 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & 0 & -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \end{bmatrix},$$

so that $\mu = [1111]^T$ and

$$S = \begin{bmatrix} \frac{8}{3} & \frac{2}{3} & 0 \\ \frac{2}{3} & \frac{2}{3} & 0 \\ 0 & 0 & \frac{2}{3} \end{bmatrix}.$$

Our first goal is to identify a non-orthodecomposable basis inside of Φ. Note that the existence of such a basis is equivalent to the connectivity of $\gamma(\Phi)$. We set

$$B = [\phi_2 \quad \phi_3 \quad \phi_4] = \begin{bmatrix} 1 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & -\frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \end{bmatrix}.$$

Our next gadget is a rooted tree on $\gamma(B)$. We simply set 4 to be the root of this tree, and 2 and 3 are the children. Let T denote this tree. We have chosen T in this manner so as to illustrate typical behavior.

Now, the permutation matrix P in Theorem 4.4 is then chosen so that P^T moves all of the "free" vectors to the left side of ΦP^T, and also so that if i is a child of j in T, then the ith vector precedes the jth vector. By our choice of Φ and T, we simply have that $P = Id$. Next, we fix the alignment matrices.

The alignment matrices of the "free" vectors are simply chosen so that $Q_i e_1 = \phi_i / \|\phi_i\|$. Choosing the alignment matrices for the basis is more complicated. In our case, $Q_1 = Id$ since $\phi_1 = e_1$. We now choose Q_2 so that

$$[\phi_2 \quad \phi_4 \quad \phi_3] = [\phi_2 \quad \phi_3 \quad \phi_4] P_{(23)} = Q_2 R_2$$

is the QR decomposition of B after we permute the second and third column. Note that we have permuted so that the ϕ_2 is followed by the vector whose index is the parent of 2 in T. It is simple to check that

$$\begin{bmatrix} 1 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & \frac{\sqrt{3}}{3} & -\frac{\sqrt{3}}{3} \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix} \begin{bmatrix} 1 & \frac{\sqrt{3}}{3} & \frac{\sqrt{3}}{3} \\ 0 & \frac{\sqrt{6}}{3} & 0 \\ 0 & 0 & \frac{\sqrt{6}}{3} \end{bmatrix},$$

and hence

$$Q_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ 0 & -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{bmatrix}.$$

The final two alignment matrices are always set to the identity, so $Q_3 = Q_4 = Id$.

Now, we set about demonstrating that the projection from this theorem is injective. Suppose that $X \in \mathbf{Q}^T \star T_\Phi \mathscr{F}_{\mu,S}$ satisfies $\pi(X) = 0$, and hence

$$X = \begin{bmatrix} x_{11} & x_{12} & x_{13} & x_{14} \\ 0 & x_{22} & x_{23} & x_{24} \\ 0 & 0 & x_{33} & x_{34} \end{bmatrix}.$$

Because $\gamma(\Phi)$ is connected, we know that

$$T_\Phi \mathscr{F}_{\mu,S} = \left\{ Y : Y = \Phi Z, Z = -Z^t, \operatorname{diag}\left(\Phi^T \Phi Z\right) = 0 \right\}.$$

In particular, we have that $X = \mathbf{Q}^T \star (\Phi Z)$ for some $Z = -Z^T$. We shall show that $Z = 0$ by induction though its columns. First, we show that we may choose Z so that $z_1 = 0$. We first note that $x_1 = \Phi z_1$. Since $\operatorname{diag}(\Phi^T \Phi Z) = 0$, we have

$$0 = \phi_1^T \Phi z_1 = e_1^T \Phi z_1 = e_1^T x_1 = x_{11}.$$

Consequently, $x_1 = 0$ and it turns out that we can assume $z_1 = 0$ in this case. The details of this are described in the full proof, but one may think of this as saying that any motion that fixes "free" vectors only needs to know how it is acting on the basis. Now, we show that $z_2 = 0$. We have that

$$P_{(23)} \begin{bmatrix} 0 \\ z_{32} \\ z_{42} \end{bmatrix} = R_2^{-1} x_2 = \begin{bmatrix} 1 & -\frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ 0 & \frac{\sqrt{6}}{2} & 0 \\ 0 & 0 & \frac{\sqrt{6}}{2} \end{bmatrix} \begin{bmatrix} x_{12} \\ x_{22} \\ 0 \end{bmatrix} = \begin{bmatrix} x_{12} - \frac{\sqrt{2}}{2} x_{22} \\ \frac{\sqrt{6}}{2} x_{22} \end{bmatrix}.$$

This means that

$$z_2 = \begin{bmatrix} 0 \\ 0 \\ 0 \\ z_{42} \end{bmatrix}.$$

Now, the other condition on $T_\Phi \mathscr{F}_{\mu,S}$, $\operatorname{diag}(\Phi^T \Phi Z) = 0$ implies that $\phi_2^T \Phi z_2 = 0$. But since we have $z_2 = z_{42} e_4$, this condition reduces to

$$z_{42} \phi_2^T \phi_4 = 0.$$

In the spanning tree of the correlation network, 4 is the parent of 2, so we have that $\phi_2^T \phi_4 \neq 0$. Therefore $z_{42} = 0$, and hence $z_2 = 0$. Repeating this trick gives us

$z_3 = 0$ as well; the last three entries of z_3 are $\Phi^{-1}x_3$, which implies that the only nonzero entry of z_3 is z_{43}, and the diagonal condition ensures that $z_{43} = 0$. Finally, $z_4 = 0$ since $Z = -Z^T$.

We have shown that $Z = 0$, so it follows that $X = 0$ and π is injective. After counting dimensions, we invoke the real-analytic inverse function theorem to obtain unique, analytic, locally well-defined coordinates. This guarantees us that our formal solutions to this system are locally valid. We now proceed to elucidate the explicit construction of formal solutions.

4.2.4 Deriving Explicit Coordinates on $\mathscr{F}_{\mu,S}$

Using the same Φ in our last example, we set

$$\phi_1 = \begin{bmatrix} \sqrt{1 - \phi_{21}^2 - \phi_{31}^2} \\ \phi_{21} \\ \phi_{31} \end{bmatrix}.$$

Our only condition imposed upon the "free" vector is that it remain in its sphere. However, as we move ϕ_1, the frame operator of the basis $[\phi_2\ \phi_3\ \phi_4]$ must change to maintain the overall frame operator. Explicitly, we want to enforce the constraint

$$S = \phi_1\phi_1^T + \phi_2\phi_2^T + \phi_3\phi_3^T + \phi_4\phi_4^T,$$

so we must have that

$$BB^T = \phi_2\phi_2^T + \phi_3\phi_3^T + \phi_3\phi_3^T = S - \phi_1\phi_1^T.$$

Since B is invertible, we can rearrange to obtain

$$B^T\left(S - \phi_1\phi_1^T\right)^{-1}B = Id.$$

By rearranging in this manner, all of the conditions on the basis become conditions on the columns. This is the central trick that supplies us with a strategy for carrying out the full derivation of the explicit coordinate systems. With this rearrangement, it is now possible to solve the entire system in a column-by-column fashion.

For this trick to work we must compute $(S - \phi_1\phi_1^T)^{-1}$. The entries of this inverse are analytic functions, but they are already complicated. While we may be able to fit this expression on a page, we only have one "free" vector to consider. With an arbitrary number of "free" vectors, one can easily see that this inverse has a very dense representation. Even if we were simply solving a linear system involving the basis and this inverse, the full expression would be vast. In our situation, we're going to solve two quadratic equations and a linear system. This dramatically inflates the complexity of the explicit expressions.

Forgoing the explicit form of $(S - \phi_1\phi_1^T)$, we now consider the conditions that must be imposed upon just ϕ_2:

$$\phi_2^T\phi_2 = 1 \quad \text{and} \quad \phi_2^T\left(S - \phi_1\phi_1^T\right)^{-1}\phi_2 = 1.$$

The first is a spherical constraint, and the second is an ellipsoidal constraint. In general, the solution set in $\mathbb{R}^3$ bears a striking resemblance to the boundary of a Pringles chip. Because of the alignment structure, we set $\phi_2 = Q_2\psi$ and solve

$$\psi^T\psi = 1 \quad \text{and} \quad \psi^T Q_2^T\left(S - \phi_1\phi_1^T\right)^{-1}Q_2\psi = 1,$$

where

$$\psi = \begin{bmatrix} \psi_1(t, \phi_1) \\ \psi_2(t, \phi_1) \\ t \end{bmatrix}.$$

As in our first example, these are two quadratic constraints and we may apply the Bézout determinant trick to obtain explicit expressions for ψ_1 and ψ_2. The resulting expressions are entirely dependent upon ϕ_1 and t. We may then set $\phi_2 = Q_2^T\psi$. With ϕ_2 in hand, we can then solve for ϕ_3 and ϕ_4 just like we did in our first example. The astute reader will recognize that we obtain numerous branches from solving these equations. However, we may prune these branches by considering the condition $\Phi(0, 0, 0) = \Phi$.

While we may write explicit expressions for these coordinate systems, these expressions will necessarily involve solutions to quartic equations, which are unwieldy when expressed in their full form. For our example, some of the expressions are so vast that they exceed LaTeX's allowed buffer. Nevertheless, computer algebra packages can manage the expressions. For a full technical derivation of these coordinates and a full proof that there is a unique branch with local validity, the reader is referred to [21].

Since the expressions for our example are too large to fit on a page, we conclude this section with Fig. 4.3, which depicts the motion that frame vectors experience as we traverse the local coordinate system. In this figure, we allow ϕ_1 to move along a great circle and allow t to vary fully. Consequently, we observe the individual basis vectors articulating along two-dimensional sheets inside the unit sphere. There are of course three degrees of freedom for this example, but it is much harder to visualize the behavior obtained by submerging a three-dimensional space in a sphere.

4.3 Grassmannians

In this section we will study a family of well-known varieties called Grassmannians. These results originally appeared in [5]. The *Grassmannian* is defined as the set $\{N\text{-dimensional subspaces of } \mathscr{H}^M\}$ and will be denoted by $Gr(M, N)$. It is not clear from this definition how this set forms a variety, but this will be explained

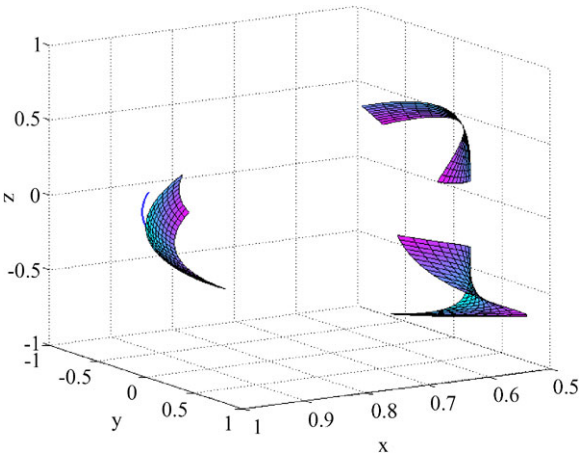


Fig. 4.3 In this figure, we have allowed ϕ_1 (the *small blue curve* near the sheet on the *left side*) to vary along a fixed curve, and the movement of ϕ_2 controls the single degree of freedom inside the basis. Consequently, ϕ_2, ϕ_3, and ϕ_4 carve out two-dimensional sheets on the unit sphere

shortly. The motivation for the use of Grassmannians in frame theory comes from the following proposition (see [1, 15]).

Proposition 4.3 *Two frames are isomorphic if and only if their corresponding analysis operators have the same image.*

Therefore a point on a Grassmannian corresponds to an entire isomorphism class of frames, but many properties of frames are invariant under isomorphisms, so Grassmannians can give a useful way to discuss families of frames with certain properties.

In this first section we will explain some basic properties of Grassmannians. Most of this material is well known, so we will provide appropriate references for technical details that are not included here.

First we will be concerned with the Grassmannian as a metric space. If $\mathscr{X}, \mathscr{Y} \in Gr(M, N)$ then $\|P_{\mathscr{X}} - P_{\mathscr{Y}}\|$ defines a metric on $Gr(M, N)$, where $P_{\mathscr{X}}$ denotes the orthogonal projection of $\mathscr{H}^M$ onto $\mathscr{X}$, and $\|\cdot\|$ denotes the usual operator norm. This metric has a geometric interpretation in terms of the "angle" between $\mathscr{X}$ and $\mathscr{Y}$. Define the N-tuple $(\sigma_1, \ldots, \sigma_k)$ as follows:

$$\sigma_1 = \max\big\{\langle x, y\rangle : x \in \mathscr{X}, y \in \mathscr{Y}, \|x\| = \|y\| = 1\big\} = \langle x_1, y_1\rangle,$$

and

$$\begin{aligned}\sigma_i &= \max\big\{\langle x, y\rangle : x \in \mathscr{X}, y \in \mathscr{Y}, \|x\| = \|y\| = 1, \langle x, x_j\rangle = \langle y, y_j\rangle = 0, j < i\big\} \\ &= \langle x_i, y_i\rangle\end{aligned}$$

for $i > 1$. Now define $\theta_i(\mathscr{X}, \mathscr{Y}) = \cos^{-1}(\sigma_i)$. The N-tuple $\theta(\mathscr{X}, \mathscr{Y}) = (\theta_1, \ldots, \theta_N)$ is called the *principal angles* between $\mathscr{X}$ and $\mathscr{Y}$ (some authors call these the *canonical angles*). Let X and Y be $N \times M$ matrices whose rows form orthonormal bases for $\mathscr{X}$ and $\mathscr{Y}$ respectively. It turns out that the σ_i's are precisely the singular values of XY^*. We also have that $\|P_{\mathscr{X}} - P_{\mathscr{Y}}\| = \sin(\theta_N(\mathscr{X}, \mathscr{Y}))$. In fact, there are many metrics that can be defined in terms of the principal angles. Justifications for these three facts can be found in [15].

We now proceed to explain a particular embedding, known as the Plücker embedding, of $Gr(M, N)$ into $\mathbb{P}^{\binom{M}{N}-1}$ which will be used extensively in this section. Let $\mathscr{X} \in Gr(M, N)$ and let $X^{(1)}$ be any $N \times M$ matrix whose rows form a basis for $\mathscr{X}$. Let $X^{(1)}_{i_1 \cdots i_N}$ be the $N \times N$ minor consisting of the columns indexed by $i_1, \ldots, i_N$ of $X^{(1)}$. Then the $\binom{M}{N}$-tuple $Plu(X^{(1)}) = (\det(X^{(1)}_{i_1 \cdots i_N}))_{1 \le i_1 < \cdots < i_N \le M}$ is called the Plücker coordinates of $\mathscr{X}$. Note that if $X^{(2)}$ is any other $N \times M$ matrix whose rows span $\mathscr{X}$, then there exists an invertible $N \times N$ matrix A such that $(X^{(2)}) = A(X^{(1)})$. It follows that $Plu(X^{(2)}) = \det(A) Plu(X^{(1)})$. Thus the mapping $\mathscr{X} \mapsto Plu(\mathscr{X})$ is a well-defined injective mapping of $Gr(M, N)$ into $\mathbb{P}^{\binom{M}{N}-1}$. In most cases this mapping is not onto, however the image of this mapping is known to be a projective variety; see [11] for more details. In particular, the vanishing locus of the polynomials

$$x_{i_1 \ldots i_N} x_{j_1 \ldots j_N} - \sum_{k=1}^{N} x_{j_k i_2 \ldots i_N} x_{j_1 \ldots j_{k-1} i_1 j_{k+1} \ldots j_N}$$

(where $x_{\sigma(i_1) \ldots \sigma(i_N)} = \operatorname{sign}(\sigma) x_{i_1 \ldots i_N}$ for any permutation σ) is precisely in the image of the Plücker embedding. We use the symbol $Plu(M, N)$ to denote this set of polynomials.

By abuse of notation let $Plu(\mathscr{X})$ denote a unit vector in $\mathscr{H}^{\binom{M}{N}}$. Then we have that $|\langle Plu(\mathscr{X}), Plu(\mathscr{Y})\rangle|$ is well defined for any $\mathscr{X}, \mathscr{Y} \in Gr(M, N)$. We define the Plücker angle between $\mathscr{X}$ and $\mathscr{Y}$ to be

$$\Theta(\mathscr{X}, \mathscr{Y}) = \cos^{-1} \left| \left\langle Plu(\mathscr{X}), Plu(\mathscr{Y}) \right\rangle \right|.$$

We have the following relationship between $\Theta(\mathscr{X}, \mathscr{Y})$ and $\theta(\mathscr{X}, \mathscr{Y})$, see [14], [17], and [18]:

Proposition 4.4

$$\cos\big(\Theta(\mathscr{X}, \mathscr{Y})\big) = \prod_{i=1}^{k} \cos\big(\theta_i(\mathscr{X}, \mathscr{Y})\big). \tag{4.6}$$

Proof Let X and Y be $N \times M$ matrices whose rows form orthonormal bases for $\mathscr{X}$ and $\mathscr{Y}$ respectively. Then

$$\begin{aligned}\cos\big(\Theta(\mathscr{X},\mathscr{Y})\big) &= \big|\langle Plu(\mathscr{X}), Plu(\mathscr{Y})\rangle\big| \\ &= \left|\sum_{1\le i_1<\cdots<i_N\le M} \det(X_{i_1\ldots i_N})\det(Y_{i_1\ldots i_N})\right| \\ &= \left|\sum_{1\le i_1<\cdots<i_N\le M} \det\big(X_{i_1\ldots i_N}Y^*_{i_1\ldots i_N}\big)\right| \\ &= \big|\det\big(XY^*\big)\big| = \big|\det(U\Sigma V)\big| = \det(\Sigma) \\ &= \prod_{i=1}^{N}\sigma_i = \prod_{i=1}^{N}\cos\big(\theta_i(\mathscr{X},\mathscr{Y})\big),\end{aligned}$$

where $XY^* = U\Sigma V$ is a singular value decomposition, and where we have employed the Cauchy-Binet formula for the fourth equality. □

In particular, $\theta_i(\mathscr{X},\mathscr{Y}) \le \Theta(\mathscr{X},\mathscr{Y})$ for every $i = 1, 2, \ldots, N$, $\Theta(\mathscr{X},\mathscr{Y}) = \frac{\pi}{2}$ if and only if $\theta_N(\mathscr{X},\mathscr{Y}) = \frac{\pi}{2}$, and $\Theta(\mathscr{X},\mathscr{Y}) = 0$ if and only if $\theta_N(\mathscr{X},\mathscr{Y}) = 0$. Also, we have the following new metric on $Gr(M,N)$:

$$d(\mathscr{X},\mathscr{Y}) = \big\|Plu(\mathscr{X}) - Plu(\mathscr{Y})\big\| = 2\sin\left(\frac{\Theta(\mathscr{X},\mathscr{Y})}{2}\right),$$

which we call the *Plücker metric*.

We now describe a particular way of breaking up the Grassmannian into subsets known as the *matroid stratification* of the Grassmannian (see [4]). First we define matroids (note that there are many equivalent ways of defining matroids, we state the one that we will use here).

Definition 4.4 A *matroid* is an ordered pair $([M], \mathscr{B})$ where $\mathscr{B} \subseteq 2^{[M]}$ satisfies:

(B1) $\mathscr{B} \neq \emptyset$
(B2) $A, B \in \mathscr{B}, a \in A\backslash B \Rightarrow \exists b \in B\backslash A$ such that $(A\backslash\{a\}) \cup \{b\} \in \mathscr{B}$.

$[M]$ is called the *ground set* of $\mathscr{M}$, and the elements of $\mathscr{B}$ are called the *bases* of $\mathscr{M}$.

For more background on matroid theory we refer to [19]. The main reason we care about matroids is summarized in the following proposition, which can be found in [19].

Proposition 4.5 *Let $[M]$ be the set of column labels of an $N \times M$ matrix F over a field $\mathbb{F}$, and let $\mathscr{B}$ be the collection of subsets $I \subseteq [M]$ for which the set of columns labeled by I is a basis for $\mathbb{F}^k$. Then $\mathscr{M}(F) := ([M], \mathscr{B})$ is a matroid.*

Matroids encode linear independence; determinants are a measure for this. In particular, observe that $Plu(\mathscr{X})$ associates to each $\mathscr{X} \in Gr(M,N)$ a matroid

$\mathscr{M}(\mathscr{X})$ as follows: A set $\{i_1, \ldots, i_N\} \subseteq [M]$ is a basis of $\mathscr{M}(\mathscr{X})$ if and only if $Plu(\mathscr{X})_{i_1 \ldots i_N} \neq 0$. Thus, to each matroid $\mathscr{M}$ we can associate the subset of $Gr(M, N)$:

$$\mathscr{R}(\mathscr{M}) = \left\{\mathscr{X} \in Gr(M, N) : \mathscr{M}(\mathscr{X}) = \mathscr{M}\right\}.$$

Thus, $Gr(M, N)$ can be written as a disjoint union of sets of this type. We will use this stratification later to prove that generic Parseval frames are dense in the set of Parseval frames.

4.3.1 Frames and Plücker Coordinates

Let $\Phi = \{\varphi_i\}_{i=1}^M$ be a frame for $\mathscr{H}^N$. We define

$$Plu(\Phi) = \left(\det(\Phi_{i_1 \ldots i_N})\right)_{1 \leq i_1 < \cdots < i_N \leq M},$$

and note that $Plu(\Phi)$ is a point in $\mathscr{H}^{\binom{M}{N}}$. By Proposition 4.3 we have that $Plu(\Phi) = \lambda Plu(\Psi)$ if and only if there is an invertible operator T so that $\varphi_i = T\psi_i$ for every $i = 1, 2, \ldots, M$, in which case $\lambda = \det(T)$. An argument similar to the proof of Proposition (4.4) yields the following.

Proposition 4.6 $\|Plu(\Phi)\|^2 = \det(S)$, *where S is the corresponding frame operator.*

One important consequence of Proposition 4.6 is the following corollary, which will be used extensively later.

Corollary 4.3 *If Φ is a Parseval frame, then $\|Plu(\Phi)\| = 1$.*

However, note that Proposition 4.6 also says that the converse of the above corollary is not true. To see this, let S be a positive, self-adjoint operator such that $\det(S) = 1$, and let $\Phi = \{\varphi_i\}_{i=1}^M$ be a Parseval frame. Now consider the frame $S^{1/2}\Phi = \{S^{1/2}\varphi_i\}_{i=1}^M$ which has S as its frame operator. If S is not the identity operator then $S^{1/2}\Phi$ is not a Parseval frame, but we still have that $\|Plu(S^{1/2}\Phi)\| = 1$.

For notational convenience we denote by $\Pi(\Phi)$ the image of the analysis operator corresponding to the frame Φ. Thus, $Plu(\Pi(\Phi))$ is a point in projective space. Given a point $\mathscr{X} \in Gr(M, N)$ we use the symbol $\Pi^{-1}(\mathscr{X})$ to denote the entire isomorphism class of frames whose analysis operator has $\mathscr{X}$ as its image. We can now prove the following result, which says that close subspaces are necessarily images of analysis operators of close Parseval frames.

Theorem 4.5 *Let $\mathscr{X}, \mathscr{Y} \in Gr(M, N)$, and let $\epsilon > 0$. Suppose that $\Theta(\mathscr{X}, \mathscr{Y}) < \frac{\epsilon}{2\sqrt{N}}$ and that $\{\varphi_i\}_{i=1}^M \in \Pi^{-1}(\mathscr{X})$ is a Parseval frame. Then there is a Parseval frame $\{\psi_i\}_{i=1}^M \in \Pi^{-1}(\mathscr{Y})$ such that $\|\varphi_i - \psi_i\| < \epsilon$ for every $i = 1, 2, \ldots, M$.*

Proof First note that $\theta_N(\mathscr{X},\mathscr{Y}) \leq \Theta(\mathscr{X},\mathscr{Y}) < \frac{\epsilon}{2\sqrt{N}}$. We can find orthonormal bases $\{a_j\}_{j=1}^N$ for $\mathscr{X}$ and $\{b_j\}_{j=1}^N$ for $\mathscr{Y}$ such that $\langle a_j, b_j\rangle = \cos(\theta_j)$ for every $j = 1,\ldots,N$. Therefore, we have

$$\|a_j - b_j\| = 2\sin\left(\frac{\theta_j}{2}\right) \leq 2\sin\left(\frac{\theta_N}{2}\right) < \frac{\epsilon}{\sqrt{N}}$$

for every $j = 1,\ldots,N$. Now let A and B be the $N \times M$ matrices whose jth columns are a_j and b_j respectively. Let a_{ij} be the ith entry of a_j and let f_i be the ith row of A, similarly let b_{ij} be the ith entry of b_j and let g_i be the ith row of B. Then we have

$$\sum_{i=1}^M (a_{ij} - b_{ij})^2 < \frac{\epsilon^2}{N} \quad \text{for every } j = 1,\ldots,N$$

which means that

$$(a_{ij} - b_{ij})^2 < \frac{\epsilon^2}{N} \quad \text{for every } j = 1,\ldots,N,\ i = 1,\ldots,M$$

which further implies that

$$\sum_{j=1}^N (a_{ij} - b_{ij})^2 = \|f_i - g_i\|^2 < \epsilon^2.$$

Now since the columns of A form an orthonormal basis for $\mathscr{X}$, we know that $\{f_i\}_{i=1}^M$ is a Parseval frame which is isomorphic to $\{\varphi_i\}_{i=1}^M$. This means there is some unitary $T : \mathscr{H}^N \to \mathscr{H}^N$ such that $Tf_i = \varphi_i$ for every $i = 1,\ldots,M$. The Parseval frame $\{\psi_i\}_{i=1}^M = \{Tg_i\}_{i=1}^M$ can now be seen to have the desired properties. □

The same argument can be used to prove a similar result for different combinations of metrics on the Grassmannian and metrics on frames.

Theorem 4.6 *Let $\mathscr{X},\mathscr{Y} \in Gr(M,N)$, and let $\epsilon > 0$. Suppose that $\sum_{j=1}^N \sin^2(\theta_j(\mathscr{X},\mathscr{Y})) < \epsilon$ and that $\{\varphi_i\}_{i=1}^M \in \Pi^{-1}(\mathscr{X})$ is a Parseval frame. Then there is a Parseval frame $\{\psi_i\}_{i=1}^M \in \Pi^{-1}(\mathscr{Y})$ such that $\sum_{i=1}^M \|\varphi_i - \psi_i\|^2 < \epsilon$.*

We can also use a similar argument to generalize Theorem 4.5 in the case that we care about frames that may not be Parseval frames.

Theorem 4.7 *Let $\mathscr{X},\mathscr{Y} \in Gr(M,N)$, and let $\epsilon > 0$. Let $\{\varphi_i\}_{i=1}^M \in \Pi^{-1}(\mathscr{X})$ with frame operator S and assume that*

$$\Theta(\mathscr{X},\mathscr{Y}) < \frac{\epsilon}{\|S^{\frac{1}{2}}\|2\sqrt{N}}.$$

Then there is a frame $\{\psi_i\}_{i=1}^M \in \Pi^{-1}(\mathscr{Y})$ such that $\|\varphi_i - \psi_i\| < \epsilon$ for every $i = 1,2,\ldots,M$. Furthermore, if $\{\varphi_i\}_{i=1}^M$ is a Parseval frame, then $\{\psi_i\}_{i=1}^M$ can be chosen to be a Parseval frame as well.

4.3.2 Generic Frames

A frame is said to be robust to m erasures if the removal of any m vectors leaves a frame. Clearly a frame consisting of M vectors in $\mathscr{H}^N$ can be robust to at most $M - N$ erasures. We call such a frame a *generic frame*. Generic frames have appeared previously in the literature under the name *maximally robust frame* (see [20]). However, we shall see that this is a very weak measure of the robustness of a given frame to erasures. In particular, we will show that there is an open dense set of frames that are robust to $M - N$ erasures, so we believe the name "maximally robust" should be reserved for robustness in a more numerical sense. In this section we will study the set of generic frames. We begin this section with the following fairly simple observation.

Proposition 4.7 *Let $\{\varphi_i\}_{i=1}^M$ be a frame, and $\epsilon > 0$. Then there is a generic frame $\{\psi_i\}_{i=1}^M$ such that*

$$\|\varphi_i - \psi_i\| < \epsilon$$

for every $i = 1, \ldots, M$.

Proof If $\{\varphi_i\}_{i=1}^M$ is generic then there is nothing to prove, so assume $\{\varphi_i\}_{i=1}^M$ is not generic. Let $\{\varphi_{i_j}\}_{j=1}^m$ be a minimal dependent set; note that $\dim(\text{span}\{\varphi_{i_j}\}_{j=1}^m) = m - 1$. Choose some $\varphi_{i_{j_0}}$ and let B be an open ball of radius ϵ centered at $\varphi_{i_{j_0}}$. Now let $\mathscr{W}$ be the set of hyperplanes (i.e., codimension 1 subspaces) spanned by any combination of vectors in $\{\varphi_i\}_{i=1}^M$ that do not include $\varphi_{i_{j_0}}$. Notice that $\mathscr{H}^N \backslash \mathscr{W}$ is an open dense set in $\mathscr{H}^N$ since $\mathscr{W}$ consists of a finite number of hyperplanes, so $B \cap (\mathscr{H}^N \backslash \mathscr{W}) \neq \emptyset$. Choose any x in this set and replace $\varphi_{i_{j_0}}$ with x. This ensures that $\dim(\text{span}\{\varphi_{i_j}\}_{j \neq j_0} \cup \{x\}) = m$ and that we have not created any new dependent sets of cardinality less than or equal to N. After repeating this process finitely many times, we can ensure that we arrive at a generic frame with the desired properties. □

Now if $\{\varphi_i\}_{i=1}^M$ is a Parseval frame, can $\{\psi_i\}_{i=1}^M$ be chosen to be a Parseval frame? The answer is yes, but to prove this we need to use the results of the previous section. Before proving this we need to explain some further properties of the matroid stratification of the Grassmannian.

Choose $1 \leq i_1 < \cdots < i_N \leq M$ and consider the set $\mathscr{V}_{i_1 \ldots i_N} = \{\mathscr{X} \in Gr(M, N) : Plu(\mathscr{X})_{i_1 \ldots i_N} = 0\} = \bigcup\{\mathscr{R}(\mathscr{M}) : \{i_1, \ldots, i_N\} \text{ is not a basis of } \mathscr{M}\}$. Now observe that $\mathscr{V}_{i_1 \ldots i_N}$ is a proper closed subvariety of $Gr(M, N)$. This tells us that $\mathscr{V}_{i_1 \ldots i_N}$ is a closed subset of $Gr(M, N)$ in the Zariski topology, which implies it is also closed in the Euclidean topology (the topology induced by the Plücker metric), so in particular $Gr(M, N) \backslash \mathscr{V}_{i_1 \ldots i_N}$ is an open and dense subset of $Gr(M, N)$ (in both topologies). The *uniform matroid of rank N on $[M]$* is the matroid whose bases consist of all subsets of $[M]$ of cardinality N; we use the symbol $\mathscr{U}_{M,N}$ to denote this matroid. Now observe that

$$\mathscr{R}(\mathscr{U}_{M,N}) = \bigcap_{1 \leq i_1 < \cdots < i_N \leq M} Gr(M, N) \backslash \mathscr{V}_{i_1 \ldots i_N},$$

which means that $\mathscr{R}(\mathscr{U}_{M,N})$ is an open and dense subset of $Gr(n,k)$. Now we can prove our result.

Theorem 4.8 *Let $\{\varphi_i\}_{i=1}^M$ be a Parseval frame, and $\epsilon > 0$. Then there is a generic Parseval frame $\{\psi_i\}_{i=1}^M$ such that $\|\varphi_i - \psi_i\| < \epsilon$ for every $i = 1, \ldots, M$.*

Proof First note that Φ is generic if and only if $\Pi(\Phi) \in \mathscr{R}(\mathscr{U}_{M,N})$, so we may assume $\Pi(\Phi) \notin \mathscr{R}(\mathscr{U}_{M,N})$. By the above remarks we can find a point $\mathscr{Y} \in \mathscr{R}(\mathscr{U}_{M,N})$ such that $\Theta(\Pi(\Phi), \mathscr{Y}) < \frac{\epsilon}{2\sqrt{k}}$, so the result follows from Theorem 4.5. □

Now that we have established that almost every frame is generic, we would like to come up with a numerical measure of the genericity of a frame and construct the Parseval frames that are somehow the "most generic." Since we have seen how to associate points on the Grassmannian to frames, and we know how to compute distance on the Grassmannian, one reasonable way to measure the genericity of a given frame is to find the shortest distance on the Grassmannian to an isomorphism class of frames that is not generic. However, there are many ways to measure distance on the Grassmannian, so we will choose one reasonable way.

We pose the following optimization problem:

$$\min_{\mathscr{X} \in Gr(M,N)} \max\left\{\Theta(\mathscr{X}, \mathscr{E}_{i_1 \cdots i_N}) : 1 \leq i_1 < \cdots < i_N \leq M\right\}, \tag{4.7}$$

where $\mathscr{E}_{i_1 \cdots i_N} = \text{span}\{e_{i_1}, \ldots, e_{i_N}\}$ and $\{e_i\}_{i=1}^M$ is the standard orthonormal basis of $\mathscr{H}^M$. Recall that by Corollary 4.3 the Plücker norm of any Parseval frame is 1, so we would like to find the unit vectors on the (Plücker embedding of the) Grassmannian whose smallest (in absolute value) Plücker coordinate is as big as possible. Intuitively, a small Plücker coordinate says that the corresponding subset is "barely" a basis.

Clearly, if the Plücker embedding is onto, then these would be the points whose Plücker coordinates (in absolute value) were all equal to $\binom{M}{N}^{-1/2}$. However, these points are only in the image of the Plücker embedding when $N = 1$ or $N = M - 1$, i.e., every sequence of unit-modulus scalars is optimal for $N = 1$ and every simplex is optimal for $N = M - 1$. For other choices of M and N we want to find the points on the Grassmannian that are as close (in the regular Euclidean sense) to these points as possible. An equivalent task is to solve the following optimization problem:

$$\begin{aligned} \text{maximize:} \quad & \sum_{1 \leq i_1 < \cdots < i_N \leq M} |x_{i_1 \cdots i_N}| \\ \text{subject to:} \quad & Plu(M,N), \\ & \sum_{1 \leq i_1 < \cdots < i_N \leq M} x_{i_1 \cdots i_N}^2 = 1. \end{aligned}$$

We will illustrate this with the first nontrivial example, $Gr(4,2)$. In this case $Plu(4,2)$ contains only the polynomial $x_{12}x_{34} - x_{13}x_{24} + x_{14}x_{23}$, so the above op-

timization problem becomes

$$\begin{aligned} \text{maximize:} \quad & |x_{12}| + |x_{13}| + |x_{14}| + |x_{23}| + |x_{24}| + |x_{34}| \\ \text{subject to:} \quad & x_{12}x_{34} - x_{13}x_{24} + x_{14}x_{23} = 0, \\ & x_{12}^2 + x_{13}^2 + x_{14}^2 + x_{23}^2 + x_{24}^2 + x_{34}^2 = 1. \end{aligned}$$

For simplicity, we will only look for solutions in the first orthant (i.e., where all Plücker coordinates are positive), so we can drop the absolute values. Using the method of Lagrange multipliers we arrive at the following system of equations:

$$\begin{aligned} 2\lambda_1 x_{12} + \lambda_2 x_{34} &= 1, \\ 2\lambda_1 x_{34} + \lambda_2 x_{12} &= 1, \\ 2\lambda_1 x_{14} + \lambda_2 x_{23} &= 1, \\ 2\lambda_1 x_{23} + \lambda_2 x_{14} &= 1, \\ 2\lambda_1 x_{13} - \lambda_2 x_{24} &= 1, \\ 2\lambda_1 x_{24} - \lambda_2 x_{13} &= 1. \end{aligned}$$

Together, the first two equations imply that

$$\begin{aligned} & 2\lambda_1 x_{12} + \lambda_2 x_{34} = 2\lambda_1 x_{34} + \lambda_2 x_{12} \\ \Rightarrow \quad & (2\lambda_1 - \lambda_2)x_{12} = (2\lambda_1 - \lambda_2)x_{34} \\ \Rightarrow \quad & x_{12} = x_{34} \quad \text{as long as } \lambda_1 \neq \frac{\lambda_2}{2}. \end{aligned}$$

Similarly, the third and fourth equations imply $x_{14} = x_{23}$ as long as $\lambda_1 \neq \frac{\lambda_2}{2}$, and the last two equations imply $x_{13} = x_{24}$ as long as $\lambda_1 \neq -\frac{\lambda_2}{2}$. This reduces our system of six equations to the following system of three equations:

$$\begin{aligned} (2\lambda_1 + \lambda_2)x_{12} &= 1, \\ (2\lambda_1 + \lambda_2)x_{14} &= 1, \\ (2\lambda_1 - \lambda_2)x_{13} &= 1. \end{aligned}$$

But the first two equations of this system now imply $x_{12} = x_{14}$ (under our assumptions on λ_1 and λ_2). The Plücker relation now becomes

$$2x_{12}^2 - x_{13}^2 = 0.$$

Now we can use our unit norm constraint to find the solutions:

$$x_{12} = x_{14} = x_{23} = x_{34} = \pm\frac{\sqrt{2}}{4}, \qquad x_{13} = x_{24} = \pm\frac{1}{2}.$$

Thus, we wish to find a 4×2 matrix whose Plücker coordinates are (a scalar multiple of) $(\frac{\sqrt{2}}{4}, \frac{1}{2}, \frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{4}, \frac{1}{2}, \frac{\sqrt{2}}{4})$. The easiest way to do this is to make the first Plücker coordinate equal to 1:

$$\frac{4}{\sqrt{2}}\left(\frac{\sqrt{2}}{4}, \frac{1}{2}, \frac{\sqrt{2}}{4}, \frac{\sqrt{2}}{4}, \frac{1}{2}, \frac{\sqrt{2}}{4}\right) = (1, \sqrt{2}, 1, 1, \sqrt{2}, 1),$$

and find a matrix of the following form:

$$\begin{bmatrix} 1 & 0 & a & b \\ 0 & 1 & c & d \end{bmatrix}.$$

For example, since $x_{13} = \sqrt{2}$ we see that $c = \sqrt{2}$. Similarly, we can solve for a, b and d and we arrive at the following matrix:

$$\begin{bmatrix} 1 & 0 & -1 & -\sqrt{2} \\ 0 & 1 & \sqrt{2} & 1 \end{bmatrix}.$$

Finally, we perform the Gram-Schmidt process to the columns of this matrix so that they become an orthonormal basis for their span in $\mathbb{R}^4$, and that means the columns should form the Parseval frame that we were looking for:

$$\begin{bmatrix} \frac{1}{2} & 0 & -\frac{1}{2} & -\frac{\sqrt{2}}{2} \\ \frac{1}{2} & \frac{\sqrt{2}}{2} & \frac{1}{2} & 0 \end{bmatrix}.$$

4.3.3 Signal Reconstruction Without Phase

In this section we will discuss a problem known as *phaseless reconstruction*. The results of this section originally appeared in [2]. Suppose we are given a frame $\Phi = \{\varphi_i\}_{i=1}^M$ for $\mathscr{H}^N$. We want to know if we can recover $x \in \mathscr{H}^N$ up to a scalar multiple of modulus one if we are just given the vector of absolute values of inner products with the frame vectors. To be more precise, we define the mappings

$$f_\Phi^a : \mathscr{H}^N \to \mathbb{R}^M, \qquad f_\Phi^a(x) = \big(|\langle x, \varphi_1\rangle|, \ldots, |\langle x, \varphi_M\rangle|\big)$$

and

$$f_\Phi : \mathscr{H}^N/\sim \to \mathbb{R}^M, \qquad f_\Phi(\hat{x}) = \big(|\langle x, \varphi_1\rangle|, \ldots, |\langle x, \varphi_M\rangle|\big), \quad x \in \hat{x},$$

where $x, y \in \hat{x} \in \mathscr{H}^N/\sim$ if there is a scalar λ with $x = \lambda y$ and $|\lambda| = 1$. So we would like to find conditions on the frame Φ which guarantee that f_Φ is injective. We will analyze the real and complex cases separately.

We start with the real case. In this case the domain of f_Φ is $\mathbb{R}^N/\sim$, where $x, y \in \hat{x} \in \mathbb{R}^N/\sim$ if and only if $x = \pm y$. Before stating our results, we need to fix

some notation. Given a subset $I \subseteq [M]$, by abuse of notation use the same symbol I to denote the characteristic function of this set, i.e., for $i \in [M]$, $I(i) = 1$ if $i \in I$ and $I(i) = 0$ if $i \notin I$. Define a mapping $\sigma_I : \mathbb{R}^M \to \mathbb{R}^M$ by

$$\sigma_I(a_1, \ldots, a_M) = \big((-1)^{I(1)} a_1, \ldots, (-1)^{I(M)} a_M\big).$$

Note that $\sigma_I^2 = I$, and $\sigma_{I^c} = -\sigma_I$. Also, let $L_I = \{(a_1, \ldots, a_M) : a_i = 0 \text{ for } i \in I\}$. Then we have that $\sigma_I(u) = u$ if and only if $u \in L_I$ and $\sigma_I(u) = -u$ if and only if $u \in L_{I^c}$.

We need one more definition before stating our theorem.

Definition 4.5 Let $\mathscr{M}$ be a matroid with ground set $[M]$. We say that $\mathscr{M}$ has the *complement property* if for every $I \subseteq [M]$ either I contains a basis of $\mathscr{M}$ or I^c contains a basis of $\mathscr{M}$.

Theorem 4.9 *For a frame $\Phi = \{\varphi_i\}_{i=1}^M \subseteq \mathbb{R}^N$ the following are equivalent*:

(1) *f_Φ is injective.*
(2) *For every nonempty proper subset $I \subseteq [M]$ and every $u \in \Pi(\Phi) \backslash (L_I \cup L_{I^c})$, $\sigma_I(u) \notin \Pi(\Phi)$.*
(3) *If there is a nonempty proper subset $I \subseteq [M]$ for which $\Pi(\Phi) \cap L_I \neq \emptyset$, then $\Pi(\Phi) \cap L_{I^c} = \emptyset$.*
(4) *$\Pi(\Phi) \in \mathscr{R}(\mathscr{M})$ for some matroid $\mathscr{M}$ with the complement property.*

Proof (1)⇒(2) Suppose there is a nonempty proper subset $I \subseteq [M]$ and a $u \in \Pi(\Phi) \backslash (L_I \cup L_{I^c})$ for which $\sigma_I(u) \in \Pi(\Phi)$. Since $u \notin L_I \cup L_{I^c}$ we know $\sigma_I(u) \neq \pm u$. Now there are $x, y \in \mathbb{R}^N$ such that $\langle x, \varphi_i \rangle = u(i)$ and $\langle y, \varphi_i \rangle (-1)^{I(i)} u(i)$ for every $i = 1, \ldots, M$. But then $f_\Phi^a(x) = f_\Phi^a(y)$ and since $\sigma_I(u) \neq \pm u$ we know that $x \neq \pm y$, so f_Φ is not injective.

(2)⇒(3) Suppose there is a nonempty proper subset $I \subseteq [M]$ for which both $\Pi(\Phi) \cap L_I \neq \emptyset$ and $\Pi(\Phi) \cap L_{I^c} \neq \emptyset$. Choose $v \in \Pi(\Phi) \cap L_I$ and $w \in \Pi(\Phi) \cap L_{I^c}$. Then $v + w \in \Pi(\Phi) \backslash (L_I \cup L_{I^c})$, but $\sigma_I(v + w) = v - w \in \Pi(\Phi)$.

(3)⇒(4) Suppose there is a subset $I \subseteq [M]$ for which neither $\{\varphi_i\}_{i \in I}$ nor $\{\varphi_i\}_{i \in I^c}$ spans $\mathbb{R}^N$. Choose $x \perp \operatorname{span}\{\varphi_i\}_{i \in I}$ and $y \perp \operatorname{span}\{\varphi_i\}_{i \in I^c}$. Then $T(x) \in L_I$ and $T(y) \in L_{I^c}$.

(4)⇒(1) Suppose $x, y \in \mathbb{R}^N$ are such that $|\langle x, \varphi_i \rangle| = |\langle y, \varphi_i \rangle|$ for every $i = 1, \ldots, M$. Let $I = \{i : \langle x, \varphi_i \rangle = -\langle y, \varphi_i \rangle\}$ and observe that $x + y \perp \operatorname{span}\{\varphi_i\}_{i \in I}$ and $x - y \perp \operatorname{span}\{\varphi_i\}_{i \in I^c}$. But we know that either $\operatorname{span}\{\varphi_i\}_{i \in I} = \mathbb{R}^N$ or $\operatorname{span}\{\varphi_i\}_{i \in I^c} = \mathbb{R}^N$ by assumption, so we have either $x + y = 0$ or $x - y = 0$, i.e., $x = \pm y$ and f_Φ is injective. □

Corollary 4.4

(1) *If $M \geq 2N - 1$ then f_Φ is injective for almost every frame $\Phi = \{\varphi_i\}_{i=1}^M \subseteq \mathbb{R}^N$.*
(2) *If $M < 2N - 1$ then f_Φ is not injective.*

Proof To see the first statement, just observe that for $M \geq 2N-1$ the uniform matroid $\mathscr{U}_{M,N}$ has the complement property, so if Φ is generic then f_Φ is injective. For the second statement, let $I \subseteq [M]$ be such that $|I| = N-1$ and note that $|I^c| \leq N-1$. Therefore it is impossible for any matroid of rank N whose ground set is $[M]$ to have the complement property. □

We now shift our attention to the complex case. In this case the domain of f_Φ is $\mathbb{C}^N/\sim$ where $x \sim y$ if and only if there is a $\lambda \in \mathbb{T}$ so that $x = \lambda y$, where $\mathbb{T}$ is the unit circle on the complex plane. At this time the complex case is not understood nearly as well as the real case, however we can still prove the existence of a large family of frames for which f_Φ is injective.

Theorem 4.10 *Suppose that $M \geq 4N-2$. Then there is an open and dense set of frames for $\mathbb{C}^N$ with M elements for which f_Φ is injective.*

Proof First note that f_Φ is injective if and only if there do not exist nonparallel vectors $v, w \in \Pi(\Phi)$ such that $|v(i)| = |w(i)|$ for every $i = 1, \ldots, M$. So we will show that the set of subspaces that have this property is a Zariski open subset of $Gr(M,N)$. Denote the complement of this set by $\mathscr{A}$, and choose any $\mathscr{X} \in \mathscr{A}$. Without loss of generality we may assume we have a basis $\{u_j\}_{j=1}^N$ for $\mathscr{X}$ so that $u_j(i) = 1$ if $j = i$ and $u_j(i) = 0$ when $j \neq i \leq N$; for $i > N$ $u_j(i)$ is undetermined. Therefore, in a neighborhood of $\mathscr{X}$ we see that $Gr(M,N)$ has dimension $2N(M-N)$ as a real variety.

Now since $\mathscr{X} \in \mathscr{A}$ we can choose nonparallel $v, w \in \mathscr{X}$ with $|v(i)| = |w(i)|$ for every i. Our choice of basis guarantees that at least one of the first N entries of v (and therefore w) is nonzero, which we can assume to be the first entry without loss of generality, so after rescaling we have that $v(i) = w(i) = 1$. Since v and w are nonparallel we know that for some $2 \leq i \leq N$ we have that $v(i) \neq w(i) \neq 0$, and again without loss of generality we can assume this happens for $i = 2$.

Now we have that there are $\lambda_2, \ldots, \lambda_M \in \mathbb{T}$ with $\lambda_2 \neq 1$ such that $w(i) = \lambda_i v(i)$ for every $i = 2, \ldots, M$ (and $v(1) = w(1) = 1$). For $i > N$ we have $v(i) = \sum_{j-1}^N v(j)u_j(i)$ and $w(j) = \sum_{j=1}^N \lambda_j v(j) u_j(i)$. Thus we have

$$\left| \sum_{j=1}^N v(j)u_j(i) \right| = \left| \sum_{j=1}^N \lambda_j v(j) u_j(i) \right|. \tag{4.8}$$

Consider the variety of all tuples $(\mathscr{Y}, v(1), \ldots, v(N), \lambda_2, \ldots, \lambda_N)$ with $\mathscr{Y} \in Gr(M,N)$ and $v(i)$ and λ_i as above. This variety is locally isomorphic to $\mathbb{C}^{N(M-N)} \times (\mathbb{C}\backslash\{0\}) \times \mathbb{C}^{N-2} \times (\mathbb{T}\backslash\{1\}) \times \mathbb{T}^{N-2}$ which has dimension $2N(M-N) + 3N - 3$ as a real variety. We also have that $\mathscr{A}$ is the image under projection onto the first factor of this variety cut out by the $M-N$ equations (4.8). Now observe that for a fixed $0 \neq v(2), \ldots, v_N$ and $1 \neq \lambda_2, \ldots, \lambda_N$ these equations are nondegen-

erate. Since $u_1(i), \ldots, u_N(i)$ appear in exactly one equation, it follows that these equations define a subspace of $\mathbb{C}^{N(M-N)}$ of real codimension at least $M-N$. Since this is true for all choices of the $v(i)$'s and the λ_i's, it follows that these equations are independent.

We can now conclude that $\mathscr{A}$ is a real variety of (local) dimension $2N(M-N)+3N-3-(M-N)$. Therefore if $3N-3-(M-N)<0$, i.e., $M \geq 4N-2$, then $\mathscr{A}$ is a proper subvariety of $Gr(M,N)$ and so its complement is open in the Zariski topology. □

It is not known whether the value $M=4N-2$ is optimal, i.e., we do not know if it is possible for f_Φ to be injective for a frame consisting of fewer than $4N-2$ vectors.

References

1. Balan, R.V.: Equivalence relations and distances between Hilbert frames. Proc. Am. Math. Soc. **127**, 2353–2366 (1999)
2. Balan, R.V., Casazza, P.G., Edidin, D.: On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**, 345–356 (2006)
3. Benedetto, J.J., Fickus, M.: Finite normalized tight frames. Adv. Comput. Math. **18**, 357–385 (2003)
4. Björner, A., Las Vergnas, M., Sturmfels, B., White, N., Ziegler, G.M.: Oriented Matroids. Cambridge University Press, Cambridge (1999)
5. Cahill, J.: Flags, frames, and Bergman spaces. Master's Thesis, San Francisco State University (2009)
6. Cahill, J., Casazza, P.G.: The Paulsen problem in operator theory (2011). arXiv:1102.2344
7. Casazza, P.G., Leon, M.T.: Existence and construction of finite frames with a given frame operator. Int. J. Pure Appl. Math. **63**, 149–158 (2010)
8. Casazza, P.G., Tremain, J.C.: The Kadison–Singer problem in mathematics and engineering. Proc. Natl. Acad. Sci. **103**, 2032–2039 (2006)
9. Dykema, K., Strawn, N.: Manifold structure of spaces of spherical tight frames. Int. J. Pure Appl. Math. **28**, 217–256 (2006)
10. Fraenkel, A.S., Yesha, Y.: Complexity of problems in games, graphs, and algebraic equations. Discrete Appl. Math. **1**, 15–30 (1979)
11. Fulton, W.: Young Tableaux—With Applications to Representation Theory and Geometry. Cambridge University Press, Cambridge (1997)
12. Guillemin, V., Pollack, A.: Differential Topology—History, Theory, and Applications. Prentice-Hall, Englewood Cliffs (1974)
13. Hartshorne, R.: Algebraic Geometry. Springer, New York (1997)
14. Jiang, S.: Angles between Euclidean subspaces. Geom. Dedic. **63**, 113–121 (1996)
15. Jordan, C.: Essai sur la géométrie á n dimensions. Bull. Soc. Math. Fr. **3**, 103–174 (1875)
16. Krantz, S.G., Parks, H.R.: The Implicit Function Theorem—History, Theory, and Applications. Birkhäuser, Boston (2002)
17. Miao, J.M., Ben-Israel, A.: On principal angles between subspaces in $\mathbb{R}^n$. Linear Algebra Appl. **171**, 81–98 (1992)
18. Miao, J.M., Ben-Israel, A.: Product cosines of angles between subspaces. Linear Algebra Appl, 237–238:71–81 (1996)
19. Oxley, J.G.: Matroid Theory. Oxford University Press, New York (1992)

20. Püschel, M., Kovačević, J.: Real tight frames with maximal robustness to erasures. In: Proc. IEEE Data Comput. Conf., pp. 63–72 (2005)
21. Strawn, N.: Finite frame varieties: nonsingular points, tangent spaces, and explicit local parameterizations. J. Fourier Anal. Appl. **17**, 821–853 (2011)
22. Weaver, N.: The Kadison–Singer problem in discrepancy theory. Discrete Math. **278**, 227–239 (2004)

Chapter 5
Group Frames

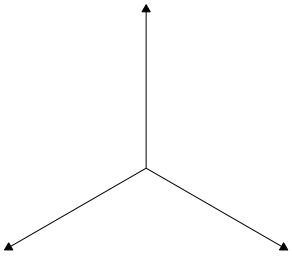
Shayne Waldron

Abstract The prototypical example of a tight frame, the *Mercedes-Benz frame* can be obtained as the orbit of a single vector under the action of the group generated by rotation by $\frac{2\pi}{3}$, or the dihedral group of symmetries of the triangle. Many frames used in applications are constructed in this way, often as the orbit of a single vector (akin to a mother wavelet). Most notable are the *harmonic frames* (finite abelian groups) used in signal analysis, and the equiangular *Heisenberg frames*, or *SIC–POVMs* (discrete Heisenberg group) used in quantum information theory. Other examples include tight frames of multivariate orthogonal polynomials sharing symmetries of the weight function, and the *highly symmetric tight frames* which can be viewed as the vertices of highly regular polytopes. We will describe the basic theory of such *group frames*, and some of the constructions that have been found so far.

Keywords Group frame · G-frame · Harmonic frames · SIC–POVM · Heisenberg frame · Highly symmetric tight frame · Symmetry group of a frame · Heisenberg frame · Group matrix · Unitary representation · Equiangular frames · Zauner's conjecture

5.1 The Symmetries of a Frame (Its Dual and Complement)

The *symmetries* of the Mercedes-Benz frame



S. Waldron (✉)
Department of Mathematics, University of Auckland, Private Bag 92019, Auckland, New Zealand
e-mail: waldron@math.auckland.ac.nz

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_5,

are those rotations and reflections (unitary maps) which permute its vectors. We now formalise this idea, with the key features of the *symmetry group* (see [19] for full proofs) being:

- It is defined for *all* finite frames as a group of permutations on the index set.
- It is simple to calculate from the Gramian of the canonical tight frame.
- The symmetry groups of similar frames are equal. In particular, a frame, its dual frame and canonical tight frame have the same symmetry group.
- The symmetry groups of various combinations of frames, such as tensor products and direct sums, are related to those of the constituent frames in a natural way.
- The symmetry group of a frame and that of its complementary frame are equal.

Let S_M be the (symmetric group of) permutations on $\{1, 2, \ldots, M\}$, and $\mathrm{GL}(\mathscr{H})$ be the (general linear group of) linear maps $\mathscr{H} \to \mathscr{H}$.

Definition 5.1 The *symmetry group* of a finite frame $\Phi = (\varphi_j)_{j=1}^M$ for $\mathscr{H} = \mathbb{F}^N$ is

$$\mathrm{Sym}(\Phi) := \left\{\sigma \in S_M : \exists L_\sigma \in \mathrm{GL}(\mathscr{H}) \text{ with } L_\sigma \varphi_j = \varphi_{\sigma j},\ j = 1, \ldots, M\right\}.$$

Let Φ^{can} denote the canonical tight frame $(\Phi\Phi^*)^{-1/2}\Phi$ of Φ.

Theorem 5.1 *If Φ and Ψ are similar frames, i.e., $\Phi = Q\Psi$, $Q \in \mathrm{GL}(\mathscr{H})$, or are complementary frames, i.e., $G_{\Phi^{\mathrm{can}}} + G_{\Psi^{\mathrm{can}}} = Id$, then*

$$\mathrm{Sym}(\Psi) = \mathrm{Sym}(\Phi).$$

In particular, a frame, its dual frame and its canonical tight frame have the same symmetry group.

Proof It suffices to show one inclusion. Suppose $\sigma \in \mathrm{Sym}(\Phi)$, i.e., $L_\sigma \varphi_j = \varphi_{\sigma j}$, $\forall j$. Since $\varphi_j = Q\psi_j$, this gives $Q^{-1}L_\sigma Q\psi_j = \psi_{\sigma j}$, $\forall j$, i.e., $\sigma \in \mathrm{Sym}(\Psi)$. □

Example 5.1 Let Φ be the Mercedes-Benz frame. Since its vectors add to zero, $\Psi = ([1], [1], [1])$ is the complementary frame for $\mathbb{R}$. Clearly, $\mathrm{Sym}(\Psi) = S_3$, and so $\mathrm{Sym}(\Phi) = S_3$ (which is isomorphic to the dihedral group of triangular symmetries).

Since a finite frame Φ is determined up to similarity by $G_{\Phi^{\mathrm{can}}}$, the Gramian of the canonical tight frame, it is possible to compute $\mathrm{Sym}(\Phi)$ from $G_{\Phi^{\mathrm{can}}}$. This is most easily done as follows.

Proposition 5.1 *Let Φ be a finite frame. Then*

$$\sigma \in \mathrm{Sym}(\Phi) \quad \Longleftrightarrow \quad P_\sigma^* G_{\Phi^{\mathrm{can}}} P_\sigma = G_{\Phi^{\mathrm{can}}},$$

where P_σ is the permutation matrix given by $P_\sigma e_j = e_{\sigma j}$.

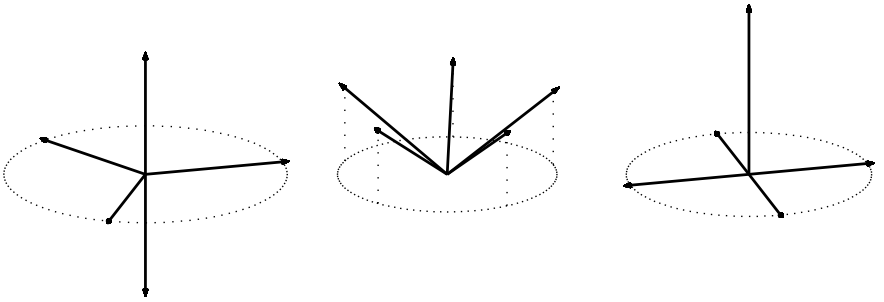


Fig. 5.1 The most symmetric tight frames of five distinct nonzero vectors in $\mathbb{R}^3$. The vertices of the trigonal bipyramid (12 symmetries), five equally spaced vectors lifted (10 symmetries), and four equally spaced vectors and one orthogonal (8 symmetries)

Since $\mathrm{Sym}(\Phi)$ is a subgroup of S_M, it follows that there are *maximally symmetric* frames of M vectors in $\mathbb{F}^N$, i.e., those with the largest possible symmetry groups.

Example 5.2 The M equally spaced vectors in $\mathbb{R}^2$ have the dihedral group of order $2M$ as symmetries. This is not always the most symmetric frame of M vectors in $\mathbb{C}^2$; e.g., if M is even, the (harmonic) tight frame given by the M distinct vectors

$$\left\{\begin{pmatrix}1\\1\end{pmatrix},\begin{pmatrix}\omega\\-\omega\end{pmatrix},\begin{pmatrix}\omega^2\\\omega^2\end{pmatrix},\begin{pmatrix}\omega^3\\-\omega^3\end{pmatrix},\begin{pmatrix}\omega^4\\\omega^4\end{pmatrix},\cdots\begin{pmatrix}\omega^{M-2}\\\omega^{M-2}\end{pmatrix},\begin{pmatrix}\omega^{M-1}\\-\omega^{M-1}\end{pmatrix}\right\},$$

$$\omega := e^{\frac{2\pi i}{M}}$$

has a symmetry group of order $\frac{1}{2}M^2$ (see [10] for details).

Example 5.3 The most symmetric tight frames of five vectors in $\mathbb{R}^3$ are as shown in Fig. 5.1.

The symmetry group of a combination of frames behaves as one would expect.

Proposition 5.2 *The symmetry groups of a finite frame satisfy*

1. $\mathrm{Sym}(\Phi)\times\mathrm{Sym}(\Psi)\subset\mathrm{Sym}(\Phi\cup\Psi)$ (*union of frames*),
2. $\mathrm{Sym}(\Phi)\times\mathrm{Sym}(\Psi)\subset\mathrm{Sym}(\Phi\otimes\Psi)$ (*tensor product*),
3. $\mathrm{Sym}(\Phi)\cap\mathrm{Sym}(\Psi)\subset\mathrm{Sym}(\Phi\oplus\Psi)$ (*direct sum*).

Here

$$\Phi\cup\Psi := \left(\begin{pmatrix}\varphi_j\\0\end{pmatrix},\begin{pmatrix}0\\\psi_k\end{pmatrix}\right),\qquad \Phi\otimes\Psi=(\varphi_j\otimes\psi_k),$$

$$\Phi\oplus\Psi := \left(\begin{pmatrix}\varphi_j\\\psi_k\end{pmatrix}\right),\quad \text{where } \sum_j\langle f,\varphi_j\rangle\psi_j=0,\ \forall f.$$

Since linear maps are determined by their action on a spanning set, it follows that if $\sigma \in \mathrm{Sym}(\Phi)$, then there is a unique $L_\sigma \in \mathrm{GL}(\mathscr{H})$ with $L_\sigma f_j = f_{\sigma j}, \forall j$. Further,

$$\mathrm{Sym}(\Phi) \to \mathrm{GL}(\mathscr{H}) : \sigma \mapsto L_\sigma \tag{5.1}$$

is a group homomorphism, i.e., a *representation* of $G = \mathrm{Sym}(\Phi)$. If the symmetry group acts transitively on Φ under this action, i.e., Φ is the orbit of any one vector, e.g., the Mercedes-Benz frame, then we have what is called a *G-frame*.

5.2 Representations and *G*-Frames

The Mercedes-Benz frame is the orbit under its symmetry group of a single vector. Formally, the symmetry group is a group of permutations (an abstract group) which acts as unitary transformations. This is a fundamental notion in abstract algebra.

Definition 5.2 A *representation* of a finite group G is a group homomorphism

$$\rho : G \mapsto \mathrm{GL}(\mathscr{H}),$$

i.e., a linear action of G on $\mathscr{H} = \mathbb{F}^N$, usually abbreviated $gv = \rho(g)v$, $v \in \mathscr{H}$.

Representations are a convenient way to study groups which appear as linear transformations, whilst being able to appeal to abstract group theory (cf. [12]).

Example 5.4 If Φ is a frame, then we have already observed that the action of $\mathrm{Sym}(\Phi)$ on $\mathscr{H}$ given by (5.1) is a representation. If Φ is *tight*, then this action is *unitary*. We will build this into our definition of a *group frame*.

Definition 5.3 Let G be a finite group. A *group frame* or *G-frame* for $\mathscr{H}$ is a frame $\Phi = (\varphi_g)_{g\in G}$ for which there exists a unitary representation $\rho : G \to \mathscr{U}(\mathscr{H})$ with

$$g\varphi_h := \rho(g)\varphi_h = \varphi_{gh}, \quad \forall g, h \in G.$$

This definition implies that a G-frame Φ is the orbit of a single vector $v \in \mathscr{H}$, i.e.,

$$\Phi = (gv)_{g\in G},$$

and so is an *equal norm* frame.

Example 5.5 An early example of group frames is the vertices of the *regular M-gon* or the *platonic solids* (see Fig. 5.2). These were some of the first examples of frames considered (see [3]). The *highly symmetric tight frames* (see Sect. 5.7) are a variation on this theme.

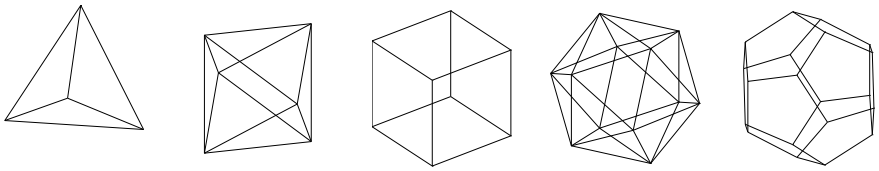


Fig. 5.2 The vertices of the platonic solids are examples of group frames

In the remaining sections, we outline the basic properties and constructions for G-frames. In particular, we will see that:

- There is a *finite* number of G-frames of M vectors in $\mathbb{F}^N$ for abelian groups G. These are known as *harmonic frames* (see Sect. 5.5)
- There is an *infinite* number of G-frames of M vectors in $\mathbb{F}^N$ for nonabelian G, most notably, the *Heisenberg frames* (see Sect. 5.9) of $M = N^2$ vectors in $\mathbb{C}^N$, which provide equiangular tight frames with the maximal number of vectors.

5.3 Group Matrices and the Gramian of a G-Frame

Since the representation defining a G-frame is unitary, i.e.,

$$\rho(g)^* = \rho(g)^{-1} = \rho\big(g^{-1}\big), \quad \text{so that } g^{-1}v = g^*v,$$

the Gramian of a G-frame $\Phi = (\varphi_g)_{g\in G} = (gv)_{g\in G}$ has a special form:

$$\langle \varphi_g, \varphi_h\rangle = \langle gv, hv\rangle = \big\langle v, g^*hv\big\rangle = \big\langle v, g^{-1}hv\big\rangle = \eta\big(g^{-1}h\big), \quad \text{where } \eta : G \to \mathbb{F}.$$

Thus the Gramian of a G-frame is a *group matrix* or *G-matrix*, i.e., a matrix A, with entries indexed by elements of a group G, which has the form

$$A = \big[\eta\big(g^{-1}h\big)\big]_{g,h\in G}.$$

One important consequence of the fact that the Gramian of a G-frame is a group matrix is that it has a small number of angles: $\{\eta(g) : g \in G\}$, which makes them good candidates for equiangular tight frames (see Sect. 5.9). We have the characterisation [18]:

Theorem 5.2 *Let G be a finite group. Then $\Phi = (\varphi_g)_{g\in G}$ is a G-frame (for its span $\mathscr{H}$) if and only if its Gramian G_Φ is a G-matrix.*

Proof If Φ is a G-frame, then we observed that its Gramian is a G-matrix.

Conversely, suppose that the Gramian of a frame Φ for $\mathscr{H}$ is a G-matrix. Let $\tilde{\Phi} = (\tilde{\phi}_g)_{g\in G}$ be the dual frame, so that

$$f = \sum_{g\in G} \langle f, \tilde{\phi}_g\rangle \phi_g, \quad \forall f \in \mathscr{H}. \tag{5.2}$$

For each $g \in G$, define a linear operator $U_g : \mathscr{H} \to \mathscr{H}$ by

$$U_g(f) := \sum_{h_1 \in G} \langle f, \tilde{\phi}_{h_1} \rangle \phi_{gh_1}, \quad \forall f \in \mathscr{H}.$$

Since $\mathrm{Gram}(\Phi) = [\langle \phi_h, \phi_g \rangle]_{g,h \in G}$ is a G-matrix, we have

$$\langle \phi_{gh_1}, \phi_{gh_2} \rangle = v\big((gh_2)^{-1} gh_1\big) = v\big(h_2^{-1} h_1\big) = \langle \phi_{h_1}, \phi_{h_2} \rangle. \tag{5.3}$$

It follows from (5.2) and (5.3) that U_g is unitary by the calculation

$$\begin{aligned}
\big\langle U_g(f_1), U_g(f_2) \big\rangle &= \left\langle \sum_{h_1 \in G} \langle f_1, \tilde{\phi}_{h_1} \rangle \phi_{gh_1}, \sum_{h_2 \in G} \langle f_2, \tilde{\phi}_{h_2} \rangle \phi_{gh_2} \right\rangle \\
&= \sum_{h_1 \in G} \sum_{h_2 \in G} \langle f_1, \tilde{\phi}_{h_1} \rangle \overline{\langle f_2, \tilde{\phi}_{h_2} \rangle} \langle \phi_{gh_1}, \phi_{gh_2} \rangle \\
&= \sum_{h_1 \in G} \sum_{h_2 \in G} \langle f_1, \tilde{\phi}_{h_1} \rangle \overline{\langle f_2, \tilde{\phi}_{h_2} \rangle} \langle \phi_{h_1}, \phi_{h_2} \rangle \\
&= \left\langle \sum_{h_1 \in G} \langle f_1, \tilde{\phi}_{h_1} \rangle \phi_{h_1}, \sum_{h_2 \in G} \langle f_2, \tilde{\phi}_{h_2} \rangle \phi_{h_2} \right\rangle = \langle f_1, f_2 \rangle.
\end{aligned}$$

Similarly, we have

$$U_g \phi_h = \sum_{h_1 \in G} \langle \phi_h, \tilde{\phi}_{h_1} \rangle \phi_{gh_1} = \sum_{h_1 \in G} \langle \phi_{gh}, \tilde{\phi}_{gh_1} \rangle \phi_{gh_1} = \phi_{gh}.$$

This implies $\rho : G \to \mathscr{U}(\mathscr{H}) : g \mapsto U_g$ is a group homomorphism, since

$$U_{g_1 g_2} \phi_h = \phi_{g_1 g_2 h} = U_{g_1} \phi_{g_2 h} = U_{g_1} U_{g_2} \phi_h, \qquad \mathscr{H} = \mathrm{span}\{\phi_h\}_{h \in G}.$$

Thus ρ is a representation of G with

$$\rho(g)\phi_h = \phi_{gh}, \quad \forall g, h \in G,$$

i.e., Φ is a G-frame for $\mathscr{H}$. □

5.4 The Characterisation of All Tight G-Frames

A complete characterisation of which G-frames are tight, i.e., which orbits $(gv)_{g \in G}$ under a unitary action of G give a tight frame, was given in [17]. Before stating the general theorem, we give a special case with an instructive proof.

Theorem 5.3 *Let $\rho : G \to \mathscr{U}(\mathscr{H})$ be a unitary representation, which is irreducible, i.e.,*

$$\operatorname{span}\{gv : g \in G\} = \mathscr{H}, \quad \forall v \in \mathscr{H}, v \neq 0.$$

Then every orbit $\Phi = (gv)_{g\in G}$, $v \neq 0$ is a tight frame.

Proof Let $v \neq 0$, so that $\Phi = (gv)_{g\in G}$ is a frame. Recall that the frame operator S_Φ is positive definite, so there is an eigenvalue $\lambda > 0$ with corresponding eigenvector w. Since the action is unitary, we calculate

$$S_\Phi(gw) = \sum_{h\in G} \langle gw, hv\rangle hv = g \sum_{h\in G} \langle w, g^{-1}hv\rangle g^{-1}hv = gS_\Phi(w) = \lambda(gw),$$

so that $S_\Phi = \lambda(Id)$ on $\operatorname{span}\{gw : g \in G\} = \mathscr{H}$, i.e., Φ is tight. □

Example 5.6 The symmetry groups of the five platonic solids acting on $\mathbb{R}^3$ as unitary transformations give irreducible representations, as do the dihedral groups acting on $\mathbb{R}^2$. Thus the vertices of the platonic solids and the M equally spaced vectors in $\mathbb{R}^2$ are tight G-frames.

For a given representation, if there exists a G-frame $\Phi = (gv)_{g\in G}$, i.e., $\operatorname{span}\{gv : g \in G\} = \mathscr{H}$, then the canonical tight frame is a tight G-frame. To describe all such tight G-frames, we need a little more terminology.

Definition 5.4 Let G be a finite group. We say that $\mathscr{H}$ is an $\mathbb{F}G$-module if there is a unitary action $(g, v) \mapsto gv$ of G on $\mathscr{H}$, i.e., a representation $G \to \mathscr{U}(\mathscr{H})$.

A linear map $\sigma : V_j \to V_k$ between $\mathbb{F}G$-modules is said to be an $\mathbb{F}G$-*homomorphism* if $\sigma g = g\sigma$, $\forall g \in G$, and an $\mathbb{F}G$-*isomorphism* if σ is a bijection. An $\mathbb{F}G$-module is *irreducible* if the corresponding representation is, and it is *absolutely irreducible* if it is irreducible when thought of as a $\mathbb{C}G$-module in the natural way.

We can now generalise Theorem 5.3.

Theorem 5.4 *Let G be a finite group which acts on $\mathscr{H}$ as unitary transformations, and let*

$$\mathscr{H} = V_1 \oplus V_2 \oplus \cdots \oplus V_m$$

be an orthogonal direct sum of irreducible $\mathbb{F}G$-modules for which repeated summands are absolutely irreducible. Then $\Phi = (gv)_{g\in G}$, $v = v_1 + \cdots + v_m$, $v_j \in V_j$ is a tight G-frame if and only if

$$\frac{\|v_j\|^2}{\|v_k\|^2} = \frac{\dim(V_j)}{\dim(V_k)}, \quad \forall j, k,$$

and $\langle \sigma v_j, v_k\rangle = 0$ when V_j is $\mathbb{F}G$-isomorphic to V_k via $\sigma : V_j \to V_k$. By Schur's lemma there is at most one σ to check.

This result is readily applied; indeed if there is a G-frame, then there is a tight one.

Proposition 5.3 *Let G be a finite group which acts on $\mathscr{H}$ as unitary transformations. If there is a $v \in \mathscr{H}$ for which $(gv)_{g\in G}$ is a frame, i.e., that spans $\mathscr{H}$, then the associated canonical tight frame is a tight G-frame for $\mathscr{H}$.*

This can be used as an alternative way to construct tight G-frames, but requires calculation of the square root of the frame operator.

Example 5.7 One situation where Theorem 5.4 applies is to orthogonal polynomials of several variables for a weight function with some symmetries G, e.g., the inner product on bivariate polynomials given by integration over a triangle. By analogy with the univariate orthogonal polynomials, the orthogonal polynomials of degree k in N variables are those polynomials of degree k which are orthogonal to all the polynomials of degree $< k$. It is natural to seek a G-invariant tight frame for this space of dimension $\binom{k+N-1}{N-1}$. Using Theorem 5.4, G-invariant tight frames with one orbit, i.e., G-frames, can be constructed; e.g., [17] gives an orthonormal basis for the quadratic orthogonal polynomials on the triangle (with constant weight), which is invariant under the action of the dihedral group of symmetries of the triangle.

Example 5.8 For G abelian, all irreducible representations are one dimensional, and it follows that there are only finitely many tight G-frames which can be constructed from these 'characters'. We discuss the resulting *harmonic frames* next.

5.5 Harmonic Frames

The $M \times M$ *Fourier matrix*

$$\frac{1}{\sqrt{M}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega & \omega^2 & \cdots & \omega^{M-1} \\ 1 & \omega^2 & \omega^4 & \cdots & \omega^{2(M-1)} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & \omega^{M-1} & \omega^{2(M-1)} & \cdots & \omega^{(M-1)(M-1)} \end{bmatrix}, \quad \omega := e^{\frac{2\pi i}{M}} \tag{5.4}$$

is a unitary matrix, and so its columns (or rows) form an orthonormal basis for $\mathbb{C}^M$.

Since the projection of an orthonormal basis is a tight frame, an equal norm tight frame for $\mathbb{C}^M$ can be obtained as the columns of any submatrix obtained by taking N rows of the Fourier transform matrix. Tight frames of this type are the most commonly used in applications, due to their simplicity of construction and flexibility (various choices for the rows can be made). They date back at least to [9]; early applications include [8, 11], and have been called *harmonic* or *geometrically uniform tight frames*. They provide a nice example of unit norm tight frames.

Proposition 5.4 *Equal norm tight frames of $M \geq N$ vectors in $\mathbb{C}^N$ exist. Indeed, harmonic ones can be constructed by taking any N rows of the Fourier matrix* (5.4).

For G an abelian group, the irreducible representations are one dimensional, and are usually called (*linear*) *characters* $\xi : G \to \mathbb{C}$. If $G = \mathbb{Z}_M$, the cyclic group of order M, then the M characters are

$$\xi_j : k \mapsto \left(\omega^j\right)^k, \quad j \in \mathbb{Z}_M,$$

i.e., the rows (or columns) of the Fourier matrix (5.4). Thus it follows from Theorem 5.4 that all $\mathbb{Z}_M$-frames for $\mathbb{C}^N$ are obtained by taking N rows (or columns) of the Fourier transform matrix. We now present the general form of this result.

Let G be a finite abelian group of order M, and let $\hat{G}$ be the *character group*, i.e., the set of M characters of G which forms a group under pointwise multiplication. The groups G and $\hat{G}$ are isomorphic, which is easily seen for $G = \mathbb{Z}_M$, though not in a canonical way. The *character table* of G is the table with rows given by the characters of G. Thus the Fourier matrix is, up to a normalising factor, the character table of $\mathbb{Z}_M$, and taking N rows corresponds to taking n characters, or taking N columns corresponds to restricting the characters to N elements of $\mathbb{Z}_M$.

Definition 5.5 Let G be a finite abelian group of order M. We call the G-frame for $\mathbb{C}^N$ obtained by taking N rows or columns of the character table of G, i.e.,

$$\Phi = \left(\left(\xi_j(g)\right)_{j=1}^N\right)_{g\in G}, \quad \xi_1, \ldots, \xi_N \in \hat{G}, \text{ or}$$
$$\Phi = \left(\left(\xi(g_j)\right)_{j=1}^N\right)_{\xi\in \hat{G}}, \quad g_1, \ldots, g_N \in G,$$

a *harmonic frame*.

It is easy to verify that the frames given in this definition are G and $\hat{G}$ frames, respectively. We now characterise the G-frames for G abelian (see [17] for details).

Theorem 5.5 *Let Φ be an equal norm finite tight frame for $\mathbb{C}^N$. Then the following are equivalent*:

1. *Φ is a G-frame, where G is an abelian group.*
2. *Φ is harmonic (obtained from the character table of G).*

Since there is a *finite* number of abelian groups of order M, we conclude the following.

Corollary 5.1 *Fix $M \geq N$. There is a finite number of tight frames of M vectors for $\mathbb{C}^N$ (up to unitary equivalence) which are given by the orbit of an abelian group of $N \times N$ matrices, namely the harmonic frames.*

Example 5.9 Taking the second and last rows of (5.4) gives the following harmonic frame for $\mathbb{C}^2$:

$$\Phi = \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} \omega \\ \bar{\omega} \end{bmatrix}, \begin{bmatrix} \omega^2 \\ \bar{\omega}^2 \end{bmatrix}, \ldots, \begin{bmatrix} \omega^{M-1} \\ \bar{\omega}^{M-1} \end{bmatrix} \right).$$

This is unitarily equivalent to the M equally spaced unit vectors in $\mathbb{R}^2$, via

$$U := \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ -i & i \end{bmatrix}, \qquad \frac{1}{\sqrt{2}} U \begin{bmatrix} \omega^j \\ \overline{\omega}^j \end{bmatrix} = \begin{bmatrix} \cos \frac{2\pi j}{n} \\ \sin \frac{2\pi j}{n} \end{bmatrix}, \quad \forall j.$$

By taking rows in complex conjugate pairs, as in the example above, and the row of 1's when N is odd, we get the following.

Corollary 5.2 *There exists a real harmonic frame of $M \geq N$ vectors for $\mathbb{R}^N$.*

Example 5.10 The smallest noncyclic abelian group is $\mathbb{Z}_2 \times \mathbb{Z}_2$. Its character table can be calculated as the *Kronecker product* of that for $\mathbb{Z}_2$ with itself, giving

$$\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \otimes \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{bmatrix}.$$

Taking any pair of the last three rows gives the harmonic frame

$$\left\{ \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ 1 \end{bmatrix}, \begin{bmatrix} -1 \\ -1 \end{bmatrix}, \begin{bmatrix} 1 \\ -1 \end{bmatrix} \right\},$$

of four equally spaced vectors in $\mathbb{R}^2$, which is also given by $\mathbb{Z}_4$ (see Example 5.9). Taking the first row and any other gives two copies of an orthogonal basis.

Thus, harmonic frames may be given by the character tables of different abelian groups; frames which arise from cyclic groups are called *cyclic* harmonic frames. There exist harmonic frames of M vectors which are *not* cyclic. These seem to be common (see Table 5.1 for when noncyclic abelian groups of order M exist).

The calculations in Table 5.1 come from [10]. Even more efficient algorithms for calculating the numbers of harmonic frames (up to unitary equivalence) can be based on the following result (see [5] for full details).

Definition 5.6 We say that subsets J and K of a finite group G are *multiplicatively equivalent* if there is an automorphism $\sigma : G \to G$ for which $K = \sigma(J)$.

Definition 5.7 We say that two G-frames Φ and Ψ are *unitarily equivalent via an automorphism* if

$$\varphi_g = cU\psi_{\sigma g}, \quad \forall g \in G,$$

Table 5.1 The numbers of inequivalent *non*cyclic, *cyc*lic harmonic frames of $M \leq 35$ distinct vectors for $\mathbb{C}^N$, $N = 2, 3, 4$ when a nonabelian group of order M exists

$N=2$				$N=3$				$N=4$			
M	Non	Cyc	Total	M	Non	Cyc	Total	M	Non	Cyc	Total
4	0	3	3	4	0	3	3	4	0	1	1
8	1	7	8	8	5	16	21	8	8	21	29
9	1	6	7	9	3	15	18	9	5	23	28
12	2	13	15	12	11	57	68	12	30	141	171
16	4	13	17	16	28	74	102	16	139	228	367
18	2	18	20	18	19	121	140	18	80	494	574
20	3	19	22	20	29	137	166	20	154	622	776
24	6	27	33	24	89	241	330	24	604	1349	1953
25	1	15	16	25	8	115	123	25	37	636	673
27	3	18	21	27	33	159	192	27	202	973	1175
28	4	25	29	28	57	255	312	28	443	1697	2140
32	9	25	34	32	158	278	436	32	1379	2152	3531

where $c > 0$, U is unitary, and $\sigma : G \to G$ is an automorphism.

Theorem 5.6 *Let G be a finite abelian group, $J, K \subset G$. The following are equivalent.*

1. *The subsets J and K are multiplicatively equivalent.*
2. *The harmonic frames given by J, K are unitarily equivalent via an automorphism.*

To make effective use of this result, it is convenient to have the following theorem.

Theorem 5.7 [5] *Let G be an abelian group of order M, and let $\Phi = \Phi_J = (\xi|_J)_{\xi \in \hat{G}}$ be the harmonic frame of M vectors for $\mathbb{C}^N$ given by $J \subset G$, where $|J| = N$. Then*

- *Φ has distinct vectors if and only if J generates G.*
- *Φ is a real frame if and only if J is closed under taking inverses.*
- *Φ is a lifted frame if and only if the identity is an element of J.*

Example 5.11 Seven vectors in $\mathbb{C}^3$. For $G = \mathbb{Z}_7$, the seven multiplicative equivalence classes of subsets of size three have representatives

$$\{1,2,6\},\ \{1,2,3\},\ \{0,1,2\},\ \{0,1,3\},\ \{1,2,5\} \quad \text{(class size 6)},$$

$$\{0,1,6\} \quad \text{(class size 3)}, \qquad \{1,2,4\} \quad \text{(class size 2)}.$$

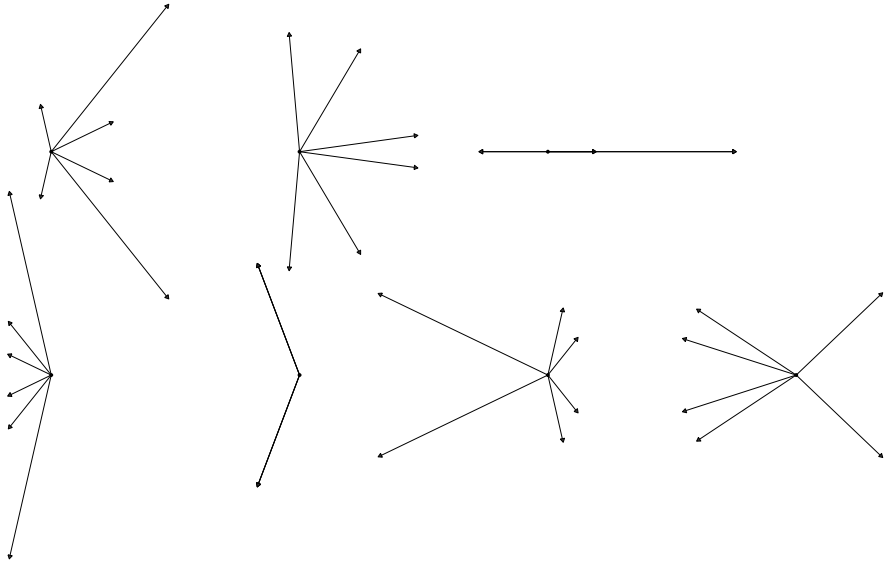


Fig. 5.3 The angle sets $\{\langle\varphi_0, \varphi_j\rangle : j \in G, j \neq 0\} \subset \mathbb{C}$ of the seven inequivalent harmonic frames of seven vectors in $\mathbb{C}^3$. Note that one is real, and three are equiangular

Each gives an harmonic frame of distinct vectors (nonzero elements generate G). None of these are unitarily equivalent since their angles are different (see Fig. 5.3).

Example 5.12 For $G = \mathbb{Z}_8$ there are 17 multiplicative equivalence classes of subsets of 3 elements. Only two of these give frames with the same angles, namely

$$\big\{\{1,2,5\},\{3,6,7\}\big\}, \qquad \big\{\{1,5,6\},\{2,3,7\}\big\}.$$

The common angle multiset is

$$\{-1, i, i, -i, -i, -2i-1, 2i-1\}.$$

These frames are unitarily equivalent, but not via an automorphism.

Due to examples such as this, there is not a complete description of all harmonic frames up to unitary equivalence. There is ongoing work to classify the cyclic harmonic frames. These are the building blocks for all harmonic frames, since abelian groups are products of cyclic groups, and we have the following (see [19]).

Theorem 5.8 *Harmonic frames can be combined as follows*:

- *The direct sum of disjoint harmonic frames is a harmonic frame.*
- *The tensor product of harmonic frames is a harmonic frame.*
- *The complement of a harmonic frame is a harmonic frame.*

5.6 Equiangular Harmonic Frames and Difference Sets

We have seen in Example 5.11 that there exist harmonic frames which are equiangular. These are characterised by the existence of a *difference set* for an abelian group, which leads to some infinite families of equiangular tight frames.

Definition 5.8 An N-element subset J of a finite group G of order M is said to be an (M, N, λ)-*difference set* if every nonidentity element of G can be written as a difference $a - b$ of two elements $a, b \in J$ in exactly λ ways.

Equiangular harmonic frames are in 1–1 correspondence with difference sets.

Theorem 5.9 [20] *Let G be an abelian group of order M. Then the frame of M vectors for $\mathbb{C}^N$ obtained by restricting the characters of G to $J \subset G$, $|J| = N$ is an equiangular tight frame if and only if J is an (M, N, λ)-difference set for G.*

The parameters of a difference set satisfy

$$1 \leq \lambda = \frac{N^2 - N}{M - 1},$$

and so an equiangular harmonic frame of M vectors for $\mathbb{C}^N$ satisfies

$$M \leq N^2 - N + 1.$$

The cyclic case has been used in applications; see, e.g., [13, 21].

Example 5.13 For $G = \mathbb{Z}_7$ three of the seven harmonic frames in Example 5.11 are equiangular, i.e., the ones given by the (multiplicatively inequivalent) difference sets

$$\{1, 2, 4\}, \quad \{1, 2, 6\}, \quad \{0, 1, 3\}.$$

Example 5.14 The *La Jolla Difference Set Repository*

http://www.ccrwest.org/diffsets/diff_sets/

has numerous examples of difference sets.

5.7 Highly Symmetric Tight Frames (and Finite Reflection Groups)

For G abelian, we have seen that there are *finitely* many G-frames. For G *non-abelian*, there are infinitely many. This follows from Theorem 5.4, but is most easily understood by an example. Let $G = D_3$ be the dihedral group of symmetries of the triangle ($|G| = 6$), acting on $\mathbb{R}^2$, so as to express the Mercedes-Benz frame as the

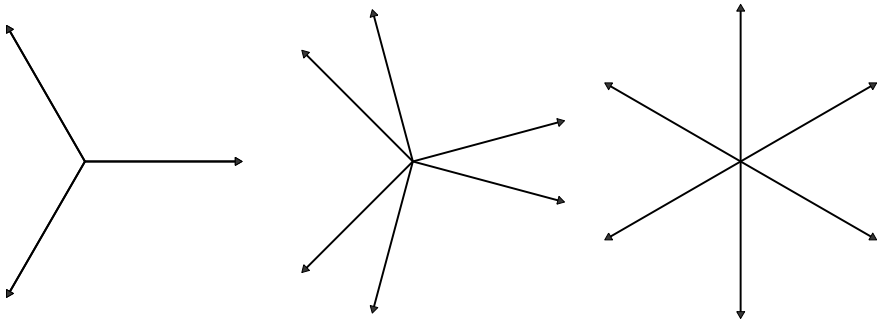


Fig. 5.4 Unitarily inequivalent tight D_3-frames for $\mathbb{R}^2$ given by the orbit of a vector v

orbit of a vector v which is fixed by a reflection. If v is not fixed by a reflection, then its orbit is a tight frame (by Theorem 5.3), and it is easily seen that infinitely many unitarily inequivalent tight D_3-frames of six distinct vectors for $\mathbb{R}^2$ can be obtained in this way (see Fig. 5.4).

All is not lost! We now consider two ways in which a finite class of G-frames can be obtained from a nonabelian (abstract) group G. The first seeks to identify the distinguishing feature of the Mercedes-Benz frame amongst the possibilities indicated by Fig. 5.4, and the second (Sect. 5.8) generalises the notion of a harmonic frame.

Motivated by the Mercedes-Benz example, we have the following definition.

Definition 5.9 A finite frame Φ of distinct vectors is *highly symmetric* if the action of its symmetry group $\mathrm{Sym}(\Phi)$ is irreducible, transitive, and the stabiliser of any one vector (and hence all) is a nontrivial subgroup which fixes a space of dimension exactly one.

Example 5.15 The standard orthonormal basis $\{e_1, \ldots, e_N\}$ is not a highly symmetric tight frame for $\mathbb{F}^N$, since its symmetry group fixes the vector $e_1 + \cdots + e_N$. However, the vertices of the regular simplex always are (the Mercedes-Benz frame is the case $N = 2$). Since both of these frames are harmonic, we conclude that a harmonic frame may or may not be highly symmetric. Moreover, for many harmonic frames of M vectors the symmetry group has order M (cf. [10]), which implies that they are not highly symmetric.

Example 5.16 The vertices of the platonic solids in $\mathbb{R}^3$, and the M equally spaced unit vectors in $\mathbb{R}^2$ are highly symmetric tight frames.

Theorem 5.10 *Fix $M \geq N$. There is a finite number of highly symmetric Parseval frames of M vectors for $\mathbb{F}^N$ (up to unitary equivalence).*

Proof Suppose that Φ is a highly symmetric Parseval frame of M vectors for $\mathbb{F}^N$. Then it is determined, up to unitary equivalence, by the representation induced by

$\mathrm{Sym}(\Phi)$, and a subgroup H which fixes only the one-dimensional subspace spanned by some vector in Φ. There is a finite number of choices for $\mathrm{Sym}(\Phi)$ since its order is $\leq |S_M| = M!$, and hence (by Maschke's theorem) a finite number of possible representations. As there is only a finite number of choices for H, it follows that the class of such frames is finite. □

The highly symmetric tight frames have only recently been defined in [4], where those corresponding to the Shephard–Todd classification of the *finite reflection groups* and *complex polytopes* were enumerated. We give a couple of examples [4].

Example 5.17 Let $G = G(1,1,8) \cong S_8$, a member of one of the three infinite families of *imprimitive irreducible complex reflection groups* acting as permutations of the indices of a vector $x \in \mathbb{C}^8$ in the subspace consisting of vectors with $x_1 + \cdots + x_8 = 0$. The orbit of the vector

$$v = 3w_2 = (3, 3, -1, -1, -1, -1, -1, -1)$$

gives an equiangular tight frame of 28 vectors for a 7-dimensional space.

Example 5.18 The *Hessian* is the regular complex polytope with 27 vertices and Schläfli symbol $3\{3\}3\{3\}3$. Its symmetry group (Shephard–Todd) ST 25 (of order 648) is generated by the following three reflections of order 3:

$$R_1 = \begin{pmatrix} \omega & & \\ & 1 & \\ & & 1 \end{pmatrix}, \quad R_2 = \frac{1}{3}\begin{pmatrix} \omega+2 & \omega-1 & \omega-1 \\ \omega-1 & \omega+2 & \omega-1 \\ \omega-1 & \omega-1 & \omega+2 \end{pmatrix},$$

$$R_3 = \begin{pmatrix} 1 & & \\ & 1 & \\ & & \omega \end{pmatrix}, \quad \omega = e^{\frac{2\pi i}{3}},$$

and it has $v = (1, -1, 0)$ as a vertex (cf. [6]). These vertices are the H-orbit of v, with H the Heisenberg group, which is a *Heisenberg frame* (see Sect. 5.9). In particular, they are a highly symmetric tight frame. We observe that H is normal in $G = \langle R_1, R_2, R_3 \rangle$.

The classification of all highly symmetric tight frames is in its infancy.

5.8 Central G-Frames

To narrow down the class of unitarily inequivalent G-frames for G nonabelian (which is infinite), we impose an additional symmetry condition.

Definition 5.10 A G-frame $\Phi = (\varphi_g)_{g \in G}$ is said to be *central* if $\nu : G \to \mathbb{C}$ defined by

$$\nu(g) := \langle \varphi_1, \varphi_g \rangle = \langle \varphi_1, g\varphi_1 \rangle$$

is a class function, i.e., is constant on the conjugacy classes of G.

It is easy to see that being central is equivalent to the *symmetry condition*

$$\langle g\varphi, h\varphi\rangle = \langle g\psi, h\psi\rangle, \quad \forall g, h \in G,\ \forall \varphi, \psi \in \Phi.$$

Example 5.19 For G abelian, all G-frames are central, since the conjugacy classes of an abelian group are singletons.

Thus central G-frames generalise harmonic frames to G nonabelian.

Definition 5.11 Let $\rho : G \to \mathscr{U}(\mathscr{H})$ be a representation of a finite group G. The *character* of ρ is the map $\chi = \chi_\rho : G \to \mathbb{C}$ defined by

$$\chi(g) := \operatorname{trace}\big(\rho(g)\big).$$

We now characterise all central Parseval G-frames in terms of the Gramian. In particular, it turns out that the class of central G-frames is *finite*.

Theorem 5.11 [18] *Let G be a finite group with irreducible characters $\chi_1, \dots, \chi_r$. Then $\Phi = (\varphi_g)_{g\in G}$ is a central Parseval G-frame if and only if its Gramian is given by*

$$\operatorname{Gram}(\Phi)_{g,h} = \sum_{i\in I} \frac{\chi_i(1)}{|G|}\overline{\chi_i}\big(g^{-1}h\big), \tag{5.5}$$

for some $I \subset \{1, \dots, r\}$.

The central G-frames can be constructed from the irreducible characters of G, in a similar way to the harmonic frames.

Corollary 5.3 *Let G be a finite group with irreducible characters $\chi_1, \dots, \chi_r$. Choose Parseval G-frames Φ_i for $\mathscr{H}_i$, $i = 1, \dots, r$, with*

$$\operatorname{Gram}(\Phi_i) = \frac{\chi_i(1)}{|G|} M(\overline{\chi_i}), \qquad \dim(\mathscr{H}_i) = \chi_i(1)^2,$$

e.g., take the columns of $\operatorname{Gram}(\Phi_i)$. *Then the unique (up to unitary equivalence) central Parseval G-frame with Gramian (5.5) is given by the direct sum*

$$\bigoplus_{i\in I} \Phi_i \subset \mathscr{H} := \bigoplus_{i\in I} \mathscr{H}_i.$$

Further, if $\rho_i : G \to U(\mathbb{C}^{d_i})$ is a representation with character χ_i, then Φ_i can be given as

$$\Phi_i := \sqrt{\frac{\chi_i(1)}{|G|}}\big(\rho_i(g)\big)_{g\in G} \subset U\big(\mathbb{C}^{d_i}\big) \subset \mathbb{C}^{d_i\times d_i} \approx \mathbb{C}^{d_i^2}, \tag{5.6}$$

where the inner product on the space of $d_i \times d_i$ matrices is $\langle A, B\rangle := \operatorname{trace}(B^*A)$.

Example 5.20 Let $G = D_3 \cong S_3$ be the dihedral group (symmetric group) of order 6,

$$G = D_3 = \langle a, b : a^3 = 1, b^2 = 1, b^{-1}ab = a^{-1} \rangle,$$

and write class functions and G-matrices with respect to the order $1, a, a^2, b, ab, a^2b$. The conjugacy classes are $\{1\}, \{a, a^2\}, \{b, ab, a^2b\}$, and the irreducible characters are

$$\chi_1 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ 1 \\ 1 \\ 1 \end{bmatrix}, \qquad \chi_2 = \begin{bmatrix} 1 \\ 1 \\ 1 \\ -1 \\ -1 \\ -1 \end{bmatrix}, \qquad \chi_3 = \begin{bmatrix} 2 \\ -1 \\ -1 \\ 0 \\ 0 \\ 0 \end{bmatrix}.$$

Corresponding to each of these, there is a central Parseval G-frame Φ_i for a space of dimension $\chi_i(1)^2$. Since χ_1 and χ_2 are one dimensional, (5.6) gives

$$\Phi_1 = \frac{1}{\sqrt{6}}(1, 1, 1, 1, 1, 1), \qquad \Phi_2 = \frac{1}{\sqrt{6}}(1, 1, 1, -1, -1, -1).$$

A representation $\rho : D_3 \to U(\mathbb{C}^2) \subset \mathbb{C}^{2\times 2} \approx \mathbb{C}^4$ with $\operatorname{trace}(\rho) = \chi_3$ is given by

$$\rho(1) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} \approx \begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \qquad \rho(a) = \begin{pmatrix} \omega & 0 \\ 0 & \omega^2 \end{pmatrix} \approx \begin{bmatrix} \omega \\ 0 \\ 0 \\ \omega^2 \end{bmatrix},$$

$$\rho(a^2) = \begin{pmatrix} \omega^2 & 0 \\ 0 & \omega \end{pmatrix} \approx \begin{bmatrix} \omega^2 \\ 0 \\ 0 \\ \omega \end{bmatrix}, \qquad \rho(b) = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \approx \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix},$$

$$\rho(ab) = \begin{pmatrix} 0 & \omega \\ \omega^2 & 0 \end{pmatrix} \approx \begin{bmatrix} 0 \\ \omega \\ \omega^2 \\ 0 \end{bmatrix}, \qquad \rho(a^2b) = \begin{pmatrix} 0 & \omega^2 \\ \omega & 0 \end{pmatrix} \approx \begin{bmatrix} 0 \\ \omega^2 \\ \omega \\ 0 \end{bmatrix},$$

and so we obtain from (5.6)

$$\Phi_3 = \frac{1}{\sqrt{3}} \left(\begin{bmatrix} 1 \\ 0 \\ 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \omega \\ 0 \\ 0 \\ \omega^2 \end{bmatrix}, \begin{bmatrix} \omega^2 \\ 0 \\ 0 \\ \omega \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \omega \\ \omega^2 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \omega^2 \\ \omega \\ 0 \end{bmatrix} \right).$$

Thus there are seven central Parseval D_3-frames, namely

$$\Phi_1, \Phi_2 \subset \mathbb{C}, \qquad \Phi_1 \oplus \Phi_2 \subset \mathbb{C}^2, \qquad \Phi_3 \subset \mathbb{C}^4$$
$$\Phi_1 \oplus \Phi_3, \Phi_2 \oplus \Phi_3 \subset \mathbb{C}^5, \qquad \Phi_1 \oplus \Phi_2 \oplus \Phi_3 \subset \mathbb{C}^6.$$

5.9 Heisenberg Frames (SIC–POVMs) Zauner's Conjecture

The Mercedes-Benz frame gives three equiangular lines in $\mathbb{R}^2$. The search for such sets of equiangular lines in $\mathbb{R}^N$ has a long history, and effectively spawned the area of *algebraic graph theory* (see [7]).

Recently, sets of $M = N^2$ equiangular lines in $\mathbb{C}^N$, equivalently equiangular tight frames of $M = N^2$ vectors in $\mathbb{C}^N$, have been constructed numerically, and, in some cases, analytically. We note that N^2 is the maximum number of vectors possible for an equiangular tight frame for $\mathbb{C}^N$ [15]. Such frames are known as *SIC-POVMs* (symmetric informationally complete positive operator valued measures) in quantum information theory (see [15]), where they are of considerable interest. The claim that they exist for all N is usually known as *Zauner's conjecture* (see [22]).

We now explain how such equiangular tight frames have been, and are expected to be constructed, as the orbit of a (Heisenberg) group.

Fix $N \geq 1$, and let ω be the primitive N-th root of unity

$$\omega := e^{2\pi i/N}.$$

Let $T \in \mathbb{C}^{N\times N}$ be the cyclic shift matrix, and $\Omega \in \mathbb{C}^{N\times N}$ the diagonal matrix

$$T := \begin{bmatrix} 0 & 0 & 0 & \cdots & 0 & 1 \\ 1 & 0 & 0 & \cdots & 0 & 0 \\ 0 & 1 & 0 & \cdots & 0 & 0 \\ \cdot & & \cdot & & & \cdot \\ \cdot & & & \cdot & & \cdot \\ 0 & 0 & 0 & & 1 & 0 \end{bmatrix}, \qquad \Omega := \begin{bmatrix} 1 & 0 & 0 & \cdot & \cdot & 0 \\ 0 & \omega & 0 & \cdot & \cdot & 0 \\ 0 & 0 & \omega^2 & & & 0 \\ \cdot & \cdot & & \cdot & & \\ \cdot & \cdot & & & \cdot & \\ 0 & 0 & 0 & & & \omega^{N-1} \end{bmatrix}.$$

These have order N, i.e., $T^N = \Omega^N = Id$, and satisfy the *commutativity relation*

$$\Omega^k T^j = \omega^{jk} T^j \Omega^k. \tag{5.7}$$

In particular, the group generated by T and Ω contains the scalar matrices $\omega^r Id$.

Definition 5.12 The group $H = \langle T, \Omega \rangle$ generated by the matrices T and Ω is called the *discrete Heisenberg group modulo N*, or for short the *Heisenberg group*.

In view of (5.7), the Heisenberg group has order N^3, and is given explicitly by

$$H = \left\{ \omega^r T^j \Omega^k \,:\, 0 \leq r, j, k \leq N-1 \right\}.$$

Since ω, T, Ω have order N, it is convenient to allow the indices of $\omega^r T^j \Omega^k$ to be integers modulo N. Since T and Ω are unitary, H is a group of unitary matrices.

The action of H on $\mathbb{C}^N$ is irreducible, and so by Theorem 5.3, every orbit $(gv)_{g\in H}$, $v \neq 0$ is a tight frame for $\mathbb{C}^N$. For j, k fixed, the N vectors $\omega^r T^j \Omega^k v$, $0 \leq r \leq N-1$ are scalar multiples of each other, which we identify together. It is

in this sense that the orbit of H is interpreted as a set of N^2 (hopefully equiangular) vectors:

$$\Phi := \left\{T^j \Omega^k v\right\}_{(j,k)\in\mathbb{Z}_N\times\mathbb{Z}_N}. \tag{5.8}$$

This Φ is the *Gabor system* given by the subset $\Lambda = \mathbb{Z}_N \times \mathbb{Z}_N \cong G \times \hat{G}$, $G = \mathbb{Z}_N$ (see Chap. 6—Gabor frames).

Definition 5.13 We call a tight frame Φ of the form (5.8) a *Heisenberg frame* if it is an equiangular tight frame, i.e., a SIC–POVM, and the v a *generating vector*.

Example 5.21 The vector

$$v = \frac{1}{\sqrt{6}} \begin{pmatrix} \sqrt{3+\sqrt{3}} \\ e^{\frac{\pi}{4}i}\sqrt{3-\sqrt{3}} \end{pmatrix}$$

generates a Heisenberg frame of 4 equiangular vectors for $\mathbb{C}^2$. To date (see [16]), there are known analytic solutions for $N = 2, 3, \ldots, 15, 19, 24, 35, 48$.

Starting with [15], there have been numerous attempts to find generating vectors v for various dimensions N, starting from numerical solutions. The current state of affairs is summarised in [16]. We now outline some of the salient points.

The key ideas for finding generating vectors are as follows.

- Solve an equivalent simplified set of equations.
- Find a generating vector with special properties.
- Understand the relationship between generating vectors.

For a unit vector $v \in \mathbb{C}^N$, the condition that it generate a Heisenberg frame is

$$\left|\langle gv, hv\rangle\right| = \frac{1}{\sqrt{N+1}}, \quad j \neq k \quad \Longleftrightarrow \quad \left|\langle v, T^j \Omega^k v\rangle\right| = \frac{1}{\sqrt{N+1}}, \quad j,k \in \mathbb{Z}_N.$$

This is not amenable to numerical calculation. In [15], the *second frame potential*,

$$f(v) = \sum_{j=0}^{N-1}\sum_{k=0}^{N-1} \left|\langle v, T^j \Omega^k v\rangle\right|^4,$$

was minimised over all v satisfying $g(v) = \|v\|^2 = 1$. A minimiser of this constrained optimisation problem with

$$f(v) = 1 + \left(N^2 - 1\right)\frac{1}{(\sqrt{N+1})^4} = \frac{2N}{N+1}$$

is a generating vector. Various simplified equations for finding generators have been proposed, most notably (see [1, 2, 14]) the following.

Theorem 5.12 *A vector $v = (z_j)_{j\in\mathbb{Z}_N}$ is a generating vector for a Heisenberg frame if and only if*

$$\sum_{j\in\mathbb{Z}_N} z_j\overline{z}_{j+s}\overline{z}_{t+j}z_{j+s+t} = \begin{cases} 0, & s,t\neq 0; \\ \frac{1}{N+1}, & s\neq 0, t=0,\ s=0, t\neq 0; \\ \frac{2}{N+1}, & (s,t)=(0,0). \end{cases}$$

If v generates a Heisenberg frame, and b is a unitary matrix which normalises the Heisenberg group, then bv is also a generating vector, since

$$\big|\langle (bv), g(bv)\rangle\big| = \big|\langle v, b^*gbv\rangle\big| = \big|\langle v, b^{-1}gbv\rangle\big| = \frac{1}{\sqrt{N+1}}, \quad g\in H, g\neq Id.$$

The normaliser of H in the unitary matrices is often called the *Clifford group*. This group contains the Fourier matrix, since

$$F^{-1}\big(T^j\Omega^k\big)F = \omega^{-jk}T^k\Omega^{-j}\in H,$$

and the matrix Z given by

$$(Z)_{jk} := \frac{1}{\sqrt{d}}\mu^{j(j+d)+2jk}, \quad \mu := e^{\frac{2\pi i}{2N}} = \omega^{\frac{1}{2}},$$

since

$$Z^{-1}\big(T^j\Omega^k\big)Z = \mu^{j(d+j-2k)}T^{k-j}\Omega^{-j}.$$

A scalar multiple of Z has order 3, i.e., $Z^3 = \sqrt{i}^{1-d}$, $\sqrt{i} := e^{\frac{2\pi i}{8}}$. The strong form of Zauner's conjecture is as follows.

Conjecture 5.1 (Zauner) *Every generating vector for a Heisenberg frame (up to unitary equivalence) is an eigenvector of Z.*

All known generating vectors (both numerical and analytic) support this conjecture. Indeed, many were found as eigenvectors of Z. Without doubt, the solution of Zauner's conjecture, and the construction of equiangular tight frames in general, is one of the central problems in the construction of tight frames via groups. This field in still in its infancy: frames given as the orbit of more than one vector (G-invariant fusion frames) have scarcely been studied.

References

1. Appleby, D.M., Dang, H.B., Fuchs, C.A.: Symmetric informationally-complete quantum states as analogues to orthonormal bases and minimum-uncertainty states (2007). arXiv:0707.2071v2 [quant-ph]

2. Bos, L., Waldron, S.: Some remarks on Heisenberg frames and sets of equiangular lines. N.Z. J. Math. **36**, 113–137 (2007)
3. Brauer, R., Coxeter, H.S.M.: A generalization of theorems of Schöhardt and Mehmke on polytopes. Trans. R. Soc. Canada Sect. III **34**, 29–34 (1940)
4. Broome, H., Waldron, S.: On the construction of highly symmetric tight frames and complex polytopes. Preprint (2010)
5. Chien, T., Waldron, S.: A classification of the harmonic frames up to unitary equivalence. Appl. Comput. Harmon. Anal. **30**, 307–318 (2011)
6. Coxeter, H.S.M.: Regular Complex Polytopes. Cambridge University Press, Cambridge (1991)
7. Godsil, C., Royle, G.: Algebraic Graph Theory. Springer, New York (2001)
8. Goyal, V.K., Kovačević, J., Kelner, J.A.: Quantized frame expansions with erasures. Appl. Comput. Harmon. Anal. **10**, 203–233 (2001)
9. Goyal, V.K., Vetterli, M., Thao, N.T.: Quantized overcomplete expansions in $\mathbb{R}^n$: analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory **44**, 16–31 (1998)
10. Hay, N., Waldron, S.: On computing all harmonic frames of n vectors in $\mathbb{C}^d$. Appl. Comput. Harmon. Anal. **21**, 168–181 (2006)
11. Hochwald, B., Marzetta, T., Richardson, T., Sweldens, W., Urbanke, R.: Systematic design of unitary space-time constellations. IEEE Trans. Inf. Theory **46**, 1962–1973 (2000)
12. James, G., Liebeck, M.: Representations and Characters of Groups. Cambridge University Press, Cambridge (1993)
13. Kalra, D.: Complex equiangular cyclic frames and erasures. Linear Algebra Appl. **419**, 373–399 (2006)
14. Khatirinejad, M.: On Weyl-Heisenberg orbits of equiangular lines. J. Algebr. Comb. **28**, 333–349 (2008)
15. Renes, J.M., Blume–Kohout, R., Scott, A.J., Caves, C.M.: Symmetric informationally complete quantum measurements. J. Math. Phys. **45**, 2171–2180 (2004)
16. Scott, A.J., Grassl, M.: SIC-POVMs: A new computer study (2009). arXiv:0910.5784v2 [quant-ph]
17. Vale, R., Waldron, S.: Tight frames and their symmetries. Constr. Approx. **21**, 83–112 (2005)
18. Vale, R., Waldron, S.: Tight frames generated by finite nonabelian groups. Numer. Algorithms **48**, 11–27 (2008)
19. Vale, R., Waldron, S.: The symmetry group of a finite frame. Linear Algebra Appl. **433**, 248–262 (2010)
20. Waldron, S.: An Introduction to Finite Tight Frames. Springer, New York (2011)
21. Xia, P., Zhou, S., Giannakis, G.B.: Achieving the Welch bound with difference sets. IEEE Trans. Inf. Theory **51**, 1900–1907 (2005)
22. Zauner, G.: Quantendesigns—Grundzüge einer nichtkommutativen Designtheorie. Doctorial thesis, University of Vienna, Vienna, Austria (1999)

Chapter 6
Gabor Frames in Finite Dimensions

Götz E. Pfander

Abstract Gabor frames have been extensively studied in time-frequency analysis over the last 30 years. They are commonly used in science and engineering to synthesize signals from, or to decompose signals into, building blocks which are localized in time and frequency. This chapter contains a basic and self-contained introduction to Gabor frames on finite-dimensional complex vector spaces. In this setting, we give elementary proofs of the central results on Gabor frames in the greatest possible generality; that is, we consider Gabor frames corresponding to lattices in arbitrary finite Abelian groups. In the second half of this chapter, we review recent results on the geometry of Gabor systems in finite dimensions: the linear independence of subsets of its members, their mutual coherence, and the restricted isometry property for such systems. We apply these results to the recovery of sparse signals, and discuss open questions on the geometry of finite-dimensional Gabor systems.

Keywords Gabor analysis on finite Abelian groups · Linear independence · Coherence · Restricted isometry constants of Gabor frames · Applications to compressed sensing · Erasure channel error correction · Channel identification

6.1 Introduction

In his seminal 1946 paper "Theory of Communication," Dennis Gabor suggested the decomposition of the time-frequency *information area* of a communications channel into the smallest possible boxes that allow exactly one information-carrying coefficient to be transmitted per box [41]. He refers to Heisenberg's uncertainty principle to argue that the smallest time-frequency boxes are achieved using time-frequency shifted copies of *probability functions*, that is, of Gaussians. In summary, he proposes transmitting the information-carrying complex-valued sequence $\{c_{nk}\}$ in the

G.E. Pfander (✉)
School of Engineering and Science, Jacobs University, 28759 Bremen, Germany
e-mail: g.pfander@jacobs-university.de

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_6,

form of the signal

$$\psi(t) = \sum_{n=-\infty}^{\infty} \sum_{k=-\infty}^{\infty} c_{nk}\, e^{-\pi \frac{(t-n\Delta t)^2}{2(\Delta t)^2}} e^{2\pi i \frac{kt}{\Delta t}},$$

where the parameter $\Delta t > 0$ can be chosen depending on physical consideration and the application at hand. Denoting *modulation operators* by

$$M_\nu g(t) = e^{2\pi i \nu t} g(t), \quad \nu \in \mathbb{R},$$

and translation operators by

$$T_\tau g(t) = g(t-\tau), \quad \tau \in \mathbb{R},$$

Gabor proposed to transmit on the carriers $\{M_{k/\Delta t} T_{n\Delta t} g_0\}_{n,k\in\mathbb{Z}}$, where g_0 is the *Gaussian window function* $g_0(t) = e^{-\pi \frac{t^2}{2(\Delta t)^2}}$.

In the second half of the twentieth century, the suggestion of Gabor, and in general the interplay of information density in time and in frequency, was studied extensively; see, for example, [24, 25, 33, 38, 61–63, 88]. This line of work focuses on functional analytic properties of function systems such as the ones suggested by Gabor. Apart from the following historical remarks, functional analysis will not play a role in this chapter. Janssen, for instance, analyzed in detail in which sense $\{M_{k/\Delta t} T_{n\Delta t} g_0\}_{n,k\in\mathbb{Z}}$ can be used to represent functions and distributions. He showed that while being complete in the Hilbert space of square integrable functions on the real line, the set suggested by Gabor is not a Riesz basis for this space [53].[1] Balian and Low then established independently from one another that any function φ which is *well concentrated* in time and in frequency does not give rise to a Riesz basis of the form $\{M_{k/\Delta t} T_{n\Delta t} \varphi\}_{n,k\in\mathbb{Z}}$ [5, 10, 11, 66]. This apparent failure of systems structured as suggested by Gabor was then rectified by resorting to the concept of frames that had been introduced by Duffin and Shaeffer [30]. Indeed, $\{M_{k\Delta\nu} T_{n\Delta t} g_0\}_{n,k\in\mathbb{Z}}$ is a frame if $\Delta\nu < 1/\Delta t$ [67, 85, 86]. Since then the theory of Gabor systems has been intimately related to the theory of frames, and many problems in frame theory find their origins in Gabor analysis. For example, the Feichtinger conjecture (see Sect. 11.2.3 and references therein), and what are called localized frames were first considered in the realm of Gabor frames [3, 4, 19, 48].

In engineering, Gabor's idea flourished over the last decade due to the increasing use of *orthogonal frequency division multiplexing* (OFDM) structured communication systems. Indeed, the carriers used in OFDM are $\{M_{k\Delta\nu} T_{n\Delta t} \varphi_0\}_{n\in\mathbb{Z},k\in K}$, where φ_0 is the characteristic function $\chi_{[0,1/\Delta\nu]}$ (or a mollified and/or cyclically extended copy thereof) and $K = \{-K_2, -K_2+1, \ldots, -K_1,\ K_1+1, \ldots, K_2\}$ is introduced to respect transmission band limitations.

[1] Prior to the work of Gabor, von Neumann postulated that the function family which is now referred to as the *Gaussian Gabor* system is complete [70] (see the respective discussions in [46, 49]).

While originally constructed on the real line, Gabor systems can be analogously defined on any locally compact Abelian group [21, 34, 37, 45]. Functions on finite Abelian groups form finite-dimensional vector spaces; hence, Gabor systems on finite groups have been studied first in the realm of numerical linear algebra. In particular, efficient matrix factorizations for Gabor analysis, Gabor synthesis, and Gabor frame operators are discussed in the literature; see, for example, [6, 79, 80, 91].

Gabor systems on finite cyclic groups have also been studied numerically in order to better understand properties of Gabor systems on the real line. The relationships between Gabor systems on the real line, on the integers, and on cyclic groups are studied based on sampling and periodization arguments in [55, 56, 71, 89, 90].

Over the last two decades it became apparent that the structure of Gabor frames on finite Abelian groups allows for the construction of finite frames with remarkable geometric properties. Most noteworthy may be the fact that many equiangular frames have been constructed as Gabor frames (for references and details, see Sect. 5.9). Also, finite Gabor systems have been considered in the study of *constant amplitude zero autocorrelation* (CAZAC) sequences [8, 9, 43, 87] and to construct spreading sequences and error-correcting codes in radar and communications [51].

This chapter serves multiple purposes. In Sects. 6.2 and 6.3 we give an elementary introduction to Gabor analysis on $\mathbb{C}^N$. Section 6.2 focuses on basic definitions, and in Sect. 6.3 we describe the fundamental ideas that make Gabor frames useful to analyze or synthesize signals with varying frequency components.

In Sect. 6.4, we define and discuss Gabor frames on finite Abelian groups. The case of Gabor frames on general finite Abelian groups is only more technically involved than the setup chosen in Sect. 6.2. This is due to the fundamental theorem of finite Abelian groups: it states that every finite Abelian group is isomorphic to the product of finite cyclic groups.

We prove fundamental results for Gabor frames on finite Abelian groups in Sect. 6.5. The properties discussed are well known, but the proofs contained in the literature involve nontrivial concepts from representation theory which we will replace with simple arguments from linear algebra.

The results in Sect. 6.5 are phrased for general finite Abelian groups, but we expect that some readers may want to skip Sect. 6.4 and simply assume in Sects. 6.5–6.9 that the group G is cyclic, as was done in Sects. 6.2 and 6.3.

We discuss geometric properties of Gabor frames in Sects. 6.6–6.9. In Sect. 6.6, we address the question of whether Gabor frames that are in general linear position, meaning any N vectors of a Gabor system are linearly independent in the underlying N-dimensional ambient space, can be constructed. As one of the byproducts of our discussion, we will establish the existence of a large class of unimodular tight Gabor frames which are maximally robust to erasures. In Sect. 6.7, we address the coherence of Gabor systems, and in Sect. 6.8 we state estimates for the probability that a randomly chosen Gabor window generates a Gabor frame which has useful *restricted isometry constants* (RICs). In Sect. 6.9, we state some results on Gabor frames in the framework of compressed sensing.

Throughout the chapter, we will not discuss multiwindow Gabor frames. For details on the structure of multiwindow Gabor frames, see [35, 65] and references therein.

6.2 Gabor Frames for $\mathbb{C}^N$

For reasons that become apparent in Sect. 6.4, we index the components of a vector $x \in \mathbb{C}^N$ by $\{0, 1, 2, \ldots, N-2, N-1\}$, namely, by the N-element *cyclic group* $\mathbb{Z}_N = \mathbb{Z}/N\mathbb{Z}$. Moreover, to avoid algebraic operations on indices, we write $x(k)$ rather than x_k for the k-th component of the column vector x. That is, we write

$$x = (x_0, x_1, x_2, \ldots, x_{N-2}, x_{N-1})^T = \big(x(0), x(1), x(2), \ldots, x(N-2), x(N-1)\big)^T,$$

where x^T denotes the transpose of the vector x.

The *(discrete) Fourier transform* $\mathscr{F} : \mathbb{C}^N \longrightarrow \mathbb{C}^N$ plays a fundamental role in Gabor analysis. It is given pointwise by

$$\mathscr{F}x(m) = \widehat{x}(m) = \sum_{n=0}^{N-1} x(n)\, e^{-2\pi i mn/N}, \quad m = 0, 1, \ldots, N-1. \tag{6.1}$$

Throughout this chapter, operators are defined by their action on column vectors, and we will not distinguish between an operator and its matrix representation with respect to the Euclidean basis $\{e_k\}_{k=0,1,\ldots,N-1}$, where $e_k(n) = \delta(k-n) = 1$ if $k = n$ and $e_k(n) = \delta(k-n) = 0$ else.

In matrix notation, the discrete Fourier transform (6.1) is represented by the *Fourier matrix* $W_N = (\omega^{-rs})_{r,s=0}^{N-1}$ with $\omega = e^{2\pi i/N}$. For example, we have

$$W_4 = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -i & -1 & i \\ 1 & -1 & 1 & -1 \\ 1 & i & -1 & -i \end{pmatrix},$$

$$W_6 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 \\ 1 & e^{-2\pi i 1/6} & e^{-2\pi i 1/3} & e^{-2\pi i 1/2} & e^{-2\pi i 2/3} & e^{-2\pi i 5/6} \\ 1 & e^{-2\pi i 1/3} & e^{-2\pi i 2/3} & 1 & e^{-2\pi i 1/3} & e^{-2\pi i 2/3} \\ 1 & e^{-2\pi i 1/2} & 1 & e^{-2\pi i 3/6} & 1 & e^{-2\pi i 1/2} \\ 1 & e^{-2\pi i 2/3} & e^{-2\pi i 1/3} & 1 & e^{-2\pi i 2/3} & e^{-2\pi i 1/3} \\ 1 & e^{-2\pi i 5/6} & e^{-2\pi i 2/3} & e^{-2\pi i 1/2} & e^{-2\pi i 1/3} & e^{-2\pi i 1/6} \end{pmatrix}.$$

The *fast Fourier transform* (FFT) provides an efficient algorithm to compute matrix vector products of the form $W_N x$ [14, 23, 58, 81].

The most important properties of the Fourier transform are the *Fourier inversion formula* (6.2), the *Parseval–Plancherel formula* (6.3), and the *Poisson summation formula* (6.5).

Theorem 6.1 *The normalized harmonics* $\frac{1}{\sqrt{N}} e^{2\pi i m(\cdot)/N}$, $m = 0, 1, \ldots, N-1$, *form an orthonormal basis of* $\mathbb{C}^N$ *and, hence, we have*

$$x = \frac{1}{N} \sum_{m=0}^{N-1} \widehat{x}(m) e^{2\pi i m(\cdot)/N}, \quad x \in \mathbb{C}^N, \tag{6.2}$$

and

$$\langle x, y \rangle = \frac{1}{N} \langle \widehat{x}, \widehat{y} \rangle, \quad x, y \in \mathbb{C}^N. \tag{6.3}$$

Moreover, for natural numbers a *and* b *with* $ab = N$ *we have*

$$\sum_{n=0}^{b-1} e^{2\pi i a m n/N} = \begin{cases} b, & \text{if } m \text{ is a multiple of } b, \\ 0, & \text{otherwise}, \end{cases} \tag{6.4}$$

and

$$a \sum_{n=0}^{b-1} x(an) = \sum_{m=0}^{a-1} \widehat{x}(bm), \quad x \in \mathbb{C}^N. \tag{6.5}$$

Proof We first prove (6.4). If m is a multiple of b, then $e^{2\pi i a m n/N} = 1$ for all $n = 0, 1, \ldots, b-1$, and (6.4) holds. Else, $z = e^{2\pi i a m/N} \neq 1$, and using the geometric sum formula, we obtain

$$\sum_{n=0}^{b-1} e^{2\pi i a m n/N} = \sum_{n=0}^{b-1} z^n = \left(1 - z^b\right)/(1 - z) = (1 - 1)/(1 - z) = 0.$$

Setting $a = 1$ and $b = N$ in (6.4) implies the orthonormality of the normalized harmonics, in fact,

$$\left\langle \frac{1}{\sqrt{N}} e^{2\pi i m(\cdot)/N}, \frac{1}{\sqrt{N}} e^{2\pi i m'(\cdot)/N} \right\rangle = \frac{1}{N} \sum_{n=0}^{N-1} e^{2\pi i (m-m')n/N} \overset{(6.4)}{=} \begin{cases} 1, & \text{if } m = m', \\ 0, & \text{otherwise}, \end{cases}$$

and the reconstruction formula (6.2) and Parseval–Plancherel (6.3) follow.

To obtain (6.5) and thereby complete the proof, we compute

$$\sum_{n=0}^{b-1} x(an) \overset{(6.2)}{=} \sum_{n=0}^{b-1} \frac{1}{N} \sum_{m=0}^{N-1} \widehat{x}(m) e^{2\pi i m n/N} = \frac{1}{N} \sum_{m=0}^{N-1} \widehat{x}(m) \sum_{n=0}^{b-1} e^{2\pi i m n/N}$$

$$\overset{(6.4)}{=} \frac{b}{N} \sum_{m=0}^{a-1} \widehat{x}(mb).$$

□

The Fourier inversion formula (6.2) shows that any x can be written as a linear combination of harmonics. While $|x(n)|^2$ quantifies the energy of the signal

x at time n, the Fourier coefficient $\widehat{x}(m)$ indicates that the harmonic $e^{2\pi i m(\cdot)/N}$ is contained in x with energy $\frac{1}{N}|\widehat{x}(m)|^2$. Indeed, setting $x = y$ in (6.3) implies conservation of energy, namely,

$$\sum_{n=0}^{N-1} \big|x(n)\big|^2 = \frac{1}{N} \sum_{m=0}^{N-1} \big|\widehat{x}(m)\big|^2, \quad x \in \mathbb{C}^N.$$

Mathematically speaking, Gabor analysis is centered on the interplay of the Fourier transform, translation operators, and modulation operators. The *cyclic shift operator* $T : \mathbb{C}^N \longrightarrow \mathbb{C}^N$ is given by

$$Tx = T\big(x(0), x(1), \ldots, x(N-1)\big)^T = \big(x(N-1), x(0), x(1), \ldots, x(N-2)\big)^T.$$

Translation T_k by $k \in \{0, 1, \ldots, N-1\}$ is given by

$$T_k x(n) = T^k x(n) = x(n-k), \quad n = 0, 1, \ldots, N-1,$$

that is, T_k simply repositions the entries of x, for instance, $x(0)$ is the k-th entry of $T_k x$. Note that the difference $n-k$ is taken modulo N, which agrees with considering the indices of $\mathbb{C}^N$ as elements of the cyclic group $\mathbb{Z}_N = \mathbb{Z}/N\mathbb{Z}$. In Sect. 6.4 we will consider Gabor frames for $\mathbb{C}^G$, that is, on the vector space where the components are indexed by a finite Abelian group G that is not necessarily cyclic.

Modulation operators $M_\ell : \mathbb{C}^N \longrightarrow \mathbb{C}^N$, $\ell = 0, 1, \ldots, N-1$, are given by

$$M_\ell x = \big(e^{2\pi i \ell 0/N} x(0), e^{2\pi i \ell 1/N} x(1), \ldots, e^{2\pi i \ell (N-1)/N} x(N-1)\big)^T, \quad x \in \mathbb{C}^N,$$

that is, the modulation operator M_ℓ simply performs a pointwise product of the input vector $x = x(\cdot)$ with the harmonic $e^{2\pi i \ell(\cdot)/N}$.

Translation operators are commonly referred to as *time-shift operators*. Moreover, modulation operators are *frequency-shift operators*. Indeed, we have

$$\begin{aligned}\widehat{M_\ell x}(m) = \mathscr{F} M_\ell x(m) &= \sum_{n=0}^{N-1} \big(e^{2\pi i \ell n/N} x(n)\big) e^{-2\pi i m n/N} = \sum_{n=0}^{N-1} x(n) e^{-2\pi i (m-\ell) n/N} \\ &= \widehat{x}(m-\ell).\end{aligned}$$

Applying the Fourier inversion formula to both sides gives

$$M_\ell = \mathscr{F}^{-1} T_\ell \mathscr{F}.$$

A *time-frequency shift operator* $\pi(k, \ell)$ combines translation by k and modulation by ℓ, that is,

$$\pi(k, \ell) : \mathbb{C}^N \longrightarrow \mathbb{C}^N, \qquad x \mapsto \pi(k, \ell) x = M_\ell T_k x.$$

For example, for $G = \mathbb{Z}_4$ the operators T_1, M_2, and $\pi(N-1, 3)$ are given by the matrices

$$\begin{pmatrix} 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & e^{2\pi i 3/4} & 0 & 0 \\ 0 & 0 & e^{2\pi i 2/4} & 0 \\ 0 & 0 & 0 & e^{2\pi i 1/4} \end{pmatrix},$$

$$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 0 & 0 & e^{2\pi i 3/4} & 0 \\ 0 & 0 & 0 & e^{2\pi i 2/4} \\ e^{2\pi i 1/4} & 0 & 0 & 0 \end{pmatrix}.$$

The following observation greatly simplifies Gabor analysis on $\mathbb{C}^N$. Recall that the space of linear operators on $\mathbb{C}^N$ forms an N^2-dimensional Hilbert space with *Hilbert–Schmidt space* inner product given independently of the chosen orthonormal basis $\{e_n\}_{n=0,1,\dots,N-1}$ by

$$\langle A, B\rangle_{\mathrm{HS}} = \sum_{\widetilde{n}=0}^{N-1} \sum_{n=0}^{N-1} \langle Ae_n, e_{\widetilde{n}}\rangle \overline{\langle Be_n, e_{\widetilde{n}}\rangle}.$$

Proposition 6.1 *The set of normalized time-frequency shift operators* $\{1/\sqrt{N}\ \pi(k,\ell)\}_{k,\ell=0,1,\dots,N-1}$ *is an orthonormal basis for the Hilbert–Schmidt space of linear operators on* $\mathbb{C}^N$.

Proof Consider $A = (a_{\widetilde{n}n})$ and $B = (b_{\widetilde{n}n})$ as matrices with respect to the Euclidean basis. We have

$$\big\langle (a_{\widetilde{n}n}), (b_{\widetilde{n}n})\big\rangle_{\mathrm{HS}} = \sum_{\widetilde{n}=0}^{N-1} \sum_{n=0}^{N-1} a_{\widetilde{n}n} \overline{b_{\widetilde{n}n}}.$$

Clearly, $\langle \pi(k,\ell), \pi(\widetilde{k}, \widetilde{\ell})\rangle_{\mathrm{HS}} = 0$ if $k \neq \widetilde{k}$ as the matrices $\pi(k,\ell)$ and $\pi(\widetilde{k}, \widetilde{\ell})$ then have disjoint support. Moreover, Theorem 6.1 implies that

$$\begin{aligned} \big\langle 1/\sqrt{N}\ \pi(k,\ell),\, 1/\sqrt{N}\ \pi(k,\widetilde{\ell})\big\rangle_{\mathrm{HS}} &= \big\langle 1/\sqrt{N}\ e^{2\pi i \ell(\cdot)/N},\ 1/\sqrt{N}\ e^{2\pi i \widetilde{\ell}(\cdot)/N}\big\rangle \\ &= \delta(\ell - \widetilde{\ell}). \end{aligned}$$

□

We now define Gabor systems on $\mathbb{C}^N$. For $\varphi \in \mathbb{C}^N \setminus \{0\}$ and $\Lambda \subseteq \{0, 1, \dots, N-1\} \times \{0, 1, \dots, N-1\}$ we call

$$(\varphi, \Lambda) = \big\{\pi(k,\ell)\varphi\big\}_{(k,\ell)\in\Lambda}$$

the *Gabor system* generated by the *window function* φ and the set Λ. A Gabor system which spans $\mathbb{C}^N$ is a frame and is referred to as a *Gabor frame*.

For instance, the Gabor system $((1,2,3,4)^T, \{0,1,2,3\} \times \{0,1,2,3\})$ in $\mathbb{C}^4$ consists of the columns in the matrix

$$\left(\begin{array}{cccc|cccc|cccc|cccc} 1 & 1 & 1 & 1 & 4 & 4 & 4 & 4 & 3 & 3 & 3 & 3 & 2 & 2 & 2 & 2 \\ 2 & 2i & -2 & -2i & 1 & i & -1 & -i & 4 & 4i & -4 & -4i & 3 & 3i & -3 & -3i \\ 3 & -3 & 3 & -3 & 2 & -2 & 2 & -2 & 1 & -1 & 1 & -1 & 4 & -4 & 4 & -4 \\ 4 & -4i & -4 & 4i & 3 & -3i & -3 & 3i & 2 & -2i & -2 & 2i & 1 & -i & -1 & i \end{array}\right),$$

while the elements of $((1,2,3,4,5,6)^T, \{0,2,4\} \times \{0,3\})$ are listed in

$$\left(\begin{array}{cc|cc|cc} 1 & 1 & 5 & 5 & 3 & 3 \\ 2 & 2i & 6 & 6i & 4 & 4i \\ 3 & 3 & 1 & 1 & 5 & 5 \\ 4 & 4i & 2 & 2i & 6 & 6i \\ 5 & 5 & 3 & 3 & 1 & 1 \\ 6 & 6i & 4 & 4i & 2 & 2i \end{array}\right).$$

The *short-time Fourier transform* $V_\varphi : \mathbb{C}^N \longrightarrow \mathbb{C}^{N\times N}$ with respect to the window $\varphi \in \mathbb{C}^N \backslash \{0\}$ is given by

$$V_\varphi x(k,\ell) = \big\langle x, \pi(k,\ell)\varphi \big\rangle = \mathscr{F}(xT_k\overline{\varphi})(\ell) = \sum_{n=0}^{N-1} x(n)\overline{\varphi(n-k)}e^{-2\pi i \ell n/N},$$

$$x \in \mathbb{C}^N,$$

[34, 35, 46, 47]. Observe that $V_\varphi x(k,\ell) = \mathscr{F}(xT_k\varphi)(\ell)$ indicates that the short-time Fourier transform on $\mathbb{C}^N$ can be efficiently computed using an FFT. This representation also indicates why short-time Fourier transforms are commonly referred to as *windowed Fourier transforms*: a window function φ centered at 0 is translated by k, the pointwise product with x selects a portion of x centered at k, and this portion is analyzed using a (fast) Fourier transform.

The short-time Fourier transform treats time and frequency almost symmetrically. In fact, using Parseval–Plancherel we obtain

$$\begin{aligned} V_\varphi x(k,\ell) &= \big\langle x, \pi(k,\ell)\varphi \big\rangle = \langle \widehat{x}, \widehat{M_\ell T_k \varphi} \rangle = \langle \widehat{x}, T_\ell M_{-k} \widehat{\varphi} \rangle \\ &= e^{-2\pi i k\ell/N} \langle \widehat{x}, M_{-k} T_\ell \widehat{\varphi} \rangle = e^{-2\pi i k\ell/N} V_{\widehat{\varphi}} \widehat{x}(\ell, -k), \quad x \in \mathbb{C}^N. \end{aligned} \tag{6.6}$$

While the short-time Fourier transform plays a distinct role in Gabor analysis on the real line—it is defined on $\mathbb{R} \times \widehat{\mathbb{R}}$ while Gabor frames are indexed by discrete subgroups of $\mathbb{R} \times \widehat{\mathbb{R}}$—in the finite-dimensional setting, the short-time Fourier transform reduces to the analysis map with respect to the *full Gabor system* $(\varphi, \{0,1,\dots,N-1\} \times \{0,1,\dots,N-1\})$, that is, a Gabor system with $\Lambda = \{0,1,\dots,N-1\} \times \{0,1,\dots,N-1\}$. Hence, the inversion formula for the short-time

Fourier transform

$$x(n) = \frac{1}{N\|\varphi\|_2^2} \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} V_\varphi x(k,\ell)\varphi(n-k)e^{-2\pi i \ell n/N}$$

$$= \frac{1}{N\|\varphi\|_2^2} \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \big\langle x, \pi(k,\ell)\varphi\big\rangle\, \pi(k,\ell)\varphi(n), \quad x \in \mathbb{C}^N, \tag{6.7}$$

simply states that for all $\varphi \neq 0$, the system $(\varphi, \{0,1,\ldots,N-1\} \times \{0,1,\ldots,N-1\})$ is an $N\|\varphi\|^2$-tight Gabor frame. Equation (6.7) is a trivial consequence of Corollary 6.2 below. It characterizes tight Gabor frames (φ, Λ) for the case that summation over $\{0,1,\ldots,N-1\} \times \{0,1,\ldots,N-1\}$ in (6.7) is replaced by summation over a subgroup Λ of $\mathbb{Z}_N \times \mathbb{Z}_N = \{0,1,\ldots,N-1\} \times \{0,1,\ldots,N-1\}$.

Not all Gabor frames are tight, meaning that the dual frame of a frame (φ, Λ) is not necessarily (φ, Λ). The following outstanding property of Gabor frames ensures that the canonical dual frame of a Gabor frame is again a Gabor frame. A similar property does not hold for other similarly structured frames; for example, canonical dual frames of wavelet frames are in general not wavelet frames.

Proposition 6.2 *The canonical dual frame of a Gabor frame (φ, Λ) with frame operator S is the Gabor frame $(S^{-1}\varphi, \Lambda)$.*

Proof We will show that $\pi(k,\ell) \circ S = S \circ \pi(k,\ell)$ for all $(k,\ell) \in \Lambda$. Then, $S^{-1} \circ \pi(k,\ell) = \pi(k,\ell) \circ S^{-1}$ and the members of the dual frame of (φ, Λ) are of the form $s^{-1}(\pi(k,\ell)\varphi) = \pi(k,\ell)(S^{-1}\varphi)$, $(k,\ell) \in \Lambda$.

The result is stated and proven in greater generality in Proposition 6.5 below. For simplicity we consider here only the case $\Lambda = \{0, a, 2a, \ldots, N-a\} \times \{0, b, 2b, \ldots, N-b\}$ where a and b divide N.

The following elementary computation completes the proof.

$$S \circ \pi(k,\ell)x(n) = \sum_{\widetilde{k}=0}^{N/a-1} \sum_{\widetilde{\ell}=0}^{N/b-1} \big\langle \pi(k,\ell)x, \pi(\widetilde{k},\widetilde{\ell})\varphi\big\rangle \pi(\widetilde{k},\widetilde{\ell})\varphi$$

$$= \sum_{\widetilde{k}=0}^{N/a-1} \sum_{\widetilde{\ell}=0}^{N/b-1} \sum_{\widetilde{n}=0}^{N-1} e^{2\pi i \ell b\widetilde{n}/N} x(\widetilde{n}-ka) e^{-2\pi i \widetilde{\ell} b\widetilde{n}/N} \overline{\varphi(\widetilde{n}-\widetilde{k}a)}$$

$$\times\, e^{-2\pi i \widetilde{\ell} bn/N} \varphi(n-\widetilde{k}a)$$

$$= \sum_{\widetilde{k}=0}^{N/a-1} \sum_{\widetilde{\ell}=0}^{N/b-1} \sum_{\widetilde{n}=0}^{N-1} x(\widetilde{n}) e^{-2\pi i (\widetilde{\ell}-\ell) b(\widetilde{n}+ka)/N} \overline{\varphi\big(\widetilde{n}-(\widetilde{k}-k)a\big)}$$

$$\times\, e^{-2\pi i \widetilde{\ell} bn/N} \varphi(n-\widetilde{k}a)$$

$$
\begin{aligned}
&= \sum_{\widetilde{k}=0}^{N/a-1} \sum_{\widetilde{\ell}=0}^{N/b-1} \sum_{\widetilde{n}=0}^{N-1} x(\widetilde{n}) e^{-2\pi i \widetilde{\ell} b \widetilde{n}/N} \overline{\varphi(\widetilde{n}-\widetilde{k}a)}\, e^{-2\pi i(\widetilde{\ell}+\ell)bn/N} \\
&\quad \varphi\big(n-(\widetilde{k}+k)a\big) \times e^{2\pi i \ell b k a/N} \\
&= \sum_{\widetilde{k}=0}^{N/a-1} \sum_{\widetilde{\ell}=0}^{N/b-1} \big\langle x, \pi(a\widetilde{k}, b\widetilde{\ell})\varphi\big\rangle \pi(ak, b\ell)\pi(a\widetilde{k}, b\widetilde{\ell})\varphi \\
&= \pi(ak, b\ell) \circ Sx(n). \qquad \square
\end{aligned}
$$

6.3 Gabor Frames as a Time-Frequency Analysis Tool

As discussed in Sect. 6.1, Gabor systems were introduced to efficiently utilize communication channels. In this section, we will focus on a second fundamental application of Gabor systems; it concerns the time-frequency analysis of signals that are dominated by few components that are concentrated in time and/or frequency.

The Fourier transform's ability to separate a signal into its frequency components provides a powerful tool in science and mathematics. Many signals, however—for example, speech and music—have frequency contributions which appear only during short time intervals. The Fourier transform of a piano sonata may provide information on which notes dominate the score, but it falls short of enabling us to write down the score of the sonata that is needed to reproduce it on a piano. Gabor analysis addresses this shortcoming by providing information on which frequencies appear in a signal at which times.

Recall that $(\varphi, \{0, 1, \ldots, N-1\} \times \{0, 1, \ldots, N-1\})$ is an $N\|\varphi\|^2$-tight Gabor frame. Assuming $\|\varphi\|^2 = 1/N$, we obtain

$$
\sum_{n=0}^{N-1} \big|x(n)\big|^2 = \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \big|V_\varphi x(k,\ell)\big|^2 = \sum_{k=0}^{N-1} \sum_{\ell=0}^{N-1} \big|\mathscr{F}(x T_k \overline{\varphi})(\ell)\big|^2, \quad x \in \mathbb{C}^N,
$$

that is, the short-time Fourier transform V_φ distributes the energy of x on the time-frequency grid $\{0, 1, \ldots, N-1\} \times \{0, 1, \ldots, N-1\}$. Equation (6.6) implies that

$$
\big|V_\varphi x(k,\ell)\big| = \big|\langle x, M_\ell T_k \varphi\rangle\big| = \big|\langle \widehat{x}, M_{-k} T_\ell \widehat{\varphi}\rangle\big| \le \min\big\{\big\langle |x|, T_k|\varphi|\big\rangle, \big\langle |\widehat{x}|, T_\ell|\widehat{\varphi}|\big\rangle\big\}.
$$

Hence, any φ with φ and $\widehat{\varphi}$ being well localized at 0, meaning $|\varphi(n)|$, $|\widehat{\varphi}(m)|$ are small for n, m and $N-n, N-m$ large, implies that the energy captured in the *spectrogram* value $SPEC_\varphi(k,\ell) = |V_\varphi x(k,\ell)|^2$ is only large if frequencies close to ℓ have a large presence in x around time k. Unfortunately, *Heisenberg's uncertainty principle* implies that φ and $\widehat{\varphi}$ cannot be simultaneously arbitrarily well localized at 0. The simplest realization of this principle is the following result attributed to Donoho and Stark [29, 69]. In the following, we set $\|x\|_0 = |\{n :\ x(n) \neq 0\}|$.

Proposition 6.3 *Let $x \in \mathbb{C}^N \backslash \{0\}$; then $\|x\|_0 \cdot \|\widehat{x}\|_0 \ge N$.*

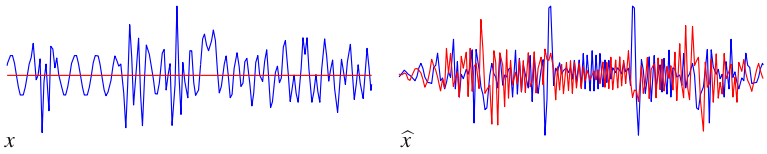


Fig. 6.1 The test signal x given in (6.8) and used in Figs. 6.2–6.6 and Fig. 6.9 as well as its Fourier transform. Here and in the following, the real part of a signal is given in *blue*, and its imaginary part is given in *red*

Proof For $x \in \mathbb{C}^N$, $x \neq 0$, and $A = \max\{|\widehat{x}(m)|,\ m = 0, 1, \ldots, N-1\} \neq 0$, we compute

$$NA^2 \leq N\left(\sum_{n=0}^{N-1} |x(n)|\right)^2 \leq N\|x\|_0 \sum_{n=0}^{N-1} |x(n)|^2 = \|x\|_0 \sum_{m=0}^{N-1} |\widehat{x}(m)|^2 \leq \|x\|_0 \|\widehat{x}\|_0 A^2. \qquad \square$$

Theorem 6.12 below strengthens Proposition 6.3 in the case that N is prime.

To illustrate the use of Gabor frames in time-frequency analysis, we will use various Gabor windows to analyze the multicomponent signal $x \in \mathbb{C}^{200}$ given by

$$\begin{aligned} x(n) = {} & \chi_{\{0,\ldots,49\}}(n) \sin(2\pi 20n/200) + \chi_{\{150,\ldots,199\}}(n) \sin\big(2\pi 50(n-150)/200\big) \\ & + \chi_{\{50,\ldots,149\}}(n) \sin\big(2\pi \big(30(n-50)^2/200^2 + 20(n-50)/200\big)\big) \\ & + 1.2\chi_{\{80,\ldots,99\}}(n)(1 + \cos\big(2\pi(10n/200 - 1/2)\big) \cos(2\pi 60n/200) \\ & + 1.2\chi_{\{60,\ldots,79\}}(n)(1 + \cos\big(2\pi(10n/200 - 1/2)\big) \cos(2\pi 50n/200) \\ & + .5\chi_{\{100,\ldots,199\}}(n)(1 + \cos\big(2\pi(2n/200 - 1/2)\big) \cos(2\pi 20n/200) \\ & + \chi_{\{20,\ldots,31\}}(n)(1 + \cos\big(2\pi(12n/200 - 1/2)\big) \cos(2\pi 20n/200) \\ & + 1.1\chi_{\{100,\ldots,109\}}(n)(1 + \cos\big(2\pi(20n/200 - 1/2)\big), \quad n = 0, 1, \ldots, 199, \end{aligned} \tag{6.8}$$

where $\chi_A(n) = 1$ if $n \in A$ and 0 otherwise. The signal and its Fourier transform are displayed in Fig. 6.1. Note that x is real-valued, so its Fourier transform has even symmetry. As we will also use real-valued window functions below, we obtain short-time Fourier transforms which are symmetric in frequency and it suffices to display $SPEC_\varphi$ in Figs. 6.2–6.9 only for frequencies 0 to 100.[2]

In Figs. 6.2 and 6.3, we use orthogonal Gabor systems generated by characteristic functions. In Fig. 6.2 we choose as Gabor window the normalized characteristic function given by $\varphi(n) = 1/\sqrt{20}$ for $n = 191, 192, \ldots, 199, 0, 1, \ldots, 10$ and $\varphi(n) = 0$ for $n = 11, 12, \ldots, 190$. The spectrogram $SPEC_\varphi x = |V_\varphi x|^2$ in Fig. 6.2

[2]Our treatment is unit-free. The reader may assume that n counts seconds, then m counts hertz, or that n represents milliseconds, in which case m represents megahertz.

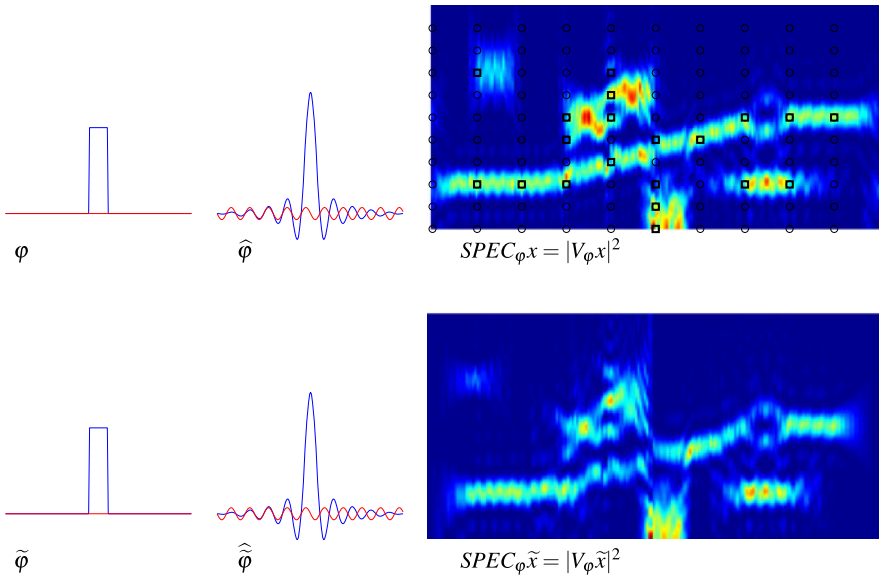


Fig. 6.2 Gabor frame analysis of the multicomponent signal (6.8) displayed in Fig. 6.1. We use the Gabor system (φ, Λ) with $\varphi(n) = 1/\sqrt{20}$ for $n = 191, 192, \ldots, 199, 0, 1, \ldots, 10$ and $\varphi(n) = 0$ for $n = 11, 12, \ldots, 190$. The Gabor system forms an orthonormal basis of $\mathbb{C}^{200}$ and is therefore self-dual; that is $\varphi = \widetilde{\varphi}$. We display φ, $\widehat{\varphi}$, $\widetilde{\varphi}$, $\widehat{\widetilde{\varphi}}$ as well as the spectrogram of x and of its approximation $\widetilde{x}$. The *circles* on $SPEC_\varphi x$ depict Λ; they mark frame coefficients of the frame (φ, Λ). The *squares* denote the 20 biggest frame coefficients, which are then used to construct the approximation $\widetilde{x}$ to x

shows that the signal has as dominating frequency 20 in the beginning and frequency 50 toward the end, with a linear transition in between. In addition, the five additional frequency clusters of x appear at five different time instances.

The picture shows some vertical ringing artifacts. These are due to the sidelobes of the Fourier transform $\widehat{\varphi}$ of φ. They imply that components well localized in frequency have an effect on $|V_\varphi x(k, \ell)|^2$ for a large range of ℓ.

The values of the short-time Fourier transform $V_\varphi x$ allow us to reconstruct x using (6.7). Doing so requires the use of N^2 coefficients to reconstruct a signal in $\mathbb{C}^N$. Clearly, it is more efficient to use only the values of $V_\varphi x$ on a lattice Λ that allows for (φ, Λ) being a frame of cardinality not exceeding the dimension of the ambient space N.

In Fig. 6.2, we circle the values of $|V_\varphi x(k, \ell)|^2$ with $(k, \ell) \in \Lambda = \{0, 20, \ldots, 180\} \times \{0, 10, \ldots, 190\}$. It is easy to see that (φ, Λ) is an orthonormal basis; hence, we can reconstruct the signal x using only values of the short-time Fourier transform that correspond to the circled values. Note that, in general, whenever (φ, Λ) is a frame with dual frame $(\widetilde{\varphi}, \Lambda)$, we can reconstruct x by means of

$$x = \sum_{(k,\ell)\in\Lambda} \langle x, \pi(k, \ell)\varphi\rangle \pi(k, \ell)\widetilde{\varphi}.$$

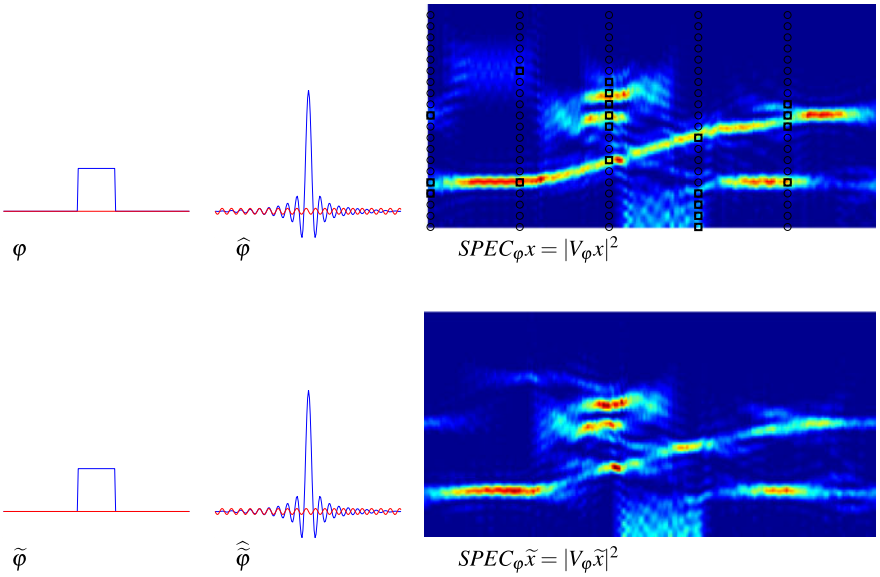


Fig. 6.3 Gabor frame analysis of the multicomponent signal displayed in Fig. 6.1. We use the orthonormal Gabor system (φ, Λ) with $\varphi(n) = 1/\sqrt{40}$ for $n = 181, 192, \ldots, 199, 0, 1, \ldots, 20$ and $\varphi(n) = 0$ for $n = 21, 12, \ldots, 180$. We display φ, $\widehat{\varphi}$, $\widetilde{\varphi}$, $\widehat{\widetilde{\varphi}}$, $SPEC_\varphi x$, and $SPEC_\varphi \widetilde{x}$. The circles on $SPEC_\varphi x$ mark frame coefficients of the frame (φ, Λ); the squares denote the 20 coefficients used to construct $\widetilde{x}$

However, in many applications, one would like to reduce the amount of information that is first stored and then used to reproduce the signal to below the dimension N of the ambient space. Rather than reproducing x perfectly, we are satisfied to obtain an approximation

$$\widetilde{x} = \sum_{(k,\ell)\in\Lambda} R\big(\langle x, \pi(k,\ell)\varphi\rangle\big)\pi(k,\ell)\widetilde{\varphi},$$

which captures the key features of x.

Here, we illustrate the effect of a rather simplistic compression algorithm. Namely, we use only the 40 largest coefficients (20 in the depicted half of the spectrogram) to produce an approximation $\widetilde{x}$ to x. That is, $R(\langle x, \pi(k,\ell)\varphi\rangle) = \langle x, \pi(k,\ell)\varphi\rangle$ for the 40 largest coefficients and $R(\langle x, \pi(k,\ell)\varphi\rangle) = 0$ otherwise. The locations in time and frequency of the chosen coefficients are marked by squares.

Graphic comparisons of $\widetilde{x}$ with x and of $\widehat{\widetilde{x}}$ with $\widehat{x}$ are not very illuminating. Instead, we compare the spectrogram of $\widetilde{x}$ with the spectrogram of the original signal x. This demonstrates well the effect of our compression procedure; most of the features of x are in fact preserved.

The setup chosen to generate Fig. 6.3 differs from the one used to obtain Fig. 6.2 only in the choice of window function φ. Here, we choose a wider window function; this leads to a better localized $\widehat{\varphi}$. Specifically, we choose $\varphi(n) = 1/\sqrt{40}$ for $n =$

$181, 192, \ldots, 199, 0, 1, \ldots, 20$ and $\varphi(n) = 0$ for $n = 21, 22, \ldots, 180$. As a lattice we choose $\Lambda = \{0, 40, 80, \ldots, 160\} \times \{0, 5, 10, \ldots, 195\}$ and observe that (φ, Λ) is again an orthonormal basis.

Comparing the spectrogram of x in Fig. 6.3 with the one of x in Fig. 6.2, we observe a reduced ringing effect and slightly better localization in frequency at the price of losing localization in time. Unfortunately, a comparison of $SPEC_\varphi x$ with $SPEC_\varphi \widetilde{x}$ shows that the canonical choice of lattice does not seem to work well in conjunction with our compression algorithm. The large gaps between lattice notes in time cause part of the frequency transition not to be preserved by our simplistic compression algorithm.

In Figs. 6.4–6.6 we choose as window functions Gaussians. In Fig. 6.4 we choose

$$\varphi(n) = ce^{-(n/6)^2}$$

where c normalizes φ and as lattice $\Lambda = \{0, 8, 16, \ldots, 192\} \times \{0, 20, 40, \ldots, 180\}$. For Fig. 6.5 we select

$$\varphi(n) = ce^{-(n/14)^2}$$

where c again normalizes φ. We let $\Lambda = \{0, 20, 40, \ldots, 180\} \times \{0, 8, 16, \ldots, 192\}$. We perform the same naive compression procedure used above to obtain Figs. 6.2 and 6.3. Note that the lattices in Figs. 6.5 and 6.6 contain 250 elements, and in fact, the Gabor frame (φ, Λ) is overcomplete.

Choosing a Gaussian window function has the benefit of removing the sidelobes and of providing an easily readable spectrogram. But our compression procedure is harmed by two facts. First of all, we are now picking 40 out of 250 coefficients; these are clustered in the dominating area, so secondary time-frequency components of x are also overlooked. Clearly, our algorithm does not benefit from the redundancy of the Gabor frame in use. Second, the good localization in frequency of φ implies that some of the components fall between lattice values. Therefore, they are overlooked.

A comparison of Figs. 6.4 and 6.5 again shows the tradeoff between good time and good frequency resolution.

In Fig. 6.6 we choose the same Gaussian window as in Fig. 6.4, but we choose a lattice which is not the product of two lattices in $\{0, 1, \ldots, 199\}$. In fact, we have

$$\begin{aligned}\Lambda = 7\{0, 40, \ldots, 160\} &\times \{0, 8, \ldots, 192\} \cup \{20, 60, 100, 140, 180\} \\ &\times \{4, 12, 20, \ldots, 196\}.\end{aligned}$$

But deviating from rectangular lattices offers little help. Moreover, even though we are choosing a lattice of the same redundancy, namely, we choose a frame with 250 elements in a 200-dimensional space, the dual window has poor frequency localization. This significantly reduces the quality of reconstruction when using the compressed version $\widetilde{x}$ of the signal x, as the dual window used for synthesis smears out the frequency signature of the signal.

Similar discussions on the use of Gabor frames to analyze discrete one-dimensional signals and discrete images can be found in [22, 52, 68, 72, 89, 90].

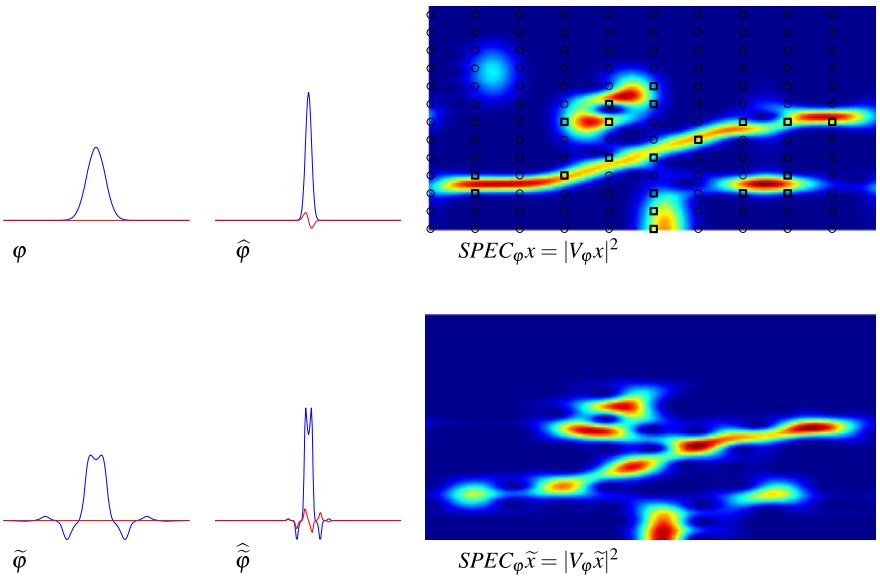


Fig. 6.4 Gabor frame analysis of the signal in Fig. 6.1. As the Gabor window we choose a normalized version of the Gaussian $\varphi(n) = ce^{-(n/6)^2}$, $n = 0, 1, \ldots, 199$. We display again φ, $\widehat{\varphi}$, $\widetilde{\varphi}$, $\widehat{\widetilde{\varphi}}$, $SPEC_\varphi x$, and $SPEC_\varphi \widetilde{x}$, where Λ is marked on $SPEC_\varphi x$ by circles. As before, the squares denote the 20 largest coefficients. Unmarked frame coefficients are not used to construct $\widetilde{x}$

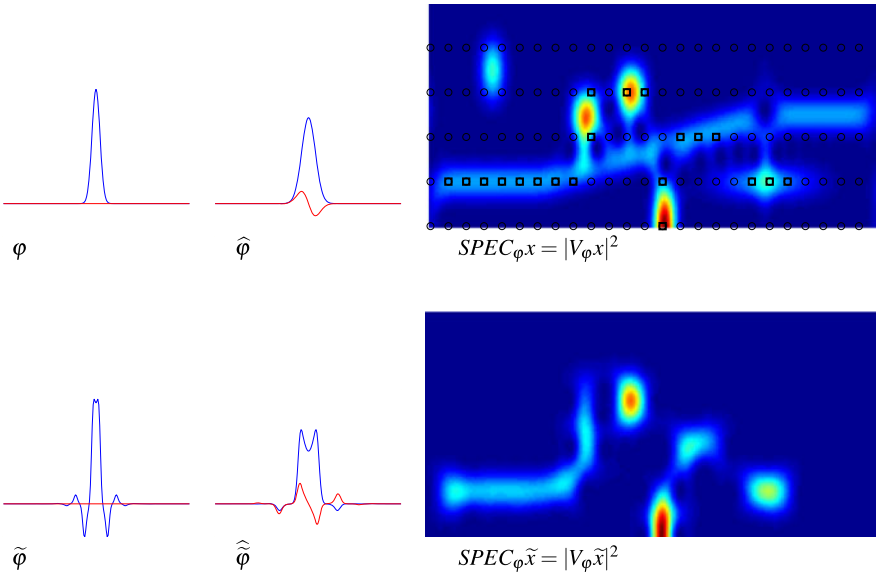


Fig. 6.5 Here, we use as the Gabor window a normalized version of $\varphi(n) = ce^{-(n/14)^2}$, $n = 0, 1, \ldots, 199$. As before, φ, $\widehat{\varphi}$, $\widetilde{\varphi}$, $\widehat{\widetilde{\varphi}}$, $SPEC_\varphi x$, and $SPEC_\varphi \widetilde{x}$ are shown, and Λ as well as the 20 largest coefficients used to construct $\widetilde{x}$ are marked on $SPEC_\varphi x$

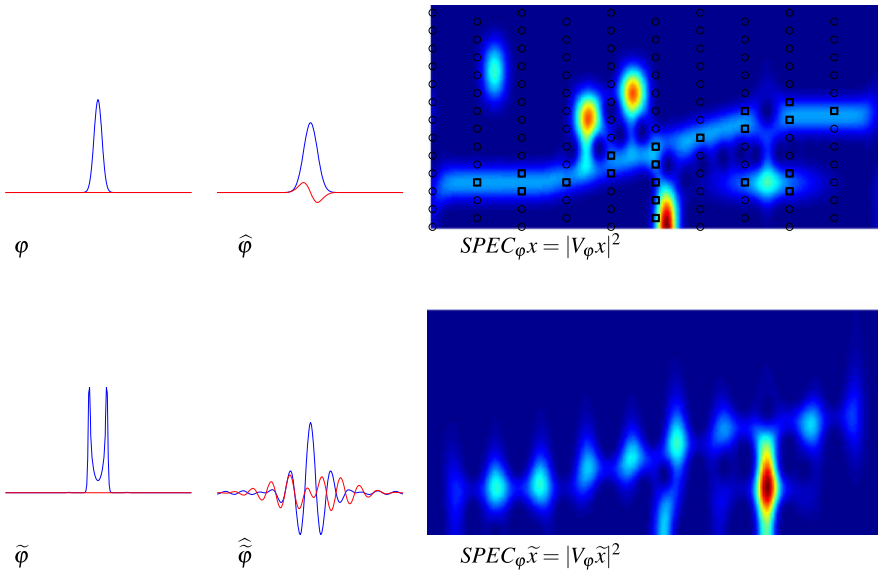


Fig. 6.6 We use the same window function as in Fig. 6.4, but a different lattice. This changes the displayed dual window $\widetilde{\varphi}$ and its Fourier transform $\widehat{\widetilde{\varphi}}$. $SPEC_{\varphi}x$ and $SPEC_{\varphi}\widetilde{x}$ and therefore x and $\widetilde{x}$ vary greatly. The lattice Λ and its 20 largest coefficients are marked as in Figs. 6.2–6.5 above

6.4 Gabor Analysis on Finite Abelian Groups

In Sect. 6.2 we defined Gabor systems in $\mathbb{C}^N$. Implicitly we considered vectors in $\mathbb{C}^N$ as vectors defined on the *cyclic group* $\mathbb{Z}_N = \mathbb{Z}/N\mathbb{Z}$. For example, the translation operator T_k was defined by $T_k x(n) = x(n-k)$ where $n-k$ was taken modulus N; that is, n and k were considered to be elements in the cyclic group $\mathbb{Z}_N$.

In this section, we will develop Gabor systems with an arbitrary finite Abelian group G in place of $\mathbb{Z}_N$. We thereby obtain results on Gabor systems on the finite-dimensional vector space

$$\mathbb{C}^G = \{x : G \longrightarrow \mathbb{C}\},$$

that is, $\mathbb{C}^G$ is a $|G|$-dimensional vector space with vector entries indexed by elements in the group G. We will continue to write $\mathbb{C}^N$ rather than $\mathbb{C}^{\mathbb{Z}_N}$ if $G = \mathbb{Z}_N$.

The group structure of the index set G allows us to define unitary *translation operators* $T_k : \mathbb{C}^G \longrightarrow \mathbb{C}^G$, $k \in G$, by

$$T_k x(n) = x(n-k), \quad n \in G.$$

Modulation operators on $\mathbb{C}^G$ are pointwise products with characters on the finite Abelian group G. A *character* $\xi \in \mathbb{C}^G$ is a group homomorphism mapping G into the multiplicative group $S^1 = \{z \in \mathbb{C} : \ |z| = 1\}$ [7, 57, 84, 94]. The set of characters on G forms a group under pointwise multiplication. This group is called the *dual group* of G and is denoted by $\widehat{G}$.

In summary, for $\xi \in \widehat{G}$, the *modulation operator* $M_\xi : \mathbb{C}^G \longrightarrow \mathbb{C}^G$ is given by

$$M_\xi x(n) = \xi(n)x(n), \quad n \in G.$$

For $\lambda = (k, \xi) \in G \times \widehat{G}$, we define the *time-frequency shift operator* $\pi(\lambda)$ by

$$\pi(\lambda) : \mathbb{C}^G \longrightarrow \mathbb{C}^G, \qquad x \mapsto \pi(\lambda)x = \pi(k, \xi)x = M_\xi T_k x = \xi(\cdot)x(\cdot - k).$$

We are now in position to define Gabor systems on $\mathbb{C}^G$ where G is a finite Abelian group with dual group $\widehat{G}$. Let Λ be a subset of the product group $G \times \widehat{G}$ and let $\varphi \in \mathbb{C}^G \setminus \{0\}$. The respective *Gabor system* is then given by

$$(\varphi, \Lambda) = \left\{\pi(\lambda)\varphi\right\}_{\lambda \in \Lambda}.$$

A Gabor system which spans $\mathbb{C}^G$ is a frame and is called a *Gabor frame*. In many cases, we will consider Gabor systems with Λ being a subgroup of $G \times \widehat{G}$.

The *short-time Fourier transform* $V_\varphi : \mathbb{C}^G \longrightarrow \mathbb{C}^{G \times \widehat{G}}$ with respect to the window $\varphi \in \mathbb{C}^G$ is given by

$$V_\varphi x(k, \xi) = \left\langle x, \pi(k, \xi)\varphi \right\rangle = \mathscr{F}(x T_k \overline{\varphi})(\xi) = \sum_{n \in G} x(n)\overline{\varphi(n-k)}\overline{\langle \xi, x \rangle}, \quad x \in \mathbb{C}^G,$$

where $\mathscr{F}$ is defined below [34, 35, 46, 47]. The inversion formula for the short-time Fourier transform

$$x(n) = \frac{1}{|G|\|\varphi\|_2^2} \sum_{(k,\xi) \in G \times \widehat{G}} V_\varphi x(k, \xi)\varphi(n-k)\langle \xi, k \rangle, \quad x \in \mathbb{C}^G,$$

holds for all $\varphi \neq 0$, as we will see in Corollary 6.2 below. As in the case $G = \mathbb{Z}_N$, we conclude that the system $(\varphi, G \times \widehat{G})$ is a $|G| \cdot \|\varphi\|^2$-tight Gabor frame.

Before continuing our discussion of Gabor systems on finite Abelian groups in Sect. 6.4.2, we will prove the harmonic analysis results that lie at the basis of Gabor analysis on finite Abelian groups.

6.4.1 Harmonic Analysis on Finite Abelian Groups

As mentioned above, a character on a finite Abelian group is a group homomorphism mapping G into the multiplicative circle group $S^1 = \{z \in \mathbb{C},\ |z| = 1\}$. The set of characters is denoted by $\widehat{G}$, which is a finite Abelian group under pointwise multiplication, meaning with composition $(\xi_1 + \xi_2)(n) = \xi_1(n)\xi_2(n)$.

In order to explicitly describe characters on finite Abelian groups, we will combine simple results on characters on cyclic groups with the *fundamental theorem of finite Abelian groups*. It states that every finite Abelian group is isomorphic to the product of cyclic groups.

Theorem 6.2 *For every finite Abelian group G there exist $N_1, N_2, \ldots, N_d \in \mathbb{N}$ with*

$$G \cong \mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_d}. \tag{6.9}$$

The factorization and the number of factors in (6.9) are not unique, but there exist a unique set of primes $\{p_1, \ldots, p_d\}$ and a unique set of natural numbers $\{r_1, \ldots, r_d\}$ so that (6.9) holds with $N_1 = p_1^{r_1}$, $N_2 = p_2^{r_2}, \ldots, N_d = p_d^{r_d}$.

Proof For our purpose it is only relevant that a factorization as given in (6.9) exists. We will outline an inductive proof of this fact.

Recall that $|G|$ is called the order of the group G, $\langle n \rangle$ denotes the group generated by $n \in G$, and the order of $n \in G$ is $|\langle n \rangle|$.

If $|G| = 1$ then $G = \{0\}$ and the claim holds trivially. Suppose that all groups of order $|G| < N$ satisfy (6.9). Let now G be given with $|G| = N$. We need to distinguish two cases.

If $N = p^s$ with p prime, choose $n \in G$ with maximal order. If its order is $|G|$, then $G = \langle n \rangle$ and $G \cong \mathbb{Z}_N$. If its order is less than $|G|$, then a short sequence of algebraic arguments shows that there exists a subgroup H with $G \cong \langle n \rangle \times H$. We obtain (6.9) for G by applying the induction hypothesis to H.

If $N = rp^s$ with p prime, $r \geq 2$ relatively prime with p, and $s \geq 1$. Then

$$G \cong \{n\text{: the order of } n \text{ is a power of } p\} \times \{n\text{: the order of } n \text{ is not divisible by } p\}$$

can be shown to be a factorization of G into two subgroups of smaller order, and we can again apply the induction hypothesis. □

As mentioned above, representations of finite groups as products of cyclic groups are not unique; for example, we have $\mathbb{Z}_{KL}$ is isomorphic to $\mathbb{Z}_K \times \mathbb{Z}_L$ if (and only if) K and L are relatively prime.

Any group isomorphism induces a group isomorphism between the respective dual groups. Theorem 6.2 therefore implies that for our study of characters on general finite Abelian groups it suffices to study characters on products of cyclic groups. Hence, we may assume

$$G = \mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_d}$$

in the following.

Observe that for the cyclic group $G = \mathbb{Z}_N = \{0, 1, \ldots, N-1\}$, a character ξ is fully determined by $\xi(1)$. Since

$$1 = \xi(0) = \xi(N) = \xi(1 + \cdots + 1) = \xi(1)^N,$$

we have $\xi(1) \in \{e^{2\pi i m/N},\ m = 0, 1, \ldots, N-1\}$. We conclude that $\widehat{\mathbb{Z}_N}$ contains exactly N characters; they are

$$\xi_m = \left(e^{2\pi i m(\cdot)0/N}, e^{2\pi i m(\cdot)1/N}, e^{2\pi i m(\cdot)2/N}, \ldots, e^{2\pi i m(\cdot)(N-1)/N}\right)^T,$$
$$m = 0, 1, \ldots, N-1.$$

The modulation operators for cyclic groups that are defined abstractly here therefore coincide with the definition of modulation operators on $\mathbb{C}^N$ given in Sect. 6.2.

Observe that under pointwise multiplication, the group of characters $\widehat{\mathbb{Z}_N}$ is cyclic and has N elements, that is, $\widehat{\mathbb{Z}_N} \cong \mathbb{Z}_N$, a fact that we will use below.

For $G = \mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_d}$, observe that any character ξ on G induces a character on the component groups $\mathbb{Z}_{N_1}, \mathbb{Z}_{N_2}, \ldots, \mathbb{Z}_{N_d}$. Hence, we can associate to any character ξ on G an $m = (m_1, m_2, \ldots, m_d)$ with

$$\xi(e_r) = \xi\big((0, \ldots, 0, 1, 0, \ldots, 0)\big) = e^{2\pi i m_r/N_1}, \quad r = 1, \ldots, d.$$

Clearly, as ξ is a group homomorphism, it is fully described by m and we have

$$\begin{aligned} \xi(n_1, n_2, \ldots, n_d) &= \xi_{m_1}(n_2) \ldots \xi_{m_1}(n_d) \\ &= e^{2\pi i m_1 n_1/N_1} e^{2\pi i m_2 n_2/N_2} \ldots e^{2\pi i m_d n_d/N_d} \\ &= e^{2\pi i (m_1 n_1/N_1 + m_2 n_2/N_2 + \cdots + m_d n_d/N_d)}. \end{aligned} \tag{6.10}$$

For notational simplicity, we will identify ξ with the derived m and write

$$\langle m, n\rangle = \xi(n) = e^{2\pi i (m_1 n_1/N_1 + m_2 n_2/N_2 + \cdots + m_d n_d/N_d)}. \tag{6.11}$$

We observe that

$$\widehat{G} = (\mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_d})\widehat{\ } \cong \widehat{\mathbb{Z}_{N_1}} \times \widehat{\mathbb{Z}_{N_2}} \times \cdots \times \widehat{\mathbb{Z}_{N_d}}.$$

Clearly, then $\widehat{\widehat{G}} \cong \widehat{G} \cong G$; in addition, G can be canonically identified with $\widehat{\widehat{G}}$ by means of the group homomorphism $n : m \mapsto \langle m, n\rangle$, thereby justifying the duality notation used in (6.11).

In the finite Abelian group setting, the *Fourier transform* $\mathscr{F} : \mathbb{C}^G \longrightarrow \mathbb{C}^{\widehat{G}}$ is given by

$$\begin{aligned} \mathscr{F}x(m) = \widehat{x}(m) &= \sum_{n \in G} x(n) \langle m, n\rangle \\ &= \sum_{n_1=0}^{N_1-1} \sum_{n_2=0}^{N_2-1} \cdots \sum_{n_d=0}^{N_d-1} x(n_1, n_2, \ldots, n_d) \\ &\quad \times e^{-2\pi i (m_1 n_1/N_1 + m_2 n_2/N_2 + \cdots + m_d n_d/N_d)}, \\ m = (m_1, m_2, \ldots, m_d) &\in \widehat{G}. \end{aligned}$$

Theorem 6.1 above implies that the normalized characters on $\mathbb{Z}_N$ form an orthonormal basis of $\mathbb{C}^N$. Combining this with (6.10) shows that the normalized characters on any finite Abelian group G form an orthonormal system of cardinality $|G| = N_1 \cdots N_d = \dim \mathbb{C}^G$. We conclude that the normalized characters form an orthonormal basis of $\mathbb{C}^G$. This simple observation generalizes (6.2) and (6.3) to the

general finite Abelian group setting. For example, the *Fourier inversion formula* (6.2) becomes

$$\begin{aligned} x(n) &= \frac{1}{|G|} \sum_{m \in \widehat{G}} \widehat{x}(m) \overline{\langle m, n \rangle} \\ &= \frac{1}{|G|} \sum_{m_1=0}^{N_1-1} \sum_{m_2=0}^{N_2-1} \cdots \sum_{m_d=0}^{N_d-1} \widehat{x}(m_1, m_2, \ldots, m_d) \\ &\qquad \times e^{2\pi i (m_1 n_1/N_1 + m_2 n_2/N_2 + \cdots + m_d n_d/N_d)}, \\ &\qquad\qquad n = (n_1, n_2, \ldots, n_d) \in G. \end{aligned}$$

To state and prove the *Poisson summation formula* (6.13) for the Fourier transform on $\mathbb{C}^G$, we define for any subgroup H of G the *annihilator subgroup*

$$H^\perp = \big\{ m \in \widehat{G} : \ \langle m, n \rangle = 1 \text{ for all } n \in H \big\}.$$

Clearly, $H^\perp$ is a subgroup of $\widehat{G}$. In Gabor and harmonic analysis, discrete subgroups of G are commonly referred to as *lattices* and their annihilators as their *dual lattices*.

Theorem 6.3 *Let H be a subgroup (lattice) of G and let $H^\perp$ be its annihilator subgroup (dual lattice). Then*

$$\sum_{n \in H} \langle m, n \rangle = \begin{cases} |H|, & \text{if } m \in H^\perp, \\ 0, & \text{otherwise}, \end{cases} \qquad \sum_{m \in H^\perp} \langle m, n \rangle = \begin{cases} |H^\perp|, & \text{if } n \in H, \\ 0, & \text{otherwise}, \end{cases} \tag{6.12}$$

and

$$\big|H^\perp\big| \sum_{n \in H} x(n) = \sum_{m \in H^\perp} \widehat{x}(m), \quad x \in \mathbb{C}^G. \tag{6.13}$$

Proof Let $m \in \widehat{G}$. Then $n \mapsto \langle m, n \rangle$ for $n \in H$ defines a character on H. This character is identical or orthogonal to the trivial character on H, namely, $0 : n \mapsto 1$ for $n \in H$, and hence

$$\sum_{n \in H} \langle m, n \rangle = \sum_{n \in H} \langle m, n \rangle \overline{\langle 0, n \rangle} = \begin{cases} |H|, & \text{if } m = 0 \text{ on } H, \\ 0, & \text{otherwise}, \end{cases} = \begin{cases} |H|, & \text{if } m \in H^\perp, \\ 0, & \text{otherwise}. \end{cases}$$

The second equality in (6.12) follows from the first equality in (6.12) by observing that $H^\perp$ is a subgroup of $\widehat{G}$ and that $(H^\perp)^\perp \subseteq \widehat{\widehat{G}}$ can be canonically identified with $H \subseteq G$.

The interchange of summation argument used to obtain (6.5) in Theorem 6.1 can be used again to prove (6.13). □

The fact that $G \cong \mathbb{Z}_{N_1} \times \mathbb{Z}_{N_2} \times \cdots \times \mathbb{Z}_{N_d}$ for any finite Abelian group G implies that the discrete Fourier matrix W_G can be expressed as the Kronecker product of

the Fourier matrices for the cyclic groups $\mathbb{Z}_{N_1}, \mathbb{Z}_{N_2}, \ldots, \mathbb{Z}_{N_d}$, that is, $W_G = W_{N_1} \otimes W_{N_2} \otimes \cdots \otimes W_{N_d}$. For example, we have

$$W_{\mathbb{Z}_2 \times \mathbb{Z}_2} = W_{\mathbb{Z}_2} \otimes W_{\mathbb{Z}_2} = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 1 & 1 \\ 1 & -1 & 1 & -1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}.$$

6.4.2 Examples of and Further Remarks on Gabor Systems on Finite Abelian Groups

In Sect. 6.4.1 it was shown that the study of finite Abelian groups coincides with the study of finite products of cyclic groups. Moreover, we described in detail characters on products of cyclic groups and thereby modulation operators acting on functions on such groups.

For example, for $G = \mathbb{Z}_2 \times \mathbb{Z}_2$, the operators $T_{(1,0)}$, and $M_{(1,1)}$ are in matrix form

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix},$$

$$\begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & 0 \\ 0 & 0 & -1 & 0 \\ 0 & 0 & 0 & 1 \end{pmatrix},$$

and $\pi((1,0),(1,1))$ is

$$\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} \otimes \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \begin{pmatrix} 0 & 1 & 0 & 0 \\ -1 & 0 & 0 & 0 \\ 0 & 0 & 0 & -1 \\ 0 & 0 & 1 & 0 \end{pmatrix}.$$

Proposition 6.1 above generalizes to the following result.

Proposition 6.4 *The normalized time-frequency shift operators* $\{1/\sqrt{|G|} \times \pi(\lambda)\}_{\lambda \in G \times \widehat{G}}$ *form an orthonormal basis for the space of linear operators on* $\mathbb{C}^G$ *equipped with the Hilbert–Schmidt inner product.*

Proof This follows from direct computation or by simply using the fact that the tensors of orthonormal bases form an orthonormal basis of the tensor space. □

Consider again $G = \mathbb{Z}_2 \times \mathbb{Z}_2$. Then

$$G \times \widehat{G} = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \widehat{\mathbb{Z}_2 \times \mathbb{Z}_2} = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \widehat{\mathbb{Z}_2} \times \widehat{\mathbb{Z}_2} = \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2$$

and the Gabor system $((1,2,3,4)^T, \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2 \times \mathbb{Z}_2)$ consists of the columns of

$$\left(\begin{array}{cccc|cccc|cccc|cccc} 1 & 1 & 1 & 1 & 2 & 2 & 2 & 2 & 3 & 3 & 3 & 3 & 4 & 4 & 4 & 4 \\ 2 & -2 & 2 & -2 & 1 & -1 & 1 & -1 & 4 & -4 & 4 & -4 & 3 & -3 & 3 & -3 \\ 3 & 3 & -3 & -3 & 4 & 4 & -4 & -4 & 1 & 1 & -1 & -1 & 2 & 2 & -2 & -2 \\ 4 & -4 & -4 & 4 & 3 & -3 & -3 & 3 & 2 & -2 & -2 & 2 & 1 & -1 & -1 & 1 \end{array}\right).$$

Note that the Gabor system above is not the tensor product of two Gabor systems on the finite Abelian group $\mathbb{Z}_2$ because $(1,2,3,4)^T$ is not a simple tensor; that is, it does not have the form $v \otimes w$ for $v, w \in \mathbb{C}^{\mathbb{Z}_2}$. Certainly, Gabor systems on product groups can be generated by tensoring Gabor systems on the component groups; that is, for finite Abelian groups G_1 and G_2 with subsets $\Lambda_1 \subseteq G_1$ and $\Lambda_2 \subseteq G_2$, and $\varphi_1 \in \mathbb{C}^{G_1}$ and $\varphi_2 \in \mathbb{C}^{G_2}$, we obtain the $\mathbb{C}^{G_1 \times G_2}$ Gabor system

$$(\varphi_1, \Lambda_1) \otimes (\varphi_2, \Lambda_2) = (\varphi_1 \otimes \varphi_2, \Lambda_1 \times \Lambda_2).$$

See, for example, [22, 32].

Every Gabor system (φ, Λ), $\varphi \neq 0$, with $\Lambda = G \times \widehat{G}$ is a tight frame for $\mathbb{C}^G$, but certainly other algebraic and geometric properties of (φ, Λ) depend on the group G and the window function φ, as we will discuss below.

6.5 Elementary Properties of Gabor Frames and of the Gabor Frame Operator

In this section we derive central properties of Gabor frames for $\mathbb{C}^G$. Throughout this chapter, the reader may choose to assume $\mathbb{C}^G = \mathbb{C}^N = \mathbb{C}^{\{0,1,\ldots,N-1\}}$, as considered in Sect. 6.2. Indeed, Sect. 6.2 reflects the special case $G = \widehat{G} = \mathbb{Z}_N = \{0, 1, \ldots, N-1\}$.

Gabor frames are derived from group frames as described in Definition 5.3 in Sect. 5.2, a fact responsible for the Gabor system $(\varphi, G \times \widehat{G})$ being a tight frame for all $\varphi \in \mathbb{C}^G \setminus \{0\}$ (see Sect. 5.4 and [34, 35, 45, 46]). Gabor frames (φ, Λ) with Λ being a subgroup of $G \times \widehat{G}$ share a number of remarkable properties that are rooted in the fact that $\pi : G \times \widehat{G} \longrightarrow \mathscr{L}(\mathbb{C}^G, \mathbb{C}^G)$, $\lambda \mapsto \pi(\lambda)$, is a projective representation [35]. (It is, in fact, up to isomorphisms, the only irreducible faithful projective representation of $G \times \widehat{G}$ on $\mathbb{C}^G$ [35].)

The results proven below have been derived in the setting of general finite Abelian groups in [34] and [35]. There, the authors use nontrivial facts from representation theory. Our aim remains to give a self-contained treatment of Gabor frames in finite dimensions, so we present elementary linear algebra proofs instead.

The following simple observation forms the foundation for most fundamental results in Gabor analysis. In abstract terms, (6.14) and (6.15) represent the previously mentioned fact that π is a projective representation.

Proposition 6.5 *For $\lambda, \mu \in G \times \widehat{G}$ exists $c_{\lambda,\mu}, c_{\mu,\lambda}$ in $\mathbb{C}$, $|c_{\lambda,\mu}| = |c_{\mu,\lambda}| = 1$, with*

$$\pi(\lambda)\pi(\mu) = c_{\lambda,\mu}\pi(\lambda+\mu) = c_{\lambda,\mu}\overline{c_{\mu,\lambda}}\pi(\mu)\pi(\lambda) \tag{6.14}$$

and

$$\pi(\lambda)^{-1} = \pi(\lambda)^* = c_{\lambda,\lambda}\pi(-\lambda). \tag{6.15}$$

If Λ is a subgroup of $G \times \widehat{G}$, then the time-frequency shifts $\pi(\mu)$, $\mu \in \Lambda$, commute with the (φ, Λ) Gabor frame operator

$$S : \mathbb{C}^G \longrightarrow \mathbb{C}^G, \qquad x \mapsto \sum_{\lambda\in\Lambda} \langle x, \pi(\lambda)\varphi\rangle \pi(\lambda)\varphi$$

for every $\varphi \in \mathbb{C}^G$.

Proof For $G = \mathbb{Z}_N$, a direct computation shows that $c_{(k,\ell)(\widetilde{k},\widetilde{\ell})} = e^{-2\pi i k\widetilde{\ell}/N}$. This implies (6.14) and (6.15) in the case of cyclic groups. The general case follows from the facts that any finite Abelian group is the product of cyclic groups and that time-frequency shift operators on $\mathbb{C}^G$ are tensor products of time-frequency shift operators on $\mathbb{C}^{\mathbb{Z}_N}$.

To show that $S\pi(\mu) = \pi(\mu)S$ for $\mu \in \Lambda$, we compute

$$\begin{aligned}
\pi(\mu)^* S\pi(\mu)x &= \sum_{\lambda\in\Lambda} \langle \pi(\mu)f, \pi(\lambda)\varphi\rangle \pi(\mu)^*\pi(\lambda)\varphi \\
&= \sum_{\lambda\in\Lambda} \langle x, c_{\mu,\mu}\pi(-\mu)\pi(\lambda)\varphi\rangle c_{\mu,\mu}\pi(-\mu)\pi(\lambda)\varphi \\
&= |c_{\mu,\mu}|^2 \sum_{\lambda\in\Lambda} \langle x, c_{\mu(-\lambda)}\pi(\lambda-\mu)\varphi\rangle c_{\mu(-\lambda)}\pi(\lambda-\mu)\varphi \\
&= \sum_{\lambda\in\Lambda} \langle x, \pi(\lambda-\mu)\varphi\rangle |c_{\mu(-\lambda)}|^2 \pi(\lambda-\mu)\varphi \\
&= \sum_{\lambda\in\Lambda} \langle x, \pi(\lambda)\varphi\rangle \pi(\lambda)\varphi \;=\; Sx.
\end{aligned}$$

The substitution in the last step utilizes the fact that $\mu \in \Lambda$ and Λ is a group. □

As the first consequence of Proposition 6.5, we derive *Janssen's representation* (6.17) of the Gabor frame operator [54].

To this end, define the *adjoint subgroup* of the subgroup $\Lambda \subseteq G \times \widehat{G}$ to be

$$\Lambda^\circ = \big\{\mu \in G \times \widehat{G} :\ \pi(\lambda)\pi(\mu) = \pi(\mu)\pi(\lambda) \text{ for all } \lambda \in \Lambda\big\}.$$

Similarly to $(\Lambda^\perp)^\perp = \Lambda$, we have $(\Lambda^\circ)^\circ = \Lambda$. For illustrative purposes, we depict some lattices, their duals, and their adjoints in Fig. 6.7.

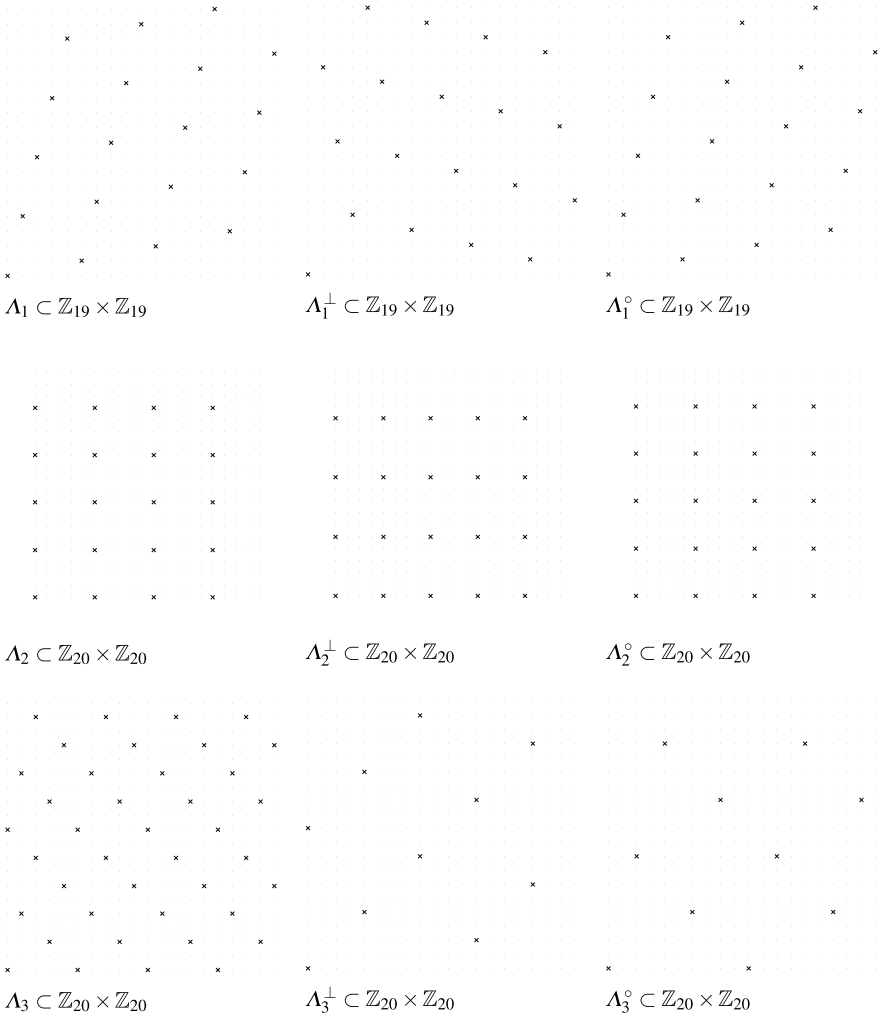


Fig. 6.7 Examples of lattices, their dual lattices, and their adjoint lattices. The lattice $\Lambda_1 \subset \mathbb{Z}_{19} \times \mathbb{Z}_{19}$ is the smallest subgroup of $\mathbb{Z}_{19} \times \mathbb{Z}_{19}$ containing $(1, 4)$, $\Lambda_2 \subset \mathbb{Z}_{20} \times \mathbb{Z}_{20}$ is generated by $(1, 2)$, and $\Lambda_3 \subset \mathbb{Z}_{20} \times \mathbb{Z}_{20}$ is the subgroup generated by the set $\{(1, 4), (0, 10)\}$

Theorem 6.4 *Let Λ be a subgroup of $G \times \widehat{G}$ and let $\varphi, \widetilde{\varphi} \in \mathbb{C}^G$. Then*

$$\sum_{\lambda \in \Lambda} \langle x, \pi(\lambda)\varphi \rangle \pi(\lambda)\widetilde{\varphi} = |\Lambda|/|G| \sum_{\mu \in \Lambda^\circ} \langle \widetilde{\varphi}, \pi(\mu)\varphi \rangle \pi(\mu)x, \quad x \in \mathbb{C}^G. \tag{6.16}$$

In particular, the (φ, Λ) Gabor frame operator S has the form

$$S = |\Lambda|/|G| \sum_{\mu \in \Lambda^\circ} \langle \varphi, \pi(\mu)\varphi \rangle \pi(\mu). \tag{6.17}$$

Setting $K = \{k : (k, \ell) \in \Lambda \text{ for some } \ell \in \widehat{G}\}$, we note that the matrix representing the frame operator with respect to the Euclidean orthonormal basis has support in the union of $|K|$ (off) diagonals. Walnut's representation (6.21) below will give additional insight on this canonical matrix representation of Gabor frame operators.

Proof Recall Proposition 6.4, namely the fact that $\{1/\sqrt{|G|}\,\pi(\lambda)\}_{\lambda\in G\times\widehat{G}}$ forms an orthonormal basis for the space of linear operators on $\mathbb{C}^G$ which is equipped with the Hilbert–Schmidt inner product. Hence, for $\varphi, \widetilde{\varphi} \in \mathbb{C}^G$, the operator

$$S : x \mapsto \sum_{\lambda\in\Lambda} \big\langle x, \pi(\lambda)\varphi\big\rangle \pi(\lambda)\widetilde{\varphi}$$

has a unique representation:

$$S = \sum_{\mu\in G\times\widehat{G}} \eta_\mu \pi(\mu).$$

Applying Proposition 6.5 gives for any $\lambda \in \Lambda$

$$\sum_{\mu\in G\times\widehat{G}} \eta_\mu \pi(\mu) = S = \pi(\lambda)^* S \pi(\lambda) = \sum_{\mu\in G\times\widehat{G}} \eta_\mu \pi(\lambda)^* \pi(\mu)\pi(\lambda).$$

Equations (6.14) and (6.15) in Proposition 6.5 imply that $\pi(\lambda)^*\pi(\mu)\pi(\lambda)$ is a scalar multiple of $\pi(\mu)$. As the coefficients η_μ, $\mu \in G \times \widehat{G}$, are unique, we have for each $\mu \in G \times \widehat{G}$ either $\eta_\mu = 0$ or $\pi(\lambda)^*\pi(\mu)\pi(\lambda) = \pi(\mu)$ for all $\lambda \in \Lambda$, that is, $\mu \in \Lambda^\circ$. We conclude that $\eta_\mu = 0$ if $\mu \notin \Lambda^\circ$.

It remains to show that for $\mu \in \Lambda^\circ$, we have $\eta_\mu = |\Lambda|/|G|\ \langle \widetilde{\varphi}, \pi(\mu)\varphi\rangle$. To this end, note that the rank one operator $x \mapsto \langle x, \varphi\rangle\widetilde{\varphi}$ is represented by the matrix $\widetilde{\varphi}\overline{\widetilde{\varphi}}^T$. Its Hilbert–Schmidt inner product with a matrix M satisfies $\langle \widetilde{\varphi}\overline{\widetilde{\varphi}}^T, M\rangle_{\mathrm{HS}} = \langle \widetilde{\varphi}, M\varphi\rangle$. Consequently, for $\mu \in \Lambda^\circ$, we have

$$\begin{aligned}
\eta_\mu &= 1/|G|\ \big\langle S, \pi(\mu)\big\rangle_{\mathrm{HS}} = 1/|G|\ \sum_{\lambda\in\Lambda} \big\langle \pi(\lambda)\widetilde{\varphi}\overline{\pi(\lambda)\varphi}^T, \pi(\mu)\big\rangle_{\mathrm{HS}} \\
&= 1/|G|\ \sum_{\lambda\in\Lambda} \big\langle \pi(\lambda)\widetilde{\varphi},\ \pi(\mu)\pi(\lambda)\varphi\big\rangle = 1/|G|\ \sum_{\lambda\in\Lambda} \big\langle \pi(\lambda)\widetilde{\varphi},\ \pi(\lambda)\pi(\mu)\varphi\big\rangle \\
&= 1/|G|\ \sum_{\lambda\in\Lambda} \big\langle \widetilde{\varphi},\ \pi(\mu)\varphi\big\rangle = |\Lambda|/|G|\ \big\langle \widetilde{\varphi},\ \pi(\mu)\varphi\big\rangle.
\end{aligned}$$

□

Taking inner products of the left-hand and right-hand sides of (6.16) with $\widetilde{x} \in \mathbb{C}^G$ shows that Janssen's representation implies the *fundamental identity in Gabor analysis* (FIGA) (6.18) below; see also [36, 46].

Corollary 6.1 *Let Λ be a subgroup of $G \times \widehat{G}$. Then*

$$\sum_{\lambda\in\Lambda} V_\varphi x(\lambda)\ \overline{V_{\widetilde{\varphi}}\widetilde{x}(\lambda)} = |\Lambda|/|G| \sum_{\lambda\in\Lambda^\circ} V_\varphi\widetilde{\varphi}(\lambda)\ \overline{V_x\widetilde{x}(\lambda)}, \quad x, \widetilde{x}, \varphi, \widetilde{\varphi} \in \mathbb{C}^G. \tag{6.18}$$

An additional important consequence of Proposition 6.5 is the fact that the canonical duals of Gabor frames are again Gabor frames; that is, the canonical dual frame of a Gabor frame inherits the time-frequency structure of the original frame.

Theorem 6.5 *Let Λ be a subgroup of $G \times \widehat{G}$, and let the Gabor system (φ, Λ) span $\mathbb{C}^G$. The canonical dual frame of (φ, Λ) has the form $(\widetilde{\varphi}, \Lambda)$; that is, for appropriate $\widetilde{\varphi} \in \mathbb{C}^G$ we have*

$$x = \sum_{\lambda \in \Lambda} \langle x, \pi(\lambda)\widetilde{\varphi}\rangle \pi(\lambda)\varphi = \sum_{\lambda \in \Lambda} \langle x, \pi(\lambda)\varphi\rangle \pi(\lambda)\widetilde{\varphi}, \quad x \in \mathbb{C}^G.$$

Proof Proposition 6.5 states that the (φ, Λ) frame operator

$$S : \mathbb{C}^G \longrightarrow \mathbb{C}^G, \qquad x \mapsto \sum_{\lambda \in \Lambda} \langle x, \pi(\lambda)\varphi\rangle \pi(\lambda)\varphi,$$

and, consequently, its inverse S^{-1}, commute with $\pi(\mu)$ for $\mu \in \Lambda$. Hence, the elements of the canonical dual frame of (φ, Λ) are of the form

$$\gamma_\lambda = S^{-1}\pi(\lambda)\varphi = \pi(\lambda)S^{-1}\varphi = \pi(\lambda)\widetilde{\varphi}, \quad \lambda \in \Lambda. \qquad \square$$

For overcomplete Gabor frames, that is, Gabor frames that span $\mathbb{C}^G$ and have cardinality larger than $N = |G|$, the dual window is not unique. In fact, choosing dual frames different from the canonical dual frame may allow us to reduce the computational complexity needed to compute the coefficients of a Gabor expansion [91].

Gabor frames $(\widetilde{\varphi}, \Lambda)$ that are dual to (φ, Λ) are characterized by the following *Wexler–Raz criterion* (see [35, 98] and references therein). It is a direct consequence of Theorem 6.4.

Theorem 6.6 *Let Λ be a subgroup of $G \times \widehat{G}$. For the Gabor systems (φ, Λ) and $(\widetilde{\varphi}, \Lambda)$, we have*

$$x = \sum_{\lambda \in \Lambda} \langle x, \pi(\lambda)\widetilde{\varphi}\rangle \pi(\lambda)\varphi, \quad x \in \mathbb{C}^G, \tag{6.19}$$

if and only if

$$\langle \varphi, \pi(\mu)\widetilde{\varphi}\rangle = |G|/|\Lambda|\, \delta_{\mu,0}, \quad \mu \in \Lambda^\circ. \tag{6.20}$$

Proof Equation (6.19) implies that the operator $S : x \mapsto \sum_{\lambda \in \Lambda} \langle x, \pi(\lambda)\varphi\rangle \pi(\lambda)\varphi$ is the identity; that is, by Theorem 6.4 we have

$$\pi(0) = Id = S = |\Lambda|/|G| \sum_{\mu \in \Lambda^\circ} \langle \varphi, \pi(\mu)\widetilde{\varphi}\rangle \pi(\mu).$$

As the operators $\{\pi(\mu)\}$ are linearly independent by Proposition 6.4, we conclude that $|\Lambda|/|G|\ \langle\varphi, \pi(\mu)\widetilde{\varphi}\rangle = \delta_{\mu,0}$, which is (6.20).

The reverse implication follows trivially from Janssen's representation. □

Corollary 6.2 *If Λ is a subgroup of $G \times \widehat{G}$, then (φ, Λ) is a tight frame for $\mathbb{C}^G$ if and only if (φ, Λ°) is an orthogonal set.*

Proof The result follows from choosing $\widetilde{\varphi} = \varphi$ in (6.19) and (6.20). □

Moreover, the Wexler–Raz criterion Theorem 6.6 implies the following *Ron–Shen duality* result [35, 83].

Theorem 6.7 *Let Λ be a subgroup of $G \times \widehat{G}$. The system (φ, Λ) is a frame for $\mathbb{C}^G$ if and only if (φ, Λ°) is a linear independent set.*

Proof If (φ, Λ) is a frame, then Theorem 6.6 implies the existence of a dual window $\widetilde{\varphi}$ with $\langle\pi(\lambda)\varphi, \pi(\mu)\widetilde{\varphi}\rangle = \delta_{\lambda,\mu}$ for $\lambda, \mu \in \Lambda^\circ$. But then $0 = \sum_{\lambda\in\Lambda^\circ} c_\lambda \pi(\lambda)\varphi$ implies for $\mu \in \Lambda^\circ$ that

$$0 = \left\langle \sum_{\lambda\in\Lambda^\circ} c_\lambda \pi(\lambda)\varphi, \pi(\mu)\widetilde{\varphi} \right\rangle = \sum_{\lambda\in\Lambda^\circ} c_\lambda \langle \pi(\lambda)\varphi, \pi(\mu)\widetilde{\varphi}\rangle = c_\mu \langle \pi(\mu)\varphi, \pi(\mu)\widetilde{\varphi}\rangle,$$

and we conclude that $c_\mu = 0$ for all $\mu \in \Lambda^\circ$. Hence, (φ, Λ°) is linearly independent.

On the other hand, if (φ, Λ°) is a linear independent set, then there exists a unique vector $\widetilde{\varphi}$ in $\operatorname{span}\{\pi(\mu)\varphi\}_{\mu\in\Lambda^\circ}$ which is orthogonal to $\operatorname{span}\{\pi(\mu)\varphi\}_{\mu\in\Lambda^\circ\setminus\{0\}}$ and $\langle\varphi, \pi(\mu)\widetilde{\varphi}\rangle = \delta_{\mu,0}$ for all $\mu \in \Lambda^\circ$. Theorem 6.6 implies that (φ, Λ) is a frame. □

We close this section with a novel general version of *Walnut's representation* of the Gabor frame operator in the finite-dimensional setting.

Theorem 6.8 *For a subgroup Λ of $G \times \widehat{G}$, set $H_0 = \{\ell : (0, \ell) \in \Lambda\}$ and $K = \{k : (k, \ell) \in \Lambda$ for some $\ell\}$. For each $k \in K$ choose an ℓ_k with $(k, \ell_k) \in \Lambda$. The (φ, Λ) Gabor frame operator matrix $(S_{\widetilde{n}n})$ satisfies*

$$S_{\widetilde{n}n} = |H_0|\ \chi_{H_0^\perp}(\widetilde{n} - n) \sum_{k\in K} \varphi(\widetilde{n} - k)\overline{\varphi(n - k)}\ \langle \ell_k, \widetilde{n} - n\rangle \tag{6.21}$$

where $H_0^\perp = \{\ell \in G : \langle\ell, k\rangle = 1$ for all $k \in H_0\}$ denotes the annihilator subgroup of H_0. If $\Lambda = \Lambda_1 \times \Lambda_2$, then (6.21) reduces to

$$S_{\widetilde{n}n} = |\Lambda_1|\ \chi_{\Lambda_2^\perp}(\widetilde{n} - n) \sum_{k\in\Lambda_1} \varphi(\widetilde{n} - k)\overline{\varphi(n - k)}. \tag{6.22}$$

Proof For $k \in K$, let H_k denote the k-section of Λ, that is, $H_k = \{\ell : (k, \ell) \in \Lambda$ for some $\ell \in \widehat{G}\}$. Clearly, $\ell, \widetilde{\ell} \in H_k$ if and only if $\widetilde{\ell} - \ell \in H_0$. Hence, $H_k = H_0 + \ell_k$ for any $\ell_k \in H_k \subseteq \widehat{G}$.

We compute

$$\begin{aligned} S_{\widetilde{n}n} &= \sum_{\lambda\in\Lambda} \pi(\lambda)\varphi(\widetilde{n})\big(\pi(\lambda)\varphi(n)\big)^* \\ &= \sum_{k\in K}\sum_{\ell\in H_k} \varphi(\widetilde{n}-k)\langle \ell,\widetilde{n}\rangle\ \overline{\varphi(n-k)\langle \ell, n\rangle} \\ &= \sum_{k\in K} \varphi(\widetilde{n}-k)\overline{\varphi(n-k)}\ \sum_{\ell\in H_0}\langle \ell+\ell_k,\widetilde{n}-n\rangle \\ &= \sum_{k\in K} \varphi(\widetilde{n}-k)\overline{\varphi(n-k)}\ \langle \ell_k,\widetilde{n}-n\rangle\ \sum_{\ell\in H_0}\langle \ell,\widetilde{n}-n\rangle \\ &\overset{(6.12)}{=} \sum_{k\in K} \varphi(\widetilde{n}-k)\overline{\varphi(n-k)}\ \langle \ell_k,\widetilde{n}-n\rangle\ |H_0|\ \chi_{H_0^\perp}(\widetilde{n}-n). \end{aligned}$$

Equation (6.22) follows directly from (6.21) by observing that $K = \Lambda_1$, $H_0 = H_k = \Lambda_2$, and $\ell_k = 0$ for $k \in \Lambda_1$. □

Equation (6.22) implies that for real-valued φ the frame operator S for $(\varphi, \Lambda_1 \times \Lambda_2)$ restricts to $\mathbb{R}^G$ and, in particular, the dual frame generating window $\gamma = S^{-1}\varphi$ is then real-valued as well. The band structure of Gabor frame operators that is displayed in (6.21) and (6.22) is also observed in Janssen's representation (6.17). It shows that at most $|H_0^\perp||G| = |G|/|H_0|$ entries of S are nonzero. This observation is in particular valuable if H_0, respectively Λ_2, is a large subgroup of $\widehat{G}$.

6.6 Linear Independence

A traditional and frequent task in Gabor analysis on the real line is to show that a given Gabor system is a Riesz basis in, or a frame for, the Hilbert space of complex valued square integrable functions $L^2(\mathbb{R})$. Simple linear independence of Gabor systems in $L^2(\mathbb{R})$ was first considered by Heil, Ramanathan, and Topiwala [50]. Their conjecture that the members of every Gabor system are linearly independent in $L^2(\mathbb{R})$ remains open to this date. In fact, it is unknown whether for all window functions φ in $L^2(\mathbb{R})$, the four functions

$$\varphi(t),\ \varphi(t-1),\ e^{2\pi i t}\varphi(t),\ e^{2\pi i\sqrt{2}t}\varphi\big(t-\sqrt{2}\big)$$

are linearly independent [26, 50].

In finite dimensions, a family of vectors is a Riesz basis for its span if and only if the vectors are linearly independent. Similarly, a family of vectors is a frame if and only if they span the finite-dimensional ambient space. Clearly, the dimension of the ambient space limits the number of linearly independent vectors, and in this section, we address the question of whether the vectors of a Gabor system in $\mathbb{C}^G$

are in *general linear position*. That is, we ask which Gabor frames (φ, Λ) have the property that every selection of less than or equal to $|G| = \dim \mathbb{C}^G$ vectors from (φ, Λ) is linearly independent.

As before, for a vector x in a finite-dimensional space let

$$\|x\|_0 = |\operatorname{supp} x|$$

count the nonzero entries of x. Also, recall that the *spark* of a matrix M is given by $\min\{\|c\|_0,\ c \neq 0,\ Mc = 0\}$. Rephrasing the above, we ask the question: For which φ and Λ is the spark of the (φ, Λ) synthesis operator equal to $|G| + 1$? Note that in complementary work, upper bounds on the spark of certain Gabor synthesis operators were obtained [99].

Before stating the main results from [59, 64], we will motivate the line of work presented here by describing its relevance to information transmission in erasure channels and in operator identification [59]. As a byproduct of our analysis, we obtain a large family of unimodular tight frames that are maximally robust to erasures [18].

In generic communication systems, information in the form of a vector $x \in \mathbb{C}^G$ is not transmitted directly. First, it is coded in a way that allows for the recovery of x at the receiver, regardless of errors that may be introduced by the communications channel. To achieve some robustness against errors, we can choose a frame $\{\varphi_k\}_{k\in K}$ for $\mathbb{C}^G$ and transmit x in the form of coefficients $\{\langle x, \varphi_k\rangle\}_{k\in K}$. At the receiver, a dual frame $\{\widetilde{\varphi}_k\}$ of $\{\varphi_k\}$ can be used to recover x via the frame reconstruction formula $x = \sum_k \langle x, \varphi_k\rangle \widetilde{\varphi}_k$.

In the case of an *erasure channel*, some of the transmitted coefficients may be lost. If only the coefficients $\{\langle x, \varphi_k\rangle\}_{k\in K'}$, $K' \subseteq K$, are received, then the original vector x can still be recovered[3] if and only if the subset $\{\varphi_k\}_{k\in K'}$ remains a frame for $\mathbb{C}^G$. Of course, this requires $|K'| \geq |G| = \dim \mathbb{C}^G$.

Definition 6.1 A frame $\Phi = \{\varphi_k\}_{k\in K}$ in $\mathbb{C}^G$ is maximally robust to erasures if the removal of any $L \leq |K| - |G|$ vectors from $\mathscr{F}$ leaves a frame.

By definition, a frame is maximally robust to erasures if and only if the frame vectors are in general linear position.

Another important application is the problem of identifying linear time-varying operators.

Definition 6.2 A linear space of operators $\mathscr{H} \subseteq \{H : \mathbb{C}^G \to \mathbb{C}^G,\ H \text{ linear}\}$ is identifiable with identifier φ if the linear map $E_\varphi : \mathscr{H} \to \mathbb{C}^G$, $H \mapsto H\varphi$, is injective.

A time-varying communication channel is frequently modeled as a linear combination of time-frequency shift operators. The idea behind this model is that the

[3] Here we assume that the receiver knows which coefficients have been erased and which coefficients have been received.

transmitted signal reaches the receiver through a small number of paths, each path causing a path-specific delay k, a path-specific frequency shift ℓ (due to Doppler effects), and a path-specific gain factor $c_{k,\ell}$. If we have a priori knowledge of the time-frequency shifts Λ caused by the paths the signals travel, then we aim to obtain knowledge of the gain factors, that is, we aim to identify operators from the class

$$\mathscr{H}_\Lambda = \left\{ \sum_{\lambda\in\Lambda} c_\lambda \pi(\lambda),\ c_\lambda \in \mathbb{C} \right\}, \quad \Lambda \subseteq G \times \hat{G}.$$

Clearly, knowing the channel is a crucial prerequisite for a successful transmission of information; see [20, 59, 74].

Often, the time delays and the modulation parameters are not known, but we may have an upper bound on the number of paths the signal may travel to the receiver. Then, we aim to identify the class of operators

$$\mathscr{H}_s = \left\{ \sum_{\lambda\in\Lambda} c_\lambda \pi(\lambda), c_\lambda \in \mathbb{C},\ \Lambda \subseteq G \times \hat{G} \text{ with } |\Lambda| \le s \right\}. \tag{6.23}$$

The following result relates the concepts discussed above.

Theorem 6.9 *The following are equivalent for* $\varphi \in \mathbb{C}^G \backslash \{0\}$:

1. *The Gabor system* $(\varphi, G \times \widehat{G})$ *is in general linear position.*
2. *The Gabor system* $(\varphi, G \times \widehat{G})$ *forms an equal norm tight frame which is maximally robust to erasures.*
3. *For all* $x \in \mathbb{C}^G \setminus \{0\}$, $\|V_\varphi x\|_0 \ge |G|^2 - |G| + 1$.
4. *For all* $x \in \mathbb{C}^G$, $V_\varphi x$ *and, therefore,* x *is completely determined by its values on any set* Λ *with* $|\Lambda| = |G|$.
5. $\mathscr{H}_\Lambda$ *is identifiable by* φ *if and only if* $|\Lambda| \le |G|$.

If $|G|$ *is even, then statements* 1–5 *are equivalent to statement* 6 *below, for* $|G|$ *odd, statements* 1–5 *imply statement* 6:

6. $\mathscr{H}_s$ *is identifiable by* φ *if and only if* $s \le |G|/2$.

Proof The equivalence of statements 1–5 follows from standard linear algebra arguments [59, 64]. Note in addition that to deduce statement 2 from any of the other statements, we can use that *a priori* $(\varphi, G \times \widehat{G})$ is an equal norm tight frame as long as $\varphi \neq 0$.

For illustrative purposes, we give below a proof of statement 1 implies statement 6. Assume that the vectors in $(\varphi, G \times \widehat{G})$ are in general position and $s \le |G|/2$. Then $H\varphi = \widetilde{H}\varphi$ for $H, \widetilde{H}$ implies

$$0 = \sum_{\lambda\in\Lambda} c_\lambda \pi(\lambda)\varphi - \sum_{\widetilde{\lambda}\in\widetilde{\Lambda}} \widetilde{c}_{\widetilde{\lambda}} \pi(\widetilde{\lambda})\varphi.$$

Note that the right-hand side is a linear combination of elements from $(\varphi, \Lambda \cup \widetilde{\Lambda}) \subseteq (\varphi, G \times \widehat{G})$ with $|(\varphi, \Lambda \cup \widetilde{\Lambda})| \le |\Lambda \cup \widetilde{\Lambda}| \le 2|G|/2 = |G|$. Statement 1 implies linear

independence of $(\varphi, \Lambda \cup \widetilde{\Lambda})$; hence, all coefficients are 0 or cancel out. We conclude that $H = \widetilde{H}$.

A similar argument shows that, in general, $\mathscr{H}_s$ is not identifiable if $s > |G|/2$. □

Theorem 6.9 leads to the question of whether a φ satisfying statements 1–6 in Theorem 6.9 exists. For the special case $|G|$ prime, the answer is affirmative [59, 64].

Theorem 6.10 *If $G = \mathbb{Z}_p$, p prime, then φ exists in $\mathbb{C}^G$ such that statements* 1–6 *in Theorem* 6.9 *are satisfied. Moreover, we can choose the vector φ to be unimodular.*

Proof A complete proof is given in [64]. It is nontrivial, and we will only recall some of its central ideas.

Consider the Gabor window consisting of p complex variables $z_0, z_1, \ldots, z_{p-1}$. Take $\Lambda \subseteq G \times \widehat{G}$ with $|\Lambda| = p$ and form a matrix from the p vectors in the Gabor system (z, Λ). The determinant of the matrix is a homogeneous polynomial P_Λ in $z_0, z_1, \ldots, z_{p-1}$ of degree p. We have to show that $P_\Lambda \neq 0$. This is achieved by observing that at least one monomial appears in the polynomial P_Λ with a coefficient which is not 0. Indeed, it can be shown that there exists at least one monomial whose coefficient is the product of minors of the Fourier matrix W_p. We can apply Chebotarev's theorem on roots of unity (see Theorem 6.12). It states that every minor of the Fourier matrix W_p, p prime, is nonzero [31, 40, 93], a property that does not hold for groups with $|G|$ composite. Hence, $P_\Lambda \neq 0$.

We conclude that for each $\Lambda \subseteq G \times \widehat{G}$ with $|\Lambda| = p$, the determinant P_Λ vanishes only on the nontrivial algebraic variety $E_\Lambda = \{z = (z_0, z_1, \ldots, z_{p-1}) : \ P_\Lambda(z) = 0\}$. E_Λ has Lebesgue measure 0; hence, any generic φ, that is,

$$\varphi \in \mathbb{C}^G \Big\backslash \Big(\bigcup_{\Lambda \subseteq G \times \widehat{G},\ |\Lambda| = p} E_\Lambda \Big),$$

generates a Gabor system $(\varphi, G \times \widehat{G})$ in general linear position.

To show that we can choose a unimodular φ, it suffices to demonstrate that the set of unimodular vectors is not contained in $\bigcup_{\Lambda \subseteq G \times \widehat{G},\ |\Lambda| = p} E_\Lambda$ [59]. □

Theorem 6.10 is complemented by the following simple observation.

Theorem 6.11 *If $G = \mathbb{Z}^2 \times \mathbb{Z}^2$, then there exists no φ in $\mathbb{C}^G$ such that the vectors in $(\varphi, G \times \widehat{G})$ are in general linear position.*

Proof For a generic $\varphi = (c_0, c_1, c_2, c_3)^T$, we compute the determinant of the matrix with columns φ, $\pi((0,0),(1,0))\varphi$, $\pi((1,1),(0,0))\varphi$, and $\pi((1,1),(0,1))\varphi$, that is,

$$
\begin{aligned}
&\det \begin{pmatrix} c_0 & c_0 & c_3 & c_3 \\ c_1 & c_1 & c_2 & -c_2 \\ c_2 & -c_2 & c_1 & c_1 \\ c_3 & -c_3 & c_0 & -c_0 \end{pmatrix} \\
&= \det \begin{pmatrix} 0 & 2c_0 & 0 & 2c_3 \\ 0 & 2c_1 & 2c_2 & 0 \\ 2c_2 & 0 & 0 & 2c_1 \\ 2c_3 & 0 & 2c_0 & 0 \end{pmatrix} \\
&= -16c_0 \det \begin{pmatrix} 0 & c_2 & 0 \\ c_2 & 0 & c_1 \\ c_3 & c_0 & 0 \end{pmatrix} - 16c_3 \det \begin{pmatrix} 0 & c_1 & c_2 \\ c_2 & 0 & 0 \\ c_3 & 0 & c_0 \end{pmatrix} \\
&= -16c_0c_1c_2c_3 + 16c_0c_1c_2c_3 = 0.
\end{aligned}
$$

We conclude that for all φ, the four vectors φ, $\pi((0,0),(1,0))\varphi$, $\pi((1,1),(0,0))\varphi$, and $\pi((1,1),(0,1))\varphi$ are linearly dependent. □

In [59], numerical results show that a vector which satisfies statement 2, and therefore all statements in Theorem 6.9 for $G = \mathbb{Z}_4, \mathbb{Z}_6$, exists (see Fig. 6.8). This observation leads to the following open question [59].

Question 6.1 For the cyclic group $G = \mathbb{Z}_N$, $N \in \mathbb{N}$, does there exist a window φ in $\mathbb{C}^G$ with $(\varphi, G \times \widehat{G})$ in general linear position?

The numerical procedure applied to resolve the cases $G = \mathbb{Z}_4$ and $\mathbb{Z}_6$ is unfortunately not applicable to larger cyclic groups of composite order. In fact, to answer Question 6.1 for the group $G = \mathbb{Z}_8$ numerically would require the computation of 64 choose 8, which is 4,426,165,368 determinants of 8 by 8 matrices. (Using symmetries, the amount of computation can be reduced, but not enough to allow for a numerical solution of the problem at hand.)

The proof of Theorem 6.10 outlined above is not constructive. In fact, with the exception of small primes 2, 3, 5, 7, we cannot test numerically whether a given vector φ satisfies the statements in Theorem 6.9. Again, a naive direct approach to check whether the system $(\varphi, \mathbb{Z}_{11} \times \widehat{\mathbb{Z}_{11}})$ is in general linear position requires the computation of 121 choose 11, that is, 1,276,749,965,026,536 determinants of 11 by 11 matrices.

Question 6.2 For $G = \mathbb{Z}_p$, p prime, does there exist an explicit construction of φ in $\mathbb{C}^G$ such that the vectors in $(\varphi, G \times \widehat{G})$ are in general linear position?

The truth is that for $G = \mathbb{Z}_p$, p prime, it is known that almost every vector φ generates a system $(\varphi, G \times \widehat{G})$ in general linear position, but aside from groups of order less than or equal to 7, not a single vector φ with $(\varphi, G \times \widehat{G})$ in general linear position is known.

As illustrated by Theorem 6.9, a positive answer to Questions 6.1 and 6.2 would have far-reaching applications. For example, to our knowledge, the only previously known *equal norm tight frames that are maximally robust to erasures* are harmonic frames, that is, frames consisting of columns of Fourier matrices where some rows have been removed. (See, for example, the conclusions section in [18].) Similarly, Theorem 6.10 together with Theorem 6.9 provide us with equal norm tight frames with p^2 elements in $\mathbb{C}^N$ for $N \leq p$: we can choose a unimodular φ satisfying the conclusions of Theorem 6.10 and remove uniformly $p - N$ components of the equal norm tight frame $(\varphi, G \times \widehat{G})$ in order to obtain an equal norm tight frame for $\mathbb{C}^N$ which is maximally robust to erasure. Obviously, the removal of components does not leave a Gabor frame proper. Alternatively, eliminating some vectors from a Gabor frame satisfying the conclusions of Theorem 6.10 leaves an equal norm Gabor frame which is maximally robust to erasure but which might not be tight.

We point out that a positive answer to Question 6.1 would imply the generalization of sampling of operator results that hold on the space of square integrable functions on the real line to operators defined on square integrable functions on Euclidean spaces of higher dimensions [78].

In the remainder of this section, we describe an observation that might be helpful to establish a positive answer to Question 6.1. *Chebotarev's theorem* can be phrased in the form of an uncertainty principle, that is, as a manifestation of the principle that x and $\widehat{x}$ cannot both be well localized at the same time [93]. Recall that $\|x\|_0 = |\operatorname{supp} x|$.

Theorem 6.12 *For $G = \mathbb{Z}_p$, p prime, we have*

$$\|x\|_0 + \|\widehat{x}\|_0 \geq |G| + 1 = p + 1, \quad x \in \mathbb{C}^p \backslash \{0\}.$$

The corresponding *time-frequency uncertainty* result for the short-time Fourier transform is the following [59, 64].

Theorem 6.13 *Let $G = \mathbb{Z}_p$, p prime. For appropriately chosen $\varphi \in \mathbb{C}^p$,*

$$\|x\|_0 + \|V_\varphi x\|_0 \geq |G \times \widehat{G}| + 1 = p^2 + 1, \quad x \in \mathbb{C}^p \setminus \{0\}.$$

Theorems 6.12 and 6.13 are sharp in the sense that all pairs (u, v) satisfying the respective bound will correspond to the support size pair of a vector and its Fourier transform, respectively, its short-time Fourier transform. In particular, for almost every φ, we have that for all $1 \leq u \leq |G|$, $1 \leq v \leq |G|^2$ with $u + v \geq |G|^2 + 1$ there exists x with $\|x\|_0 = u$ and $\|V_\varphi x\|_0 = v$. Comparing Theorems 6.12 and 6.13, we observe that for $a, b \in \mathbb{Z}_p$, the pair of numbers $(a, p^2 - b)$ can be realized as $(\|x\|_0, \|V_\varphi x\|_0)$ if and only if $(a, p - b)$ can be realized as $(\|x\|_0, \|\widehat{x}\|_0)$. This observation leads to the following question [59].

Question 6.3 [64] For G cyclic, that is, $G = \mathbb{Z}_N$, $N \in \mathbb{N}$, does there exist φ in $\mathbb{C}^N$ such that

$$\left\{(\|x\|_0, \|V_\varphi x\|_0),\ x \in \mathbb{C}^N\right\} = \left\{\left(\|x\|_0, |G|^2 - |G| + \|\widehat{x}\|_0\right),\ x \in \mathbb{C}^N\right\}?$$

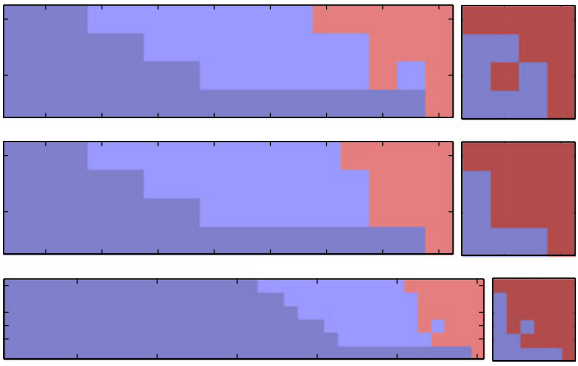


Fig. 6.8 The set $\{(\|x\|_0, \|V_\varphi x\|_0),\ x \in \mathbb{C}^G\backslash\{0\}\}$ for appropriately chosen $\varphi \in \mathbb{C}^G\backslash\{0\}$ for $G = \mathbb{Z}_2 \times \mathbb{Z}_2, \mathbb{Z}_4, \mathbb{Z}_6$. For comparison, the right column shows the set $\{(\|x\|_0, \|\widehat{x}\|_0),\ x \in \mathbb{C}^G\backslash\{0\}\}$. *Dark red/blue* implies that it is proven analytically in [59] that the respective pair (u, v) is achieved/is not achieved, where φ is a generic window. *Light red/blue* implies that it was shown numerically that the respective pair (u, v) is achieved/is not achieved

Figure 6.8 compares the achievable support size pairs $(\|x\|_0, \|V_\varphi x\|_0)$, φ chosen appropriately, and $(\|x\|_0, \|\widehat{x}\|_0)$ for the groups $\mathbb{Z}_2 \times \mathbb{Z}_2$, $\mathbb{Z}_4$, and $\mathbb{Z}_6$.

Note that any vector φ satisfying statements 1–6 in Theorem 6.9 has the property that $\|\varphi\|_0 = \|\widehat{\varphi}\|_0 = |G|$ [64]. For arbitrary $\varphi \neq 0$, it is easily observed that

$$\|V_\varphi x\|_0 \geq |G|, \quad x \in \mathbb{C}^G, \tag{6.24}$$

and stronger qualitative statements on $\|V_\varphi x\|_0$ depending on $\|\varphi\|_0$, $\|\widehat{\varphi}\|_0$, $\|x\|_0$, $\|\widehat{x}\|_0$ are provided in [59].

Ghobber and Jaming obtained quantitative versions of (6.24) and Theorem 6.13. For example, the result below estimates the energy of x that can be captured by a small number of components of $V_\varphi x$ [42].

Theorem 6.14 *Let $G = \mathbb{Z}_N$, $N \in \mathbb{N}$. For φ with $\|\varphi\| = 1$ and $\Lambda \subseteq G \times \widehat{G}$ with $|\Lambda| < |G| = N$, we have*

$$\sum_{\lambda\in\Lambda} \big|V_\varphi x(\lambda)\big|^2 \leq \big(1 - \big(1 - |\Lambda|/|G|\big)^2/8\big)\|x\|^2, \quad x \in \mathbb{C}^G.$$

6.7 Coherence

The analysis of the coherence of Gabor systems has a twofold motivation. First of all, many equiangular frames have been constructed as Gabor frames and, second, a number of algorithms aimed at solving underdetermined system $Ax = b$ for a sparse vector x succeed if the coherence of columns in A is sufficiently small; see Sect. 6.8 and [28, 44, 95–97].

The *coherence* of a unit norm frame $\Phi = \{\varphi_k\}$ is given by

$$\mu(\Phi) = \max_{k \neq \widetilde{k}} \left| \langle \varphi_k, \varphi_{\widetilde{k}} \rangle \right|.$$

That is, the coherence of a unit norm frame $\Phi = \{\varphi_k\}$ is the cosine of the smallest angle between elements from the frame. A unit norm frame $\Phi = \{\varphi_k\}$ with $|\langle \varphi_k, \varphi_{\widetilde{k}} \rangle| =$ constant for $k \neq \widetilde{k}$ is called an *equiangular frame*. It is easily seen that, among all unit norm frames with K elements in $\mathbb{C}^N$, the equiangular frames are those with minimal coherence.

If $\|\varphi\| = 1$, then the Gabor system (φ, Λ) is unit norm and, if Λ is a subgroup of $G \times \widehat{G}$, then Proposition 6.5 implies that the coherence of (φ, Λ) is

$$\mu(\varphi, \Lambda) = \max_{\lambda \in \Lambda \setminus \{0\}} \left| \langle \varphi, \pi(\lambda)\varphi \rangle \right| = \max_{\lambda \in \Lambda \setminus \{0\}} \left| V_\varphi \varphi(\lambda) \right|.$$

In frame theory, it is a well-known fact that for any unit norm frame Φ of K vectors in $\mathbb{C}^N$, we have

$$\mu(\Phi) \geq \sqrt{\frac{K - N}{N(K-1)}}; \tag{6.25}$$

see, for example, [92] and references therein. For tight frames, (6.25) follows from a simple estimate of the magnitude of the off-diagonal entries of the Gram matrix $(\langle \varphi_k, \varphi_{\widetilde{k}} \rangle)$:

$$\begin{aligned}(K-1)K\mu(\Phi)^2 &\geq \sum_{k \neq \widetilde{k}} \left| \langle \varphi_k, \varphi_{\widetilde{k}} \rangle \right|^2 = \sum_{k=1}^{K} \left(-\left| \langle \varphi_k, \varphi_k \rangle \right|^2 + \sum_{\widetilde{k}=1}^{K} \left| \langle \varphi_k, \varphi_{\widetilde{k}} \rangle \right|^2 \right) \\ &= \sum_{k=1}^{K} \left(-1 + \frac{K}{N} \|\varphi_k\|^2 \right) = \frac{K^2}{N} - K. \end{aligned} \tag{6.26}$$

This computation also shows that any tight frame with equality in (6.25) is equiangular. Note that equiangularity necessitates $K \leq N^2$, a result which holds for all unit norm frames [92].

The Gabor frame $(\varphi, G \times \widehat{G})$ has $|G|^2$ elements; hence, (6.25) simplifies to

$$\mu(\varphi, G \times \widehat{G}) \geq \sqrt{\frac{|G|^2 - |G|}{|G|(|G|^2 - 1)}} = \sqrt{\frac{|G| - 1}{|G|^2 - 1}} = 1/\sqrt{|G| + 1}.$$

Alltop considered the window $\varphi_A \in \mathbb{C}^p$, $p \geq 5$ prime, with entries

$$\varphi_A(k) = p^{-1/2} e^{2\pi i k^3/p}, \quad k = 0, 1, \ldots, p-1. \tag{6.27}$$

For the *Alltop window* function, we have [1, 92]

$$\mu(\varphi_A, \mathbb{Z}_p \times \widehat{\mathbb{Z}_p}) = 1/\sqrt{p},$$

which is close to the optimal lower bound $1/\sqrt{p+1}$. In fact, φ_A being unimodular implies that $(\varphi_A, G \times \widehat{G})$ is the union of $|G|$ orthonormal bases. A minor adjustment to the argument in (6.26) shows that whenever Φ is the union of N orthonormal bases for $\mathbb{C}^N$, we have necessarily $\mu(\Phi) \geq 1/\sqrt{N}$.

The Alltop window for $G = \mathbb{Z}_N$, N not prime, does not guarantee good coherence. For illustrative purposes, we display $|V_{\varphi_A}\varphi_A(\lambda)| = |\langle \varphi_A, \pi(\lambda)\varphi_A\rangle|$, $\lambda \in \mathbb{Z}_N \times \widehat{\mathbb{Z}_N}$, for $N = 6, 7, 8$,

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 \end{pmatrix}, \quad \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ u & u & u & u & u & u & u \\ u & u & u & u & u & u & u \\ u & u & u & u & u & u & u \\ u & u & u & u & u & u & u \\ u & u & u & u & u & u & u \\ u & u & u & u & u & u & u \end{pmatrix},$$

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0.5 & 0 & 0.5 \\ 1/\sqrt{2} & 0 & 0 & 0 & 1/\sqrt{2} & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0.5 & 0 & 0.5 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0.5 & 0 & 0.5 \\ 1/\sqrt{2} & 0 & 0 & 0 & 1/\sqrt{2} & 0 & 0 & 0 \\ 0 & 0.5 & 0 & 0.5 & 0 & 0.5 & 0 & 0.5 \end{pmatrix}, \tag{6.28}$$

where $u = 1/\sqrt{7} \approx 0.3880$.

Gabor systems $(\varphi_A, G \times \widehat{G})$ employing the Alltop window for $G = \mathbb{Z}_N$, $N \in \mathbb{N}$, were also analyzed numerically in [2] in terms of chirp sensing codes. In fact, the *frames of chirps* considered there are of the form

$$\Phi_{\text{chirps}} = \left\{\phi_\lambda(x) = \phi_{(k,\ell)}(x) = e^{2\pi i k x^2/N} e^{2\pi i \ell x/N},\ \lambda = (k,\ell) \in G \times \widehat{G}\right\}.$$

We have

$$\begin{aligned} \pi(k,\ell)\varphi_A(x) &= e^{2\pi i \ell x/N} e^{2\pi i (x-k)^3/N} = e^{2\pi i \ell x/N} e^{2\pi i (x^3 - 3x^2 k + 3xk^2 - k^3)/N} \\ &= e^{-2\pi i k^3/N} e^{2\pi i x^3/N} e^{2\pi i (\ell - k^2)x/N} e^{-2\pi i 3kx^2/N} \\ &= \overline{e^{2\pi i k^3/N} \varphi_A(x)} \phi_{(3k,\ell-k^2)}(x), \end{aligned}$$

and if N is not divisible by 3, then Φ_{chirps} is, aside from renumbering, the unitary image of a Gabor frame with an Alltop window. Hence, for N not divisible by 3, the coherence results on $(\varphi_A, G \times \widehat{G})$ are identical to the coherence results on Φ_{chirp}. Also, the restricted isometry constants (see Sect. 6.8) for $(\varphi_A, G \times \widehat{G})$ and Φ_{chirp} are identical for the same reason.

As an alternative to the Alltop sequence, J.J. Benedetto, R.L. Benedetto, and Woodworth used results from number theory such as Andre Weil's exponential sum

bounds to estimate the coherence of Gabor frames based on Björck sequences as Gabor window functions [8, 12, 13]. Note that any Björck sequence φ_B is a *constant amplitude zero autocorrelation* (CAZAC) sequence; therefore, we have

$$\langle T_k\varphi_B, \varphi_B\rangle = 0 = \langle M_\ell\varphi_B, \varphi_B\rangle, \quad (k,\ell) \in G \times \widehat{G}.$$

Accounting again for the zero entries in the CAZAC Gabor frame Gram matrices, we observe that the smallest achievable coherence is $1/\sqrt{|G|-1}$.

For $p \geq 5$ prime with $p = 1 \bmod 4$, the *Björck sequence* $\varphi_B \in \mathbb{C}^{\mathbb{Z}_p}$ is given by

$$\varphi_B(x) = \frac{1}{\sqrt{p}} \begin{cases} 1, & \text{for } x = 0, \\ e^{i \arccos(1/(1+\sqrt{p}))}, & x = m^2 \bmod p \text{ for some } m = 1, \ldots, p-1, \\ e^{-i \arccos(1/(1+\sqrt{p}))}, & \text{otherwise,} \end{cases}$$

and for $p \geq 3$ prime with $p = 3 \bmod 4$, we set

$$\varphi_B(x) = \frac{1}{\sqrt{p}} \begin{cases} e^{i \arccos((1-p)/(1+p))/p}, & x \neq m^2 \bmod p \text{ for all } m = 0, 1, \ldots, p-1, \\ 1, & \text{otherwise.} \end{cases}$$

Then [8]

$$\mu(\varphi_B, \mathbb{Z}_p \times \widehat{\mathbb{Z}_p}) < \frac{2}{\sqrt{p}} + \begin{cases} \frac{4}{p}, & p = 1 \bmod 4; \\ \frac{4}{p^{3/2}}, & p = 3 \bmod 4. \end{cases}$$

In comparison to (6.28) , the rounded values of $|V_{\varphi_B}\varphi_B(\lambda)| = |\langle \varphi_B, \pi(\lambda)\varphi_B\rangle|$, $\lambda \in \mathbb{Z}_N \times \widehat{\mathbb{Z}_N}$ for $N = 7$ are

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0.2955 & 0.3685 & 0.5991 & 0.1640 & 0.4489 & 0.4354 \\ 0 & 0.3685 & 0.1640 & 0.4354 & 0.2955 & 0.5991 & 0.4489 \\ 0 & 0.5991 & 0.4354 & 0.3685 & 0.4489 & 0.2955 & 0.1640 \\ 0 & 0.1640 & 0.2955 & 0.4489 & 0.3685 & 0.4354 & 0.5991 \\ 0 & 0.4489 & 0.5991 & 0.2955 & 0.4354 & 0.1640 & 0.3685 \\ 0 & 0.4354 & 0.4489 & 0.1640 & 0.5991 & 0.3685 & 0.2955 \end{pmatrix}.$$

To study the generic behavior of the coherence of Gabor systems $\mu(\varphi, \mathbb{Z}_N \times \widehat{\mathbb{Z}_N})$ for $N \in \mathbb{N}$, we turn to random windows. To this end, we let ϵ denote a random variable uniformly distributed on the torus $\{z \in \mathbb{C}, |z| = 1\}$. For $N \in \mathbb{N}$, we let φ_R be the random window function with entries

$$\varphi_R(x) = \frac{1}{\sqrt{N}}\epsilon_x, \quad x = 0, \ldots, N-1, \tag{6.29}$$

where the ϵ_x are independent copies of ϵ. In short, φ_R is a *normalized random Steinhaus sequence*.

For $N = 8$, the rounded values of $|V_{\varphi_R}\varphi_R(\lambda)|$, $\lambda \in \mathbb{Z}_N \times \widetilde{\mathbb{Z}_N}$, for a sample φ_R, are

$$\begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.1915 & 0.5266 & 0.3831 & 0.1418 & 0.1269 & 0.4575 & 0.5410 & 0.0341 \\ 0.0520 & 0.2736 & 0.2872 & 0.7912 & 0.2384 & 0.1880 & 0.0741 & 0.3411 \\ 0.3712 & 0.5519 & 0.2569 & 0.2757 & 0.5049 & 0.3123 & 0.2200 & 0.1215 \\ 0.0968 & 0.2423 & 0.6019 & 0.2632 & 0.1005 & 0.2632 & 0.6019 & 0.2423 \\ 0.3712 & 0.1215 & 0.2200 & 0.3123 & 0.5049 & 0.2757 & 0.2569 & 0.5519 \\ 0.0520 & 0.3411 & 0.0741 & 0.1880 & 0.2384 & 0.7912 & 0.2872 & 0.2736 \\ 0.1915 & 0.0341 & 0.5410 & 0.4575 & 0.1269 & 0.1418 & 0.3831 & 0.5266 \end{pmatrix}.$$

Here and in the following, $\mathbb{E}$ denotes expectation and $\mathbb{P}$ the probability of an event. In this context, a slight adjustment of the proof of Proposition 4.6 in [59] implies that, for p prime,

$$\mathbb{P}\big((\varphi_R, \mathbb{Z}_p \times \widehat{\mathbb{Z}_p}) \text{ is a unimodular tight frame maximally robust to erasures}\big) = 1.$$

The following result on the expected coherence of Gabor systems is given in [76]. Aside from the factor α, the coherence in Theorem 6.15 resembles with high probability the coherence $1/\sqrt{N}$ of the Alltop window and in this sense is close to the lower coherence bound $1/\sqrt{N+1}$.

Theorem 6.15 *Let $N \in \mathbb{N}$ and let φ_R be the random vector with entries*

$$\varphi_R(x) = \frac{1}{\sqrt{N}}\epsilon_x, \quad x = 0, \ldots, N-1, \tag{6.30}$$

where the ϵ_x are independent and uniformly distributed on the torus $\{z \in \mathbb{C}, |z| = 1\}$. Then for $\alpha > 0$ and N even,

$$\mathbb{P}\left(\mu(\varphi_R, \mathbb{Z}_N \times \widehat{\mathbb{Z}_N}) \geq \frac{\alpha}{\sqrt{N}}\right) \leq 4N(N-1)e^{-\alpha^2/4},$$

while for N odd,

$$\mathbb{P}\left(\mu(\varphi_R, \mathbb{Z}_N \times \widehat{\mathbb{Z}_N}) \geq \frac{\alpha}{\sqrt{N}}\right) \leq 2N(N-1)\big(e^{-\frac{N-1}{N}\alpha^2/4} + e^{-\frac{N+1}{N}\alpha^2/4}\big).$$

For example, a window $\varphi \in \mathbb{C}^{10,000}$ chosen according to (6.30) generates a Gabor frame with coherence less than $8.6/\sqrt{10,000} = 0.086$ with probability exceeding $10,000 \cdot 9,999 \cdot e^{-8.6^2/4} \approx 0.0671$. Note that our result does not guarantee the existence of a Gabor frame for $\mathbb{C}^{10,000}$ with coherence 0.085. The Alltop window, though, provides us with a Gabor frame for $\mathbb{C}^{9,973}$ with coherence ≈ 0.0100.

Proof The result is proven in full in [76]; here, we will simply give an outline of the proof in the case that N is even.

To estimate $\langle \varphi_R, \pi(\lambda)\varphi_R\rangle = \langle \varphi_R, M_\ell T_k \varphi_R\rangle$ for $\lambda = (k,\ell) \in G \times \widehat{G} \setminus \{0\}$, note first that if $k = 0$, then $\langle \varphi_R, M_\ell \varphi_R\rangle = \langle |\varphi_R|^2, M_\ell 1\rangle = 0$ for $\ell \neq 0$.

For the case $k \neq 0$, choose first $\omega_q \in [0,1)$ in $\epsilon_q = e^{2\pi i \omega_q}$ and observe that

$$\overline{\langle \varphi_R, \pi(\lambda)\varphi_R\rangle} = \langle \pi(\lambda)\varphi_R, \varphi_R\rangle = \frac{1}{N}\sum_{q\in G} e^{2\pi i \frac{q\ell}{N}} \epsilon_{q-p}\overline{\epsilon_q} = \frac{1}{N}\sum_{q\in G} e^{2\pi i(\omega_{q-p}-\omega_q+\frac{q\ell}{N})}.$$

The random variables

$$\delta_q^\lambda = e^{2\pi i(k_{q-p}-\omega_q+\frac{q\ell}{n})}$$

are uniformly distributed on the torus $\mathbb{T}$, but they are not jointly independent. As demonstrated in [76], these random variables can be split into two subsets of jointly independent random variables $\Lambda_1, \Lambda_2 \subseteq G$ with $|\Lambda_1| = |\Lambda_2| = N/2$.

The complex Bernstein inequality [97, Proposition 15], [73], implies that for an independent sequence $\epsilon_q, q = 0,\dots,N-1$, of random variables that are uniformly distributed on the torus, we have

$$\mathbb{P}\left(\left|\sum_{q=0}^{N-1}\epsilon_q\right| \geq Nu\right) \leq 2e^{-Nu^2/2}. \tag{6.31}$$

Using the pigeonhole principle and the inequality (6.31) leads to

$$\begin{aligned}\mathbb{P}\big(\big|\langle \pi(\lambda)\varphi_R, \varphi_R\rangle\big| \geq t\big) &\leq \mathbb{P}\left(\left|\sum_{q\in\Lambda^1}\delta_q^{(p,\ell)}\right| \geq Nt/2\right) + \mathbb{P}\left(\left|\sum_{q\in\Lambda^2}\delta_q^{(p,\ell)}\right| \geq Nt/2\right)\\ &\leq 4\exp\big(-Nt^2/4\big).\end{aligned}$$

Applying the union bound over all possible $\lambda \in G \times \widehat{G} \setminus \{(0,0)\}$ and choosing $t = \alpha/\sqrt{N}$ concludes the proof. □

Remark 6.1 A Gabor system (φ, Λ) which is in general linear position, which has small coherence, or which satisfies the restricted isometry property, is generally not useful for time-frequency analysis as described in Sect. 6.3. Recall that in order to obtain meaningful spectrograms of time-frequency localized signals, we chose windows which were well localized in time and in frequency; that is, we chose windows so that $V_\varphi\varphi(k,\ell) = \langle \varphi, \pi(k,\ell)\varphi\rangle$ was small for k,ℓ far from 0 (in the cyclic group $\mathbb{Z}_N$). To achieve a good coherence, though, we attempt to seek φ such that $V_\varphi\varphi(k,\ell)$ is close to being a constant function on all of the time-frequency plane.

To illustrate how inappropriate it is to use windows as discussed in Sects. 6.6–6.9, we perform in Fig. 6.9 the analysis carried out in Figs. 6.2–6.6 with a window chosen according to (6.30).

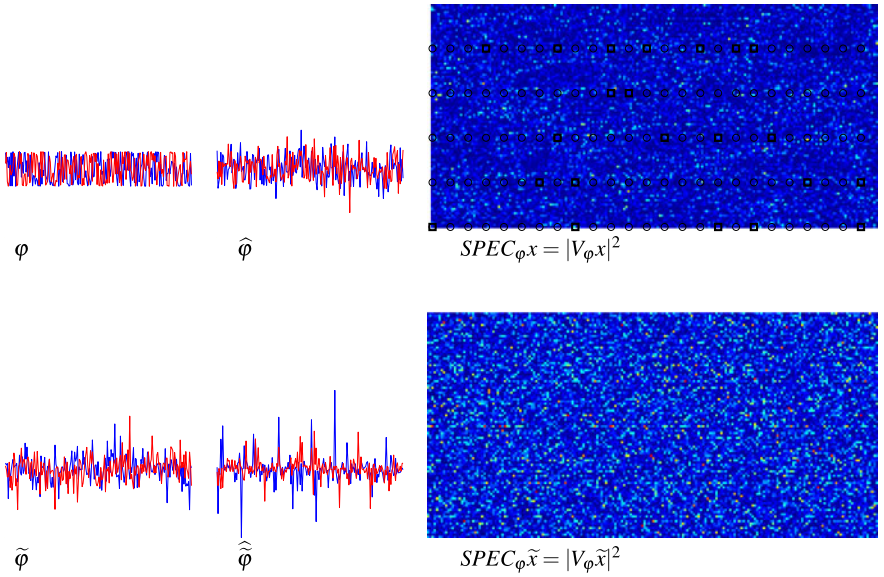


Fig. 6.9 We carry out the same analysis of the signal in Fig. 6.1 as in Figs. 6.2–6.6. The Gabor system uses as the window $\varphi = \varphi_R$ given in (6.30). The functions $\varphi, \widetilde{\varphi}$ are both not localized to an area in time or in frequency; in fact, this serves as an advantage in compressed sensing. We display again only the lower half of the spectrogram of x and of its approximation $\widetilde{x}$. Both are of little use. The lattice used is given by $\Lambda = \{0, 8, 16, \ldots, 192\} \times \{0, 20, 40, \ldots, 180\}$ and is marked by circles. Those of the 40 biggest frame coefficients in the part of the spectrogram shown are marked by *squares*

6.8 Restricted Isometry Constants

The coherence of a unit norm frame measures the smallest angle between two distinct elements of the frame. In the theory of compressed sensing, it is crucial to understand the geometry of subsets of frames that contain a small number of elements, but more than just two elements. The coherence of unit norm frames can be used to control the behavior of small subsets, but the compressed sensing results achieved in this manner are rather weak. To capture the geometry of small families of vectors, the concept of *restricted isometry constants* (RICs) has been developed. This leads to useful results in the area of compressed sensing [15, 16, 39, 82].

The *restricted isometry constant* $\delta_s(\Phi) = \delta_s$, $2 \le s \le N$, of a frame Φ of M elements in $\mathbb{C}^N$, is the smallest $0 < \delta_s < 1$ that satisfies

$$(1-\delta_s)\sum_{i=1}^{M}|c_i|^2 \le \left\|\sum_{i=1}^{M} c_i\varphi_i\right\|_2^2 \le (1+\delta_s)\sum_{i=1}^{M}|c_k|^2 \quad \text{for all } c \text{ with } \|c\|_0 \le s. \tag{6.32}$$

A simple computation shows that the coherence of a unit norm frame Φ satisfies $\mu(\Phi) = \delta_2(\Phi)$.

Statement (6.32) implies that every subfamily of s vectors forms a Riesz system with Riesz bounds $(1-\delta_s)$, $(1+\delta_s)$. In particular, the existence of a restricted isometry constant implies that any s vectors in Φ are linearly independent.

Frames with small restricted isometry constants for s sufficiently large are difficult to construct. A trick to bypass the problem of having to do an intricate study of all possible selections of s vectors from a frame Φ with M elements, $M \gg s$, is to introduce randomness in the definition of the frame. For example, if each component of each vector in a frame is generated independently by a fixed random process, then every family of s vectors is structured identically and the probability that a target δ_s fails can be estimated using a union bound argument.

To obtain results on restricted isometry constants of generic Gabor systems, we will choose again as window function φ_R, namely, the normalized random Steinhaus sequence defined in (6.29). The following is the main result in [77].

Theorem 6.16 *Let $G = \mathbb{Z}_N$ and let φ_R be a normalized Steinhaus sequence.*

1. *The expectation of the restricted isometry constant δ_s of $(\varphi_R, G \times \widehat{G})$, $s \le N$, satisfies*

$$\mathbb{E}\delta_s \le \max\left\{C_1\sqrt{\frac{s^{3/2}}{N}}\log s\sqrt{\log N},\ C_2\frac{s^{3/2}\log^{3/2} N}{N}\right\},$$

   where $C_1, C_2 > 0$ are universal constants.
2. *For $0 \le \lambda \le 1$, we have*

$$\mathbb{P}\big(\delta_s \ge \mathbb{E}[\delta_s] + \lambda\big) \le e^{-\lambda^2/\sigma^2}, \qquad \text{where } \sigma^2 = \frac{C_3 s^{\frac{3}{2}}\log N \log^2 s}{N}$$

   with $C_3 > 0$ being a universal constant.

The result remains true when generating the entries of φ by any Gaussian or sub-Gaussian random variable. In particular, the result holds true if the entries of φ are generated with a Bernoulli process; in this case, the Shannon entropy of the generated $N \times N^2$ matrix is remarkably small, namely, N bits. The bounds in Theorem 6.16 have been improved in [60].

6.9 Gabor Synthesis Matrices for Compressed Sensing

The problem of determining a signal in a high-dimensional space by combining *a priori* nonlinear information on a vector or on its Fourier transform with a small number of linear measurements appears frequently in the natural sciences and engineering. Here, we will address the problem of determining a vector $F \in \mathbb{C}^M$ by N linear measurements under the assumption that

$$\|F\|_0 = \big|\{n:\ F(n) \neq 0\}\big| \le s, \quad s \ll N \ll M.$$

This topic is treated in general terms in Chap. 9; we will focus entirely on the case where the linear measurements are achieved through the application of a Gabor frame synthesis matrix.

In detail, with T_φ^* denoting the $(\varphi, G \times \widehat{G})$ synthesis operator and

$$\Sigma_s = \left\{ F \in \mathbb{C}^{G \times \widehat{G}} : \ \|F\|_0 \leq s \right\},$$

we ask the question: For which s, can every vector $F \in \Sigma_s \subseteq \mathbb{C}^{G \times \widehat{G}}$ be recovered efficiently from

$$T_\varphi^* F = \sum_{\lambda \in G \times \widehat{G}} F_\lambda \pi(\lambda)\varphi \in \mathbb{C}^G ?$$

The problem of finding the sparse vector $F \in \Sigma_s$ from $T_\varphi^* F$ is identical to the problem of identifying $\mathscr{H}_s$ as defined in (6.23) from the observation of $H\varphi = \sum_{\lambda \in G \times \widehat{G}} \eta_\lambda \pi(\lambda)\varphi$. This holds since $\{\pi(\lambda)\}_{\lambda \in G \times \widehat{G}}$ is a linear independent set in the space of linear operators on $\mathbb{C}^G$, and, hence, the coefficient vector η is in one-to-one correspondence with the respective channel operator [76].

In addition, the problem at hand can be rephrased as follows. Suppose we know that a vector $x \in \mathbb{C}^G$ has the form $x = \sum_{\lambda \in \Lambda} c_\lambda \pi(\lambda)\varphi$ with $|\Lambda| \leq s$; that is, x is the linear combination of at most s frame elements from (φ, Λ). Can we compute the coefficients c_λ? Obviously, x can be expanded in $(\varphi, G \times \widehat{G})$ in many ways, for example, by using

$$x = \sum_{\lambda \in \Lambda} \langle x, \pi(\lambda)\widetilde{\varphi} \rangle \pi(\lambda)\varphi, \tag{6.33}$$

where $(\widetilde{\varphi}, \Lambda)$ is a dual frame of (φ, Λ). The coefficients in (6.33) are optimal in the sense that they have the lowest possible ℓ^2-norm. In this section, though, the goal is to find the expansion involving the fewest nonzero coefficients.

Theorem 6.10 implies that for $G = \mathbb{Z}_p$, p prime, there exists φ with the elements of $(\varphi, G \times \widehat{G})$ being in general linear position. Consequently, if $s \leq p/2$, then T_φ^* is injective on Σ_s and recovering F from $T_\varphi^* F$ is always possible, but this may not be computationally feasible, as every one of the $|G \times \widehat{G}|$ chosen s possible subsets of $G \times \widehat{G}$ sets of F would have to be considered as support sets of F.

To obtain a numerically feasible problem, we have to reduce s, and indeed, for small s, the literature contains a number of criteria on the measurement matrix M to allow the computation of F from MF by algorithms such as *basis pursuit* (BP) (see Sect. 9.2.2.1) and *orthogonal matching pursuit* (OMP) (see Sect. 9.2.2.2).

It is known that the success of BP and OMP for small s can be guaranteed if the coherence of the columns of a measurement matrix is small, in our setting, if $\mu(\varphi, G \times \widehat{G}) < 1/(2s-1)$ [27, 95]. In fact, combining this result with our coherence results in Sect. 6.7—in particular, the coherence of the Alltop frame $(\varphi_A, G \times \widehat{G})$ for $G = \mathbb{Z}_p$, p prime—leads to the following results [76].

Theorem 6.17 *Let $G = \mathbb{Z}_p$, p prime, and let φ_A be the Alltop window given in (6.27). If $s < \frac{\sqrt{p}+1}{2}$ then BP recovers F from $T^*_{\varphi_A} F$ for every $F \in \Sigma_s \subseteq G \times \widehat{G}$.*

In the case of Steinhaus sequences, Theorem 6.15 implies the following theorem [76].

Theorem 6.18 *Let $G = \mathbb{Z}_N$, N even. Let φ_R be the random unimodular window in (6.29). Let $t > 0$ and*

$$s \leq \frac{1}{4}\sqrt{\frac{N}{2\log N + \log 4 + t}} + \frac{1}{2}.$$

Then with probability $1 - e^{-t}$, BP recovers F from $T^*_{\varphi_R} F$ for every $F \in \Sigma_s$.

Note that in Theorems 6.17 and 6.18, the number of measurements N required to guarantee the recovery of every s-sparse vector scales as s^2. This can be improved if we are satisfied to recover an s-sparse vector with high probability [75].

Theorem 6.19 *Let $G = \mathbb{Z}_N$, $N \in \mathbb{N}$. There exists $C > 0$ so that whenever $s \leq CN/\log(N/\epsilon)$, the following holds: for $F \in \Sigma_s$ choose φ_R according to (6.30), then with probability at least $1 - \epsilon$ BP recovers F from $T^*_{\varphi_R} F$.*

Clearly, in Theorem 6.19 s scales as $N/\log(N)$, but we recover the vector F only with high probability.

The estimates on the restricted isometry constants in Theorem 6.16 imply in fact that with high probability the Gabor synthesis matrix $T^*_{\varphi_R}$ guarantees that BP recovers every $F \in \Sigma_s$ from $T^*_{\varphi_R} F$ if s is of the order $N^{2/3}/\log^2 N$ [77]. This follows from the fact that BP recovers $F \in \Sigma_s$ if $\delta_{2s}(\varphi_R, G \times \widehat{G}) \leq 3/(4 + \sqrt{6})$ [15, 17].

Numerical simulations show that the recoverability guarantees given above are rather pessimistic. In fact, the performance of Gabor synthesis matrices with Alltop window φ_A and with random window φ_R as measurement matrices seem to perform similarly well as, for example, random Gaussian matrices [76].

For related Gabor frame results aimed at recovering signals that are only well approximated by s-sparse vectors, see [75–77].

Acknowledgements The author acknowledges support under the *Deutsche Forschungsgemeinschaft* (DFG) grant 50292 DFG PF-4 (Sampling of Operators).

Parts of this paper were written during a sabbatical of the author at the Research Laboratory for Electronics and the Department of Mathematics at the Massachusetts Institut of Technology. He is grateful for the support and the stimulating research environment.

References

1. Alltop, W.O.: Complex sequences with low periodic correlations. IEEE Trans. Inf. Theory **26**(3), 350–354 (1980)

2. Applebaum, L., Howard, S.D., Searle, S., Calderbank, R.: Chirp sensing codes: deterministic compressed sensing measurements for fast recovery. Appl. Comput. Harmon. Anal. **26**(2), 283–290 (2009)
3. Balan, R., Casazza, P.G., Heil, C., Landau, Z.: Density, overcompleteness, and localization of frames. I: theory. J. Fourier Anal. Appl. **12**(2), 105–143 (2006)
4. Balan, R., Casazza, P.G., Heil, C., Landau, Z.: Density, overcompleteness, and localization of frames. II: Gabor systems. J. Fourier Anal. Appl. **12**(3), 307–344 (2006)
5. Balian, R.: Un principe d'incertitude fort en théorie du signal on en mécanique quantique. C. R. Acad. Sci. Paris **292**, 1357–1362 (1981)
6. Bastiaans, M.J., Geilen, M.: On the discrete Gabor transform and the discrete Zak transform. Signal Process. **49**(3), 151–166 (1996)
7. Benedetto, J.J.: Harmonic Analysis and Applications. Studies in Advanced Mathematics. CRC Press, Boca Raton (1997)
8. Benedetto, J.J., Benedetto, R.L., Woodworth, J.T.: Optimal ambiguity functions and Weil's exponential sum bound. J. Fourier Anal. Appl. **18**(3), 471–487 (2012)
9. Benedetto, J.J., Donatelli, J.J.: Ambiguity function and frame theoretic properties of periodic zero autocorrelation waveforms. In: IEEE J. Special Topics Signal Process, vol. 1, pp. 6–20 (2007)
10. Benedetto, J.J., Heil, C., Walnut, D.: Remarks on the proof of the Balian-Low theorem. Annali Scuola Normale Superiore, Pisa (1993)
11. Benedetto, J.J., Heil, C., Walnut, D.F.: Gabor systems and the Balian-Low theorem. In: Feichtinger, H., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications, pp. 85–122. Birkhäuser, Boston (1998)
12. Björck, G.: Functions of modulus one on $\mathbf{Z}_p$ whose Fourier transforms have constant modulus. In: A. Haar memorial conference, Vol. I, II, Budapest, 1985. Colloq. Math. Soc. János Bolyai, vol. 49, pp. 193–197. North-Holland, Amsterdam (1987)
13. Björck, G.: Functions of modulus 1 on Z_n whose Fourier transforms have constant modulus, and "cyclic n-roots". In: Recent Advances in Fourier Analysis and Its Applications, Il Ciocco, 1989. NATO Adv. Sci. Inst. Ser. C Math. Phys. Sci., vol. 315, pp. 131–140. Kluwer Academic, Dordrecht (1990)
14. Brigham, E. (ed.): The Fast Fourier Transform. Prentice Hall, Englewood Cliffs (1974)
15. Candès, E., Romberg, J., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
16. Candès, E., Tao, T.: Near optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
17. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. C. R. Math. Acad. Sci. Paris **346**(9–10), 589–592 (2008)
18. Casazza, P., Kovačević, J.: Equal-norm tight frames with erasures. Adv. Comput. Math. **18**(2–4), 387–430 (2003)
19. Casazza, P., Pfander, G.E.: Infinite dimensional restricted invertibility, preprint (2011)
20. Chiu, J., Demanet, L.: Matrix probing and its conditioning. SIAM J. Numer. Anal. **50**(1), 171–193 (2012)
21. Christensen, O.: Atomic decomposition via projective group representations. Rocky Mt. J. Math. **26**(4), 1289–1313 (1996)
22. Christensen, O., Feichtinger, H.G., Paukner, S.: Gabor analysis for imaging. In: Handbook of Mathematical METHODS in Imaging, vol. 3, pp. 1271–1307. Springer, Berlin (2010)
23. Cooley, J., Tukey, J.: An algorithm for the machine calculation of complex Fourier series. Math. Comput. **19**, 297–301 (1965)
24. Daubechies, I.: The wavelet transform, time-frequency localization and signal analysis. IEEE Trans. Inf. Theory **36**(5), 961–1005 (1990)
25. Daubechies, I.: Ten Lectures on Wavelets. CBMS-NSF Reg. Conf. Series in Applied Math. Society for Industrial and Applied Mathematics, Philadelphia (1992)
26. Demeter, C., Zaharescu, A.: Proof of the HRT conjecture for (2,2) configurations, preprint

27. Donoho, D.L., Elad, M.: Optimally sparse representations in general (non-orthogonal) dictionaries via ℓ^1 minimization. Proc. Natl. Acad. Sci. **100**, 2197–2202 (2002)
28. Donoho, D.L., Elad, M., Temlyakov, V.N.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inf. Theory **52**(1), 6–18 (2006)
29. Donoho, D.L., Stark, P.B.: Uncertainty principles and signal recovery. SIAM J. Appl. Math. **49**(3), 906–931 (1989)
30. Duffin, R.J., Schaeffer, A.C.: A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**(2), 341–366 (1952)
31. Evans, R.J., Isaacs, I.M.: Generalized Vandermonde determinants and roots of unity of prime order. Proc. Am. Math. Soc. **58**, 51–54 (1976)
32. Feichtinger, H.G., Gröchenig, K.: Theory and practice of irregular sampling. In: Benedetto, J.J., Frazier, M.W. (eds.) Wavelets: Mathematics and Applications. CRC Press, Boca Raton (1994)
33. Feichtinger, H.G., Gröchenig, K.: Gabor frames and time-frequency analysis of distributions. J. Funct. Anal. **146**(2), 464–495 (1996)
34. Feichtinger, H.G., Kozek, W.: Quantization of TF-lattice invariant operators on elementary LCA groups. In: Feichtinger, H.G., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications, pp. 233–266. Birkhäuser, Boston (1998)
35. Feichtinger, H.G., Kozek, W., Luef, F.: Gabor analysis over finite abelian groups. Appl. Comput. Harmon. Anal. **26**(2), 230–248 (2009)
36. Feichtinger, H.G., Luef, F.: Wiener amalgam spaces for the Fundamental Identity of Gabor Analysis. Collect. Math. **57**, 233–253 (2006) (Extra Volume)
37. Feichtinger, H.G., Strohmer, T., Christensen, O.: A group-theoretical approach to Gabor analysis. Opt. Eng. **34**(6), 1697–1704 (1995)
38. Folland, G., Sitaram, A.: The uncertainty principle: a mathematical survey. J. Fourier Anal. Appl. **3**(3), 207–238 (1997)
39. Fornasier, M., Rauhut, H.: Compressive sensing. In: Scherzer, O. (ed.) Handbook of Mathematical Methods in Imaging, pp. 187–228. Springer, Berlin (2011)
40. Frenkel, P.: Simple proof of Chebotarevs theorem on roots of unity, preprint (2004). math.AC/0312398
41. Gabor, D.: Theory of communication. J. IEE, London **93**(3), 429–457 (1946)
42. Ghobber, S., Jaming, P.: On uncertainty principles in the finite dimensional setting. Linear Algebra Appl. **435**(4), 751–768 (2011)
43. Golomb, S., Gong, G.: Signal Design for Good Correlation: For Wireless Communication, Cryptography, and Radar. Cambridge University Press, Cambridge (2005)
44. Gribonval, R., Vandergheynst, P.: On the exponential convergence of matching pursuits in quasi-incoherent dictionaries. IEEE Trans. Inf. Theory **52**(1), 255–261 (2006)
45. Gröchenig, K.: Aspects of Gabor analysis on locally compact abelian groups. In: Feichtinger, H., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications, pp. 211–231. Birkhäuser, Boston (1998)
46. Gröchenig, K.: Foundations of Time-Frequency Analysis. Applied and Numerical Harmonic Analysis. Birkhäuser, Boston (2001)
47. Gröchenig, K.: Uncertainty principles for time-frequency representations. In: Feichtinger, H., Strohmer, T. (eds.) Advances in Gabor Analysis, pp. 11–30. Birkhäuser, Boston (2003)
48. Gröchenig, K.: Localization of frames, Banach frames, and the invertibility of the frame operator. J. Fourier Anal. Appl. **10**(2), 105–132 (2004)
49. Heil, C.: History and evolution of the density theorem for Gabor frames. J. Fourier Anal. Appl. **12**, 113–166 (2007)
50. Heil, C., Ramanathan, J., Topiwala, P.: Linear independence of time–frequency translates. Proc. Amer. Math. Soc. **124**(9), 2787–2795 (1996)
51. Howard, S.D., Calderbank, A.R., Moran, W.: The finite Heisenberg-Weyl groups in radar and communications. EURASIP J. Appl. Signal Process. (Frames and overcomplete representations in signal processing, communications, and information theory), Art. ID 85,685, 12 (2006)

52. Jaillet, F., Balazs, P., Dörfler, M., Engelputzeder, N.: Nonstationary Gabor Frames. In: SAMPTA'09, Marseille, May 18–22. ARI; Gabor; NuHAG; NHG-coop (2009)
53. Janssen, A.J.E.M.: Gabor representation of generalized functions. J. Math. Anal. Appl. **83**, 377–394 (1981)
54. Janssen, A.J.E.M.: Duality and biorthogonality for Weyl-Heisenberg frames. J. Fourier Anal. Appl. **1**(4), 403–436 (1995)
55. Janssen, A.J.E.M.: From continuous to discrete Weyl-Heisenberg frames through sampling. J. Fourier Anal. Appl. **3**(5), 583–596 (1997)
56. Kaiblinger, N.: Approximation of the Fourier transform and the dual Gabor window. J. Fourier Anal. Appl. **11**(1), 25–42 (2005)
57. Katznelson, Y.: An Introduction to Harmonic Analysis. Dover, New York (1976)
58. Keiner, J., Kunis, S., Potts, D.: Using NFFT 3—a software library for various nonequispaced fast Fourier transforms. ACM Trans. Math. Softw. **36**(4), 30 (2009), Art. 19
59. Krahmer, F., Pfander, G.E., Rashkov, P.: Uncertainty in time-frequency representations on finite abelian groups and applications. Appl. Comput. Harmon. Anal. **25**(2), 209–225 (2008)
60. Krahmer, F., Mendelson, S., Rauhut, H.: Suprema of chaos processes and the restricted isometry property (2012)
61. Landau, H.: Necessary density conditions for sampling an interpolation of certain entire functions. Acta Math. **117**, 37–52 (1967)
62. Landau, H.: On the density of phase-space expansions. IEEE Trans. Inf. Theory **39**(4), 1152–1156 (1993)
63. Landau, H., Pollak, H.: Prolate spheroidal wave functions, Fourier analysis and uncertainty. II. Bell Syst. Tech. J. **40**, 65–84 (1961)
64. Lawrence, J., Pfander, G.E., Walnut, D.F.: Linear independence of Gabor systems in finite dimensional vector spaces. J. Fourier Anal. Appl. **11**(6), 715–726 (2005)
65. Li, S.: Discrete multi-Gabor expansions. IEEE Trans. Inf. Theory **45**(6), 1954–1967 (1999)
66. Low, F.: Complete sets of wave packets. In: DeTar, C. (ed.) A Passion for Physics—Essay in Honor of Geoffrey Chew, pp. 17–22. World Scientific, Singapore (1985)
67. Lyubarskii, Y.I.: Frames in the Bargmann space of entire functions. Adv. Sov. Math. **429**, 107–113 (1992)
68. Manjunath, B.S., Ma, W.: Texture features for browsing and retrieval of image data. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI—Special issue on Digital Libraries) **18**(8), 837–842 (1996)
69. Matusiak, E., Özaydın, M., Przebinda, T.: The Donoho-Stark uncertainty principle for a finite abelian group. Acta Math. Univ. Comen. (N.S.) **73**(2), 155–160 (2004)
70. von Neumann, J.: Mathematical Foundations of Quantum Mechanics. Princeton University Press, Princeton (1932), (1949) and (1955)
71. Orr, R.: Derivation of the finite discrete Gabor transform by periodization and sampling. Signal Process. **34**(1), 85–97 (1993)
72. Pei, S.C., Yeh, M.H.: An introduction to discrete finite frames. IEEE Signal Process. Mag. **14**(6), 84–96 (1997)
73. Pevskir, G., Shiryaev, A.N.: The Khintchine inequalities and martingale expanding sphere of their action. Russ. Math. Surv. **50**(5), 849–904 (1995)
74. Pfander, G.E.: Note on sparsity in signal recovery and in matrix identification. Open Appl. Math. J. **1**, 21–22 (2007)
75. Pfander, G.E., Rauhut, H.: Sparsity in time-frequency representations. J. Fourier Anal. Appl. **16**(2), 233–260 (2010)
76. Pfander, G.E., Rauhut, H., Tanner, J.: Identification of matrices having a sparse representation. IEEE Trans. Signal Process. **56**(11), 5376–5388 (2008)
77. Pfander, G.E., Rauhut, H., Tropp, J.A.: The restricted isometry property for time-frequency structured random matrices. Probab. Theory Relat. Fields (to appear)
78. Pfander, G.E., Walnut, D.: Measurement of time-variant channels. IEEE Trans. Inf. Theory **52**(11), 4808–4820 (2006)

79. Qiu, S.: Discrete Gabor transforms: the Gabor-Gram matrix approach. J. Fourier Anal. Appl. **4**(1), 1–17 (1998)
80. Qiu, S., Feichtinger, H.: Discrete Gabor structure and optimal representation. IEEE Trans. Signal Process. **43**(10), 2258–2268 (1995)
81. Rao, K.R., Kim, D.N., Hwang, J.J.: Fast Fourier Transform: Algorithms and Applications. Signals and Communication Technology. Springer, Dordrecht (2010)
82. Rauhut, H.: Compressive sensing and structured random matrices. In: Fornasier, M. (ed.) Theoretical Foundations and Numerical Methods for Sparse Recovery. Radon Series Comp. Appl. Math., vol. 9, pp. 1–92. de Gruyter, Berlin (2010)
83. Ron, A., Shen, Z.: Weyl-Heisenberg frames and Riesz bases in $\ell_2(\mathbb{R}^d)$. Tech. Rep. 95-03, University of Wisconsin, Madison (WI) (1995)
84. Rudin, W.: Fourier Analysis on Groups. Interscience Tracts in Pure and Applied Mathematics, vol. 12. Interscience Publishers (a division of John Wiley & Sons), New York–London (1962)
85. Seip, K., Wallstén, R.: Density theorems for sampling and interpolation in the Bargmann-Fock space. I. J. Reine Angew. Math. **429**, 91–106 (1992)
86. Seip, K., Wallstén, R.: Density theorems for sampling and interpolation in the Bargmann-Fock space. II. J. Reine Angew. Math. **429**, 107–113 (1992)
87. Skolnik, M.: Introduction to Radar Systems. McGraw-Hill, New York (1980)
88. Slepian, D., Pollak, H.O.: Prolate spheroidal wave functions, Fourier analysis and uncertainty. I. Bell Syst. Tech. J. **40**, 43–63 (1961)
89. Soendergaard, P.L., Torresani, B., Balazs, P.: The linear time frequency analysis toolbox. Int. J. Wavelets Multi. 10(4), 27 pp.
90. Søndergaard, P.L.: Gabor frames by sampling and periodization. Adv. Comput. Math. **27**(4), 355–373 (2007)
91. Strohmer, T.: Numerical algorithms for discrete Gabor expansions. In: Feichtinger, H., Strohmer, T. (eds.) Gabor Analysis and Algorithms: Theory and Applications, pp. 267–294. Birkhäuser, Boston (1998)
92. Strohmer, T., Heath, R.W. Jr.: Grassmannian frames with applications to coding and communication. Appl. Comput. Harmon. Anal. **14**(3), 257–275 (2003)
93. Tao, T.: An uncertainty principle for groups of prime order. Math. Res. Lett. **12**, 121–127 (2005)
94. Terras, A.: Fourier Analysis on Finite Groups and Applications. London Mathematical Society Student Texts, vol. 43. Cambridge University Press, Cambridge (1999)
95. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inf. Theory **50**(10), 2231–2242 (2004)
96. Tropp, J.A.: Just relax: convex programming methods for identifying sparse signals. IEEE Trans. Inf. Theory **51**(3), 1030–1051 (2006)
97. Tropp, J.A.: On the conditioning of random subdictionaries. Appl. Comput. Harmon. Anal. **25**(1), 1–24 (2008)
98. Wexler, J., Raz, S.: Discrete Gabor expansions. Signal Process. **21**(3), 207–221 (1990)
99. Xia, X.G., Qian, S.: On the rank of the discrete Gabor transform matrix. Signal Process. **52**(3), 1083–1087 (1999)

Chapter 7
Frames as Codes

Bernhard G. Bodmann

Abstract This chapter reviews the development of finite frames as codes for erasures and additive noise. These types of errors typically occur when analog signals are transmitted in an unreliable environment. The use of frames allows one to recover the signal with controllable accuracy from part of the encoded, noisy data. While linear binary codes have a long history in information theory, frames as codes over the real or complex numbers have only been examined since the 1980s. In the encoding process, a vector in a finite-dimensional real or complex Hilbert space is mapped to the sequence of its inner products with frame vectors. An erasure occurs when part of these frame coefficients is no longer accessible after the transmission. Additive noise can arise from the encoding process, such as when coefficients are rounded, or from the transmission. This chapter covers two of the most popular recovery algorithms: blind reconstruction, where missing coefficients are set to zero, and active error correction, which aims to recover the signal perfectly based on the known coefficients. The erasures can be modeled as either having a deterministic or a random occurrence pattern. In the deterministic regime it has been customary to optimize the frame performance in the worst-case scenario. Optimality for a small number of erasures then leads to geometric conditions such as the class of equiangular tight frames. Random erasure models are often used in conjunction with performance measures based on averaged reconstruction errors, such as the mean-squared error. Frames as codes for erasures are also closely related to recent results on sparse recovery. Finally, fusion frames and packet erasures introduce an additional structure which imposes constraints on the construction of optimal frames.

Keywords Frames · Parseval frames · Codes · Erasures · Worst-case error · Mean-squared error · Random frames · Protocols · Fusion frames · Packet erasures · Equi-isoclinic fusion frames · Equidistance fusion frames

B.G. Bodmann (✉)
Mathematics Department, University of Houston, 651 Philip G. Hoffman Hall, Houston, TX 77204-3008, USA
e-mail: bgb@math.uh.edu

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_7,

7.1 Introduction

Digital signal communications are omnipresent, ranging from cell phone transmissions to streaming media such as Voice over Internet Protocol telephony, satellite radio, or Internet TV. In principle, digital error correction protocols can guarantee nearly faultless transmissions in the presence of noise and data loss. Much of the development of these protocols is inspired by Shannon's seminal work of more than 60 years ago [42–44], in which he founded the theory of transmitting data through unreliable analog channels. However, today we typically face a problem outside of Shannon's immediate concern: transmitting analog data such as audio or video through a somewhat unreliable *digital* channel, the Internet. After the error of converting from analog to digital, network outages and buffer overflows are the main problem in digital transmissions. This means that the typical reconstruction error is composed of additive noise created in the digitization and partial data loss from the transmission. Another difference between Shannon's communication theory and today's practice is that latency issues are not part of computing Shannon's channel capacity, whereas for cell phones or Voice over Internet Protocol telephony, they become essential concerns. The simplest way to control latency is to work with block codes of a given size. The question is then to what extent imperfections in the transmission can be suppressed. This topic, called rate-distortion theory, was developed by Shannon for digital transmissions. His work pioneered the tandem strategy of digitization and subsequent channel coding, which allowed for distortion measures in the digital as well as in the analog domain. In the latter case, it is natural to consider an alternative to the tandem strategy by adding redundancy at the analog level [38, 39]. If the encoding is linear at the analog level, then this amounts to using frames as codes. Simply put, frames act as *block codes*, which replace the coefficients of a vector with respect to an orthonormal basis by a larger number of linear coefficients in the expansion with a stable spanning set, thereby incorporating redundancy in the representation and providing error suppression capabilities.

This strategy has been applied to suppress quantization errors, meaning the rounding of frame coefficients, see, e.g., [2–4, 8, 9, 12], and erasures and data loss in the course of transmission [7, 17, 28–30, 32, 34, 45, 47]. The generalization of these results to the suppression of errors due to lost *packets* in frame-based encoding was studied in [5, 19, 35, 40]. This model assumes that frame coefficients are partitioned into subsets, often taken to be of equal size, and if an erasure occurs, then it makes the contents of an entire subset of coefficients inaccessible. For a related problem of robust nonorthogonal subspace decompositions, the concept of frames for subspaces was introduced by Casazza and Kutyniok [18], later referred to as fusion frames in an application to distributed processing [20].

Finally, correcting erasures has also been discussed in the context of sparse representations and compressed sensing [14, 15], building on groundbreaking works by Donoho, Stark, and Huo [22, 23]. Although there are probabilistic proofs for the existence of good frames for recovery, even if a (sufficiently small) fraction of frame coefficients is lost and a certain level of noise is added to the remaining ones, there is currently no deterministic construction with matching error correction capabilities [13, 16]. In the fusion frame literature, there are even more open problems,

in particular the construction of optimal fusion frames in large dimensions. We can only cover a few aspects of this exciting topic, which have been chosen to present a consistent theme throughout this chapter. Unfortunately, recent results on optimal dual frames [37] and on structured erasures [11] could not be covered here.

Section 7.2 compiles the material on frames for suppressing erasures and additive noise. This includes a discussion of different performance measures, hierarchical and generic error models, and a discussion of random matrices for coding. Section 7.3 repeats some of the discussion in the context of fusion frames and generalizes the characterization of optimality with a more differentiated error measure.

7.2 Frames for the Encoding of Analog Data

A finite frame $\Phi = \{\varphi_j\}_{j=1}^M$ is a spanning family of vectors in an N-dimensional real or complex Hilbert space $\mathcal{H}$. If the Parseval-type identity

$$\|x\|^2 = \frac{1}{A}\sum_{j=1}^{M} \left|\langle x, \varphi_j\rangle\right|^2$$

is true for all $x \in \mathcal{H}$ with some constant $A > 0$, then Φ is called *A-tight*. If A=1, then we say that Φ is a *Parseval frame*. In this case, we also call it an (M, N)-frame, in analogy with the literature on block codes. The analysis operator of a frame Φ is the map $T : \mathcal{H} \to \ell^2(\{1, 2, \ldots, M\})$, $(Tx)_j = \langle x, \varphi_j\rangle$. If Φ is a Parseval frame, then T is an isometry. Frames are often classified by geometric properties: If all the frame vectors have the same norm, then the frame is called *equal norm*. If the frame is tight and there is $c \geq 0$ such that for all $j \neq l$, $|\langle \varphi_j, \varphi_l\rangle| = c$, then the frame is called *equiangular and tight*. The significance of the geometric characteristics of frames is that they are related to optimality of frame designs in certain situations. This will be reviewed in the following material.

The general model for frame-coded transmissions contains three parts: (1) the linear encoding of a vector in terms of its frame coefficients, (2) the transmission which may alter the frame coefficients, and (3) the reconstruction algorithm. The input vectors to be transmitted can either be assumed to have some distribution, or one can attempt to minimize the reconstruction error among all possible inputs of a given norm. The same can be applied to the errors occurring in the transmission. Our discussion restricts the treatment of input vectors to the worst-case scenario, or to the average over the uniform probability distribution among all unit-norm input vectors. The channel models are taken to be either the worst case or a uniform erasure distribution, possibly together with the addition of independently distributed random components to the frame coefficients which model the digitization noise for the input. We refer the reader to a more detailed treatment of the digitization errors in [2–4, 8, 9, 12] or in the chapter on quantization of this book. Among all possible reconstruction algorithms, we concentrate on linear ones, which may or may not depend on the type of error that occurred in the course of the transmission.

7.2.1 Frames for Erasure Channels

A standard assumption in network models is that a *sequence* of vectors is transmitted in the form of their frame coefficients. These coefficients are sent in parallel streams to the receiver; see [29, Example 1.1] and [34]. If one of the nodes in the network experiences a buffer overflow or a wireless outage, then the streams passing through this node are corrupted. The integrity of each coefficient in a transmission is typically protected by some error correction scheme, so for practical purposes one may assume that coefficients passing through the affected node are not used in the reconstruction process. The linear reconstruction of a vector from a subset of its frame coefficients amounts to setting the lost coefficients to zero; this is called an erasure error. This type of error has been the subject of many works [7, 28–30, 32, 34]. In our formulation, the encoded vector is given by its frame coefficients Tx, and the erasure acts by applying a diagonal projection matrix E to Tx, before linear reconstruction is attempted.

We can either reconstruct by an erasure-dependent linear transform, performing active error correction, or use blind reconstruction, which ignores the fact that some coefficients have been set to zero. For now, we focus on the second alternative and add some comments about active error correction later.

Definition 7.1 Let Φ be an (M, N)-frame for a real or complex Hilbert space $\mathcal{H}$, with analysis operator T. The *blind reconstruction error* for an input vector $x \in \mathcal{H}$ and an erasure of frame coefficients with indices $\mathbb{K} = \{j_1, j_2, \ldots, j_m\} \subset \mathbb{J} = \{1, 2, \ldots, M\}$, $m \leq M$, is given by

$$\left\| T^* E_{\mathbb{K}} T x - x \right\| = \left\| \left(T^* E_{\mathbb{K}} T - I \right) x \right\| = \left\| T^* (I - E_{\mathbb{K}}) T x \right\|$$

where E is the diagonal $M \times M$ matrix with $E_{j,j} = 0$ if $j \in \mathbb{K}$ and $E_{j,j} = 1$ otherwise. The residual error after performing active error correction is defined as $\|W E_{\mathbb{K}} T x - x\|$, where W is the Moore-Penrose pseudoinverse of $E_{\mathbb{K}} T$. If $W E_{\mathbb{K}} T = I$, then we say that the erasure of coefficients indexed by $\mathbb{K}$ is *correctable*.

Depending on the type of input and transmission model, the performance of a frame can be measured in deterministic or probabilistic ways. One measure is the worst case for the reconstruction, which is the maximal error norm among all reconstructed vectors. Since the error is proportional to the norm of the input vector, the operator norm $\|T^*(I - E_{\mathbb{K}})T\|$ can be chosen as a measure for the worst-case error among all normalized inputs [7, 17, 32, 33]. Another possibility is a statistical performance measure such as the mean-squared error, where the average is either over unit-norm input vectors for specific erasures or over the combination of such input vectors and random erasures. We combine these performance measures in a unified notation; see, e.g., [11].

Definition 7.2 Let $\mathbb{S}$ be the unit sphere in a real or complex N-dimensional Hilbert space $\mathcal{H}$, and let $\Omega = \{0, 1\}_{j=1}^{M}$ be the space of binary sequences of length M. Given

a binary sequence $\omega = \{\omega_1, \omega_2, \ldots, \omega_M\}$, we let the associated operator $E(\omega)$ be the diagonal $M \times M$ matrix with $E(\omega)_{j,j} = \omega_j$ for all $j \in \mathbb{J} = \{1, 2, \ldots, M\}$. Let μ be a probability measure on the space $\mathbb{S} \times \Omega$, which is the product of the uniform probability measure on $\mathbb{S}$ and a probability measure on Ω. The p-th power error norm is given by

$$e_p(\Phi, \mu) = \left(\int_{\mathbb{S} \times \Omega} \left\| T^* E(\omega) T x - x \right\|^p d\mu(x, \omega) \right)^{1/p}$$

with the understanding that when $p = \infty$ it is the usual sup-norm. The quantity $e_\infty(\Phi, \mu)$ has also been called the worst-case error norm and $e_2(\Phi, \mu)^2$ is commonly referred to as the mean-squared error.

We conclude this section with remarks concerning the relationship between passive and active error correction for erasures when $p = \infty$. In principle, active error correction either results in perfect reconstruction or in an error that can only be controlled by the norm of the input, because $W E_{\mathbb{K}} T$ is an orthogonal projection if W is the Moore-Penrose pseudoinverse of $E_{\mathbb{K}} T$. This may make it seem as if the only relevant question for active error correction is whether $E_{\mathbb{K}} T$ has a left inverse.

However, even in cases where an erasure is correctable, numerical stability against roundoff errors and other additive noise is desirable. This will be examined in more detail in Sect. 7.2.3. We prepare the discussion there by a comparison of an error measure based on the Moore-Penrose pseudoinverse with the p-th power error norm. It turns out that if all erasures in Ω are correctable, then achieving optimality with respect to e_∞ is equivalent to minimizing the maximal operator norm among all Moore-Penrose pseudoinverses of $E(\omega)T$, $\omega \in \Omega$.

Definition 7.3 Let $\mathbb{J}$, $\mathbb{S}$, and Ω be as above, and let ν be the uniform probability measure on $\mathbb{S} \times \Omega$. Let $W(\omega)$ be the Moore-Penrose pseudoinverse of $E(\omega)T$; then we define

$$a_p(\Phi, \nu) = \left(\int_{\mathbb{S} \times \Omega} \left\| W(\omega) y \right\|^p d\nu(y, \omega) \right)^{1/p}$$

with the understanding that when $p = \infty$ it is the usual sup-norm.

Proposition 7.1 *Let $\mathcal{H}$ be an N-dimensional real or complex Hilbert space. For any set of erasures $\Gamma \subset \Omega$, let μ_Γ be the probability measure which is invariant with respect to the product of unitaries and permutations on $\mathbb{S} \times \Gamma$. Similarly, let ν be the probability measure on the unit sphere of $\ell^2(\{1, 2, \ldots, M\}) \times \Gamma$, which is invariant with respect to the product of unitaries and permutations. If all erasures in Γ are correctable for a closed subset $\mathcal{S}$ of (M, N)-frames, then a frame Φ achieves the minimal worst-case error norm $e_\infty(\Phi, \mu) = \min_{\Psi \in \mathcal{S}} e_\infty(\Psi, \mu)$ if and only if it achieves the minimum $a_\infty(\Phi, \nu) = \min_{\Psi \in \mathcal{S}} a_\infty(\Psi, \nu)$.*

Proof Let us fix an erasure E corresponding to a choice of $\omega \in \Gamma$. Given an isometry T (analysis operator of a Parseval frame), the left inverse to T with smallest

operator norm is the (Hilbert) adjoint T^*. Given a Parseval frame and a diagonal projection E, then the operator norm of $T^*ET - I$ is the largest eigenvalue of $I - T^*ET$, because $T^*ET - I = T^*(E - I)T$ is negative definite. Factoring ET by polar decomposition into $VA = ET$, where A is nonnegative and V is an isometry, so the operator norm of $A^{-1}T^*$ is $\|A^{-1}V^*VA^{-1}\|^{1/2} = \|A^{-2}\|^{1/2} = \|A^{-1}\|$, by positivity of A the inverse of the smallest eigenvalue of A, $a_{\min}$.

This means that minimizing a_∞ with a fixed set of erasures amounts to maximizing the smallest eigenvalue appearing among the set of operators $\{A(\omega) : \omega \in \Gamma\}$. Comparing this with the error for blind reconstruction gives

$$\big\|\big(T^*ET - I\big)\big\| = \big\|I - A^*V^*VA\big\| = 1 - a_{\min}^2.$$

Minimizing this error over Γ also amounts to maximizing the smallest eigenvalue. Thus, the minimization of e_∞ or a_∞ for a fixed set of erasures Γ is equivalent. □

7.2.1.1 Hierarchical error models

Often it is assumed that losing one coefficient in the transmission process is rare, and that the occurrence of two lost coefficients is much less likely. A similar hierarchy of probabilities usually holds for a higher number of losses. This motivates the design of frames following an inductive scheme: We require perfect reconstruction when no data is lost. Among the protocols giving perfect reconstruction, we want to minimize the maximal error in the case of one lost coefficient. Generally, we continue by choosing among the frames which are optimal for m erasures those performing best for $m + 1$ erasures. For an alternative approach, which does not assume a hierarchy of errors, see maximally robust encoding [41] and the section on random Parseval frames later in this chapter.

Definition 7.4 Let $\mathcal{H}$ be an N-dimensional real or complex Hilbert space. We denote by $\mathcal{F}(M, N)$ the set of all (M, N)-frames, equipped with the natural topology from $\mathcal{H}^M$. Using as Γ_m the set of all m-erasures, $\Gamma_m = \{\omega \in \Omega : \sum_{j=1}^M \omega_j = m\}$, we let μ_m denote the product of uniform probability measures on $\mathbb{S} \times \Gamma_m$. We let $e_p^{(1)}(M, N) = \min\{e_p(\Phi, \mu_1) : \Phi \in \mathcal{F}(M, N)\}$ and $\mathcal{E}_p^{(1)}(M, N) = \{\Phi \in \mathcal{F}(M, N) : e_p(\Phi, \mu_m) = e_p^{(1)}(M, N)\}$. Proceeding inductively, we set for $1 \le m \le M$, $e_p^{(m)}(M, N) = \min\{e_m^p(\Phi, \mu_m) : \Phi \in \mathcal{E}_p^{(m-1)}(M, N)\}$ and define the optimal *m-erasure frames for* e_p to be the nonempty compact subset $\mathcal{E}_p^{(m)}(M, N)$ of $\mathcal{E}_p^{(m-1)}(M, N)$ where the minimum of $e_p^{(m)}$ is attained.

In this manner, we obtain a decreasing family of frames which can be characterized in a geometric fashion. Results by Casazza and Kovačević [17] as reviewed in [32, Proposition 2.1] and slightly extended in [7] can be interpreted as the statement that, among all Parseval frames, the equal-norm ones minimize the worst-case reconstruction error for one erasure.

Proposition 7.2 *For $1 < p \leq \infty$, the set $\mathcal{E}_p^{(1)}(M,N)$ coincides with the family of equal-norm (M,N)-frames. Consequently, for $1 < p \leq \infty$, $e_p^{(1)}(M,N) = N/M$.*

Proof Given an (M,N)-frame $\Phi = \{\varphi_1, \ldots, \varphi_M\}$ with analysis operator T, and a diagonal projection matrix D with one nonzero entry $D_{j,j}$, then $\|T^*DT\| = \|DTT^*D\| = \|\varphi_j\|^2$. If Φ is a Parseval frame, then $\sum_{j=1}^M \|f\|^2 = \operatorname{tr} TT^* = \operatorname{tr} T^*T = trI_N = N$, so minimizing the maximum norm among all frame vectors is achieved if and only if they all have the same norm. In this case, $\|\varphi_j\|^2 = N/M$, and thus $e_p^{(1)}(M,N) = N/M$ for any $p > 1$. □

Strohmer and Heath as well as Holmes and Paulsen [32, 45] showed that when they exist, equiangular Parseval frames are optimal for up to two erasures with respect to $e_\infty^{(2)}$. As stated by Holmes and Paulsen, if Φ is an equiangular (M,N)-frame, then TT^* is a self-adjoint rank-N projection that can be written in the form $TT^* = aI + c_{M,N}Q$ where $a = N/M$, $c_{M,N} = (\frac{N(M-N)}{M^2(M-1)})^{1/2}$, and the signature matrix $Q = (Q_{i,j})$ is a self-adjoint $M \times M$ matrix satisfying $Q_{i,i} = 0$ for all i and for $i \neq j$, $|Q_{i,j}| = 1$. The proof of optimality uses that if D is a diagonal projection matrix with a 1 in the i-th and j-th diagonal entries and T is the analysis operator for an equal norm (M,N)-frame $\Phi = \{\varphi_1, \ldots, \varphi_M\}$, then $\|T^*DT\| = \|DTT^*D\| = N/M + |\langle \varphi_i, \varphi_j \rangle|$. Since $\sum_{j \neq l} |\langle \varphi_j, \varphi_l \rangle|^2 = \operatorname{tr}[(TT^*)^2] - \sum_{j=1}^M (TT^*)_{j,j}^2 = N - N^2/M$, the maximum magnitude among all the inner products cannot be below the average value, which gives a lower bound for the worst-case 2-erasure. This bound is saturated if and only if all inner products have the same magnitude. Welch had established this inequality for unit-norm vector sequences [48].

The characterization of equiangular Parseval frames as optimal 2-erasure frames was extended to all sufficiently large values of p in [7].

Theorem 7.1 [7] *If equiangular frames exist among the equal norm (M,N)-frames and if $p > 2 + (\frac{5N(M-1)}{M-N})^{1/2}$, then $\mathcal{E}_p^{(2)}(M,N)$ consists precisely of these equiangular frames.*

The existence of such equiangular Parseval frames for real Hilbert spaces depends on the existence of a matrix of ± 1's which satisfies certain algebraic equations. Thanks to the discovery in [45] of the connection between equiangular frames and the earlier work of Seidel and his collaborators in graph theory, much of the work on existence and construction of real equiangular tight frames benefits from known techniques. The construction of equiangular Parseval frames in the complex case was investigated with number-theoretic tools, see [49] and [33], as well as with a numerical scheme [46]. Recently, Seidel's combinatorial approach was extended to the complex case by considering signature matrices whose entries are roots of unity [6, 10, 31].

An averaging argument similar to the inequality by Welch was derived for the case of 3 erasures [5], in the context of fusion frames. We present the consequences for the special case of equiangular Parseval frames.

Theorem 7.2 *Let $M \geq 3$, $M > N$ and let Φ be an equiangular (M, N)-frame; then*

$$e_\infty^{(3)}(M, N) \geq \frac{N}{M} + 2c_{M,N}\cos(\theta/3)$$

where $\theta \in [-\pi, \pi]$ observes $\cos\theta = \frac{M-2N}{M(M-2)c_{M,N}}$. Equality holds if and only if $\mathrm{Re}[Q_{i,j}Q_{j,l}Q_{l,i}] = \cos(\theta)$ for all $i \neq j \neq l \neq i$, where Q is the signature matrix of Φ.

Proof The operator norm of the submatrix of Q with rows and columns indexed by $\{i, j, l\}$ is $2\cos(\theta/3)$, with $\mathrm{Re}[Q_{i,j}Q_{j,l}Q_{l,i}] = \cos(\theta)$. However, the sum of all triple products is the constant

$$\sum_{i,j,l=1}^{M} Q_{i,j}Q_{j,l}Q_{l,i} = \frac{(M-1)(M-2N)}{c_{M,N}}$$

so the largest real part among all the triple products cannot be smaller than the average, which yields the desired inequality. □

Corollary 7.1 *Let M, N be such that $M > N$ and an equiangular (M, N)-frame exists with constant triple products $\mathrm{Re}[Q_{i,j}Q_{j,l}Q_{l,i}] = \cos(\theta)$ for some $\theta \in [-\pi, \pi]$; then the set $\mathcal{E}_\infty^{(3)}(M, N)$ contains precisely these frames.*

Remark 7.1 In [5], only $(M, M-1)$-frames were mentioned as examples for such 3-erasure optimal frames. However, recently Hoffman and Solazzo [31] found a family of examples which are not of this trivial type. We present one of their examples. It is the complex equiangular $(8, 4)$-frame with signature matrix

$$Q = \begin{pmatrix}
0 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
1 & 0 & -i & -i & -i & i & i & i \\
1 & i & 0 & -i & i & -i & -i & i \\
1 & i & i & 0 & -i & -i & i & -i \\
1 & i & -i & i & 0 & i & -i & -i \\
1 & -i & i & i & -i & 0 & -i & i \\
1 & -i & i & -i & i & i & 0 & -i \\
1 & -i & -i & i & i & -i & i & 0
\end{pmatrix}$$

In the real case Bodmann and Paulsen showed that, in fact, this 3-optimality condition is satisfied if and only if the frame is an equiangular $(M, M-1)$ or $(M, 1)$-frame. Thus, in order to differentiate between frames, they had to examine the case of equiangular tight frames in more detail. Bodmann and Paulsen [7] related the performance of these frames in the presence of higher numbers of erasures to graph-theoretic quantities.

To this end, they established an upper bound for the error and characterized cases of equality in graph-theoretic terms: Let Φ be a real equiangular (M, N)-frame.

Then $e_m^{\infty}(F) \leq N/M + (m-1)c_{M,N}$ with equality if and only if the signature matrix Q associated with Φ is the Seidel adjacency matrix of a graph that contains an induced complete bipartite subgraph on m vertices. The *Seidel adjacency matrix* of a graph G of M vertices is defined to be the $M \times M$ matrix $A = (a_{i,j})$, where $a_{i,j}$ is -1 when i and j are adjacent, it is $+1$ when i and j are not adjacent, and 0 when $i = j$. In certain cases Bodmann and Paulsen showed that as the size of a graph grows beyond some number, then among all the induced subgraphs of size up to 5 there is at least one complete bipartite graph. For such graphs, the worst-case m-erasure error is known up to $m = 5$. To differentiate between such types of equiangular Parseval frames, one needs to look beyond the 5-erasure error. Graph-theoretic criteria allowed the characterization of optimality by considering induced subgraphs of larger sizes [7].

7.2.1.2 Robustness of equiangular tight frames against erasures

We recall that when a frame with analysis operator T is used for the encoding, then an erasure is called correctable if ET has a left inverse, where E is the diagonal projection which sets the erased frame coefficients to zero. In this case, the left inverse effectively recovers any encoded vector x from the remaining set of nonzero frame coefficients.

The matrix ET has a left inverse if and only if all of its singular values are nonzero, or equivalently, whenever T^*ET is invertible. For Parseval frames this amounts to $\|I - T^*ET\| = \|T^*(I-E)T\| < 1$. This condition applies verbatim to sets of erasures, for example, the set of diagonal projection matrices with m zeros on the diagonal representing erasures of m frame coefficients.

Definition 7.5 A Parseval frame Φ with analysis operator T is robust against m erasures if

$$\left\| T^*ET - I \right\| < 1$$

for each diagonal projection E with $\operatorname{tr} E = M - m$.

A sufficient criterion for robustness of real and complex equiangular Parseval frames uses the following error estimate, which is a special case of a result for fusion frames [5].

Theorem 7.3 *Let Φ be an equiangular (M, N)-frame with signature matrix Q; then*

$$e_{\infty}^{(m)}(M, N) \leq N/M + (m-1)c_{M,N}$$

with equality if and only if there exists a diagonal $M \times M$ unitary matrix Y such that Y^*QY contains an $m \times m$ principal submatrix with off-diagonal entries that are all 1's.

Proof The largest eigenvalue of any $m \times m$ compression of Q determines the largest eigenvalue of the corresponding compression of TT^*. For each choice of m indices $\mathbb{K} = \{j_1, j_2, \ldots, j_m\}$, a normalized eigenvector x of $(Q_{j,l})_{j,l\in\mathbb{K}}$ belonging to the largest eigenvalue maximizes $q(x) = \langle Qx, x\rangle$ among all unit vectors with support contained in $\mathbb{K}$. Using the Cauchy-Schwarz inequality gives $q(x) \leq \sum_{j\neq l} |x_j x_l| \leq (m-1)$, and equality occurs if and only if $Q_{j,l}x_l\overline{x_j} = 1/m$ for all $j \neq l$ in $\mathbb{K}$. Now we can pick Y such that $Y_{j,j} = x_j$ if $j \in \mathbb{K}$, and then Y^*QY is seen to have the claimed form. □

Corollary 7.2 *If* $\frac{N}{M} + (m-1)c_{M,N} < 1$, *then any equiangular* (M, N)*-frame is robust against* m *erasures.*

Although this theorem permits the design of frames which correct a large number of erased coefficients, the fraction of erasures is only allowed to grow proportional to the square root of the number of frame vectors in order to guarantee correctability. A random choice of a frame corrects this deficiency with high probability, which we see in the next section.

7.2.2 Random Frames and Robustness Against Erasures

The results in the last section are qualitatively different from the usual results on binary coding, which guarantee the existence of codes that provide perfect recovery for memoryless channels [43] if the coding rate (here N/M) is below a certain value, the channel capacity. This can be traced back to the hierarchy among errors, which imposes the rigidity in the construction of optimizers.

In this section, we follow a complementary approach in which we wish to ensure correctability of m erasures, when m is chosen to be a fixed fraction of M, called the error rate. With the help of random frames, error correction for a fixed error rate is possible with arbitrarily large frames. To generate these frames, we choose an orthonormal sequence on N vectors in an M-dimensional real or complex Hilbert space, and transform this sequence by applying a random unitary (or orthogonal matrix, in the real case). The probability measure on the unitaries is the usual, normalized Haar measure.

Lemma 7.1 (Dasgupta and Gupta [21]) *Let* $0 < \epsilon < 1$. *Let* x *be a unit vector in a real Hilbert space* $\mathcal{H}$ *of dimension* M. *If* V *is a subspace of dimension* $N < M$, *drawn uniformly at random, and* P_V *is the orthogonal projection onto* V, *then*

$$\sqrt{\frac{M}{N}}\|P_V x\| \leq \frac{1}{1-\epsilon} \tag{7.1}$$

holds on a set of measure

$$\mathbb{P}\big(\{V : (7.1)\ \textit{holds}\}\big) \geq 1 - e^{-N\epsilon^2}.$$

Proof Dasgupta and Gupta prove that for $\beta > 1$,

$$\mathbb{P}\left(\frac{M}{N}\|P_V x\|^2 \geq \beta\right) \leq e^{\frac{N}{2}(1-\beta+\ln\beta)}.$$

Choosing $\beta = (1-\epsilon)^{-2}$ and comparing the terms in the Taylor expansions for $(1-\epsilon)^{-2}$ and $2\log(1-\epsilon)$ about $\epsilon = 0$ gives the desired bound. □

We pair this lemma with an argument similar to the exposition in Baraniuk et al. [1].

Lemma 7.2 *Let P_V be a random orthogonal projection onto a subspace of dimension N in $\mathbb{R}^M$, $M > N$, and let $W = \operatorname{span}\{e_{j_1}, e_{j_2}, \ldots, e_{j_s}\}$ be spanned by s vectors from an orthonormal basis, with $s < N$, and let $0 < \delta < 2$; then*

$$\sqrt{\frac{M}{N}}\|P_V x\| \leq \frac{1}{1-\delta+\delta^2/4}\|x\| \quad \textit{for all } x \in W \tag{7.2}$$

for a set of subspaces with probability

$$\mathbb{P}\big(\{V : (7.2)\ \textit{holds}\}\big) \geq 1 - 2\left(1+\frac{8}{\delta}\right)^s e^{-N\delta^2/4}.$$

Proof By scaling, we only have to show that (7.2) holds for $\|x\| = 1$, $x \in W$. Using the Minkowski inequality and Lipschitz continuity of the norm we can bootstrap from a net S with $\min_{y\in S}\|x-y\| \leq \frac{\delta}{4}$ for all $x \in W$, $\|x\| = 1$. By a volume inequality for sphere packings, we know there is such an S with cardinality

$$|S| \leq \left(1+\frac{8}{\delta}\right)^s.$$

Applying the upper bound from the Johnson-Lindenstrauss lemma in the version by Dasgupta and Gupta [21] we get a set of V's with measure as described for $\epsilon = \delta/2$, such that

$$\sqrt{\frac{M}{N}}\|P_V x\| \leq \frac{1}{1-\frac{\delta}{2}}\|x\|$$

for all $x \in S$. Now let a be the smallest number such that $\sqrt{\frac{M}{N}}\|P_V x\| \leq \frac{1}{1-a}\|x\|$ holds for all $x \in W$.

We show $a \leq \delta - \delta^2/4$. To see this, let $x \in W$, $\|x\| = 1$ and pick $y \in S$, $\|y-x\| \leq \frac{\delta}{4}$.

Then, using Minkowski's inequality yields

$$\sqrt{\frac{M}{N}}\|P_V x\| \leq \sqrt{\frac{M}{N}}\|P_V y\| + \sqrt{\frac{M}{N}}\|P_V(x-y)\| \leq \frac{1}{1-\frac{\delta}{2}} + \frac{1}{1-a}\frac{\delta}{4}.$$

Since the right-hand side of the inequality chain is independent of x, according to the definition of a we obtain

$$\frac{1}{1-a} \leq \frac{1}{1-\frac{\delta}{2}} + \frac{1}{1-a}\frac{\delta}{4}.$$

Solving for $(1-a)^{-1}$ and further estimating gives

$$\frac{1}{1-a} \leq \frac{1}{1-\frac{\delta}{2}}\frac{1}{1-\frac{\delta}{4}} \leq \frac{1}{(1-\frac{\delta}{2})^2}.$$

Inverting both sides results in

$$1-a \geq 1-\delta+\frac{\delta^2}{4}.$$

Hence, $a \leq \delta - \delta^2/4$. □

Since any M-dimensional complex Hilbert space can be interpreted as a $2M$-dimensional real Hilbert space, we conclude the following theorem.

Theorem 7.4 *Let $\mathcal{H}$ be an M-dimensional real or complex Hilbert space. Let P_V be a random orthogonal projection onto a subspace of dimension N in $\mathcal{H}$, $N < M$, let $0 < \delta < 2$, and let $\mathcal{W}$ be the union of all sets of subspaces spanned by s vectors from an orthonormal basis of $\mathcal{H}$; then*

$$\sqrt{\frac{M}{N}}\|P_V x\| \leq \frac{1}{1-\delta+\delta^2/4}\|x\| \quad \textit{for all } x \in \mathcal{W} \tag{7.3}$$

with probability

$$\mathbb{P}\big(\{V : (7.3)\ \textit{holds}\}\big) \geq 1 - \left(1+\frac{8}{\delta}\right)^{\tilde{s}}\left(\frac{eM}{s}\right)^{s} e^{-\delta^2 N/4},$$

where $\tilde{s} = s$ in the real case and $\tilde{s} = 2s$ in the complex case.

Proof There are $\binom{M}{s}$ choices for the subspaces spanned by s orthonormal basis vectors. Stirling's approximation gives $\binom{M}{s} \leq (eM/s)^s$. In the real case, the result follows directly from the preceding lemma by a further union bound. In the complex case, we identify each s-dimensional subspace with a $2s$-dimensional real subspace in a $2M$-dimensional real Hilbert space. The sphere packing argument then yields a net S of size $|S| \leq (1+8/\delta)^{\tilde{s}}$. Using the same union bound as in the real case then yields the desired result. □

The exponential term in the failure probability ensures that for a fixed, sufficiently small coding ratio N/M there is an erasure ratio s/M which can be corrected with overwhelming probability. This result was established in a discussion with Gitta Kutyniok, and we are grateful for the opportunity to present it here.

Theorem 7.5 *Let $0 < c < 1$ and $M \geq 3$. Let Φ be a random Parseval frame consisting of M vectors for a real or complex Hilbert space with dimension N, and some δ such that*

$$0 < \delta < 2\left(1 - \sqrt[4]{\frac{N}{M}}\right)$$

and

$$\frac{s}{N}\left(1 + 2\ln\left(1 + \frac{8}{\delta}\right) + \ln\frac{M}{s}\right) < c\frac{\delta^2}{4}.$$

Then the probability that any s erasures are not correctable decays exponentially fast in the number of frame coefficients.

Proof This is a result of the preceding theorem, together with the requirement for correctability.

All sets of s erased coefficients can be corrected if the union $\mathcal{W}$ of the subspaces spanned by s basis vectors satisfies

$$\|P_V x\| \leq \sqrt{\frac{N}{M}}\frac{1}{1 - \delta + \delta^2/4} < 1 \quad \text{for all } x \in \mathcal{W}, \|x\| = 1.$$

If we assume $\delta < 2(1 - \sqrt[4]{\frac{N}{M}})$, then $\frac{1}{1-\delta+\delta^2/4} < \sqrt{\frac{M}{N}}$ and the set of V's such that this fails has measure bounded above by

$$\mathbb{P}\big[\|P_V x\| = \|x\| \text{ for some } x \in \mathcal{W}\big] \leq e^{2s\ln(1+8/\delta)+s(1+\ln(M/s))-\delta^2 N/4}.$$

Finally, if there is $0 < c < 1$ and if the exponent is bounded by $2s\ln(1+8/\delta)+s(1+\ln(M/s)) - \delta^2 N/4 \leq (c-1)\delta^2 N/4$, then we achieve overwhelming probability for correcting any such s erasures as $N \to \infty$. □

7.2.3 Erasures and Additive Noise

If erasures are present and the encoded vector is subject to additional noise in its coefficients, then the error estimates must be modified to account for this. However, deriving upper bounds is relatively simple if the noise is assumed to be independent of the input vector in the case of either the worst-case or the mean-squared error. We first examine the performance of blind reconstruction.

7.2.3.1 Passive error correction

A natural measure for performance when the reconstruction is performed with T^* is an L^p-norm for the reconstruction error, where the underlying measure models the

distribution of input vectors, erasures, and the additive noise. The function whose L^p-norm is computed is the reconstruction error $(x, \omega, y) \mapsto \|T^*E(\omega)(Tx + y) - x\|$ where $x \in \mathcal{H}$ is the vector to be encoded, an erasure given by $\omega \in \Omega$ occurs, and the noise $y \in \ell^2(\{1, 2, \ldots, M\})$ is added to the frame coefficients.

Definition 7.6 If the input vectors and erasures are governed by the uniform probability measure μ on $\mathbb{S} \times \Gamma$, with $\mathbb{S}$ the unit sphere in $\mathcal{H}$ and $\Gamma \subset \Omega$, and the frame coefficients are subject to additive noise distributed according to a probability measure ν, then the error for blind reconstruction is

$$e_p(\Phi, \mu, \nu) = \left(\int_{\ell^2(\mathbb{J})} \int_{\mathbb{S}\times\Omega} \left\| T^*E(\omega)(Tx + y) - x \right\|^p d\mu(x, \omega)\, d\nu(y) \right)^{1/p}$$

We prepare estimates for e_p by examining the noiseless case.

Lemma 7.3 *Let Φ be an (M, N)-frame. The input-averaged mean-squared error $e_2(\Phi, \mu)$ has the form*

$$e_2(\Phi, \mu) = \sum_{j,l=1}^{M} w_{j,l} \left| \left(TT^*\right)_{j,l} \right|^2$$

with some $w_{j,l} = w_{l,j} \geq 0$, so it is the square of a weighted Frobenius norm of the Gram matrix.

Proof We have that for fixed E, $\int \|T^*(E-I)Tx\|^2 d\mu(x) = \mathrm{tr}[T^*(E-I)TT^*(E-I)T]/N = \mathrm{tr}[((E-I)TT^*(E-I))^2]/N$. This is the square of the Frobenius norm of the submatrix of TT^* corresponding to the erased coefficients. Next, taking convex combinations of this expression when averaging over the erasures preserves this form. □

Proposition 7.3 *Let ν be the uniform probability measure on the sphere of radius $\sigma > 0$ in $\ell^2(\{1, 2, \ldots, M\})$, and let Φ be an (M, N)-frame and μ be as above; then we have the inequality*

$$e_\infty(\Phi, \mu, \nu) \leq e_\infty(\Phi, \mu) + \sigma$$

and the Pythagoras-like identity

$$e_2(\Phi, \mu, \nu)^2 = e_2(\Phi, \mu)^2 + \sigma^2 e_2(\Phi, \overline{\mu})^2.$$

Here, $\overline{\mu}$ denotes the measure on Ω which is induced by μ under the map $\omega \mapsto \overline{\omega}$, $\overline{\omega}_j = 1 - \omega_j$.

Proof When performing the average over y, the mixed term in the expansion

$$\begin{aligned} &\left\| T^*(E-I)Tx + T^*Ey \right\|^2 \\ &\quad = \left\| T^*(E-I)Tx \right\|^2 + 2\mathrm{Re}\left[\left\langle T^*(E-I)Tx, T^*Ey \right\rangle \right] + \left\| T^*Ey \right\|^2 \end{aligned}$$

does not contribute because $\langle T^*(E-I)Tx, T^*Ey\rangle = \langle ETT^*(E-I)Tx, y\rangle$ and y averages to zero. Therefore, the erasures and noise are additive for the mean-squared error. Averaging the noise term gives

$$\int_{\sigma\mathbb{S}} \left\|T^*Ey\right\|^2 d\nu(y) = \frac{\sigma^2}{M}\operatorname{tr}\left[ETT^*E\right].$$

This is the error expression that applies when the complement of the erasure given by E occurs. □

7.2.3.2 Active error correction

If active error correction is used to compensate erasures in the presence of additive noise, then the operator norm of the Moore-Penrose pseudoinverse determines how much the error affecting the frame coefficients contributes to the reconstruction error. The worst-case reconstruction error among all unit-norm input vectors, a set of erasures, and the additive noise to be considered is the essential supremum of the function $(x, \omega, y) \mapsto \|W(\omega)E(\omega)(Tx + y) - x\|$, where $W(\omega)$ is the Moore-Penrose pseudoinverse of $E(\omega)T$. As before, we assume that the additive error is distributed uniformly on a sphere of radius $\sigma > 0$ in $\ell^2(\{1, 2, \ldots, M\})$. In this case, it turns out that the previously introduced quantity $a_\infty(\Phi, \mu)$ determines the performance.

Proposition 7.4 *Let Φ be an (M, N)-frame and μ be as above. If every erasure in Ω is correctable in the absence of noise, and additive noise is distributed uniformly on a sphere of radius $\sigma > 0$ in $\ell^2(\{1, 2, \ldots, M\})$ then the worst-case reconstruction error for active error correction is given by*

$$\max_{\|y\|=\sigma, \|x\|=1, \omega\in\Omega} \left\|W(\omega)E(\omega)(Tx + y) - x\right\| = \sigma a_\infty(\Phi, \nu \times \mu),$$

where a_∞ is the measure for active error correction from Definition 7.3 and ν is the uniform measure on the unit sphere $\mathbb{S}$ in $\ell^2(\{1, 2, \ldots, M\})$.

Proof If for every $\omega \in \Omega$, the erasure can be corrected, then $W(\omega)E(\omega)T = I$. This means that the expression for the worst-case error simplifies to

$$\max_{\|y\|=\sigma, \omega\in\Omega} \left\|W(\omega)E(\omega)y\right\| = \sigma a_\infty(\Phi, \nu \times \mu).$$

In the last step, the homogeneity of the norm is used, and the operator norm is replaced by the L^∞-norm appearing in the definition of a_∞. □

7.3 Fusion Frames for Packet Encoding

As discussed in the preceding section, a finite frame can be interpreted as a block code for analog signals. Instead of blocks of bits (or strings of some length), the

analysis operator transforms a vector x, which can be characterized by its expansion in a given orthonormal basis of size N, into a sequence of M frame coefficients $\{\langle x, \varphi_j\rangle\}_{j=1}^M$. Similarly, the analysis operator of a fusion frame is given by a family of linear maps $\{T_j\}_{j=1}^M$ which transform a vector $x \in \mathcal{H}$ into its image $\bigoplus_j T_j x \in \bigoplus_j \mathcal{K}_j$ in the direct sum of spaces $\mathcal{K}_j$ containing the range of each T_j. We will refer to each vector $T_j x \in \mathcal{K}_j$ as a component of x.

Frames can be understood as a special case of fusion frames when the rank of each T_j is one. Thus, many insights for frames have an analogue for fusion frames. In finite dimensions, the condition for having a frame or a fusion frame is simply the spanning property of $\{\varphi_j\}_{j=1}^M$ or of the ranges $\{\operatorname{ran} T_j^*\}_{j=1}^M$, respectively. This implies for the number of frame vectors, $M \geq N$, and for the ranks $\sum_{j=1}^M \operatorname{rank} T_j \geq N$. In both cases, the redundancy which is incorporated by the analysis operator can be exploited to compensate errors which may corrupt the frame coefficients or components in the course of a transmission or when they are stored. This is the goal of designing frames or fusion frames as codes. The purpose of optimal design is generally to use the redundancy to suppress the effect of errors maximally.

We know sharp estimates for the Euclidean reconstruction error and corresponding optimal designs for deterministic and random signals. The deterministic, worst-case scenario was examined for one lost component [19] and for a higher number of losses [5]. The averaged performance has been discussed as well, in the probabilistic treatment for the mean-squared error occurring for random input vectors [35].

Definition 7.7 Let $\mathcal{H}$ be a real or complex Hilbert space, and let $\{T_j\}_{j=1}^M$ be a finite family of linear maps $T_j : \mathcal{H} \to \mathcal{K}$ into a Hilbert space $\mathcal{K}$. We say that $\{T_j\}_{j=1}^M$ is a set of *coordinate operators* on $\mathcal{H}$ if they form a *resolution of the identity*

$$\sum_{j=1}^M T_j^* T_j = I.$$

We will also say that the family $\{T_j\}_{j=1}^M$ forms a *Parseval fusion frame*, although this is, according to the usual definition [18], only the case if each T_j is a multiple of a partial isometry.

We observe that the *analysis operator* T formed by combining the blocks $\{T_j\}_{j=1}^M$ as rows in an isometry

$$T : \mathcal{H} \longrightarrow \bigoplus_{j\in\mathbb{J}} \mathcal{K}, (Tx)_j = T_j x$$

has its adjoint T^* as a left inverse. We will often abbreviate $\bigoplus_{j=1}^M \mathcal{K} = \mathcal{K}^M$.

Definition 7.8 We call a family $\{T_j\}_{j\in\mathbb{J}}$ of coordinate operators of a Parseval fusion frame *equal norm* provided there is a constant $c > 0$ so that the operator norm $\|T_j\| = c$ for all $j \in \mathbb{J}$.

Definition 7.9 We shall let $\mathcal{V}(M, L, N)$ denote the collection of all families $\{T_j\}_{j=1}^{M}$ consisting of $M \in \mathbb{N}$ coordinate operators $T_j : \mathcal{H} \to \mathcal{K}$ of maximal rank $L \in \mathbb{N}$ that provide a resolution of the identity for the N-dimensional real or complex Hilbert space $\mathcal{H}$, $N \in \mathbb{N}$. We call the analysis operator T of such a family $\{T_j\} \in \mathcal{V}(M, L, N)$ an (M, L, N)-*protocol*.

The ratio ML/N we shall refer to as the *redundancy ratio* of the encoding.

As in the frames case, among all possible left inverses of T, the analysis operator of a Parseval fusion frame, we have that T^* is the unique left inverse that minimizes both the operator norm and the Hilbert-Schmidt norm.

7.3.1 Packet Erasures and Performance Measures

The problem we consider is that in the process of transmission some of the packets $(T_j x)$ are lost, or their content has become inaccessible because of some transmission error.

Definition 7.10 Let $\mathbb{K} \subset \mathbb{J} = \{1, 2, \ldots, M\}$ be a subset of size $|\mathbb{K}| = m \in \mathbb{N}$. The *packet erasure* matrix $E_{\mathbb{K}}$ on $\bigoplus_{j\in\mathbb{J}} \mathcal{K}$ is given by

$$E_{\mathbb{K}} : \bigoplus_{j=1}^{M} \mathcal{K} \longrightarrow \bigoplus_{j=1}^{M} \mathcal{K}, \quad (E_{\mathbb{K}} y)_j = \begin{cases} y_j, & j \notin \mathbb{K}, \\ 0, & j \in \mathbb{K}. \end{cases}$$

The operator $E_{\mathbb{K}}$ can be thought of as erasing the coordinates $(T_j x)_{j\in\mathbb{K}}$ in the terminology of [29]. The main goal of this section is to characterize when the norms of these error operators are in some sense minimized for a given number of lost packets, independent of which packets are lost. Of course, there are many ways that one could define optimality in this setting. Here, we only pursue two possibilities: optimality for the worst-case error and for the input-averaged mean-squared error. Optimizing with respect to the second performance measure is equivalent to finding frames which give the best statistical recovery with Wiener filtering [35].

Definition 7.11 Let $T : \mathcal{H} \to \bigoplus_{j\in\mathbb{J}} \mathcal{K}$ be an (M, L, N)-protocol, and let $\mu = \sigma \times \rho$ be a product of probability measures governing the inputs as well as the packet erasures. We chose σ to be the uniform probability measure on the sphere of $\mathcal{H}$ and ρ a probability measure supported on a subset Γ of Ω. We define the reconstruction error by

$$e_{p,\infty}(T, \mu) = \left(\max_{\omega\in\Gamma} \int_{\mathbb{S}} \left\| T^* \big(I - E(\omega)\big) T x \right\|^p d\sigma(x) \right)^{1/p}$$

and focus mostly on the *worst-case error* $e_{\infty,\infty}(T, \mu)$ as well as the *worst-case input-averaged mean-squared error* $e_{2,\infty}(\Phi, \mu)^2$.

The expressions for these two types of error measures can be simplified by replacing the average over the input vectors with matrix norms. Because of the nature of the sup-norm, the dependence of $e_{p,\infty}(T, \sigma \times \rho)$ on the measure ρ is only through its support, $\Gamma \subset \Omega$.

Proposition 7.5 *Let $T : \mathcal{H} \to \bigoplus_{j\in\mathbb{J}} \mathcal{K}$ be an (M, L, N)-protocol. If $\mu = \sigma \times \rho$, where σ is the uniform probability measure on the sphere in the Hilbert space $\mathcal{H}$, and ρ is a probability measure supported on $\Gamma \subset \Omega$, then*

$$e_{\infty,\infty}(T, \mu) = \max\big\{\big\|\big(I - E(\omega)\big)TT^*\big(I - E(\omega)\big)\big\| : \omega \in \Gamma\big\},$$

and

$$e_{2,\infty}(T, \mu) = \max_{\omega\in\Gamma} \operatorname{tr}\big[\big(I - E(\omega)\big)TT^*\big(I - E(\omega)\big)^2\big]/N.$$

Proof For a fixed erasure and the corresponding matrix E, by the positivity of $T^*(I - E)T$, the eigenvector $x \in \mathbb{S}$ belonging to the maximal eigenvalue gives the operator norm. This, in turn, equals $\|T^*(I - E)T\| = \|(I - E)TT^*(I - E)\|$.

Averaging the square of the reconstruction error over all normalized input vectors gives

$$\begin{aligned}\int_{\mathbb{S}} \big\|T^*(I - E)Tx\big\|^2 \, d\sigma(x) &= \operatorname{tr}\big[\big(T^*(I - E)T\big)^2\big]/N \\ &= \operatorname{tr}\big[\big((I - E)TT^*(I - E)\big)^2\big]/N,\end{aligned}$$

where we have used that $I - E$ is an orthogonal projection. □

The input-averaged mean-squared error for a fixed erasure is therefore proportional to the square of the Frobenius norm of $[TT^*]_{\mathbb{K}} = (T_i T_j^*)_{i,j\in\mathbb{K}}$, the submatrix of the Gram matrix consisting of the rows and columns indexed by the erased packets.

7.3.2 Optimality for Hierarchical Error Models

We denote, similarly as in the frames case, by μ_m the product of the uniform probability measure on $\mathbb{S} \times \Gamma_m$, where $\mathbb{S}$ is the unit sphere in $\mathcal{H}$ and Γ_m is the subset $\{\omega \in \{0, 1\}^M : \sum_{j=1}^M \omega_j = m\}$. Since the set $\mathcal{V}(M, L, N)$ of all (M, L, N)-protocols is a compact set, the value

$$e^{(1)}_{p,\infty}(M, L, N) = \inf\big\{e_{p,\infty}(T, \mu_1) : T \in \mathcal{V}(M, L, N)\big\}$$

is attained, and we define the set of *1-erasure optimal protocols* to be the nonempty compact set $\mathcal{V}^{(1)}_p(M, L, N)$ where this infimum is attained, i.e.,

$$\mathcal{V}^{(1)}_p(M, L, N) = \big\{T \in \mathcal{V}(M, L, N) : e_{p,\infty}(T, \mu_1) = e^{(1)}_{p,\infty}(M, L, N)\big\}.$$

Proceeding inductively, we now set for $2 \le m \le M$,

$$e_{p,\infty}^{(m)}(M, L, N) = \min\left\{e_{p,\infty}(T, \mu_m) : T \in \mathcal{V}_p^{(m-1)}(M, L, N)\right\}$$

and define the *optimal m-erasure protocols* to be the nonempty compact subset $\mathcal{V}_p^{(m)}(M, L, N)$ of $\mathcal{V}_p^{(m-1)}(M, L, N)$ where this minimum is attained.

7.3.2.1 Worst-case analysis

Next, we discuss optimality for the worst case occurring with one lost packet, as presented in [5]. The proofs are relatively straightforward extensions of the frames case.

Proposition 7.6 [5] *If the coordinate operators* $\{T_j : \mathcal{H} \to \mathcal{K}\}$ *belong to an* (M, L, N)*-protocol on a Hilbert space* $\mathcal{H}$*, then*

$$\max_j \left\| T_j^* T_j \right\| \ge \frac{N}{ML}$$

and equality holds if and only if for all $j \in \{1, 2, \ldots, m\}$ *we have* $T_j^* T_j = \frac{N}{ML} P_j$*, where* P_j *is a self-adjoint rank-L projection operator.*

Proof Comparing the operator norm of $T_j^* T_j$ with its trace gives

$$\max_j \left\| T_j^* T_j \right\| \ge \frac{1}{ML} \sum_{j=1}^{m} \operatorname{tr}\left[T_j^* T_j\right] = \frac{N}{ML}.$$

If equality holds, then for each j, $L\|T_j^* T_j\| = \operatorname{tr}[T_j^* T_j]$, so each $T_j^* T_j$ is rank L and has only one nonzero eigenvalue. Dividing by this eigenvalue gives the self-adjoint projection $P_j = ML T_j^* T_j / N$. □

Corollary 7.3 [5] *Let* $M, L, N \in \mathbb{N}$*, and let* $T : \mathcal{H} \to \bigoplus_{j=1}^{M} \mathcal{K}$ *be an* (M, L, N)*-protocol. Then*

$$e_{\infty,\infty}^{(1)}(T, \mu_1) \ge \frac{N}{ML}$$

and equality holds if and only if the coordinate operators $\{T_j : \mathcal{H} \to \mathcal{K}\}_{j=1}^{M}$ *satisfy that for all* $j \in \{1, 2, \ldots, M\}$,

$$T_j^* T_j = \frac{N}{ML} P_j$$

with self-adjoint rank-L projections $\{P_j\}_{j=1}^{M}$ *on* $\mathcal{H}$.

Proof If the largest operator norm achieves the lower bound $\max_j \|T_j^* T_j\| = N/ML$, then the preceding proposition shows that $\{T_j\}_{j=1}^M$ provides an equal norm (M, L, N)-protocol. □

The consequence of this characterization of optimality is that if there exist m uniformly weighted rank-L projections resolving the identity on a Hilbert space $\mathcal{H}$ of dimension N, then the equal-norm (M, L, N)-protocols are precisely the 1-erasure optimal ones. These protocols are also known as Parseval frames for subspaces [18] or Parseval fusion frames [20].

We now turn to the case of two lost packets. We abbreviate

$$c_{M,L,N} = \sqrt{\frac{N(ML-N)}{M^2L^2(M-1)}}.$$

A form of the bound by Welch [48] gives a characterization of optimality.

Lemma 7.4 [5] *If $\{T_j\}_{j=1}^M$, $M \geq 2$, is a family of uniformly weighted rank-L coordinate operators of a projective resolution of the identity on a Hilbert space $\mathcal{H}$ of dimension N, then*

$$\max_{i \neq j} \left\| T_i T_j^* \right\| \geq c_{M,L,N}$$

and equality holds if and only if for all $i \neq j$, $T_i T_j^* = c_{M,L,N} Q_{i,j}$ with a unitary $Q_{i,j}$ on $\mathcal{K}$.

A block-matrix version of the estimate for the spectrum of $(I-E)TT^*(I-E)$ gives an expression for the worst-case two-packet erasure error.

Theorem 7.6 [5] *Let $M, L, N \in \mathbb{N}$. If $T : \mathcal{H} \to \bigoplus_{j=1}^M \mathcal{K}$ is a uniform (M, L, N)-protocol, then if $M \geq 2$,*

$$e_2(T) \geq \frac{N}{ML} + c_{M,L,N}$$

and equality holds if and only if for each pair $i, j \in \{1, 2, \ldots, M\}$, $i \neq j$, we have $T_i T_j^* = c_{M,L,N} Q_{i,j}$ with a unitary $\{Q_{i,j}\}_{i \neq j}$ on $\mathcal{K}$.

The case when this bound is saturated describes a set of protocols which can be characterized in geometric terms.

Definition 7.12 We call a linear map $T : \mathcal{H} \to \bigoplus_{j=1}^m \mathcal{K}$ an *equi-isoclinic (M, L, N)-protocol* provided that the coordinate operators of T are uniform and in addition there is a constant $c > 0$ such that $\|T^*(I-E)T\| = c$ for all two-packet erasure operators E.

The fact that for $i \neq j$, $T_i T_j^* = c_{M,L,N} Q_{i,j}$ with a unitary $Q_{i,j}$ on $\mathcal{K}$ means that for every $x \in \mathcal{K}$, $\|T_i T_j^* x\| = c_{M,L,N} \|x\|$. However, T_i^* and T_j^* are isometries, so for

any $y \in \mathcal{H}$ in the range of T_j^*, we have $\|T_i^* T_i y\| = c_{M,L,N}\|y\|$. This means, for all $i \neq j$, projecting any vector in the range of P_i onto the range of P_j changes its length by the scalar multiple $c_{M,L,N}$. Such a family of subspaces is called *equi-isoclinic* [27, 36], and we have named the corresponding protocols accordingly.

Definition 7.13 Given an equi-isoclinic (M, L, N)-protocol $T : \mathcal{H} \to \bigoplus_{j=1}^m \mathcal{K}$, then $TT^* = aI + c_{M,L,N} Q$ is a projection on $\bigoplus_j \mathcal{K}$ where $a = N/ML$, $c_{M,L,N}$ is the lower bound in Lemma 7.4, and $Q = (Q_{i,j})_{i,j=1}^m$ is a self-adjoint matrix containing the zero operator $Q_{i,i} = 0$ on $\mathcal{K}$ for all $i \in \{1, 2, \ldots, m\}$ and unitaries $Q_{i,j}$ on $\mathcal{K}$ for off-diagonal entries indexed by $i \neq j$. We call this self-adjoint matrix of operators Q the *signature matrix* of T.

Since TT^* has two eigenvalues, so does the signature matrix. This fact reduces the construction of equi-isoclinic (M, L, N)-protocols to producing matrices Q satisfying a quadratic equation.

Lemmens and Seidel [36] describe constructions to obtain examples of real equi-isoclinic subspaces and thus of real signature matrices. Godsil and Hensel [27] show how to obtain such subspaces from distance-regular antipodal covers of the complete graph. It is an open problem to find a graph-theoretic characterization of equivalence classes of equi-isoclinic protocols for real Hilbert spaces. Even less seems to be known about generic constructions and an analogue of the graph-theoretic characterization of two-uniform protocols in the complex case [24–26].

Next, for given dimensions M, L and $N \in \mathbb{N}$, we want to minimize the worst-case Euclidean reconstruction error for three lost packets among two-uniform (M, L, N)-protocols.

For any three-element subset of indices $\mathbb{K} = \{h, i, j\} \subset \mathbb{J} = \{1, 2, \ldots, m\}$, we denote the compression of an $M \times M$ (block) matrix A to the corresponding rows and columns as

$$[A]_{\mathbb{K}} = \begin{pmatrix} A_{h,h} & A_{h,i} & A_{h,j} \\ A_{i,h} & A_{i,i} & A_{i,j} \\ A_{j,h} & A_{j,i} & A_{j,j} \end{pmatrix}.$$

The following theorem gives a lower bound for e_3 among all two-uniform (M, L, N)-protocols. If $\mathcal{H}$ is a real Hilbert space and $\mathcal{K} = \mathbb{R}$, then it can be reduced to a known statement [7, Sect. 5.2].

Theorem 7.7 [5] *Let $M, L, N \in \mathbb{N}$, $M \geq 3$ and $N \leq ML$. Let $T : \mathcal{H} \to \bigoplus_{j=1}^M \mathcal{K}$ be a two-uniform (M, L, N)-protocol. Then*

$$e_3(T) \geq \frac{N}{ML} + 2c_{M,L,N} \cos(\theta/3)$$

where $\theta \in [-\pi, \pi]$ observes

$$\cos\theta = \frac{ML - 2N}{ML(M-2)c_{M,L,N}}.$$

When $N < ML$, the protocol T saturates the lower bound for e_3 if and only if the signature matrix Q of T satisfies that for all $\{h, i, j\} \subset \{1, 2, \ldots, m\}$, the largest eigenvalue of $Q_{h,i}Q_{i,j}Q_{j,h} + Q_{h,j}Q_{j,i}Q_{i,h}$ is $2\cos(\theta)$.

7.3.2.2 Correctability of equi-isoclinic protocols for a higher number of lost packets

If the largest eigenvalue among all $\{Q_{h,i}Q_{i,j}Q_{j,h} + Q_{h,j}Q_{j,i}Q_{i,h}, h \neq i \neq j \neq h\}$ is 2 for the signature matrix of an equi-isoclinic (M, L, N)-protocol, then this protocol maximizes the worst-case norm of the reconstruction error for $m = 3$ lost packets. We characterize the analogue of this situation for higher values of m.

If $\mathcal{H}$ is a real Hilbert space and $\mathcal{K} = \mathbb{R}$, then the "unitaries" $Q_{i,j}$ are scalars ± 1, and the presence of a covariant vector amounts to partitioning $\mathbb{K}$ into two subsets, such that $Q_{i,j} = -1$ whenever i and j belong to different subsets. This can be restated in graph-theoretic terminology, which is the basis for the derivation of error bounds [7] in this special case. Here, we state an analogous result for packet encoding.

Theorem 7.8 [5] *Let $M, L, N, m \in \mathbb{N}$. If T is a two-uniform (M, L, N)-protocol with signature matrix Q, then*

$$e^{(m)}(M, L, N) \leq \frac{N}{ML} + (m-1)c_{M,L,N}.$$

We use this theorem to derive a sufficient condition for correctability of packet losses. If the upper bound is strictly less than one, then the content of any m lost packets can be recovered.

Corollary 7.4 *If T is a two-uniform (M, L, N)-protocol, $N \leq ML$, then any m-packet loss operator is correctable if $1 \leq m < 1 + \sqrt{\frac{(M-1)(ML-N)}{N}}$.*

7.3.2.3 Optimality for the input-averaged mean-squared error

The characterization of optimality is changed slightly when we change the performance measure to the input-averaged mean-squared error. This measure is not the same as the mean-squared error for linear reconstruction with Wiener filtering as discussed by Kutyniok et al. [35], but the optimizers are identical.

Proposition 7.7 *If the coordinate operators $\{T_j : \mathcal{H} \to \mathcal{K}\}$ belong to an (M, L, N)-protocol on a Hilbert space $\mathcal{H}$, then*

$$\max_j \operatorname{tr}\left[\left(T_j^* T_j\right)^2\right] \geq \frac{N^2}{M^2 L}$$

and equality holds if and only if for all $j \in \{1, 2, \ldots, m\}$ *we have* $T_j^* T_j = \frac{N}{ML} P_j$, *where* P_j *is a self-adjoint rank-L projection operator.*

Proof The maximum square of the Frobenius norm is larger than the average,

$$\max_j \operatorname{tr}\left[\left(T_j^* T_j\right)^2\right] \geq \frac{1}{M} \sum_{j=1}^{M} \operatorname{tr}\left[\left(T_j^* T_j\right)^2\right].$$

In terms of the eigenvalues, this is simply the square of an ℓ^2-norm. However, the ℓ^1-norm is fixed, $\sum_j \operatorname{tr}[T_j^* T_j] = N$, so the minimum is achieved when all eigenvalues are equal to N/ML. This gives $\operatorname{tr}[(T_j^* T_j)^2] = L(N/ML)^2$ and $\max_j \operatorname{tr}[(T_j^* T_j)^2] \geq N^2/M^2 L$. If equality holds, then each $T_j^* T_j$ is rank L and has only one nonzero eigenvalue. Dividing by this eigenvalue gives the self-adjoint projection $P_j = MLT_j^* T_j/N$. □

Corollary 7.5 *Let* $M, L, N \in \mathbb{N}$, *and let* $T : \mathcal{H} \to \bigoplus_{j=1}^{M} \mathcal{K}$ *be an* (M, L, N)-*protocol. Then*

$$e_{2,\infty}^{(1)}(T, \mu_1) \geq \frac{N}{M^2 L}$$

and equality holds if and only if the coordinate operators $\{T_j : \mathcal{H} \to \mathcal{K}\}_{j=1}^{M}$ *satisfy that for all* $j \in \{1, 2, \ldots, M\}$,

$$T_j^* T_j = \frac{N}{ML} P_j$$

with self-adjoint rank-L projections $\{P_j\}_{j=1}^{M}$ *on* $\mathcal{H}$.

Proof The proof is immediate from the expression of $e_{2,\infty}^{(1)}$ in terms of the square of the Frobenius norms among all diagonal elements of the block matrix $(T_i T_j^*)_{i,j=1}^{M}$.

If the Frobenius norm achieves the lower bound, then as before we have $\max_j \|T_j^* T_j\| = N/ML$, and the preceding proposition shows that $\{T_j\}_{j=1}^{M}$ provides an equal-norm (M, L, N)-protocol. □

To summarize, if equal norm fusion frames exist, then their analysis operators are optimal protocols for the worst-case and for the input-averaged error.

The 2-erasure optimality for the mean-squared error is qualitatively different from the worst-case analysis.

Proposition 7.8 *Let* T *be the analysis operator of an equal-norm* (M, L, N)-*protocol*; *then*

$$e_{2,\infty}^{(2)}(T, \mu_2) \geq 2\frac{N}{M^2 L} + 2\frac{ML - N}{M^2(M-1)L}.$$

Proof The sum of all the traces is

$$\sum_{i,j=1}^{M} \mathrm{tr}\big[T_i T_j^* T_j T_i^*\big] = N,$$

and subtracting the diagonal gives

$$\sum_{i\neq j} \mathrm{tr}\big[T_i T_j^* T_j T_i^*\big] = N - \frac{N^2}{ML} = \frac{NML - N^2}{ML}.$$

The maximum among the $M(M-1)$ terms cannot be smaller than the average, so

$$\max_{i\neq j} \mathrm{tr}\big[\big(T_i T_j^*\big)^2\big] \geq \frac{NML - N^2}{M^2(M-1)L}.$$

Now adding the contribution of two diagonal blocks and two off-diagonal blocks in the expression for $e_{2,\infty}^{(2)}$ gives the desired estimate. □

Corollary 7.6 *An equal-norm (M, L, N)-protocol achieves the lower bound in the preceding proposition if and only if there is a constant $c \geq 0$ such that for any pair $i \neq j$, we have*

$$\mathrm{tr}\big[T_i^* T_i T_j^* T_j\big] = c.$$

One could interpret this as the Hilbert-Schmidt inner product between $T_i^* T_i$ and $T_j^* T_j$ and define a distance between these two operators, or equivalently, between their ranges. The identity then means that all pairs of subspaces have an equal distance. For this reason, the associated fusion frames have been called equidistance fusion frames [35].

Acknowledgements Special thanks go to Gitta Kutyniok and to Pete Casazza for the helpful comments in the course of preparing this chapter. The research presented here was partially supported by NSF grant DMS-1109545 and AFOSR grant FA9550-11-1-0245.

References

1. Baraniuk, R., Davenport, M., DeVore, R., Wakin, M.: A simple proof of the restricted isometry property for random matrices. Constr. Approx. **28**(3), 253–263 (2008). doi:10.1007/s00365-007-9003-x
2. Benedetto, J., Yilmaz, O., Powell, A.: Sigma-delta quantization and finite frames. In: IEEE International Conference on Acoustics, Speech, and Signal Processing, 2004. Proceedings. (ICASSP'04), vol. 3, pp. iii, 937–940 (2004). doi:10.1109/ICASSP.2004.1326700
3. Benedetto, J.J., Powell, A.M., Yılmaz, Ö.: Second-order sigma-delta ($\Sigma\Delta$) quantization of finite frame expansions. Appl. Comput. Harmon. Anal. **20**(1), 126–148 (2006). doi:10.1016/j.acha.2005.04.003

4. Benedetto, J.J., Powell, A.M., Yılmaz, Ö.: Sigma-delta ($\Sigma\Delta$) quantization and finite frames. IEEE Trans. Inf. Theory **52**(5), 1990–2005 (2006). doi:10.1109/TIT.2006.872849
5. Bodmann, B.G.: Optimal linear transmission by loss-insensitive packet encoding. Appl. Comput. Harmon. Anal. **22**(3), 274–285 (2007). doi:10.1016/j.acha.2006.07.003
6. Bodmann, B.G., Elwood, H.J.: Complex equiangular Parseval frames and Seidel matrices containing pth roots of unity. Proc. Am. Math. Soc. **138**(12), 4387–4404 (2010). doi:10.1090/S0002-9939-2010-10435-5
7. Bodmann, B.G., Paulsen, V.I.: Frames, graphs and erasures. Linear Algebra Appl. **404**, 118–146 (2005). doi:10.1016/j.laa.2005.02.016
8. Bodmann, B.G., Paulsen, V.I.: Frame paths and error bounds for sigma-delta quantization. Appl. Comput. Harmon. Anal. **22**(2), 176–197 (2007). doi:10.1016/j.acha.2006.05.010
9. Bodmann, B.G., Paulsen, V.I., Abdulbaki, S.A.: Smooth frame-path termination for higher order sigma-delta quantization. J. Fourier Anal. Appl. **13**(3), 285–307 (2007). doi:10.1007/s00041-006-6032-y
10. Bodmann, B.G., Paulsen, V.I., Tomforde, M.: Equiangular tight frames from complex Seidel matrices containing cube roots of unity. Linear Algebra Appl. **430**(1), 396–417 (2009). doi:10.1016/j.laa.2008.08.002
11. Bodmann, B.G., Singh, P.K.: Burst erasures and the mean-square error for cyclic Parseval frames. IEEE Trans. Inf. Theory **57**(7), 4622–4635 (2011). doi:10.1109/TIT.2011.2146150
12. Bölcskei, H., Hlawatsch, F.: Noise reduction in oversampled filter banks using predictive quantization. IEEE Trans. Inf. Theory **47**(1), 155–172 (2001). doi:10.1109/18.904519
13. Bourgain, J., Dilworth, S., Ford, K., Konyagin, S., Kutzarova, D.: Explicit constructions of rip matrices and related problems. Duke Math. J. **159**(1), 145–185 (2011). doi:10.1215/00127094-1384809
14. Candes, E., Rudelson, M., Tao, T., Vershynin, R.: Error correction via linear programming. In: 46th Annual IEEE Symposium on Foundations of Computer Science. FOCS 2005, pp. 668–681 (2005). doi:10.1109/SFCS.2005.5464411
15. Candès, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006). doi:10.1002/cpa.20124
16. Candes, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006). doi:10.1109/TIT.2006.885507
17. Casazza, P.G., Kovačević, J.: Equal-norm tight frames with erasures. Adv. Comput. Math. **18**(2–4), 387–430 (2003). Frames. doi:10.1023/A:1021349819855
18. Casazza, P.G., Kutyniok, G.: Frames of subspaces. In: Wavelets, Frames and Operator Theory. Contemp. Math., vol. 345, pp. 87–113. Am. Math. Soc., Providence (2004)
19. Casazza, P.G., Kutyniok, G.: Robustness of fusion frames under erasures of subspaces and of local frame vectors. In: Radon Transforms, Geometry, and Wavelets. Contemp. Math., vol. 464, pp. 149–160. Am. Math. Soc., Providence (2008)
20. Casazza, P.G., Kutyniok, G., Li, S.: Fusion frames and distributed processing. Appl. Comput. Harmon. Anal. **25**(1), 114–132 (2008). doi:10.1016/j.acha.2007.10.001
21. Dasgupta, S., Gupta, A.: An elementary proof of a theorem of Johnson and Lindenstrauss. Random Struct. Algorithms **22**(1), 60–65 (2003). doi:10.1002/rsa.10073
22. Donoho, D.L., Huo, X.: Uncertainty principles and ideal atomic decomposition. IEEE Trans. Inf. Theory **47**(7), 2845–2862 (2001). doi:10.1109/18.959265
23. Donoho, D.L., Stark, P.B.: Uncertainty principles and signal recovery. SIAM J. Appl. Math. **49**(3), 906–931 (1989). doi:10.1137/0149053
24. Et-Taoui, B.: Equi-isoclinic planes of Euclidean spaces. Indag. Math. (N.S.) **17**(2), 205–219 (2006). doi:10.1016/S0019-3577(06)80016-9
25. Et-Taoui, B.: Equi-isoclinic planes in Euclidean even dimensional spaces. Adv. Geom. **7**(3), 379–384 (2007). doi:10.1515/ADVGEOM.2007.023
26. Et-Taoui, B., Fruchard, A.: Sous-espaces équi-isoclins de l'espace euclidien. Adv. Geom. **9**(4), 471–515 (2009). doi:10.1515/ADVGEOM.2009.029

27. Godsil, C.D., Hensel, A.D.: Distance regular covers of the complete graph. J. Comb. Theory, Ser. B **56**(2), 205–238 (1992). doi:10.1016/0095-8956(92)90019-T
28. Goyal, V.K., Kelner, J.A., Kovačević, J.: Multiple description vector quantization with a coarse lattice. IEEE Trans. Inf. Theory **48**(3), 781–788 (2002). doi:10.1109/18.986048
29. Goyal, V.K., Kovačević, J., Kelner, J.A.: Quantized frame expansions with erasures. Appl. Comput. Harmon. Anal. **10**(3), 203–233 (2001). doi:10.1006/acha.2000.0340
30. Goyal, V.K., Vetterli, M., Thao, N.T.: Quantized overcomplete expansions in $\mathbf{R}^N$: analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory **44**(1), 16–31 (1998). doi:10.1109/18.650985
31. Hoffman, T.R., Solazzo, J.P.: Complex equiangular tight frames and erasures, preprint, available at arxiv:1107.2267
32. Holmes, R.B., Paulsen, V.I.: Optimal frames for erasures. Linear Algebra Appl. **377**, 31–51 (2004). doi:10.1016/j.laa.2003.07.012
33. Kalra, D.: Complex equiangular cyclic frames and erasures. Linear Algebra Appl. **419**(2–3), 373–399 (2006). doi:10.1016/j.laa.2006.05.008
34. Kovačević, J., Dragotti, P.L., Goyal, V.K.: Filter bank frame expansions with erasures. IEEE Trans. Inf. Theory **48**(6), 1439–1450 (2002). Special issue on Shannon theory: perspective, trends, and applications. doi:10.1109/TIT.2002.1003832
35. Kutyniok, G., Pezeshki, A., Calderbank, R., Liu, T.: Robust dimension reduction, fusion frames, and Grassmannian packings. Appl. Comput. Harmon. Anal. **26**(1), 64–76 (2009). doi:10.1016/j.acha.2008.03.001
36. Lemmens, P.W.H., Seidel, J.J.: Equi-isoclinic subspaces of Euclidean spaces. Nederl. Akad. Wetensch. Proc. Ser. A 76=Indag. Math. **35**, 98–107 (1973)
37. Lopez, J., Han, D.: Optimal dual frames for erasures. Linear Algebra Appl. **432**(1), 471–482 (2010). doi:10.1016/j.laa.2009.08.031
38. Marshall, T.G. Jr.: Coding of real-number sequences for error correction: a digital signal processing problem. IEEE J. Sel. Areas Commun. **2**(2), 381–392 (1984). doi:10.1109/JSAC.1984.1146063
39. Marshall, T.: Fourier transform convolutional error-correcting codes. In: Twenty-Third Asilomar Conference on Signals, Systems and Computers, vol. 2, pp. 658–662 (1989). doi:10.1109/ACSSC.1989.1200980
40. Massey, P.G.: Optimal reconstruction systems for erasures and for the q-potential. Linear Algebra Appl. **431**(8), 1302–1316 (2009). doi:10.1016/j.laa.2009.05.001
41. Püschel, M., Kovačević, J.: Real, tight frames with maximal robustness to erasures. In: Data Compression Conference. Proceedings. DCC 2005, pp. 63–72 (2005). doi:10.1109/DCC.2005.77
42. Shannon, C.E.: A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948). 623–656
43. Shannon, C.E.: Communication in the presence of noise. Proc. I.R.E. **37**, 10–21 (1949)
44. Shannon, C.E., Weaver, W.: The Mathematical Theory of Communication. The University of Illinois Press, Urbana (1949)
45. Strohmer, T., Heath, R.W. Jr.: Grassmannian frames with applications to coding and communication. Appl. Comput. Harmon. Anal. **14**(3), 257–275 (2003). doi:10.1016/S1063-5203(03)00023-X
46. Tropp, J.A., Dhillon, I.S., Heath, R.W. Jr., Strohmer, T.: Designing structured tight frames via an alternating projection method. IEEE Trans. Inf. Theory **51**(1), 188–209 (2005). doi:10.1109/TIT.2004.839492
47. Vershynin, R.: Frame expansions with erasures: an approach through the non-commutative operator theory. Appl. Comput. Harmon. Anal. **18**(2), 167–176 (2005). doi:10.1016/j.acha.2004.12.001
48. Welch, L.: Lower bounds on the maximum cross correlation of signals (corresp.). IEEE Trans. Inf. Theory **20**(3), 397–399 (1974). doi:10.1109/TIT.1974.1055219
49. Xia, P., Zhou, S., Giannakis, G.: Achieving the Welch bound with difference sets. IEEE Trans. Inf. Theory **51**(5), 1900–1907 (2005). doi:10.1109/TIT.2005.846411

Chapter 8
Quantization and Finite Frames

Alexander M. Powell, Rayan Saab, and Özgür Yılmaz

Abstract Frames are a tool for providing stable and robust signal representations in a wide variety of pure and applied settings. Frame theory uses a set of frame vectors to discretely represent a signal in terms of its associated collection of frame coefficients. Dual frames and frame expansions allow one to reconstruct a signal from its frame coefficients—the use of redundant or overcomplete frames ensures that this process is robust against noise and other forms of data loss. Although frame expansions provide discrete signal decompositions, the frame coefficients generally take on a continuous range of values and must also undergo a lossy step to discretize their amplitudes so that they may be amenable to digital processing and storage. This analog-to-digital conversion step is known as quantization. We shall give a survey of quantization for the important practical case of finite frames and shall give particular emphasis to the class of Sigma-Delta algorithms and the role of noncanonical dual frame reconstruction.

Keywords Digital signal representations · Noncanonical dual frame · Quantization · Sigma-Delta quantization · Sobolev duals

8.1 Introduction

Data representation is crucial in modern signal processing applications. Among other things, one seeks signal representations that are numerically stable, robust against noise and data loss, computationally tractable, and well adapted to specific

A.M. Powell (✉)
Department of Mathematics, Vanderbilt University, Nashville, TN 37240, USA
e-mail: alexander.m.powell@vanderbilt.edu

R. Saab
Department of Mathematics, Duke University, Durham, NC 27708, USA
e-mail: rayans@math.duke.edu

Ö. Yılmaz
Department of Mathematics, University of British Columbia, Vancouver, BC Canada V6T 1Z2
e-mail: oyilmaz@math.ubc.ca

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_8,

applied problems. Frame theory has emerged as an important tool for meeting these requirements. Frames use redundancy or overcompleteness to provide robustness and design flexibility, and the linearity of frame expansions makes them simple to use in practice.

The linear representations given by frame expansions are a cornerstone of frame theory. If $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ is a frame for $\mathbb{R}^N$ and if $(\psi_i)_{i=1}^M \subset \mathbb{R}^N$ is any associated dual frame, then the following frame expansion holds:

$$\forall x \in \mathbb{R}^N, \quad x = \sum_{i=1}^{M} \langle x, \varphi_i \rangle \psi_i. \tag{8.1}$$

Equivalently, if Φ^* is the analysis operator associated to $(\varphi_i)_{i=1}^M$ and Ψ is the synthesis operator associated to $(\psi_i)_{i=1}^M$, then

$$\forall x \in \mathbb{R}^N, \quad x = \Psi \Phi^* x. \tag{8.2}$$

The frame expansion (8.1) discretely *encodes* $x \in \mathbb{R}^N$ by the frame coefficients $(\langle x, \varphi_i \rangle)_{i=1}^M$. Consequently, frame expansions can be interpreted as generalized sampling formulas, where frame coefficients play the role of samples of the underlying object. Our technology nowadays is overwhelmingly digital, and therefore for a sampling theory to be practicable, it needs to be accompanied by a *quantization theory*. In general, quantization refers to the process by which one converts an object in the continuum into a finite bitstream, i.e., into a finite sequence of elements in $\{0, 1\}$. Typically, this is done by replacing the underlying object by an element from a finite set called the quantization alphabet (since the alphabet is finite, its elements can ultimately be given a binary encoding).

Our survey of quantization for finite frames will cover several different quantization methods for finite frames. Our main emphasis will be on the following methods, which will be covered in detail.

- *Memoryless scalar quantization (MSQ)*: This is a simple classical method but is not particularly adept at exploiting the redundancy present in frames.
- *First order Sigma-Delta* ($\Sigma\Delta$) *quantization*: This is a more sophisticated low complexity approach which effectively exploits redundancy but still leaves much room for theoretical improvements.
- *Higher order Sigma-Delta* ($\Sigma\Delta$) *quantization*: This method yields strong error bounds at the cost of increased complexity by exploiting a class of noncanonical dual frames known as Sobolev duals.

Before discussing the above methods we begin with a hands-on formulation of two quantization problems of interest.

8.1.1 Quantization Problem: Synthesis Formulation

Fix a frame $\Psi = (\psi_i)_{i=1}^M$ for $\mathbb{R}^N$. From here on, N denotes the dimension of the ambient space and $M \geq N$ is the number of frame vectors. Furthermore, we abuse the notation and use Ψ to denote both the frame $(\psi_i)_{i=1}^M$ and the associated $N \times M$ synthesis matrix. Let $\mathscr{A}$ be a finite set which is called the *quantization alphabet*. The goal is to represent a given $x \in \mathbb{R}^N$ via an expansion of the form (8.1) where the coefficients $\langle x, \varphi_i \rangle$ are replaced by elements of $\mathscr{A}$. More precisely, we quantize $x \in \mathbb{R}^N$ by replacing it with an element of the *constellation* $\Gamma(F, \mathscr{A}) := \{\Psi q : q \in \mathscr{A}^M\}$. In this setting, the objective is as follows.

> *QP-Synthesis*. Given a bounded set $\mathscr{B} \in \mathbb{R}^N$ and a frame Ψ for $\mathbb{R}^N$, find a map $\mathscr{Q} : \mathbb{R}^N \mapsto \mathscr{A}^M$—the quantizer—such that the distortion $\mathscr{E}(x) := \|x - \Psi \mathscr{Q}(x)\|$ is "small" on $\mathscr{B}$ in some norm (deterministic setting) or in expectation (probabilistic setting).

Consequently, the *optimal* quantizer (for a given norm $\|\cdot\|$) is defined by

$$\mathscr{Q}_{\text{opt}}(x; \Psi, \mathscr{A}) = \arg\min_{q \in \mathscr{A}^M} \big\{ \|x - \Psi q\| \big\}.$$

This formulation (QP-Synthesis) of the frame quantization problem arises in efforts to reduce "computational noise" in the calculation of fast Fourier transforms (FFTs) by using algebraic integers in the computation [13, 28]. In particular, the proposed approach is based on solving QP-Synthesis with $N = 2$ and the underlying frames Ψ_M that are given by the Mth roots of unity, i.e., $\Psi_M = (\psi_j)_{j=1}^M$ with $\psi_j = [\cos \frac{2\pi}{M} j, \sin \frac{2\pi}{M} j]^T$. In Fig. 8.1 we show the set $\Gamma(\Psi_M, \mathscr{A})$ and $\mathscr{A} = \{\pm 1\}$ for various values of M. In this specific instance of QP-Synthesis, there are partial results. For example, when M is an integer power of 2, [13] shows that the distortion $\mathscr{E}$ decays exponentially as M increases, at least for certain alphabets. Furthermore, [13] also proposes an algorithm that (nearly) implements $\mathscr{Q}_{\text{opt}}$. However, for general M, both these problems—computationally tractable implementation of $\mathscr{Q}_{\text{opt}}$ and decay rate of optimal accuracy $\mathscr{E}$ as M grows—are, to our knowledge, open even in the case of the roots-of-unity frames Ψ_M. Our focus in the remainder of the chapter will be on the analysis formulation of the quantization problem, which we describe next.

8.1.2 Quantization Problem: Analysis Formulation

Let Φ be a frame for $\mathbb{R}^N$, again with M frame vectors. Suppose that we have access to frame coefficients y_i where $y = [y_1, \ldots, y_M]^T = \Phi^* x$. In practice, y_i could be sensor readings [25], samples of a continuous function on a finite interval [22], or "compressive samples" of a sparse object—see [34]. Here y_i are in general real numbers that are assumed to be known with infinite accuracy. To be amenable to

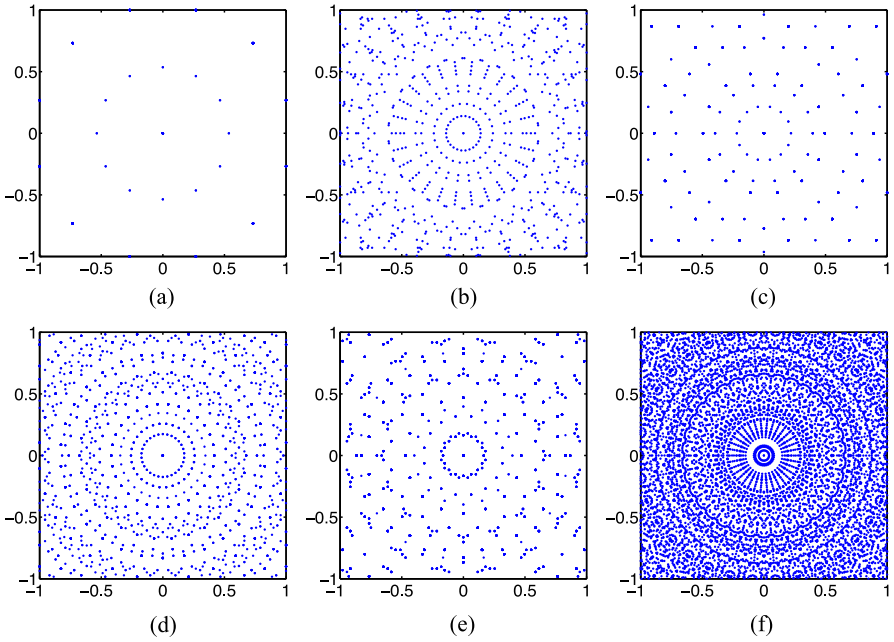


Fig. 8.1 The set $\Gamma(\Psi_M, \mathscr{A}) \cap [-1, 1]^2$ where Ψ_M is the frame for $\mathbb{R}^2$ given by the Mth roots of unity—$M = 9, \ldots, 14$ in (a),...,(f)

digital processing and storage, the frame coefficients y_i must further be quantized to lie in a given alphabet $\mathscr{A}$. Since the ultimate goal is to obtain a digital approximation of x, one straightforward approach is to reconstruct x *before quantization* via $x = \Psi y$. Ψ is a dual of Φ so the reconstruction is exact. Once we have x, we can compute an orthonormal-basis expansion of x and round off each coefficient to an accuracy level that depends on the overall bit budget. This approach would bypass the difficulties that arise from the fact that the original frame is possibly oblique and redundant, and it generally yields a more accurate digital approximation of x than any known method where the original frame coefficients are replaced with quantized ones in (8.1).

Unfortunately, the above-described approach is not viable in typical practical settings for several reasons. First of all, it requires sophisticated high-accuracy analog computations, which is generally not practicable. Furthermore, in applications where the frame coefficients are obtained in a sequential manner, e.g., when sampling a continuous function or when collecting measurements from a large number of sensors, this approach requires that a large number (M) of analog quantities, i.e., real numbers, be stored in memory on analog hardware, which is often not feasible in practice. Finally, in many applications redundant frames are preferred over orthonormal bases because the built-in redundancy makes the associated expansions more robust to various sources of errors, such as additive noise, partial loss of data (in the case of transmission over erasure channels) [8, 15, 29], and quantization

with imperfect circuit elements, for example, in the case of oversampled bandlimited functions [21, 22, 45]. Taking into account all these factors, it is important to consider the following formulation of the frame quantization problem.

QP-Analysis. Given a bounded set $\mathscr{B} \in \mathbb{R}^N$ and a frame Φ for $\mathbb{R}^N$ with M vectors, find a map $\mathscr{Q} : \mathbb{R}^M \mapsto \mathscr{A}^M$—the quantizer—such that

1. $\mathscr{Q}$ acts on the frame coefficients of $x \in \mathbb{R}^N$, given by $y = \Phi^* x$.
2. $\mathscr{Q}$ is "causal"; i.e., the quantized value $q_j = \mathscr{Q}(y)_j$ depends only on $y_1, \ldots, y_j$ and $q_1, \ldots, q_{j-1}$. To avoid the need to use a large number of analog memory elements, one may further enforce that $\mathscr{Q}$ depends only on y_j and on $r \ll M$ previously computed elements, i.e., on an r-dimensional "state vector."
3. The distortion $\widetilde{\mathscr{D}}(x) := \|x - \mathscr{G}\mathscr{Q}(\Phi^* x)\|$ is "small" on $\mathscr{B}$ in some norm (deterministic setting) or in expectation (probabilistic setting). Here $\mathscr{G} : \mathscr{A}^M \mapsto \mathbb{R}^N$ is some decoder possibly tailored to the quantizer $\mathscr{Q}$ and the frame Φ. A natural choice for such a decoder is motivated by frame theory and is given by Ψ, some dual of Φ (also possibly tailored to the quantizer $\mathscr{Q}$). Such a decoder corresponds to linear reconstruction of x from its quantized coefficients.

8.1.3 Stylized Example: Memoryless Scalar Quantization

To illustrate the challenges posed by QP-Analysis, consider the following example. Let Φ_M be the frame for $\mathbb{R}^2$ given by the Mth roots of unity, let $\mathscr{B} \subset \mathbb{R}^2$ be the unit disk, and consider the 1-bit quantization alphabet $\mathscr{A}_1 = \{\pm 1\}$. First, we quantize the frame coefficients $y = \Phi_{12}^* x$ for any $x \in \mathscr{B}$ using a *memoryless scalar quantizer (MSQ)*; i.e., each y_j is quantized to the element of $\mathscr{A}$ that is closest to it, which in this particular case corresponds to $q_j = \text{sign}(y_j)$. Note that q_j only depends on the jth frame coefficient; hence the quantizer is memoryless. In Fig. 8.2(a), we show the quantizer cells that correspond to the quantizer described above, i.e., a 1-bit MSQ. Every cell, identified with a distinct color, consists of vectors with identical quantized coefficients. In other words, after quantization we cannot distinguish between the vectors in the same cell, and therefore the diameters of these cells reflect the ultimate distortion bounds for the given quantizer, in this case $\mathscr{Q}_{\text{MSQ}}^{\mathscr{A}_1}$ with $\mathscr{A}_1 = \{\pm 1\}$ as described above. Note that out of 2^{12} possible 1-bit quantized sequences, $\mathscr{Q}_{\text{MSQ}}^{\mathscr{A}_1}$ uses only 12 distinct ones. Furthermore, since the cells are convex, ideally all points in a given cell should be quantized to a "representative point" that falls inside the respective cell and is at a location that minimizes the distortion (in a norm of choice). In Fig. 8.2(a), we also show the reconstructed vector for each cell obtained via $x_{\text{rec}} = \Psi \mathscr{Q}_{\text{MSQ}}^{\mathscr{A}_1}(\Phi_{12}^* x)$, where $\Psi = \frac{1}{12}\Phi_{12}$ is the canonical dual of Φ_{12}. Such a reconstruction method is referred to as *linear reconstruction* using the canonical dual.

In Fig. 8.2(b), we repeat the experiment described above with a 2-bit MSQ; i.e., the alphabet in this case is $\mathscr{A}_2 := \{\pm 1, \pm\frac{1}{3}\}$. We observe that the number of cells

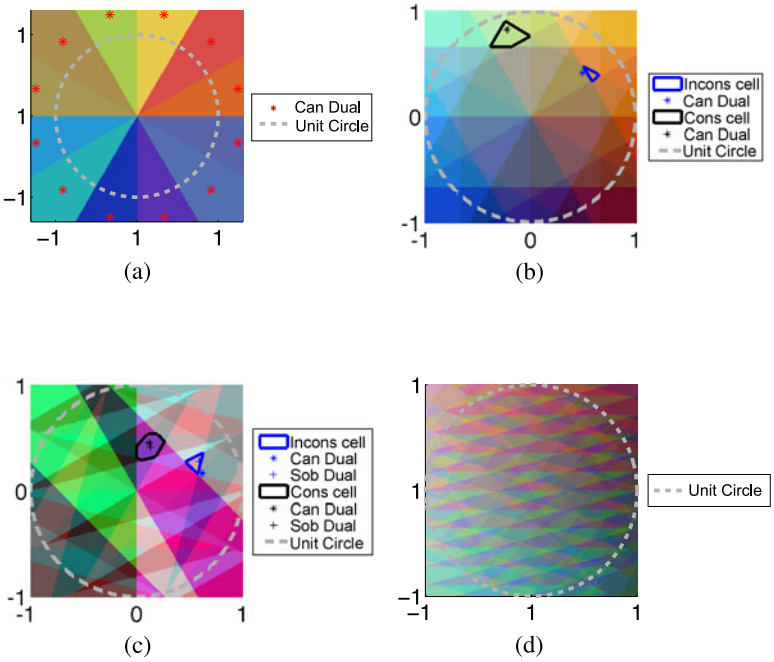


Fig. 8.2 12 cells for 1-bit MSQ, 204 cells for 1-bit $\Sigma\Delta$, 84 cells for 2-bit MSQ, 1844 cells for 2-bit $\Sigma\Delta$ (empirical counts)

increases substantially, from 12 to 84 distinct cells. However, still a very small fraction of 4^{12} possible quantized sequences are utilized. Another important point is that with 2-bit MSQ some cells are not *consistent* under linear reconstruction using the canonical dual. That is, points in these cells are quantized to a point that is outside the cell, and consequently, for x in such a cell and $\hat{x} = \Psi\mathcal{Q}(\Phi^* x)$, we have $\mathcal{Q}(\Phi^*(\hat{x})) \neq \mathcal{Q}(\Phi^* x)$. In Fig. 8.2(b) we marked two cells that exemplify consistent and inconsistent cells. Of course, alternative *nonlinear* reconstruction techniques could be used to enforce *consistent reconstruction*, which we discuss in detail in Sect. 8.2.3.

8.1.4 Stylized Example: $\Sigma\Delta$ Quantization

The strikingly small number of cells in Fig. 8.2(a) and Fig. 8.2(b) hint that MSQ may not be well suited to quantize frame expansions (in the QP-Analysis sense)—see Sect. 8.2 for a survey of MSQ in frame quantization and its performance limitations.

An alternative approach to frame quantization is to use the Sigma-Delta ($\Sigma\Delta$) quantizers. $\Sigma\Delta$ quantizers are widely used in analog-to-digital (A/D) conversion for oversampled bandlimited functions, and recently have been adopted for quantizing arbitrary frame expansions; see, e.g., [2]. Most of our chapter is dedicated to

the analysis of the performance of these quantizers in addressing QP-Analysis for various families of frames—see Sects. 8.3–8.5. It turns out that these quantizers outperform MSQ (even with optimal consistent reconstruction) if the underlying frame is sufficiently redundant; see, e.g., [3, 5, 42]. Here we repeat the above-described experiments with $\Sigma\Delta$ quantizers in place of MSQ. Figures 8.2(c) and 8.2(d) show the quantization cells corresponding to 1-bit (with alphabet $\mathscr{A}_1$) and 2-bit (with alphabet $\mathscr{A}_2$) first order $\Sigma\Delta$ quantizers, respectively. Even though the schemes use the same alphabets as the 1-bit and 2-bit MSQ, the number of distinct cells is significantly larger in the case of $\Sigma\Delta$ schemes: 204 cells for 1-bit $\Sigma\Delta$ (cf. 12 cells for 1-bit MSQ) and 1844 cells for 2-bit $\Sigma\Delta$ (cf. 84 cells for 2-bit MSQ). In Fig. 8.2(c), we again show a consistent cell and an inconsistent cell, together with the linear reconstructions using the canonical dual of Φ_{12}. In addition, we show alternative linear reconstructions obtained using the *Sobolev dual* of Φ_{12}. Sobolev duals are alternate duals that are designed to reduce the quantization error in the specific case of $\Sigma\Delta$ quantizers—see Sect. 8.4. Note that for the "inconsistent cell" in Fig. 8.2, while the canonical dual reconstruction is not consistent, the reconstruction obtained using the Sobolev dual is consistent (although this is by no means guaranteed in general).

QP-Analysis is relevant in various practical applications. The quintessential example is high resolution A/D conversion of bandlimited signals. To overcome physical constraints that limit the accuracy of the "binary decision elements," one common strategy is to use noise shaping analog-to-digital converters (ADCs). These ADCs—mostly based on $\Sigma\Delta$ quantization—first oversample the bandlimited function, effectively collecting frame coefficients with respect to a redundant frame. Then this redundancy is exploited to design quantization strategies that are robust with respect to implementation errors. Specifically, the family of $\Sigma\Delta$ quantizers achieve this goal successfully: they can be implemented with low accuracy circuit elements and still yield a high bit depth.

Another example of QP-Analysis arises in compressed sensing, where the classically separate steps of measurement (encoding) and data compression are combined into a single step to provide efficient digital representations of sparse signals in high dimensions. We shall briefly describe some connections between $\Sigma\Delta$ quantization, noncanonical dual frames, and compressed sensing in Sect. 8.4.3.

In this chapter we shall provide a survey of quantization for finite frames, where our main focus is on QP-Analysis. Consequently, we contain our discussion to a framework with three main steps: encoding, quantization, and reconstruction. These three steps discussed above may be summarized as follows.

$$\begin{aligned}
&\text{Encoding:} && x \in \mathbb{R}^N \longmapsto (\langle x, \varphi_i \rangle)_{i=1}^M \in \mathbb{R}^M \\
&\text{Quantization:} && (\langle x, \varphi_i \rangle)_{i=1}^M \in \mathbb{R}^M \longmapsto (q_i)_{i=1}^M \in \mathscr{A}^M \\
&\text{Reconstruction:} && (q_i)_{i=1}^M \in \mathscr{A}^M \longmapsto \widetilde{x} \in \mathbb{R}^N
\end{aligned}$$

Throughout this chapter the encoding step will be done using a finite frame to compute frame coefficients. The redundancy introduced by encoding $x \in \mathbb{R}^N$ with frame coefficients $(\langle x, \varphi_i \rangle)_{i=1}^M$, generally with $M > N$, will play an important role in mitigating the losses incurred by quantization. Our survey of the quantization step will

primarily focus on two different methods that we considered in the stylized examples above: (i) memoryless scalar quantization and (ii) $\Sigma\Delta$ quantization. The reconstruction step is intimately linked to both the encoding and quantization steps. Motivated by frame theoretic considerations, our discussion of the reconstruction step will mainly focus on linear methods and shall describe the important role that various choices of dual frames play in quantization problems.

In particular, we will see that while MSQ may be attractive for its simplicity, it does not exploit the redundancy implicit in frame representations and thus does not provide error guarantees that decay well with oversampling. On the other hand, $\Sigma\Delta$ quantization is only slightly more complex computationally, but it exploits redundancy. Thus it provides error guarantees that decay well with oversampling, particularly when higher order schemes are used in conjunction with appropriate reconstruction methods.

8.2 Memoryless Scalar Quantization

The *scalar quantizer* is a basic component of quantization algorithms. Given a finite set $\mathscr{A} \subset \mathbb{R}$, called a *quantization alphabet*, the associated scalar quantizer is the function $Q : \mathbb{R} \to \mathscr{A}$ defined by

$$Q(u) = \arg\min_{a \in \mathscr{A}} |u - a|. \tag{8.3}$$

In other words, Q quantizes real numbers by rounding them to the nearest element of the quantization alphabet. There will be finitely many values of $u \in \mathbb{R}$, i.e., midpoints of quantization bins, for which the minimizer defining $Q(u)$ is not unique. In this case, there will be two possible choices of the minimizer, and one may arbitrarily pick one for the definition of $Q(u)$.

For concreteness of our discussion it will be convenient to work with a specific uniform quantization alphabet throughout this chapter. Fix a positive integer L and $\delta > 0$ and define the $(2L + 1)$ level midtread quantization alphabet with stepsize δ as the finite set of numbers

$$\mathscr{A} = \mathscr{A}_L^\delta = \{-L\delta, \ldots, -\delta, 0, \delta, \ldots, L\delta\}. \tag{8.4}$$

Unless otherwise stated, throughout this chapter we will work with the *midtread* alphabet (8.4), but in most cases other alphabets work equally well. For example, the closely related $2L$ level *midrise* alphabet with stepsize δ defined by

$$\left\{-(2L + 1)\delta/2, \ldots, -\delta/2, \delta/2, \ldots, (2L + 1)\delta/2\right\} \tag{8.5}$$

is also commonly used, especially in coarse quantization with a 1-bit alphabet such as $\{-1, +1\}$.

8.2.1 Memoryless Scalar Quantization of Frame Coefficients

Let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ be a frame for $\mathbb{R}^N$. The most basic approach to quantizing frame coefficients $(\langle x, \varphi_i\rangle)_{i=1}^M$ is to individually quantize each coefficient $y_i = \langle x, \varphi_i\rangle$ by

$$q_i = Q(y_i) = Q\big(\langle x, \varphi_i\rangle\big). \tag{8.6}$$

This step is referred to as *memoryless scalar quantization* (MSQ). A simple method for signal reconstruction from the MSQ quantized coefficients $(q_i)_{i=1}^M$ is to fix a dual frame $(\psi_i)_{i=1}^M \subset \mathbb{R}^N$ associated to $(\varphi_i)_{i=1}^M$ and use linear reconstruction by

$$\widetilde{x} = \sum_{i=1}^M q_i \psi_i. \tag{8.7}$$

With the alphabet $\mathscr{A}_L^\delta$ we may quantify the reconstruction error $\|x - \widetilde{x}\|$ associated to MSQ with linear reconstruction in (8.6) and (8.7) as follows. Let $C = \max_{1\le i\le M} \|\varphi_i\|$. If $x \in \mathbb{R}^N$ satisfies $\|x\| < (L+1/2)/C$ then $|y_i| = |\langle x, \varphi_i\rangle| \le (L+1/2)$ and the quantizer remains unsaturated, i.e., the following holds:

$$\forall 1 \le i \le M, \quad |y_i - q_i| = \big|y_i - Q(y_i)\big| \le \delta/2. \tag{8.8}$$

Consequently, the linear reconstruction $\widetilde{x} \in \mathbb{R}^N$ by (8.7) satisfies the simple bound

$$\|x - \widetilde{x}\| = \left\| \sum_{i=1}^M \big(\langle x, \varphi_i\rangle - q_i\big)\psi_i \right\| \le \frac{\delta}{2} \sum_{i=1}^M \|\psi_i\|. \tag{8.9}$$

In the special case where $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ is a unit norm tight frame and $\psi_i = \frac{N}{M}\varphi_i$ is taken as the canonical dual frame, then the error bound (8.9) reduces to

$$\|x - \widetilde{x}\| \le \frac{\delta N}{2}. \tag{8.10}$$

As one would expect, this error bound shows that a finer quantization alphabet, i.e., taking $\delta > 0$ smaller, results in more accurate quantization. However, the role of the frame size M is conspicuously absent in this bound.

It will become apparent in later sections that, in general, neither MSQ nor linear reconstruction is optimal for quantization in any sense. However, for the special case when $(b_i)_{i=1}^N \subset \mathbb{R}^N$ is an orthonormal basis and $\psi_i = b_i$ is the (in this case unique) dual frame, then it follows from Parseval's equality that MSQ is optimal if one insists on linear reconstruction. In particular, if $(q_i)_{i=1}^N \subset \mathscr{A}$ is *arbitrary* and $\widetilde{x} = \sum_{i=1}^N q_i b_i$, then

$$\|x - \widetilde{x}\|^2 = \left\| \sum_{i=1}^N \big(\langle x, b_i\rangle - q_i\big) b_i \right\|^2 = \sum_{i=1}^N \big|\langle x, b_i\rangle - q_i\big|^2. \tag{8.11}$$

This error is minimized by taking $q_i = Q(\langle x, b_i \rangle)$, which shows that MSQ is optimal for orthonormal bases when linear reconstruction is used. On the other hand, the simple upper bound in (8.10) is not sharp even for orthonormal bases since in this case (8.11) yields

$$\|x - \widetilde{x}\| \leq \frac{\delta\sqrt{N}}{2}. \tag{8.12}$$

From the point of view of frame theory, an important shortcoming of the bound (8.9) is that it does not utilize a frame's redundancy. The redundancy of a frame can very directly translate into increased robustness against noise, but the upper bound (8.9) does not improve if the frame is taken to be more redundant; i.e., (8.9) does not improve when the dimension N is fixed and the frame size M increases. This indicates that MSQ is not particularly well suited for quantizing redundant collections of frame coefficients. For an intuitive understanding of this, note that MSQ nonadaptively quantizes each frame coefficient $\langle x, e_i \rangle$ without any regard for how other frame coefficients are quantized, and thus MSQ is not able to make very effective use of the correlations present among frame coefficients.

It is easy to produce concrete examples where frame redundancy does not improve the performance of MSQ. Let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ be any unit norm frame and assume that one uses the scalar quantizer with the midtread alphabet $\mathscr{A}_L^\delta$ given by (8.4). For any $x \in \mathbb{R}^N$ with $\|x\| < \delta/2$ it then holds that $q_i = Q(\langle x, \varphi_i \rangle) = 0$. In particular, for any dual frame $(\psi_i)_{i=1}^M$ the linear reconstruction (8.7) gives $\widetilde{x} = 0$. Thus, regardless of how redundant the frame $(\varphi_i)_{i=1}^M$ is, the associated quantization error satisfies $\|x - \widetilde{x}\| = \|x\|$. This example is overly simple but nonetheless illustrates some basic shortcomings of MSQ. The recent work in [53] provides a thorough and detailed technical investigation into the difficulties that MSQ faces in utilizing frame redundancy. See also [14, 26] for work on quantization in Banach spaces.

8.2.2 Noise Models and Dual Frames

Error bounds such as (8.9) and (8.10) provide *worst case* upper bounds on quantization error and only suggest that quantization error can be decreased by choosing a quantizer with finer stepsize $\delta > 0$. Worst case error bounds play an important role in quantizer analysis, but in practice one often also observes an *average error* that is much smaller than the worst case predictions. The uniform noise model is an important tool for understanding average quantization error.

8.2.2.1 The uniform noise model

Let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ be a frame for $\mathbb{R}^N$ and let $x \in \mathbb{R}^N$ have frame coefficients $y_i = \langle x, \varphi_i \rangle$, for $1 \leq i \leq M$. When the midtread scalar quantizer (8.4) operates in

its unsaturated regime, Eq. (8.8) shows that the individual coefficient quantization errors $\eta_i = y_i - Q(y_i)$ satisfy

$$\eta_i = y_i - Q(y_i) \in [-\delta/2, \delta/2].$$

Uniform noise models go one step further than this and posit that $(\eta_i)_{i=1}^M$ should *on average* be quite uniformly spread out in $[-\delta/2, \delta/2]$. This leads one to randomly model the individual coefficient quantization errors $(\eta_i)_{i=1}^M$ as independent identically distributed (i.i.d.) uniform random variables on $[-\delta/2, \delta/2]$.

> *Uniform noise model*: Treat the quantization errors $\eta_i = y_i - Q(y_i)$, $1 \leq i \leq M$, as i.i.d. uniform random variables on $[-\delta/2, \delta/2]$.

The uniform noise model has a long history that dates back to Bennett's 1940s work in [4] and has been widely used as a tool in the engineering literature. The uniform noise model has been observed to be empirically reasonable, but it also suffers from known theoretical shortcomings; see, e.g., [39]. Since quantization is a deterministic process, some additional assumptions will be needed to justify the introduction of randomness in the uniform noise model. We shall briefly discuss two general approaches commonly used to justify the uniform noise model: (i) dithering and (ii) high resolution asymptotics.

Dithering is the process of deliberately injecting noise into a quantization system to beneficially reshape the properties of the individual errors $(\eta_i)_{i=1}^M$. For an overview of the large theoretical and applied literature on dithering, see [7, 31] and the references therein. We shall briefly discuss one particular dithering method known as *subtractive dither*, which is used to justify the uniform noise model. For the quantization step we assume that we have available a sequence $(\varepsilon_i)_{i=1}^M$ of i.i.d. uniform random variables on $[-\delta/2, \delta/2]$. This sequence $(\varepsilon_i)_{i=1}^M$ is known as the *dither sequence*. To quantize a sequence of frame coefficients $(y_i)_{i=1}^M$ one uses MSQ to quantize the dithered coefficients $y_i + \varepsilon_i$ as $q_i = Q(y_i + \varepsilon_i)$. This quantized sequence $(q_i)_{i=1}^M$ provides the desired digital representation of the coefficients $(y_i)_{i=1}^M$. To reconstruct a signal from $(q_i)_{i=1}^M$ in a manner that respects the uniform noise model one must first subtractively remove the dither sequence to obtain $\widetilde{y}_i = q_i - \varepsilon_i$. The individual coefficient quantization errors $y_i - \widetilde{y}_i$ then satisfy

$$y_i - \widetilde{y}_i = y_i - (q_i - \varepsilon_i) = (y_i + \varepsilon_i) - Q(y_i + \varepsilon_i).$$

In particular, if $(y_i)_{i=1}^M$ is any deterministic sequence, it follows that $(y_i - \widetilde{y}_i)_{i=1}^M$ are i.i.d. uniform random variables on $[-\delta/2, \delta/2]$. An obvious practical issue with this method is that it requires (infinite precision) knowledge of the dither sequence at both the quantizer and reconstruction stages.

High resolution asymptotics provide a different approach to justifying the uniform noise model. Here one introduces randomness by assuming that the signal $x \in \mathbb{R}^N$ that will be quantized is an absolutely continuous random vector supported on the unit ball of $\mathbb{R}^N$. We let Q_δ denote the midtread quantizer with stepsize δ

and with $L = 1/\lfloor\delta\rfloor$. Let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ be a frame for $\mathbb{R}^N$ and consider the M-dimensional random vector of normalized quantization errors

$$V_\delta = \delta^{-1}\big[\langle x, e_1\rangle - Q_\delta\big(\langle x, e_1\rangle\big), \ldots, \langle x, \varphi_M\rangle - Q_\delta\big(\langle x, \varphi_M\rangle\big)\big]. \tag{8.13}$$

It is proven in [39] that under suitable conditions on the frame $(\varphi_i)_{i=1}^M$ the normalized error vector V_δ converges in distribution to the uniform distribution on $[-1/2, 1/2]^M$ as $\delta \to 0$. This provides a rigorous justification of the uniform noise model in the high resolution limit as $\delta \to 0$. For related work in the setting of lattice quantizers, see [11]. On the other hand, this approach generally only holds asymptotically, since it is shown in [39] that for fixed $\delta > 0$ and $M > N$ the entries of V_δ are never independent. Moreover, while high resolution asymptotics provide elegant and mathematically rigorous results, they may not always be easy to apply to specific practical settings, since the frame $(\varphi_i)_{i=1}^M$ is required to be held fixed. For example, if one wishes to understand how the performance of a quantizer changes when increasingly redundant frames are used, then high resolution asymptotics might not be appropriate.

8.2.2.2 Dual frames and MSQ

We now consider frame theoretic issues which arise when analyzing MSQ under the uniform noise model. We shall freely use the uniform noise model in this section, but the reader should keep in mind the noise model's mathematical limitations and the issues involved in rigorously justifying it. The results obtained under the uniform noise model are a valuable source of intuition on quantization.

Let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ be a frame for $\mathbb{R}^N$ and suppose that the frame coefficients $y_i = \langle x, \varphi_i\rangle$ are quantized to $q_i = Q(y_i)$ using MSQ. We assume that the sequence $\eta_i = y_i - q_i$, with $1 \le i \le M$, satisfies the uniform noise model. Suppose that one reconstructs $\widetilde{x} \in \mathbb{R}^N$ from the quantized coefficients $(q_i)_{i=1}^M$ using a dual frame $(\psi_i)_{i=1}^M$ of $(\varphi_i)_{i=1}^M$ as follows:

$$\widetilde{x} = \sum_{i=1}^M q_i \psi_i. \tag{8.14}$$

A simple computation shows that the mean squared error (*MSE*) satisfies

$$MSE = \mathbb{E}\|x - \widetilde{x}\|^2 = \sum_{i=1}^M \sum_{j=1}^M \mathbb{E}[\eta_i \eta_j]\langle \psi_i, \psi_j\rangle = \frac{\delta^2}{12} \sum_{i=1}^M \|\psi_i\|^2. \tag{8.15}$$

In particular, if $(\varphi_i)_{i=1}^M$ is a unit norm tight frame and $\psi_i = \widetilde{e}_i = \frac{N}{M}\varphi_i$ is its canonical dual frame, then

$$\mathbb{E}\|x - \widetilde{x}\|^2 = \frac{N^2\delta^2}{12M}. \tag{8.16}$$

In contrast to the worst case bound (8.10), the mean squared error (8.16) decreases when a more redundant unit norm tight frame is used, i.e., when M increases. This shows that frame theory and redundancy play an important role in error reduction in quantization problems, and it hints at the more rigorous and more precise error bounds possible with sophisticated algorithms such as $\Sigma\Delta$ quantization—see Sects. 8.3–8.5.

The mean squared error bound (8.15) depends strongly on the choice of dual frame $(\psi_i)_{i=1}^M$. It is natural to ask which choice of dual frame is best for the linear reconstruction in (8.14). The following classical proposition shows that the canonical dual frame is optimal for memoryless scalar quantization under the uniform noise model; e.g., see [5, 29].

Proposition 8.1 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathbb{R}^N$. Consider the minimization problem*

$$\min\left\{\sum_{i=1}^M \|\psi_i\|^2 : (\psi_i)_{i=1}^M \text{ a dual frame associated to } (\varphi_i)_{i=1}^M\right\}. \tag{8.17}$$

The dual frame $(\psi_i)_{i=1}^M$ is a minimizer of (8.17) if and only if $(\psi_i)_{i=1}^M$ is the canonical dual frame of $(\varphi_i)_{i=1}^M$.

The frame problem (8.17) may equivalently be stated in matrix form using the $M \times N$ analysis operator Φ^* and the $N \times M$ synthesis operator Ψ associated to the respective frame $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ and dual frame $(\psi_i)_{i=1}^M$. In matrix form, (8.17) becomes

$$\min\left\{\|\Psi\|_{\text{Frob}}^2 : \Psi\Phi^* = I\right\}. \tag{8.18}$$

In this form, Proposition 8.1 now states that: the matrix Ψ is a minimizer of (8.18) if and only if $\Psi = (\Phi^*)^\dagger = (\Phi\Phi^*)^{-1}\Phi$ is the canonical left inverse of Φ^*.

When the canonical dual frame $\psi_i = \widetilde{\varphi}_i$ is used in (8.14) the mean squared error bound (8.15) becomes

$$\mathbb{E}\|x - \widetilde{x}\|^2 = \frac{\delta^2}{12}\sum_{i=1}^M \|\widetilde{\varphi}_i\|^2. \tag{8.19}$$

At this point, having established that the canonical dual frame is optimal for the reconstruction step, the error bound (8.19) still depends strongly on the original frame $(\varphi_i)_{i=1}^M$ through its canonical dual frame. A natural follow-up question to Proposition 8.1 is to ask which frames $(\varphi_i)_{i=1}^M$ are optimal for the encoding step. For this question to be meaningful one must impose some restrictions on the norms of the frame involved. Otherwise, rescaling a fixed frame $(\varphi_i)_{i=1}^M$ trivially allows $\sum_{i=1}^M \|\widetilde{\varphi}_i\|^2$ in (8.19) to become arbitrarily close to zero. More precisely, if $(\varphi_i)_{i=1}^M$ has canonical dual frame $(\widetilde{\varphi}_i)_{i=1}^M$, then the rescaled frame $(c\varphi_i)_{i=1}^M$ has canonical dual frame $(c^{-1}\widetilde{\varphi}_i)_{i=1}^M$.

The following theorem shows that if one restricts the encoding frame to be unit norm and uses the (optimal) canonical dual for reconstruction, then for MSQ under

the uniform noise model an optimal choice for the encoding frame is to take any unit norm tight frame, see [29].

Theorem 8.1 *Let M and N be fixed, and consider the minimization problem*

$$\min\left\{\sum_{i=1}^{M}\|\widetilde{\varphi}_i\|^2 : (\varphi_i)_{i=1}^M \subset \mathbb{R}^N \text{ is a unit norm frame}\right\}. \tag{8.20}$$

A unit norm frame $(\varphi_i)_{i=1}^M$ is a minimizer of (8.20) if and only if $(\varphi_i)_{i=1}^M$ is a unit norm tight frame for $\mathbb{R}^N$.

The frame problem (8.20) may equivalently be stated in matrix form using the $M \times N$ analysis operator Φ^* associated to $(\varphi_i)_{i=1}^N \subset \mathbb{R}^N$ and the $N \times M$ canonical left inverse $(\Phi^*)^\dagger = (\Phi\Phi^*)^{-1}\Phi$ as follows:

$$\min\left\{\left\|\left(\Phi^*\right)^\dagger\right\|_{\text{Frob}}^2 : \operatorname{rank}(\Phi) = N \text{ and } \operatorname{diag}\left(\Phi^*\Phi\right) = I\right\}. \tag{8.21}$$

Here $\|\cdot\|_{\text{Frob}}$ denotes the Frobenius norm. In this form, Theorem 8.1 now states: the full rank matrix Φ with $\operatorname{diag}(\Phi^*\Phi) = I$ is a minimizer of (8.21) if and only if $\Phi\Phi^* = (\frac{N}{M})I$ and $\operatorname{diag}(\Phi^*\Phi) = I$.

Consequently, combining Proposition 8.1 and Theorem 8.1 shows that for a fixed frame size M in dimension N, MSQ under the uniform noise model performs optimally when a unit norm tight frame is used for the encoding step and the canonical dual frame is used for linear reconstruction. Moreover, in this case the associated optimal error bound is $\mathbb{E}\|x - \widetilde{x}\|^2 = \frac{N^2\delta^2}{12M}$; see (8.16).

8.2.3 Consistent Reconstruction

The error bounds for MSQ presented in the previous sections all make use of linear reconstruction methods. If an optimal encoding frame and optimal dual frame are used, then MSQ (under the uniform noise model) achieves the mean squared error

$$\mathbb{E}\|x - \widetilde{x}\|^2 = \frac{N^2\delta^2}{12M}. \tag{8.22}$$

In this section we briefly discuss the role of more general *nonlinear* reconstruction methods for MSQ with a focus on theoretical limitations and on concrete algorithmic approaches. Our main interest will be on how well reconstruction methods for MSQ are able to utilize frame redundancy as reflected by their dependence on M in bounds such as (8.22). In other words, how much information can be squeezed out of a set of MSQ quantized frame coefficients? This is of great interest for frame theory since it quantifies the extent to which MSQ is suitable for redundant frames, and it will motivate the need for alternatives such as $\Sigma\Delta$ quantization.

We begin by stating a main theoretical obstruction against significantly improving the reconstruction bound (8.22). There are various lower bounds in the literature which show that even with nonlinear reconstruction methods it is not possible for MSQ to achieve a mean squared error rate that is better than $1/M^2$. For example, the work in [30] assumes that the signal $x \in \mathbb{R}^N$ is a suitable nondegenerate random vector and that the frame $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ is selected from a suitable family of frames for $\mathbb{R}^N$. It is shown in [30] that if

$$R : (q_i)_{i=1}^M = \big(Q\big(\langle x, \varphi_i\rangle\big)\big)_{i=1}^M \longmapsto \widetilde{x} \in \mathbb{R}^N$$

is any (potentially nonlinear) reconstruction map for recovering x from the MSQ quantized coefficients $q_i = Q(\langle x, \varphi_i\rangle)$, then there exists a constant $C > 0$ such that

$$\mathbb{E}\|x - \widetilde{x}\|^2 = \mathbb{E}\big\|x - R\big(\big(Q\big(\langle x, \varphi_i\rangle\big)\big)_{i=1}^M\big)\big\|^2 > \frac{C}{M^2}. \tag{8.23}$$

This result does not use the uniform noise model and the expectation is taken over the random vector x. The constant C does not depend on the frame size M but may depend on the dimension N and the family of frames being considered.

One might expect that less restrictive lower bounds than (8.23) are possible if one uses the uniform noise model since the noise model is often more optimistic than deterministic reality. However, this is not the case, and there is work in [49] which proves a similar $1/M^2$ lower bound even under the uniform noise model.

There is a gap between the theoretical lower bounds of order $1/M^2$ for general reconstruction methods and the upper bounds of order $1/M$ obtainable with linear reconstruction. *Consistent reconstruction* is a technique that closes this gap. The basic idea behind consistent reconstruction is that if one observes a quantized frame coefficient $q_i = Q(\langle x, \varphi_i\rangle)$, then the true signal x must lie in the set

$$H_i = \big\{u \in \mathbb{R}^N : \big|\langle u, \varphi_i\rangle - q_i\big| \le \delta/2\big\}.$$

Consistent reconstruction simply selects any $\widetilde{x}$ in the intersection of the sets H_i, $1 \le i \le M$, by taking $\widetilde{x} \in \mathbb{R}^N$ as a solution to the system of linear inequality constraints:

$$\forall 1 \le i \le M, \quad \big|\langle \widetilde{x}, \varphi_i\rangle - q_i\big| \le \delta/2. \tag{8.24}$$

Consistent reconstruction can be efficiently implemented using linear programming methods. It has been shown that in appropriate settings consistent reconstruction achieves the mean squared error bound

$$\mathbb{E}\|x - \widetilde{x}\|^2 \le \frac{C}{M^2}. \tag{8.25}$$

As with the matching theoretical lower bounds, upper bounds of order $1/M^2$ in (8.25) have been proven in various settings under different sets of assumptions. Early results of this type appear in [51] in the context of sampling for bandlimited signals. The work in [30] proves deterministic upper bounds of order $1/M^2$ for

certain harmonic frames without using the uniform noise model, and the work in [19], cf. [20], obtains upper bounds of order $1/M^2$ with high probability for random frames but without the uniform noise model. The work in [48] proves (8.25) under the uniform noise model for certain classes of random frames, and quantifies the dimension dependence of the constant C using methods from stochastic geometry. The main point of these various error bounds is to highlight the ability of consistent reconstruction to outperform linear reconstruction. Moreover, since the $1/M^2$ bound for consistent reconstruction matches the order of the theoretical lower bound, consistent reconstruction is essentially considered to be an optimal recovery method for MSQ.

Consistent reconstruction globally enforces the full set of constraints in (8.24). Motivated by considerations of computational efficiency, there also exist iterative algorithms which proceed by locally enforcing consistency constraints. For example, given quantized frame coefficients $q_i = Q(\langle x, \varphi_i \rangle)$, $1 \leq i \leq M$, the *Rangan-Goyal algorithm* iteratively produces estimates $x_i \in \mathbb{R}^N$ of x by using

$$x_i = x_{i-1} + \frac{\varphi_i}{\|\varphi_i\|^2} S_{\delta/2}\big(q_i - \langle x_{i-1}, \varphi_i \rangle\big), \tag{8.26}$$

where the iteration is run for $i = 1, \ldots, M$, and $x_0 \in \mathbb{R}^N$ is an arbitrarily chosen initial estimate. Here, for fixed $t > 0$, $S_t(\cdot)$ denotes the soft thresholding function defined by

$$S_t(u) = \begin{cases} u - t, & \text{if } u > t, \\ 0, & \text{if } |u| \leq t, \\ u + t, & \text{if } u < -t. \end{cases} \tag{8.27}$$

Similar to consistent reconstruction, the Rangan-Goyal algorithm has been proven to achieve a mean squared error of order $1/M^2$, see [46, 49], for certain random or appropriately ordered deterministic frames. A key point is that the convergence of the Rangan-Goyal algorithm will depend strongly on the order in which it processes quantized frame coefficients.

We may summarize the results of this section for MSQ as follows. Reconstruction methods for MSQ that are based on consistent reconstruction are able to achieve a mean squared error of the optimal order $1/M^2$. In particular, consistent reconstruction and its variants outperform dual frame linear reconstruction which are only able to achieve a mean squared error of the order $1/M$.

8.3 First Order $\Sigma\Delta$ Quantization

$\Sigma\Delta$ quantization is an alternative approach to MSQ that is specifically designed to efficiently utilize redundancy in the quantization process. $\Sigma\Delta$ algorithms were first developed in the 1960s in the context of quantizing oversampled bandlimited signals [38], but the algorithms are quite generally applicable and have been shown to be particularly well adapted to the class of finite frames; see, e.g., [2]. $\Sigma\Delta$ quantization

uses the fact that if $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ is a frame with $M > N$, then there are correlations among the frame vectors $(\varphi_i)_{i=1}^M$ that can be used to compensate for errors during the quantization process. This section will focus on a specific first order $\Sigma\Delta$ quantizer. This will allow us to quickly highlight the mechanics and key features of $\Sigma\Delta$ algorithms without being slowed down by the technical issues that will later arise in higher order methods.

Given frame coefficients $y_i = \langle x, \varphi_i \rangle$, $1 \le i \le M$, the *first order* $\Sigma\Delta$ *quantizer* produces quantized coefficients $(q_i)_{i=1}^M$ by running the following iteration for $i = 1, \ldots, M$:

$$\begin{aligned} q_i &= Q(u_{i-1} + y_i), \\ u_i &= u_{i-1} + y_i - q_i. \end{aligned} \tag{8.28}$$

Here $(u_i)_{i=0}^M \subset \mathbb{R}$ is an internal sequence of state variables which, for convenience, we shall always initialize with $u_0 = 0$. The $\Sigma\Delta$ quantizer (8.28) has the following important stability property, e.g., [2, 21], that relates boundedness of the input sequence $y = (y_i)_{i=1}^M$ to boundedness of the state variables $u = (u_i)_{i=1}^M$:

$$\|y\|_\infty < L\delta \quad \Longrightarrow \quad \|u\|_\infty \le \delta/2. \tag{8.29}$$

Here, $\|\cdot\|_\infty$ denotes the usual ℓ^∞ norm of a finite or infinite sequence. Stability plays an important role in the error analysis of $\Sigma\Delta$ quantizers, but it also ensures that $\Sigma\Delta$ quantizers can be implemented in circuitry with operating parameters that remain in a practical range.

Linear reconstruction is the simplest method for recovering a signal $\widetilde{x} \in \mathbb{R}^N$ from a set of $\Sigma\Delta$ quantized frame coefficients. Suppose that $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ is a frame and that $(\psi_i)_{i=1}^M \subset \mathbb{R}^N$ is any associated dual frame. Suppose that $x \in \mathbb{R}^N$ and that the frame coefficients $y_i = \langle x, \varphi_i \rangle$ are used as input to the $\Sigma\Delta$ quantizer and that $(q_i)_{i=1}^M$ is the resulting quantized output. We may then reconstruct $\widetilde{x}$ as

$$\widetilde{x} = \sum_{i=1}^M q_i \psi_i. \tag{8.30}$$

We then have the following $\Sigma\Delta$ error formula [2].

Proposition 8.2 *Suppose that first order* $\Sigma\Delta$ *quantization is used to quantize frame coefficients of the frame* $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ *and that the dual frame* $(\psi_i)_{i=1}^M \subset \mathbb{R}^N$ *is used for linear reconstruction in* (8.30). *The* $\Sigma\Delta$ *quantization error satisfies*

$$x - \widetilde{x} = \sum_{i=1}^{M-1} u_i(\psi_i - \psi_{i+1}) + u_M \psi_M. \tag{8.31}$$

Proof The proof follows from an application of summation by parts:

$$\begin{aligned} x - \widetilde{x} &= \sum_{i=1}^{M} \langle x, \varphi_i \rangle \psi_i - \sum_{i=1}^{M} q_i \psi_i \\ &= \sum_{i=1}^{M} (y_i - q_i) \psi_i \\ &= \sum_{i=1}^{M} (u_i - u_{i-1}) \psi_i \\ &= \sum_{i=1}^{M-1} u_i (\psi_i - \psi_{i+1}) + u_M \psi_M - u_0 \psi_1. \end{aligned}$$

□

The $\Sigma\Delta$ quantization error $\|x - \widetilde{x}\|$ depends strongly on the order in which the frame coefficients $(\langle x, \varphi_i \rangle)_{i=1}^{M}$ are entered into the $\Sigma\Delta$ algorithm. In (8.31) this dependence appears in the state variable sequence $(u_i)_{i=1}^{M}$ (this sequence changes if the order of the input sequence changes) and also appears in the ordering of the dual frame $(\psi_i)_{i=1}^{M}$ associated to $(\varphi_i)_{i=1}^{M}$ via the terms $(\psi_i - \psi_{i+1})$. To help quantify the dependence on the ordering of the dual frame sequence we will make use of the *frame variation* $\sigma((\psi_i)_{i=1}^{M})$ defined by

$$\sigma\big((\psi_i)_{i=1}^{M}\big) = \sum_{i=1}^{M-1} \|\psi_i - \psi_{i+1}\|. \tag{8.32}$$

The frame variation can be used to give the following $\Sigma\Delta$ error bound [2].

Theorem 8.2 *Suppose that the frame* $(\varphi_i)_{i=1}^{M} \subset \mathbb{R}^N$ *satisfies* $\sup_{1 \le i \le M} \|\varphi_i\| \le C$ *and that* $x \in \mathbb{R}^N$ *satisfies* $\|x\| < \delta L C^{-1}$. *Then the* $\Sigma\Delta$ *error satisfies*

$$\|x - \widetilde{x}\| \le \frac{\delta}{2}\big(\sigma\big((\psi_i)_{i=1}^{M}\big) + \|\psi_M\|\big).$$

Proof The result will follow from Proposition 8.2. Since $\|x\| < \delta L/M$ we have that $|y_i| \le |\langle x, \varphi_i \rangle| \le \|x\| \|\varphi_i\| < \delta L C^{-1} C = L\delta$. Using the stability bound (8.29) and that $u_0 = 0$, it follows from (8.31) that

$$\|x - \widetilde{x}\| \le \frac{\delta}{2}\left(\sum_{i=1}^{M-1} \|\psi_i - \psi_{i+1}\| + \|\psi_M\|\right) = \frac{\delta}{2}\big(\sigma\big((\psi_i)_{i=1}^{M}\big) + \|\psi_M\|\big).$$

□

The following corollary addresses the important special case when $(\varphi_i)_{i=1}^{M} \subset \mathbb{R}^N$ is a unit norm tight frame and $\psi_i = \frac{N}{M}\varphi_i$ is the canonical dual frame.

Corollary 8.1 *If $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ is a unit norm tight frame and $(\psi_i)_{i=1}^M$ is the associated canonical dual frame, then for every $x \in \mathbb{R}^N$ with $\|x\| < \delta L$ the $\Sigma\Delta$ quantization error satisfies*

$$\|x - \widetilde{x}\| \leq \frac{\delta N(\sigma((\varphi_i)_{i=1}^M) + 1)}{2M}.$$

A practical consequence of Corollary 8.1 is that for a wide variety of finite frames the $\Sigma\Delta$ quantization error $\|x - \widetilde{x}\|$ is of order $1/M$. The following example illustrates this phenomenon for a particular family of unit norm tight frames in $\mathbb{R}^2$.

Example 8.1 Let $(\varphi_i^M)_{i=1}^M \subset \mathbb{R}^2$ be the unit norm tight frame for $\mathbb{R}^2$ given by the Mth roots of unity in the natural ordering

$$1 \leq j \leq M, \quad \varphi_j^M = \big(\cos(2\pi j/M), \sin(2\pi j/M)\big). \tag{8.33}$$

It can be shown that the frame variation satisfies the following upper bound that is independent of the frame size M:

$$\sigma\big((\varphi_j)_{j=1}^M\big) \leq 2\pi. \tag{8.34}$$

Thus, Corollary 8.1 yields the following $\Sigma\Delta$ error bound:

$$\|x - \widetilde{x}\| \leq \frac{\delta(2\pi + 1)}{M}. \tag{8.35}$$

The error bound (8.35) of order $1/M$ is in no way specific to the roots-of-unity frame; it simply requires a class of finite frames for which the frame variation can be bounded independently of the frame size M. See [2, 9] for similar results with more general classes of frames in $\mathbb{R}^N$, such as harmonic frames and frames generated by frame paths. We shall consider this issue more deeply in the next section on higher order $\Sigma\Delta$ quantization.

The constant $(2\pi + 1)$ in (8.35) arose from having used the frame variation to derive a $\Sigma\Delta$ error bound. Upper bounds obtained using the frame variation are convenient but are generally not optimal. The work in [9] improves the constants in first order $\Sigma\Delta$ error bounds by working with a suitably generalized variant of the frame variation. There are also refined error bounds in [2] which show that the $1/M$ error rate can sometimes be improved. For example, for the roots-of-unity frame and certain frames generated by a frame path, there are circumstances when the $\Sigma\Delta$ error satisfies a refined bound of order $M^{-5/4}\log M$; see [2]. The refined $\Sigma\Delta$ error bounds for finite frames in [2] are motivated by the refined bounds for sampling expansions in [33], but there are technical differences in the order of estimates that are obtained in these two settings. The work in [1] carefully compares the pointwise performance of $\Sigma\Delta$ quantization with MSQ, and the work in [52] makes interesting connections between $\Sigma\Delta$ error analysis and the traveling salesman problem.

8.4 Higher Order Sigma-Delta Quantization

The first order $\Sigma\Delta$ quantizer (8.28) sits at the heart of a rich class of algorithms. We have already seen that first order $\Sigma\Delta$ quantization can achieve accuracy $\|x - \widetilde{x}\| \leq C/M$ using just the simple single loop feedback mechanism of (8.28). Moreover, first order $\Sigma\Delta$ error bounds such as (8.35) are deterministic (requiring no noise models) and thus outperform MSQ even if optimal MSQ reconstruction methods are used. The first order $\Sigma\Delta$ quantizer (8.28) is the tip of the algorithmic iceberg. The algorithm (8.28) can be broadly generalized to provide quantization that is dramatically superior to both first order $\Sigma\Delta$ quantization and MSQ, and in some cases it performs near optimally; e.g., see the high precision methods in Sect. 8.5.

There are several directions along which one can generalize first order $\Sigma\Delta$ quantization. For example, it is common among engineering practitioners to study general $\Sigma\Delta$ quantizers in the context of spectral noise shaping or in the framework of error diffusion algorithms; see, e.g., [12]. In this section, we follow a purely structural approach to generalization which builds on the fact that (8.28) expresses the coefficient quantization errors $y_i - q_i$ as a *first order* difference $(\Delta u)_i = u_i - u_{i-1}$ of state variables u_i with a uniform stability bound given by (8.29). Specifically, rth order $\Sigma\Delta$ quantization will generalize the relation

$$(\Delta u)_i = y_i - q_i$$

by using *higher order* difference operators Δ^r.

With this in mind, let us define the class of higher order difference operators which will be needed. Let $(u_i)_{i=1}^M \subset \mathbb{R}$ be a given sequence which we extend to nonpositive indices using the convention $u_i = 0$ for $i \leq 0$. The standard first order backward difference operator $\Delta = \Delta^1$ acts on the sequence $(u_i)_{i=1}^M$ by $(\Delta u)_i = u_i - u_{i-1}$ for all $1 \leq i \leq M$. For each positive integer r we may recursively define the rth order backward difference operator Δ^r by $(\Delta^r u)_i = (\Delta \circ \Delta^{r-1} u)_i$ or by the following equivalent closed-form expression for $i = 1, \ldots, M$:

$$\left(\Delta^r u\right)_i = \sum_{j=0}^{r} (-1)^j \binom{r}{j} u_{i-j}. \tag{8.36}$$

8.4.1 Higher Order $\Sigma\Delta$ Quantization for Finite Frames

An rth order $\Sigma\Delta$ quantizer takes a sequence $(y_i)_{i=1}^M \subset \mathbb{R}$ as its input and produces the quantized output sequence $(q_i)_{i=1}^M$ by iteratively satisfying the following equations for $i = 1, \ldots, M$:

$$\begin{aligned} q_i &= Q\big(R(u_{i-1}, \ldots, u_{i-T}, y_i, \ldots, y_{i-S})\big), \\ \left(\Delta^r u\right)_i &= y_i - q_i. \end{aligned} \tag{8.37}$$

Here S, T are fixed positive integers and $R : \mathbb{R}^{T+S+1} \to \mathbb{R}$ is a fixed function known as the *quantization rule*. As with the first order $\Sigma\Delta$ quantizer, $(u_i)_{i=1-T}^{M} \subset \mathbb{R}$ is a sequence of state variables. For simplicity, we always assume that the state variable sequence is initialized by $u_0 = u_{-1} = \cdots = u_{1-T} = 0$ and, if needed, define $y_i = 0$ for $i \leq 0$. As in previous sections, Q denotes the scalar quantizer associated to the $(2L+1)$ level midtread quantization alphabet $\mathscr{A}_L^\delta$ with stepsize $\delta > 0$.

There is a great deal of flexibility in the choice of the quantization rule R; see [54] for some typical choices. The most important (and difficult) factor in selecting R is that the associated $\Sigma\Delta$ algorithm should be *stable* in the sense that there exist constants $C_1, C_2 > 0$, independent of M, such that the input sequence $y = (y_i)_{i=1}^{M}$ and state variable sequence $u = (u_i)_{i=1}^{M}$ satisfy

$$\|y\|_\infty \leq C_1 \quad \Longrightarrow \quad \|u\|_\infty \leq C_2.$$

In contrast to the bound (8.29) for the first order algorithm (8.28), stability can be a technically challenging issue for higher order $\Sigma\Delta$ quantizers, especially in the case of 1-bit quantizers. In fact, it was only recently proven in [21] that stable 1-bit rth order $\Sigma\Delta$ quantizers actually exist for each positive integer r. Proving that a particular higher order $\Sigma\Delta$ quantizer is stable often leads to delicate issues from the theory of dynamical systems. For example, $\Sigma\Delta$ quantization has close connections to the ergodic dynamics of piecewise affine dynamical systems [37, 54], and to geometric tiling properties of invariant sets [37].

For concreteness and to avoid technical issues involving $\Sigma\Delta$-stability, we shall restrict our discussion to the following particular rth order $\Sigma\Delta$ quantizer, known as the *greedy* $\Sigma\Delta$ quantizer:

$$\begin{aligned} q_i &= Q\left(\sum_{j=1}^{r}(-1)^{j-1}\binom{r}{j}u_{i-j} + y_i\right), \\ u_i &= \sum_{j=1}^{r}(-1)^{j-1}\binom{r}{j}u_{i-j} + y_i - q_i. \end{aligned} \tag{8.38}$$

It is easy to check that with this rule, e.g., [34], if the input sequence $y = (y_i)_{i=1}^{M}$ satisfies $\|y\|_\infty < \delta(L - 2^{r-1} - 3/2)$, then one has the stability bounds

$$|u_i| \leq 2^{-1}\delta \quad \text{and} \quad |y_i - q_i| \leq 2^{r-1}\delta. \tag{8.39}$$

Note that each iteration of the rth order $\Sigma\Delta$ quantizer (8.38) requires more computation and more memory (access to several state variables u_{i-j}) than the standard first order $\Sigma\Delta$ quantizer in (8.28). As a trade-off for this increased computational burden we shall later see that higher order $\Sigma\Delta$ quantization produces increasingly accurate signal representations.

Let $(\varphi_i)_{i=1}^{M} \subset \mathbb{R}^N$ be a frame for $\mathbb{R}^N$ and let $(\psi_i)_{i=1}^{M}$ be any associated dual frame with synthesis operator Ψ. Suppose that $x \in \mathbb{R}^N$, that the frame coefficients $y_i = \langle x, \varphi_i \rangle$ are used as input to the rth order $\Sigma\Delta$ quantizer (8.38), and that $(q_i)_{i=1}^{M}$

are the resulting $\Sigma\Delta$ quantized frame coefficients. Let q denote the $M \times 1$ column vector with $(q_i)_{i=1}^M$ as its entries. The simplest approach to recovering a signal $\widetilde{x} \in \mathbb{R}^N$ from the $\Sigma\Delta$ quantized coefficients $q = (q_i)_{i=1}^M$ is to linearly reconstruct with the dual frame $(\psi_i)_{i=1}^M$ by using

$$\widetilde{x} = \Psi q = \sum_{i=1}^{M} q_i \psi_i. \tag{8.40}$$

Our discussion of higher order $\Sigma\Delta$ quantization will only address reconstruction using the linear approach (8.40), but it is important to point out that nonlinear alternatives such as consistent reconstruction can also be very effective, e.g., [50], at the cost of increased complexity.

For the remainder of this section $\widetilde{x}$ will denote the linear reconstruction (8.40). The $\Sigma\Delta$ error $(x - \widetilde{x})$ can be compactly represented in matrix form using the $M \times M$ matrix D defined by

$$D_{ij} := \begin{cases} 1, & \text{if } i = j, \\ -1, & \text{if } i = j+1, \\ 0, & \text{otherwise.} \end{cases} \tag{8.41}$$

Letting u denote the $M \times 1$ column vector of state variables $u = (u_i)_{i=1}^M$, we have the following $\Sigma\Delta$ error formula; see [5, 34, 42].

Lemma 8.1 *The rth order $\Sigma\Delta$ quantization error $(x - \widetilde{x})$ satisfies*

$$x - \widetilde{x} = \sum_{i=1}^{M} (y_i - q_i)\psi_i = \Psi D^r u. \tag{8.42}$$

If $x \in \mathbb{R}^N$ and the frame coefficients $y_i = \langle x, \varphi_i \rangle$ satisfy $|y_i| < \delta(L - 2^{r-1} - 3/2)$, then the stability bound (8.39) for the $\Sigma\Delta$ quantizer (8.38) gives that

$$\|u\| \leq \sqrt{M}\|u\|_\infty \leq 2^{-1}\delta\sqrt{M}. \tag{8.43}$$

A typical way to ensure $|y_i| = |\langle x, \varphi_i \rangle| < \delta(L - 2^{r-1} - 3/2)$ is to assume that $x \in \mathbb{R}^N$ satisfies $\|x\| < \delta(L - 2^{r-1} - 3/2)C^{-1}$, where $C = \sup_{1 \leq i \leq M} \|\varphi_i\|$. The stability bound (8.43) together with Lemma 8.1 gives the following upper bound on the $\Sigma\Delta$ error.

For the remainder of the chapter, if $T : \mathbb{R}^{d_1} \to \mathbb{R}^{d_2}$ is a linear operator, then $\|T\|_{\text{op}} = \|T\|_{\ell_2 \to \ell_2}$ denotes the operator norm of T when $\mathbb{R}^{d_1}$ and $\mathbb{R}^{d_2}$ are both endowed with the standard Euclidean ℓ_2 norm.

Corollary 8.2 *If $x \in \mathbb{R}^N$ and the frame coefficients $y_i = \langle x, \varphi_i \rangle$ satisfy*

$$\|y\|_\infty < \delta\big(L - 2^{r-1} - 3/2\big)$$

then the rth order $\Sigma\Delta$ quantization error satisfies

$$\|x - \widetilde{x}\| = \left\|\Psi D^r u\right\| \leq \|u\| \left\|\Psi D^r\right\|_{\text{op}} \leq 2^{-1}\delta\sqrt{M}\left\|\Psi D^r\right\|_{\text{op}}. \tag{8.44}$$

8.4.2 Sobolev Dual Frames

Our goal in this section is to obtain quantitative estimates on how small the rth order $\Sigma\Delta$ error $\|x - \widetilde{x}\|$ is for certain specific families of finite frames. To do this we will need a clearer understanding of the error bound (8.44) in Corollary 8.2. Similar to bounds such as (8.16) and (8.35), we are especially interested in quantifying how small the $\Sigma\Delta$ error is as a function of the frame size M.

It will be helpful to give some perspective on the type of error bounds that one might hope for. The groundbreaking work in [21] studied rth order $\Sigma\Delta$ quantization in the setting of *bandlimited sampling expansions* and showed error bounds of the form

$$\|h - \widetilde{h}\|_{L^\infty(\mathbb{R})} \lesssim \frac{1}{\lambda^r}, \tag{8.45}$$

where $\widetilde{h}$ is obtained from the bandlimited function h by $\Sigma\Delta$ quantization, and λ denotes the oversampling rate. The full details on bandlimited sampling are not essential here, but (8.45) illustrates the point that higher order $\Sigma\Delta$ algorithms can make increasingly effective use of redundancy (oversampling) as the order r of the algorithm increases. We wish to show that similar results hold in the setting of finite frames. For first order $\Sigma\Delta$ quantization of certain finite frames we have already seen such a result in (8.35).

Toward obtaining quantitative $\Sigma\Delta$ error bounds, Corollary 8.2 shows that one can decouple the roles of the state variable sequence u and the dual frame Ψ. Since stability bounds give direct control over the state variable sequence, the main issue for bounding $\|x - \widetilde{x}\|$ will be to clarify the role of the dual frame Ψ and to understand the size of the operator norm $\|\Psi D^r\|_{\text{op}}$; see (8.44). For a redundant frame Φ, the choice of dual frame Ψ is highly nonunique, so a closely related issue is to address which particular dual frames are most suitable for reconstructing signals from $\Sigma\Delta$ quantized coefficients.

Given a frame $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ with analysis operator Φ^*, we wish to determine which dual frames $(\psi_i)_{i=1}^M$ work well when the linear reconstruction (8.40) is used to reconstruct signals from their $\Sigma\Delta$ quantized frame coefficients. We seek a choice of dual frame that does not depend on the specific signal $x \in \mathbb{R}^N$ being quantized. The widespread use of canonical dual frames makes it reasonable to give some initial consideration to canonical dual frame reconstruction in higher order $\Sigma\Delta$ quantization. We have already seen in Sect. 8.2 that the canonical dual frame is optimally suited for MSQ and works well for first order $\Sigma\Delta$ problems such as Example 8.1; see also [2, 3]. Unfortunately, the canonical dual frame can perform quite poorly for higher order $\Sigma\Delta$ problems. An example of this phenomenon is shown in [42]: For

rth order $\Sigma\Delta$ quantization of the roots-of-unity frame (8.33), if $r \geq 3$ then canonical dual frame reconstruction cannot robustly achieve quantization error $\|x - \widetilde{x}\|$ of order better than $1/M^2$. This means that proper choices of dual frames are very important for higher order $\Sigma\Delta$ quantization of finite frames. For comparison, this issue does not arise in the infinite dimensional setting of $\Sigma\Delta$ quantization of bandlimited sampling expansions in [21].

The following result addresses how to choose dual frames which minimize the quantity $\|\Psi D^r\|_{\text{op}}$. In view of Corollary 8.2, these dual frames will be natural candidates for $\Sigma\Delta$ signal reconstruction.

Proposition 8.3 *Let Φ be a given $N \times M$ matrix with full rank and let D be the $M \times M$ matrix defined by* (8.41). *Consider the following minimization problem taken over all $N \times M$ matrices Ψ:*

$$\min\left\{\left\|\Psi D^r\right\|_{\text{op}} : \Psi\Phi^* = I\right\}. \tag{8.46}$$

The minimizer $\Psi = \Psi_{\text{r,Sob}}$ of (8.46) *is given by*

$$\Psi_{\text{r,Sob}} = \left(D^{-r}\Phi^*\right)^{\dagger} D^{-r} = \left(\Phi\left(D^*\right)^{-r} D^{-r}\Phi^*\right)^{-1}\Phi\left(D^*\right)^{-r} D^{-r}. \tag{8.47}$$

We refer to $\Psi_{\text{r,Sob}}$ in (8.47) *as the* rth order Sobolev dual *associated to Φ. Using the notation of frames, if $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ is a frame with analysis operator Φ^*, then the dual frame $(\psi_i)_{i=1}^M$ with synthesis operator $\Psi_{\text{r,Sob}}$ is referred to as the* rth order Sobolev dual frame.

It is worth mentioning that D and D^* do not commute. Readers of [5] should consult the errata [6] to avoid an unfortunate notational error in the definition of Sobolev dual in [5] caused by this noncommutativity.

Up to this point, we have shown that Sobolev duals minimize the $\Sigma\Delta$ error term $\|\Psi D^r\|_{\text{op}}$, but it remains to give precise quantitative bounds on this expression. For this, it will be convenient to consider the class of frames that are generated by frame paths.

Definition 8.1 A vector-valued function $\Phi : [0, 1] \to \mathbb{R}^N$ given by

$$\Phi(t) = \left(\varphi_1(t), \varphi_2(t), \ldots, \varphi_N(t)\right)$$

is a *piecewise-C^1 uniformly sampled frame path* if the following three conditions hold:

(a) $\forall 1 \leq i \leq M$, the map $\varphi_i : [0, 1] \to \mathbb{R}$ is piecewise-C^1.
(b) The functions $(\varphi_i)_{i=1}^N$ are linearly independent.
(c) $\exists M_0$ such that $\forall M \geq M_0$ the collection $(\Phi(i/M))_{i=1}^M$ is a frame for $\mathbb{R}^N$.

Many standard finite frames arise from frame path constructions; for example, see [5]. The simplest example of a frame path is given by the function

$$\Phi(t) = \left(\cos(2\pi t), \sin(2\pi t)\right).$$

This frame path recovers the family of unit norm tight frames in (8.33) by

$$\varphi_k^M = \Phi(k/M) = \big(\cos(2\pi k/M), \sin(2\pi k/M)\big),$$

so that for each $M \geq 3$ the set $(E(k/M))_{k=1}^M$ is a unit norm tight frame for $\mathbb{R}^2$.

We require the following slightly lengthy setup for the next theorem. Let $\Phi : [0, 1] \to \mathbb{R}^N$ be a piecewise-C^1 uniformly sampled frame path, and for each $M \geq M_0$, let $(\psi_i^M)_{i=1}^M$ be the rth order Sobolev dual frame associated to the frame $(\Phi(i/M))_{i=1}^M \subset \mathbb{R}^N$. If $x \in \mathbb{R}^N$ then, for each $M \geq M_0$, the signal x has the frame coefficients $y_i^M = \langle x, \Phi(i/M)\rangle$, $1 \leq i \leq M$. Assume that the frame coefficients all satisfy $|y_i^M| \leq \delta(K - 2^{r-1} - 3/2)$. For each $M \geq M_0$, rth order $\Sigma\Delta$ quantization is applied to the frame coefficients $(y_i^M)_{i=1}^M$ to obtain the quantized coefficients $(q_i^M)_{i=1}^M$. Finally, the Sobolev dual frame $(\psi_i^M)_{i=1}^M$ is used to linearly reconstruct a signal $\widetilde{x}_M$ from $(q_i^M)_{i=1}^M$.

Theorem 8.3 *Consider rth order $\Sigma\Delta$ quantization of a C^1 uniformly sampled frame path and assume the setup of the preceding paragraph. Then there exists a constant $C_{r,\Phi}$, depending only on r and the frame path Φ, such that the $\Sigma\Delta$ quantization error using rth order Sobolev dual frame reconstruction satisfies*

$$\forall M \geq M_0, \quad \|x - \widetilde{x}_M\| \leq \frac{C_{r,\Phi}}{M^r}. \tag{8.48}$$

Theorem 8.3, for example, applies to the root-of-unity frame for $\mathbb{R}^2$ in (8.33), harmonic frames in $\mathbb{R}^N$, and tight frames obtained by repeating an orthonormal basis, see [5], and in each case ensures that rth order $\Sigma\Delta$ quantization using Sobolev duals achieves accuracy $\|x - \widetilde{x}\| \leq c/M^r$. It is important to emphasize again that this error performance is generally not possible if the canonical dual frame is used instead of a Sobolev dual; see [42].

Example 8.2 Let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^2$ be the roots-of-unity unit norm tight frame for $\mathbb{R}^2$ with $M = 256$ given by (8.33). Figure 8.4(a) shows the frame vectors $(\varphi_i)_{i=1}^M$, and Fig. 8.4(b) shows the associated canonical dual frame vectors given by $\widetilde{\varphi}_i = (\frac{2}{256})\varphi_i$ with $1 \leq n \leq 256$. Figure 8.4(c) shows the associated Sobolev dual frame of order $r = 2$. Note that each of these figures has been scaled differently to optimize visibility.

Example 8.3 30 points in $\mathbb{R}^2$ are randomly chosen according to the uniform distribution on the unit square. For each of the 30 points, the corresponding frame coefficients with respect to the roots-of-unity frame (8.33) are quantized using a particular third order $\Sigma\Delta$ scheme from [21]. Linear reconstruction is then performed with each of the 30 sets of quantized coefficients using both the canonical dual frame and the third order Sobolev dual. Candual(M) and Altdual(M) will denote the largest of the 30 errors obtained using the canonical dual frame and Sobolev dual, respectively. Figure 8.3 shows a log-log plot of Altdual(M) and Candual(M) against the

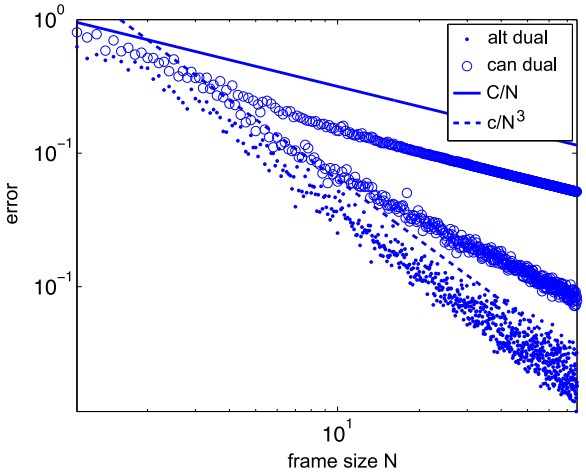


Fig. 8.3 Log-log plot of third order $\Sigma\Delta$ quantization errors against frame size M, for the Mth roots-of-unity family of frames for $\mathbb{R}^2$. The figure compares the use of the canonical dual frame and the third order Sobolev dual frame for reconstruction, and illustrates the superior accuracy provided by Sobolev duals

frame size M. For comparison, log-log plots of $1/M^3$ and $1/M$ are also given. Note that Sobolev duals yield a smaller reconstruction error than canonical dual frames. Further details on this example can be found in [5].

The Sobolev dual frames illustrate the importance of noncanonical representations in quantization problems. More generally, the use of alternative dual frames is a valuable technique in several other problems on mathematical signal processing. For example, [23, 41] use noncanonical Gabor frames to provide improved time-frequency localization. The work in [16–18] uses noncanonical representations to provide desirable support, smoothness, and structural properties in the settings of Gabor and shift invariant systems. See [27, 43, 44] for work on noise reduction and other properties of noncanonical representations.

In practice, one may not always have full control over the encoding frame $(\varphi_i)_{i=1}^M$ that is used to compute frame coefficients $y_i = \langle x, \varphi_i \rangle$. For example, this might be the case if the frame Φ corresponds to a physical measurement device and the coefficients $(y_i)_{i=1}^M$ are observed measurements. A valuable feature of the Sobolev dual frame method is that it places relatively few constraints on the encoding frame Φ, and it is entirely self-contained to the signal reconstruction step after encoding and quantization have already taken place. This modularity makes Sobolev duals a flexible tool. A different approach which has also proven fruitful is to custom build special frames for $\Sigma\Delta$ quantization which are specifically designed to work well with canonical linear reconstruction; see [10, 40]. By necessity, this approach places very strong restrictions on the encoding frame (for example, it excludes any unit norm frames), but one gains a simplified reconstruction step involving tight frame expansions. Nonetheless, similar to Sobolev duals, a key issue in these constructions is to design frames that terminate smoothly at the origin.

8.4.3 Sobolev Duals of Random Frames

We have seen in the previous section that higher order $\Sigma\Delta$ algorithms are able to make judicious use of the correlations among frame vectors to provide accurate quantization. The frame path structure used in Theorem 8.3 guarantees sufficient correlations among frame vectors since they lie along a piecewise smooth path and hence vary slowly, ensuring that nearby frame vectors are highly correlated. In view of this, it is perhaps surprising that $\Sigma\Delta$ quantization also performs well even if highly *unstructured random frames* are used.

Let Φ be an $N \times M$ random matrix with i.i.d. standard normal $\mathcal{N}(0, 1)$ entries, and let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ be the collection of random vectors with synthesis operator Φ. Since Φ has full rank with probability one, we shall refer to $(\varphi_i)_{i=1}^M$ as a *Gaussian random frame* for $\mathbb{R}^N$. The following theorem addresses the performance of $\Sigma\Delta$ quantization with Sobolev dual frames when a Gaussian random frame is used [34–36].

Theorem 8.4 *Let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^N$ be a Gaussian random frame and let $(\psi_i)_{i=1}^M$ be the associated Sobolev dual frame with synthesis operator $\Psi = \Psi_{\mathrm{r,Sob}}$. Let $\lambda = M/N$.*

For any $\alpha \in (0, 1)$, if $\lambda \geq c(\log M)^{1/(1-\alpha)}$, then with probability at least

$$1 - \exp\bigl(-c' M \lambda^{-\alpha}\bigr),$$

the following holds:

$$\bigl\| \Psi D^r \bigr\|_{op} \lesssim_r \lambda^{-\alpha(r-\frac{1}{2})} M^{-1/2}. \tag{8.49}$$

Consequently, the following error bound holds for rth order $\Sigma\Delta$ quantization of Gaussian random frames:

$$\|x - \widetilde{x}_M\|_2 \lesssim_r \lambda^{-\alpha(r-\frac{1}{2})} \delta. \tag{8.50}$$

Example 8.4 Let $(\varphi_i)_{i=1}^M \subset \mathbb{R}^2$ be a Gaussian random frame of size $M = 256$. Figure 8.4(d) shows the frame vectors $(\varphi_i)_{i=1}^M$ and Fig. 8.4(e) shows the associated canonical dual frame vectors. Note that the Gaussian random frame is approximately tight; e.g., see [30]. Figure 8.4(f) shows the associated Sobolev dual frame of order $r = 4$. Note that each of these figures has been scaled differently to optimize visibility.

Theorem 8.4 has important implications for compressed sensing in [34–36]. In contrast to frame theory, compressed sensing involves a nonlinear signal space (the collection of s-sparse signals in $\mathbb{R}^N$) and the high dimensional nature of compressed sensing often places a premium cost on oversampling. Nonetheless, frame theory implicitly plays an important role in many compressed sensing problems. When combined with appropriate support recovery methods, Theorem 8.4 directly implies

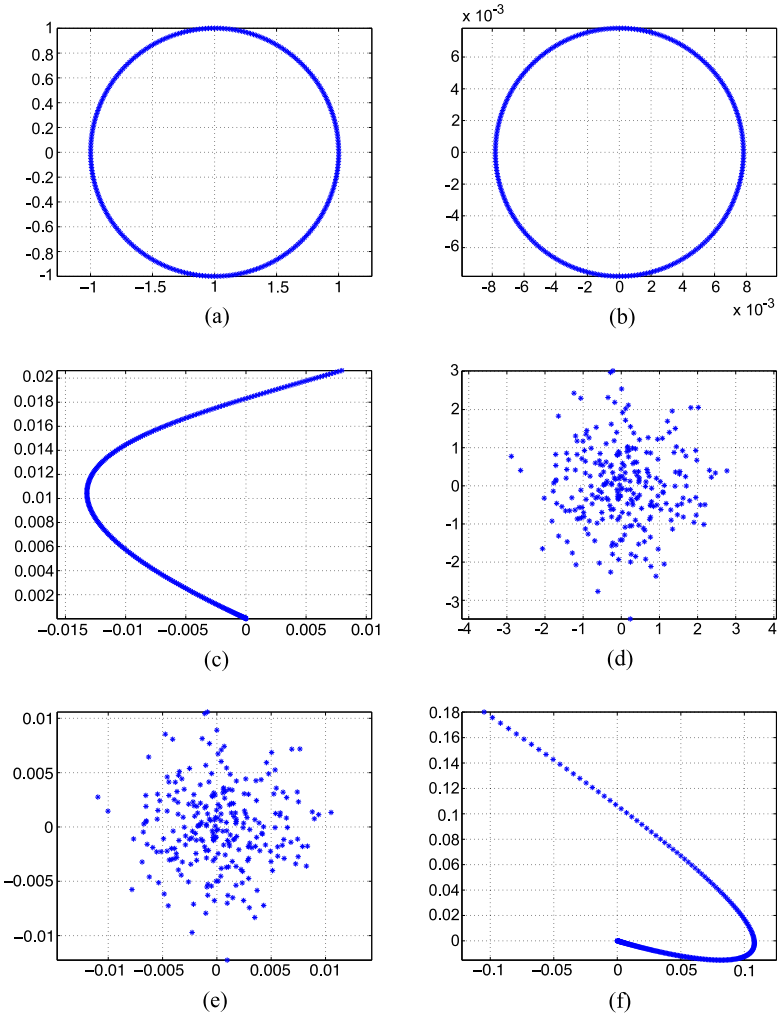


Fig. 8.4 Figure (**a**) shows the roots-of-unity frame with $M = 256$, Fig. (**b**) shows the associated canonical dual frame, and Fig. (**c**) shows the associated Sobolev dual frame of order $r = 2$. Figure (**d**) shows a Gaussian random frame of size $M = 256$, Fig. (**e**) shows the associated canonical dual frame, and Fig. (**f**) shows the associated Sobolev dual frame of order $r = 4$. Note that the axes are scaled differently in the various figures for visibility

that $\Sigma\Delta$ quantization is an effective strategy for quantizing compressed sensing measurements, and that Sobolev duals are a useful tool for extracting information from quantized data in high dimensions; see [34–36]. See [47] for the use of $\Sigma\Delta$ algorithms with other random sampling geometries in the context of randomly interleaved sampling of bandlimited signals.

8.5 Root-Exponential Accuracy

The preceding discussion on frame quantization has assumed a particular paradigm: Given an appropriate frame Φ with *oversampling rate* $\lambda := M/N$, one fixes an rth order $\Sigma\Delta$ quantization scheme $\mathscr{Q}_r$ to quantize the frame expansion, i.e., $q := \mathscr{Q}_r(\Phi^* x)$. Subsequently, one approximates x with $\widetilde{x} = \Psi_r q$, where Ψ_r is the rth order Sobolev dual of Φ. Under this paradigm, where the order r is fixed, we have seen that, for example, if Φ is a Gaussian random frame, the approximation error behaves like an inverse polynomial (in λ); specifically, we have $\|x - \widetilde{x}\|_2 \lesssim C(r)\lambda^{-r}$.

Next, we shall diverge from the above paradigm and treat the order r of the $\Sigma\Delta$ quantization scheme as a parameter. Using this approach, we shall show that "root-exponential" error rates can be obtained (when decoding is done via linear reconstruction with Sobolev duals). Specifically, if we optimize the order r as a function of λ, we show that the reconstruction error satisfies $\|x - \widetilde{x}\|_2 \leq Ce^{-c\sqrt{\lambda}}$, provided that the $\Sigma\Delta$ schemes and the encoding frame Φ are chosen appropriately [40].

8.5.1 Superpolynomial Accuracy and $\Sigma\Delta$ Quantization: Bandlimited Setting

The $\Sigma\Delta$ schemes we use to achieve root-exponential accuracy in the finite frame setting were originally devised for quantization of oversampled bandlimited functions. In fact, the superpolynomial error decay (as a function of the oversampling rate[1] λ) of the approximation error in $\Sigma\Delta$ quantization was first shown in the context of bandlimited functions (in L^∞) [21]. To achieve superpolynomial decay, [21] constructs a family of stable $\Sigma\Delta$ schemes of arbitrary order, with a nonlinear quantization rule involving concatenations of "sign" functions. Next, the order r of the actual quantization scheme is determined as a function of the oversampling rate λ. This way, [21] shows that the approximation error is $O(\lambda^{-c\log\lambda})$. In the same bandlimited setting, it was later shown in [32] that exponential error decay rates can be obtained. In particular, the rth order stable $\Sigma\Delta$ quantizer proposed in [32], which we shall briefly describe later in this section, uses a linear quantization rule and an auxiliary state sequence v that is updated based on r of its (non-immediate) previous values. Exponential accuracy is achieved by, again, optimally choosing r as a function of λ. Recently, [24] obtained improved exponential rates using $\Sigma\Delta$ schemes that are constructed within the framework of [32] with better stability properties.

[1] In this setting, the oversampling rate is defined as the ratio of the sampling rate to the Nyquist rate.

8.5.2 Superpolynomial Accuracy and $\Sigma\Delta$ Quantization: Finite Frame Setting

The above-described approach can be adapted to the finite frame setting. In particular, when appropriate finite frame families are considered and when appropriate (Sobolev) duals are used in the reconstruction, one can show that the approximation error decays like a "root exponential" in the oversampling rate λ [40]. The remainder of this section is dedicated to describing how this can be done.

As noted in Sect. 8.4, one can control the reconstruction error involved in rth order $\Sigma\Delta$ quantization via the product bound $\|x - \widetilde{x}\| = \|\Psi D^r u\| \leq \|\Psi D^r\|_{\text{op}} \|u\| \sqrt{M}$, where Ψ is the specific dual of Φ that is used to reconstruct $\widetilde{x}$ from the quantized coefficients. The use of stable $\Sigma\Delta$ schemes guarantees that $\|u\|$ is bounded, and Sobolev duals minimize $\|\Psi D^r\|_{\text{op}}$ over all duals of Φ. In Sect. 8.4, we saw that when the frame Φ is chosen appropriately, this technique leads to polynomial error decay in λ. Motivated by the above discussion on the bandlimited setting, we now wish to optimize r as a function of λ in the hope of obtaining faster than polynomial decay rates.

If one were to treat r as a design parameter in the quantization problem, then the precise dependence on r of constants in any upper bound on $\|F D^r\|_{\text{op}}$, e.g., (8.48) and (8.50) , becomes critical and must be computed.

To resolve this issue, one may use specialized frames such as Sobolev self-dual frames [40]. For such frames Φ, the bound on $\|\Psi D^r\|_{\text{op}}$, where Ψ is the Sobolev dual (and in fact $\Psi = \Phi$) is explicit—see Theorem 8.6. Alternatively, one may work with a given frame Φ but explicitly control the r-dependent constants in the bounds on $\|\Psi D^r\|_{\text{op}}$ where, again, Ψ is the Sobolev dual. This approach is also pursued in [40] for the case of harmonic frames.

It is also important to note that the greedy $\Sigma\Delta$ schemes in (8.38) that we have thus far used to guarantee stability, i.e., to guarantee that $\|u\|$ is bounded, require more levels as the order r increases; see, e.g., [34]. Instead of dealing in the optimization process with the interplay between λ, r, and the number of quantizer levels, one may use alternative $\Sigma\Delta$ schemes where one can choose the number of levels independent of the order r. In particular, we shall use the schemes of [32] and [24] to control $\|u\|$.

It will be convenient to use the following convolution notation. Given infinite sequences $x = (x_i)_{i=-\infty}^{\infty}$ and $y = (y_i)_{i=-\infty}^{\infty}$, the convolution sequence $x * y$ is defined componentwise by

$$\forall i \in \mathbb{Z}, \quad (x * y)_i = \sum_{k=-\infty}^{\infty} x_k y_{i-k}.$$

In the case when $(x_j)_{j=J_1}^{J_2}$ and $(y_k)_{k=K_1}^{K_2}$ are finite sequences, we extend them to infinite sequences by $x_j = 0$, $y_k = 0$ if $j \notin \{J_1, \ldots, J_2\}$, $k \notin \{K_1, \ldots, K_2\}$, and then define the convolution $x * y$ as above.

In the schemes in [24, 32], one substitutes $u = g * v$ for some fixed $g = [g_0, \ldots, g_m]$, where $m \geq r$, $g_0 = 1$, and $g_i \in \mathbb{R}$. Moreover, one sets the quantiza-

tion rule

$$\rho(v_i, v_{i-1}, \ldots, y_i, y_{i-1}, \ldots) = (h * v)_i + y_i,$$

where $h = \delta^{(0)} - \Delta^r g$ (and $\delta^{(0)}$ is the Kronecker delta). Thus, the quantization is performed according to

$$q_i = Q\big((h * v)_i + y_i\big), \tag{8.51}$$

$$v_i = (h * v)_i + y_i - q_i. \tag{8.52}$$

Since $(\Delta^r g)_0 = g_0 = 1$, we have $h_0 = 0$; thus this formula prescribes how v_i is computed from v_j, $j < i$. Here and for the remainder of this section we shall use the midrise quantization alphabet (8.5).

It can be shown (see, e.g., [24, 32]) that the above scheme is stable. The following theorem summarizes its important stability properties.

Theorem 8.5 *There exists a universal constant $C_1 > 0$ such that for any midrise quantization alphabet (8.5) with $2L$ levels and stepsize $\delta > 0$, for any order $r \in \mathbb{N}$, and for all $\mu < \delta(K - \frac{1}{2})$, there exists $g \in \mathbb{R}^m$ for some $m > r$ such that the $\Sigma\Delta$ scheme given in (8.51) is stable for all input signals y with $\|y\|_\infty \le \mu$, and*

$$\|u\|_\infty \le C_1 C_2^r r^r \frac{\delta}{2}, \tag{8.53}$$

where $u = g * v$ as above and $C_2 = (\lceil \frac{\pi^2}{(\cosh^{-1}\gamma)^2} \rceil \frac{e}{\pi})$ with $\gamma := 2K - \frac{2\mu}{\delta}$.

8.5.3 Sobolev Self-dual Frames

We are now ready to define and discuss the properties of the Sobolev self-dual frames proposed in [40]. To that end, recall that for any matrix X in $\mathbb{R}^{m\times n}$ of rank k, there exists a singular value decomposition (SVD) of the form $X = U_X S_X V_X^*$, where $U_X \in \mathbb{R}^{m\times k}$ is a matrix with orthonormal columns, $S_X \in \mathbb{R}^{k\times k}$ is a diagonal matrix with strictly nonnegative entries, and $V_X \in \mathbb{R}^{n\times k}$ is a matrix with orthonormal columns. The Sobolev self-dual frames are constructed from the left singular vectors of the matrix D^r corresponding to its smallest N singular values. Moreover, for any N, M, and r, these frames admit themselves as both canonical duals and Sobolev duals of order r. Figure 8.5 shows the first three coordinates of the first order Sobolev self-dual frame vectors $(\varphi_i)_{i=1}^{1000}$ for $\mathbb{R}^{13}$.

Theorem 8.6 *Let $U_{D^r} = [u_1|u_2|\cdots|u_M]$ be the matrix containing the left singular vectors of D^r, corresponding to the decreasing arrangement of the singular values of D^r. Let $\Phi = [u_{M-N+1}|\cdots|u_{M-1}|u_M]^*$ and denote by Ψ and $(\Phi^*)^\dagger$ the r-th order Sobolev dual and canonical dual of Φ, respectively. Then*

1. *Φ is a tight frame with frame bound* 1,

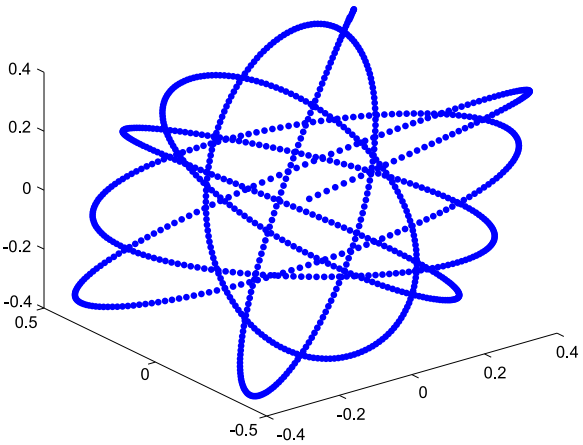


Fig. 8.5 The first three coordinates of 1000 vectors constituting a first order Sobolev self-dual frame for $\mathbb{R}^{13}$

2. $\Psi = (\Phi^*)^\dagger = \Phi$,
3. $\|\Psi D^r\|_{\mathrm{op}} \leq (2\cos(\frac{(M-N-2r+1)\pi}{2M+1}))^r$.

Combining Theorem 8.6 with Theorem 8.5 and optimizing over r, [40] proves the following result.

Theorem 8.7 *For $0 < L \in \mathbb{Z}$ and $0 < \delta \in \mathbb{R}$, let $x \in \mathbb{R}^N$ be such that $\|x\|_2 \leq \mu < \delta(L - \frac{1}{2})$. Suppose that we wish to quantize a redundant representation of x with oversampling rate $\lambda = M/N$ using the $2L$ level midrise alphabet (8.5) with step-size $\delta > 0$. If $\lambda \geq c(\log N)^2$, then there exists a Sobolev self-dual frame Φ and an associated $\Sigma\Delta$ quantization scheme $Q^{\Sigma\Delta}$, both of order $r^{\#} = r(\lambda) \approx \sqrt{\lambda}$, such that*

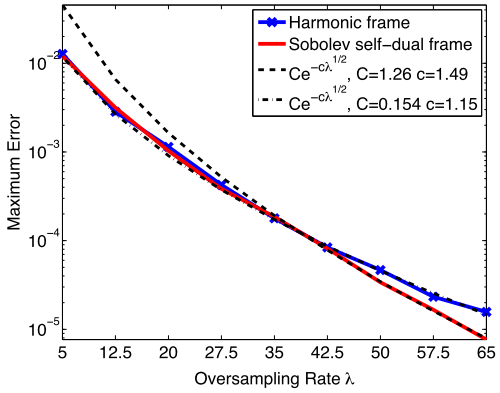
$$\left\|x - \Phi Q^{\Sigma\Delta}\left(\Phi^* x\right)\right\|_2 \leq C_1 e^{-C_2\sqrt{\lambda}}.$$

Here, c, C_1, and C_2 are constants independent of N and x.

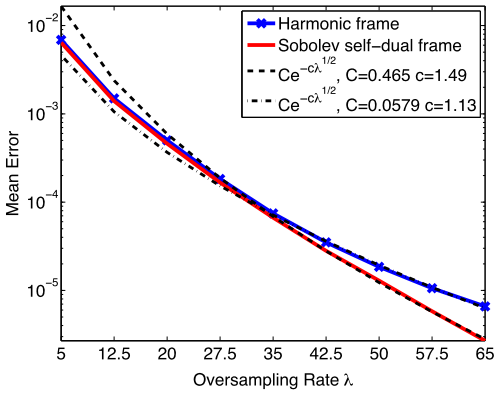
Due to the fact that the frames Φ above admit themselves as both canonical and Sobolev duals, one additionally obtains robustness to noise.

8.5.4 Harmonic Frames

In analogy with Theorem 8.6, [40] presents the following result on harmonic frames.



(a)



(b)

Fig. 8.6 The maximum (**a**) and mean (**b**) error from linear reconstruction of $\Sigma\Delta$ quantized redundant representations with $N = 20$. The error is plotted (in log scale) as a function of the oversampling rate λ

Lemma 8.2 *Let Ψ be the r-th order Sobolev dual of the harmonic frame Φ; then there exist (possibly N-dependent) constants C_1 and C_2, such that*

$$\left\|\Psi D^r\right\|_{\mathrm{op}} \leq C_1 e^{-r/2} M^{-(r+1/2)} r^{r+C_2}\left(1 + O\left(M^{-1}\right)\right).$$

As before, combining Lemma 8.2 with Theorem 8.5 and optimizing over r, [40] obtains the following theorem, showing root-exponential error decay.

Theorem 8.8 *Let $0 < L \in \mathbb{Z}$ and $x \in \mathbb{R}^N$ with $\|x\|_2 \leq \mu < \delta(L - 1/2)$. Suppose that we wish to quantize the harmonic frame expansion $\Phi^* x$ with oversampling rate $\lambda = M/N$ using the $2L$ level midrise alphabet (8.5) with stepsize $\delta > 0$. There*

exists a $\Sigma\Delta$ quantization scheme $Q^{\Sigma\Delta}$ of order $r := r(\lambda) \approx \sqrt{\lambda}$, such that

$$\left\| x - \Psi_r Q^{\Sigma\Delta}\left(\Phi^* x\right)\right\|_2 \leq C_1 e^{-C_2\sqrt{\lambda}}.$$

Here Ψ_r is the rth order Sobolev dual of Φ and the constants are independent of x, but depend on N.

Example 8.5 We run the following experiment to illustrate the results of this section. For $N = 20$ we generate 1500 random vectors $x \in \mathbb{R}^N$ (from the Gaussian ensemble) and normalize their magnitude so that $\|x\| = 2 - \cosh(\pi/\sqrt{6}) \approx 0.0584$. For each x, we obtain the redundant representation $y = \Phi^* x$ where $\Phi \in \mathbb{R}^{N\times M}$ is the harmonic frame or the Sobolev self-dual frame of order r. For $r \in \{1, \ldots, 10\}$ and several values of M, we perform 3-bit $\Sigma\Delta$ quantization on y according to the schemes in Theorem 8.5. Subsequently, we obtain an approximation of x by linear reconstruction using the rth order Sobolev dual of Φ, and the approximation error is computed. For each M, the smallest (over r) of the maximum and mean error (over the 1500 runs) is computed. The resulting error curves are illustrated in Fig. 8.6. Note that both the average and worst case behavior decay as a root exponential, indicating that with the methods and frames of this section, exponential error decay is not possible.

Acknowledgements The authors thank Sinan Güntürk, Mark Lammers, and Thao Nguyen for valuable discussions and collaborations on frame theory and quantization.

A. Powell was supported in part by NSF DMS Grant 0811086 and also gratefully acknowledges the hospitality and support of the Academia Sinica Institute of Mathematics (Taipei, Taiwan).

R. Saab was supported by a Banting Postdoctoral Fellowship, administered by the Natural Science and Engineering Research Council of Canada.

Ö. Yılmaz was supported in part by a Discovery Grant from the Natural Sciences and Engineering Research Council of Canada (NSERC). He was also supported in part by the NSERC CRD Grant DNOISE II (375142-08). Finally, Yılmaz acknowledges the Pacific Institute for the Mathematical Sciences (PIMS) for supporting a CRG in Applied and Computational Harmonic Analysis.

References

1. Benedetto, J.J., Oktay, O.: Pointwise comparison of PCM and $\Sigma\Delta$ quantization. Constr. Approx. **32**, 131158 (2010)
2. Benedetto, J.J., Powell, A.M., Yılmaz, Ö.: Sigma-Delta ($\Sigma\Delta$) quantization and finite frames. IEEE Trans. Inf. Theory **52**, 1990–2005 (2006)
3. Benedetto, J.J., Powell, A.M., Yılmaz, Ö.: Second order Sigma-Delta quantization of finite frame expansions. Appl. Comput. Harmon. Anal. **20**, 126–148 (2006)
4. Bennett, W.R.: Spectra of quantized signals. AT&T Tech. J. **27**(3), 446–472 (1947)
5. Blum, J., Lammers, M., Powell, A.M., Yılmaz, Ö.: Sobolev duals in frame theory and Sigma-Delta quantization. J. Fourier Anal. Appl. **16**, 365–381 (2010)
6. Blum, J., Lammers, M., Powell, A.M., Yılmaz, Ö.: Errata to: Sobolev duals in frame theory and Sigma-Delta quantization. J. Fourier Anal. Appl. **16**, 382 (2010)
7. Bodmann, B., Lipshitz, S.: Randomly dithered quantization and Sigma-Delta noise shaping for finite frames. Appl. Comput. Harmon. Anal. **25**, 367–380 (2008)

8. Bodmann, B., Paulsen, V.: Frames, graphs and erasures. Linear Algebra Appl. **404**, 118–146 (2005)
9. Bodmann, B., Paulsen, V.: Frame paths and error bounds for Sigma-Delta quantization. Appl. Comput. Harmon. Anal. **22**, 176–197 (2007)
10. Bodmann, B., Paulsen, V., Abdulbaki, S.: Smooth frame-path termination for higher order Sigma-Delta quantization. J. Fourier Anal. Appl. **13**, 285–307 (2007)
11. Borodachov, S., Wang, Y.: Lattice quantization error for redundant representations. Appl. Comput. Harmon. Anal. **27**, 334341 (2009)
12. Boufounos, P., Oppenheim, A.: Quantization noise shaping on arbitrary frame expansions. EURASIP J. Appl. Signal Process., Article ID 53807 (2006), 12 pp.
13. Buhler, J., Shokrollahi, M.A., Stemann, V.: Fast and precise Fourier transforms. IEEE Trans. Inf. Theory **46**, 213–228 (2000)
14. Casazza, P., Dilworth, S., Odell, E., Schlumprecht, T., Zsak, A.: Coefficient quantization for frames in Banach spaces. J. Math. Anal. Appl. **348**, 66–86 (2008)
15. Casazza, P., Kovačević, J.: Equal-norm tight frames with erasures. Adv. Comput. Math. **18**, 387–430 (2003)
16. Christensen, O., Goh, S.S.: Pairs of oblique duals in spaces of periodic functions. Adv. Comput. Math. **32**, 353–379 (2010)
17. Christensen, O., Kim, H.O., Kim, R.Y.: Gabor windows supported on $[-1, 1]$ and compactly supported dual windows. Appl. Comput. Harmon. Anal. **28**, 89–103 (2010)
18. Christensen, O., Sun, W.: Explicitly given pairs of dual frames with compactly supported generators and applications to irregular B-splines. J. Approx. Theory **151**, 155–163 (2008)
19. Cvetkovic, Z.: Resilience properties of redundant expansions under additive noise and quantization. IEEE Trans. Inf. Theory **49**, 644–656 (2003)
20. Cvetkovic, Z., Vetterli, M.: On simple oversampled A/D conversion in $L^2(\mathbb{R})$. IEEE Trans. Inf. Theory **47**, 146–154 (2001)
21. Daubechies, I., DeVore, R.: Approximating a bandlimited function using very coarsely quantized data: a family of stable Sigma-Delta modulators of arbitrary order. Ann. Math. **158**, 679–710 (2003)
22. Daubechies, I., DeVore, R.A., Güntürk, C.S., Vaishampayan, V.A.: A/D conversion with imperfect quantizers. IEEE Trans. Inf. Theory **52**, 874–885 (2006)
23. Daubechies, I., Landau, H., Landau, Z.: Gabor time-frequency lattices and the Wexler-Raz identity. J. Fourier Anal. Appl. **1**, 437–478 (1995)
24. Deift, P., Güntürk, C.S., Krahmer, F.: An optimal family of exponentially accurate one-bit Sigma-Delta quantization schemes. Commun. Pure Appl. Math. **64**, 883–919 (2011)
25. Deshpande, A., Sarma, S.E., Goyal, V.K.: Generalized regular sampling of trigonometric polynomials and optimal sensor arrangement. IEEE Signal Process. Lett. **17**, 379–382 (2010)
26. Dilworth, S., Odell, E., Schlumprecht, T., Zsak, A.: Coefficient quantization in Banach spaces. Found. Comput. Math. **8**, 703–736 (2008)
27. Eldar, Y., Christensen, O.: Characterization of oblique dual frame pairs. EURASIP J. Appl. Signal Process., Article ID 92674 (2006), 11 pp.
28. Games, R.A.: Complex approximations using algebraic integers. IEEE Trans. Inf. Theory **31**, 565–579 (1985)
29. Goyal, V., Kovačević, J., Kelner, J.: Quantized frame expansions with erasures. Appl. Comput. Harmon. Anal. **10**, 203–233 (2001)
30. Goyal, V., Vetterli, M., Thao, N.T.: Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory **44**, 16–31 (1998)
31. Gray, R., Stockham, T.: Dithered quantizers. IEEE Trans. Inf. Theory **39**, 805–812 (1993)
32. Güntürk, C.S.: One-bit Sigma-Delta quantization with exponential accuracy. Commun. Pure Appl. Math. **56**, 1608–1630 (2003)
33. Güntürk, C.S.: Approximating a bandlimited function using very coarsely quantized data: improved error estimates in Sigma-Delta modulation. J. Am. Math. Soc. **17**, 229242 (2004)
34. Güntürk, C.S., Lammers, M., Powell, A.M., Saab, R., Yılmaz, Ö.: Sobolev duals for random frames and Sigma-Delta quantization of compressed sensing measurements, preprint (2010)

35. Güntürk, C.S., Lammers, M., Powell, A.M., Saab, R., Yılmaz, Ö.: Sigma Delta quantization for compressed sensing. In: 44th Annual Conference on Information Sciences and Systems, Princeton, NJ, March (2010)
36. Güntürk, C.S., Lammers, M., Powell, A.M., Saab, R., Yılmaz, Ö.: Sobolev duals of random frames. In: 44th Annual Conference on Information Sciences and Systems, Princeton, NJ, March (2010)
37. Güntürk, C.S., Thao, N.: Ergodic dynamics in Sigma-Delta quantization: tiling invariant sets and spectral analysis of error. Adv. Appl. Math. **34**, 523–560 (2005)
38. Inose, H., Yasuda, Y.: A unity bit coding method by negative feedback. Proc. IEEE **51**, 1524–1535 (1963)
39. Jimenez, D., Wang, L., Wang, Y.: White noise hypothesis for uniform quantization errors. SIAM J. Math. Anal. **28**, 2042–2056 (2007)
40. Krahmer, F., Saab, R., Ward, R.: Root-exponential accuracy for coarse quantization of finite frame expansions. IEEE Trans. Inf. Theory **58**, 1069–1079 (2012)
41. Lammers, M., Maeser, A.: An uncertainty principle for finite frames. J. Math. Anal. Appl. **373**, 242247 (2011)
42. Lammers, M., Powell, A.M., Yılmaz, Ö.: Alternative dual frames for digital-to-analog conversion in Sigma-Delta quantization. Adv. Comput. Math. **32**, 73–102 (2010)
43. Li, S., Ogawa, H.: Optimal noise suppression: a geometric nature of pseudoframes for subspaces. Adv. Comput. Math. **28**, 141–155 (2008)
44. Li, S., Ogawa, H.: Pseudo-duals of frames with applications. Appl. Comput. Harmon. Anal. **11**, 289–304 (2001)
45. Norsworthy, S., Schreier, R., Temes, G. (eds.): Delta-Sigma Data Converters. IEEE Press, New York (1997)
46. Powell, A.M.: Mean squared error bounds for the Rangan-Goyal soft thresholding algorithm. Appl. Comput. Harmon. Anal. **29**, 251–271 (2010)
47. Powell, A.M., Tanner, J., Yılmaz, Ö., Wang, Y.: Coarse quantization for random interleaved sampling of bandlimited signals. ESAIM, Math. Model. Numer. Anal. **46**, 605–618 (2012)
48. Powell, A.M., Whitehouse, J.T.: Consistent reconstruction error bounds, random polytopes and coverage processes, preprint (2011)
49. Rangan, S., Goyal, V.: Recursive consistent estimation with bounded noise. IEEE Trans. Inf. Theory **47**, 457–464 (2001)
50. Thao, N.: Deterministic analysis of oversampled A/D conversion and decoding improvement based on consistent estimates. IEEE Trans. Signal Process. **42**, 519–531 (1994)
51. Thao, N., Vetterli, M.: Reduction of the MSE in R-times oversampled A/D conversion from $\mathcal{O}(1/R)$ to $\mathcal{O}(1/R^2)$. IEEE Trans. Signal Process. **42**, 200–203 (1994)
52. Wang, Y.: Sigma-Delta quantization errors and the traveling salesman problem. Adv. Comput. Math. **28**, 101118 (2008)
53. Wang, Y., Xu, Z.: The performance of PCM quantization under tight frame representations, preprint (2011)
54. Yılmaz, Ö.: Stability analysis for several second-order Sigma-Delta methods of coarse quantization of bandlimited functions. Constr. Approx. **18**, 599–623 (2002)

Chapter 9
Finite Frames for Sparse Signal Processing

Waheed U. Bajwa and Ali Pezeshki

Abstract Over the last decade, considerable progress has been made toward developing new signal processing methods to manage the deluge of data caused by advances in sensing, imaging, storage, and computing technologies. Most of these methods are based on a simple but fundamental observation: high-dimensional data sets are typically highly redundant and live on low-dimensional manifolds or subspaces. This means that the collected data can often be represented in a sparse or parsimonious way in a suitably selected finite frame. This observation has also led to the development of a new sensing paradigm, called compressed sensing, which shows that high-dimensional data sets can often be reconstructed, with high fidelity, from only a small number of measurements. Finite frames play a central role in the design and analysis of both sparse representations and compressed sensing methods. In this chapter, we highlight this role primarily in the context of compressed sensing for estimation, recovery, support detection, regression, and detection of sparse signals. The recurring theme is that frames with small spectral norm and/or small worst-case coherence, average coherence, or sum coherence are well suited for making measurements of sparse signals.

Keywords Approximation theory · Coherence property · Compressed sensing · Detection · Estimation · Grassmannian frames · Model selection · Regression · Restricted isometry property · Typical guarantees · Uniform guarantees · Welch bound

W.U. Bajwa (✉)
Department of Electrical and Computer Engineering, Rutgers, The State University of New Jersey, 94 Brett Rd, Piscataway, NJ 08854, USA
e-mail: waheed.bajwa@rutgers.edu

A. Pezeshki
Department of Electrical and Computer Engineering, Colorado State University, Fort Collins, CO 80523, USA
e-mail: ali.pezeshki@colostate.edu

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_9,

9.1 Introduction

It was not too long ago that scientists, engineers, and technologists were complaining about *data starvation*. In many applications, there never was sufficient data available to reliably carry out various inference and decision-making tasks in real time. Technological advances during the last two decades, however, have changed all of that—so much in fact that *data deluge*, instead of data starvation, is now becoming a concern. If left unchecked, the rate at which data is being generated in numerous applications will soon overwhelm the associated systems' computational and storage resources.

During the last decade or so, there has been a surge of research activity in the signal processing and statistics communities to deal with the problem of data deluge. The proposed solutions to this problem rely on a simple but fundamental principle of *redundancy*. Massive data sets in the real world may live in high-dimensional spaces, but the information embedded within these data sets almost always lives near low-dimensional (often linear) manifolds. There are two ways in which the principle of redundancy can help us better manage the sheer abundance of data. First, we can represent the collected data in a parsimonious (or *sparse*) manner in carefully designed bases and frames. Sparse representations of data help reduce their (computational and storage) footprint and constitute an active area of research in signal processing [12]. Second, we can redesign the *sensing systems* to acquire only a small number of measurements by exploiting the low-dimensional nature of the signals of interest. The term *compressed sensing* has been coined for the area of research that deals with rethinking the design of sensing systems under the assumption that the signal of interest has a sparse representation in a *known* basis or frame [1, 16, 27].

There is a fundamental difference between the two aforementioned approaches to dealing with the data deluge; the former deals with the collected data while the latter deals with the collection of data. Despite this difference, however, there exists a great deal of mathematical similarity between the areas of sparse signal representation and compressed sensing. Our primary focus in this chapter will be on the compressed sensing setup and the role of finite frames in its development. However, many of the results discussed in this context can be easily restated for sparse signal representation. We will therefore use the generic term *sparse signal processing* in this chapter to refer to the collection of these results.

Mathematically, sparse signal processing deals with the case when a highly redundant frame $\Phi = (\varphi_i)_{i=1}^M$ in $\mathscr{H}^N$ is used to make (possibly noisy) measurements of sparse signals.[1] Consider an arbitrary signal $x \in \mathscr{H}^M$ that is K-sparse: $\|x\|_0 := \sum_{i=1}^M 1_{\{x_i \neq 0\}}(x) \leq K < N \ll M$. Instead of measuring x directly, sparse signal processing uses a small number of linear measurements of x, given by $y = \Phi x + n$, where $n \in \mathscr{H}^N$ corresponds to deterministic perturbation or stochastic noise. Given measurements y of x, the fundamental problems in sparse signal

[1] The sparse signal processing literature often uses the terms *sensing matrix*, *measurement matrix*, and *dictionary* for the frame Φ in this setting.

processing include: (i) recovering/estimating the sparse signal x, (ii) estimating x for linear regression, (iii) detecting the locations of the nonzero entries of x, and (iv) testing for the presence of x in noise. In all of these problems, certain geometrical properties of the frame Φ play crucial roles in determining the optimality of the end solutions. In this chapter, our goal is to make explicit these connections between the geometry of frames and sparse signal processing.

The four geometric measures of frames that we focus on in this chapter include the *spectral norm*, *worst-case coherence*, *average coherence*, and *sum coherence*. Recall that the spectral norm $\|\Phi\|$ of a frame Φ is simply a measure of its tightness and is given by the maximum singular value: $\|\Phi\| = \sigma_{\max}(\Phi)$. The worst-case coherence μ_Φ, defined as

$$\mu_\Phi := \max_{\substack{i,j\in\{1,\dots,M\}\\ i\neq j}} \frac{|\langle\varphi_i,\varphi_j\rangle|}{\|\varphi_i\|\|\varphi_j\|}, \tag{9.1}$$

is a measure of the similarity between different frame elements. On the other hand, the average coherence is a new notion of frame coherence, introduced recently in [2, 3] and analyzed further in [4]. In words, the average coherence ν_Φ, defined as

$$\nu_\Phi := \frac{1}{M-1} \max_{i\in\{1,\dots,M\}} \left| \sum_{\substack{j=1\\ j\neq i}}^{M} \frac{\langle\varphi_i,\varphi_j\rangle}{\|\varphi_i\|\|\varphi_j\|} \right|, \tag{9.2}$$

is a measure of the spread of normalized frame elements $(\varphi_i/\|\varphi_i\|)_{i=1}^M$ in the unit ball. The sum coherence, defined as

$$\sum_{j=2}^{M}\sum_{i=1}^{j-1} \frac{|\langle\varphi_i,\varphi_j\rangle|}{\|\varphi_i\|\|\varphi_j\|}, \tag{9.3}$$

is a notion of coherence that arises in the context of detecting the presence of a sparse signal in noise [76, 77].

In the following sections, we show that different combinations of these geometric measures characterize the performance of a multitude of sparse signal processing algorithms. In particular, a theme that emerges time and again throughout this chapter is that frames with small spectral norm and/or small worst-case coherence, average coherence, or sum coherence are particularly well suited for the purposes of making measurements of sparse signals.

Before proceeding further, we note that the signal x in some applications is sparse in the identity basis, in which case Φ represents the measurement process itself. In other applications, however, x can be sparse in some other orthonormal basis or an overcomplete dictionary Ψ. In this case, Φ corresponds to a composition of Θ, the frame resulting from the measurement process, and Ψ, the sparsifying dictionary, i.e., $\Phi = \Theta\Psi$. We do not make a distinction between the two formulations in this chapter. In particular, while the reported results are most readily interpretable in a physical setting for the former case, they are easily extendable to the latter case.

We note that this chapter provides an overview of only a small subset of current results in sparse signal processing literature. Our aim is simply to highlight the central role that finite frame theory plays in the development of sparse signal processing theory. We refer the interested reader to [34] and the references therein for a more comprehensive review of the sparse signal processing literature.

9.2 Sparse Signal Processing: Uniform Guarantees and Grassmannian Frames

Recall the fundamental system of equations in sparse signal processing: $y = \Phi x + n$. Given the measurements y, our goal in this section is to specify conditions on the frame Φ and accompanying computational methods that enable reliable inference of the high-dimensional sparse signal x from the low-dimensional measurements y. There has been a lot of work in this direction in the sparse signal processing literature. Our focus in this section is on providing an overview of some of the key results in the context of performance guarantees for *every* K-sparse signal in $\mathscr{H}^M$ using a *fixed* frame Φ. It is shown in the following that *uniform performance guarantees* for sparse signal processing are directly tied to the worst-case coherence of frames. In particular, the closer a frame is to being a *Grassmannian frame*—defined as one that has the smallest worst-case coherence for given N and M—the better its performance is in the uniform sense.

9.2.1 Recovery of Sparse Signals via ℓ_0 Minimization

We consider the simplest of setups in sparse signal processing, corresponding to the recovery of a sparse signal x from noiseless measurements $y = \Phi x$. Mathematically speaking, this problem is akin to solving an *underdetermined* system of linear equations. Although an underdetermined system of linear equations has infinitely many solutions in general, one of the surprises of sparse signal processing is that recovery of x from y remains a well-posed problem for large classes of random and deterministic frames because of the underlying sparsity assumption. Since we are looking to solve y for a K-sparse x, an intuitive way of obtaining a candidate solution from y is to search for the sparsest solution $\widehat{x}_0$ that satisfies $y = \Phi \widehat{x}_0$. Mathematically, this solution criterion can be expressed in terms of the following ℓ_0 minimization program:

$$\widehat{x}_0 = \arg \min_{z \in \mathscr{H}^M} \|z\|_0 \quad \text{subject to} \quad y = \Phi z. \tag{P_0}$$

Despite the apparent simplicity of (P_0), the conditions under which it can be claimed that $\widehat{x}_0 = x$ for any $x \in \mathscr{H}^M$ are not immediately obvious. Given that

(P_0) is a highly nonconvex optimization, there is in fact little reason to expect that $\widehat{x}_0$ should be unique to begin with. It is because of these roadblocks that a rigorous mathematical understanding of (P_0) alluded researchers for a long time. These mathematical challenges were eventually overcome through surprisingly elementary mathematical tools in [28, 41]. In particular, it is argued in [41] that a property termed the *unique representation property* (URP) of Φ is the key to understanding the behavior of the solution obtained from (P_0).

Definition 9.1 (Unique Representation Property) A frame $\Phi = (\varphi_i)_{i=1}^M$ in $\mathscr{H}^N$ is said to have the unique representation property of order K if any K frame elements of Φ are linearly independent.

It has been shown in [28, 41] that the URP of order $2K$ is both a necessary and a sufficient condition for the equivalence of $\widehat{x}_0$ and x.[2]

Theorem 9.1 [28, 41] *An arbitrary K-sparse signal x can be uniquely recovered from $y = \Phi x$ as a solution to (P_0) if and only if Φ satisfies the URP of order $2K$.*

The proof of Theorem 9.1 is simply an exercise in elementary linear algebra. It follows from the simple observation that K-sparse signals in $\mathscr{H}^M$ are mapped injectively into $\mathscr{H}^N$ if and only if the nullspace of Φ does not contain nontrivial $2K$-sparse signals. In order to understand the significance of Theorem 9.1, note that random frames with elements distributed uniformly at random on the unit sphere in $\mathscr{H}^N$ will almost surely have the URP of order $2K$ as long as $N \geq 2K$. This is rather powerful, since this signifies that sparse signals can be recovered from a number of random measurements that are only linear in the *sparsity* K of the signal, rather than the ambient dimension M. Despite this powerful result, however, Theorem 9.1 is rather opaque in the case of arbitrary (not necessarily random) frames. The reason is that the URP is a local geometric property of Φ, and explicitly verifying the URP of order $2K$ requires a combinatorial search over all $\binom{M}{2K}$ possible collections of frame elements. Nevertheless, it is possible to replace the URP in Theorem 9.1 with the worst-case coherence of Φ, which is a global geometric property of Φ that can be easily computed in polynomial time. The key to this is the classical *Geršgorin circle theorem* [40], which can be used to relate the URP of a frame Φ to its worst-case coherence.

Lemma 9.1 (Geršgorin) *Let $t_{i,j}, i, j = 1, \ldots, M$, denote the entries of an $M \times M$ matrix T. Then every eigenvalue of T lies in at least one of the M circles defined*

[2] Theorem 9.1 has been stated in [28] using the terminology of *spark*, instead of the URP. The spark of a frame Φ is defined in [28] as the smallest number of frame elements of Φ that are linearly dependent. In other words, Φ satisfies the URP of order K if and only if $\text{spark}(\Phi) \geq K + 1$.

below:

$$\mathscr{D}_i(T) = \left\{ z \in \mathbb{C} : |z - t_{i,i}| \leq \sum_{\substack{j=1 \\ j \neq i}}^{M} |t_{i,j}| \right\}, \quad i = 1, \ldots, M. \tag{9.4}$$

The Geršgorin circle theorem seems to have first appeared in 1931 in [40], and its proof can be found in any standard text on matrix analysis such as [50]. This theorem allows one to relate the worst-case coherence of Φ to the URP as follows.

Theorem 9.2 [28] *Let Φ be a unit norm frame and $K \in \mathbb{N}$. Then Φ satisfies the URP of order K as long as $K < 1 + \mu_\Phi^{-1}$.*

The proof of this theorem follows by bounding the minimum eigenvalue of any $K \times K$ principal submatrix of the Gramian matrix G_Φ using Lemma 9.1. We can now combine Theorem 9.1 with Theorem 9.2 to obtain the following theorem that relates the worst-case coherence of Φ to the sparse signal recovery performance of (P_0).

Theorem 9.3 *An arbitrary K-sparse signal x can be uniquely recovered from $y = \Phi x$ as a solution to (P_0), provided*

$$K < \frac{1}{2}\left(1 + \mu_\Phi^{-1}\right). \tag{9.5}$$

Theorem 9.3 states that ℓ_0 minimization enables unique recovery of *every* K-sparse signal measured using a frame Φ as long as $K = O(\mu_\Phi^{-1})$.[3] This dictates that frames that have small worst-case coherence are particularly well suited for measuring sparse signals. It is also instructive to understand the fundamental limitations of Theorem 9.3. In order to do so, we recall the following fundamental lower bound on the worst-case coherence of unit norm frames.

Lemma 9.2 (The Welch Bound [75]) *The worst-case coherence of any unit norm frame $\Phi = (\varphi_i)_{i=1}^M$ in $\mathscr{H}^N$ satisfies the inequality $\mu_\Phi \geq \sqrt{\frac{M-N}{N(M-1)}}$.*

It can be seen from the Welch bound that $\mu_\Phi = \Omega(N^{-1/2})$ as long as $M > N$. Therefore, we have from Theorem 9.3 that even in the best of cases ℓ_0 minimization yields unique recovery of every sparse signal as long as $K = O(\sqrt{N})$. This implication is weaker than the $K = O(N)$ scaling that we observed earlier for random frames. A natural question to ask therefore is whether Theorem 9.3 is weak

[3] Recall, with big-O notation, that $f(n) = O(g(n))$ if there exist positive C and n_0 such that for all $n > n_0$, $f(n) \leq Cg(n)$. Also, $f(n) = \Omega(g(n))$ if $g(n) = O(f(n))$, and $f(n) = \Theta(g(n))$ if $f(n) = O(g(n))$ and $g(n) = O(f(n))$.

in terms of the relationship between K and μ_Φ. The answer to this question however is in the negative, since there exist frames such as union of identity and Fourier bases [30] and Steiner equiangular tight frames [36] that have certain collections of frame elements with cardinality $O(\sqrt{N})$ that are linearly dependent. We therefore conclude from the preceding discussion that Theorem 9.3 is tight from the frame-theoretic perspective and, in general, frames with small worst-case coherence are better suited for recovery of sparse signals using (P_0). In particular, this highlights the importance of Grassmannian frames in the context of sparse signal recovery in the uniform sense.

9.2.2 *Recovery and Estimation of Sparse Signals via Convex Optimization and Greedy Algorithms*

The implications of Sect. 9.2.1 are quite remarkable. We have seen that it is possible to recover a K-sparse signal x using a small number of measurements that is proportional to μ_Φ^{-1}; in particular, for large classes of frames such as *Gabor frames* [3], we see that $O(K^2)$ number of measurements suffice to recover a sparse signal using ℓ_0 minimization. This can be significantly smaller than the $N = M$ measurements dictated by classical signal processing when $K \ll M$. Despite this, however, sparse signal recovery using (P_0) is something that one cannot be expected to use for practical purposes. The reason for this is the computational complexity associated with ℓ_0 minimization; in order to solve (P_0), one needs to exhaustively search through all possible sparsity levels. The complexity of such exhaustive search is clearly exponential in M, and it has been shown in [54] that (P_0) is in general an NP-hard problem. Alternate methods of solving $y = \Phi x$ for a K-sparse x that are also computationally feasible therefore have been of great interest to the practitioners. The recent interest in the literature on sparse signal processing partly stems from the fact that significant progress has been made by numerous researchers in obtaining various practical alternatives to (P_0). Such alternatives range from convex optimization-based methods [18, 22, 66] to greedy algorithms [25, 51, 55]. In this subsection, we review the performance guarantees of two such seminal alternative methods that are widely used in practice and once again highlight the role Grassmannian frames play in sparse signal processing.

9.2.2.1 Basis pursuit

A common heuristic approach taken in solving nonconvex optimization problems is to approximate them with a convex problem and solve the resulting optimization program. A similar approach can be taken to *convexify* (P_0) by replacing the ℓ_0 "norm" in (P_0) with its closest convex approximation, the ℓ_1 norm: $\|z\|_1 = \sum_i |z_i|$. The resulting optimization program, which seems to have been first proposed as a

heuristic in [59], can be formally expressed as follows:

$$\widehat{x}_1 = \arg \min_{z \in \mathscr{H}^M} \|z\|_1 \quad \text{subject to} \quad y = \Phi z. \tag{P_1}$$

The ℓ_1 minimization program (P_1) is termed *basis pursuit* (BP) [22] and is in fact a linear optimization program [11]. A number of numerical methods have been proposed for solving BP in an efficient manner; we refer the reader to [72] for a survey of some of these methods.

Even though BP has existed in the literature since at least the mid-1980s [59], it is only in the last decade that results concerning its performance have been reported. Below, we present one such result that is expressed in terms of the worst-case coherence of the frame Φ [28, 42].

Theorem 9.4 [28, 42] *An arbitrary K-sparse signal x can be uniquely recovered from $y = \Phi x$ as a solution to (P_1) provided*

$$K < \frac{1}{2}\left(1 + \mu_\Phi^{-1}\right). \tag{9.6}$$

The reader will notice that the sparsity requirements in both Theorem 9.3 and Theorem 9.4 are the same. However, this does not mean that (P_0) and (P_1) always yield the same solution, because the sparsity requirements in the two theorems are only sufficient conditions. Regardless, it is rather remarkable that one can solve an underdetermined system of equations $y = \Phi x$ for a K-sparse x in polynomial time as long as $K = O(\mu_\Phi^{-1})$. In particular, we can once again conclude from Theorem 9.4 that frames with small worst-case coherence in general and Grassmannian frames in particular are highly desirable in the context of recovery of sparse signals using BP.

9.2.2.2 Orthogonal matching pursuit

BP is arguably a highly practical scheme for recovering a K-sparse signal x from the set of measurements $y = \Phi x$. In particular, depending upon the particular implementation, the computational complexity of convex optimization methods like BP for general frames is typically $O(M^3 + NM^2)$, which is much better than the complexity of (P_0), assuming P $\neq$ NP. Nevertheless, BP can be computationally demanding for large-scale sparse recovery problems. Fortunately, there do exist greedy alternatives to optimization-based approaches for sparse signal recovery. The oldest and perhaps the most well-known among these greedy algorithms goes by the name of *orthogonal matching pursuit* (OMP) in the literature [51]. Note that just like BP, OMP has been in practical use for a long time, but it is only recently that its performance has been characterized by the researchers.

The OMP algorithm obtains an estimate $\widehat{\mathscr{K}}$ of the indices of the frame elements $\{\varphi_i : x_i \neq 0\}$ that contribute to the measurements $y = \sum_{i:x_i \neq 0} \varphi_i x_i$. The final OMP

Algorithm 1 Orthogonal Matching Pursuit

Input: Unit norm frame Φ and measurement vector y
Output: Sparse OMP estimate $\widehat{x}_{\text{OMP}}$
Initialize: $i = 0$, $\widehat{x}^0 = 0$, $\widehat{\mathscr{K}} = \emptyset$, and $r^0 = y$

while $\|r^i\| \geq \epsilon$ **do**
- $i \leftarrow i + 1$ {Increment counter}
- $z \leftarrow \Phi^* r^{i-1}$ {Form signal proxy}
- $\ell \leftarrow \arg\max_j |z_j|$ {Select frame element}
- $\widehat{\mathscr{K}} \leftarrow \widehat{\mathscr{K}} \cup \{\ell\}$ {Update the index set}
- $\widehat{x}^i_{\widehat{\mathscr{K}}} \leftarrow \Phi^\dagger_{\widehat{\mathscr{K}}} y$ and $\widehat{x}^i_{\widehat{\mathscr{K}}^c} \leftarrow 0$ {Update the estimate}
- $r^i \leftarrow y - \Phi \widehat{x}^i$ {Update the residue}

end while
return $\widehat{x}_{\text{OMP}} = \widehat{x}^i$

estimate $\widehat{x}_{\text{OMP}}$ then corresponds to a least-squares estimate of x using the frame elements $\{\varphi_i\}_{i \in \widehat{\mathscr{K}}}$: $\widehat{x}_{\text{OMP}} = \Phi^\dagger_{\widehat{\mathscr{K}}} y$, where $(\cdot)^\dagger$ denotes the Moore–Penrose pseudoinverse. In order to estimate the indices, the OMP starts with an empty set and greedily expands that set by one additional frame element in each iteration. A formal description of the OMP algorithm is presented in Algorithm 1, in which $\epsilon > 0$ is a stopping threshold. The power of OMP stems from the fact that if the estimate delivered by the algorithm has exactly K nonzeros then its computational complexity is only $O(NMK)$, which is typically much better than the computational complexity of $O(M^3 + NM^2)$ for convex optimization-based approaches. We are now ready to state a theorem characterizing the performance of the OMP algorithm in terms of the worst-case coherence of frames.

Theorem 9.5 [29, 68] *An arbitrary K-sparse signal x can be uniquely recovered from $y = \Phi x$ as a solution to the OMP algorithm with $\epsilon = 0$, provided*

$$K < \frac{1}{2}\left(1 + \mu_\Phi^{-1}\right). \tag{9.7}$$

Theorem 9.5 shows that the guarantees for the OMP algorithm in terms of the worst-case coherence match those for both (P_0) and BP; OMP too requires that $K = O(\mu_\Phi^{-1})$ in order for it to successfully recover a K-sparse x from $y = \Phi x$. However, it cannot be emphasized enough that once $K = \Omega(\mu_\Phi^{-1})$, we start to see a difference in the empirical performance of (P_0), BP, and OMP. Nevertheless, the basic insight of Theorems 9.3–9.5 that frames with smaller worst-case coherence improve the recovery performance remains valid in all three cases.

9.2.2.3 Estimation of sparse signals

Our focus in this section has so far been on recovery of sparse signals from the measurements $y = \Phi x$. In practice, however, it is seldom the case that one obtains

measurements of a signal without any additive noise. A more realistic model for measurement of sparse signals in this case can be expressed as $y = \Phi x + n$, where n represents either deterministic or random noise. In the presence of noise, one's objective changes from sparse signal recovery to sparse signal estimation; the goal being an estimate $\widehat{x}$ that is to close to the original sparse signal x in an ℓ_2 sense.

It is clear from looking at (P_1) that BP in its current form should not be used for estimation of sparse signals in the presence of noise, since $y \neq \Phi x$ in this case. However, a simple modification of the constraint in (P_1) allows us to gracefully handle noise in sparse signal estimation problems. The modified optimization program can be formally described as

$$\widehat{x}_1 = \arg \min_{z \in \mathscr{H}^M} \|z\|_1 \quad \text{subject to} \quad \|y - \Phi z\| \leq \epsilon \tag{P_1^ϵ}$$

where ϵ is typically chosen to be equal to the noise magnitude: $\epsilon = \|n\|$. The optimization (P_1^ϵ) is often called *basis pursuit with inequality constraint* (BPIC). It is easy to check that BPIC is also a convex optimization program, although it is no longer a linear program. Performance guarantees based upon the worst-case coherence for BPIC in the presence of deterministic noise alluded researchers for quite some time. The problem was settled recently in [29], and the solution is summarized in the following theorem.

Theorem 9.6 [29] *Suppose that an arbitrary K-sparse signal x satisfies the sparsity constraint $K < \frac{1+\mu_\Phi^{-1}}{4}$. Given $y = \Phi x + n$, BPIC with $\epsilon = \|n\|$ can be used to obtain an estimate $\widehat{x}_1$ of x such that*

$$\|x - \widehat{x}_1\| \leq \frac{2\epsilon}{\sqrt{1 - \mu_\Phi(4K-1)}}. \tag{9.8}$$

Theorem 9.6 states that BPIC with an appropriate ϵ results in a stable solution, *despite* the fact that we are dealing with an underdetermined system of equations. In particular, BPIC also handles sparsity levels that are $O(\mu_\Phi^{-1})$ and results in a solution that differs from the true signal x by $O(\|n\|)$.

In contrast with BP, OMP in its original form can be run for both noiseless sparse signal recovery and noisy sparse signal estimation. The only thing that changes in OMP in the latter case is the value of ϵ, which typically should also be set equal to the noise magnitude. The following theorem characterizes the performance of OMP in the presence of noise [29, 67, 69].

Theorem 9.7 [29, 67, 69] *Suppose that $y = \Phi x + n$ for an arbitrary K-sparse signal x and OMP is used to obtain an estimate $\widehat{x}_{\text{OMP}}$ of x with $\epsilon = \|n\|$. Then the OMP solution satisfies*

$$\|x - \widehat{x}_{\text{OMP}}\| \leq \frac{\epsilon}{\sqrt{1 - \mu_\Phi(K-1)}} \tag{9.9}$$

provided x satisfies the sparsity constraint

$$K < \frac{1+\mu_{\Phi}^{-1}}{2} - \frac{\epsilon \cdot \mu_{\Phi}^{-1}}{x_{\min}}. \tag{9.10}$$

Here, $x_{\min}$ denotes the smallest (in magnitude) nonzero entry of x: $x_{\min} = \min_{i:x_i \neq 0} |x_i|$.

It is interesting to note that, unlike the case of sparse signal recovery, OMP in the noisy case does *not* have guarantees similar to that of BPIC. In particular, while the estimation error in OMP is still $O(\|n\|)$, the sparsity constraint in the case of OMP becomes restrictive as the smallest (in magnitude) nonzero entry of x decreases.

The estimation error guarantees provided in Theorem 9.6 and Theorem 9.7 are near-optimal for the case when the noise n follows an adversarial (or deterministic) model. This happens because the noise n under the adversarial model can always be aligned with the signal x, making it impossible to guarantee an estimation error smaller than the size of n. However, if one is dealing with stochastic noise, then it is possible to improve upon the estimation error guarantees for sparse signals. In order to do that, we first define a Lagrangian relaxation of (P_1^{ϵ}), which can be formally expressed as

$$\widehat{x}_{1,2} = \arg \min_{z \in \mathscr{H}^M} \frac{1}{2}\|y - \Phi x\| + \tau \|z\|_1. \tag{$P_{1,2}$}$$

The mixed-norm optimization program ($P_{1,2}$) goes by the name of *basis pursuit denoising* (BPDN) [22] as well as *least absolute shrinkage and selection operator* (LASSO) [66]. In the following, we state estimation error guarantees for both the LASSO and OMP under the assumption of an *additive white Gaussian noise* (AWGN): $n \sim \mathscr{N}(0, \sigma^2 Id)$.

Theorem 9.8 [6] *Suppose that $y = \Phi x + n$ for an arbitrary K-sparse signal x, the noise n is distributed as $\mathscr{N}(0, \sigma^2 Id)$, and the LASSO is used to obtain an estimate $\widehat{x}_{1,2}$ of x with $\tau = 4\sqrt{\sigma^2 \log(M-K)}$. Then under the assumption that x satisfies the sparsity constraint $K < \frac{\mu_{\Phi}^{-1}}{3}$, the LASSO solution satisfies support$(\widehat{x}_{1,2}) \subset$ support(x) and*

$$\|x - \widehat{x}_{1,2}\| \leq \left(\sqrt{3} + 3\sqrt{4\log(M-K)}\right)^2 K\sigma^2 \tag{9.11}$$

with probability exceeding $(1 - \frac{1}{(M-K)^2})(1 - e^{-K/7})$.

A few remarks are in order now concerning Theorem 9.8. First, note that the results of the theorem hold with high probability since there exists a small probability that the Gaussian noise aligns with the sparse signal. Second, (9.11) shows that the estimation error associated with the LASSO solution is $O(\sqrt{\sigma^2 K \log M})$. This estimation error is within a logarithmic factor of the *best unbiased* estimation

error $O(\sqrt{\sigma^2 K})$ that one can obtain in the presence of stochastic noise.[4] Ignoring the probabilistic aspect of Theorem 9.8, it is also worth comparing the estimation error of Theorem 9.6 with that of the LASSO. It is a tedious but simple exercise in probability to show that $\|n\| = \Omega(\sqrt{\sigma^2 M})$ with high probability. Therefore, if one applies Theorem 9.6 directly to the case of stochastic noise, then one obtains that the square of the estimation error scales linearly with the ambient dimension M of the sparse signal. On the other hand, Theorem 9.8 yields that the square of the estimation error scales linearly with the sparsity (modulo a logarithmic factor) of the sparse signal. This highlights the differences that exist between guarantees obtained under a deterministic noise model versus a stochastic (random) noise model.

We conclude this subsection by noting that it is also possible to obtain better OMP estimation error guarantees for the case of stochastic noise *provided* one inputs the sparsity of x to the OMP algorithm and modifies the halting criterion in Algorithm 1 from $\|r^i\| \geq \epsilon$ to $i \leq K$ (i.e., the OMP is restricted to K iterations only). Under this modified setting, the guarantees for the OMP algorithm can be stated in terms of the following theorem.

Theorem 9.9 [6] *Suppose that $y = \Phi x + n$ for an arbitrary K-sparse signal x, the noise n is distributed as $\mathcal{N}(0, \sigma^2 Id)$, and the OMP algorithm is input the sparsity K of x. Then under the assumptions that x satisfies the sparsity constraint*

$$K < \frac{1 + \mu_\Phi^{-1}}{2} - \frac{2\sqrt{\sigma^2 \log M} \cdot \mu_\Phi^{-1}}{x_{\min}}, \tag{9.12}$$

the OMP solution obtained by terminating the algorithm after K iterations satisfies $support(\widehat{x}_{\mathrm{OMP}}) = support(x)$ and

$$\|x - \widehat{x}_{\mathrm{OMP}}\| \leq 4\sqrt{\sigma^2 K \log M} \tag{9.13}$$

with probability exceeding $1 - \frac{1}{M\sqrt{2\pi \log M}}$. Here, $x_{\min}$ again denotes the smallest (in magnitude) nonzero entry of x.

9.2.3 Remarks

Recovery and estimation of sparse signals from a small number of linear measurements $y = \Phi x + n$ is an area of immense interest to a number of communities such as signal processing, statistics, and harmonic analysis. In this context, numerous reconstruction algorithms based upon either optimization techniques or greedy methods have been proposed in the literature. Our focus in this section has primarily been

[4] We point out here that if one is willing to tolerate some bias in the estimate, then the estimation error can be made smaller than $O(\sqrt{\sigma^2 K})$; see, e.g., [18, 31].

on two of the most well-known methods in this regard, namely, BP (and BPIC and LASSO) and OMP. Nevertheless, it is important for the reader to realize that there exist other methods in the literature, such as the Dantzig selector [18], CoSaMP [55], subspace pursuit [25], and iterative hard thresholding (IHT) [7], that can also be used for recovery and estimation of sparse signals. These methods primarily differ from each other in terms of computational complexity and explicit constants, but offer error guarantees that appear very similar to the ones in Theorems 9.4–9.9.

We conclude this section by noting that our focus here has been on providing uniform guarantees for sparse signals and relating those guarantees to the worst-case coherence of frames. The most important lesson of the preceding results in this regard is that there exist many computationally feasible algorithms that enable recovery/estimation of arbitrary K-sparse signals as long as $K = O(\mu_\Phi^{-1})$. There are two important aspects of this lesson. First, frames with small worst-case coherence are particularly well suited for making observations of sparse signals. Second, even Grassmannian frames cannot be guaranteed to work well if $K = O(N^{1/2+\delta})$ for $\delta > 0$, which follows trivially from the Welch bound. This second observation seems overly restrictive, and there exists literature based upon other properties of frames that attempts to break this "square-root" bottleneck. One such property, which has found widespread use in the compressed sensing literature, is termed the *restricted isometry property* (RIP) [14].

Definition 9.2 (Restricted Isometry Property) A unit norm frame $\Phi = (\varphi_i)_{i=1}^M$ in $\mathscr{H}^N$ is said to have the RIP of order K with parameter $\delta_K \in (0, 1)$ if for every K-sparse x, the following inequalities hold:

$$(1 - \delta_K)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_K)\|x\|_2^2. \tag{9.14}$$

The RIP of order K is essentially a statement concerning the minimum and maximum singular values of *all* $N \times K$ submatrices of Φ. However, even though the RIP has been used to provide guarantees for numerous sparse recovery/estimation algorithms such as BP, BPDN, CoSaMP, and IHT, explicit verification of this property for arbitrary frames appears to be computationally intractable. In particular, the only frames that are known to break the square-root bottleneck (using the RIP) for uniform guarantees are random (Gaussian, random binary, randomly subsampled partial Fourier, etc.) frames.[5] Still, it is possible to verify the RIP indirectly through the use of the Geršgorin circle theorem [5, 44, 71]. Doing so, however, yields results that match the ones reported above in terms of the sparsity constraint: $K = O(\mu_\Phi^{-1})$.

[5] Recently Bourgain et al. in [10] have reported a deterministic construction of frames that satisfies the RIP of $K = O(N^{1/2+\delta})$. However, the constant δ in there is so small that the scaling can be considered $K = O(N^{1/2})$ for all practical purposes.

9.3 Beyond Uniform Guarantees: Typical Behavior

The square-root bottleneck in sparse recovery/estimation problems is hard to overcome in part because of our insistence that the results hold uniformly for all K-sparse signals. In this section, we take a departure from uniform guarantees and instead focus on the *typical* behavior of various methods. In particular, we demonstrate in the following that the square-root bottleneck can be shattered by (i) imposing a statistical prior on the support and/or the nonzero entries of sparse signals and (ii) considering additional geometric measures of frames in conjunction with the worst-case coherence. In the following, we will focus on recovery, estimation, regression, and support detection of sparse signals using a multitude of methods. In all of these cases, we will assume that the support $\mathscr{K} \subset \{1, \ldots, M\}$ of x is drawn uniformly at random from all $\binom{M}{K}$ size-K subsets of $\{1, \ldots, M\}$. In some sense, this is the simplest statistical prior one can put on the support of x; in words, this assumption simply states that all supports of size K are equally likely.

9.3.1 Typical Recovery of Sparse Signals

In this section, we focus on typical recovery of sparse signals and provide guarantees for both ℓ_0 and ℓ_1 minimization (cf. (P_0) and (P_1)). The statistical prior we impose on the nonzero entries of sparse signals for this purpose however will differ for the two optimization schemes. We begin by providing a result for typical recovery of sparse signals using (P_0). The following theorem is due to Tropp and follows from combining results of [70] and [71].

Theorem 9.10 [70, 71] *Suppose that $y = \Phi x$ for a K-sparse signal x whose support is drawn uniformly at random and whose nonzero entries have a jointly continuous distribution. Further, let the frame Φ be such that $\mu_\Phi \leq (c_1 \log M)^{-1}$ for numerical constant $c_1 = 240$. Then under the assumption that x satisfies the sparsity constraint*

$$K < \min\left\{\frac{\mu_\Phi^{-2}}{\sqrt{2}}, \frac{M}{c_2^2 \|\Phi\|^2 \log M}\right\}, \tag{9.15}$$

the solution of (P_0) satisfies $\widehat{x}_0 = x$ with probability exceeding $1 - M^{-2\log 2}$. Here, $c_2 = 148$ is another numerical constant.

In order to understand the significance of Theorem 9.10, let us focus on the case of an approximately tight frame Φ: $\|\Phi\|^2 \approx \Theta(\frac{M}{N})$. In this case, ignoring the logarithmic factor, we have from (9.15) that ℓ_0 minimization can recover a K-sparse signal with high probability as long as $K = O(\mu_\Phi^{-2})$. This is in stark contrast to Theorem 9.3, which only allows $K = O(\mu_\Phi^{-1})$; in particular, Theorem 9.10 implies recovery of "most" K-sparse signals with $K = O(N/\log M)$ using frames such as

Gabor frames. In essence, shifting our focus from uniform guarantees to typical guarantees allows us to break the square-root bottleneck for arbitrary frames.

Even though Theorem 9.10 allows us to obtain near-optimal sparse recovery results, it is still a statement about the computationally infeasible ℓ_0 optimization. We now shift our focus to the computationally tractable BP optimization and present guarantees concerning its typical behavior. Before proceeding further, we point out that typicality in the case of ℓ_0 minimization is defined by a uniformly random support and a continuous distribution of the nonzero entries. In contrast, typicality in the case of BP will be defined in the following by a uniformly random support *but* nonzero entries whose phases are independent and uniformly distributed on the unit circle $\mathscr{C} = \{w \in \mathbb{C} : |w| = 1\}$.[6] The following theorem is once again due to Tropp and follows from combining results of [70] and [71].

Theorem 9.11 [70, 71] *Suppose that $y = \Phi x$ for a K-sparse signal x whose support is drawn uniformly at random and whose nonzero entries have independent phases distributed uniformly on $\mathscr{C}$. Further, let the frame Φ be such that $\mu_\Phi \leq (c_1 \log M)^{-1}$. Then under the assumption that x satisfies the sparsity constraint*

$$K < \min\left\{ \frac{\mu_\Phi^{-2}}{16 \log M}, \frac{M}{c_2^2 \|\Phi\|^2 \log M} \right\}, \tag{9.16}$$

the solution of BP satisfies $\widehat{x}_1 = x$ with probability exceeding $1 - M^{-2\log 2} - M^{-1}$. Here, c_1 and c_2 are the same numerical constants specified in Theorem 9.10.

It is worth pointing out that there exists another variant of Theorem 9.11 that involves sparse signals whose nonzero entries are independently distributed with zero median. Theorem 9.11 once again provides us with a powerful typical behavior result. Given approximately tight frames, it is possible to recover with high probability K-sparse signals using BP as long as $K = O(\mu_\Phi^{-2}/\log M)$. It is interesting to note here that, unlike Sect. 9.2, which dictates the use of Grassmannian frames for best uniform guarantees, both Theorem 9.10 and Theorem 9.11 dictate the use of Grassmannian frames that are also approximately tight for best typical guarantees. Heuristically speaking, insisting on tightness of frames is what allows us to break the square-root bottleneck in the typical case.

9.3.2 Typical Regression of Sparse Signals

Instead of shifting the discussion to typical sparse estimation, we now focus on another important problem in the statistics literature, namely, sparse linear regression

[6] Recall the definition of the phase of a number $r \in \mathbb{C}$: $\operatorname{sgn}(r) = \frac{r}{|r|}$.

[32, 38, 66]. We will return to the problem of sparse estimation in Sect. 9.3.4. Given $y = \Phi x + n$ for a K-sparse vector $x \in \mathbb{R}^M$, the goal in sparse regression is to obtain an estimate $\widehat{x}$ of x such that the *regression error* $\|\Phi x - \Phi\widehat{x}\|_2$ is small. It is important to note that the only nontrivial result that can be provided for sparse linear regression is in the presence of noise, since the regression error in the absence of noise is always zero. Our focus in this section will be once again on the AWGN n with variance σ^2, and we will restrict ourselves to the LASSO solution (cf. ($P_{1,2}$)). The following theorem provides guarantees for the typical behavior of the LASSO as reported in a recent work of Candès and Plan [15].

Theorem 9.12 [15] *Suppose that $y = \Phi x + n$ for a K-sparse signal $x \in \mathbb{R}^M$ whose support is drawn uniformly at random and whose nonzero entries are jointly independent with zero median. Further, let the noise n be distributed as $\mathcal{N}(0, \sigma^2 Id)$, let the frame Φ be such that $\mu_\Phi \leq (c_3 \log M)^{-1}$, and let x satisfy the sparsity constraint $K \leq \frac{M}{c_4 \|\Phi\|^2 \log M}$ for some positive numerical constants c_3 and c_4. Then the solution $\widehat{x}_{1,2}$ of the LASSO computed with $\tau = 2\sqrt{2\sigma^2 \log M}$ satisfies*

$$\|\Phi x - \Phi\widehat{x}\|_2 \leq c_5 \sqrt{2\sigma^2 K \log M} \tag{9.17}$$

with probability at least $1 - 6M^{-2\log 2} - M^{-1}(2\pi \log M)^{-1/2}$. Here, the constant c_5 may be taken as $8(1+\sqrt{2})^2$.

There are two important things to note about Theorem 9.12. First, it states that the regression error of the LASSO is $O(\sqrt{\sigma^2 K \log M})$ with very high probability. This regression error is in fact very close to the near-ideal regression error of $O(\sqrt{\sigma^2 K})$. Second, the performance guarantees of Theorem 9.12 are a strong function of $\|\Phi\|$ but only a weak function of the worst-case coherence μ_Φ. In particular, Theorem 9.12 dictates that the sparsity level accommodated by the LASSO is primarily a function of $\|\Phi\|$, provided μ_Φ is not too large. If, for example, Φ was an approximately tight frame, then the LASSO can handle $K \approx O(N/\log M)$ regardless of the value of μ_Φ, provided $\mu_\Phi = O(1/\log M)$. In essence, the above theorem signifies the use of approximately tight frames with small-enough coherence in regression problems. We conclude this subsection by noting that some of the techniques used in [15] to prove this theorem can in fact be used to also relax the dependence of BP on μ_Φ and obtain BP guarantees that primarily require small $\|\Phi\|$.

9.3.3 Typical Support Detection of Sparse Signals

It is often the case in many signal processing and statistics applications that one is interested in obtaining locations of the nonzero entries of a sparse signal x from a small number of measurements. This problem of *support detection* or *model selection* is of course trivial in the noiseless setting; exact recovery of sparse signals

in this case implies exact recovery of the signal support: $support(\widehat{x}) = support(x)$. Given $y = \Phi x + n$ with nonzero noise n, however, the support detection problem becomes nontrivial. This happens because a small estimation error in this case does not necessarily imply a small support detection error. Both exact support detection ($support(\widehat{x}) = support(x)$) and partial support detection ($support(\widehat{x}) \subset support(x)$) in the case of deterministic noise are very challenging (perhaps impossible) tasks. In the case of stochastic noise, however, both these problems become feasible, and we alluded to them in Theorem 9.8 and Theorem 9.9 in the context of uniform guarantees. In this subsection, we now focus on typical support detection in order to overcome the square-root bottleneck.

9.3.3.1 Support detection using the LASSO

The LASSO is arguably one of the standard tools used for support detection by the statistics and signal processing communities. Over the years, a number of theoretical guarantees have been provided for the LASSO support detection in [53, 73, 79]. The results reported in [53, 79] established that the LASSO asymptotically identifies the correct support under certain conditions on the frame Φ and the sparse signal x. Later, Wainwright in [73] strengthened the results of [53, 79] and made explicit the dependence of exact support detection using the LASSO on the smallest (in magnitude) nonzero entry of x. However, apart from the fact that the results reported in [53, 73, 79] are only asymptotic in nature, the main limitation of these works is that explicit verification of the conditions (such as the *irrepresentable condition* of [79] and the *incoherence condition* of [73]) that an arbitrary frame Φ needs to satisfy is computationally intractable for $K = \Omega(\mu_\Phi^{-1-\delta})$, $\delta > 0$.

The support detection results reported in [53, 73, 79] suffer from the square-root bottleneck because of their focus on uniform guarantees. Recently, Candès and Plan reported typical support detection results for the LASSO that overcome the square-root bottleneck of the prior work in the case of exact support detection [15].

Theorem 9.13 [15] *Suppose that $y = \Phi x + n$ for a K-sparse signal $x \in \mathbb{R}^M$ whose support is drawn uniformly at random and whose nonzero entries are jointly independent with zero median. Further, let the noise n be distributed as $\mathcal{N}(0, \sigma^2 Id)$, let the frame Φ be such that $\mu_\Phi \leq (c_6 \log M)^{-1}$, and let x satisfy the sparsity constraint $K \leq \frac{M}{c_7 \|\Phi\|^2 \log M}$ for some positive numerical constants c_6 and c_7. Finally, let $\mathcal{K}$ be the support of x and suppose that*

$$\min_{i \in \mathcal{K}} |x_i| > 8\sqrt{2\sigma^2 \log M}. \tag{9.18}$$

Then the solution $\widehat{x}_{1,2}$ of LASSO computed with $\tau = 2\sqrt{2\sigma^2 \log M}$ satisfies

$$support(\widehat{x}_{1,2}) = support(x) \quad \text{and} \quad \operatorname{sgn}(\widehat{x}_{\mathcal{K}}) = \operatorname{sgn}(x_{\mathcal{K}}) \tag{9.19}$$

with probability at least $1 - 2M^{-1}((2\pi \log M)^{-1/2} + KM^{-1}) - O(M^{-2\log 2})$.

Algorithm 2 The One-Step Thresholding (OST) Algorithm for Support Detection

Input: Unit norm frame Φ, measurement vector y, and a threshold $\lambda > 0$
Output: Estimate of signal support $\widehat{\mathscr{K}} \subset \{1, \ldots, M\}$

$z \leftarrow \Phi^* y$ {Form signal proxy}
$\widehat{\mathscr{K}} \leftarrow \{i \in \{1, \ldots, M\} : |z_i| > \lambda\}$ {Select indices via OST}

This theorem states that if the nonzero entries of the sparse signal x are significant in the sense that they roughly lie (modulo the logarithmic factor) above the noise floor σ, then the LASSO successfully carries out exact support detection for sufficiently sparse signals. Of course if any nonzero entry of the signal lies below the noise floor, then it is impossible to tell that entry apart from the noise itself. Theorem 9.13 is nearly optimal for exact model selection in this regard. In terms of the sparsity constraints, the statement of this theorem matches that of Theorem 9.12. Therefore, we once again see that frames that are approximately tight and have worst-case coherence that is not too large are particularly well suited for sparse signal processing when used in conjunction with the LASSO.

9.3.3.2 Support detection using one-step thresholding

Although the support detection results reported in Theorem 9.13 are near-optimal, it is desirable to investigate alternative solutions to the problem of typical support detection, because:

1. The LASSO requires the minimum singular value of the subframe of Φ corresponding to the support $\mathscr{K}$ to be bounded away from zero [15, 53, 73, 79]. While this is a plausible condition for the case when one is interested in estimating x, it is arguable whether this condition is necessary for the case of support detection.
2. Theorem 9.13 still lacks guarantees for $K = \Omega(\mu_\Phi^{-1-\delta}), \delta > 0$ in the case of deterministic nonzero entries of x.
3. The computational complexity of the LASSO for arbitrary frames tends to be $O(M^3 + NM^2)$. This makes the LASSO computationally demanding for large-scale model-selection problems.

In light of these concerns, a few researchers recently revisited the much older (and oft-forgotten) method of thresholding for support detection [2, 3, 37, 39, 57, 61]. The *one-step thresholding* (OST) algorithm, described in Algorithm 2, has a computational complexity of only $O(NM)$ and it has been known to be nearly optimal for $M \times M$ orthonormal bases [31]. In this subsection, we focus on a recent result of Bajwa et al. [2, 3] concerning typical support detection using OST. The forthcoming theorem in this regard relies on a notion of the *coherence property*, defined below.

Definition 9.3 (The Coherence Property [2, 3]) We say that a unit norm frame Φ satisfies the coherence property if

$$\text{(CP-1)} \quad \mu_\Phi \leq \frac{0.1}{\sqrt{2\log M}} \quad \text{and} \quad \text{(CP-2)} \quad \nu_\Phi \leq \frac{\mu_\Phi}{\sqrt{N}}.$$

In words, (CP-1) roughly states that the frame elements of Φ are not too similar, while (CP-2) roughly states that the frame elements of a unit norm Φ are somewhat distributed within the N-dimensional unit ball. Note that the coherence property (i) does not require the singular values of the submatrices of Φ to be bounded away from zero, and (ii) can be verified in polynomial time since it simply requires checking $\|G_\Phi - Id\|_{\max} \leq (200\log M)^{-1/2}$ and $\|(G_\Phi - Id)1\|_\infty \leq \|G_\Phi - Id\|_{\max}(M-1)N^{-1/2}$.

The implications of the coherence property are described in the following theorem. Before proceeding further, however, we first define some notation. We use $\mathsf{SNR} \doteq \|x\|^2/\mathbb{E}[\|n\|^2]$ to denote the *signal-to-noise ratio* associated with the support detection problem. Also, we use $x_{(\ell)}$ to denote the ℓ-th largest (in magnitude) nonzero entry of x. We are now ready to state the typical support detection performance of the OST algorithm.

Theorem 9.14 [3] *Suppose that $y = \Phi x + n$ for a K-sparse signal $x \in \mathbb{C}^M$ whose support $\mathscr{K}$ is drawn uniformly at random. Further, let $M \geq 128$, let the noise n be distributed as complex Gaussian with mean 0 and covariance $\sigma^2 Id$, $n \sim \mathscr{CN}(0, \sigma^2 Id)$, and let the frame Φ satisfy the coherence property. Finally, fix a parameter $t \in (0,1)$ and choose the threshold*

$$\lambda = \max\left\{\frac{1}{t}10\mu_\Phi\sqrt{N\cdot\mathsf{SNR}}, \frac{1}{1-t}\sqrt{2}\right\}\sqrt{2\sigma^2\log M}.$$

Then, under the assumption that $K \leq N/(2\log M)$, the OST algorithm (Algorithm 2) guarantees with probability exceeding $1 - 6M^{-1}$ that $\widehat{\mathscr{K}} \subset \mathscr{K}$ and $|\mathscr{K} \setminus \widehat{\mathscr{K}}| \leq (K - L)$, where L is the largest integer for which the following inequality holds:

$$x_{(L)} > \max\left\{c_8\sigma, c_9\mu_\Phi\|x\|\right\}\sqrt{\log M}. \tag{9.20}$$

Here, $c_8 \doteq 4(1-t)^{-1}$, $c_9 \doteq 20\sqrt{2}\,t^{-1}$, and the probability of failure is with respect to the true model $\mathscr{K}$ and the Gaussian noise n.

In order to put the significance of Theorem 9.14 into perspective, we recall the thresholding results obtained by Donoho and Johnstone [31]—which form the basis of ideas such as wavelet denoising—for the case of $M \times M$ orthonormal bases. It was established in [31] that if Φ is an orthonormal basis, then hard thresholding the entries of $\Phi^* y$ at $\lambda = \Theta(\sqrt{\sigma^2\log M})$ results in oracle-like performance in the sense that one recovers (with high probability) the locations of all the nonzero entries of x that are above the noise floor (modulo $\log M$).

Now the first thing to note regarding Theorem 9.14 is the intuitively pleasing nature of the proposed threshold. Specifically, assume that Φ is an orthonormal basis and notice that, since $\mu_\Phi = 0$, the threshold $\lambda = \Theta(\max\{\mu_\Phi\sqrt{N \cdot \mathsf{SNR}}, 1\} \times \sqrt{\sigma^2 \log M})$ proposed in the theorem reduces to the threshold proposed in [31] *and* Theorem 9.14 guarantees that thresholding recovers (with high probability) the locations of all the nonzero entries of x that are above the noise floor. The reader can become convinced of this assertion by noting that $x_{(\ell)} = \Omega(\sqrt{\sigma^2 \log M}) \Rightarrow \ell \in \widehat{\mathscr{K}}$ in the case of orthonormal bases. Now consider instead frames that are not necessarily orthonormal but which satisfy $\mu_\Phi = O(N^{-1/2})$ and $\nu_\Phi = O(N^{-1})$. Then we have from the theorem that OST identifies (with high probability) the locations of the nonzero entries of x whose energies are greater than both the noise variance (modulo $\log M$) and the average energy per nonzero entry: $x_{(\ell)}^2 = \Omega(\max\{\sigma^2 \log M, \|x\|^2/K\}) \Rightarrow \ell \in \widehat{\mathscr{K}}$. It is then easy to see in this case that if either the noise floor is high enough or the nonzero entries of x are roughly of the same magnitude then the simple OST algorithm leads to recovery of the locations of all the nonzero entries that are above the noise floor. Stated differently, the OST in certain cases has the oracle property in the sense of Donoho and Johnstone [31] *without* requiring the frame Φ to be an orthonormal basis.

9.3.4 Typical Estimation of Sparse Signals

Our goal in this section is to provide typical guarantees for the reconstruction of sparse signals from noisy measurements $y = \Phi x + n$, where the entries of the noise vector $n \in \mathbb{C}^N$ are independent, identical complex Gaussian random variables with mean zero and variance σ^2. The reconstruction algorithm we analyze here is an extension of the OST algorithm described earlier for support detection. This OST algorithm for reconstruction is described in Algorithm 3, and has been recently analyzed in [4]. The following theorem is due to Bajwa et al. [4] and shows that the OST algorithm leads to a near-optimal reconstruction error for certain important classes of sparse signals.

Before a formal statement of the theorem, however, we need to define some more notation. We use $\mathscr{T}_\sigma(t) := \{i : |x_i| > \frac{2\sqrt{2}}{1-t}\sqrt{2\sigma^2 \log M}\}$ for any $t \in (0, 1)$ to denote the locations of all the entries of x that, roughly speaking, lie above the *noise floor* σ.

Algorithm 3 One-Step Thresholding (OST) for Sparse Signal Reconstruction

Input: Unit norm frame Φ, measurement vector y, and a threshold $\lambda > 0$
Output: Sparse OST estimate $\widehat{x}^{\mathrm{OST}}$

$\widehat{x}^{\mathrm{OST}} \leftarrow 0$	{Initialize}
$z \leftarrow \Phi^* y$	{Form signal proxy}
$\widehat{\mathscr{K}} \leftarrow \{i : \lvert z_i \rvert > \lambda\}$	{Select indices via OST}
$\widehat{x}^{\mathrm{OST}}_{\widehat{\mathscr{K}}} \leftarrow (\Phi_{\widehat{\mathscr{K}}})^\dagger y$	{Reconstruct signal via least-squares}

Also, we use $\mathscr{T}_\mu(t) := \{i : |x_i| > \frac{20}{t}\mu_\Phi \|x\|\sqrt{2\log M}\}$ to denote the locations of entries of x that, roughly speaking, lie above the *self-interference floor* $\mu_\Phi\|x\|$. Finally, we also need a stronger version of the coherence property for reconstruction guarantees.

Definition 9.4 (The Strong Coherence Property [3]) We say a unit norm frame Φ satisfies the *strong coherence property* if

$$\text{(SCP-1)} \quad \mu_\Phi \le \frac{1}{164\log M} \quad \text{and} \quad \text{(SCP-2)} \quad \nu_\Phi \le \frac{\mu_\Phi}{\sqrt{N}}.$$

Theorem 9.15 [4] *Take a unit norm frame Φ which satisfies the strong coherence property, pick $t \in (0,1)$, and choose $\lambda = \sqrt{2\sigma^2\log M}\ \max\{\frac{10}{t}\mu_\Phi\sqrt{N\ \mathsf{SNR}}, \frac{\sqrt{2}}{1-t}\}$. Further, suppose $x \in \mathbb{C}^M$ has support $\mathscr{K}$ drawn uniformly at random from all possible K-subsets of $\{1,\dots,M\}$. Then provided*

$$K \le \frac{M}{c_{10}^2\|\Phi\|^2\log M}, \tag{9.21}$$

Algorithm 3 produces $\widehat{\mathscr{K}}$ such that $\mathscr{T}_\sigma(t)\cap\mathscr{T}_\mu(t) \subseteq \widehat{\mathscr{K}} \subseteq \mathscr{K}$ and $\widehat{x}^{\mathrm{OST}}$ such that

$$\left\|x - \widehat{x}^{\mathrm{OST}}\right\| \le c_{11}\sqrt{\sigma^2|\widehat{\mathscr{K}}|\log M} + c_{12}\|x_{\mathscr{K}\setminus\widehat{\mathscr{K}}}\| \tag{9.22}$$

with probability exceeding $1 - 10M^{-1}$. Finally, defining $T := |\mathscr{T}_\sigma(t)\cap\mathscr{T}_\mu(t)|$, we further have

$$\|x - \widehat{x}\| \le c_{11}\sqrt{\sigma^2 K\log M} + c_{12}\|x - x_T\| \tag{9.23}$$

in the same probability event. Here, $c_{10} = 37\mathrm{e}$, $c_{11} = \frac{2}{1-\mathrm{e}^{-1/2}}$, and $c_{12} = 1 + \frac{\mathrm{e}^{-1/2}}{1-\mathrm{e}^{-1/2}}$ are numerical constants.

A few remarks are in order now for Theorem 9.15. First, if Φ satisfies the strong coherence property *and* Φ is nearly tight, then OST handles sparsity that is almost linear in N: $K = O(N/\log M)$ from (9.21). Second, the ℓ_2 error associated with the OST algorithm is the near-optimal (modulo the log factor) error of $\sqrt{\sigma^2 K\log M}$ *plus* the best T-term approximation error caused by the inability of the OST algorithm to recover signal entries that are smaller than $O(\mu_\Phi\|x\|\sqrt{2\log M})$. In particular, if the K-sparse signal x, the worst-case coherence μ_Φ, and the noise n together satisfy $\|x - x_T\| = O(\sqrt{\sigma^2 K\log M})$, then the OST algorithm succeeds with a near-optimal ℓ_2 error of $\|x - \widehat{x}\| = O(\sqrt{\sigma^2 K\log M})$. To see why this error is near-optimal, note that a K-dimensional vector of random entries with mean zero and variance σ^2 has expected squared norm $\sigma^2 K$; in here, the OST pays an additional log factor to find the locations of the K nonzero entries among the entire M-dimensional signal. It is important to recognize that the optimality condition $\|x - x_T\| = O(\sqrt{\sigma^2 K\log M})$ depends on the signal class, the noise variance, and

the worst-case coherence of the frame; in particular, the condition is satisfied whenever $\|x_{\mathscr{K}\setminus\mathscr{T}_\mu(t)}\| = O(\sqrt{\sigma^2 K \log M})$, since

$$\|x - x_T\| \le \|x_{\mathscr{K}\setminus\mathscr{T}_\sigma(t)}\| + \|x_{\mathscr{K}\setminus\mathscr{T}_\mu(t)}\| = O\left(\sqrt{\sigma^2 K \log M}\right) + \|x_{\mathscr{K}\setminus\mathscr{T}_\mu(t)}\|. \tag{9.24}$$

We conclude this subsection by stating a lemma from [4] that provides classes of sparse signals which satisfy $\|x_{\mathscr{K}\setminus\mathscr{T}_\mu(t)}\| = O(\sqrt{\sigma^2 K \log M})$ given sufficiently small noise variance and worst-case coherence.

Lemma 9.3 *Take a unit norm frame Φ with worst-case coherence $\mu_\Phi \le \frac{c_{13}}{\sqrt{N}}$ for some $c_{13} > 0$, and suppose that $K \le \frac{M}{c_{14}^2 \|\Phi\|^2 \log M}$ for some $c_{14} > 0$. Fix a constant $\beta \in (0, 1]$, and suppose the magnitudes of βK nonzero entries of x are some $\alpha = \Omega(\sqrt{\sigma^2 \log M})$, while the magnitudes of the remaining $(1-\beta)K$ nonzero entries are not necessarily the same, but are smaller than α and scale as $O(\sqrt{\sigma^2 \log M})$. Then $\|x_{\mathscr{K}\setminus\mathscr{T}_\mu(t)}\| = O(\sqrt{\sigma^2 K \log M})$, provided $c_{13} \le \frac{t c_{14}}{20\sqrt{2}}$.*

In words, Lemma 9.3 states that OST is near-optimal for those K-sparse signals whose entries above the noise floor have roughly the same magnitude. This subsumes a very important class of signals that appears in applications such as multi-label prediction [47], in which all the nonzero entries take values $\pm\alpha$.

9.4 Finite Frames for Detecting the Presence of Sparse Signals

In the previous sections, we discussed the role of frame theory in recovering and estimating sparse signals in different settings. We now consider a different problem: detecting the presence of a sparse signal in noise. In the simplest form, the problem is to decide whether an observed data vector is a realization from a hypothesized noise-only model or from a hypothesized signal-plus-noise model, where in the latter model the signal is sparse but the indices and the values of its nonzero elements are unknown. The problem is a binary hypothesis test of the form

$$\begin{cases} \mathscr{H}_0 : y = \Phi n, \\ \mathscr{H}_1 : y = \Phi(x + n), \end{cases} \tag{9.25}$$

where $x \in \mathbb{R}^M$ is a deterministic but unknown K-sparse signal, the measurement matrix $\Phi = \{\varphi_i\}_{i=1}^M$ is a frame for $\mathbb{R}^N$, $N \le M$, which we get to design, and $n \in \mathbb{R}^M$ is a white Gaussian noise vector with covariance matrix $\mathbb{E}[nn^T] = (\sigma_n^2/M)Id$.

We assume here that the number of measurements N allowed for detection is fixed and prespecified. We wish to decide whether the measurement vector $y \in \mathbb{R}^N$ belongs to model $\mathscr{H}_0$ or $\mathscr{H}_1$. This problem is fundamentally different from that of estimating a sparse signal, as the objective in detection typically is to maximize the probability of detection, while maintaining a low false alarm rate, or to minimize the total error probability or a Bayes risk, rather than to find the sparsest signal that fits

a linear observation model. Unlike the signal estimation problem, the detection of sparse signals has received very little attention so far, with notable exceptions being [45, 56, 74]. But in particular, the design of optimal or near-optimal compressive measurement matrices for detection of sparse signals has scarcely been addressed [76, 77]. In this section, we provide an overview of selected results by Zahedi et al. [76, 77], concerning the necessary and sufficient conditions for a frame Φ to optimize a measure of detection performance.

We look at the general problem of designing the measurement frame Φ to maximize the measurement SNR, under $\mathscr{H}_1$, which is given by

$$\mathsf{SNR} = \frac{\|\Phi x\|^2}{\sigma_n^2/M}. \tag{9.26}$$

This is motivated by the fact that for the class of linear log-likelihood ratio detectors, where the log-likelihood ratio is a linear function of the data, the detection performance is improved by increasing the SNR. In particular, for a Neyman–Pearson detector (see, e.g., [60]) with false alarm rate $P_F \leq \gamma$, the probability of detection

$$P_d = Q\big(Q^{-1}(\gamma) - \sqrt{\mathsf{SNR}}\big) \tag{9.27}$$

is monotonically increasing in SNR, where $Q(\cdot)$ is the Q-function, given by

$$Q(z) = \int_z^\infty e^{-w^2/2}\, dw. \tag{9.28}$$

In addition, maximizing SNR leads to maximum detection probability at a prespecified false alarm rate in an energy detector, which simply tests the energy of the measured vector y against a threshold. Without loss of generality, we assume that $\sigma_n^2 = 1$ and $\|x\|^2 = 1$, and we design Φ to maximize the measured signal energy $\|\Phi x\|^2$. To avoid coloring the noise vector n, that is, to keep the noise vector white, we constrain the measurement frame Φ to be Parseval, or tight with frame bound equal to one. That is, we only consider frames for which the frame operator $S_\Phi = \Phi\Phi^T$ is identity. From here on we simply refer to these frames as tight frames, but it is understood that all tight frames we consider in this section are in fact Parseval.

In solving the problem, one approach is to assume a value for the sparsity level K and design the measurement frame Φ based on this assumption. This approach, however, runs the risk that the true sparsity level might be different. An alternative approach is not to assume any specific sparsity level. Instead, when designing Φ, we prioritize the level of importance of different values of sparsity. In other words, we first find a set of solutions that are optimal for a K_1-sparse signal. Then, within this set, we find a subset of solutions that are also optimal for K_2-sparse signals. We follow this procedure until we find a subset that contains a family of optimal solutions for sparsity levels $K_1, K_2, K_3, \ldots$. This approach is known as a *lexicographic optimization* method (see, e.g., [33, 43, 48]). The measurement frame design naturally depends on one's assumptions about the unknown vector x. In the following sections, we review two different design problems, namely a *worst-case SNR design* and an *average SNR design*, following the developments of [76, 77].

We note that lexicographic optimizations have been employed earlier in [46] in the design of frames that have maximal robustness to erasures of frame coefficients. The analysis used in deriving the main results for the worst-case SNR design is similar in nature to that used in [46].

9.4.1 Worst-Case SNR Design

In the worst-case design for a sparsity level K, we consider the vector x that minimizes the SNR among all K-sparse signals and design the frame Φ to maximize this minimum SNR. Of course, when minimizing the SNR with respect to x, we have to find the minimum SNR with respect to both the locations and the values of the nonzero entries in x. To combine this with the lexicographic approach, we design the matrix Φ to maximize the *worst-case* detection SNR, where the worst case is taken over all subsets of size K_i of elements of x, where K_i is the sparsity level considered at the ith level of lexicographic optimization. This is a design for robustness with respect to the worst sparse signal that can be produced.

Consider the Kth step of the lexicographic approach. In this step, the vector x is assumed to have up to K nonzero entries, and we assume $\|x\|^2 = 1$. But otherwise, we do not impose any constraints on the locations and the values of the nonzero entries of x. We wish to maximize the minimum (worst-case) SNR, produced by assigning the worst possible locations and values to the nonzero entries of the K-sparse vector x. Since we assume $\sigma_n^2 = 1$, this corresponds to a worst-case design for maximizing the signal energy $\|\Phi x\|^2$.

Let $\mathscr{B}_0$ be the set containing all $(N \times M)$ tight frames. We recursively define the set $\mathscr{B}_K$, $K = 1, 2, \ldots$, as the set of solutions to the following worst-case optimization problem [77]:

$$\begin{aligned} &\max_{\Phi} \min_{x} \|\Phi x\|^2, \\ &\text{s.t.} \quad \Phi \in \mathscr{B}_{K-1}, \\ &\qquad \|x\| = 1, \\ &\qquad x \text{ is } K\text{-sparse}. \end{aligned} \tag{9.29}$$

The optimization problem for the Kth stage (9.29) involves a worst-case objective restricted to the set of solutions $\mathscr{B}_{K-1}$ from the $(K-1)$th problem. So, $\mathscr{B}_K \subset \mathscr{B}_{K-1} \subset \cdots \subset \mathscr{B}_0$.

Now let $\Omega = \{1, 2, \ldots, M\}$, and define Ω_K to be $\Omega_K = \{\omega \subset \Omega : |\omega| = K\}$. For any $\mathscr{T} \in \Omega_K$, let $x_{\mathscr{T}}$ be the subvector of size $(K \times 1)$ that contains all the components of x corresponding to indices in $\mathscr{T}$. Similarly, given a frame Φ, let $\Phi_{\mathscr{T}}$ be the $(N \times K)$ submatrix consisting of all columns of Φ whose indices are in $\mathscr{T}$. Note that the vector $x_{\mathscr{T}}$ may have zero entries and hence is not necessarily the same as the support of x. Given $\mathscr{T} \in \Omega_K$, the product Φx can be replaced by $\Phi_{\mathscr{T}} x_{\mathscr{T}}$ instead. To consider the worst-case design, for any $\mathscr{T}$ we need to consider the $x_{\mathscr{T}}$ that minimizes $\|\Phi_{\mathscr{T}} x_{\mathscr{T}}\|^2$ and then also find the worst $\mathscr{T} \in \Omega_K$. Using this

notation and after some simple algebra, the worst-case problem (9.29) can be posed as the following max-min problem [77]:

$$(\mathscr{P}_K) \quad \begin{cases} \max_{\Phi} \min_{\mathscr{T}} \lambda_{\min}(\Phi_{\mathscr{T}}^T \Phi_{\mathscr{T}}), \\ \text{s.t.} \quad \Phi \in \mathscr{B}_{K-1}, \\ \qquad \mathscr{T} \in \Omega_K, \end{cases} \tag{9.30}$$

where $\lambda_{\min}(\Phi_{\mathscr{T}}^T \Phi_{\mathscr{T}})$ denotes the smallest eigenvalue of the frame sub-Gramian $G_{\Phi_{\mathscr{T}}} = \Phi_{\mathscr{T}}^T \Phi_{\mathscr{T}}$.

To solve the worst-case design problem, we first find the solution set $\mathscr{B}_1$ for problem $(\mathscr{P}_1)$. Then, we find a subset $\mathscr{B}_2 \subset \mathscr{B}_1$ as the solution for $(\mathscr{P}_2)$. We continue this procedure for general sparsity level K.

Sparsity level $K = 1$ If $K = 1$, then any $\mathscr{T}$ such that $|\mathscr{T}| = 1$ can be written as $\mathscr{T} = \{i\}$ with $i \in \Omega$, and $\Phi_{\mathscr{T}} = \varphi_i$ consists of only the ith column of Φ. Therefore, $\Phi_{\mathscr{T}}^T \Phi_{\mathscr{T}} = \|\varphi_i\|^2$, and $\mathscr{P}_1$ simplifies to

$$\begin{aligned} &\max_{\Phi} \min_{i} \|\varphi_i\|^2, \\ &\text{s.t.} \quad \Phi \in \mathscr{B}_0, \\ &\qquad i \in \Omega. \end{aligned} \tag{9.31}$$

We have the following result.

Theorem 9.16 [77] *The optimal value of the objective function of the max-min problem* (9.31) *is* N/M, *and a necessary and sufficient condition for* $\hat{\Phi} \in \mathscr{B}_0$ *to lie in the solution set* $\mathscr{B}_1$ *is for* $\hat{\Phi} = \{\hat{\varphi}_i\}_{i=1}^M$ *to be an equal norm tight frame with* $\|\hat{\varphi}_i\| = \sqrt{N/M}$, *for* $i = 1, 2, \ldots, M$.

Sparsity level $K = 2$ The next step is to solve $(\mathscr{P}_2)$. Given $\mathscr{T} \in \Omega_2$, the matrix $\Phi_{\mathscr{T}}$ consists of two columns, say, φ_i and φ_j. So, the matrix $\Phi_{\mathscr{T}}^T \Phi_{\mathscr{T}}$ in the max-min problem $(\mathscr{P}_2)$ is a (2×2) matrix:

$$\Phi_{\mathscr{T}}^T \Phi_{\mathscr{T}} = \begin{bmatrix} \langle \varphi_i, \varphi_i \rangle & \langle \varphi_i, \varphi_j \rangle \\ \langle \varphi_i, \varphi_j \rangle & \langle \varphi_j, \varphi_j \rangle \end{bmatrix}.$$

The solution for this case must lie among the family of optimal solutions for $K = 1$. In other words, the optimal solution $\hat{\Phi}$ must be an equal norm tight frame with $\|\hat{\varphi}_i\| = \sqrt{N/M}$, for $i = 1, 2, \ldots, M$. Therefore, we have

$$\Phi_{\mathscr{T}}^T \Phi_{\mathscr{T}} = (N/M) \begin{bmatrix} 1 & \cos\alpha_{ij} \\ \cos\alpha_{ij} & 1 \end{bmatrix},$$

where α_{ij} is the angle between vectors φ_i and φ_j. The minimum possible eigenvalue of this matrix is

$$\lambda_{\min}\left(\Phi_{\mathscr{T}}^T \Phi_{\mathscr{T}}\right) = (N/M)(1 - \mu_{\Phi}), \tag{9.32}$$

where μ_Φ is the worst-case coherence of the frame $\Phi = \{\varphi_i\}_{i=1}^M \in \mathscr{B}_1$, as defined in (9.1).

Now, let $\mu_{\min}$ be the minimum worst-case coherence

$$\mu_{\min} = \min_{\Phi \in \mathscr{B}_1} \mu_\Phi \tag{9.33}$$

for all frames in $\mathscr{B}_1$. We refer to the element of $\mathscr{B}_1$ that has the worst-case coherence $\mu_{\min}$ as a *Grassmannian equal norm tight frame*.

We have the following theorem.

Theorem 9.17 [77] *The optimal value of the objective function of the max-min problem* $(\mathscr{P}_2)$ *is* $(N/M)(1-\mu_{\min})$. *A frame* $\hat{\Phi}$ *is in* $\mathscr{B}_2$ *if and only if the columns of* $\hat{\Phi}$ *form an equal norm tight frame with norm values* $\sqrt{N/M}$ *and* $\mu_{\hat{\Phi}} = \mu_{\min}$. *In other words, the solution to* $(\mathscr{P}_2)$ *is an* $N \times M$ *Grassmannian equal norm tight frame.*

Sparsity level $K > 2$ We now consider the case where $K > 2$. In this case, $\mathscr{T} \in \Omega_K$ can be written as $\mathscr{T} = \{i_1, i_2, \ldots, i_K\} \subset \Omega$. From the previous results, we know that an optimal frame $\hat{\Phi} \in \mathscr{B}_K$ must be a Grassmannian equal norm tight frame, with norms $\sqrt{N/M}$ and worst-case coherence $\mu_{\min}$. Taking this into account, the $(K \times K)$ matrix $\hat{\Phi}_{\mathscr{T}}^T \hat{\Phi}_{\mathscr{T}}$ in $(\mathscr{P}_K)$, $K > 2$, can be written as $\hat{\Phi}_{\mathscr{T}}^T \hat{\Phi}_{\mathscr{T}} = (N/M)[Id + A_{\mathscr{T}}]$ where $A_{\mathscr{T}}$ is given by

$$A_{\mathscr{T}} = \begin{bmatrix} 0 & \cos\hat{\alpha}_{i_1 i_2} & \ldots & \cos\hat{\alpha}_{i_1 i_k} \\ \cos\hat{\alpha}_{i_1 i_2} & 0 & \ldots & \cos\hat{\alpha}_{i_2 i_k} \\ \vdots & \vdots & \ddots & \vdots \\ \cos\hat{\alpha}_{i_1 i_k} & \cos\hat{\alpha}_{i_2 i_k} & \ldots & 0 \end{bmatrix}, \tag{9.34}$$

and $\cos\hat{\alpha}_{i_h i_f}$ is the cosine of the angle between frame elements $\hat{\varphi}_{i_h}$ and $\hat{\varphi}_{i_f}$, $i_h \neq i_f \in \mathscr{T}$. It is easy to see that

$$\lambda_{\min}\big(\hat{\Phi}_{\mathscr{T}}^T \hat{\Phi}_{\mathscr{T}}\big) = (N/M)\big(1 + \lambda_{\min}(A_{\mathscr{T}})\big). \tag{9.35}$$

So, the problem $(\mathscr{P}_K)$, $K > 2$, simplifies to

$$(\mathscr{P}_K) \quad \begin{cases} \max_\Phi \min_{\mathscr{T}} \lambda_{\min}(A_{\mathscr{T}}), \\ \text{s.t.} \quad \Phi \in \mathscr{B}_{K-1}, \\ \qquad \mathscr{T} \in \Omega_K. \end{cases} \tag{9.36}$$

Solving the above problem however is not trivial. But we can at least bound the optimum value. Given $\mathscr{T} \in \Omega_K$, let $\hat{\delta}_{i_h i_f}$ and $\Delta_{\min}$ be

$$\hat{\delta}_{i_h i_f} = \mu_{\min} - |\cos\alpha_{i_h i_f}|, \quad i_h \neq i_f \in \mathscr{T}, \tag{9.37}$$

$$\Delta_{\min} = \min_{\mathscr{T} \in \Omega_K} \sum_{i_h \neq i_f \in \mathscr{T}} \hat{\delta}_{i_h i_f}. \tag{9.38}$$

Also, define $\hat{\Delta}$ in the following way:

$$\hat{\Delta} = \min_{\mathscr{T} \in \Omega_K} \sum_{i_h \neq i_f \in \mathscr{T}} \hat{\delta}_{i_h i_f}.$$

We have the following theorem.

Theorem 9.18 [77] *The optimal value of the objective function of the max-min problem* $(\mathscr{P}_K)$ *for* $K > 2$ *lies between* $(N/M)(1 - \binom{K}{2}\mu_{\min} + \Delta_{\min})$ *and* $(N/M) \times (1 - \mu_{\min})$.

Before we conclude the worst-case SNR design, a few remarks are in order.

1. Examples of uniform tight frames and their methods of construction can be found in [8, 13, 19, 20] and the references therein.
2. In the case where $K = 2$, $\hat{\Phi}_{\mathscr{T}}^T \hat{\Phi}_{\mathscr{T}}$ associated with the frame $\hat{\Phi}$ identified in Theorem 9.17 has the largest minimum eigenvalue $(N/M)(1 - \mu_{\min})$ and the smallest maximum eigenvalue $(N/M)(1 + \mu_{\min})$ among all $\Phi \in \mathscr{B}_1$ and $\mathscr{T} \in \Omega_2$. This means that the solution $\hat{\Phi}$ to $(\mathscr{P}_2)$ is an RIP matrix of order 2 with optimal RIC $\delta_2 = \mu_{\min}$.
3. In general, the minimum worst-case coherence $\mu_{\min}$ of the solution $\hat{\Phi}$ to $(\mathscr{P}_K)$, $K \geq 2$, is bounded below by the Welch bound (see Lemma 9.2). However, when $1 \leq N \leq M - 1$ and

$$M \leq \min\big\{N(N+1)/2, (M-N)(M-N+1)/2\big\}, \tag{9.39}$$

the Welch bound can be met [64]. For such a case, all frame angles are equal and the solution to $(\mathscr{P}_K)$ for $K \geq 2$ is an *equiangular equal norm tight frame*. Such frames are *Grassmannian line packings* (see, e.g., [8, 21, 24, 49, 52, 58, 63–65]).

9.4.2 Average-Case Design

Let us now assume that in (9.25) the locations of nonzero entries of x are random, but their values are deterministic and unknown. We wish to find the frame Φ that maximizes the expected value of the minimum SNR. The expectation is taken with respect to a random index set with uniform distribution over the set of all possible subsets of size K_i of the index set $\{1, 2, \ldots, M\}$ of elements of x. The minimum SNR, whose expected value we wish to maximize, is calculated with respect to the values of the entries of the vector x for each realization of the random index set.

Let $\mathscr{T}_K$ be a random variable that is uniformly distributed over Ω_K. Then $p_{\mathscr{T}_K}(t) = 1/\binom{M}{K}$ is the probability that $\mathscr{T}_K = t$ for $t \in \Omega_K$. Our goal is to find a measurement frame Φ that maximizes the expected value of the minimum SNR, where the expectation is taken with respect to the random $\mathscr{T}_K$, and the minimum is taken with respect to the entries of the vector x on $\mathscr{T}_K$. Taking into account

the simplifying steps used earlier for the worst-case problem and also adopting the lexicographic approach, the problem of maximizing the average SNR can then be formulated in the following way.

Let $\mathcal{N}_0$ be the set containing all $(N \times M)$ tight frames. Then for $K = 1, 2, \ldots,$ recursively define the set $\mathcal{N}_K$ as the solution set to the following optimization problem:

$$\begin{cases} \max_{\Phi} \mathbb{E}_{\mathcal{T}_K} \min_{x_K} \|\Phi_{\mathcal{T}_K} x_K\|^2, \\ \text{s.t.} \quad \Phi \in \mathcal{N}_{K-1}, \\ \qquad \|x_K\| = 1, \end{cases} \tag{9.40}$$

where $\mathbb{E}_{\mathcal{T}_K}$ is the expectation with respect to $\mathcal{T}_K$. As before, the $(N \times K)$ matrix $\Phi_{\mathcal{T}_K}$ is a submatrix of Φ whose column indices are in $\mathcal{T}_K$. This problem can be simplified to the following [77]:

$$(\mathcal{F}_K) \quad \begin{cases} \max_{\Phi} \mathbb{E}_{\mathcal{T}_K} \lambda_{\min}(\Phi_{\mathcal{T}_K}^T \Phi_{\mathcal{T}_K}), \\ \text{s.t.} \quad \Phi \in \mathcal{N}_{K-1}. \end{cases} \tag{9.41}$$

To solve the lexicographic problems $(\mathcal{F}_K)$, we follow the same method we used earlier for the worst-case problem; i.e., we begin by solving problem $(\mathcal{F}_1)$. Then, from the solution set $\mathcal{N}_1$, we find optimal solutions for the problem $(\mathcal{F}_2)$, and so on.

Sparsity level $K = 1$ Assume that the signal x is 1-sparse. So, there are $\binom{M}{1} = M$ different possibilities to build the matrix $\Phi_{\mathcal{T}_1}$ from the matrix Φ. The expectation in problem $(\mathcal{F}_1)$ can be written as:

$$\mathbb{E}_{\mathcal{T}_1} \lambda_{\min}\big(\Phi_{\mathcal{T}_1}^T \Phi_{\mathcal{T}_1}\big) = \sum_{t \in \Omega_1} p_{\mathcal{T}_1}(t) \lambda_{\min}\big(\Phi_t^T \Phi_t\big) = \sum_{i=1}^{M} p_{\mathcal{T}_1}\big(\{i\}\big) \|\varphi_i\|^2 = \frac{N}{M}. \tag{9.42}$$

The following result holds.

Theorem 9.19 [77] *The optimal value of the objective function of problem* $(\mathcal{F}_1)$ *is* N/M. *This value is obtained by using any* $\Phi \in \mathcal{N}_0$, *i.e., any tight frame.*

Theorem 9.19 shows that, unlike the worst-case problem, any tight frame is an optimal solution for the problem $(\mathcal{F}_1)$. Next, we study the case where the signal x is 2-sparse.

Sparsity level $K = 2$ For problem $(\mathcal{F}_2)$, the expected value term $\mathbb{E}_{\mathcal{T}_2} \lambda_{\min}(\Phi_{\mathcal{T}_2}^T \Phi_{\mathcal{T}_2})$ is equal to

$$\sum_{t \in \Omega_2} p_{\mathcal{T}_2}(t) \lambda_{\min}\big(\Phi_t^T \Phi_t\big) = \frac{2}{M(M-1)} \sum_{j=2}^{M} \sum_{i=1}^{j-1} \lambda_{\min}\big(\Phi_{\{i,j\}}^T \Phi_{\{i,j\}}\big). \tag{9.43}$$

In general, solving the family of problems $(\mathcal{F}_K)$, $K = 2, 3, \ldots,$ is not trivial. However, if we constrain ourselves to the class of equal norm tight frames, which

also arise in solving the worst-case problem, we can establish necessary and sufficient conditions for optimality. These conditions are different from those for the worst-case problem and, as we will show next, the optimal solution here is an equal norm tight frame for which a cumulative measure of coherence is minimal.

Let $\mathscr{M}_1$ be defined as $\mathscr{M}_1 = \{\Phi : \Phi \in \mathscr{N}_1, \|\varphi_i\| = \sqrt{N/M}, \forall i \in \Omega\}$. Also, for $K = 2, 3, \ldots$, recursively define the set $\mathscr{M}_K$ as the solution set to the following optimization problem:

$$\left(\mathscr{F}_K^{'}\right) \quad \begin{cases} \max_{\Phi} \mathbb{E}_{\mathscr{T}_K} \lambda_{\min}(\Phi_{\mathscr{T}_K}^T \Phi_{\mathscr{T}_K}), \\ \text{s.t.} \quad \Phi \in \mathscr{M}_{K-1}. \end{cases} \tag{9.44}$$

We will concentrate on solving the above family of problems instead of $(\mathscr{F}_K)$, $K = 2, 3, \ldots$. We have the following results.

Theorem 9.20 [77] *The frame Φ is in $\mathscr{M}_2$ if and only if the sum coherence of Φ, i.e., $\sum_{j=2}^{M} \sum_{i=1}^{j-1} |\langle \varphi_i, \varphi_j \rangle| / (\|\varphi_i\| \|\varphi_j\|)$, is minimized.*

Theorem 9.20 shows that for problem $(\mathscr{F}_2^{'})$, angles between elements of the equal norm tight frame Φ should be designed in a different way than for the worst-case problem. For example, an equiangular tight frame of $M = 2N$ in N dimensions, with vectors of equal norm $\sqrt{1/2}$, has worst-case coherence $1/(2\sqrt{2N-1})$ and sum coherence $N\sqrt{2N-1}/2$, while two copies of an orthonormal basis form a frame with worst-case coherence $1/2$ and sum coherence $N/2$. While it is not clear whether copies of orthonormal bases form tight frames with minimal sum coherence, this example certainly illustrates that Grassmannian frames do not, in general, result in minimal sum coherence. To the best of our knowledge, no general method for constructing tight frames with minimal sum coherence has been proposed so far.

The following lemma provides bounds on the sum coherence of an equal norm tight frame.

Lemma 9.4 [77] *For an equal norm tight frame Φ with norm values $\sqrt{N/M}$, the following inequalities hold*:

$$c\left|(M/N-1) - 2(M-1)\mu_{\Phi}^2\right| \leq \sum_{j=2}^{M} \sum_{i=1}^{j-1} \left|\langle \varphi_i, \varphi_j \rangle\right| \leq c(M-1)\mu_{\Phi}^2,$$

where

$$c = \left(\frac{(N/M)^2}{1-2(N/M)}\right)\left(\frac{M(M-2)}{2}\right).$$

Sparsity level $K > 2$ Similar to the worst-case problem, solving problems $(\mathscr{F}_K^{'})$ for $K > 2$ is not trivial—the solution sets for these problems all lie in $\mathscr{M}_2$, and $(\mathscr{F}_2^{'})$ is still an open problem. The following lemma provides a lower bound for the optimal objective function of $(\mathscr{F}_K^{'})$, $K > 2$.

Lemma 9.5 [77] *The optimal value of the objective function for problem* $(\mathscr{F}'_K)$, $K > 2$, *is bounded below by* $(N/M)(1 - (K(K-1)/2)\mu_\Phi)$.

We conclude this section by giving a summary. In the worst-case SNR problem, the optimal measurement matrix is a Grassmannian equal norm tight frame for most—and an equal norm tight frame for all—sparse signals. In the average SNR problem, we limited ourselves to the class of equal norm tight frames and showed that the optimal measurement frame is an equal norm tight frame that has minimum sum coherence.

9.5 Other Topics

As mentioned earlier, this chapter covers only a small subset of the results in the sparse signal processing literature. Our aim has been to simply highlight the central role that finite frames and their geometric measures, such as spectral norm, worst-case coherence, average coherence, and sum coherence, play in the development of sparse signal processing methods. But many developments, which also involve finite frames, have not been covered. For example, there is a large body of work on signal processing of *compressible* signals. These are signals that are not sparse, but whose entries decay in magnitude according to a particular power law. Many of the results covered in this chapter on estimating sparse signals have counterparts for compressible signals. The reader is referred to [17, 23, 26, 27] for examples of such results. Another example is the estimation and recovery of *block-sparse* signals, where the nonzero entries of the signal to be estimated are either clustered or the signal has a sparse representation in a fusion frame. Again, the majority of the results on the estimation and recovery of sparse signals can be extended to block-sparse signals. The reader is referred to [9, 35, 62, 78] and the references therein.

References

1. IEEE Signal Processing Magazine, special issue on compressive sampling (2008)
2. Bajwa, W.U., Calderbank, R., Jafarpour, S.: Model selection: two fundamental measures of coherence and their algorithmic significance. In: Proc. IEEE Intl. Symp. Information Theory (ISIT'10), Austin, TX, pp. 1568–1572 (2010)
3. Bajwa, W.U., Calderbank, R., Jafarpour, S.: Why Gabor frames? Two fundamental measures of coherence and their role in model selection. J. Commun. Netw. **12**(4), 289–307 (2010)
4. Bajwa, W.U., Calderbank, R., Mixon, D.G.: Two are better than one: fundamental parameters of frame coherence. Appl. Comput. Harmon. Anal. **33**(1), 58–78 (2012)
5. Bajwa, W.U., Haupt, J., Raz, G., Nowak, R.: Compressed channel sensing. In: Proc. 42nd Annu. Conf. Information Sciences and Systems (CISS'08), Princeton, NJ, pp. 5–10 (2008)
6. Ben-Haim, Z., Eldar, Y.C., Elad, M.: Coherence-based performance guarantees for estimating a sparse vector under random noise. IEEE Trans. Signal Process. **58**(10), 5030–5043 (2010)
7. Blumensath, T., Davies, M.E.: Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. **27**(3), 265–274 (2009)

8. Bodmann, B.G., Paulsen, V.I.: Frames, graphs and erasures. Linear Algebra Appl. **404**, 118–146 (2005)
9. Boufounos, P., Kutynio, G., Rahut, H.: Sparse recovery from combined fusion frame measurements. IEEE Trans. Inf. Theory **57**(6), 3864–3876 (2011)
10. Bourgain, J., Dilworth, S.J., Ford, K., Konyagin, S.V., Kutzarova, D.: Breaking the k^2 barrier for explicit RIP matrices. In: Proc. 43rd Annu. ACM Symp. Theory Computing (STOC'11), San Jose, California, pp. 637–644 (2011)
11. Boyd, S., Vandenberghe, L.: Convex Optimization. Cambridge University Press, Cambridge (2004)
12. Bruckstein, A.M., Donoho, D.L., Elad, M.: From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009)
13. Calderbank, R., Casazza, P., Heinecke, A., Kutyniok, G., Pezeshki, A.: Sparse fusion frames: existence and construction. Adv. Comput. Math. **35**, 1–31 (2011)
14. Candès, E.J.: The restricted isometry property and its implications for compressed sensing. In: C. R. Acad. Sci., Ser. I, Paris, vol. 346, pp. 589–592 (2008)
15. Candès, E.J., Plan, Y.: Near-ideal model selection by ℓ_1 minimization. Ann. Stat. **37**(5A), 2145–2177 (2009)
16. Candès, E.J., Romberg, J., Tao, T.: Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inform. Theory **52**(2), 489–509 (2006)
17. Candès, E.J., Tao, T.: Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inform. Theory **52**(12), 5406–5425 (2006)
18. Candès, E.J., Tao, T.: The Dantzig selector: statistical estimation when p is much larger than n. Ann. Stat. **35**(6), 2313–2351 (2007)
19. Casazza, P., Fickus, M., Mixon, D., Wang, Y., Zhou, Z.: Constructing tight fusion frames. Appl. Comput. Harmon. Anal. **30**, 175–187 (2011)
20. Casazza, P., Leon, M.: Existence and construction of finite tight frames. J. Concr. Appl. Math. **4**(3), 277–289 (2006)
21. Casazza, P.G., Kovačević, J.: Equal-norm tight frames with erasures. Appl. Comput. Harmon. Anal. **18**(2–4), 387–430 (2003)
22. Chen, S.S., Donoho, D.L., Saunders, M.A.: Atomic decomposition by basis pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1998)
23. Cohen, A., Dahmen, W., Devore, R.A.: Compressed sensing and best k-term approximation. J. Am. Math. Soc. **22**(1), 211–231 (2009)
24. Conway, J.H., Hardin, R.H., Sloane, N.J.A.: Packing lines, planes, etc.: packings in Grassmannian spaces. Exp. Math. **5**(2), 139–159 (1996)
25. Dai, W., Milenkovic, O.: Subspace pursuit for compressive sensing signal reconstruction. IEEE Trans. Inform. Theory **55**(5), 2230–2249 (2009)
26. Devore, R.A.: Nonlinear approximation. In: Iserles, A. (ed.) Acta Numerica, vol. 7, pp. 51–150. Cambridge University Press, Cambridge (1998)
27. Donoho, D.L.: Compressed sensing. IEEE Trans. Inform. Theory **52**(4), 1289–1306 (2006)
28. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via ℓ^1 minimization. Proc. Natl. Acad. Sci. **100**(5), 2197–2202 (2003)
29. Donoho, D.L., Elad, M., Temlyakov, V.N.: Stable recovery of sparse overcomplete representations in the presence of noise. IEEE Trans. Inform. Theory **52**(1), 6–18 (2006)
30. Donoho, D.L., Huo, X.: Uncertainty principles and ideal atomic decomposition. IEEE Trans. Inform. Theory **47**(7), 2845–2862 (2001)
31. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. Biometrika **81**(3), 425–455 (1994)
32. Efron, B., Hastie, T., Johnstone, I., Tibshirani, R.: Least angle regression. Ann. Stat. **32**(2), 407–451 (2004)
33. Ehrgott, M.: Multicriteria Optimization, 2nd edn. Springer, Berlin (2005)
34. Eldar, Y., Kutyniok, G.: Compressed Sensing: Theory and Applications, 1st edn. Cambridge University Press, Cambridge (2012)

35. Eldar, Y.C., Kuppinger, P., Bölcskei, H.: Block-sparse signals: uncertainty relations and efficient recovery. IEEE Trans. Signal Process. **58**(6), 3042–3054 (2010)
36. Fickus, M., Mixon, D.G., Tremain, J.C.: Steiner equiangular tight frames. Linear Algebra Appl. **436**(5), 1014–1027 (2012). doi:10.1016/j.laa.2011.06.027
37. Fletcher, A.K., Rangan, S., Goyal, V.K.: Necessary and sufficient conditions for sparsity pattern recovery. IEEE Trans. Inform. Theory **55**(12), 5758–5772 (2009)
38. Foster, D.P., George, E.I.: The risk inflation criterion for multiple regression. Ann. Stat. **22**(4), 1947–1975 (1994)
39. Genovese, C.R., Jin, J., Wasserman, L., Yao, Z.: A comparison of the lasso and marginal regression. J. Mach. Learn. Res. **13**, 2107–2143 (2012)
40. Geršgorin, S.A.: Über die Abgrenzung der Eigenwerte einer Matrix. Izv. Akad. Nauk SSSR Ser. Fiz.-Mat. **6**, 749–754 (1931)
41. Gorodnitsky, I.F., Rao, B.D.: Sparse signal reconstruction from limited data using FOCUSS: a re-weighted minimum norm algorithm. IEEE Trans. Signal Process. **45**(3), 600–616 (1997)
42. Gribonval, R., Nielsen, M.: Sparse representations in unions of bases. IEEE Trans. Inform. Theory **49**(12), 3320–3325 (2003)
43. Hajek, B., Seri, P.: Lex-optimal online multiclass scheduling with hard deadlines. Math. Oper. Res. **30**(3), 562–596 (2005)
44. Haupt, J., Bajwa, W.U., Raz, G., Nowak, R.: Toeplitz compressed sensing matrices with applications to sparse channel estimation. IEEE Trans. Inform. Theory **56**(11), 5862–5875 (2010)
45. Haupt, J., Nowak, R.: Compressive sampling for signal detection. In: Proc. IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), vol. 3, pp. III-1509–III-1512 (2007)
46. Holmes, R.B., Paulsen, V.I.: Optimal frames for erasures. Linear Algebra Appl. **377**(15), 31–51 (2004)
47. Hsu, D., Kakade, S., Langford, J., Zhang, T.: Multi-label prediction via compressed sensing. In: Advances in Neural Information Processing Systems, pp. 772–780 (2009)
48. Isermann, H.: Linear lexicographic optimization. OR Spektrum **4**(4), 223–228 (1982)
49. Kutyniok, G., Pezeshki, A., Calderbank, R., Liu, T.: Robust dimension reduction, fusion frames, and Grassmannian packings. Appl. Comput. Harmon. Anal. **26**(1), 64–76 (2009)
50. Lancaster, P., Tismenetsky, M.: The Theory of Matrices, 2nd edn. Academic Press, Orlando (1985)
51. Mallat, S.G., Zhang, Z.: Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
52. Malozemov, V.N., Pevnyi, A.B.: Equiangular tight frames. J. Math. Sci. **157**(6), 789–815 (2009)
53. Meinshausen, N., Bühlmann, P.: High-dimensional graphs and variable selection with the Lasso. Ann. Stat. **34**(3), 1436–1462 (2006)
54. Natarajan, B.K.: Sparse approximate solutions to linear systems. SIAM J. Comput. **24**(2), 227–234 (1995)
55. Needell, D., Tropp, J.A.: CoSaMP: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**(3), 301–321 (2009)
56. Paredes, J., Wang, Z., Arce, G., Sadler, B.: Compressive matched subspace detection. In: Proc. 17th European Signal Processing Conference, Glasgow, Scotland, pp. 120–124 (2009)
57. Reeves, G., Gastpar, M.: A note on optimal support recovery in compressed sensing. In: Proc. 43rd Asilomar Conf. Signals, Systems and Computers, Pacific Grove, CA (2009)
58. Renes, J.: Equiangular tight frames from Paley tournaments. Linear Algebra Appl. **426**(2–3), 497–501 (2007)
59. Santosa, F., Symes, W.W.: Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Statist. Comput. **7**(4), 1307–1330 (1986)
60. Scharf, L.L.: Statistical Signal Processing. Addison-Wesley, Cambridge (1991)
61. Schnass, K., Vandergheynst, P.: Average performance analysis for thresholding. IEEE Signal Process. Lett. **14**(11), 828–831 (2007)

62. Stojnic, M., Parvaresh, F., Hassibi, B.: On the representation of block-sparse signals with an optimal number of measurements. IEEE Trans. Signal Process. **57**(8), 3075–3085 (2009)
63. Strohmer, T.: A note on equiangular tight frames. Linear Algebra Appl. **429**(1), 326–330 (2008)
64. Strohmer, T., Heath, R.W. Jr.: Grassmannian frames with applications to coding and communication. Appl. Comput. Harmon. Anal. **14**(3), 257–275 (2003)
65. Sustik, M., Tropp, J.A., Dhillon, I.S., Heath, R.W. Jr.: On the existence of equiangular tight frames. Linear Algebra Appl. **426**(2–3), 619–635 (2007)
66. Tibshirani, R.: Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B **58**(1), 267–288 (1996)
67. Tropp, J., Gilbert, A., Muthukrishnan, S., Strauss, M.: Improved sparse approximation over quasiincoherent dictionaries. In: Proc. IEEE Conf. Image Processing (ICIP'03), pp. 37–40 (2003)
68. Tropp, J.A.: Greed is good: algorithmic results for sparse approximation. IEEE Trans. Inform. Theory **50**(10), 2231–2242 (2004)
69. Tropp, J.A.: Just relax: convex programming methods for identifying sparse signals in noise. IEEE Trans. Inform. Theory **52**(3), 1030–1051 (2006)
70. Tropp, J.A.: Norms of random submatrices and sparse approximation. In: C. R. Acad. Sci., Ser. I, Paris, vol. 346, pp. 1271–1274 (2008)
71. Tropp, J.A.: On the conditioning of random subdictionaries. Appl. Comput. Harmon. Anal. **25**, 1–24 (2008)
72. Tropp, J.A., Wright, S.J.: Computational methods for sparse solution of linear inverse problems. Proc. IEEE **98**(5), 948–958 (2010)
73. Wainwright, M.J.: Sharp thresholds for high-dimensional and noisy sparsity recovery using ℓ_1-constrained quadratic programming (Lasso). IEEE Trans. Inform. Theory **55**(5), 2183–2202 (2009)
74. Wang, Z., Arce, G., Sadler, B.: Subspace compressive detection for sparse signals. In: IEEE Int. Conf. Acoust., Speech, Signal Process. (ICASSP), pp. 3873–3876 (2008)
75. Welch, L.: Lower bounds on the maximum cross correlation of signals. IEEE Trans. Inform. Theory **20**(3), 397–399 (1974)
76. Zahedi, R., Pezeshki, A., Chong, E.K.P.: Robust measurement design for detecting sparse signals: equiangular uniform tight frames and Grassmannian packings. In: Proc. 2010 American Control Conference (ACC), Baltimore, MD (2010)
77. Zahedi, R., Pezeshki, A., Chong, E.K.P.: Measurement design for detecting sparse signals. Phys. Commun. **5**(2), 64–75 (2012). doi:10.1016/j.phycom.2011.09.007
78. Zelnik-Manor, L., Rosenblum, K., Eldar, Y.C.: Sensing matrix optimization for block-sparse decoding. IEEE Trans. Signal Process. **59**(9), 4300–4312 (2011)
79. Zhao, P., Yu, B.: On model selection consistency of Lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)

Chapter 10
Finite Frames and Filter Banks

Matthew Fickus, Melody L. Massar, and Dustin G. Mixon

Abstract Filter banks are fundamental tools of signal and image processing. A filter is a linear operator which computes the inner products of an input signal with all translates of a fixed function. In a filter bank, several filters are applied to the input, and each of the resulting signals is then downsampled. Such operators are closely related to frames, which consist of equally spaced translates of a fixed set of functions. In this chapter, we highlight the rich connections between frame theory and filter banks. We begin with the algebraic properties of related operations, such as translation, convolution, downsampling, the discrete Fourier transform, and the discrete Z-transform. We then discuss how basic frame concepts, such as frame analysis and synthesis operators, carry over to the filter bank setting. The basic theory culminates with the representation of a filter bank's synthesis operator in terms of its polyphase matrix. This polyphase representation greatly simplifies the process of constructing a filter bank frame with a given set of properties. Indeed, we use this representation to better understand the special case in which the filters are modulations of each other, namely Gabor frames.

Keywords Filter · Convolution · Translation · Polyphase · Gabor

10.1 Introduction

Frame theory is intrinsically linked to the study of filter banks, with the two fields sharing a great deal of common history. Indeed, much of the modern terminology of frames, such as analysis and synthesis operators, was borrowed from the filter bank literature. And though frames were originally developed for the study of nonharmonic Fourier series, much of their recent popularity stems from their use in

M. Fickus (✉) · M.L. Massar
Department of Mathematics, Air Force Institute of Technology, Wright-Patterson AFB, OH 45433, USA
e-mail: Matthew.Fickus@afit.edu

D.G. Mixon
Program in Applied and Computational Mathematics, Princeton University, Princeton, NJ 08544, USA

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_10,

Gabor (time-frequency) and wavelet (time-scale) analysis; both Gabor and wavelet transforms are examples of filter banks.

In this chapter, we highlight the connections between frames and filter banks. Specifically, we discuss how analysis and synthesis filter banks correspond to the analysis and synthesis operators of a certain class of frames. We then discuss the *polyphase* representation of a filter bank—a key tool in filter bank design—which reduces the problem of constructing a high-dimensional filter bank frame to that of constructing a low-dimensional frame for a space of polynomials. For the signal processing researcher, these results show how to build filter banks that possess the hallmarks of any good frame: robustness against noise and flexibility with respect to redundancy. Meanwhile, for the frame theorist, these results show how to construct many explicit examples of frames, and also pose many new, interesting problems regarding the generalization of frame theory to spaces of polynomials.

Like frames, filter banks are essentially sequences of vectors in a Hilbert space. But, whereas the vectors in a frame are somewhat arbitrary, the vectors in a filter bank are, by definition, obtained by taking all evenly spaced translates of the vectors from some given collection. As such, we only consider filter banks in Hilbert spaces on which a translation operator can be defined. In the signal processing literature, the Hilbert space of choice is usually

$$\ell^2(\mathbb{Z}) := \left\{ x : \mathbb{Z} \to \mathbb{C} \;\middle|\; \sum_{k\in\mathbb{Z}} |x[k]|^2 < \infty \right\},$$

namely the space of all finite-energy complex-valued sequences over the integers. Here, the *translation by k* operator is $\mathrm{T}^k : \ell^2(\mathbb{Z}) \to \ell^2(\mathbb{Z})$, $(\mathrm{T}^k x)[k'] := x[k' - k]$. Electrical engineers like to use this space despite its infinite dimension since it naturally corresponds to the discrete samples of an analog signal defined over a real-variable time axis. Such signals naturally arise in a variety of real-world applications.

For instance, in classical radar, one transmits an electromagnetic pulse which we model as a function of time φ. This pulse travels through the atmosphere until it encounters a target, such as an aircraft. The pulse then bounces off the target and returns to a receiver which is located alongside the transmitter. Here, the measured return signal x can be modeled as $x[k'] = \alpha\varphi[k' - k] + \nu[k']$, where α relates to the fraction of the transmitted energy that was received, k corresponds to the time lag incurred by φ as it traveled to the target and back again, and $\nu[k']$ is noise, such as background radiation. The radar operator then processes the received signal $x = \alpha\mathrm{T}^k\varphi + \nu$ with the goal of estimating k: multiplying this time lag by one-half the speed of light gives the distance to the target. The standard method for such processing is known as *matched filtering*, which computes the inner products of x with all possible translates of φ:

$$\left\langle x, \mathrm{T}^{k'}\varphi \right\rangle = \left\langle \alpha\mathrm{T}^k\varphi + \nu, \mathrm{T}^{k'}\varphi \right\rangle = \alpha\left\langle \varphi, \mathrm{T}^{k-k'}\varphi \right\rangle + \left\langle \nu, \mathrm{T}^{k'}\varphi \right\rangle.$$

Here, the Cauchy-Schwarz inequality gives

$$\left|\left\langle \varphi, \mathrm{T}^{k-k'}\varphi \right\rangle\right| \leq \|\varphi\| \left\|\mathrm{T}^{k-k'}\varphi\right\| = \|\varphi\|^2 = \left\langle \varphi, \mathrm{T}^{k-k}\varphi \right\rangle,$$

and so it's reasonable to believe that the desired parameter k can be approximated by the k' that maximizes $|\langle x, \mathrm{T}^{k'}\varphi\rangle|$, provided the magnitude of the noise ν is relatively small. Here, the term "matched" in "matched filtering" means that one analyzes the returned signal x in terms of the transmitted one φ.

More generally, regardless of the relationship between x and φ, the act of computing the inner products of x with all translates of φ is known as *filtering* x. To be precise, in the language of frames, this operation corresponds to applying the frame analysis operator of $\{\mathrm{T}^k\varphi\}_{k\in\mathbb{Z}}$, and so we refer to it as the *analysis filter* corresponding to the *filter* φ. *Filter banks* arise from a collection $\{\varphi_n\}_{n=0}^{N-1}$ of such filters. In particular, an *analysis filter bank* is the frame analysis operator of $\{\mathrm{T}^k\varphi_n\}_{n=0,k\in\mathbb{Z}}^{N-1}$, namely a transform which, given x, computes $\langle x, \mathrm{T}^k\varphi_n\rangle$ for all k and n. Such filter banks arise naturally in applications. For instance, in radar one often uses Gabor filter banks in which the φ_n's correspond to distinct modulations of the transmitted waveform φ; by computing the indices k and n for which $|\langle x, \mathrm{T}^k\varphi_n\rangle|$ is maximal, one estimates not only the distance to the target via k, but also the speed at which the target is approaching the radar via n, as a consequence of the Doppler effect.

Similar rationales have led to the use of filter banks in many real-world applications. In short, they are natural tools for detecting the times or locations at which a given fixed set of features appear in a given signal. As filter banks have risen in popularity, more attention has been paid to their subtle details. In particular, with the rise of wavelets, attention shifted to the case where one does not compute inner products of x with every translate of φ_n, but rather only with a subcollection $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,p\in\mathbb{Z}}^{N-1}$ of equally spaced translates; this helps one compensate for the greater amount of computation required as N grows large. Attention has further shifted toward the sensitivity of filter banks to noise, as well their use for signal reconstruction; both topics led to the advent of frames, namely a desire to find *frame bounds* A and B such that

$$A\|x\|^2 \le \sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}}\left|\left\langle x, \mathrm{T}^{Mp}\varphi_n\right\rangle\right|^2 \le B\|x\|^2, \quad \forall x \in \ell^2(\mathbb{Z}).$$

As discussed in previous chapters, such frame expansions are more robust to noise and conducive to stable reconstruction provided A is close to B. The fundamental frame-theoretic properties of filter banks over $\ell^2(\mathbb{Z})$ are given in [4, 8].

This book is specifically about *finite* frames. As such, in this chapter we cannot make direct use of the infinite-dimensional results of [4, 8]. Rather, we follow the approach of [7, 11], in which the results of [4, 8] are generalized to the context of the finite-dimensional Hilbert space:

$$\ell(\mathbb{Z}_P) := \left\{x : \mathbb{Z} \to \mathbb{C} \;\middle|\; x[p+P] = x[p]\ \forall p \in \mathbb{Z}\right\}, \tag{10.1}$$

namely the space of all P-periodic complex-valued sequences over the integers, where P is any fixed positive integer. This space is a Hilbert space under the stan-

dard inner product

$$\langle x_1, x_2 \rangle := \sum_{p \in \mathbb{Z}_P} x_1[p] \big(x_2[p]\big)^*,$$

where ζ^* denotes the complex conjugate of a number $\zeta \in \mathbb{C}$ while the indexing "$p \in \mathbb{Z}_p$" means choosing one representative from each of the P cosets of the subgroup $P\mathbb{Z}$ of the integers; one may for example take $p = 0, \ldots, P-1$. For each $p \in \mathbb{Z}$, consider the δ-Dirac function $\delta_p \in \ell(\mathbb{Z}_P)$:

$$\delta_p\big[p'\big] = \begin{cases} 1, & p = p' \mod P, \\ 0, & p \neq p' \mod P. \end{cases}$$

One can quickly show that $\{\delta_p\}_{p \in \mathbb{Z}_P}$ is an orthonormal basis in $\ell(\mathbb{Z}_P)$—called the *standard basis*—and as such $\ell(\mathbb{Z}_P)$ is a P-dimensional Hilbert space. It is therefore isometric to $\mathbb{C}^P$; in fact the only "difference" between vectors in $\mathbb{C}^P$ and those in $\ell(\mathbb{Z}_P)$ is that the indices of vectors in $\mathbb{C}^P$ are typically taken to be $p = 1, \ldots, P$ whereas the indices of vectors in $\ell(\mathbb{Z}_P)$ can be regarded as elements of the cyclic group $\mathbb{Z}_P := \mathbb{Z}/P\mathbb{Z} = \{0, \ldots, P-1\}$.

The translation operator over $\ell(\mathbb{Z}_P)$ is defined similarly to its infinite-dimensional cousin, namely $\mathrm{T} : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$, $(\mathrm{T}x)[p] := x[p-1]$. However, these two translation operators behave differently due to the periodic nature of signals in $\ell(\mathbb{Z}_P)$. Indeed, viewing x as a $P \times 1$ column vector indexed from 0 to $P-1$, $\mathrm{T}x$ is obtained by shifting the entries of the vector down by one entry *and* cycling the $(P-1)$th entry of x up into the zero index: $(\mathrm{T}x)[0] = x[0-1] = x[P-1]$. More generally, p' repeated applications of T correspond to cyclic translation by p': $(\mathrm{T}^{p'}x)[p] = x[p-p']$, where the P-periodicity of x implies that the subtraction $p - p'$ may be performed modulo P. In particular, this cyclic translation operator satisfies $(\mathrm{T}^P x)[p] = x[p-P] = x[p]$ for all x and so $\mathrm{T}^P = \mathrm{I}$. This stands in contrast to the translation operator on $\ell^2(\mathbb{Z})$ which satisfies $\mathrm{T}^m \neq \mathrm{I}$ for all nonzero integers m.

Working over $\ell(\mathbb{Z}_P)$ instead of $\ell^2(\mathbb{Z})$ has both advantages and disadvantages. One can easily argue that $\ell^2(\mathbb{Z})$ is often the more realistic signal model, since many real-world signals, such as electromagnetic waves and images, are usually not periodic. At the same time, $\ell(\mathbb{Z}_P)$ is a more realistic setting from the point of view of computation: a computer can only perform a finite number of algebraic operations in any fixed period of time. Also, from the point of view of the mathematics itself, working over $\ell(\mathbb{Z}_P)$ makes filter banks a purely algebraic topic, while working over $\ell^2(\mathbb{Z})$ requires functional analysis. In any case, with regard to this chapter, this point is moot: our focus is the special topic of how filter banks are examples of *finite* frames, and as such, we must work with finite-dimensional filter banks. That said, for one to become a true filter bank expert, both in theory and application, one must understand them in both settings; comprehensive, mathematician-accessible introductions to filter banks over $\ell^2(\mathbb{Z})$ from the engineering perspective are given in [23, 26]. Much of the finite-dimensional presentation of this chapter is taken from [7] and [11].

In the next section, we discuss basic concepts of frames and filters, with a particular emphasis on the signal processing tools, such as convolutions, upsampling, discrete Z-transforms and discrete Fourier transforms, that we will need later on. In Sect. 10.3, we discuss the fundamental relationships between frames and filter banks. In particular, we see how the frame analysis and synthesis operators of certain collections of vectors are analysis and synthesis filter banks, respectively. In Sect. 10.4, we discuss the polyphase representation of a filter bank, and use it to provide an efficient method for computing the optimal frame bounds of a filter bank. In the fifth and final section, we then exploit this polyphase representation to delve briefly into the theory of discrete Gabor frames.

10.2 Frames and Filters

Before discussing filter banks, let's review the basics of finite frame theory in the context of the P-dimensional Hilbert space $\ell(\mathbb{Z}_P)$ defined in (10.1). Let $\mathcal{N}$ be an index set of N elements, and let $\ell(\mathcal{N}) = \{y : \mathcal{N} \to \mathbb{C}\}$ denote the set of complex-valued functions over $\mathcal{N}$. The *synthesis operator* of a sequence of vectors $\Phi = \{\varphi_n\}_{n=1}^N$ in $\ell(\mathbb{Z}_P)$ is $\Phi : \ell(\mathcal{N}) \to \ell(\mathbb{Z}_P)$, $\Phi y := \sum_{n\in\mathcal{N}} y[n]\varphi_n$. Essentially, Φ is the $P \times N$ matrix whose columns are the φ_n's. Note that here and throughout, we make no notational distinction between the vectors themselves and the synthesis operator they induce. The *analysis operator* of Φ is its adjoint $\Phi^* : \ell(\mathbb{Z}_P) \to \ell(\mathcal{N})$ defined by $(\Phi^* x)[n] := \langle x, \varphi_n\rangle$ for all $n \in \mathcal{N}$. The vectors Φ are said to be a *frame* for $\ell(\mathbb{Z}_P)$ if there exist *frame bounds* $0 < A \le B < \infty$ such that $A\|x\|^2 \le \|\Phi^* x\|^2 \le B\|x\|^2$ for all $x \in \ell(\mathbb{Z}_P)$. The optimal frame bounds A and B of an arbitrary Φ are the least and greatest eigenvalues of the *frame operator* $\Phi\Phi^* : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$ given by

$$\Phi\Phi^* = \sum_{n\in\mathcal{N}} \varphi_n\varphi_n^*,$$

respectively, where the "row vector" φ_n^* is the linear functional $\varphi_n^* : \ell(\mathbb{Z}_P) \to \mathbb{C}$, $\varphi_n^* x := \langle x, \varphi_n\rangle$. In particular, Φ is a frame if and only if the φ_n's span $\ell(\mathbb{Z}_P)$, which necessitates $P \le N$. Frames provide overcomplete decompositions of vectors; if Φ is a frame for $\ell(\mathbb{Z}_P)$, then any $x \in \ell(\mathbb{Z}_P)$ can be decomposed as

$$x = \Phi\Psi^* x = \sum_{n\in\mathcal{N}} \langle x, \psi_n\rangle\varphi_n,$$

where $\Psi = \{\psi_n\}_{n\in\mathcal{N}}$ is a *dual frame* of Φ, meaning it satisfies $\Phi\Psi^* = \mathrm{I}$. Any frame has at least one dual, namely the *canonical* dual given by the pseudoinverse $\Psi = (\Phi\Phi^*)^{-1}\Phi$. Note that computing a canonical dual involves the inversion of the frame operator. As such, when designing a frame for a given application, it is important to retain control over the spectrum of $\Phi\Phi^*$.

10.2.1 Filters

The remainder of the material in this section is classical, being a finite-dimensional version of the well-known theory of filters [20, 24]. A filter bank is a special type of frame in which the frame elements are required to be translates of each other. Before studying filter banks in general, it helps to first consider the special case in which the frame consists of every cyclic translate of a single vector φ in $\ell(\mathbb{Z}_P)$. To be precise, recall from the introduction that the pth *cyclic translate* of φ is $(\mathrm{T}^p\varphi)[p'] := \varphi[p' - p]$. Due to the fact that $\mathrm{T}^P = \mathrm{I}$, we do not consider $\{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}}$ but rather $\{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}_P}$. Here, the indexing set $\mathcal{N}$ is $\mathbb{Z}_P$, and so the analysis and synthesis operators of $\{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}_P}$ map from $\ell(\mathbb{Z}_P)$ into itself. In particular, the synthesis operator $\Phi : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$, which is also known as the *synthesis filter* in this context, is

$$(\Phi y)[p] = \sum_{p'\in\mathbb{Z}_P} y\big[p'\big]\big(\mathrm{T}^{p'}\varphi\big)[p] = \sum_{p'\in\mathbb{Z}_P} y\big[p'\big]\varphi\big[p - p'\big]. \tag{10.2}$$

Harmonic analysts will recognize the right-hand side of (10.2). Indeed, in general the *convolution* of $y_1, y_2 \in \ell(\mathbb{Z}_P)$ is $y_1 * y_2 \in \ell(\mathbb{Z}_P)$ defined by

$$(y_1 * y_2)[p] := \sum_{p'\in\mathbb{Z}_P} y_1\big[p'\big]y_2\big[p - p'\big],$$

and so the synthesis filter (10.2) is the operator that convolves a given input y with φ. The following easily verified result gives several useful properties of convolution.

Proposition 10.1 *For any* $y_1, y_2, y_3 \in \ell(\mathbb{Z}_P)$,

(a) *Convolution is associative*: $(y_1 * y_2) * y_3 = y_1 * (y_2 * y_3)$.
(b) *Convolution is commutative*: $y_1 * y_2 = y_2 * y_1$.
(c) *Convolution's multiplicative identity is* δ_0: $y_1 * \delta_0 = y_1$.
(d) *Convolution distributes over addition*: $(y_1 + y_2) * y_3 = (y_1 * y_3) + (y_2 * y_3)$.
(e) *Convolution distributes over scalar multiplication*: $(\alpha y_1) * y_2 = \alpha(y_1 * y_2)$.

In general, a linear operator $\Phi : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$ is referred to as a *time-invariant filter* precisely when there exists φ in $\ell(\mathbb{Z}_P)$ such that $\Phi y = y * \varphi$ for all $y \in \ell(\mathbb{Z}_P)$. Though succinct, this definition of a filter is not very intuitive. The next result gives a better understanding of what a time-invariant filter truly is: a linear operator that commutes with translation, that is, $\Phi\mathrm{T} = \mathrm{T}\Phi$. In other words, filters are linear operators on $\ell(\mathbb{Z}_P)$ for which delaying one's input into Φ by a given time results in an equal delay in output. We further see that this is equivalent to having Φ be a linear combination of powers of the translation operator itself.

Proposition 10.2 *The following are equivalent*:

(a) Φ *is a time-invariant filter.*
(b) Φ *is linear and commutes with translation.*

(c) *Φ is a linear combination of the operators* $\{\mathrm{T}^p\}_{p\in\mathbb{Z}_P}$.

Moreover, for such Φ we have $\Phi y = y * \varphi$ and $\Phi = \sum_{p\in\mathbb{Z}_p} \varphi[p]\mathrm{T}^p$ where $\varphi = \Phi\delta_0$.

Proof $(a \Rightarrow c)$ Let Φ be a filter. By definition, there exists $\varphi \in \ell(\mathbb{Z}_P)$ such that

$$(\Phi y)\big[p'\big] = (y * \varphi)\big[p'\big] = (\varphi * y)\big[p'\big] = \sum_{p\in\mathbb{Z}_P} \varphi[p] y\big[p' - p\big]$$

$$= \sum_{p\in\mathbb{Z}_P} \varphi[p]\big(\mathrm{T}^p y\big)\big[p'\big],$$

for any $y \in \ell(\mathbb{Z}_P)$ and $p' \in \mathbb{Z}_P$, and so $\Phi = \sum_{p\in\mathbb{Z}_P} \varphi[p]\mathrm{T}^p$ as claimed.

$(c \Rightarrow b)$ Letting $\Phi = \sum_{p\in\mathbb{Z}_P} \varphi[p]\mathrm{T}^p$, we immediately have that Φ is linear. Moreover,

$$\Phi\mathrm{T} = \sum_{p\in\mathbb{Z}_P} \varphi[p]\mathrm{T}^p\mathrm{T} = \sum_{p\in\mathbb{Z}_P} \varphi[p]\mathrm{T}^{p+1} = \sum_{p\in\mathbb{Z}_P} \varphi[p]\mathrm{T}\mathrm{T}^p = \mathrm{T}\sum_{p\in\mathbb{Z}_P} \varphi[p]\mathrm{T}^p = \mathrm{T}\Phi.$$

$(b \Rightarrow a)$ Let Φ be linear and satisfy $\Phi\mathrm{T} = \mathrm{T}\Phi$. Letting $\varphi = \Phi\delta_0$, we therefore have that $\Phi\delta_p = \Phi\mathrm{T}^p\delta_0 = \mathrm{T}^p\Phi\delta_0 = \mathrm{T}^p\varphi$ for all $p \in \mathbb{Z}_P$. As such, for any $y \in \ell(\mathbb{Z}_P)$,

$$(\Phi y)\big[p'\big] = \left(\Phi \sum_{p\in\mathbb{Z}_P} y[p]\delta_p\right)\big[p'\big]$$

$$= \sum_{p\in\mathbb{Z}_P} y[p](\Phi\delta_p)\big[p'\big]$$

$$= \sum_{p\in\mathbb{Z}_P} y[p]\big(\mathrm{T}^p\varphi\big)\big[p'\big]$$

$$= \sum_{p\in\mathbb{Z}_P} y[p]\varphi\big[p' - p\big]$$

$$= (y * \varphi)\big[p'\big],$$

and so $\Phi y = y * \varphi$ as claimed. □

Here, an illustrative example is helpful.

Example 10.1 Let $P = 8$. We can represent any $x \in \ell(\mathbb{Z}_8)$ as a column vector in $\mathbb{C}^8$, provided we take the indexing of this vector to begin at 0. The entries of this column are the inner products of x with the elements of the standard basis $\{\delta_p\}_{p=0}^7$. This representation induces an 8×8 matrix representation of any linear operator Φ from $\ell(\mathbb{Z}_8)$ into itself: let the pth column of the matrix be the column vector

representation of $\Phi\delta_p$. In particular, the translation operator is represented as

$$\mathrm{T} = \begin{bmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}. \tag{10.3}$$

Let's use Proposition 10.2 to compute the matrix representation of a filter Φ defined by $\Phi y = y * \varphi$ where φ is taken to be of the form $a\delta_0 + b\delta_1 + c\delta_2 + d\delta_3$ for the sake of simplicity, where a, b, c, and d are some arbitrarily chosen complex numbers. By Proposition 10.2, Φ is of the form $\Phi = a\mathrm{T}^0 + b\mathrm{T}^1 + c\mathrm{T}^2 + d\mathrm{T}^3$. That is, Φ is a linear combination of the translation by zero, one, two, and three operators whose matrix representations are:

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

where for the sake of readability we have suppressed all zero entries. Combining these four matrices with coefficients a, b, c, and d yields the matrix representation of the filter Φ:

$$\Phi = \begin{bmatrix} a & & & & & d & c & b \\ b & a & & & & & d & c \\ c & b & a & & & & & d \\ d & c & b & a & & & & \\ & d & c & b & a & & & \\ & & d & c & b & a & & \\ & & & d & c & b & a & \\ & & & & d & c & b & a \end{bmatrix}. \tag{10.4}$$

Note that Φ is constant along diagonals, and moreover that these diagonals wrap around from left to right and top to bottom. That is, the matrix representation of Φ satisfies $\Phi[p, p'] = \Phi[p+1, p'+1]$ for all p and p', where the index arithmetic is performed modulo P. Such matrices are termed *circulant*. Every circulant matrix corresponds to a filter Φ where φ is given by the first column of the matrix. In particular, for a fully general filter on $\ell(\mathbb{Z}_8)$ we have $\varphi = a\delta_0 + b\delta_1 + c\delta_2 + d\delta_3 +$

$e\delta_4 + f\delta_5 + g\delta_6 + h\delta_7$ which corresponds to placing the values h, g, f, and e on the first, second, third, and fourth circulant superdiagonals of (10.4), respectively.

Applying (10.4) to an input column vector y yields the following output vector Φy:

$$\begin{aligned}
(y * \varphi)[0] &= ay[0] + by[7] + cy[6] + dy[5],\\
(y * \varphi)[1] &= ay[1] + by[0] + cy[7] + dy[6],\\
(y * \varphi)[2] &= ay[2] + by[1] + cy[0] + dy[7],\\
(y * \varphi)[3] &= ay[3] + by[2] + cy[1] + dy[0],\\
(y * \varphi)[4] &= ay[4] + by[3] + cy[2] + dy[1],\\
(y * \varphi)[5] &= ay[5] + by[4] + cy[3] + dy[2],\\
(y * \varphi)[6] &= ay[6] + by[5] + cy[4] + dy[3],\\
(y * \varphi)[7] &= ay[7] + by[6] + cy[5] + dy[4].
\end{aligned} \tag{10.5}$$

Here, we see what filtering really does: it computes "rolling" inner products of the input signal with coefficients from the filter. In particular, if the entries of φ are nonnegative and sum to one, then filtering y with φ produces a sequence of rolling averages of the values of y. For other choices of φ, such as $\varphi = \delta_0 - \delta_1$, filtering becomes akin to taking a discrete derivative. In the next subsection, we use the discrete Fourier transform to get an even better intuitive understanding of filtering.

We now use the previous example to give a few notes on terminology. Though the vector φ is occasionally referred to as a "filter," technically speaking, this term should be reserved for the operation of convolving with φ; in the signal processing literature, φ is known as the *impulse response* of the filter since it is the output one receives after passing the *impulse* δ_0 through Φ, that is, $\varphi = \delta_0 * \varphi = \Phi\delta_0$.

The number K of nonzero values of φ is known as its number of *taps*; one often seeks to design filters in which K is small, since a direct computation of (10.2) at any fixed p requires K multiplications. For example, when a, b, c, and d are nonzero, the filter (10.4) is called a 4-tap filter. In general, since we are working in $\ell(\mathbb{Z}_P)$ this number of taps is at most P. In particular, K is finite. However, in the standard signal processing literature, φ is taken to be a member of the infinite-dimensional space $\ell^2(\mathbb{Z})$, and there one must draw a distinction between those φ's with a finite number of taps and those with an infinite number, namely *finite impulse response* (FIR) filters and *infinite impulse response* (IIR) filters, respectively. Though the concept of FIR versus IIR does not carry over to $\ell(\mathbb{Z}_P)$, one nevertheless tries to keep the number of taps K as small as possible, subject to the other constraints that one needs φ to satisfy for a given application.

Causal filters are another important concept in the signal processing literature. To be precise, a filter φ in $\ell^2(\mathbb{Z})$ is *causal* if $\varphi[k] = 0$ for all $k < 0$. Causality is only a significant issue for signals whose input axis corresponds to time, such as audio signals, as opposed to images which have two spatial input axes. Indeed, for signals of time, causal filtering means that the filter requires no precognition: at any time,

the value of the filtered signal only depends on the values of the input signal at that and previous times. Such ideas do not immediately generalize to the $\ell(\mathbb{Z}_P)$ setting, as the requirement that $k < 0$ has no meaning in $\mathbb{Z}_P$. Nevertheless, we can mimic casuality by requiring that φ is supported on those integers which are equivalent to $\{0, \ldots, K-1\}$ modulo P; under this hypothesis (10.2) becomes

$$(\Phi y)[p] = (y * \varphi)[p] = (\varphi * y)[p] = \sum_{p' \in \mathbb{Z}_P} \varphi[p']y[p-p'] = \sum_{p'=0}^{K-1} \varphi[p']y[p-p'],$$

and so $(\Phi y)[p] = \varphi[0]y[p] + \varphi[1]y[p-1] + \cdots + \varphi[K-1]y[p-K+1]$, as desired. For example, as the impulse response φ of the filter Φ given in (10.2) is supported over the indices $\{0, 1, 2, 3\}$, then the value of $\Phi y = y * \varphi$ at any given time p depends only on the values of y at times p, $p-1$, $p-2$, and $p-3$, as validated in (10.5).

Having discussed filters at length, we now examine their frame-theoretic properties. We have already seen that the synthesis operator $\Phi : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$ of $\{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}_P}$ is given by $\Phi y = y * \varphi$. Meanwhile, the corresponding analysis operator $\Phi^* : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$ is given by

$$\big(\Phi^* x\big)[p] = \big\langle x, \mathrm{T}^p\varphi\big\rangle = \sum_{p' \in \mathbb{Z}_P} x[p']\big(\varphi[p'-p]\big)^* = \sum_{p' \in \mathbb{Z}_P} x[p']\tilde{\varphi}[p-p'] = x * \tilde{\varphi},$$

where $\tilde{\varphi}$ is the *involution* (conjugate reversal) of φ defined as $(\tilde{\varphi})[p] := (\varphi[-p])^*$. In particular, we see that the adjoint of filtering with φ is filtering with $\tilde{\varphi}$. We refer to Φ^* as the *analysis filter* of φ. For example, for the synthesis filter Φ over $\ell(\mathbb{Z}_8)$ of (10.4) whose impulse response is $\varphi = a\delta_0 + b\delta_1 + c\delta_2 + d\delta_3$, taking the conjugate transpose of (10.4) yields

$$\Phi^* = \begin{bmatrix} a^* & b^* & c^* & d^* & & & & \\ & a^* & b^* & c^* & d^* & & & \\ & & a^* & b^* & c^* & d^* & & \\ & & & a^* & b^* & c^* & d^* & \\ & & & & a^* & b^* & c^* & d^* \\ d^* & & & & & a^* & b^* & c^* \\ c^* & d^* & & & & & a^* & b^* \\ b^* & c^* & d^* & & & & & a^* \end{bmatrix},$$

namely the analysis filter Φ^* whose impulse response is

$$\tilde{\varphi} = a^*\delta_0 + b^*\delta_{-1} + c^*\delta_{-2} + d^*\delta_{-3} = a^*\delta_0 + d^*\delta_5 + c^*\delta_6 + b^*\delta_7.$$

Moreover, since both the analysis and synthesis operators of $\Phi = \{\mathrm{T}^P\varphi\}_{p\in\mathbb{Z}_P}$ are filters, then so is the frame operator due to the associativity of convolution:

$$\Phi\Phi^* x = \Phi^*(x * \varphi) = (x * \varphi) * \tilde{\varphi} = x * (\varphi * \tilde{\varphi}). \tag{10.6}$$

That is, the analysis, synthesis, and frame operators of $\Phi = \{\mathrm{T}^P\varphi\}_{p\in\mathbb{Z}_P}$ correspond to filtering with φ, $\tilde{\varphi}$, and $\varphi * \tilde{\varphi}$, respectively. The function $\varphi * \tilde{\varphi}$ is known as the *autocorrelation* of φ, as its value gives the correlation between φ and the translates of itself: $(\varphi * \tilde{\varphi})[p] = \langle \varphi, \mathrm{T}^p\varphi\rangle$.

Now that we have expressions for the canonical frame operators of $\Phi = \{\mathrm{T}^P\varphi\}_{p\in\mathbb{Z}_P}$, our next goals are to determine the conditions on φ that are needed to ensure that Φ is a frame for $\ell(\mathbb{Z}_P)$, and in this case, to determine dual frames Ψ. Now, the optimal frame bounds for $\Phi = \{\mathrm{T}^P\varphi\}_{p\in\mathbb{Z}_P}$ are given by the extreme eigenvalues of the frame operator $\Phi\Phi^*$. Since $\Phi\Phi^*$ is a filter, we first find the eigenvalues of any filter $\Phi y = y * \varphi$ and then apply this result where φ is replaced with $\varphi * \tilde{\varphi}$; the next subsection contains the tools needed to accomplish this task.

10.2.2 The Z-Transform and the Discrete Fourier Transform

The Z-transform is a standard tool in signal processing, and relates convolution to polynomial multiplication. To be precise, when working in the infinite-dimensional space $\ell^2(\mathbb{Z})$, the Z-transform of $\varphi \in \ell^1(\mathbb{Z})$ is the Laurent series

$$(\mathrm{Z}\varphi)(z) := \sum_{k=-\infty}^{\infty} \varphi[k] z^{-k}.$$

Note that the assumption that $\varphi \in \ell^1(\mathbb{Z})$ guarantees that this series converges absolutely on the unit circle. Further note that the Z-transform of an FIR filter is but a rational function, while the transform of a causal filter is a power series.

As our goal is to understand frames of translates in the finite-dimensional space $\ell(\mathbb{Z}_P)$, we must generalize this notion of a Z-transform. Mathematically speaking, this frees us from needing analysis, while, at the same time, it forces us to consider more exotic algebra. To be precise, the Z-*transform* of $y \in \ell(\mathbb{Z}_P)$ is

$$(\mathrm{Z}y)(z) := \sum_{p\in\mathbb{Z}_P} y[p] z^{-p}, \tag{10.7}$$

which lies within the ring of polynomials $\mathbb{P}_P[z] := \mathbb{C}[z]/\langle z^P - 1\rangle$. Here, $\mathbb{C}[z]$ denotes the ring of all polynomials with complex coefficients where addition and multiplication are defined in the standard way, and $\langle z^P - 1\rangle$ denotes the *ideal generated by* $z^P - 1$ which consists of all polynomial multiples of $z^P - 1$. Defining two polynomials in $\mathbb{C}[z]$ to be *equivalent* if their difference is divisible by $z^P - 1$, the quotient ring $\mathbb{P}_P[z]$ is the set of all corresponding equivalence classes. Essentially, $\mathbb{P}_P[z]$ is the set of all polynomials whose exponents of z are regarded modulo P; apart from this strangeness, polynomial addition and multiplication are defined in the usual manner. For example, recalling Example 10.1 in which $\varphi = a\delta_0 + b\delta_1 + c\delta_2 + d\delta_3$ is considered in $\ell(\mathbb{Z}_8)$, we have

$$(\mathrm{Z}\varphi)(z) = a + bz^{-1} + cz^{-2} + dz^{-3}.$$

Here, the exponents of z are only defined modulo 8, and as such we could have also written $(\mathrm{Z}\varphi)(z) = az^8 + bz^7 + cz^{14} + dz^{-11}$, for example.

Note that the Z-transform is a bijection from $\ell(\mathbb{Z}_P)$ onto $\mathbb{P}_P[z]$, with each signal y leading to a unique polynomial $(\mathrm{Z}y)(z)$ and vice versa. As with its infinite-dimensional cousin, the usefulness of this finite-dimensional Z-transform is the natural way in which it represents convolution as polynomial multiplication.

Proposition 10.3 *For any* $y, \varphi \in \ell(\mathbb{Z}_P)$, $[\mathrm{Z}(y * \varphi)](z) = (\mathrm{Z}y)(z)(\mathrm{Z}\varphi)(z)$.

Proof By definition,

$$\begin{aligned}(\mathrm{Z}y)(z)(\mathrm{Z}\varphi)(z) &= \sum_{p\in\mathbb{Z}_P} y[p]z^{-p} \sum_{p'\in\mathbb{Z}_P} \varphi\big[p'\big]z^{-p'} \\ &= \sum_{p\in\mathbb{Z}_P} \sum_{p'\in\mathbb{Z}_P} y[p]\varphi\big[p'\big]z^{-(p+p')}.\end{aligned}$$

For any fixed p, replacing the variable p' with $p'' = p + p'$ gives the result:

$$\begin{aligned}(\mathrm{Z}y)(z)(\mathrm{Z}\varphi)(z) &= \sum_{p\in\mathbb{Z}_P} \sum_{p''\in\mathbb{Z}_P} y[p]\varphi\big[p'' - p\big]z^{-p''} \\ &= \sum_{p''\in\mathbb{Z}_P} \Bigg(\sum_{m\in\mathbb{Z}_P} y[p]\varphi\big[p'' - p\big] \Bigg) z^{-p''} \\ &= \sum_{p''\in\mathbb{Z}_P} (y * \varphi)\big[p''\big]z^{-p''} \\ &= \big[\mathrm{Z}(y * \varphi)\big](z).\end{aligned}$$

□

For example, when $P = 8$, multiplying the Z-transform of a given signal y with that of $\varphi = a\delta_0 + b\delta_1 + c\delta_2 + d\delta_3$ and collecting common terms—identifying the exponents of z modulo 8—yields:

$$\begin{aligned}&(\mathrm{Z}y)(z)(\mathrm{Z}\varphi)(z) \\ &\quad= \big(y[0] + y[1]z^{-1} + y[2]z^{-2} + y[3]z^{-3} + y[4]z^{-4} + y[5]z^{-5} + y[6]z^{-6} \\ &\qquad + y[7]z^{-7}\big) \times \big(a + bz^{-1} + cz^{-2} + dz^{-3}\big) \\ &\quad= \big(ay[0] + by[7] + cy[6] + dy[5]\big) \\ &\qquad + \big(ay[1] + by[0] + cy[7] + dy[6]\big)z^{-1} \\ &\qquad + \big(ay[2] + by[1] + cy[0] + dy[7]\big)z^{-2} \\ &\qquad + \big(ay[3] + by[2] + cy[1] + dy[0]\big)z^{-3}\end{aligned}$$

$$
\begin{aligned}
&+\big(ay[4]+by[3]+cy[2]+dy[1]\big)z^{-4}\\
&+\big(ay[5]+by[4]+cy[3]+dy[2]\big)z^{-5}\\
&+\big(ay[6]+by[5]+cy[4]+dy[3]\big)z^{-6}\\
&+\big(ay[7]+by[6]+cy[5]+dy[4]\big)z^{-7}
\end{aligned}
$$

which is precisely the Z-transform of $y * \varphi$, directly computed in (10.5).

Now, since the exponents of z of a polynomial $(\mathrm{Z}y)(z)$ in $\mathbb{P}_P[z]$ are only well defined modulo P, one cannot hope to evaluate this polynomial over the entire complex plane. Indeed, the polynomial z^3 is equivalent to z^0 in $\mathbb{P}_3[z]$, but inserting $\zeta = -1$ into each yields distinct values of $(-1)^3 = -1$ and $(-1)^0 = 1$, respectively. In fact, the evaluation of a polynomial in the quotient ring $\mathbb{P}_P[z] = \mathbb{C}[z]/\langle z^P - 1\rangle$ is only well defined at points $\zeta \in \mathbb{C}$ which are roots of the generator of the ideal. That is, for $y \in \ell(\mathbb{Z}_P)$, $(\mathrm{Z}y)(\zeta)$ is only well defined at ζ which satisfy $\zeta^P - 1 = 0$, namely, the *Pth roots of unity* $\{\mathrm{e}^{2\pi \mathrm{i} p/P}\}_{p\in\mathbb{Z}_P}$. The evaluation of $\mathrm{Z}y$ at these points is a type of Fourier transform. To be precise, the *discrete Fourier transform* of $y \in \ell(\mathbb{Z}_P)$ is the operator $\mathrm{F}^* : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P)$ defined by

$$
\big(\mathrm{F}^* y\big)[p] := \frac{1}{\sqrt{P}}(\mathrm{Z}y)\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big) = \frac{1}{\sqrt{P}} \sum_{p'\in\mathbb{Z}_P} y\big[p'\big]\mathrm{e}^{-2\pi \mathrm{i} pp'/P}. \tag{10.8}
$$

The $\frac{1}{\sqrt{P}}$ term in (10.8) is a normalization factor which makes the Fourier transform a unitary operator. To see this, consider the *discrete Fourier basis* $\{\mathrm{f}_p\}_{p\in\mathbb{Z}_P}$ in $\ell(\mathbb{Z}_P)$ whose pth vector is $\mathrm{f}_p[p'] = \frac{1}{\sqrt{P}}\mathrm{e}^{2\pi \mathrm{i} pp'/P}$. We immediately observe that the Fourier transform is the analysis operator of this basis:

$$
\big(\mathrm{F}^* y\big)[p] = \frac{1}{\sqrt{P}} \sum_{p'\in\mathbb{Z}_P} y\big[p'\big]\mathrm{e}^{-2\pi \mathrm{i} pp'/P} = \sum_{p'\in\mathbb{Z}_P} y\big[p'\big]\big(\mathrm{f}_p\big[p'\big]\big)^* = \langle y, \mathrm{f}_p\rangle.
$$

Moreover, the geometric sum formula gives that this basis is orthonormal:

$$
\langle \mathrm{f}_p, \mathrm{f}_{p'}\rangle = \frac{1}{P} \sum_{p''\in\mathbb{Z}_P} \big[\mathrm{e}^{2\pi \mathrm{i}(p-p')/P}\big]^{p''} = \begin{cases} 1, & p = p' \mod P,\\ 0, & p \neq p'' \mod P. \end{cases}
$$

Being the analysis operator of an orthonormal basis, the Fourier transform is necessarily unitary, meaning that the *inverse Fourier transform* is given by the corresponding synthesis operator:

$$
(\mathrm{F}x)\big[p'\big] = \sum_{p\in\mathbb{Z}_p} x[p]\mathrm{f}_p\big[p'\big] = \frac{1}{\sqrt{P}} \sum_{p\in\mathbb{Z}_P} x[p]\mathrm{e}^{2\pi \mathrm{i} pp'/P}.
$$

The relationship between the Z-transform, the Fourier transform, and the Fourier basis is the key to understanding the eigenvalues and eigenvectors of a filter, and

moreover, the significance of the term "filter" itself. To be precise, evaluating the result of Proposition 10.3 at any pth root of unity gives

$$\begin{aligned}\big[\mathrm{F}^*(y*\varphi)\big]\big[p'\big] &= \frac{1}{\sqrt{P}}\big[\mathrm{Z}(y*\varphi)\big]\big(\mathrm{e}^{2\pi \mathrm{i} p'/P}\big)\\ &= \frac{1}{\sqrt{P}}(\mathrm{Z}y)\big(\mathrm{e}^{2\pi \mathrm{i} p'/P}\big)(\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p'/P}\big)\\ &= \big(\mathrm{F}^* y\big)\big[p'\big](\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p'/P}\big).\end{aligned}$$

For any fixed p, letting y be the pth element of the Fourier basis then gives

$$\big[\mathrm{F}^*(\mathrm{f}_p*\varphi)\big]\big[p'\big] = \big(\mathrm{F}^*\mathrm{f}_p\big)\big[p'\big](\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p'/P}\big) = \big\langle \mathrm{f}_p, \mathrm{f}'_p\big\rangle(\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p'/P}\big).$$

Since the Fourier basis is orthonormal, taking inverse Fourier transforms of this relation then yields

$$\begin{aligned}\mathrm{f}_p*\varphi &= \mathrm{F}\mathrm{F}^*(\mathrm{f}_p*\varphi)\\ &= \sum_{p'\in\mathbb{Z}_P}\big[\mathrm{F}^*(\mathrm{f}_p*\varphi)\big]\big[p'\big]\mathrm{f}_{p'}\\ &= \sum_{p'\in\mathbb{Z}_P}\big\langle \mathrm{f}_p, \mathrm{f}'_p\big\rangle(\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p'/P}\big)\mathrm{f}_{p'}\\ &= (\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big)\mathrm{f}_p.\end{aligned}$$

Thus, the operator $\Phi y := y*\varphi$ satisfies $\Phi \mathrm{f}_p = (\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i} p/P})\mathrm{f}_p$ and so f_p is an eigenvector for Φ with eigenvalue $(\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i} p/P})$. We summarize this result as follows.

Proposition 10.4 *If Φ is a filter on $\ell(\mathbb{Z}_P)$ with impulse response φ, then each member f_p of the Fourier basis is an eigenvector for Φ with eigenvalue $(\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i} p/P})$.*

Note that the above result gives $\Phi\mathrm{F} = \mathrm{F}D$, where D is a diagonal (pointwise multiplication) operator whose pth diagonal entry is $(\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i} p/P})$. Since F is unitary, this is equivalent to having $\Phi = \mathrm{F}D\mathrm{F}^*$; this is the famous result that every filter (circulant matrix) can be unitarily diagonalized using the Fourier transform. Moreover, as we now explain, this result justifies the use of the term "filter." Indeed, since F is unitary, every $y \in \ell(\mathbb{Z}_P)$ can be decomposed in terms of the Fourier basis:

$$y = \mathrm{F}\mathrm{F}^* y = \sum_{p\in\mathbb{Z}_p}\langle y, \mathrm{f}_p\rangle \mathrm{f}_p. \tag{10.9}$$

For any p, note that $\sqrt{P}\mathrm{f}_p[p'] = \mathrm{e}^{2\pi \mathrm{i} p p'/P}$ consists of the discrete samples of a complex wave $\mathrm{e}^{2\pi \mathrm{i} p t}$ of *frequency* p over $[0, 1]$. As such, the decomposition (10.9)

indicates how to break up the input signal y into an ensemble of waves, each wave with its own constant distinct frequency. The magnitude and argument of the complex scalar $\langle y, \mathrm{f}_p \rangle$ are the amplitude and phase shift of the pth wave, respectively. By Proposition 10.4, applying a filter Φ to (10.9) produces

$$\Phi y = \sum_{p \in \mathbb{Z}_p} \langle y, \mathrm{f}_p \rangle \Phi \mathrm{f}_p = \sum_{p \in \mathbb{Z}_p} \langle y, \mathrm{f}_p \rangle (\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big) \mathrm{f}_p. \tag{10.10}$$

By comparing (10.9) and (10.10) we see the effect of the filter Φ: each component wave f_p has had its magnitudes/phase-shift factor $\langle y, \mathrm{f}_p \rangle$ multiplied by $(\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i} p/P})$. In particular, for values p for which $(\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i} p/P})$ is large, the pth frequency component of Φy is much larger than that of y. Similarly, for values p for which $(\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i} p/P})$ is small, the corresponding frequencies will be less apparent in Φy than they are in y. Essentially, Φ acts like an equalizer on one's home stereo system: the input sound y is modified according to frequency to produce a more desirable output sound Φy. In particular, by carefully designing φ, one can create a filter Φ that amplifies the bass or another that amplifies the treble; such filters are referred to as *low-pass* or *high-pass*, respectively, as they allow low or high frequencies to pass through while filtering out undesired frequencies. Pairs of low- and high-pass filters are essential to the theory of wavelets, as discussed in greater detail in the following sections.

Having found the eigenvalues of an arbitrary filter, we then apply this result to determine the frame properties of a sequence of translates $\Phi = \{\mathrm{T}^p \varphi\}_{p \in \mathbb{Z}_P}$. Recall that the corresponding synthesis, analysis, and frame operators correspond to filtering with φ, $\tilde{\varphi}$, and $\varphi * \tilde{\varphi}$, respectively, where $\tilde{\varphi}[p] := (\varphi[-p])^*$. We already know that the eigenvalues of the synthesis filter are given by evaluating $(\mathrm{Z}\varphi)(z)$ at the pth roots of unity. As such, the eigenvalues of the analysis filter of φ—which is the synthesis filter of $\tilde{\varphi}$—are given by evaluating

$$(\mathrm{Z}\tilde{\varphi})(z) = \sum_{p \in \mathbb{Z}_p} \tilde{\varphi}[p] z^{-p} = \sum_{p \in \mathbb{Z}_p} \big(\varphi[-p]\big)^* z^{-p} = \sum_{p \in \mathbb{Z}_p} \big(\varphi[p]\big)^* z^p = \big(\mathrm{Z}\varphi^*\big)\big(z^{-1}\big)$$

at these same points. This can be further simplified by noting that $\zeta^* = \zeta^{-1}$ whenever $|\zeta| = 1$; as such for any pth root of unity ζ we have

$$(\mathrm{Z}\tilde{\varphi})(\zeta) = \sum_{p \in \mathbb{Z}_p} \big(\varphi[p]\big)^* \zeta^p = \left(\sum_{p \in \mathbb{Z}_p} \varphi[p] \zeta^{-p} \right)^* = \big[\mathrm{Z}\varphi(\zeta)\big]^*.$$

Thus, the eigenvalues of the analysis filter Φ^* are the conjugates of those of Φ. This fact along with Proposition 10.3 gives that the pth eigenvalue of $\Phi\Phi^*$ is

$$\begin{aligned} \big[\mathrm{Z}(\varphi * \tilde{\varphi})\big]\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big) &= (\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big)(\mathrm{Z}\tilde{\varphi})\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big) \\ &= (\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big)\big[(\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big)\big]^* \\ &= \big|(\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big)\big|^2. \end{aligned}$$

Thus, the optimal frame bounds for $\Phi = \{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}_P}$ are the extreme values of the squared modulus of the Z-transform of φ over all Pth roots of unity:

$$A = \min_{p\in\mathbb{Z}_P} \left|(\mathrm{Z}\varphi)\left(\mathrm{e}^{2\pi \mathrm{i}p/P}\right)\right|^2, \qquad B = \max_{p\in\mathbb{Z}_P} \left|(\mathrm{Z}\varphi)\left(\mathrm{e}^{2\pi \mathrm{i}p/P}\right)\right|^2,$$

meaning that such a Φ is a frame if and only if $(\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i}p/P}) \neq 0$ for all p. Moreover, Φ is a tight frame if and only if the Fourier transform of φ is flat, namely,

$$\frac{A}{P} = \frac{1}{P}\left|(\mathrm{Z}\varphi)\left(\mathrm{e}^{2\pi \mathrm{i}p/P}\right)\right|^2 = \left|\left(\mathrm{F}^*\varphi\right)[p]\right|^2$$

for all p. As such, since $\Phi = \{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}_P}$ consists of P vectors in a P-dimensional space, we see that the set of all translates of φ is an orthonormal basis for $\ell(\mathbb{Z}_P)$, namely Φ is unitary, if and only if $|(\mathrm{F}^*\varphi)[p]|^2 = \frac{1}{P}$ for all p.

Now recall that $\Phi = \{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}_P}$ can be written as $\Phi = \mathrm{F}D\mathrm{F}^*$, where the pth diagonal entry of D is $(\mathrm{Z}\varphi)(\mathrm{e}^{2\pi \mathrm{i}p/P})$. Whenever Φ is a frame, these diagonal entries are nonzero, and we see that the canonical dual frame Ψ is itself a filter:

$$\Psi = \left(\Phi\Phi^*\right)^{-1}\Phi = \left(\mathrm{F}D\mathrm{F}^*\mathrm{F}D^*\mathrm{F}^*\right)^{-1}\mathrm{F}D\mathrm{F}^* = \mathrm{F}\left(DD^*\right)^{-1}D\mathrm{F}^* = \mathrm{F}\left(D^*\right)^{-1}\mathrm{F}^*.$$

Writing $\Psi = \{\mathrm{T}^p\psi\}_{p\in\mathbb{Z}_P}$ where $\psi := \Psi\delta_0$ is the impulse response of Ψ, note that this canonical dual satisfies $\Phi\Psi^* = \mathrm{I}$. This means that filtering by $\tilde{\psi}$ may be undone by filtering by φ and vice versa; every $x \in \ell(\mathbb{Z}_P)$ can be decomposed as

$$x = \sum_{p\in\mathbb{Z}_P} \left\langle x, \mathrm{T}^p\psi\right\rangle \mathrm{T}^p\varphi.$$

Note that in this setting Φ is square and so $\Psi^* = \Phi^{-1}$; this also follows immediately from having $\Psi = \mathrm{F}(D^*)^{-1}\mathrm{F}^*$. Further note that the canonical dual filter ψ is obtained by *deconvolving* φ by its autocorrelation $\varphi * \tilde{\varphi}$:

$$\psi = \Psi\delta_0 = \left(\Phi\Phi^*\right)^{-1}\Phi\delta_0 = \left(\Phi\Phi^*\right)^{-1}\varphi.$$

In the Z-transform domain, such deconvolution corresponds to polynomial division:

$$(\mathrm{Z}\psi)(z) = \frac{(\mathrm{Z}\varphi)(z)}{[\mathrm{Z}(\varphi * \tilde{\varphi})](z)} = \frac{(\mathrm{Z}\varphi)(z)}{(\mathrm{Z}\varphi)(z)(\mathrm{Z}\varphi^*)(z^{-1})} = \frac{1}{(\mathrm{Z}\varphi^*)(z^{-1})}.$$

This division implies that ψ will often not be a "nice" filter even when φ is "nice." In particular, when working in the infinite-dimensional setting $\ell^2(\mathbb{Z})$, the reciprocal of a finite-term rational function is often an infinite-term Laurent series. A similar principle holds in the finite-dimensional setting of $\ell(\mathbb{Z}_P)$: when $\Phi = \{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}_P}$ is a frame and φ has a small number of taps, its canonical dual filter ψ will often have a large number of taps, the exception to this rule being when φ has one tap, meaning $(\mathrm{Z}\varphi)(z)$ is a monomial. Though algebraically nice, such one-tap filters leave much to be desired from the point of view of applications: their Fourier transforms are

flat, meaning they do not truly "filter" a signal in the sense of emphasizing certain frequencies over others. Rather, they simply delay the signal. That is, any single filter is unable to give us what we truly want: frequency selectivity *and* a small number of taps for both the filter and its dual. To obtain these capabilities, we thus generalize the preceding theory to operators that consist of multiple filters.

10.3 Filter Banks

A *filter bank* is an operator consisting of multiple filters. Such operators afford greater design possibilities than any single filter. Though long a subject of interest, filter banks became particularly popular during the heyday of wavelets [10]. Recall that the synthesis filter corresponding to φ is the synthesis operator of the set $\Phi = \{\mathrm{T}^p\varphi\}_{p\in\mathbb{Z}_P}$ of all translates of φ, namely, $\Phi y = y * \varphi$. Filter banks arise as a natural generalization of this idea: consider the synthesis operator of the set of all translates of multiple φ's.

To be precise, given a sequence of N desired impulse responses $\{\varphi_n\}_{n=0}^{N-1}$ in $\ell(\mathbb{Z}_P)$, we can generalize the theory of the previous section to systems of the form $\{\mathrm{T}^p\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$. Note that this system consists of NP vectors in P-dimensional space, and therefore necessarily has integer redundancy $\frac{NP}{P} = N$. In order to be more flexible with respect to redundancy, we further generalize these notions to systems of translates by a *subgroup* of all possible translates. Specifically, given any positive integer M and any $\{\varphi_n\}_{n=0}^{N-1}$ in $\ell(\mathbb{Z}_{MP})$, consider the set of all M-translates of the φ_n's, namely $\Phi = \{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$. Note here that we have changed the underlying space from $\ell(\mathbb{Z}_P)$ to $\ell(\mathbb{Z}_{MP})$; in the theory that follows, the spacing of the translates M must divide the length of the signal, and not making this change would burden us with writing $\frac{P}{M}$ instead of simply P.

The synthesis operator of $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ is an operator over the NP-dimensional space $[\ell(\mathbb{Z}_P)]^N$, namely, the direct sum of N copies of $\ell(\mathbb{Z}_P)$. We write any Y in $[\ell(\mathbb{Z}_P)]^N$ as $Y = \{y_n\}_{n=0}^{N-1}$ where y_n lies in $\ell(\mathbb{Z}_P)$ for all n. Under this notation, the synthesis operator $\Phi : [\ell(\mathbb{Z}_P)]^N \to \ell(\mathbb{Z}_{MP})$ of $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ is given by

$$\left(\Phi\{y_n\}_{n=0}^{N-1}\right)[k] = \left(\sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P} y_n[p]\mathrm{T}^{Mp}\varphi_n\right)[k] = \sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P} y_n[p]\varphi_n[k-Mp]. \tag{10.11}$$

We want to write this expression for Φ in terms of convolutions in order to take advantage of the rich theory of filters. Here, the issue is that the argument "p" of y_n in (10.11) does not match the "Mp" term in the argument of φ_n. The solution to this problem is to *upsample* y by a factor of M, namely, to stretch the P-periodic signal y to an MP-periodic signal by inserting $M-1$ values of 0 between any two values of y. To be precise, the *upsampling by M* operator on $\ell(\mathbb{Z}_P)$ is $\uparrow : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_{MP})$

defined by

$$(\uparrow y)[k] := \begin{cases} y[k/M], & M \mid k, \\ 0, & M \nmid k. \end{cases}$$

This concept in hand, we return to the simplification of (10.11). Making the change of variables $k' = Mp$ gives

$$\begin{aligned} \big(\Phi\{y_n\}_{n=0}^{N-1}\big)[k] &= \sum_{n=0}^{N-1} \sum_{\substack{k' \in \mathbb{Z}_{MP} \\ M \mid k'}} y_n\big[k'/M\big]\varphi_n\big[k-k'\big] \\ &= \sum_{n=0}^{N-1} \sum_{k' \in \mathbb{Z}_{MP}} (\uparrow y_n)\big[k'\big]\varphi_n\big[k-k'\big] \\ &= \sum_{n=0}^{N-1} \big((\uparrow y_n) * \varphi_n\big)[k]. \end{aligned} \tag{10.12}$$

With (10.12), we turn to writing the analysis operator of $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ in terms of convolutions. Specifically, $\Phi^* : \ell(\mathbb{Z}_{MP}) \to [\ell(\mathbb{Z}_P)]^N$ is given by

$$\begin{aligned} \big(\Phi^* x\big)_n[p] &= \big\langle x, \mathrm{T}^{Mp}\varphi_n \big\rangle_{\ell(\mathbb{Z}_{MP})} \\ &= \sum_{k \in \mathbb{Z}_{MP}} x[k]\big[\big(\mathrm{T}^{Mp}\varphi_n\big)[k]\big]^* \\ &= \sum_{k \in \mathbb{Z}_{MP}} x[k]\tilde{\varphi}_n[Mp-k] \\ &= (x * \tilde{\varphi}_n)[Mp] \\ &= \big[\downarrow (x * \tilde{\varphi}_n)\big][p], \end{aligned} \tag{10.13}$$

where $\downarrow : \ell(\mathbb{Z}_{MP}) \to \ell(\mathbb{Z}_P)$ is the *downsampling* operator $\downarrow : \ell(\mathbb{Z}_{MP}) \to \ell(\mathbb{Z}_P)$ defined by $(\downarrow x)[p] = x[Mp]$. Downsampling by M transforms an MP-periodic signal into a P-periodic signal by only retaining those indices which are divisible by M. Collecting (10.12) and (10.13), we make the following definitions.

Definition 10.1 Given filters $\{\varphi_n\}_{n=0}^{N-1} \subseteq \ell(\mathbb{Z}_{MP})$, the corresponding *synthesis filter bank* is the operator $\Phi : [\ell(\mathbb{Z}_P)]^N \to \ell(\mathbb{Z}_{MP})$ defined by

$$\Phi\{y_n\}_{n=0}^{N-1} = \sum_{n=0}^{N-1} (\uparrow y_n) * \varphi_n.$$

Meanwhile, the *analysis filter bank* $\Phi^* : \ell(\mathbb{Z}_{MP}) \to [\ell(\mathbb{Z}_P)]^N$ is defined by

$$\Phi^* x = \big\{\downarrow (x * \tilde{\varphi}_n)\big\}_{n=0}^{N-1}.$$

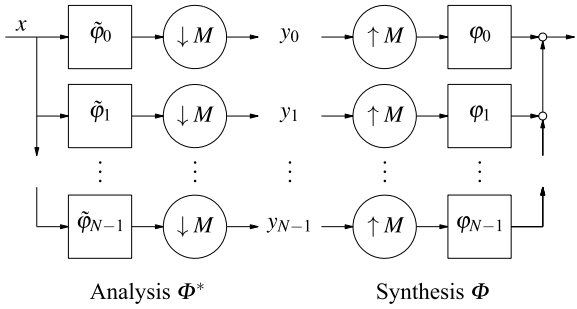


Fig. 10.1 An N-channel filter bank with a downsampling rate of M. The analysis filter bank Φ^* computes inner products of a given input signal x in $\ell(\mathbb{Z}_{MP})$ with the M-translates of each φ_n, resulting in the output signals $\Phi^* x = \{\downarrow (x * \tilde{\varphi}_n)\}_{n=0}^{N-1}$ where each signal $\downarrow (x * \tilde{\varphi}_n)$ lies in $\ell(\mathbb{Z}_P)$. Meanwhile, the synthesis filter bank Φ forms a linear combination of the M-translates of the φ_n's using the values of some $\{y_n\}_{n=0}^{N-1}$ in $[\ell(\mathbb{Z}_{MP})]^N$ as coefficients: $\Phi\{y_n\}_{n=0}^{N-1} = \sum_{n=0}^{N-1} (\uparrow y_n) * \varphi_n$. Regarding frame theory, these analysis and synthesis filter banks are the analysis and synthesis operators of $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$, and so the composition of these operators is the corresponding frame operator. In the next section, we use the polyphase representation of this filter bank to give an efficient method for computing the frame bounds of such a system

A diagram depicting these operations is given in Fig. 10.1.

Example 10.2 We conclude this section with some examples of filter banks. Indeed, let's first consider $M = N = 2$ and build from Example 10.1, considering two 4-tap filters in $\ell(\mathbb{Z}_8)$:

$$\varphi_0 = a\delta_0 + b\delta_1 + c\delta_2 + d\delta_4, \qquad \varphi_1 = e\delta_0 + f\delta_1 + g\delta_2 + h\delta_4.$$

Here, the synthesis filter bank is $\Phi : [\ell(\mathbb{Z}_4)]^2 \to \ell(\mathbb{Z}_8)$:

$$\Phi\{y_0, y_1\} = (\uparrow y_0) * \varphi_0 + (\uparrow y_1) * \varphi_1.$$

Writing the operation of filtering by φ_0 as the circulant matrix (10.4) gives

$$(\uparrow y_0) * \varphi_0 = \begin{bmatrix} a & & & & & d & c & b \\ b & a & & & & & d & c \\ c & b & a & & & & & d \\ d & c & b & a & & & & \\ & d & c & b & a & & & \\ & & d & c & b & a & & \\ & & & d & c & b & a & \\ & & & & d & c & b & a \end{bmatrix} \begin{bmatrix} y_0[0] \\ 0 \\ y_0[1] \\ 0 \\ y_0[2] \\ 0 \\ y_0[3] \\ 0 \end{bmatrix}$$

$$= \begin{bmatrix} a & & & c \\ b & & & d \\ c & a & & \\ d & b & & \\ & c & a & \\ & d & b & \\ & & c & a \\ & & d & b \end{bmatrix} \begin{bmatrix} y_0[0] \\ y_0[1] \\ y_0[2] \\ y_0[3] \end{bmatrix}. \tag{10.14}$$

Writing $(\uparrow_2 y_1) * \varphi_1$ similarly and summing the results gives

$$\Phi\{y_0, y_1\} = (\uparrow y_0) * \varphi_0 + (\uparrow y_1) * \varphi_1 = \left[\begin{array}{cccc|cccc} a & & & c & e & & & g \\ b & & & d & f & & & h \\ c & a & & & g & e & & \\ d & b & & & h & f & & \\ & c & a & & & g & e & \\ & d & b & & & h & f & \\ & & c & a & & & g & e \\ & & d & b & & & h & f \end{array}\right] \left[\begin{array}{c} y_0[0] \\ y_0[1] \\ y_0[2] \\ y_0[3] \\ \hline y_1[0] \\ y_1[1] \\ y_1[2] \\ y_1[3] \end{array}\right].$$

This makes sense, since the columns of Φ are supposed to be the frame vectors $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$, namely the 4 even translates of φ_0 and φ_1. If we add a third filter $\varphi_3 = i\delta_0 + j\delta_1 + k\delta_2 + l\delta_4$ to this filter bank, meaning we now have $M = 2$, $N = 3$, and $P = 4$, our synthesis operator becomes 8×12:

$$\Phi = \left[\begin{array}{cccc|cccc|cccc} a & & & c & e & & & g & i & & & k \\ b & & & d & f & & & h & j & & & l \\ c & a & & & g & e & & & k & i & & \\ d & b & & & h & f & & & l & j & & \\ & c & a & & & g & e & & & k & i & \\ & d & b & & & h & f & & & l & j & \\ & & c & a & & & g & e & & & k & i \\ & & d & b & & & h & f & & & l & j \end{array}\right].$$

For a general system of form $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$, the matrix representing the synthesis filter bank is of size $MP \times NP$, namely the concatenation of N blocks of size $MP \times P$, each block containing all M-translates of a given φ_n in $\ell(\mathbb{Z}_{MP})$. In order to determine the optimal frame bounds for such systems, we must necessarily compute the eigenvalues of $\Phi\Phi^*$, namely, the singular values of Φ. At first glance, this does not seem like an easy problem. At the same time, the fact that these matrices have a quasi-circulant structure gives reason to believe that they can be better understood via Z-transforms and Fourier transforms; this is the subject of the next section.

10.4 The Polyphase Representation

The previous sections discussed the relationship between filter banks and frame theory. Frame theorists often want frames with certain desirable properties, such as tightness or incoherence. As one might guess, there are also desirable filter bank properties. For example, one might wish to design analysis and synthesis filter banks with the property that the synthesis filter bank reconstructs a signal that was input into the analysis filter bank—see the discussion on *perfect reconstruction filter banks* after Theorem 10.2 below. In addition, we might want the filters in these filter banks to have few taps; that is, we'd like to convolve with vectors of small support, as this would enable faster implementation.

In this section, we introduce a representation of filter banks, called the *polyphase representation*, which is quite useful in designing filter banks with certain properties. Though this representation was first introduced to study nonredundant filter banks [22, 24, 25], it wasn't long before it was adapted to the case of filter bank frames [4, 8]. The main results of this section are all straightforward generalizations of results from [4, 8, 24] to the finite-dimensional setting [7].

Definition 10.2 For any $\varphi \in \ell(\mathbb{Z}_{MP})$, the *polyphase vector* of φ with respect to M is the $M \times 1$ vector of polynomials:

$$\varphi(z) = \begin{bmatrix} \varphi^{(0)}(z) \\ \varphi^{(1)}(z) \\ \vdots \\ \varphi^{(M-1)}(z) \end{bmatrix},$$

where each entry of $\varphi(z)$ is defined to be a Z-transform of the restriction of φ to a coset of $\mathbb{Z}_{MP}$ with respect to the subgroup $M\mathbb{Z}_P$:

$$\varphi^{(m)}(z) := \sum_{p \in \mathbb{Z}_P} \varphi[m + Mp] z^{-p} = \big[\mathrm{Z}\big(\downarrow \mathrm{T}^{-m}\varphi\big)\big](z). \tag{10.15}$$

For example, when $M = 2$ and $P = 4$, the polyphase vector of some 4-tap filter φ in $\ell(\mathbb{Z}_8)$ of the form $\varphi = a\delta_0 + b\delta_1 + c\delta_2 + d\delta_3$ is the 2×1 vector consisting of the Z-transforms of the even and odd parts of φ:

$$\varphi(z) = \begin{bmatrix} \varphi^{(0)}(z) \\ \varphi^{(1)}(z) \end{bmatrix} = \begin{bmatrix} a + cz^{-1} \\ b + dz^{-1} \end{bmatrix}.$$

This section is dedicated to explaining why this polyphase representation is the key to understanding the frame properties of filter banks.

Formally speaking, since $\varphi \in \ell(\mathbb{Z}_{MP})$ then for any m we have $\downarrow \mathrm{T}^{-m}\varphi \in \ell(\mathbb{Z}_P)$, and so its Z-transform lies in the quotient polynomial ring $\mathbb{P}_P[z] := \mathbb{C}[z]/\langle z^P - 1\rangle$ discussed in the previous section. As such, the polyphase vector $\varphi(z)$ lies in the Cartesian product of M copies of $\mathbb{P}_P[z]$, denoted $\mathbb{P}_P^M[z]$. Letting $\mathbb{T}_P$ denote the *discrete torus* that consists of the Pth roots of unity, one can show that $\mathbb{P}_P^M[z]$ is a

Hilbert space under the inner product:

$$\begin{aligned}\big\langle \varphi(z), \psi(z)\big\rangle_{\mathbb{P}_P^M[z]} &:= \frac{1}{P}\sum_{\zeta\in\mathbb{T}_P}\big\langle \varphi(\zeta), \psi(\zeta)\big\rangle_{\mathbb{C}^M}\\ &= \frac{1}{P}\sum_{p\in\mathbb{Z}_P}\big\langle \varphi\big(\mathrm{e}^{2\pi \mathrm{i}p/P}\big), \psi\big(\mathrm{e}^{2\pi \mathrm{i}p/P}\big)\big\rangle_{\mathbb{C}^M}\\ &= \frac{1}{P}\sum_{p\in\mathbb{Z}_P}\sum_{m\in\mathbb{Z}_M}\varphi^{(m)}\big(\mathrm{e}^{2\pi \mathrm{i}p/P}\big)\big[\psi^{(m)}\big(\mathrm{e}^{2\pi \mathrm{i}p/P}\big)\big]^*. \end{aligned} \tag{10.16}$$

In fact, as the next result shows, the space $\mathbb{P}_P^M[z]$ is isometric to $\ell(\mathbb{Z}_{MP})$ under the mapping $\varphi \mapsto \varphi(z)$, which is known as a *Zak transform*.

Proposition 10.5 *The Zak transform $\varphi \mapsto \varphi(z)$ is an isometry from $\ell(\mathbb{Z}_{MP})$ onto $\mathbb{P}_P^M[z]$.*

Proof For any $\varphi(z), \psi(z) \in \mathbb{P}_P^M[z]$, the definition (10.16) of the polyphase inner product gives

$$\big\langle \varphi(z), \psi(z)\big\rangle_{\mathbb{P}_P^M[z]} = \frac{1}{P}\sum_{p\in\mathbb{Z}_P}\sum_{m\in\mathbb{Z}_M}\varphi^{(m)}\big(\mathrm{e}^{2\pi \mathrm{i}p/P}\big)\big[\psi^{(m)}\big(\mathrm{e}^{2\pi \mathrm{i}p/P}\big)\big]^*. \tag{10.17}$$

Considering the Z-transform representation of the polyphase entries (10.15), we use (10.8) to rewrite $\varphi^{(m)}(\mathrm{e}^{2\pi \mathrm{i}p/P})$ in terms of Fourier transforms:

$$\varphi^{(m)}\big(\mathrm{e}^{2\pi \mathrm{i}p/P}\big) = \big[\mathrm{Z}\big(\downarrow \mathrm{T}^{-m}\varphi\big)\big]\big(\mathrm{e}^{2\pi \mathrm{i}p/P}\big) = \sqrt{P}\big[\mathrm{F}^*\big(\downarrow \mathrm{T}^{-m}\varphi\big)\big](p). \tag{10.18}$$

Obtaining a similar expression for $\psi^{(m)}(\mathrm{e}^{2\pi \mathrm{i}p/P})$, we substitute this and (10.18) into (10.17) to get

$$\big\langle \varphi(z), \psi(z)\big\rangle_{\mathbb{P}_P^M[z]} = \sum_{p\in\mathbb{Z}_P}\sum_{m\in\mathbb{Z}_M}\big[\mathrm{F}^*\big(\downarrow \mathrm{T}^{-m}\varphi\big)\big](p)\big[\big[\mathrm{F}^*\big(\downarrow \mathrm{T}^{-m}\psi\big)\big](p)\big]^*. \tag{10.19}$$

Switching the order of summation in (10.19) gives a sum of inner products in $\ell(\mathbb{Z}_P)$:

$$\big\langle \varphi(z), \psi(z)\big\rangle_{\mathbb{P}_P^M[z]} = \sum_{m\in\mathbb{Z}_M}\big\langle \mathrm{F}^*\big(\downarrow \mathrm{T}^{-m}\varphi\big), \mathrm{F}^*\big(\downarrow \mathrm{T}^{-m}\psi\big)\big\rangle_{\ell(\mathbb{Z}_P)}.$$

Here, we use the fact that the Fourier transform is unitary to get

$$\begin{aligned}\big\langle \varphi(z), \psi(z)\big\rangle_{\mathbb{P}_P^M[z]} &= \sum_{m\in\mathbb{Z}_M}\big\langle \downarrow \mathrm{T}^{-m}\varphi, \downarrow \mathrm{T}^{-m}\psi\big\rangle_{\ell(\mathbb{Z}_P)}\\ &= \sum_{m\in\mathbb{Z}_M}\sum_{p\in\mathbb{Z}_P}\big(\downarrow \mathrm{T}^{-m}\varphi\big)[p]\big[\big(\downarrow \mathrm{T}^{-m}\psi\big)[p]\big]^*\\ &= \sum_{m\in\mathbb{Z}_M}\sum_{p\in\mathbb{Z}_P}\varphi[m+Mp]\big(\psi[m+Mp]\big)^*. \end{aligned}$$

Finally, we perform a change of variables $k = m + Mp$ to obtain our claim:

$$\big\langle \varphi(z), \psi(z) \big\rangle_{\mathbb{P}_P^M[z]} = \sum_{k \in \mathbb{Z}_{MP}} \varphi[k] \big(\psi[k]\big)^* = \langle \varphi, \psi \rangle_{\mathbb{Z}_{MP}}.$$

□

Filter banks, like those depicted in Fig. 10.1, consist of N filters in $\ell(\mathbb{Z}_{MP})$, denoted $\{\varphi_n\}_{n=0}^{N-1}$. Taking the polyphase transform of each results in N polynomial vectors $\{\varphi_n(z)\}_{n=0}^{N-1}$ lying in $\mathbb{P}_P^M[z]$. As we shall see, many of the frame properties of the system $\Phi = \{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ can be understood in terms of the synthesis operator of $\Phi(z) = \{\varphi_n(z)\}_{n=0}^{N-1}$, namely, the polyphase matrix.

Definition 10.3 Given a sequence of filters $\{\varphi_n\}_{n=0}^{N-1} \subseteq \ell(\mathbb{Z}_{MP})$, the associated *polyphase matrix* is the $M \times N$ matrix whose columns are the polyphase vectors $\{\varphi_n(z)\}_{n=0}^{N-1}$:

$$\Phi(z) := \begin{bmatrix} \varphi_0^{(0)}(z) & \varphi_1^{(0)}(z) & \cdots & \varphi_{N-1}^{(0)}(z) \\ \varphi_0^{(1)}(z) & \varphi_1^{(1)}(z) & \cdots & \varphi_{N-1}^{(1)}(z) \\ \vdots & \vdots & \ddots & \vdots \\ \varphi_0^{(M-1)}(z) & \varphi_1^{(M-1)}(z) & \cdots & \varphi_{N-1}^{(M-1)}(z) \end{bmatrix}.$$

The next result, which gives a polynomial-domain interpretation of the operation of a synthesis filter bank, provides the first hints at the usefulness of the polyphase representation.

Theorem 10.1 *Let Φ be the synthesis operator (filter bank) of $\{\mathrm{T}^p\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ and let $\Phi(z)$ be the polyphase matrix of $\{\varphi_n(z)\}_{n=0}^{N-1}$. Then*

$$x = \Phi Y \quad \Longleftrightarrow \quad x(z) = \Phi(z)Y(z)$$

where $x(z)$ in $\mathbb{P}_P^M[z]$ is the $M \times 1$ polyphase vector of $x \in \ell(\mathbb{Z}_{MP})$ and $Y(z)$ denotes the $N \times 1$ vector in $\mathbb{P}_P^N[z]$ whose nth component is the Z-transform of y_n, where $Y = \{y_n\}_{n=0}^{N-1} \in [\ell(\mathbb{Z}_P)]^N$.

To prove this fact, we first prove the following results involving Z-transforms.

Proposition 10.6

(a) $[\mathrm{Z}(\uparrow y)](z) = (\mathrm{Z}y)(z^M)$ *for any* $y \in \ell(\mathbb{Z}_P)$.
(b) $(\mathrm{Z}x)(z) = \sum_{m\in\mathbb{Z}_M} z^{-m} x^{(m)}(z^M)$ *for any* $x \in \ell(\mathbb{Z}_{MP})$.

Proof For (a), note that the definition of upsampling gives

$$\big[\mathrm{Z}(\uparrow y)\big](z) = \sum_{k\in\mathbb{Z}} (\uparrow y)[k] z^{-k} = \sum_{\substack{k\in\mathbb{Z}_{MP} \\ M|k}} y[k/M] z^{-k}.$$

Performing the change of variables $k = Mp$ gives the result

$$\big[\mathrm{Z}(\uparrow y)\big](z) = \sum_{p\in\mathbb{Z}_P} y[p]z^{-Mp} = \sum_{p\in\mathbb{Z}_P} y[p]\big(z^M\big)^{-p} = (\mathrm{Z}y)\big(z^M\big).$$

For (b), making the change of variables $k = m + Mp$ gives

$$\begin{aligned}(\mathrm{Z}x)(z) &= \sum_{k\in\mathbb{Z}_{MP}} x[k]z^{-k}\\ &= \sum_{m\in\mathbb{Z}_M}\sum_{p\in\mathbb{Z}_P} x[m+Mp]z^{-(m+Mp)}\\ &= \sum_{m\in\mathbb{Z}_M} z^{-m}\sum_{p\in\mathbb{Z}_p} x[m+Mp]\big(z^M\big)^{-p}\\ &= \sum_{m\in\mathbb{Z}_M} z^{-m}x^{(m)}\big(z^M\big).\end{aligned}$$

Alternatively, one may prove (b) first, of which (a) is a special case. □

Proof of Theorem 10.1 Note that $x = \Phi Y$ if and only if $(\mathrm{Z}x)(z) = (\Phi Y)(z)$. Here, $(\mathrm{Z}x)(z)$ is given by Proposition 10.6(b). Meanwhile, Definition 10.1, the linearity of the Z-transform, and Propositions 10.3 and 10.6 give

$$\begin{aligned}(\mathrm{Z}\Phi Y)(z) &= \left(\mathrm{Z}\sum_{n=0}^{N-1}(\uparrow y_n) * \varphi_n\right)(z)\\ &= \sum_{n=0}^{N-1}\big[\mathrm{Z}(\uparrow y_n)\big](z)(\mathrm{Z}\varphi_n)(z)\\ &= \sum_{n=0}^{N-1}(\mathrm{Z}y_n)\big(z^M\big)(\mathrm{Z}\varphi_n)(z).\end{aligned}$$

We continue to rewrite $(\mathrm{Z}\Phi Y)(z)$ by applying Proposition 10.6(b) to $(\mathrm{Z}\varphi_n)(z)$:

$$\begin{aligned}(\mathrm{Z}\Phi y)(z) &= \sum_{n=0}^{N-1}(\mathrm{Z}y_n)\big(z^M\big)\sum_{m\in\mathbb{Z}_M} z^{-m}\varphi_n^{(m)}\big(z^M\big)\\ &= \sum_{m\in\mathbb{Z}_M} z^{-m}\sum_{n=0}^{N-1}\varphi_n^{(m)}\big(z^M\big)(\mathrm{Z}y_n)\big(z^M\big)\\ &= \sum_{m\in\mathbb{Z}_M} z^{-m}\big[\Phi\big(z^M\big)Y\big(z^M\big)\big]_m.\end{aligned}$$

As such, $x = \Phi Y$ if and only if

$$\sum_{m\in\mathbb{Z}_M} z^{-m} x^{(m)}\big(z^M\big) = (\mathrm{Z}x)(z) = (\mathrm{Z}\Phi y)(z) = \sum_{m\in\mathbb{Z}_M} z^{-m}\big[\Phi\big(z^M\big)Y\big(z^M\big)\big]_m.$$

Considering only those exponents which are equal modulo m, we thus have $x = \Phi Y$ if and only if $x^{(m)}(z^M) = [\Phi(z^M)Y(z^M)]_m$ for all m, meaning $x(z^M) = \Phi(z^M)Y(z^M)$. To conclude, note that the Z-transforms of x and φ invoked here lie in the ring $\mathbb{P}_{MP}[z] = \mathbb{C}[z]/\langle z^{MP} - 1\rangle$; having $x(z^M) = \Phi(z^M)Y(z^M)$ in this ring is equivalent to having $x(z) = \Phi(z)Y(z)$ in the ring $\mathbb{P}_P[z] = \mathbb{C}[z]/\langle z^P - 1\rangle$. □

We can also prove a result which is analogous to Theorem 10.1 for analysis filter banks. Here, the *para-adjoint* $\Phi^*(z)$ is the matrix of polynomials obtained by taking the conjugate transpose of $\Phi(z)$, where the variable z is regarded as an element of the unit circle $\mathbb{T} := \{\zeta \in \mathbb{C} : |\zeta| = 1\}$ and so $z^* := z^{-1}$. Formally speaking, $\Phi^*(z)$ is an $N \times M$ matrix of polynomials, each entry lying in $\mathbb{P}_P[z]$, whose (n, m)th entry is

$$\big[\Phi^*(z)\big]_{n,m} := \big(\varphi_n^*\big)^{(m)}\big(z^{-1}\big) = \sum_{p\in\mathbb{Z}_P} \varphi_n^*[m + Mp]z^p.$$

Theorem 10.2 *Let Φ^* be the analysis operator (filter bank) of $\{\mathrm{T}^p\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ and let $\Phi(z)$ be the polyphase matrix of $\{\varphi_n(z)\}_{n=0}^{N-1}$. Then*

$$Y = \Phi^* x \quad\Longleftrightarrow\quad Y(z) = \Phi^*(z)x(z).$$

Proof The nth entry of $\Phi^*(z)x(z)$ is

$$\begin{aligned}\big[\Phi^*(z)x(z)\big]_n &= \sum_{m\in\mathbb{Z}_M} \big[\Phi^*(z)\big]_{n,m} x^{(m)}(z)\\ &= \sum_{m\in\mathbb{Z}_M}\sum_{p'\in\mathbb{Z}_P} \varphi_n^*\big[m + Mp'\big]z^{p'} \sum_{p''\in\mathbb{Z}_P} x\big[m + Mp''\big]z^{-p''}\\ &= \sum_{m\in\mathbb{Z}_M}\sum_{p'\in\mathbb{Z}_P}\sum_{p''\in\mathbb{Z}_P} \varphi_n^*\big[m + Mp'\big]x\big[m + Mp''\big]z^{-(p''-p')}.\end{aligned}$$

Making two changes of variables, $p = p'' - p'$ and then $k = m + Mp''$, gives

$$\begin{aligned}\big[\Phi^*(z)x(z)\big]_n &= \sum_{p\in\mathbb{Z}_P}\Bigg(\sum_{m\in\mathbb{Z}_M}\sum_{p''\in\mathbb{Z}_P} \varphi_n^*\big[m + Mp'' - Mp\big]x\big[m + Mp''\big]\Bigg)z^{-p}\\ &= \sum_{p\in\mathbb{Z}_P}\Bigg(\sum_{k\in\mathbb{Z}_{MP}} \tilde{\varphi}_n[Mp - k]x[k]\Bigg)z^{-p}\\ &= \sum_{p\in\mathbb{Z}_P} (x * \tilde{\varphi}_n)[Mp]z^{-p}\end{aligned}$$

$$= \sum_{p\in\mathbb{Z}_P} \big[\downarrow (x * \tilde{\varphi}_n)\big][p] z^{-p}$$
$$= \big\{\mathrm{Z}\big[\downarrow (x * \tilde{\varphi}_n)\big]\big\}(z).$$

In particular, $Y(z) = \Phi^*(z)x(z)$ if and only if $(\mathrm{Z}y_n)(z) = \{\mathrm{Z}[\downarrow (x * \tilde{\varphi}_n)]\}(z)$ for all n; this occurs precisely when $y_n = \downarrow (x * \tilde{\varphi}_n)$ for all n, namely, when $Y = \Phi^* x$. □

Theorems 10.1 and 10.2 tell us something very interesting about the polyphase representation: the polyphase matrix of an analysis filter bank behaves as an analysis operator of sorts in polyphase space, and similarly for synthesis. As the remainder of this section will show, there are several properties of Φ that the polyphase representation preserves in a similar way.

For example, with Theorems 10.1 and 10.2, we can characterize an important class of filter banks. We say the pair (Ψ^*, Φ) is a *perfect reconstruction filter bank (PRFB)* if $\Phi\Psi^* = \mathrm{I}$. This is equivalent to having the corresponding frames be duals of each other. PRFBs are useful because the synthesis filter bank can be used to reconstruct whatever signal was input into the analysis filter bank. Note that combining Theorems 10.1 and 10.2 gives that

$$\Phi\Psi^* x = x \quad \text{if and only if} \quad \Phi(z)\Psi^*(z)x(z) = x(z),$$

and so we have a polyphase characterization of PRFBs: $\Phi(z)\Psi^*(z) = \mathrm{I}$. The polyphase representation can also be used to characterize other useful properties of filter banks. Before we state them, we prove the following lemma.

Lemma 10.1 *For any* $x, \varphi \in \ell(\mathbb{Z}_{MP})$,

$$\big\langle x\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big), \varphi\big(\mathrm{e}^{2\pi \mathrm{i} p/P}\big)\big\rangle_{\mathbb{C}^M} = \sqrt{P}\big(\mathrm{F}^*\langle x, \mathrm{T}^{M\bullet}\varphi\rangle_{\ell(\mathbb{Z}_{MP})}\big)[p]. \tag{10.20}$$

Proof We first show that the polyphase representation of $\mathrm{T}^{Mp}\varphi$ is $z^{-p}\varphi(z)$. For any $p \in \mathbb{Z}_P$ and $m \in \mathbb{Z}_M$,

$$\big(\mathrm{T}^{Mp}\varphi\big)^{(m)}(z) = \sum_{p'\in\mathbb{Z}_P} \big(\mathrm{T}^{Mp}\varphi\big)\big[m + Mp'\big] z^{-p'} = \sum_{p'\in\mathbb{Z}_P} \varphi\big[m + M\big(p' - p\big)\big] z^{-p'}. \tag{10.21}$$

Letting $p'' = p' - p$ in (10.21) then gives

$$\big(\mathrm{T}^{Mp}\varphi\big)^{(m)}(z) = \sum_{p''\in\mathbb{Z}_P} \varphi\big[m + Mp''\big] z^{-(p''+p)} = z^{-p}\varphi^{(m)}(z). \tag{10.22}$$

Since (10.22) holds for all $m \in \mathbb{Z}_M$, we have that $(\mathrm{T}^{Mp}\varphi)(z) = z^{-p}\varphi(z)$ for all $p \in \mathbb{Z}_P$. This fact, along with Proposition 10.5, gives

$$\big\langle x, \mathrm{T}^{Mp}\varphi\big\rangle_{\ell(\mathbb{Z}_{MP})} = \big\langle x(z), \big(\mathrm{T}^{Mp}\varphi\big)(z)\big\rangle_{\mathbb{P}_P^M[z]} = \big\langle x(z), z^{-p}\varphi(z)\big\rangle_{\mathbb{P}_P^M[z]}.$$

By the definitions of the inner product on $\mathbb{P}_P^M[z]$ and the inverse Fourier transform,

$$\begin{aligned}\langle x, \mathrm{T}^{Mp}\varphi\rangle_{\ell(\mathbb{Z}_{MP})} &= \frac{1}{P}\sum_{p'\in\mathbb{Z}_P}\langle x(\mathrm{e}^{2\pi \mathrm{i}p'/P}), \mathrm{e}^{-2\pi \mathrm{i}pp'/P}\varphi(\mathrm{e}^{2\pi \mathrm{i}p'/P})\rangle_{\mathbb{C}^M}\\ &= \frac{1}{P}\sum_{p'\in\mathbb{Z}_P}\langle x(\mathrm{e}^{2\pi \mathrm{i}p'/P}), \varphi(\mathrm{e}^{2\pi \mathrm{i}p'/P})\rangle_{\mathbb{C}^M}\,\mathrm{e}^{2\pi \mathrm{i}pp'/P}\\ &= \frac{1}{\sqrt{P}}\big(\mathrm{F}\langle x(\mathrm{e}^{2\pi \mathrm{i}\bullet/P}), \varphi(\mathrm{e}^{2\pi \mathrm{i}\bullet/P})\rangle_{\mathbb{C}^M}\big)[p], \end{aligned} \tag{10.23}$$

where "$\bullet$" denotes the variable argument of a given function. Taking Fourier transforms of (10.23) and multiplying by $\sqrt{P}$ gives the result (10.20). □

Perhaps the most straightforward PRFB arises from a unitary filter bank, in which $M = N$ and the synthesis filter bank operator is the inverse of the analysis filter bank operator, $\Phi = (\Psi^*)^{-1} = \Psi$. In this case, showing that $\Phi\Psi^* = \Phi\Phi^* = \mathrm{I}$ is equivalent to showing that $\Phi^*\Phi = \mathrm{I}$, i.e., showing that the columns of the matrix Φ are orthonormal. The following result—a finite-dimensional version of a well-known result [24]—characterizes unitary filter banks; it says that the columns of Φ are orthonormal precisely when the columns of the corresponding polyphase matrix $\Phi(\zeta)$ are orthonormal for every ζ in the Pth roots of unity $\mathbb{T}_P := \{\zeta \in \mathbb{C} : \zeta^P = 1\}$.

Theorem 10.3 *For every $\varphi, \psi \in \ell(\mathbb{Z}_{MP})$,*

(a) $\{\mathrm{T}^{Mp}\varphi\}_{p\in\mathbb{Z}_P}$ *is orthonormal if and only if* $\|\varphi(\zeta)\|^2_{\mathbb{C}^M} = 1$ *for every* $\zeta \in \mathbb{T}_P$,

(b) $\{\mathrm{T}^{Mp}\varphi\}_{p\in\mathbb{Z}_P}$ *is orthogonal to* $\{\mathrm{T}^{Mp}\psi\}_{p\in\mathbb{Z}_P}$ *if and only if* $\langle\varphi(\zeta), \psi(\zeta)\rangle_{\mathbb{C}^M} = 0$ *for every* $\zeta \in \mathbb{T}_P$.

Proof For (a), note that $\{\mathrm{T}^{Mp}\varphi\}_{p\in\mathbb{Z}_P}$ being orthonormal is equivalent to having $\langle\varphi, \mathrm{T}^{Mp}\varphi\rangle_{\ell(\mathbb{Z}_{MP})} = \delta_0[p]$ for every $p \in \mathbb{Z}_P$. Taking Fourier transforms of this relation, Lemma 10.1 equivalently gives $\|\varphi(\mathrm{e}^{2\pi \mathrm{i}p/P})\|^2_{\mathbb{C}^M} = 1$ for every $p \in \mathbb{Z}_P$. Similarly for (b), $\{\mathrm{T}^{Mp}\varphi\}_{p\in\mathbb{Z}_P}$ being orthogonal to $\{\mathrm{T}^{Mp}\psi\}_{p\in\mathbb{Z}_P}$ is equivalent to having $\langle\varphi, \mathrm{T}^{Mp}\psi\rangle_{\ell(\mathbb{Z}_{MP})} = 0$ for every $p \in \mathbb{Z}_P$. Taking Fourier transforms of this relation, Lemma 10.1 equivalently gives $\langle\varphi(\mathrm{e}^{2\pi \mathrm{i}p/P}), \psi(\mathrm{e}^{2\pi \mathrm{i}p/P})\rangle_{\mathbb{C}^M} = 0$ for every $p \in \mathbb{Z}_P$. □

Example 10.3 Let's design a pair of real 4-tap filters $\psi_0, \psi_1 \in \ell(\mathbb{Z}_{2P})$ in such a way that $\{\mathrm{T}^{2p}\psi_n\}_{n=0,\,p\in\mathbb{Z}_P}^1$ forms an orthonormal basis. Our design for these taps will be independent of $P \geq 4$; the fact that we can design such filters while keeping P arbitrary speaks to the strength of this design process. In this example, we want the filters to have common support:

$$\begin{aligned}\psi_0 &:= a\delta_0 + b\delta_1 + c\delta_2 + d\delta_3,\\ \psi_1 &:= e\delta_0 + f\delta_1 + g\delta_2 + h\delta_3.\end{aligned}$$

Since $M = 2$, the polyphase components of these filters are given by Z-transforms over the even and odd indices:

$$\Psi(z) = \begin{bmatrix} \psi_0^{(0)}(z) & \psi_1^{(0)}(z) \\ \psi_0^{(1)}(z) & \psi_1^{(1)}(z) \end{bmatrix} = \begin{bmatrix} a + cz^{-1} & e + gz^{-1} \\ b + dz^{-1} & f + hz^{-1} \end{bmatrix}.$$

To determine which choices for ψ_0 and ψ_1 make $\{\mathrm{T}^{2p}\psi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{1}$ an orthonormal basis, we appeal to Theorem 10.3, which requires $\Psi(\zeta)$ to be a unitary matrix for every $\zeta \in \mathbb{T}_P$. A square polyphase matrix $\Psi(z)$ with this property is known as a *paraunitary* matrix. Since $\Psi(\zeta)$ is a 2×2 matrix, this isn't a difficult task; the second column can be taken to be a modulated involution of the first. However, we want this property to hold for every $\zeta \in \mathbb{T}_P$, and so we'll be more careful in applying Theorem 10.3. Specifically, Theorem 10.3(b) requires the first and second columns of $\Psi(\zeta)$ to be orthogonal for every $\zeta \in \mathbb{T}_P$, and so

$$\begin{aligned} 0 &= \big(a + c\zeta^{-1}\big)\big(e + g\zeta^{-1}\big)^* + \big(b + d\zeta^{-1}\big)\big(f + h\zeta^{-1}\big)^* \\ &= \big(a + c\zeta^{-1}\big)(e + g\zeta) + \big(b + d\zeta^{-1}\big)(f + h\zeta) \\ &= (ae + bf + cg + dh) + (ce + df)\zeta^{-1} + (ag + bh)\zeta. \end{aligned} \tag{10.24}$$

The coefficient of ζ^{-1} being zero gives

$$e = \alpha d, \qquad f = -\alpha c \tag{10.25}$$

for some $\alpha \in \mathbb{R}$, while the coefficient of ζ gives

$$g = \beta b, \qquad h = -\beta a \tag{10.26}$$

for some $\beta \in \mathbb{R}$. Substituting (10.25) and (10.26) into the constant term of (10.24) then gives

$$\begin{aligned} 0 &= ae + bf + cg + dh \\ &= a(\alpha d) + b(-\alpha c) + c(\beta b) + d(-\beta a) \\ &= (\alpha - \beta)(ad - bc). \end{aligned}$$

Thus, the columns of $\Psi(\zeta)$ are always orthogonal if and only if either $ad - bc = 0$ or $\alpha = \beta$. Forcing $ad - bc = 0$ would remove a lot of freedom in our filter design, and so instead we take $\alpha = \beta$. We may now rewrite $\Psi(z)$:

$$\Psi(z) = \begin{bmatrix} a + cz^{-1} & e + gz^{-1} \\ b + dz^{-1} & f + hz^{-1} \end{bmatrix} = \begin{bmatrix} a + cz^{-1} & \alpha(d + bz^{-1}) \\ b + dz^{-1} & -\alpha(c + az^{-1}) \end{bmatrix}.$$

Next, Theorem 10.3(a) requires the columns of $\Psi(\zeta)$ to be unit norm for every $\zeta \in \mathbb{T}_P$. Notice that for each $\zeta \in \mathbb{T}_P$, the norm squared of the second column is

$$\begin{aligned}\left|\alpha\left(d+b\zeta^{-1}\right)\right|^2+\left|-\alpha\left(c+a\zeta^{-1}\right)\right|^2&=\left|\alpha\zeta^{-1}(d\zeta+b)\right|^2+\left|-\alpha\zeta^{-1}(c\zeta+a)\right|^2\\&=\alpha^2\left(\left|b+d\left(\zeta'\right)^{-1}\right|^2+\left|a+c\left(\zeta'\right)^{-1}\right|^2\right),\end{aligned}$$

where $\zeta' := \zeta^{-1}$. That is, the norm squared of the second column at $z = \zeta$ is α^2 times the norm squared of the first column at $z = \zeta^{-1}$. Thus, to satisfy Theorem 10.3(a), we must have $\alpha = \pm 1$. Picking $\alpha = 1$, we rewrite $\Psi(z)$:

$$\Psi(z)=\begin{bmatrix}a+cz^{-1} & d+bz^{-1}\\ b+dz^{-1} & -c-az^{-1}\end{bmatrix}. \tag{10.27}$$

To summarize, the columns of (10.27) are the polyphase representations of ψ_0 and ψ_1, respectively, and $\{\mathrm{T}^{2p}\psi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{1}$ forms an orthonormal basis, provided

$$\left|a+c\zeta^{-1}\right|^2+\left|b+d\zeta^{-1}\right|^2=1,\quad \forall\zeta\in\mathbb{T}_P, \tag{10.28}$$

which by Theorem 10.3(a), is equivalent to having $\{\mathrm{T}^{2p}\psi_0\}_{p\in\mathbb{Z}_P}$ be orthonormal. The remaining degrees of freedom in choosing a, b, c, and d can be used to optimize for additional desirable filter properties, such as frequency selectivity, as discussed in greater detail in the next section.

Generalizing unitary filter banks, another nice class of PRFBs arises from Parseval filter banks. In contrast to the unitary case, the condition $\Phi\Phi^* = \mathrm{I}$ is not equivalent to $\Phi^*\Phi = \mathrm{I}$ since Φ is not a square matrix. The following result—a finite-dimensional version of the main results of [4, 8]—expresses the frame bounds of Φ in terms of the frame bounds of the corresponding polyphase representation $\Phi(\zeta)$ at each $\zeta \in \mathbb{T}_P$. In particular, Φ is Parseval if and only if $\Phi(\zeta)$ is Parseval for every $\zeta \in \mathbb{T}_P$.

Theorem 10.4 *Given any filters $\{\varphi_n\}_{n=0}^{N-1}$ in $\ell(\mathbb{Z}_{MP})$, the optimal frame bounds A and B for $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\ p\in\mathbb{Z}_P}^{N-1}$ in $\ell(\mathbb{Z}_{MP})$ are*

$$A=\min_{p\in\mathbb{Z}_P}A_p,\qquad B=\max_{p\in\mathbb{Z}_P}B_p,$$

where A_p and B_p denote the optimal frame bounds for $\{\varphi_n(\mathrm{e}^{2\pi\mathrm{i}p/P})\}_{n=0}^{N-1}$ in $\mathbb{C}^M$.

Proof We first show that this A and this B are indeed frame bounds, namely,

$$A\|x\|_{\ell(\mathbb{Z}_{MP})}^2\le\sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P}\left|\left\langle x,\mathrm{T}^{Mp}\varphi_n\right\rangle_{\ell(\mathbb{Z}_{MP})}\right|^2\le B\|x\|_{\ell(\mathbb{Z}_{MP})}^2,\quad\forall x\in\ell(\mathbb{Z}_{MP}). \tag{10.29}$$

To this end, we express the middle expression of (10.29) in terms of a norm in $\ell(\mathbb{Z}_P)$, and then we use the fact that the Fourier transform is unitary:

$$\begin{aligned}\sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P}\left|\left\langle x,\mathrm{T}^{Mp}\varphi_n\right\rangle_{\ell(\mathbb{Z}_{MP})}\right|^2 &= \sum_{n=0}^{N-1}\left\|\left\langle x,\mathrm{T}^{M\bullet}\varphi_n\right\rangle_{\ell(\mathbb{Z}_{MP})}\right\|_{\ell(\mathbb{Z}_P)}^2\\ &= \sum_{n=0}^{N-1}\left\|\mathrm{F}^*\left\langle x,\mathrm{T}^{M\bullet}\varphi_n\right\rangle_{\ell(\mathbb{Z}_{MP})}\right\|_{\ell(\mathbb{Z}_P)}^2\\ &= \sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P}\left|\left(\mathrm{F}^*\left\langle x,\mathrm{T}^{M\bullet}\varphi_n\right\rangle_{\ell(\mathbb{Z}_{MP})}\right)[p]\right|^2.\end{aligned}$$

Next, we apply Lemma 10.1 and then change the order of summation:

$$\begin{aligned}\sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P}\left|\left\langle x,\mathrm{T}^{Mp}\varphi_n\right\rangle_{\ell(\mathbb{Z}_{MP})}\right|^2 &= \sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P}\left|\frac{1}{\sqrt{P}}\left\langle x\left(\mathrm{e}^{2\pi\mathrm{i}p/P}\right),\varphi_n\left(\mathrm{e}^{2\pi\mathrm{i}p/P}\right)\right\rangle_{\mathbb{C}^M}\right|^2\\ &= \frac{1}{P}\sum_{p\in\mathbb{Z}_P}\left(\sum_{n=0}^{N-1}\left|\left\langle x\left(\mathrm{e}^{2\pi\mathrm{i}p/P}\right),\varphi_n\left(\mathrm{e}^{2\pi\mathrm{i}p/P}\right)\right\rangle_{\mathbb{C}^M}\right|^2\right).\end{aligned}\tag{10.30}$$

Since A_p is a lower frame bound for $\{\varphi_n(\mathrm{e}^{2\pi\mathrm{i}p/P})\}_{n=0}^{N-1}\subseteq\mathbb{C}^M$ for each $p\in\mathbb{Z}_P$, we continue (10.30):

$$\begin{aligned}\sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P}\left|\left\langle x,\mathrm{T}^{Mp}\varphi_n\right\rangle_{\ell(\mathbb{Z}_{MP})}\right|^2 &\geq \frac{1}{P}\sum_{p\in\mathbb{Z}_P}A_p\left\|x\left(\mathrm{e}^{2\pi\mathrm{i}p/P}\right)\right\|_{\mathbb{C}^M}^2\\ &\geq \left(\min_{p\in\mathbb{Z}_P}A_p\right)\left(\frac{1}{P}\sum_{p\in\mathbb{Z}_P}\left\|x\left(\mathrm{e}^{2\pi\mathrm{i}p/P}\right)\right\|_{\mathbb{C}^M}^2\right)\\ &= A\left\|x(z)\right\|_{\mathbb{P}_P^M[z]}^2\\ &= A\|x\|_{\ell(\mathbb{Z}_{MP})}^2,\end{aligned}\tag{10.31}$$

where (10.31) uses Proposition 10.5. Similarly, we can continue (10.30) to get

$$\sum_{n=0}^{N-1}\sum_{p\in\mathbb{Z}_P}\left|\left\langle x,\mathrm{T}^{Mp}\varphi_n\right\rangle_{\ell(\mathbb{Z}_{MP})}\right|^2\leq B\|x\|_{\ell(\mathbb{Z}_{MP})}^2,$$

and so A and B are indeed frame bounds for $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\ p\in\mathbb{Z}_P}^{N-1}\subseteq\ell(\mathbb{Z}_{MP})$.

To show that A and B are *optimal* frame bounds, we need to find nontrivial x's for which the left- and right-hand inequalities of (10.29) are achieved. For the left-hand

inequality, let p' be an index such that $A_{p'}$ is minimal, and let $x_{p';\min} \in \mathbb{C}^M$ be the vector that achieves this optimal lower frame bound of $\{\varphi_n(\mathrm{e}^{2\pi \mathrm{i} p'/P})\}_{n=0}^{N-1} \subseteq \mathbb{C}^M$. We define $x_{\min} \in \ell(\mathbb{Z}_{MP})$ in terms of its polyphase components:

$$x_{\min}^{(m)}(z) := x_{p';\min}[m] \prod_{p \in \mathbb{Z}_P \setminus \{p'\}} \left(\frac{z - \mathrm{e}^{2\pi \mathrm{i} p/P}}{\mathrm{e}^{2\pi \mathrm{i} p'/P} - \mathrm{e}^{2\pi \mathrm{i} p/P}} \right).$$

Notice that $x_{\min}(\mathrm{e}^{2\pi \mathrm{i} p/P}) = \delta_{p'}[p] x_{p';\min}$, and so the inequalities in (10.31) are achieved. Similar definitions produce an $x_{\max}$ that achieves the right-hand inequality of (10.29), and so we are done. □

From the point of view of applications, the significance of Theorem 10.4 is that it facilitates the design of filter banks that have good frame bounds. To be precise, though PRFBs satisfy $\Phi\Psi^* = \mathrm{I}$ and therefore provide overcomplete decompositions, such filter banks can be poor frames. This is significant, since only good frames are guaranteed to be robust against noise. To be precise, in many signal processing applications, the goal is to reconstruct x from $y = \Psi^* x + \varepsilon$, where ε is "noise" due to transmission errors, quantization, etc. Applying the dual frame Φ to y yields the reconstruction $\Phi y = x + \Phi\varepsilon$. Clearly, the validity of this estimate of x depends on the size of $\Phi\varepsilon$ relative to that of x. In general, if this filter bank is a poor frame, it is possible for $\Phi\varepsilon$ to be large even when ε is small; the only way to prevent this is to ensure that the *condition number* $\frac{B}{A}$ of the frame is as close as possible to 1. Though computing condition numbers can be computationally expensive for large matrices, Theorem 10.4 provides a shortcut for computing this number when the frame in question corresponds to a filter bank. Note that from this perspective, the best possible PRFBs are tight frames, an example of which we now consider.

Example 10.4 We now apply Theorem 10.4 to build a PRFB which arises from a Parseval synthesis filter bank $\Phi : [\ell(\mathbb{Z}_P)]^3 \to \ell(\mathbb{Z}_{2P})$ defined by

$$\Phi\{y_n\}_{n=0}^2 = \sum_{n=0}^{2} (\uparrow y_n) * \varphi_n.$$

As before, our choice for P will remain arbitrary. Note that, by Theorem 10.4, Φ is Parseval if and only if $\Phi(\zeta)$ is Parseval for every $\zeta \in \mathbb{T}_P$. Also, $\Phi(\zeta)$ is a 2×3 matrix, and the only 2×3 equal norm Parseval frame—up to rotation—is given by scaled cubed roots of unity:

$$\frac{1}{\sqrt{6}} \begin{bmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{bmatrix}. \tag{10.32}$$

As we will see in the following section, having equal norm columns in the polyphase matrix is desirable because it implies that the corresponding filter bank is somewhat balanced in its frequency selectivity.

Let's define our polyphase matrix $\Phi(z)$ to be the matrix in (10.32). Since the columns of (10.32) are the polyphase representations of each of our filters, we read off the coefficients of the polynomials to get

$$\begin{aligned}\varphi_0 &:= \frac{2}{\sqrt{6}}\delta_0,\\ \varphi_1 &:= -\frac{1}{\sqrt{6}}\delta_0 + \frac{\sqrt{3}}{\sqrt{6}}\delta_1,\\ \varphi_2 &:= -\frac{1}{\sqrt{6}}\delta_0 - \frac{\sqrt{3}}{\sqrt{6}}\delta_1.\end{aligned}$$

Here, the fact that $\Phi(z)$ has no dependence on z implies that the filters are only supported in the first $M = 2$ entries. Further, the matrix representation of the synthesis operator is

$$\Phi = \frac{1}{\sqrt{6}}\left[\begin{array}{cccc|cccc|cccc} 2 & & & & -1 & & & & -1 & & & \\ 0 & & & & \sqrt{3} & & & & -\sqrt{3} & & & \\ & 2 & & & & -1 & & & & -1 & & \\ & 0 & & & & \sqrt{3} & & & & -\sqrt{3} & & \\ & & \ddots & & & & \ddots & & & & \ddots & \\ & & & 2 & & & & -1 & & & & -1 \\ & & & 0 & & & & \sqrt{3} & & & & -\sqrt{3} \end{array}\right].$$

Notice that the translates of a common filter have disjoint support. Also, the columns of Φ can be shuffled to form a version of $\Phi(z)$ tensored with I_P. As such, this example is rather uninteresting as a filter bank; the corresponding analysis operator only analyzes interactions between certain pairs of entries.

For various applications, there are additional desirable filter properties which we would like our PRFBs to satisfy. In order to have the freedom necessary to satisfy these properties, our filters will need more taps. This means that the polyphase matrix $\Phi(z)$ of our filter bank will no longer be constant. But (10.32) gives the only 2×3 equal norm Parseval frame up to rotation. Therefore, we will define $\Phi(z)$ by multiplying (10.32) on the left with a unitary matrix of polynomials. To this end, consider $\Psi(z)$ in (10.27) with $a, c = 2^{-\frac{5}{2}}(1 \pm \sqrt{3})$ and $b, d = 2^{-\frac{5}{2}}(3 \pm \sqrt{3})$; we choose these numbers because they correspond to a Daubechies wavelet, as described in greater detail in the following section. Then

$$\begin{aligned}\Phi(z) &:= \begin{bmatrix} a + cz^{-1} & d + bz^{-1} \\ b + dz^{-1} & -c - az^{-1} \end{bmatrix}\left(\frac{1}{\sqrt{6}}\begin{bmatrix} 2 & -1 & -1 \\ 0 & \sqrt{3} & -\sqrt{3} \end{bmatrix}\right)\\ &= \frac{2}{\sqrt{6}}\begin{bmatrix} a + cz^{-1} & -2^{-\frac{3}{2}}(1 - \sqrt{3}z^{-1}) & d - bz^{-1} \\ b + dz^{-1} & -2^{-\frac{3}{2}}(\sqrt{3} + z^{-1}) & -c + az^{-1} \end{bmatrix}. \end{aligned} \tag{10.33}$$

The filters are then found by reading off coefficients from these polynomials:

$$\begin{aligned}
\varphi_0 &:= \frac{2a}{\sqrt{6}}\delta_0 + \frac{2b}{\sqrt{6}}\delta_1 + \frac{2c}{\sqrt{6}}\delta_3 + \frac{2d}{\sqrt{6}}\delta_3,\\
\varphi_1 &:= -\frac{1}{\sqrt{12}}\delta_0 - \frac{\sqrt{3}}{\sqrt{12}}\delta_1 + \frac{\sqrt{3}}{\sqrt{12}}\delta_2 - \frac{1}{\sqrt{12}}\delta_3,\\
\varphi_2 &:= \frac{2d}{\sqrt{6}}\delta_0 - \frac{2c}{\sqrt{6}}\delta_1 - \frac{2b}{\sqrt{6}}\delta_3 + \frac{2a}{\sqrt{6}}\delta_3.
\end{aligned}$$

When $P = 4$, the synthesis operator then becomes

$$\Phi = \frac{2}{\sqrt{6}}\left[\begin{array}{cccc|cccc|cccc}
a & & & c & -\frac{1}{2\sqrt{2}} & & & \frac{\sqrt{3}}{2\sqrt{2}} & d & & & -b\\
b & & & d & -\frac{\sqrt{3}}{2\sqrt{2}} & & & -\frac{1}{2\sqrt{2}} & -c & & & a\\
c & a & & & \frac{\sqrt{3}}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & & & -b & d & & \\
d & b & & & -\frac{1}{2\sqrt{2}} & -\frac{\sqrt{3}}{2\sqrt{2}} & & & a & -c & & \\
 & c & a & & & \frac{\sqrt{3}}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & & & -b & d & \\
 & d & b & & & -\frac{1}{2\sqrt{2}} & -\frac{\sqrt{3}}{2\sqrt{2}} & & & a & -c & \\
 & & c & a & & & \frac{\sqrt{3}}{2\sqrt{2}} & -\frac{1}{2\sqrt{2}} & & & -b & d\\
 & & d & b & & & -\frac{1}{2\sqrt{2}} & -\frac{\sqrt{3}}{2\sqrt{2}} & & & a & -c
\end{array}\right].$$

By construction, Φ is a Parseval frame. Certainly, this would be cumbersome to check by hand, but it is guaranteed by the fact that $\Psi(z)$ is a unitary matrix of polynomials. In fact, Φ would be Parseval regardless of our choice for P. This illustrates the utility of designing Parseval frames of polynomials like $\Phi(z)$.

10.5 Designing Filter Bank Frames

When designing a filter bank for a given real-world application, we usually seek three things: for the filter bank to be a "good" frame, for the filters to have small support, and for the filters to have good *frequency selectivity*, as detailed below. While these three goals are not mutually exclusive, they do compete. For example, filters with a very small number of taps give very little freedom in designing a desirable frequency response; this is a type of *uncertainty principle*. Because of this, even when restated in the polyphase domain, the task of designing such nice filter banks remains a difficult problem, as well as an active area of research [6, 12, 18, 19]. Currently, a popular choice of filter bank frame is the cosine modulated filter bank introduced in [3]. For other examples, see [14, 15] and the references therein.

Here, a "good" frame is one whose frame bounds A and B are as close to each other as they can be with respect to the design constraints of the other desired properties. This is desirable since the speed and numerical stability with which we may solve for x from $y = \Phi^* x + \varepsilon$ improves the closer our condition number B/A of $\Phi\Phi^*$ is to 1. In particular, when $M = N$, we would like $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ to be an orthonormal basis for $\ell(\mathbb{Z}_{MP})$, meaning its synthesis operator is unitary. Meanwhile for $M < N$, we would hope for Φ to be a tight frame, meaning the canonical dual is but a scalar multiple of the frame vectors themselves: $\psi_n = \frac{1}{A}\varphi_n$. Sometimes other design considerations trump this one. For example, while orthogonal wavelets exists, none but the Haar wavelet exhibit even symmetry—an important property in image processing—and thus the theory of *biorthogonal wavelets*, i.e., nontight filter bank frames with $M = N = 2$, was developed.

We also usually want the filters in our filter bank to have small support, that is, to have a small number of taps. This is crucial because in most real-world signal processing applications, the filtering itself is implemented directly in the time domain. To be clear, as discussed in Sect. 10.2, filtering can be viewed as multiplication in the frequency domain: $x * \varphi$ can be computed by taking the inverse Fourier transform of the pointwise multiplication of the Fourier transforms of x and φ. Using this method, the cost of computing $x * \varphi$ is essentially that of three Fourier transforms, namely $\mathcal{O}(MP\log(MP))$ operations for $x, \varphi \in \ell(\mathbb{Z}_{MP})$. However, in order to compute $x * \varphi$ at even a single time, this frequency-domain method requires all of the values of the input signal x to be known. Time-domain filtering, on the other hand, can be done in real time. In particular, if the support of one's filter is an interval of K points, then one can directly compute $x * \varphi$ using only $\mathcal{O}(KMP)$ operations; moreover, each value of $x * \varphi$ can be computed from K neighboring values of x.

Indeed, in most real-world applications, the true benefit of the frequency-domain representation of a filter is not any computational advantage, but rather an intuitive understanding of what that filter truly does. Frequency content is an important component of many signals of interest, such as audio signals, electromagnetic signals, and images. Proper filtering can help isolate the part of a signal that a user truly cares about. For example, a low-pass filter can help remove a high-pitched whine from the background of an audio sample. To be clear, no *frame* expansion can completely eliminate any part of a signal, since otherwise that part would not be able to be reconstructed. However, a properly designed filter bank frame can separate a signal into multiple *channels*, each emphasizing a particular region of the frequency spectrum. That is, though a low-pass filter itself is a poor frame, it can become a good one if taken together with the appropriate high-pass filter.

Formally speaking, the *nth channel* of the analysis and synthesis filter banks of $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ refers to the operations $x \mapsto \downarrow (x * \tilde{\varphi}_n)$ and $y_n \mapsto (\uparrow y_n) * \varphi_n$, respectively. In terms of frequency content, it is not difficult to show that the downsampling operation *periodizes*—sums the M-translates of—the MP-periodic Fourier transform of $x * \tilde{\varphi}_n$, while the upsampling operation periodically extends the P-periodic Fourier transform of y_n. That is, the downsampling and upsampling operations have fixed, known effects on the frequency content of a signal. As such, the only true design freedom in this channel lies with our choice of φ_n. Here, we

recall material from Sect. 10.2 to find that the frequency content of the filtered signal $x * \varphi_n$ is

$$\begin{aligned}\big[F^*(x*\varphi_n)\big][k] &= \frac{1}{\sqrt{MP}}\big[\mathrm{Z}(x*\varphi_n)\big]\big(\mathrm{e}^{2\pi \mathrm{i}k/MP}\big)\\ &= \frac{1}{\sqrt{MP}}(\mathrm{Z}x)\big(\mathrm{e}^{2\pi \mathrm{i}k/MP}\big)(\mathrm{Z}\tilde{\varphi}_n)\big(\mathrm{e}^{2\pi \mathrm{i}k/MP}\big)\\ &= \big(F^*x\big)[k]\big[(\mathrm{Z}\varphi_n)\big(\mathrm{e}^{2\pi \mathrm{i}k/MP}\big)\big]^*.\end{aligned}$$

In particular, $|[F^*(x*\varphi_n)][k]|^2 = |(F^*x)[k]|^2\,|(\mathrm{Z}\varphi_n)(\mathrm{e}^{2\pi \mathrm{i}k/MP})|^2$. Here, the multiplier $|(\mathrm{Z}\varphi_n)(\mathrm{e}^{2\pi \mathrm{i}k/MP})|^2$ is known as the *frequency response* of φ_n. This frequency response indicates the degree to which filtering x with φ_n will change the strength of any given frequency component of x. We note that in the classical signal processing literature in which filters typically lie in the infinite-dimensional space $\ell^1(\mathbb{Z})$, frequency responses are usually expressed in terms of classical Fourier series:

$$\hat{\varphi}_n(\omega) := \sum_{k=-\infty}^{\infty} \varphi_n[k]\mathrm{e}^{-\mathrm{i}k\omega}.$$

We adapt this notion to periodic signals in $\ell(\mathbb{Z}_{MP})$ by only summing over the smallest coset representatives of $\mathbb{Z}_{MP}$, and treating the remaining coefficients as zero:

$$\hat{\varphi}_n(\omega) := \sum_{k=-\lfloor MP/2\rfloor}^{\lfloor (MP-1)/2\rfloor} \varphi_n[k]\mathrm{e}^{-\mathrm{i}k\omega}.$$

Under this definition, $\hat{\varphi}_n$ is well defined at any ω in $\mathbb{R}/2\pi\mathbb{Z} = [-\pi, \pi)$. Moreover, the frequency response of φ_n is

$$\left|\hat{\varphi}_n\left(\frac{2\pi k}{MP}\right)\right|^2 = \left|\sum_{k'\in\mathbb{Z}_{MP}} \varphi_n\big[k'\big]\mathrm{e}^{-2\pi \mathrm{i}kk'/MP}\right|^2 = \big|(\mathrm{Z}\varphi_n)\big(\mathrm{e}^{2\pi \mathrm{i}k/MP}\big)\big|^2.$$

For an example of frequency responses, recall the two 4-tap Daubechies filters $\{\psi_0, \psi_1\}$ used in the previous section:

$$\begin{aligned}\psi_0 &:= a\delta_0 + b\delta_1 + c\delta_2 + d\delta_3,\\ \psi_1 &:= d\delta_0 - c\delta_1 + b\delta_2 - a\delta_3,\end{aligned}$$

where $a, c = 2^{-\frac{5}{2}}(1 \pm \sqrt{3})$ and $b, d = 2^{-\frac{5}{2}}(3 \pm \sqrt{3})$. Though this choice of coefficients seems quite arbitrary, a glimpse at the frequency responses of ψ_0 and ψ_1 reveals them to be quite special. Indeed, by looking at the plots of $|\hat{\psi}_0(\omega)^2|$ and $|\hat{\psi}_1(\omega)^2|$ in Fig. 10.2 over all $\omega \in [-\pi, \pi)$, we see that ψ_0 is a *low-pass* filter, meaning that it is concentrated on the frequencies near zero, while ψ_1 is a *high-pass* filter concentrated on frequencies near $\pm\pi$. The even symmetry of these graphs is

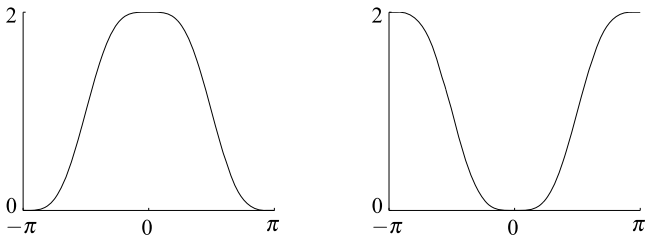


Fig. 10.2 The frequency responses $|\hat{\psi}_0(\omega)^2|$ and $|\hat{\psi}_1(\omega)^2|$ of the 4-tap low-pass Daubechies filter ψ_0 (*left*) and its high-pass cousin ψ_1 (*right*). As discussed in Sect. 10.3, the corresponding 2×2 polyphase matrix $\Psi(z)$ is unitary, meaning that the set $\{\mathrm{T}^{2p}\psi_0, \mathrm{T}^{2p}\psi_1\}_{p\in\mathbb{Z}_P}$ of all even translates of ψ_0 and ψ_1 forms an orthonormal basis for $\ell(\mathbb{Z}_P)$. The corresponding 2-channel filter bank consisting of the analysis and synthesis filter banks Ψ and Ψ^* exhibits all three of the most important properties of a good filter bank: it is a good frame (an orthonormal basis), the filters have a low number of taps (4), and the filters themselves exhibit good frequency selectivity, meaning that each of the 2 channels is concentrated on a particular range of frequencies. Taken together, these facts imply that $\Psi^* x$ can be quickly computed from x, that x can quickly be reconstructed from $\Psi^* x$ in a numerically stable fashion, and that each of the 2 output channels $y_0 :=\downarrow (x * \psi_0)$ and $y_1 :=\downarrow (x * \psi_1)$ contains a distinct frequency component of x. That is, Ψ^* nicely decomposes a signal x into its low and high frequency components

due to the fact that the coefficients of ψ_0 and ψ_1 are real. The particular values of a, b, c, and d are chosen so as to make ψ_0 as tall and flat as possible at $\omega = 0$, subject to the conditions (10.28) necessary for the resulting polyphase matrix $\Psi(z)$ of (10.27) be unitary.

A careful inspection of the two frequency responses depicted in Fig. 10.2 reveals that $|\hat{\psi}_1(\omega)^2|$ is a shift of $|\hat{\psi}_0(\omega)^2|$ by a factor of π and moreover that these two graphs sum to 2. The fact that $|\hat{\psi}_1(\omega)^2|$ is a shift of $|\hat{\psi}_0(\omega)^2|$ is an artifact of the way in which ψ_1 was constructed from ψ_0, and does not hold in general. However, as the next result shows, the constancy of the sum of $|\hat{\psi}_0(\omega)^2|$ and its shift is a consequence of the fact that the even translates of ψ_0 are orthonormal.

Theorem 10.5 *The vectors* $\{\mathrm{T}^{Mp}\varphi\}_{p\in\mathbb{Z}_P}$ *are orthonormal in* $\ell(\mathbb{Z}_{MP})$ *if and only if*

$$\sum_{m\in\mathbb{Z}_M} \left|\hat{\varphi}\left(\frac{2\pi(k-Pm)}{MP}\right)\right|^2 = M, \quad \forall k \in \mathbb{Z}_{MP}.$$

Proof From Theorem 10.3, we know that $\{\mathrm{T}^{Mp}\varphi\}_{p\in\mathbb{Z}_P}$ is orthonormal if and only if

$$1 = \sum_{m\in\mathbb{Z}_M} \left|\varphi^{(m)}(\zeta)\right|^2, \quad \forall \zeta \in \mathbb{T}_P,$$

which is equivalent to having

$$1 = \sum_{m\in\mathbb{Z}_M} \left|\varphi^{(m)}\left(\zeta^M\right)\right|^2, \quad \forall \zeta \in \mathbb{T}_{MP}. \tag{10.34}$$

Recalling from Proposition 10.6 that $(\mathrm{Z}\varphi)(z) = \sum_{m\in\mathbb{Z}_M} z^{-m}\varphi^{(m)}(z^M)$, we have

$$\begin{aligned}\sum_{m\in\mathbb{Z}_M} \left|(\mathrm{Z}\varphi)\left(\mathrm{e}^{-2\pi \mathrm{i}m/M}\zeta\right)\right|^2 &= \sum_{m\in\mathbb{Z}_M}\left|\sum_{m'\in\mathbb{Z}_M} \mathrm{e}^{2\pi \mathrm{i}mm'/M}\zeta^{-m'}\varphi^{(m')}\left(\zeta^M\right)\right|^2 \\ &= \sum_{m\in\mathbb{Z}_M} \left|\sqrt{M}\left[\mathrm{F}\left(\zeta^{-\bullet}\varphi^{(\bullet)}\right)\left(\zeta^M\right)\right][m]\right|^2,\end{aligned}$$

for any $\zeta \in \mathbb{T}_{MP}$. Since the inverse Fourier transform is unitary, it follows that

$$\sum_{m\in\mathbb{Z}_M} \left|(\mathrm{Z}\varphi)\left(\mathrm{e}^{-2\pi \mathrm{i}m/M}\zeta\right)\right|^2 = M\sum_{m\in\mathbb{Z}_M}\left|\zeta^{-m}\varphi^{(m)}\left(\zeta^M\right)\right|^2 = M\sum_{m\in\mathbb{Z}_M}\left|\varphi^{(m)}\left(\zeta^M\right)\right|^2.$$

In light of (10.34), we therefore have that $\{\mathrm{T}^{Mp}\varphi\}_{p\in\mathbb{Z}_P}$ is orthonormal if and only if

$$M = \sum_{m\in\mathbb{Z}_M} \left|(\mathrm{Z}\varphi)\left(\mathrm{e}^{-2\pi \mathrm{i}m/M}\zeta\right)\right|^2, \quad \forall \zeta \in \mathbb{T}_{MP}.$$

Writing ζ as $\mathrm{e}^{2\pi \mathrm{i}k/MP}$ then gives that $\{\mathrm{T}^{Mp}\varphi\}_{p\in\mathbb{Z}_P}$ is orthonormal if and only if

$$M = \sum_{m\in\mathbb{Z}_M} \left|(\mathrm{Z}\varphi)\left(\mathrm{e}^{-2\pi \mathrm{i}m/M}\mathrm{e}^{2\pi \mathrm{i}k/MP}\right)\right|^2 = \sum_{m\in\mathbb{Z}_M}\left|\hat{\varphi}\left(\frac{2\pi(k-Pm)}{MP}\right)\right|^2,$$
$$\forall k \in \mathbb{Z}_{MP},$$

as claimed. □

We note that for a general filter bank frame $\{\mathrm{T}^{MP}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$, we do not require $\{\mathrm{T}^{Mp}\varphi_n\}_{p\in\mathbb{Z}_P}$ to be orthonormal for each n. As such, Theorem 10.5 does not necessarily apply, meaning that for any n, the M-term periodization of $\{|\hat{\varphi}_n(\omega)|^2\}_{n=0}^{N-1}$ is not necessarily flat. At the same time, it can be advantageous to make such a restriction: having each of the M-translates of φ_n be orthonormal is equivalent to having the nth channel of the filter bank frame operator, namely the operation

$$x \to \left(\uparrow\downarrow (x * \tilde{\varphi}_n)\right) * \varphi_n = \sum_{p\in\mathbb{Z}_p} \left\langle x, \mathrm{T}^{Mp}\varphi_n\right\rangle \mathrm{T}^{Mp}\varphi_n,$$

be an orthogonal projection from $\ell(\mathbb{Z}_{MP})$ onto the subspace spanned by the M-translates of φ_n. In this case, the frame operator $\Phi\Phi^*$ of the filter bank becomes the sum of N projections—one for each channel—meaning that $\{\mathrm{T}^{Mp}\varphi_n\}_{n=0,\,p\in\mathbb{Z}_P}^{N-1}$ can be viewed as a *fusion frame*. Introduced in [5], such frames are optimal linear packet encoders [2], and are the focus of another chapter of this book. Intuitively, such a restriction ensures that each channel is of equal significance to all others. Such *filter bank fusion frames* are the focus of [7].

An example of such a filter bank was given in the previous section, namely the 3-channel, 2-downsampled Parseval filter bank frame whose 2×3 polyphase matrix

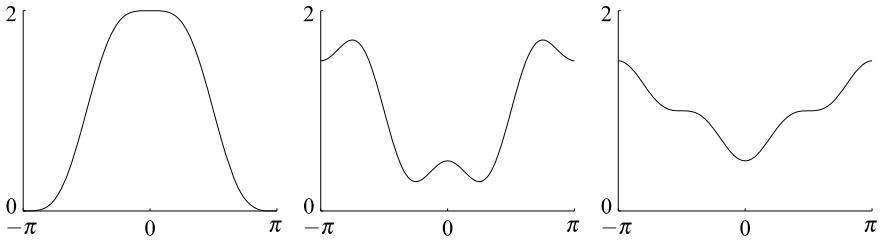


Fig. 10.3 The frequency response curves $|\hat{\varphi}_0(\omega)^2|$ (*left*), $|\hat{\varphi}_1(\omega)^2|$ (*center*), and $|\hat{\varphi}_2(\omega)^2|$ (*right*) of the 3-channel, 2-downsampled filter bank given in (10.33). This filter bank only exhibits two of the three commonly desired properties of a filter bank: it is a good frame, being tight, and each of the filters only has 4 taps. However, as the above graphs indicate, these filters do not exhibit good frequency selectivity. In particular, while φ_0—a copy of the 4-tap Daubechies low-pass filter ψ_0—does a good job of isolating the low frequencies of a signal, the other filters φ_1 and φ_2 allow all frequencies to pass through to a significant degree. As such, this filter bank would not be useful in the many signal processing applications in which frequency isolation is a primary goal

(10.33) is obtained by taking the product of the 4-tap Daubechies 2×2 paraunitary matrix $\Psi(z)$ with the fixed 2×3 Parseval Mercedes-Benz synthesis matrix. This approach to generating filter bank frames—multiplying a fixed synthesis matrix by a paraunitary matrix—was used to construct *strongly uniform tight frames* in [16]. A generalization of such frames, namely totally finite impulse response filter banks, is considered in [1]. Explicit constructions of strongly uniform tight frames are presented in [17], where the authors design these frames in a manner analogous to non-downsampled filter banks, such as the ones presented in [21] and various other works. The key attraction of this idea is that it permits us to exploit a known complete characterization of all paraunitary matrices [24].

Returning to our example, note that rescaling the columns by a factor of $\sqrt{3/2}$, the resulting matrix $\Phi(z)$ is a 2×3 unit norm tight frame of polynomials, meaning that at any $\zeta \in \mathbb{T}_P$, the three columns of $\Phi(\zeta)$ each have unit norm while its two rows are orthogonal with constant norm $\sqrt{3/2}$. The analysis filter bank decomposes any signal x in $\ell(\mathbb{Z}_{2P})$ into three component signals y_0, y_1, and y_2, each in $\ell(\mathbb{Z}_P)$. The frequency responses of the corresponding three filters are given in Fig. 10.3. They show that while this filter bank has good frame properties, it leaves much to be desired from the point of view of frequency selectivity. To construct filter bank frames with better frequency selectivity, we turn to the theory of Gabor filter banks.

10.5.1 Gabor Filter Banks

As finite Gabor frames are discussed in detail in another chapter, we only consider them briefly here, focusing on the special case of integer redundancy, meaning that the downsampling rate M divides the number of filters N. A *Gabor* filter bank is one in which the filters $\{\varphi_n\}_{n=0}^{N-1}$ are all *modulates* of a single filter φ. Such filter banks are attractive from a design perspective: here, the frequency response curves

of the various filters are all translates of each other, and so only a single "mother" filter need be designed [9]. Meanwhile, from a purely mathematical perspective, such filter banks are nice, since their polyphase representation is intimately related to the classical theory of Zak transforms [13].

To be precise, for any $p \in \mathbb{Z}$, the *modulation by* p operator on $\ell(\mathbb{Z}_P)$ is

$$\mathrm{E}^p : \ell(\mathbb{Z}_P) \to \ell(\mathbb{Z}_P), \qquad \big(\mathrm{E}^p y\big)\big[p'\big] := \mathrm{e}^{2\pi \mathrm{i} p p'/P} y\big[p'\big].$$

Let $R := N/M$ be the desired *redundancy* of the frame over $\ell(\mathbb{Z}_{MP})$, and let $Q := P/R$, meaning $N = MR$ and $P = QR$. Given any φ in $\ell(\mathbb{Z}_{MQR})$, we consider the *Gabor* system $\{\mathrm{T}^{Mp}\mathrm{E}^{Qn}\varphi\}_{p\in\mathbb{Z}_{QR}, n\in\mathbb{Z}_{MR}}$ of all M-translates and Q-modulates of φ, namely, the N-channel filter bank whose nth filter is $\varphi_n = \mathrm{E}^{Qn}\varphi$.

As before, we want a filter bank that is a good frame and has filters with a small number of taps and good frequency selectivity; using a Gabor filter bank helps us achieve these goals on all three fronts. To be precise, note that since the number of taps of $\varphi_n = \mathrm{E}^{Qn}\varphi$ equals that of φ, we only need to design a single filter φ of small support. Moreover, the frequency response of φ_n is but a shift of that of φ; we have

$$\begin{aligned}
(\mathrm{Z}\varphi_n)(z) &= \big(\mathrm{ZE}^{Qn}\varphi\big)(z) \\
&= \sum_{k\in\mathbb{Z}_{MQR}} \big(\mathrm{E}^{Qn}\varphi\big)[k] z^{-k} \\
&= \sum_{k\in\mathbb{Z}_{MQR}} \mathrm{e}^{2\pi \mathrm{i} Qnk/MQR}\varphi[k] z^{-k} \\
&= \sum_{k\in\mathbb{Z}_{MQR}} \varphi[k]\big(\mathrm{e}^{-2\pi \mathrm{i} n/MR} z\big)^{-k} \\
&= (\mathrm{Z}\varphi)\big(\mathrm{e}^{-2\pi \mathrm{i} n/MR} z\big),
\end{aligned}$$

and so the frequency response of φ_n has values

$$\begin{aligned}
\left|\hat{\varphi}_n\left(\frac{2\pi k}{MQR}\right)\right|^2 &= \big|(\mathrm{Z}\varphi_n)\big(\mathrm{e}^{2\pi \mathrm{i} k/MQR}\big)\big|^2 \\
&= \big|(\mathrm{Z}\varphi)\big(\mathrm{e}^{2\pi \mathrm{i} - n/MR}\mathrm{e}^{2\pi \mathrm{i} k/MQR}\big)\big|^2 \\
&= \big|(\mathrm{Z}\varphi)\big(\mathrm{e}\mathrm{e}^{2\pi \mathrm{i}(k-Qn)/MQR}\big)\big|^2 \\
&= \left|\hat{\varphi}\left(\frac{2\pi(k-Qn)}{MQR}\right)\right|^2. \qquad (10.35)
\end{aligned}$$

In particular, if we design φ well enough so that $|\hat{\varphi}(\omega)|^2$ is concentrated on a given band of frequencies, then the frequency response of any one of its modulates φ_n will be concentrated on one of N evenly spaced shifts of this band in $[-\pi, \pi)$.

What remains to be discussed is how using a Gabor filter bank simplifies our problem of constructing a good frame. By Theorem 10.4, the optimal frame bounds

of $\{\mathrm{T}^{Mp}\mathrm{E}^{Qn}\varphi\}_{p\in\mathbb{Z}_{QR},n\in\mathbb{Z}_{MR}}$ are obtained by computing the extreme eigenvalues of $\Phi(\zeta)[\Phi(\zeta)]^*$ over all $\zeta\in\mathbb{T}_{QR}=\{\zeta\in\mathbb{C}:\zeta^{QR}=1\}$, where $\Phi(z)$ is the polyphase matrix of $\{\varphi_n\}_{n=0}^{N-1}$. For our Gabor system, the mth component of the nth polyphase vector is

$$\begin{aligned}\varphi_n^{(m)}(z)&=\sum_{p\in\mathbb{Z}_P}\big(\mathrm{E}^{Qn}\varphi\big)[m+Mp]z^{-p}\\&=\sum_{p\in\mathbb{Z}_P}\mathrm{e}^{2\pi\mathrm{i}Qn(m+Mp)/MQR}\varphi[m+Mp]z^{-p}\\&=\mathrm{e}^{2\pi\mathrm{i}mn/MR}\sum_{p\in\mathbb{Z}_P}\varphi[m+Mp]\big(\mathrm{e}^{-2\pi\mathrm{i}n/R}z\big)^{-p}\\&=\mathrm{e}^{2\pi\mathrm{i}mn/MR}\varphi^{(m)}\big(\mathrm{e}^{-2\pi\mathrm{i}n/R}z\big).\end{aligned}$$

Remarkably, this fact implies that when the redundancy R is an integer, the rows of the polyphase matrix $\Phi(z)$ are necessarily orthogonal; for any $\zeta\in\mathbb{T}_{QR}$ and any row indices m and m', letting $n=r+Rm''$ gives

$$\begin{aligned}&\big(\Phi(\zeta)\big[\Phi(\zeta)\big]^*\big)_{m,m'}\\&\quad=\sum_{n\in\mathbb{Z}_{MR}}\varphi_n^{(m)}(\zeta)\big[\varphi_n^{(m')}(\zeta)\big]^*\\&\quad=\sum_{n\in\mathbb{Z}_{MR}}\mathrm{e}^{2\pi\mathrm{i}(m-m')n/MR}\varphi^{(m)}\big(\mathrm{e}^{-2\pi\mathrm{i}n/R}\zeta\big)\big[\varphi^{(m')}\big(\mathrm{e}^{-2\pi\mathrm{i}n/R}\zeta\big)\big]^*\\&\quad=\sum_{r\in\mathbb{Z}_R}\sum_{m''\in\mathbb{Z}_M}\mathrm{e}^{2\pi\mathrm{i}(m-m')(r+Rm'')/MR}\varphi^{(m)}\big(\mathrm{e}^{-2\pi\mathrm{i}r/R}\zeta\big)\big[\varphi^{(m')}\big(\mathrm{e}^{-2\pi\mathrm{i}r/R}\zeta\big)\big]^*\\&\quad=\sum_{r\in\mathbb{Z}_R}\varphi^{(m)}\big(\mathrm{e}^{-2\pi\mathrm{i}r/R}\zeta\big)\big[\varphi^{(m')}\big(\mathrm{e}^{-2\pi\mathrm{i}r/R}\zeta\big)\big]^*\mathrm{e}^{2\pi\mathrm{i}(m-m')r/MR}\\&\qquad\times\sum_{m''\in\mathbb{Z}_M}\mathrm{e}^{2\pi\mathrm{i}(m-m')m''/M}\\&\quad=\begin{cases}M\sum_{r\in\mathbb{Z}_R}|\varphi^{(m)}(\mathrm{e}^{-2\pi\mathrm{i}r/R}\zeta)|^2, & m=m'\mod M,\\0, & m\neq m'\mod M.\end{cases}\end{aligned}$$

In particular, since $\Phi(\zeta)[\Phi(\zeta)]^*$ is diagonal, its eigenvalues are its diagonal entries. We summarize the above discussion as follows.

Theorem 10.6 *For any positive integers M, Q, and R and any φ in $\ell(\mathbb{Z}_{MQR})$, the optimal frame bounds of the Gabor system $\{\mathrm{T}^{Mp}\mathrm{E}^{Qn}\varphi\}_{p\in\mathbb{Z}_{QR},n\in\mathbb{Z}_{MR}}$ are*

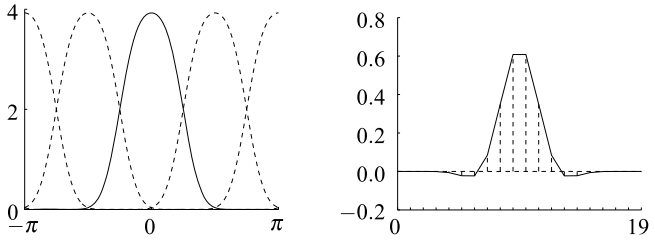


Fig. 10.4 A 20-tap max flat filter φ whose four modulates and even translates form a 4-channel, 2-downsampled tight Gabor filter bank frame. The frequency responses of φ and its three modulates are depicted in (**a**), while (**b**) depicts the filter itself as a function of time. As indicated by (10.35), these four frequency responses consist of evenly spaced translates of the frequency response of φ in $[-\pi, \pi)$. The filter coefficients were obtained by making the frequency response of φ be as small and flat as possible at $\omega = \pm\pi$, subject to the 20-tap constraint and the requirement (10.36) that the resulting frame be tight. Here, $M = R = 2$ and the even and odd parts of φ are both orthogonal to their own 2-translates. In particular, φ is orthogonal to its own 4-translates

$$A = M \min_{\zeta \in \mathbb{T}_{QR}} \min_{m \in \mathbb{Z}_M} \sum_{r \in \mathbb{Z}_R} \left|\varphi^{(m)}\left(\mathrm{e}^{-2\pi \mathrm{i} r/R}\zeta\right)\right|^2,$$

$$B = M \max_{\zeta \in \mathbb{T}_{QR}} \max_{m \in \mathbb{Z}_M} \sum_{r \in \mathbb{Z}_R} \left|\varphi^{(m)}\left(\mathrm{e}^{-2\pi \mathrm{i} r/R}\zeta\right)\right|^2.$$

In particular, when $\|\varphi\| = 1$, such Gabor systems are tight frames if and only if

$$\frac{R}{M} = \sum_{r \in \mathbb{Z}_R} \left|\varphi^{(m)}\left(\mathrm{e}^{-2\pi \mathrm{i} r/R}\zeta\right)\right|^2, \tag{10.36}$$

for all $m = 0, \ldots, M-1$ and all $\zeta \in \mathbb{T}_{QR}$. Thus, one way to construct a nice Gabor filter bank is to fix a desired number of taps K, and find a K-tap filter φ whose frequency response is concentrated about the origin subject to the constraint (10.36). An example of such a construction is given in Fig. 10.4.

An interesting consequence of (10.36) is that any φ which generates a finite tight Gabor frame with integer redundancy is necessarily orthogonal to some of its translates. This follows from the fact that (10.36) is a special case of Theorem 10.5. In particular, we have that (10.36) holds if and only if for every fixed m, the R-translates of the scaled mth coset $\sqrt{M}\varphi[m + M\bullet]$ are orthonormal in $\ell(\mathbb{Z}_{QR})$. To formally see this, note that Theorem 10.5 states that $\{\mathrm{T}^{Rq}\sqrt{M} \downarrow \mathrm{T}^{-m}\varphi\}_{q \in \mathbb{Z}_Q}$ is orthonormal if and only if

$$R = \sum_{r \in \mathbb{Z}_R} \left|\left(\mathrm{Z}\sqrt{M} \downarrow \mathrm{T}^{-m}\varphi\right)\left(\mathrm{e}^{2\pi \mathrm{i}(p-Qr)/QR}\right)\right|^2, \quad \forall p \in \mathbb{Z}_{QR}.$$

Writing $\varphi^{(m)}(z) = \sum_{p\in\mathbb{Z}_P} \varphi[m+Mp]z^{-p} = (\mathrm{Z}\downarrow_M \mathrm{T}^{-m}\varphi)(z)$ then gives this to be equivalent to having

$$\frac{R}{M} = \sum_{r\in\mathbb{Z}_R} \left|\varphi^{(m)}\left(\mathrm{e}^{2\pi\mathrm{i}(p-Qr)/QR}\right)\right|^2 = \sum_{r\in\mathbb{Z}_R} \left|\varphi^{(m)}\left(\mathrm{e}^{-2\pi\mathrm{i}r/R}\mathrm{e}^{2\pi\mathrm{i}p/QR}\right)\right|^2,$$
$$\forall p \in \mathbb{Z}_{QR},$$

namely, (10.36) where $\zeta = \mathrm{e}^{2\pi\mathrm{i}p/QR}$.

This fact in particular implies that if φ generates a finite, integer-redundant, tight Gabor frame $\{\mathrm{T}^{Mp}\mathrm{E}^{Qn}\varphi\}_{p\in\mathbb{Z}_{QR},n\in\mathbb{Z}_{MR}}$ in $\ell(\mathbb{Z}_{MQR})$, then the MR-translates of φ are necessarily orthogonal. As such, all such frames are necessarily filter bank fusion frames to some degree.

Acknowledgements The authors thank Amina Chebira and Terika Harris for useful discussions. This work was supported by NSF DMS 1042701, NSF CCF 1017278, AFOSR F1ATA01103J001, AFOSR F1ATA00183G003, and the A.B. Krongard Fellowship. The views expressed in this article are those of the authors and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the U.S. Government.

References

1. Bernardini, R., Rinaldo, R.: Oversampled filter banks from extended perfect reconstruction filter banks. IEEE Trans. Signal Process. **54**, 2625–2635 (2006)
2. Bodmann, B.G.: Optimal linear transmission by loss-insensitive packet encoding. Appl. Comput. Harmon. Anal. **22**, 274–285 (2007)
3. Bölcskei, H., Hlawatsch, F.: Oversampled cosine modulated filter banks with perfect reconstruction. IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process. **45**, 1057–1071 (1998)
4. Bölcskei, H., Hlawatsch, F., Feichtinger, H.G.: Frame-theoretic analysis of oversampled filter banks. IEEE Trans. Signal Process. **46**, 3256–3269 (1998)
5. Casazza, P.G., Kutyniok, G.: Frames of subspaces. Contemp. Math. **345**, 87–113 (2004)
6. Chai, L., Zhang, J., Zhang, C., Mosca, E.: Frame-theory-based analysis and design of oversampled filter banks: direct computational method. IEEE Trans. Signal Process. **55**, 507–519 (2007)
7. Chebira, A., Fickus, M., Mixon, D.G.: Filter bank fusion frames. IEEE Trans. Signal Process. **59**, 953–963 (2011)
8. Cvetković, Z., Vetterli, M.: Oversampled filter banks. IEEE Trans. Signal Process. **46**, 1245–1255 (1998)
9. Cvetković, Z., Vetterli, M.: Tight Weyl-Heisenberg frames in $\ell^2(\mathbb{Z})$. IEEE Trans. Signal Process. **46**, 1256–1259 (1998)
10. Daubechies, I.: Ten Lectures on Wavelets. SIAM, Philadelphia (1992)
11. Fickus, M., Johnson, B.D., Kornelson, K., Okoudjou, K.: Convolutional frames and the frame potential. Appl. Comput. Harmon. Anal. **19**, 77–91 (2005)
12. Gan, L., Ling, C.: Computation of the para-pseudo inverse for oversampled filter banks: forward and backward Greville formulas. IEEE Trans. Image Process. **56**, 5851–5859 (2008)
13. Gröchenig, K.: Foundations and Time-Frequency Analysis. Birkhäuser, Boston (2001)
14. Kovačević, J., Chebira, A.: Life beyond bases: the advent of frames (Part I). IEEE Signal Process. Mag. **24**, 86–104 (2007)
15. Kovačević, J., Chebira, A.: Life beyond bases: the advent of frames (Part II). IEEE Signal Process. Mag. **24**, 115–125 (2007)

16. Kovačević, J., Dragotti, P.L., Goyal, V.K.: Filter bank frame expansions with erasures. IEEE Trans. Inf. Theory **48**, 1439–1450 (2002)
17. Marinkovic, S., Guillemot, C.: Erasure resilience of oversampled filter bank codes based on cosine modulated filter banks. In: Proc. IEEE Int. Conf. Commun., pp. 2709–2714 (2004)
18. Mertins, A.: Frame analysis for biorthogonal cosine-modulated filterbanks. IEEE Trans. Signal Process. **51**, 172–181 (2003)
19. Motwani, R., Guillemot, C.: Tree-structured oversampled filterbanks as joint source-channel codes: application to image transmission over erasure channels. IEEE Trans. Signal Process. **52**, 2584–2599 (2004)
20. Oppenheim, A.V., Schafer, R.W.: Discrete-Time Signal Processing, 3rd edn. Pearson, Upper Saddle River (2009)
21. Selesnick, I.W., Baraniuk, R.G., Kingsbury, N.G.: The dual-tree complex wavelet transform. IEEE Signal Process. Mag. **22**, 123–151 (2005)
22. Smith, M., Barnwell, T.: Exact reconstruction techniques for tree-structured subband coders. IEEE Trans. Acoust. Speech Signal Process. **34**, 434–441 (1986)
23. Strang, G., Nguyen, T.: Wavelets and Filter Banks, 2nd edn. Cambridge Press, Wellesley (1996)
24. Vaidyanathan, P.P.: Multirate Systems and Filter Banks. Prentice Hall, Englewood Cliffs (1992)
25. Vetterli, M.: Filter banks allowing perfect reconstruction. Signal Process. **10**, 219–244 (1986)
26. Vetterli, M., Kovačević, J., Goyal, V.K.: Fourier and Wavelet Signal Processing (2011). http://www.fourierandwavelets.org/

Chapter 11
The Kadison–Singer and Paulsen Problems in Finite Frame Theory

Peter G. Casazza

Abstract We now know that some of the basic open problems in frame theory are equivalent to fundamental open problems in a dozen areas of research in both pure and applied mathematics, engineering, and others. These problems include the 1959 *Kadison–Singer problem* in C^*-algebras, the *paving conjecture* in operator theory, the *Bourgain–Tzafriri conjecture* in Banach space theory, the *Feichtinger conjecture* and the R_ϵ-conjecture in frame theory, and many more. In this chapter we will show these equivalences among others. We will also consider a slight weakening of the Kadison–Singer problem called the *Sundberg problem*. Then we will look at the recent advances on another deep problem in frame theory called the *Paulsen problem*. In particular, we will see that this problem is also equivalent to a fundamental open problem in operator theory. Namely, if a projection on a finite dimensional Hilbert space has a nearly constant diagonal, how close is it to a constant diagonal projection?

Keywords Kadison–Singer problem · Paving conjecture · State · Rieszable · Discrete Fourier transform · R_ϵ-Conjecture · Feichtinger conjecture · Bourgain–Tzafriri conjecture · Restricted invertibility principle · Sundberg problem · Paulsen problem · Principal angles · Chordal distance

11.1 Introduction

Finite frame theory is beginning to have an important impact on some of the deepest problems in both pure and applied mathematics. In this chapter we will look at two cases where finite frame theory is having a serious impact: the *Kadison–Singer problem* and the *Paulsen problem*. We warn the reader that because we are restricting ourselves to finite dimensional Hilbert spaces, a significant body of literature on the infinite dimensional versions of these problems does not appear here.

P.G. Casazza (✉)
Mathematics Department, University of Missouri, Columbia, MO 65211, USA
e-mail: casazzap@missouri.edu

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_11,

11.2 The Kadison–Singer Problem

For over 50 years the Kadison–Singer problem [32] has defied the best efforts of some of the most talented mathematicians of our time.

Kadison–Singer Problem 11.1 (KS) *Does every pure state on the* (*abelian*) *von Neumann algebra* $\mathbb{D}$ *of bounded diagonal operators on* ℓ_2, *the Hilbert space of square summable sequences on the integers*, *have a unique extension to a* (*pure*) *state on* $B(\ell_2)$, *i.e.*, *the von Neumann algebra of all bounded linear operators on the Hilbert space* ℓ_2*?*

A *state* of a von Neumann algebra $\mathcal{R}$ is a linear functional f on $\mathcal{R}$ for which $f(I) = 1$ and $f(T) \geq 0$ whenever $T \geq 0$ (whenever T is a positive operator). The set of states of $\mathcal{R}$ is a convex subset of the dual space of $\mathcal{R}$ which is compact in the ω^*-topology. By the Krein-Milman theorem, this convex set is the closed convex hull of its extreme points. The extremal elements in the space of states are called the *pure states* (of $\mathcal{R}$).

This problem arose from the very productive collaboration of Kadison and Singer in the 1950s when they were studying Dirac's Quantum Mechanics book [26] which culminated in their seminal work on triangular operator algebras.

It is now known that the 1959 Kadison–Singer problem is equivalent to fundamental unsolved problems in a dozen areas of research in pure mathematics, applied mathematics, and engineering (see [1–4, 16, 22, 23, 33] and their references). We will not develop this topic in detail here, since it is fundamentally an infinite dimensional problem and we are concentrating on finite dimensional frame theory. In this chapter we will look at a number of these finite dimensional problems which are equivalent to KS and which are impacted by finite frame theory. Most people today seem to agree with the original statement of Kadison and Singer [32] that KS will have a negative answer and so all the equivalent forms will have negative answers also.

11.2.1 The Paving Conjecture

A significant advance on KS was made by Anderson [2] in 1979 when he reformulated KS into what is now known as the *paving conjecture* (see also [3, 4]). Lemma 5 of [32] shows a connection between KS and paving. For notation, if $J \subset \{1, 2, \ldots, n\}$, the *diagonal projection* Q_J is the matrix whose entries are all zero except for the (i, i) entries for $i \in J$ which are all one. For a matrix $A = (a_{ij})_{i,j=1}^N$ let $\delta(A) = \max_{1 \leq i \leq N} |a_{ii}|$.

Definition 11.1 An operator $T \in B(\ell_2^N)$ is said to have an (r, ϵ)-*paving* if there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that

$$\|Q_{A_j} T Q_{A_j}\| \leq \epsilon \|T\|.$$

Paving Conjecture 11.1 (PC) *For every* $0 < \epsilon < 1$, *there is a natural number* r *so that for every natural number* N *and every linear operator* T *on* l_2^N *whose matrix has zero diagonal*, T *has an* (r, ϵ)*-paving*.

It is important that r not depend on N in PC. We will say that an arbitrary operator T satisfies PC if $T - D(T)$ satisfies PC where $D(T)$ is the diagonal of T.

The only large classes of operators which have been shown to be pavable are "diagonally dominant" matrices [6, 7, 9, 27], ℓ_1-localized operators [21], matrices with all entries real and positive [28], matrices with small coefficients in comparison with the dimension [12] (see [36] for a paving into blocks of constant size), and Toeplitz operators over Riemann integrable functions (see also [29]). Also, in [8] there is an analysis of the paving problem for certain Schatten C_p-norms.

Theorem 11.1 *The paving conjecture has a positive solution if any one of the following classes satisfies the paving conjecture*:

1. *Unitary operators* [23]
2. *Orthogonal projections* [23]
3. *Orthogonal projections with constant diagonal* $1/2$ [17]
4. *Positive operators* [23]
5. *Self-adjoint operators* [23]
6. *Gram matrices* $(\langle \varphi_i, \varphi_j \rangle)_{i,j \in I}$ *where* $T : \ell_2(I) \to \ell_2(I)$ *is a bounded linear operator, and* $Te_i = \varphi_i$, $\|Te_i\| = 1$ *for all* $i \in I$ [23]
7. *Invertible operators* (*or invertible operators with zero diagonal*) [23]
8. *Triangular operators* [34].

Recently, Weaver [38] provided important insight into KS by giving an equivalent problem to PC in terms of projections.

Conjecture 11.1 (Weaver) *There exist universal constants* $0 < \delta, \epsilon < 1$ *and* $r \in \mathbb{N}$ *so that for all* N *and all orthogonal projections* P *on* ℓ_2^N *with* $\delta(P) \leq \delta$, *there is a paving* $\{A_j\}_{j=1}^r$ *of* $\{1, 2, \ldots, N\}$ *so that* $\|Q_{A_j} P Q_{A_j}\| \leq 1 - \epsilon$, *for all* $j = 1, 2, \ldots, r$.

This needs some explanation, since there is nothing in [38] that looks anything like Conjecture 11.1. Weaver observes that the fact that Conjecture 11.1 implies PC follows by a minor modification of Propositions 7.6 and 7.7 of [1]. Then he introduces what he calls "Conjecture $\mathrm{KS_r}$" (see Conjecture 11.8). A careful examination of the proof of Theorem 1 of [38] reveals that Weaver shows that Conjecture $\mathrm{KS_r}$ implies Conjecture 11.1, which in turn implies KS, which (after the theorem is proved) is equivalent to $\mathrm{KS_r}$.

In [17] it was shown that PC fails for $r = 2$, even for projections with constant diagonal $1/2$. Recently [19] there appeared a frame theoretic concrete construction of non-2-pavable projections. If this construction can be generalized, we would have a counterexample to PC and KS. We now look at the construction from [19].

Definition 11.2 A family of vectors $\{\varphi_i\}_{i=1}^M$ for an N-dimensional Hilbert space $\mathcal{H}^N$ is (δ, r)-*Rieszable* if there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ so that for all $j = 1, 2, \ldots, r$ and all scalars $\{a_i\}_{i\in A_j}$ we have

$$\left\| \sum_{i\in A_j} a_i \varphi_i \right\|^2 \geq \delta \sum_{i\in A_j} |a_i|^2.$$

A projection P on $\mathcal{H}^N$ is (δ, r)-*Rieszable* if $\{Pe_i\}_{i=1}^N$ is (δ, r)-Rieszable.

We now have the following.

Proposition 11.1 *Let P be an orthogonal projection on $\mathcal{H}^N$. The following are equivalent*:

(1) *The vectors $\{Pe_i\}_{i=1}^N$ are (δ, r)-Rieszable.*
(2) *There is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that for all $j = 1, 2, \ldots, r$ and all scalars $\{a_i\}_{i\in A_j}$ we have*

$$\left\| \sum_{i\in A_j} a_i (I - P)e_i \right\|^2 \leq (1 - \delta) \sum_{i\in A_j} |a_i|^2.$$

(3) *The matrix of $I - P$ is (δ, r)-pavable.*

Proof (1) $\Leftrightarrow$ (2): For any scalars $\{a_i\}_{i\in A_j}$ we have

$$\sum_{i\in A_j} |a_i|^2 = \left\| \sum_{i\in A_j} a_i Pe_i \right\|^2 + \left\| \sum_{i\in A_j} a_i (I - P)e_i \right\|^2.$$

Hence,

$$\left\| \sum_{i\in A_j} a_i (I - P)e_i \right\|^2 \leq (1 - \delta) \sum_{i\in A_j} |a_i|^2 \quad \text{if and only if}$$

$$\left\| \sum_{i\in A_j} a_i Pe_i \right\|^2 \geq \delta \sum_{i\in A_j} |a_i|^2.$$

(2) $\Leftrightarrow$ (3): Given any partition $\{A_j\}_{j=1}^r$, any $1 \leq j \leq r$, and any $x = \sum_{i\in A_j} a_i e_i$, we have

$$\big\langle (I - P)x, x \big\rangle = \big\| (I - P)x \big\|^2 = \left\| \sum_{i\in A_j} a_i (I - P)e_i \right\|^2$$

$$\leq (1 - \delta) \sum_{i\in A_j} |a_i|^2 = \big\langle (1 - \delta)x, x \big\rangle,$$

if and only if $I - P \leq (1 - \delta)I$. $\square$

Given $N \in \mathbb{N}$, let $\omega = \exp^{(\frac{2\pi i}{N})}$; we define the discrete Fourier transform (DFT) matrix in $\mathbb{C}^N$ by

$$D_N = \sqrt{\frac{1}{N}}\left(\omega^{jk}\right)_{j,k=0}^{N-1}.$$

The main point of these D_N matrices is that they are unitary matrices for which the moduli of all the entries are equal to $\sqrt{\frac{1}{N}}$. The following is a simple observation.

Proposition 11.2 *Let $A = (a_{ij})_{i,j=1}^{N}$ be a matrix with orthogonal rows and satisfying $|a_{ij}|^2 = a$ for all i, j. If we multiply the j^{th} row of A by a constant C_j to get a new matrix B, then*:

(1) *The rows of B are orthogonal.*
(2) *The square sums of the entries of any column of B all equal*

$$a\sum_{j=1}^{N} C_j^2.$$

(3) *The square sum of the entries of the j^{th} row of B equals aC_j^2.*

To construct our example, we start with a $2N \times 2N$ DFT and multiply the first $N-1$ rows by $\sqrt{2}$ and the remaining rows by $\sqrt{\frac{2}{N+1}}$ to get a new matrix B_1. Next, we take a second $2N \times 2N$ DFT matrix and multiply the first $N-1$ rows by 0 and the remaining rows by $\sqrt{\frac{2N}{2N+1}}$ to get a matrix B_2. We then put the matrices B_1, B_2 side by side to get an $N \times 2N$ matrix B of the form

$$B = \begin{array}{|c|c|c|} \hline (N-1) \text{ Rows} & \sqrt{2} & 0 \\ \hline (N+1) \text{ Rows} & \sqrt{\frac{2}{N+1}} & \sqrt{\frac{2N}{N+1}} \\ \hline \end{array}$$

This matrix has $2N$ rows and $4N$ columns. Now we show that this matrix gives the required example.

Proposition 11.3 *The matrix B satisfies the following.*

(1) *The columns are orthogonal and the square sum of the coefficients of every column equals* 2.
(2) *The square sum of the coefficients of every row equals* 1.

The row vectors of the matrix B are not $(\delta, 2)$-Rieszable, for any δ independent of N.

Proof A direct calculation yields (1) and (2).

We will now show that the column vectors of B are not uniformly 2-Rieszable independent of N. So let $\{A_1, A_2\}$ be a partition of $\{1, 2, \ldots, 4N\}$. Without loss of generality, we may assume that $|A_1 \cap \{1, 2, \ldots, 2N\}| \geq N$. Let the column vectors of

the matrix B be $\{\varphi_i\}_{i=1}^{4N}$ as elements of $\mathbb{C}^{2N}$. Let P_{N-1} be the orthogonal projection of $\mathbb{C}^{2N}$ onto the first $N-1$ coordinates. Since $|A_1| \geq N$, there are scalars $\{a_i\}_{i\in A_1}$ so that $\sum_{i\in A_1} |a_i|^2 = 1$ and

$$P_{N-1}\Big(\sum_{i\in A_1} a_i\varphi_i\Big) = 0.$$

Also, let $\{\psi_j\}_{j=1}^{2N}$ be the orthonormal basis consisting of the original columns of the DFT_{2N}. We now have

$$\begin{aligned}
\Big\|\sum_{i\in A_1} a_i\varphi_i\Big\|^2 &= \Big\|(I-P_{N-1})\Big(\sum_{i\in A_1} a_i\varphi_i\Big)\Big\|^2 \\
&= \frac{2}{N+1}\Big\|(I-P_{N-1})\Big(\sum_{i\in A_1} a_i\psi_i\Big)\Big\|^2 \\
&\leq \frac{2}{N+1}\Big\|\sum_{i\in A_1} a_i\psi_i\Big\|^2 \\
&= \frac{2}{N+1}\sum_{i\in A_1}|a_i|^2 \\
&= \frac{2}{N+1}.
\end{aligned}$$

Letting $N\to\infty$, this class of matrices is not $(\delta,2)$-Rieszable, and hence not $(\delta,2)$-pavable for any $\delta > 0$. □

If this argument could be generalized to yield non-$(\delta,3)$-Rieszable (pavable) matrices, then such an argument should lead to a complete counterexample to PC.

11.2.2 The R_ϵ-Conjecture

In this section we will define the R_ϵ-conjecture and show that it is equivalent to the paving conjecture.

Definition 11.3 A family of vectors $\{\varphi_i\}_{i=1}^M$ is an ϵ-Riesz basic sequence for $0 < \epsilon < 1$ if for all scalars $\{a_i\}_{i=1}^M$ we have

$$(1-\epsilon)\sum_{i=1}^M |a_i|^2 \leq \Big\|\sum_{i=1}^M a_i\varphi_i\Big\|^2 \leq (1+\epsilon)\sum_{i=1}^M |a_i|^2.$$

A natural question is whether we can improve the Riesz basis bounds for a unit norm Riesz basic sequence by partitioning the sequence into subsets.

Conjecture 11.2 (R_ϵ-Conjecture) *For every $\epsilon > 0$, every unit norm Riesz basic sequence is a finite union of ϵ-Riesz basic sequences.*

This conjecture was first stated by Casazza and Vershynin and was first studied in [15], where it was shown that PC implies the conjecture. One advantage of the R_ϵ-conjecture is that it can be shown to students at the beginning of a course in Hilbert spaces.

The R_ϵ-conjecture has a natural finite dimensional form.

Conjecture 11.3 *For every $\epsilon > 0$ and every $T \in B(\ell_2^N)$ with $\|Te_i\| = 1$ for $i = 1, 2, \ldots, N$ there is an $r = r(\epsilon, \|T\|)$ and a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that for all $j = 1, 2, \ldots, r$ and all scalars $\{a_i\}_{i \in A_j}$ we have*

$$(1-\epsilon)\sum_{i\in A_j}|a_i|^2 \leq \left\|\sum_{i\in A_j} a_i T e_i\right\|^2 \leq (1+\epsilon)\sum_{i\in A_j}|a_i|^2.$$

Now we show that the R_ϵ-conjecture is equivalent to PC.

Theorem 11.2 *The following are equivalent*:

(1) *The paving conjecture.*
(2) *For $0 < \epsilon < 1$, there is an $r = r(\epsilon, B)$ so that for every $N \in \mathbb{N}$, if $T : \ell_2^N \to \ell_2^N$ is a bounded linear operator with $\|T\| \leq B$ and $\|Te_i\| = 1$ for all $i = 1, 2, \ldots, N$, then there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that for each $1 \leq j \leq r$, $\{Te_i\}_{i\in A_j}$ is an ϵ-Riesz basic sequence.*
(3) *The R_ϵ-conjecture.*

Proof (1) $\Rightarrow$ (2): Fix $0 < \epsilon < 1$. Given T as in (2), let $S = T^*T$. Since S has ones on its diagonal, by the paving conjecture there is an $r = r(\epsilon, \|T\|)$ and a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that for every $j = 1, 2, \ldots, r$ we have

$$\left\| Q_{A_j}(I - S)Q_{A_j} \right\| \leq \delta \|I - S\|$$

where $\delta = \frac{\epsilon}{\|S\|+1}$. Now, for all $x = \sum_{i=1}^N a_i e_i$ and all $j = 1, 2, \ldots, r$ we have

$$\left\|\sum_{i\in A_j} a_i T e_i\right\|^2 = \|T Q_{A_j} x\|^2 = \langle T Q_{A_j} x, T Q_{A_j} x\rangle = \left\langle T^* T Q_{A_j} x, Q_{A_j} x\right\rangle$$

$$= \langle Q_{A_j} x, Q_{A_j} x\rangle - \left\langle Q_{A_j}(I - S)Q_{A_j} x, Q_{A_j} x\right\rangle$$

$$\geq \|Q_{A_j}x\|^2 - \delta\|I - S\|\|Q_{A_j}x\|^2$$
$$\geq (1-\epsilon)\|Q_{A_j}x\|^2 = (1-\epsilon)\sum_{i\in A_j}|a_i|^2.$$

Similarly, $\|\sum_{i\in A_j} a_i T e_i\|^2 \leq (1+\epsilon)\sum_{i\in A_j}|a_i|^2$.

(2) $\Rightarrow$ (3): This is obvious.

(3) $\Rightarrow$ (1): Let $T \in B(\ell_2^N)$ with $Te_i = \varphi_i$ and $\|\varphi_i\| = 1$ for all $1 \leq i \leq N$. By Theorem 11.1 part 6, it suffices to show that the Gram operator G of $\{\varphi_i\}_{i=1}^N$ is pavable. Fix $0 < \delta < 1$ and let $\epsilon > 0$. Let $\psi_i = \sqrt{1-\delta^2}\varphi_i \oplus \delta e_i \in \ell_2^N \oplus \ell_2^N$. Then $\|\psi_i\| = 1$ for all $1 \leq i \leq N$ and for all scalars $\{a_i\}_{i=1}^N$,

$$\delta\sum_{i=1}^N |a_i|^2 \leq \left\|\sum_{i=1}^N a_i\psi_i\right\|^2 = \left(1-\delta^2\right)\left\|\sum_{i=1}^N a_i T e_i\right\|^2 + \delta^2\sum_{i=1}^N |a_i|^2$$
$$\leq \left[\left(1-\delta^2\right)\|T\|^2 + \delta^2\right]\sum_{i=1}^N |a_i|^2.$$

So $\{\psi_i\}_{i=1}^N$ is a unit norm Riesz basic sequence and $\langle\psi_i, \psi_k\rangle = (1-\delta^2)\langle\varphi_i, \varphi_k\rangle$ for all $1 \leq i \neq k \leq N$. By the R_ϵ-conjecture, there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that for all $j = 1, 2, \ldots, r$ and all $x = \sum_{i\in A_j} a_i e_i$,

$$(1-\epsilon)\sum_{i\in A_j}|a_i|^2 \leq \left\|\sum_{i\in A_j} a_i\psi_i\right\|^2 = \left\langle\sum_{i\in A_j} a_i\psi_i, \sum_{k\in A_j} a_k\psi_k\right\rangle$$
$$= \sum_{i\in A_j}|a_i|^2\|\psi_i\|^2 + \sum_{i\neq k\in A_j} a_i\overline{a_k}\langle\psi_i, \psi_k\rangle$$
$$= \sum_{i\in A_j}|a_i|^2 + \left(1-\delta^2\right)\sum_{i\neq k\in A_j} a_i\overline{a_k}\langle\varphi_i, \varphi_k\rangle$$
$$= \sum_{i\in A_j}|a_i|^2 + \left(1-\delta^2\right)\left\langle Q_{A_j}\left(G - D(G)\right)Q_{A_j}x, x\right\rangle$$
$$\leq (1+\epsilon)\sum_{i\in A_j}|a_i|^2.$$

Subtracting $\sum_{i\in A_j}|a_i|^2$ through the inequality yields

$$-\epsilon\sum_{i\in A_j}|a_i|^2 \leq \left(1-\delta^2\right)\left\langle Q_{A_j}\left(G - D(G)\right)Q_{A_j}x, x\right\rangle \leq \epsilon\sum_{i\in A_j}|a_i|^2.$$

That is,

$$\left(1-\delta^2\right)\left|\left\langle Q_{A_j}\left(G - D(G)\right)Q_{A_j}x, x\right\rangle\right| \leq \epsilon\|x\|^2.$$

Since $Q_{A_j}(G - D(G))Q_{A_j}$ is a self-adjoint operator, we have $(1-\delta^2)\|Q_{A_j}(G - D(G))Q_{A_j}\| \leq \epsilon$, i.e., $(1-\delta^2)G$ (and hence G) is pavable. □

Remark 11.1 The proof of (3) ⇒ (1) of Theorem 11.2 illustrates a standard method for turning conjectures about unit norm Riesz basic sequences $\{\psi_i\}_{i\in I}$ into conjectures about unit norm families $\{\varphi_i\}_{i\in I}$ with $T \in B(\ell_2(I))$ and $Te_i = \varphi_i$. Namely, given $\{\varphi_i\}_{i\in I}$ and $0 < \delta < 1$ let $\psi_i = \sqrt{1-\delta^2}f_i \oplus \delta e_i \in \ell_2(I) \oplus \ell_2(I)$. Then $\{\psi_i\}_{i\in I}$ is a unit norm Riesz basic sequence and, for δ small enough, ψ_i is close enough to φ_i to pass inequalities from $\{\psi_i\}_{i\in I}$ to $\{\varphi_i\}_{i\in I}$.

The R_ϵ-conjecture is different from all other conjectures in this chapter in that it does not hold for equivalent norms on the Hilbert space in general. For example, if we renorm ℓ_2 by $|\{a_i\}| = \|\{a_i\}\|_{\ell_2} + \sup_i |a_i|$, then the R_ϵ-conjecture fails for this equivalent norm. To see this, we proceed by way of contradiction and assume there is an $0 < \epsilon < 1$ and an $r = r(\epsilon, 2)$ satisfying the R_ϵ-conjecture. Let $\{e_i\}_{i=1}^{2N}$ be the unit vectors for ℓ_2^{2N} and let $x_i = \frac{e_{2i}+e_{2i+1}}{\sqrt{2}+1}$ for $1 \leq i \leq N$. This is now a unit norm Riesz basic sequence with upper Riesz bound 2. Assume we partition $\{1, 2, \ldots, 2N\}$ into sets $\{A_j\}_{j=1}^r$. Then for some $1 \leq k \leq r$ we have $|A_k| \geq \frac{N}{r}$. Let $A \subset A_k$ with $|A| = \frac{N}{r}$ and $a_i = \frac{1}{\sqrt{N}}$ for $i \in A$. Then

$$\left|\sum_{i\in A} a_i x_i\right| = \frac{1}{\sqrt{2}+1}\left(\sqrt{2} + \frac{r}{\sqrt{N}}\right).$$

Since the norm above is bounded away from one for large N, we cannot satisfy the requirements of the R_ϵ-conjecture. It follows that a positive solution to KS would imply a fundamental new result concerning "inner products," not just norms.

Another important equivalent form of PC comes from [22]. It is, at face value, a significant weakening of the R_ϵ-conjecture while it still remains equivalent to PC.

Conjecture 11.4 *There exist a constant $A > 0$ and a natural number r so that for all natural numbers N and all $T : \ell_2^N \to \ell_2^N$ with $\|Te_i\| = 1$ for all $i = 1, 2, \ldots, N$ and $\|T\| \leq 2$, there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that for all $j = 1, 2, \ldots, r$ and all scalars $\{a_i\}_{i\in A_j}$ we have*

$$\left\|\sum_{i\in A_j} a_i Te_i\right\|^2 \geq A\sum_{i\in A_j} |a_i|^2.$$

Theorem 11.3 *Conjecture 11.4 is equivalent to PC.*

Proof Since PC is equivalent to the R_ϵ-conjecture, which in turn implies Conjecture 11.4, we just need to show that Conjecture 11.4 implies Conjecture 11.1. So choose r, A satisfying Conjecture 11.4. Fix $0 < \delta \leq \frac{3}{4}$ and let P be an orthogonal projection on ℓ_2^N with $\delta(P) \leq \delta$ Now, $\langle Pe_i, e_i\rangle = \|Pe_i\|^2 \leq \delta$ implies

$\|(I-P)e_i\|^2 \geq 1-\delta \geq \frac{1}{4}$. Define $T : \ell_2^N \to \ell_2^N$ by $Te_i = \frac{(I-P)e_i}{\|(I-P)e_i\|}$. For any scalars $\{a_i\}_{i=1}^N$ we have

$$\left\| \sum_{i=1}^N a_i T e_i \right\|^2 = \left\| \sum_{i=1}^N \frac{a_i}{\|(I-P)e_i\|}(I-P)e_i \right\|^2$$

$$\leq \sum_{i=1}^N \left| \frac{a_i}{\|(I-P)e_i\|} \right|^2$$

$$\leq 4 \sum_{i=1}^N |a_i|^2.$$

So $\|Te_i\| = 1$ and $\|T\| \leq 2$. By Conjecture 11.4, there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that for all $j = 1, 2, \ldots, r$ and all scalars $\{a_i\}_{i \in A_j}$ we have

$$\left\| \sum_{i \in A_J} a_i T e_i \right\|^2 \geq A \sum_{i \in A_j} |a_i|^2.$$

Hence,

$$\left\| \sum_{i \in A_j} a_i (I-P)e_i \right\|^2 = \left\| \sum_{i \in A_j} a_i \|(I-P)e_i\| Te_i \right\|^2$$

$$\geq A \sum_{i \in A_j} |a_i|^2 \|(I-P)e_i\|^2$$

$$\geq \frac{A}{4} \sum_{i \in A_j} |a_i|^2.$$

It follows that for all scalars $\{a_i\}_{i \in A_j}$,

$$\sum_{i \in A_j} |a_i|^2 = \left\| \sum_{i \in A_j} a_i P e_i \right\|^2 + \left\| \sum_{i \in A_j} a_i (I-P)e_i \right\|^2$$

$$\geq \left\| \sum_{i \in A_j} a_i P e_i \right\|^2 + \frac{A}{4} \sum_{i \in A_j} |a_i|^2.$$

Now, for all $x = \sum_{i=1}^N a_i e_i$ we have

$$\|P Q_{A_j} x\|^2 = \left\| \sum_{i \in A_j} a_i P e_i \right\|^2 \leq \left(1 - \frac{A}{4}\right) \sum_{i \in A_j} |a_i|^2.$$

Thus,

$$\|Q_{A_j} P Q_{A_j}\| = \|P Q_{A_j}\|^2 \leq 1 - \frac{A}{4}.$$

So Conjecture 11.1 holds. □

Weaver [38] established an important relationship between frames and PC by showing that the following conjecture is equivalent to PC.

Conjecture 11.5 *There are universal constants $B \geq 4$ and $\alpha > \sqrt{B}$ and an $r \in \mathbb{N}$ so that the following holds. Whenever $\{\varphi_i\}_{i=1}^M$ is a unit norm B-tight frame for ℓ_2^N, there exists a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ so that for all $j = 1, 2, \ldots, r$ and all $x \in \ell_2^N$ we have*

$$\sum_{i \in A_j} \left|\langle x, \varphi_i \rangle\right|^2 \leq (B - \alpha)\|x\|^2. \tag{11.1}$$

Using Conjecture 11.5 we can show that the following conjecture is equivalent to PC.

Conjecture 11.6 *There is a universal constant $1 \leq D$ so that for all $T \in B(\ell_2^N)$ with $\|Te_i\| = 1$ for all $i = 1, 2, \ldots, N$, there is an $r = r(\|T\|)$ and a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that for all $j = 1, 2, \ldots, r$ and all scalars $\{a_i\}_{i \in A_j}$*

$$\left\| \sum_{i \in A_j} a_i T e_i \right\|^2 \leq D \sum_{i \in A_j} |a_i|^2.$$

Theorem 11.4 *Conjecture 11.6 is equivalent to PC.*

Proof Since Conjecture 11.3 clearly implies Conjecture 11.6, we just need to show that Conjecture 11.6 implies Conjecture 11.5. So, choose D as in Conjecture 11.6 and choose $B \geq 4$ and $\alpha > \sqrt{B}$ so that $D \leq B - \alpha$. Let $\{\varphi_i\}_{i=1}^M$ be a unit norm B-tight frame for ℓ_2^N. If $Te_i = \varphi_i$ is the synthesis operator for this frame, then $\|T\|^2 = \|T^*\|^2 = B$. So by Conjecture 11.6, there is an $r = r(\|B\|)$ and a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ so that for all $j = 1, 2, \ldots, r$ and all scalars $\{a_i\}_{i \in A_j}$,

$$\left\| \sum_{i \in A_j} a_i T e_i \right\|^2 = \left\| \sum_{i \in A_j} a_i \varphi_i \right\|^2 \leq D \sum_{i \in A_j} |a_i|^2 \leq (B - \alpha) \sum_{i \in A_j} |a_i|^2.$$

So $\|T Q_{A_j}\|^2 \leq B - \alpha$ and for all $x \in \ell_2^N$,

$$\sum_{i \in A_j} \left|\langle x, \varphi_i \rangle\right|^2 = \left\|(Q_{A_j} T)^* x\right\|^2 \leq \|T Q_{A_j}\|^2 \|x\|^2 \leq (B - \alpha)\|x\|^2.$$

This verifies that Conjecture 11.5 holds and so KS holds. □

Remark 11.1 and Conjecture 11.6 show that we only need any universal upper bound in the R_ϵ-conjecture to hold for KS.

11.2.3 The Feichtinger Conjecture

While working in time-frequency analysis, Feichtinger [15] observed that all of the Gabor frames he was using had the property that they could be divided into a finite number of subsets which were Riesz basic sequences. This led to the following conjecture.

Feichtinger Conjecture 11.1 (FC) *Every bounded frame* (*or equivalently*, *every unit norm frame*) *is a finite union of Riesz basic sequences*.

The finite dimensional form of FC is as follows.

Conjecture 11.7 (Finite Dimensional Feichtinger Conjecture) *For every* $B, C > 0$, *there is a natural number* $r = r(B, C)$ *and a constant* $A = A(B, C) > 0$ *so that whenever* $\{\varphi_i\}_{i=1}^N$ *is a frame for* $\mathcal{H}^N$ *with upper frame bound* B *and* $\|\varphi_i\| \geq C$ *for all* $i = 1, 2, \ldots, N$, *then* $\{1, 2, \ldots, N\}$ *can be partitioned into subsets* $\{A_j\}_{j=1}^r$ *so that for each* $1 \leq j \leq r$, $\{\varphi_i\}_{i \in A_j}$ *is a Riesz basic sequence with lower Riesz basis bound* A *and upper Riesz basis bound* B.

There is a significant body of work on this conjecture [6, 7, 15, 27], yet it remains open even for Gabor frames.

We now check that the Feichtinger conjecture is equivalent to PC.

Theorem 11.5 *The following are equivalent*:

(1) *The paving conjecture*.
(2) *The Feichtinger conjecture*.

Proof (1) $\Rightarrow$ (2): Part (2) of Theorem 11.2 is equivalent to PC and clearly implies FC.

(2) $\Rightarrow$ (1): We will observe that FC implies Conjecture 11.4 which is equivalent to PC by Theorem 11.3. In Conjecture 11.4, $\{Te_i\}_{i=1}^N$ is a frame for its span with upper frame bound 2. It is now immediate that the Finite Dimensional Feichtinger Conjecture above implies Conjecture 11.4. □

Another equivalent formulation of KS due to Weaver [38] is the following.

Conjecture 11.8 (KS_r) *There are universal constants* B *and* $\epsilon > 0$ *so that the following holds. Let* $\{\varphi_i\}_{i=1}^M$ *be elements of* ℓ_2^N *with* $\|\varphi_i\| \leq 1$ *for* $i = 1, 2, \ldots, M$ *and*

suppose for every $x \in \ell_2^N$,

$$\sum_{i=1}^{M} \left|\langle x, \varphi_i\rangle\right|^2 \leq B\|x\|^2. \tag{11.2}$$

Then, there is a partition $\{A_j\}_{j=1}^r$ *of* $\{1, 2, \ldots, M\}$ *so that for all* $x \in \ell_2^N$ *and all* $j = 1, 2, \ldots, r$,

$$\sum_{i\in A_j} \left|\langle x, \varphi_i\rangle\right|^2 \leq (B - \epsilon)\|x\|^2.$$

Theorem 11.6 *The following are equivalent*:

(1) *The paving conjecture.*
(2) *Conjecture* KS_r *holds for some* $r \geq 2$.

Proof Assume Conjecture KS_r is true for some fixed r, B, ϵ. We will show that Conjecture 11.1 is true. Let P be an orthogonal projection on $\mathcal{H}_M$ with $\delta(P) \leq \frac{1}{B}$. If P has rank N, then its range is an N-dimensional subspace W of $\mathcal{H}_M$. Define $\varphi_i = \sqrt{B} \cdot Pe_i \in W$ for all $1 \leq i \leq M$. We check that

$$\|\varphi_i\|^2 = B \cdot \|Pe_i\|^2 = B\langle Pe_i, e_i\rangle \leq B\delta(P) \leq 1, \quad \text{for all } i = 1, 2, \ldots, M.$$

Now, if $x \in W$ is any unit vector, then

$$\sum_{i=1}^{M} \left|\langle x, \varphi_i\rangle\right|^2 = \sum_{i=1}^{M} \left|\langle x, \sqrt{B}Pe_i\rangle\right|^2 = B \cdot \sum_{i=1}^{M} \left|\langle x, e_i\rangle\right|^2 = B.$$

By Conjecture KS_r, there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ satisfying for all $1 \leq j \leq r$ and all unit vectors $x \in W$,

$$\sum_{i\in A_j} \left|\langle x, \varphi_i\rangle\right|^2 \leq B - \epsilon.$$

Then $\sum_{j=1}^r Q_{A_j} = Id$, and for any unit vector $x \in W$ we have

$$\begin{aligned}\|Q_{A_j}Px\|^2 &= \sum_{i=1}^{M} \left|\langle Q_{A_j}Px, e_i\rangle\right|^2 = \sum_{i=1}^{M} \left|\langle x, PQ_j e_i\rangle\right|^2 \\ &= \frac{1}{B}\sum_{i\in A_j} \left|\langle x, \varphi_i\rangle\right|^2 \leq \frac{\epsilon}{B}.\end{aligned}$$

Thus Conjecture 11.1 holds and so PC holds.

Conversely, assume KS_r fails for all r. Fix $B = r \geq 2$ and let $\{\varphi_i\}_{i=1}^M$ in $\mathcal{H}^N$ be a counterexample with $\epsilon = 1$. Let $\psi_i = \frac{\varphi_i}{\sqrt{B}}$ and note that $\|\psi_i\psi_i^T\| = \|\psi_i\|^2 \leq \frac{1}{B}$, for

all $i = 1, 2, \ldots, M$ and $\sum_{i=1}^{M} \psi_i \psi_i^T \leq Id$. Then $Id - \sum_{i=1}^{M} \psi_i \psi_i^T$ is a positive finite rank operator, so we can find positive rank one operators $\psi_i \psi_i^T$ for $M+1 \leq i \leq K$ such that $\|\psi_i \psi_i^T\| \leq \frac{1}{B}$ for all $1 \leq i \leq K$ and $\sum_{i=1}^{K} \psi_i \psi_i^T = Id$.

Let T be the analysis operator for $\{\psi_i\}_{i=1}^{K}$, which is an isometry, and if P is the orthogonal projection of $\mathcal{H}_K$ with range $T(\mathcal{H}^N)$, then $Pe_i = T\psi_i$ for all $i = 1, 2, \ldots, K$. Let D be the diagonal matrix with the same diagonal as P. Then

$$\|D\| = \max_{1 \leq i \leq K} \|\psi_i\|^2 \leq \frac{1}{B}.$$

Let $\{Q_j\}_{j=1}^{r}$ be any $K \times K$ diagonal projections which sum to the identity. Define a partition $\{A_j\}_{j=1}^{r}$ of $\{1, 2, \ldots, K\}$ by letting A_j be the diagonal of Q_j. By our choice of $\{\varphi_i\}_{i=1}^{M}$, there exist $1 \leq j \leq r$ and $x \in \mathcal{H}^N$ with $\|x\| = 1$ and

$$\sum_{i \in A_j \cap \{1,2,\ldots,M\}} \left|\langle x, \varphi_i \rangle\right|^2 > B - 1.$$

Hence,

$$\sum_{i \in A_j} \left|\langle x, \psi_i \rangle\right|^2 > 1 - \frac{1}{B}.$$

It follows that for all j,

$$\begin{aligned} \left\|Q_j P(Tx)\right\|^2 &\geq \sum_{i=1}^{K} \left|\left\langle Q_j P(Tx), e_i \right\rangle\right|^2 = \sum_{i \in A_j} \left|\langle Tx, e_i \rangle\right|^2 \\ &= \sum_{i \in A_j} \left|\langle x, \psi_i \rangle\right|^2 > 1 - \frac{1}{B}. \end{aligned}$$

Thus, $\|Q_j P Q_j\| = \|Q_j P\|^2 > 1 - \frac{1}{B}$. Now, the matrix $A = P - D$ has zero diagonal and satisfies $\|A\| \leq 1 + \frac{1}{B}$, and the above shows that for any $K \times K$ diagonal projections $\{Q_j\}_{j=1}^{r}$ with $\sum_{j=1}^{r} Q_j = Id$ we have

$$\|Q_j A Q_j\| \geq \|Q_j P Q_j\| - \|Q_j D Q_j\| \geq 1 - \frac{2}{B}, \quad \text{for some } j.$$

Finally, as $B = r \to \infty$, we obtain a sequence of examples which negate the paving conjecture. □

Weaver [38] also shows that Conjecture $\mathrm{KS_r}$ is equivalent to PC if we assume equality in Eq. (11.2) for all $x \in \ell_2^M$. Weaver further shows that Conjecture $\mathrm{KS_r}$ is equivalent to PC even if we strengthen its assumptions so as to require that the vectors $\{\varphi_i\}_{i=1}^{M}$ are of equal norm and that equality holds in (11.2), but at great cost to our $\epsilon > 0$.

Conjecture 11.9 (KS_r') *There exist universal constants $B \geq 4$ and $\epsilon > \sqrt{B}$ so that the following holds. Let $\{\varphi_i\}_{i=1}^M$ be elements of ℓ_2^N with $\|\varphi_i\| = 1$ for $i = 1, 2, \ldots, M$ and suppose for every $x \in \ell_2^N$,*

$$\sum_{i=1}^{M} \left|\langle x, \varphi_i\rangle\right|^2 = B\|x\|^2. \tag{11.3}$$

Then, there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ so that for all $x \in \ell_2^M$ and all $j = 1, 2, \ldots, r$,

$$\sum_{i \in A_j} \left|\langle x, \varphi_i\rangle\right|^2 \leq (B - \epsilon)\|x\|^2.$$

We introduce one more conjecture.

Conjecture 11.10 *There exist universal constants $0 < \delta, \sqrt{\delta} \leq \epsilon < 1$ and $r \in \mathbb{N}$ so that for all N and all orthogonal projections P on ℓ_2^N with $\delta(P) \leq \delta$ and $\|Pe_i\| = \|Pe_j\|$ for all $i, j = 1, 2, \ldots, N$, there is a paving $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, N\}$ so that $\|Q_{A_j} P Q_{A_j}\| \leq 1 - \epsilon$, for all $j = 1, 2, \ldots, r$.*

Using Conjecture 11.9 we can see that PC is equivalent to Conjecture 11.10.

Theorem 11.7 *PC is equivalent to Conjecture 11.10.*

Proof It is clear that Conjecture 11.1 (which is equivalent to (PC)) implies Conjecture 11.10. So we assume that Conjecture 11.10 holds and we will show that Conjecture 11.9 holds. Let $\{\varphi_i\}_{i=1}^M$ be elements of $\mathcal{H}^N$ with $\|\varphi_i\| = 1$ for $i = 1, 2, \ldots, M$ and suppose for every $x \in \mathcal{H}^N$,

$$\sum_{i=1}^{M} \left|\langle x, \varphi_i\rangle\right|^2 = B\|x\|^2, \tag{11.4}$$

where $\frac{1}{B} \leq \delta$. It follows from Eq. (11.4) that $\{\frac{1}{\sqrt{B}}\varphi_i\}_{i=1}^M$ is an equal norm Parseval frame and so by Naimark's theorem, we may assume there is a larger Hilbert space ℓ_2^M and a projection $P : \ell_2^M \to \mathcal{H}^N$ so that $Pe_i = \varphi_i$ for all $i = 1, 2, \ldots, M$. Now $\|Pe_i\|^2 = \langle Pe_i, e_i\rangle = \frac{1}{B} \leq \delta$ for all $i = 1, 2, \ldots, N$. So by Conjecture 11.10, there is a paving $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ so that $\|Q_{A_j} P Q_{A_j}\| \leq 1 - \epsilon$, for all $j = 1, 2, \ldots, r$. Now for all $1 \leq j \leq r$ and all $x \in \ell_2^N$ we have

$$\begin{aligned}\|Q_{A_j} Px\|^2 &= \sum_{i=1}^{M} \left|\langle Q_{A_j} Px, e_i\rangle\right|^2 \\ &= \sum_{i=1}^{M} \left|\langle x, PQ_{A_j} e_i\rangle\right|^2\end{aligned}$$

$$
\begin{aligned}
&= \frac{1}{B} \sum_{i \in A_j} \left| \langle x, \varphi_i \rangle \right|^2 \\
&\leq \| Q_{A_j} P \|^2 \|x\|^2 \\
&= \| Q_{A_j} P Q_{A_j} \| \|x\|^2 \\
&\leq (1 - \epsilon) \|x\|^2.
\end{aligned}
$$

It follows that for all $x \in \ell_2^N$ we have

$$\sum_{i \in A_j} \left| \langle x \varphi_i \rangle \right|^2 \leq (B - \epsilon B) \|x\|^2.$$

Since $\epsilon B > \sqrt{B}$, we have verified Conjecture 11.9. □

11.2.4 The Bourgain–Tzafriri Conjecture

We start with a fundamental theorem of Bourgain and Tzafriri called the *restricted invertibility principle*. This theorem led to the *(strong and weak) Bourgain–Tzafriri conjectures*. We will see that these conjectures are equivalent to PC.

In 1987, Bourgain and Tzafriri [11] proved a fundamental result in Banach space theory known as the *restricted invertibility principle*.

Theorem 11.8 (Bourgain–Tzafriri) *There is a universal constant $0 < c < 1$ so that whenever $T : \ell_2^N \to \ell_2^N$ is a linear operator for which $\|Te_i\| = 1$, for $1 \leq i \leq N$, then there exists a subset $\sigma \subset \{1, 2, \ldots, N\}$ of cardinality $|\sigma| \geq cN/\|T\|^2$ so that for all choices of scalars $\{a_j\}_{j \in \sigma}$,*

$$\left\| \sum_{j \in \sigma} a_j T e_j \right\|^2 \geq c \sum_{j \in \sigma} |a_j|^2.$$

A close examination of the proof of the theorem [11] yields that c is on the order of 10^{-72}. The proof of the theorem uses probabilistic and function analytic techniques, and it is nontrivial and nonconstructive. A significant breakthrough occurred recently when Spielman and Srivastava [35] presented an *algorithm* for proving the restricted invertibility theorem. Moreover, their proof gives the best possible constants in the theorem.

Theorem 11.9 (Restricted Invertibility Theorem: Spielman–Srivastava Form) *Assume $\{v_i\}_{i=1}^M$ are vectors in ℓ_2^N with $A = \sum_{i=1}^M v_i v_i^T = I$ and $0 < \epsilon < 1$. If $L : \ell_2^N \to \ell_2^N$ is a linear operator, then there is a subset $J \subset \{1, 2, \ldots, M\}$ of size*

$|J| \geq \epsilon^2 \frac{\|L\|_F^2}{\|L\|^2}$ *for which* $\{Lv_i\}_{i\in J}$ *is linearly independent and*

$$\lambda_{min}\left(\sum_{i\in J} Lv_i(Lv_i)^T\right) > \frac{(1-\epsilon)^2\|L\|_F}{M},$$

where $\|L\|_F$ *is the Frobenius norm of* L *and* λ_{min} *is the smallest eigenvalue of the operator computed on span* $\{v_i\}_{i\in J}$.

This *generalized form* of the restricted invertibility theorem was introduced by Vershynin [37], where he studied the contact points of convex bodies using *John's decompositions* of the identity. The corresponding theorem for infinite dimensional Hilbert spaces is still open. But this case requires the set J to be large with respect to the Beurling density [6, 7]. Special cases of this problem were solved in [21, 37].

The inequality in the restricted invertibility theorem is referred to as a *lower* ℓ_2*-bound*. It is known [24, 37] that there is a corresponding close to one *upper* ℓ_2*-bound* which can be achieved in the theorem.

The corresponding theorem for infinite dimensional Hilbert spaces is still open. Special cases of this problem were solved in [21, 37].

Theorem 11.8 gave rise to a problem in the area which has received a great deal of attention [12, 22, 23].

Bourgain–Tzafriri Conjecture 11.1 (BT) *There is a universal constant* $A > 0$ *so that for every* $B > 1$ *there is a natural number* $r = r(B)$ *satisfying*: *For any natural number* N, *if* $T : \ell_2^N \to \ell_2^N$ *is a linear operator with* $\|T\| \leq B$ *and* $\|Te_i\| = 1$ *for all* $i = 1, 2, \ldots, N$, *then there is a partition* $\{A_j\}_{j=1}^r$ *of* $\{1, 2, \ldots, N\}$ *so that for all* $j = 1, 2, \ldots, r$ *and all choices of scalars* $\{a_i\}_{i\in A_j}$ *we have*

$$\left\|\sum_{i\in A_j} a_i Te_i\right\|^2 \geq A\sum_{i\in A_j}|a_i|^2.$$

Sometimes BT is called *strong BT*, since there is a weakening called *weak BT*. In weak BT we allow A to depend upon the norm of the operator T. A significant amount of effort over the years was invested in trying to show that strong and weak BT are equivalent. Casazza and Tremain finally proved this equivalence [22]. We will not have to do any work here, since we developed all of the needed results in earlier sections.

Theorem 11.10 *The following are equivalent*:

(1) *The paving conjecture.*
(2) *The Bourgain–Tzafriri conjecture.*
(3) *The* (*weak*) *Bourgain–Tzafriri conjecture.*

Proof (1) $\Rightarrow$ (2) $\Rightarrow$ (3): The paving conjecture is equivalent to the R_ϵ-conjecture, which clearly implies the Bourgain–Tzafriri conjecture, and this immediately implies the (weak) Bourgain–Tzafriri conjecture.

(3) ⇒ (1): The weak Bourgain–Tzafriri conjecture immediately implies Conjecture 11.4, which is equivalent to the paving conjecture. □

11.2.5 Partitioning Frames into Frame Subsets

A natural and frequently occurring problem in frame theory is to partition a frame into subsets, each of which has *good* frame bounds. This seemingly innocent question turns out to be much deeper than it looks, and as we will now see, it is equivalent to PC.

Conjecture 11.11 *There exists an $\epsilon > 0$ so that for large K, for all N, and all equal norm Parseval frames $\{\varphi_i\}_{i=1}^{KN}$ for ℓ_2^N, there is a nonempty set $J \subset \{1, 2, \ldots, KN\}$ so that both $\{\varphi_i\}_{i\in J}$ and $\{\varphi_i\}_{i\in J^c}$ have lower frame bounds which are greater than ϵ.*

The ideal situation would be for Conjecture 11.11 to hold for all $K \geq 2$. In order for $\{\varphi_i\}_{i\in J}$ and $\{\varphi_i\}_{i\in J^c}$ to both be frames for ℓ_2^N, they at least have to span ℓ_2^N. So the first question is whether we can partition our frame into spanning sets. This follows from a generalization of the Rado-Horn theorem. See the chapter in this book entitled: Spanning and Independence Properties of Finite Frames.

Proposition 11.4 *Every equal norm Parseval frame $\{\varphi_i\}_{i=1}^{KN+L}$, $0 \leq L < N$ for ℓ_2^N can be partitioned into K linearly independent spanning sets plus a linearly independent set of L elements.*

The natural question is whether we can make such a partition so that each of the subsets has *good* frame bounds, i.e., a universal lower frame bound for all subsets. Before addressing this question, we state another conjecture.

Conjecture 11.12 *There exist $\epsilon > 0$ and a natural number r so that for all N, all large K, and all equal norm Parseval frames $\{\varphi_i\}_{i=1}^{KN}$ in ℓ_2^N, there is a partition $\{A_j\}_{j=1}^{r}$ of $\{1, 2, \ldots, KN\}$ so that for all $j = 1, 2, \ldots, r$ the Bessel bound of $\{\varphi_i\}_{i\in A_j}$ is $\leq 1 - \epsilon$.*

We will now establish a relationship between our conjectures and PC.

Theorem 11.11

(1) *Conjecture 11.11 implies Conjecture 11.12.*
(2) *Conjecture 11.12 is equivalent to PC.*

Proof (1): Fix $\epsilon > 0$, r, K as in Conjecture 11.11. Let $\{\varphi_i\}_{i=1}^{KN}$ be an equal norm Parseval frame for an N-dimensional Hilbert space $\mathcal{H}^N$. By Naimark's theorem we may assume there is an orthogonal projection P on ℓ_2^{KN} with $Pe_i = \varphi_i$ for

all $i = 1, 2, \ldots, KN$. By Conjecture 11.11, there is a $J \subset \{1, 2, \ldots, KN\}$ so that $\{Pe_i\}_{i\in J}$ and $\{Pe_i\}_{i\in J^c}$ both have a lower frame bound of $\epsilon > 0$. Hence, for $x \in \mathcal{H}_M = P(\ell_2^{KN})$,

$$\begin{aligned}\|x\|^2 &= \sum_{i=1}^{KN} \big|\langle x, Pe_i\rangle\big|^2 = \sum_{i\in J} \big|\langle x, Pe_i\rangle\big|^2 + \sum_{i\in J^c} \big|\langle x, Pe_i\rangle\big|^2 \\ &\geq \sum_{i\in J} \big|\langle x, Pe_i\rangle\big|^2 + \epsilon\|x\|^2.\end{aligned}$$

That is, $\sum_{i\in J} |\langle x, Pe_i\rangle|^2 \leq (1-\epsilon)\|x\|^2$. So the upper frame bound of $\{Pe_i\}_{i\in J}$ (which is the norm of the analysis operator $(PQ_J)^*$ for this frame) is $\leq 1-\epsilon$. Since PQ_J is the synthesis operator for this frame, we have that $\|Q_J PQ_J\| = \|PQ_J\|^2 = \|(PQ_J)^*\|^2 \leq 1-\epsilon$. Similarly, $\|Q_{J^c} PQ_{J^c}\| \leq 1-\epsilon$. So Conjecture 11.12 holds for $r = 2$.

(2): We will show that Conjecture 11.12 implies Conjecture 11.5. Choose ϵ and r satisfying Conjecture 11.12 for all large K. In particular, choose any K with $\frac{1}{\sqrt{K}} < \alpha$. Let $\{\varphi_i\}_{i=1}^M$ be a unit norm K-tight frame for an N-dimensional Hilbert space $\mathcal{H}^N$. Then $M = \sum_{i=1}^M \|\varphi_i\|^2 = KN$. Since $\{\frac{1}{\sqrt{K}}\varphi_i\}_{i=1}^M$ is an equal norm Parseval frame, by Naimark's theorem we may assume there is an orthogonal projection P on ℓ_2^M with $Pe_i = \frac{1}{\sqrt{K}}\varphi_i$, for $i = 1, 2, \ldots, M$. By Conjecture 11.12 there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, M\}$ so that the Bessel bound $\|(PQ_{A_j})^*\|^2$ for each family $\{\varphi_i\}_{i\in A_j}$ is $\leq 1-\epsilon$. So for $j = 1, 2, \ldots, r$ and any $x \in \ell_2^N$ we have

$$\begin{aligned}\sum_{i\in A_j} \left|\left\langle x, \frac{1}{\sqrt{K}}\varphi_i\right\rangle\right|^2 &= \sum_{i\in A_j} \big|\langle x, PQ_{A_j}e_i\rangle\big|^2 = \sum_{i\in A_j} \big|\langle Q_{A_j}Px, e_i\rangle\big|^2 \leq \|Q_{A_j}Px\|^2 \\ &\leq \|Q_{A_j}P\|^2\|x\|^2 = \big\|(PQ_{A_j})^*\big\|^2\|x\|^2 \leq (1-\epsilon)\|x\|^2.\end{aligned}$$

Hence,

$$\sum_{i\in A_j} \big|\langle x, \varphi_i\rangle\big|^2 \leq K(1-\epsilon)\|x\|^2 = (K - K\epsilon)\|x\|^2.$$

Since $K\epsilon > \sqrt{K}$, we have verified Conjecture 11.5.

For the converse, choose r, δ, ϵ satisfying Conjecture 11.1. If $\{\varphi_i\}_{i=1}^{KN}$ is an equal norm Parseval frame for an N-dimensional Hilbert space $\mathcal{H}^N$ with $\frac{1}{K} \leq \delta$, by Naimark's theorem we may assume we have an orthogonal projection P on ℓ_2^{KN} with $Pe_i = \varphi_i$ for $i = 1, 2, \ldots, KN$. Since $\delta(P) = \|\varphi_i\|^2 \leq \frac{1}{K} \leq \delta$, by Conjecture 11.1 there is a partition $\{A_j\}_{j=1}^r$ of $\{1, 2, \ldots, KN\}$ so that for all $j = 1, 2, \ldots, r$,

$$\|Q_{A_j}PQ_{A_j}\| = \|PQ_{A_j}\|^2 = \big\|(PQ_{A_j})^*\big\|^2 \leq 1-\epsilon.$$

Since $\|(PQ_{A_j})^*\|^2$ is the Bessel bound for $\{Pe_i\}_{i\in A_j} = \{\varphi_i\}_{i\in A_j}$, we have that Conjecture 11.12 holds. □

11.3 The Sundberg Problem

Recently, an apparent weakening of the Kadison–Singer problem has arisen. In his work on interpolation in complex function theory, Sundberg noticed the following problem. Although this is an infinite dimensional problem, we state it here because of its connections to the Kadison–Singer problem.

Problem 11.1 (Sundberg Problem) If $\{\varphi\}_{i=1}^{\infty}$ is a unit norm Bessel sequence, can we partition $\{\varphi_i\}_{i=1}^{\infty}$ into a finite number of non-spanning sets?

This problem appears to be quite innocent, but it is surprisingly difficult. It is immediate that the Feichtinger conjecture implies the Sundberg problem.

Theorem 11.12 *A positive solution to the Feichtinger conjecture implies a positive solution to the Sundberg problem.*

Proof If $\{\varphi_i\}_{i=1}^{\infty}$ is a unit norm Bessel sequence, then by FC, we can partition the natural numbers into a finite number of sets $\{A_j\}_{j=1}^{r}$ so that $\{\varphi_i\}_{i\in A_j}$ is a Riesz sequence for all $j = 1, 2, \ldots, r$. For each $j = 1, 2, \ldots, r$ choose $i_j \in A_j$. Then neither φ_{i_j} nor $\{\varphi_i\}_{i\in A_j\setminus\{i_j\}}$ can span the space. □

11.4 The Paulsen Problem

The Paulsen problem has been intractable for over a dozen years despite receiving quite a bit of attention. In this section we look at the current state of the art on this problem. First we need two definitions.

Definition 11.4 A frame $\{\varphi_i\}_{i=1}^{M}$ for $\mathcal{H}^N$ with frame operator S is said to be ϵ-*nearly equal norm* if

$$(1-\epsilon)\frac{N}{M} \leq \|\varphi_i\|^2 \leq (1+\epsilon)\frac{N}{M}, \quad \text{for all } i = 1, 2, \ldots, M,$$

and it is ϵ-*nearly Parseval* if

$$(1-\epsilon)Id \leq S \leq (1+\epsilon)Id.$$

Definition 11.5 Given frames $\Phi = \{\varphi_i\}_{i=1}^M$ and $\Psi = \{\psi_i\}_{i=1}^M$ for $\mathcal{H}^N$, we define the distance between them by

$$d(\Phi, \Psi) = \sum_{i=1}^{M} \|\varphi_i - \psi_i\|^2.$$

This function is not exactly a distance function, since we have not taken the square root of the right-hand side of the equality. But since this formulation is standard, we will use it. We can now state the Paulsen problem.

Problem 11.2 (Paulsen Problem) How close is an ϵ-nearly equal norm and ϵ-nearly Parseval frame to an equal norm Parseval frame?

The importance of the Paulsen problem is that we have algorithms for constructing frames which are equal norm and nearly Parseval. The question is, if we work with these frames, are we sure that we are working with a frame which is close to some equal norm Parseval frame? We are looking for the function $f(\epsilon, N, M)$ so that every ϵ-nearly equal norm and ϵ-nearly Parseval frame $\Phi = \{\varphi_i\}_{i=1}^M$ satisfies

$$d(\Phi, \Psi) \leq f(\epsilon, N, M),$$

for some equal norm Parseval frame Ψ. A simple compactness argument due to Hadwin (see [10]) shows that such a function must exist.

Lemma 11.1 *The function $f(\epsilon, N, M)$ exists.*

Proof We will proceed by way of contradiction. If this fails, then there is an $0 < \epsilon$ so that for every $\delta = \frac{1}{n}$, there is a frame $\{\varphi_i^n\}_{i=1}^M$ with frame bounds $1 - \frac{1}{n}, 1 + \frac{1}{n}$ and satisfying

$$\left(1 - \frac{1}{n}\right)\frac{N}{M} \leq \|\varphi_i^n\| \leq \left(1 + \frac{1}{n}\right)\frac{N}{M},$$

while $\Phi_n = \{\varphi_i^n\}_{i=1}^M$ is a distance greater than ϵ from any equal norm Parseval frame. By compactness and switching to a subsequence, we may assume that

$$\lim_{n\to\infty} \varphi_i^n = \varphi_i, \quad \text{exists for all } i = 1, 2, \ldots, M.$$

But now $\Phi = \{\varphi_i\}_{i=1}^M$ is an equal norm Parseval frame, contradicting the fact that

$$d(\Phi_n, \Phi) \geq \epsilon > 0, \quad \text{for all } n = 1, 2, \ldots,$$

for any equal norm Parseval frame. □

The problem with this argument is that it does not give any quantitative estimate on the parameters. We do not have a good idea of what form the function $f(\epsilon, N, M)$

must have. We do not even know if M must be in the function or if it is independent of the number of frame vectors. The following example shows that the Paulsen function is certainly a function of the dimension of the space.

Lemma 11.2 *The Paulsen function satisfies*

$$f(\epsilon, N, M) \geq \epsilon^2 N.$$

Proof Fix an $\epsilon > 0$ and an orthonormal basis $\{e_j\}_{j=1}^N$ for $\mathcal{H}^N$. We define a frame $\{\varphi_i\}_{i=1}^{2N}$ by

$$\varphi_i = \begin{cases} \frac{1-\epsilon}{\sqrt{2}} e_i & \text{if } 1 \leq i \leq N, \\ \frac{1+\epsilon}{\sqrt{2}} e_{i-N} & \text{if } N+1 \leq i \leq 2N. \end{cases}$$

By the definition, $\{\varphi_i\}_{i=1}^M$ is ϵ-nearly equal norm. Also, for any $x \in \mathcal{H}^N$ we have

$$\sum_{i=1}^{2N} \big|\langle x, \varphi_i\rangle\big|^2 = \frac{(1-\epsilon)^2}{2} \sum_{i=1}^{N} \big|\langle x, e_i\rangle\big|^2 + \frac{(1+\epsilon)^2}{2} \sum_{i=1}^{N} \big|\langle x, e_i\rangle\big|^2 = \big(1+\epsilon^2\big)\|x\|^2.$$

So $\{\varphi_i\}_{i=1}^{2N}$ is a $(1+\epsilon^2)$ tight frame and hence an ϵ-nearly Parseval frame. The closest equal norm frame to $\{\varphi_i\}_{i=1}^{2N}$ is $\{\frac{e_i}{\sqrt{2}}\}_{i=1}^N \cup \{\frac{e_i}{\sqrt{2}}\}_{i=1}^N$. Also,

$$\sum_{i=1}^{N} \left\| \frac{e_i}{\sqrt{2}} - \varphi_i \right\|^2 + \sum_{i=N+1}^{2N} \left\| \frac{e_{i-N}}{\sqrt{2}} - \varphi_i \right\|^2 = \sum_{i=1}^{N} \left\| \frac{\epsilon}{\sqrt{2}} e_i \right\|^2 + \sum_{i=1}^{N} \left\| \frac{\epsilon}{\sqrt{2}} e_i \right\|^2 = \epsilon^2 N. \qquad \square$$

The main difficulty in solving the Paulsen problem is that finding a close equal norm frame to a given frame involves finding a close frame which satisfies a *geometric* condition, while finding a close Parseval frame to a given frame involves satisfying (an algebraic) *spectral condition*. At this time, we lack techniques for combining these two conditions. However, each of them individually has a known solution. That is, we do know the closest equal norm frame to a given frame [14], and we do know the closest Parseval frame to a given frame [5, 10, 14, 20, 31].

Lemma 11.3 *If $\{\varphi_i\}_{i=1}^M$ is an ϵ-nearly equal norm frame in $\mathcal{H}^N$, then the closest equal norm frame to $\{\varphi_i\}_{i=1}^M$ is*

$$\psi_i = a \frac{\varphi_i}{\|\varphi_i\|}, \quad \textit{for } i = 1, 2, \ldots, M,$$

where

$$a = \frac{\sum_{i=1}^M \|\varphi_i\|}{M}.$$

It is well known that for a frame $\{\varphi_i\}_{i=1}^M$ for $\mathcal{H}^N$ with frame operator S, the *closest* Parseval frame to $\{\varphi_i\}_{i=1}^M$ is $\{S^{-1/2}\varphi_i\}_{i=1}^M$ [5, 10, 14, 20, 31]. We will give the version from [10] here.

Proposition 11.5 *Let $\{\varphi_i\}_{i=1}^M$ be a frame for an N-dimensional Hilbert space $\mathcal{H}_N$, with frame operator $S = T^*T$. Then $\{S^{-1/2}\varphi_i\}_{i=1}^M$ is the closest Parseval frame to $\{\varphi_i\}_{i=1}^M$. Moreover, if $\{\varphi_i\}_{i=1}^M$ is an ϵ-nearly Parseval frame, then*

$$\sum_{i=1}^M \left\| S^{-1/2}\varphi_i - \varphi_i \right\|^2 \leq N(2 - \epsilon - 2\sqrt{1-\epsilon}) \leq N\epsilon^2/4.$$

Proof We first check that $\{S^{-1/2}\varphi_i\}_{i=1}^M$ is the closest Parseval frame to $\{\varphi_i\}_{i=1}^M$.

The squared ℓ^2-distance between $\{\varphi_i\}_{i=1}^M$ and any Parseval frame $\{\psi_j\}_{j=1}^n$ can be expressed in terms of their analysis operators T and T_1 as

$$\begin{aligned}\|\mathcal{F} - \mathcal{G}\|^2 &= \operatorname{Tr}\big[(T - T_1)(T - T_1)^*\big] \\ &= \operatorname{Tr}\big[TT^*\big] + \operatorname{Tr}\big[T_1T_1^*\big] - 2\Re\operatorname{Tr}\big[TT_1^*\big].\end{aligned}$$

Choosing a Parseval frame $\{\psi_i\}_{i=1}^M$ is equivalent to choosing the isometry T_1. To minimize the distance over all choices of T_1, consider the polar decomposition $T = UP$, where P is positive and U is an isometry. In fact, $S = T^*T$ implies $P = S^{1/2}$, which means its eigenvalues are bounded away from zero.

Since P is positive and bounded away from zero, the term $\operatorname{Tr}[TT_1^*] = \operatorname{Tr}[UPT_1^*] = \operatorname{Tr}[T_1^*UP]$ is an inner product between T_1 and U. Its magnitude is bounded by the Cauchy-Schwarz inequality, and thus it has a maximal real part if $T_1 = U$, which implies $T_1^*U = I$. In this case, $T = T_1P = T_1S^{1/2}$ or, equivalently, $T_1^* = S^{-1/2}T^*$, and we conclude that $\psi_i = S^{-1/2}\varphi_i$ for all $i = 1, 2, \ldots M$.

After choosing $T_1 = TS^{-1/2}$, the ℓ^2-distance is expressed in terms of the eigenvalues $\{\lambda_j\}_{j=1}^N$ of $S = T^*T$ by

$$\begin{aligned}\|\mathcal{F} - \mathcal{G}\|^2 &= \operatorname{Tr}[S] + \operatorname{Tr}[I] - 2\operatorname{Tr}\big[S^{1/2}\big] \\ &= \sum_{j=1}^N \lambda_j + N - 2\sum_{j=1}^N \sqrt{\lambda_j}.\end{aligned}$$

If $1 - \epsilon \leq \lambda_j \leq 1 + \epsilon$ for all $j = 1, 2, \ldots N$, calculus shows that the maximum of $\lambda_j - 2\sqrt{\lambda_j}$ is achieved when $\lambda_j = 1 - \epsilon$.

Consequently,

$$\|\mathcal{F} - \mathcal{G}\|^2 \leq 2N - N\epsilon - 2N\sqrt{1-\epsilon}.$$

Estimating $\sqrt{1-\epsilon}$ by the first three terms in its decreasing power series gives the inequality $\|\mathcal{F} - \mathcal{G}\|^2 \leq N\epsilon^2/4$. □

It can be shown that the estimate above is exact, and so we have separate verification that the closeness function is a function of N.

There is a simple algorithm for turning any frame into an equal norm frame with the same frame operator due to Holmes and Paulsen [30].

Proposition 11.6 *There is an algorithm for turning any frame into an equal norm frame with the same frame operator.*

Proof Let $\{\varphi_i\}_{i=1}^M$ be a frame for $\mathcal{H}^N$ with frame operator S and analysis operator T. Then

$$\sum_{i=1}^{M} \|\varphi_i\|^2 = \operatorname{Tr} S.$$

Let $\lambda = \frac{\operatorname{Tr} S}{M}$. If $\|\varphi_i\|^2 = \lambda$, for all $m = 1, 2, \ldots, M$, then we are done. Otherwise, there exists $1 \leq i \neq j \leq M$ with $\|\varphi_i\|^2 > \lambda > \|\varphi_j\|^2$. For any θ, replace the vectors φ_i, φ_j by the vectors

$$\psi_i = (\cos\theta)\varphi_i - (\sin\theta)\varphi_j, \quad \psi_j = (\sin\theta)\varphi_i + (\cos\theta)\varphi_j, \quad \psi_k = \varphi_k, \quad \text{for } k \neq i, j.$$

Now, the analysis operator for $\{\psi_i\}_{i=1}^M$ is $T_1 = UT$ for a unitary operator U on ℓ_2^N given by the Givens rotation. Hence, $T_1^* T_1 = T^* U^* U T = T^* T = S$; so the frame operator is unchanged for any value of θ. Now choose the θ yielding $\|\psi_i\|^2 = \lambda$. Repeating this process at most $M - 1$ times yields an equal norm frame with the same frame operator as $\{\varphi_i\}_{i=1}^M$. □

Using a Parseval frame in Proposition 11.6, we do obtain an equal norm Parseval frame. The problem, again, is that we do not have any quantitative measure of how close these two Parseval frames are.

There is an obvious approach toward solving the Paulsen problem. Given an ϵ-nearly equal norm ϵ-nearly Parseval frame $\{\varphi_i\}_{i=1}^M$ for $\mathcal{H}^N$ with frame operator S, we can switch to the closest Parseval frame $\{S^{-1/2}\varphi_i\}_{i=1}^M$. Then switch to the closest equal norm frame to $\{S^{-1/2}\varphi_i\}_{i=1}^M$, and call it $\{\psi_i\}_{i=1}^M$ with frame operator S_1. Now switch to $\{S_1^{-1/2}\psi_i\}_{i=1}^M$ and again switch to the closest equal norm frame and continue. Unfortunately, even if we could show that this process converges and we could check the distance traveled through this process, we would still not have an answer to the Paulsen problem, because this process does not have to converge to an equal norm Parseval frame. In particular, there is a fixed point of this process which is not an equal norm Parseval frame.

Example 11.1 Let $\{e_i\}_{i=1}^N$ be an orthonormal basis for ℓ_2^N and let $\{\varphi_i\}_{i=1}^{N+1}$ be an equiangular unit norm tight frame for ℓ_2^N. Then $\{e_i \oplus 0\}_{i=1}^N \cup \{0 \oplus \varphi_i\}_{i=1}^{N+1}$ in $\ell_2^N \oplus \ell_2^N$ is an $\epsilon = \frac{1}{N}$-nearly equal norm and $\frac{1}{N}$-nearly Parseval frame with frame operator, say S. A direct calculation shows that taking $S^{-1/2}$ of the frame vectors and switching to the closest equal norm frame leaves the frame unchanged.

The Paulsen problem has proven to be intractable for over 12 years. Recently, two partial solutions to the problem were given in [10, 18], each with its advantages. Since each of these papers is technical, we will only outline the ideas here.

In [10], a new technique is introduced. This is a system of vector-valued ordinary differential equations (ODEs) which starts with a given Parseval frame and has the property that all frames in the flow are still Parseval while approaching an equal norm Parseval frame. The authors then bound the arc length of the system of ODEs by the *frame energy*. Finally, giving an exponential bound on the frame energy, they have a quantitative estimate for the distance between the initial, ϵ-nearly equal norm and ϵ-nearly Parseval frame $\mathcal{F}$ and the equal norm Parseval frame $\mathcal{G}$. For the method to work, they must assume that the dimension N of the Hilbert space and the number M of frame vectors are relatively prime. The authors show that in *practice*, this is not a serious restriction. The main result of [10] is the following.

Theorem 11.13 *Let $N, M \in \mathbb{N}$ be relatively prime, let $0 < \epsilon < \frac{1}{2}$, and assume $\Phi = \{\varphi_i\}_{i=1}^M$ is an ϵ-nearly equal norm and ϵ-nearly Parseval frame for a real or complex Hilbert space of dimension N. Then there is an equal norm Parseval frame $\Psi = \{\psi_i\}_{i=1}^M$ such that*

$$\|\Phi - \Psi\| \leq \frac{29}{8} N^2 M (M-1)^8 \epsilon.$$

In [18], the authors present a new iterative algorithm—gradient descent of the frame potential—for increasing the degree of tightness of any finite unit norm frame. The algorithm itself is trivial to implement, and it preserves certain group structures present in the initial frame. In the special case where the number of frame elements is relatively prime to the dimension of the underlying space, they show that this algorithm converges to a unit norm tight frame at a linear rate, provided the initial unit norm frame is already sufficiently close to being tight. The main difference between this approach and the approach in [10] is that in [10], the authors start with a nearly equal norm Parseval frame and improve its closeness to an equal norm frame while maintaining Parseval, and in [18] the authors start with an equal norm nearly Parseval frame and give an algorithm for improving its *algebraic* properties while changing its transform as little as possible. The main result from [18] is the following theorem.

Theorem 11.14 *Suppose M and N are relatively prime. Pick $t \in (0, \frac{1}{2M})$, and let $\Phi_0 = \{\varphi_i\}_{i=1}^M$ be a unit norm frame with analysis operator T_0 satisfying $\|T_0^* T_0 - \frac{M}{N} I\|_{\text{HS}}^2 \leq \frac{2}{N^3}$. Now, iterate the gradient descent of the frame potential method to obtain Φ^k. Then $\Phi_\infty := \lim_k \Phi_k$ exists and is a unit norm tight frame satisfying*

$$\|\Phi_\infty - \Phi_0\|_{\text{HS}} \leq \frac{4N^{20} M^{8.5}}{1 - 2Mt} \left\| T_0^* T_0 - \frac{M}{N} I \right\|_{\text{HS}}.$$

In [10], the authors showed that there is a connection between the Paulsen problem and a fundamental open problem in operator theory.

Problem 11.3 (Projection Problem) Let $\mathcal{H}^N$ be an N-dimensional Hilbert space with orthonormal basis $\{e_i\}_{i=1}^N$. Find the function $g(\epsilon, N, M)$ satisfying the following. If P is a projection of rank M on $\mathcal{H}^N$ satisfying

$$(1-\epsilon)\frac{M}{N} \leq \|Pe_i\|^2 \leq (1+\epsilon)\frac{M}{N}, \quad \textit{for all } i = 1, 2, \ldots, N,$$

then there is a projection Q with $\|Qe_i\|^2 = \frac{M}{N}$ for all $i = 1, 2, \ldots, N$ satisfying

$$\sum_{i=1}^N \|Pe_i - Qe_i\|^2 \leq g(\epsilon, N, M).$$

In [13], it was shown that the Paulsen problem is equivalent to the projection problem and that their closeness functions are within a factor of 2 of one another. The proof of this result gives several exact connections between the distance between frames and the distance between the ranges of their analysis operators.

Theorem 11.15 *Let $\Phi = \{\varphi_i\}_{i\in I}, \Psi = \{\psi_i\}_{i\in I}$ be Parseval frames for a Hilbert space $\mathcal{H}$ with analysis operators T_1, T_2 respectively. If*

$$d(\Phi, \Psi) = \sum_{i\in I} \|\varphi_i - \psi_i\|^2 < \epsilon,$$

then

$$d\big(T_1(\Phi), T_2(\Psi)\big) = \sum_{i\in I} \|T_1\varphi_i - T_2\psi_i\|^2 < 4\epsilon.$$

Proof Note that for all $j \in I$,

$$T_1\varphi_j = \sum_{i\in I} \langle \varphi_j, \varphi_i\rangle e_i, \quad \text{and} \quad T_2\psi_j = \sum_{i\in I} \langle \psi_j, \psi_i\rangle e_i.$$

Hence,

$$\begin{aligned}
\|T_1\varphi_j - T_2\psi_j\|^2 &= \sum_{i\in I} \big|\langle \varphi_j, \varphi_i\rangle - \langle \psi_j, \psi_i\rangle\big|^2 \\
&= \sum_{i\in I} \big|\langle \varphi_j, \varphi_i - \psi_i\rangle + \langle \varphi_j - \psi_j, \psi_i\rangle\big|^2 \\
&\leq 2\sum_{i\in I} \big|\langle \varphi_j, \varphi_i - \psi_i\rangle\big|^2 + 2\sum_{i\in I} \big|\langle \varphi_j - \psi_j, \psi_i\rangle\big|^2.
\end{aligned}$$

Summing over j and using the fact that our frames Φ and Ψ are Parseval gives

$$\begin{aligned}\sum_{j\in I}\|T_1\varphi_j - T_2\psi_j\|^2 &\le 2\sum_{j\in I}\sum_{i\in I}\big|\langle\varphi_j,\varphi_i-\psi_i\rangle\big|^2 + 2\sum_{j\in I}\sum_{i\in I}\big|\langle\varphi_j-\psi_j,\psi_i\rangle\big|^2\\ &= 2\sum_{i\in I}\sum_{j\in I}\big|\langle\varphi_j,\varphi_i-\psi_i\rangle\big|^2 + 2\sum_{j\in I}\|\varphi_j-\psi_j\|^2\\ &= 2\sum_{i\in I}\|\varphi_i-\psi_i\|^2 + 2\sum_{j\in I}\|\varphi_j-\psi_j\|^2\\ &= 4\sum_{j\in I}\|\varphi_j-\psi_j\|^2.\end{aligned}$$

□

Next, we want to relate the *chordal distance* between two subspaces to the distance between their orthogonal projections. First we need to define the distance between projections.

Definition 11.6 If P, Q are projections on $\mathcal{H}^N$, we define the distance between them by

$$d(P,Q) = \sum_{i=1}^{M}\|Pe_i - Qe_i\|^2,$$

where $\{e_i\}_{i=1}^N$ is an orthonormal basis for $\mathcal{H}^N$.

The *chordal distance* between subspaces of a Hilbert space was defined in [25] and has been shown to have many uses over the years.

Definition 11.7 Given M-dimensional subspaces W_1, W_2 of a Hilbert space, define the M-tuple $(\sigma_1,\sigma_2,\ldots,\sigma_M)$ as follows:

$$\sigma_1 = \max\big\{\langle x,y\rangle : x\in Sp_{W_1},\ y\in Sp_{W_2}\big\} = \langle x_1,y_1\rangle,$$

where Sp_W is the unit sphere of the subspace W. For $2\le i\le M$,

$$\sigma_i = \max\big\{\langle x,y\rangle : \|x\|=\|y\|=1,\ \langle x_j,x\rangle = 0 = \langle y_j,y\rangle,\ \text{for } 1\le j\le i-1\big\},$$

where

$$\sigma_i = \langle x_i,y_i\rangle.$$

The M-tuple $(\theta_1,\theta_2,\ldots,\theta_M)$ with $\theta_i = \cos^{-1}(\sigma_i)$ is called the *principal angles* between W_1, W_2. The *chordal distance* between W_1, W_2 is given by

$$d_c^2(W_1,W_2) = \sum_{i=1}^{M}\sin^2\theta_i.$$

So by the definition, there exist orthonormal bases $\{a_j\}_{j=1}^M$, $\{b_j\}_{j=1}^M$ for W_1, W_2 respectively satisfying

$$\|a_j - b_j\| = 2\sin\left(\frac{\theta}{2}\right), \quad \text{for all } j = 1, 2, \ldots, M.$$

It follows that for $0 \leq \theta \leq \frac{\pi}{2}$,

$$\sin^2\theta \leq 4\sin^2\left(\frac{\theta}{2}\right) = \|a_j - b_j\|^2 \leq 4\sin^2\theta, \quad \text{for all } j = 1, 2, \ldots, M.$$

Hence,

$$d_c^2(W_1, W_2) \leq \sum_{j=1}^M \|a_j - b_j\|^2 \leq 4d_c^2(W_1, W_2). \tag{11.5}$$

We also need the following result [25].

Lemma 11.4 *If $\mathcal{H}^N$ is an N-dimensional Hilbert space and P, Q are rank M orthogonal projections onto subspaces W_1, W_2 respectively, then the chordal distance $d_c(W_1, W_2)$ between the subspaces satisfies*

$$d_c^2(W_1, W_2) = M - \operatorname{Tr} PQ.$$

Next we give a precise connection between chordal distance for subspaces and the distance between the projections onto these subspaces. This result can be found in [25] in the language of Hilbert-Schmidt norms.

Proposition 11.7 *Let $\mathcal{H}^M$ be an M-dimensional Hilbert space with orthonormal basis $\{e_i\}_{i=1}^M$. Let P, Q be the orthogonal projections of $\mathcal{H}^M$ onto N-dimensional subspaces W_1, W_2 respectively. Then the chordal distance between W_1, W_2 satisfies*

$$d_c^2(W_1, W_2) = \frac{1}{2}\sum_{i=1}^M \|Pe_i - Qe_i\|^2.$$

In particular, there are orthonormal bases $\{e_i\}_{i=1}^N$ for W_1 and $\{\tilde{e}_i\}_{i=1}^N$ for W_2 satisfying

$$\frac{1}{2}\sum_{i=1}^M \|Pe_i - Qe_i\|^2 \leq \sum_{i=1}^N \|e_i - \tilde{e}_i\|^2 \leq 2\sum_{i=1}^N \|Pe_i - Qe_i\|^2.$$

Proof We compute

$$\begin{aligned}\sum_{i=1}^M \|Pe_i - Qe_i\|^2 &= \sum_{i=1}^M \langle Pe_i - Qe_i, Pe_i - Qe_i\rangle \\ &= \sum_{i=1}^M \|Pe_i\|^2 + \sum_{i=1}^M \|Qe_i\|^2 - 2\sum_{i=1}^M \langle Pe_i, Qe_i\rangle\end{aligned}$$

$$\begin{aligned}
&= 2N - 2\sum_{i=1}^{M} \langle PQe_i, e_i \rangle \\
&= 2N - 2\,\mathrm{Tr}\, PQ \\
&= 2N - 2\big[N - d_c^2(W_1, W_2)\big] \\
&= 2d_c^2(W_1, W_2).
\end{aligned}$$

This combined with Eq. (11.5) completes the proof. □

The next problem to be addressed is to connect the distance between projections and the distance between the corresponding ranges of analysis operators for Parseval frames.

Theorem 11.16 *Let P, Q be projections of rank N on $\mathcal{H}^M$ and let $\{e_i\}_{i=1}^M$ be the coordinate basis of $\mathcal{H}^M$. Further, assume that there is a Parseval frame $\{\varphi_i\}_{i=1}^M$ for $\mathcal{H}^N$ with analysis operator T satisfying $T\varphi_i = Pe_i$ for all $i = 1, 2, \ldots, M$. If*

$$\sum_{i=1}^{M} \|Pe_i - Qe_i\|^2 < \epsilon,$$

then there is a Parseval frame $\{\psi_i\}_{i=1}^M$ for $\mathcal{H}_M$ with analysis operator T_1 satisfying

$$T_1\psi_i = Qe_i, \quad \text{for all } i = 1, 2, \ldots, M,$$

and

$$\sum_{i=1}^{M} \|\varphi_i - \psi_i\|^2 < 2\epsilon.$$

Moreover, if $\{Qe_i\}_{i=1}^M$ is equal norm, then $\{\psi_i\}_{i=1}^M$ may be chosen to be equal norm.

Proof By Proposition 11.7, there are orthonormal bases $\{a_j\}_{j=1}^M$ and $\{b_j\}_{j=1}^M$ for $W_1 = T(\mathcal{H}^N)$, $W_2 = T_1(\mathcal{H}^N)$ respectively satisfying

$$\sum_{j=1}^{M} \|a_j - b_j\|^2 < 2\epsilon.$$

Let A, B be the $N \times M$ matrices whose jth columns are a_j, b_j respectively. Let a_{ij}, b_{ij} be the (i, j) entries of A, B respectively. Finally, let $\{\varphi_i'\}_{i=1}^M, \{\psi_i'\}_{i=1}^M$ be the ith rows of A, B respectively. Then we have

$$\begin{aligned}
\sum_{i=1}^{M} \|\varphi_i' - \psi_i'\|^2 &= \sum_{i=1}^{M}\sum_{j=1}^{N} |a_{ij} - b_{ij}|^2 \\
&= \sum_{j=1}^{N}\sum_{i=1}^{M} |a_{ij} - b_{ij}|^2
\end{aligned}$$

$$= \sum_{i=1}^{N} \|a_j - b_j\|^2$$
$$\leq 2\epsilon.$$

Since the columns of A form an orthonormal basis for W_1, we know that $\{\varphi_i'\}_{i=1}^M$ is a Parseval frame which is isomorphic to $\{\varphi_i\}_{i=1}^M$. Thus there is a unitary operator $U : \mathcal{H}^M \to \mathcal{H}^M$ with $U\varphi_i' = \varphi_i$. Now let $\{\psi_i\}_{i=1}^M = \{U\psi_i'\}_{i=1}^M$. Then

$$\sum_{i=1}^{M} \|\varphi_i - U\psi_i'\|^2 = \sum_{i=1}^{M} \|U(\varphi_i') - U(\psi_i')\|^2 = \sum_{i=1}^{M} \|\varphi_i' - \psi_i'\|^2 \leq 2\epsilon.$$

Finally, if T_1 is the analysis operator for the Parseval frame $\{\psi_i\}_{i=1}^M$, then T_1 is a isometry and since $T_1\psi_i = Qe_i$, for all $i = 1, 2, \ldots, N$, if Qe_i is equal norm, so is $\{T_1\psi_i\}_{i=1}^M$ and hence so is $\{\psi_i\}_{i=1}^N$. □

Theorem 11.17 *If $g(\epsilon, N, M)$ is the function for the Paulsen problem and $f(\epsilon, N, M)$ is the function for the Projection problem, then*

$$f(\epsilon, N, M) \leq 4g(\epsilon, N, M) \leq 8f(\epsilon, N, M).$$

Proof First, assume that the projection problem holds with function $f(\epsilon, N, M)$. Let $\{\varphi_i\}_{i=1}^M$ be a Parseval frame for $\mathcal{H}^N$ satisfying

$$(1-\epsilon)\frac{N}{M} \leq \|\varphi_i\|^2 \leq (1+\epsilon)\frac{N}{M}.$$

Let T be the analysis operator of $\{\varphi_i\}_{i=1}^M$ and let P be the projection of $\mathcal{H}_M$ onto range T. So, $T\varphi_i = Pe_i$ for all $i = 1, 2, \ldots, M$. By our assumption that the projection problem holds, there is a projection Q on $\mathcal{H}_M$ with constant diagonal so that

$$\sum_{i=1}^{M} \|Pe_i - Qe_i\|^2 \leq f(\epsilon, N, M).$$

By Theorem 11.16, there is a Parseval frame $\{\psi_i\}_{i=1}^M$ for $\mathcal{H}^N$ with analysis operator T_1 so that $T_1\psi_i = Qe_i$ and

$$\sum_{i=1}^{M} \|\varphi_i - \psi_i\|^2 \leq 2f(\epsilon, N, M).$$

Since T_1 is an isometry and $\{T_1\psi_i\}_{i=1}^M$ is equal norm, it follows that $\{\psi_i\}_{i=1}^M$ is an equal norm Parseval frame satisfying the Paulsen problem.

Conversely, assume the Parseval Paulsen problem has a positive solution with function $g(\epsilon, N, M)$. Let P be an orthogonal projection on $\mathcal{H}^M$ satisfying

$$(1-\epsilon)\frac{N}{M} \leq \|Pe_i\|^2 \leq (1+\epsilon)\frac{N}{M}.$$

Then $\{Pe_i\}_{i=1}^M$ is an ϵ-nearly equal norm Parseval frame for $\mathcal{H}^N$ and, by the Paulsen problem, there is an equal norm Parseval frame $\{\psi_i\}_{i=1}^M$ so that

$$\sum_{i=1}^M \|\varphi_i - \psi_i\|^2 < g(\epsilon, N, M).$$

Let T_1 be the analysis operator of $\{\psi_i\}_{i=1}^M$. Letting Q be the projection onto the range of T_1, we have that $Qe_i = T_1\psi_i$, for all $i = 1, 2, \ldots, M$. By Theorem 11.15, we have that

$$\sum_{i=1}^M \|Pe_i - T_1\psi_i\|^2 = \sum_{i=1}^N \|Pe_i - Qe_i\|^2 \leq 4g(\epsilon, N, M).$$

Since T_1 is an isometry and $\{\psi_i\}_{i=1}^M$ is equal norm, it follows that Q is a constant diagonal projection. □

In [13] there are several generalizations of the Paulsen and projection problems.

11.5 Final Comments

We have concentrated here on some problems in pure mathematics which have finite dimensional formulations. There are many other infinite dimensional versions of these problems [22, 23] in sampling theory, harmonic analysis, and other areas which we have not covered.

Because of the long history of these problems and their connections to so many areas of mathematics, we are naturally led to consider the *decidability* of KS. Since we have finite dimensional versions of the problem, it can be reformulated in the language of pure number theory, and hence it has a property logicians call *absoluteness*. As a practical matter, the general feeling is that this means it is very unlikely to be undecidable.

Acknowledgements The author acknowledges support from NSF DMS 1008183, NSF ATD1042701, and AFOSR FA9550-11-1-0245.

References

1. Akemann, C.A., Anderson, J.: Lyapunov theorems for operator algebras. Mem. AMS **94** (1991)

2. Anderson, J.: Restrictions and representations of states on C^*-algebras. Trans. Am. Math. Soc. **249**, 303–329 (1979)
3. Anderson, J.: Extreme points in sets of positive linear maps on $B(\mathcal{H})$. J. Funct. Anal. **31**, 195–217 (1979)
4. Anderson, J.: A conjecture concerning pure states on $B(\mathcal{H})$ and a related theorem. In: Topics in Modern Operator Theory, pp. 27–43. Birkhäuser, Basel (1981)
5. Balan, R.: Equivalence relations and distances between Hilbert frames. Proc. Am. Math. Soc. **127**(8), 2353–2366 (1999)
6. Balan, R., Casazza, P.G., Heil, C., Landau, Z.: Density, overcompleteness and localization of frames. I. Theory. J. Fourier Anal. Appl. **12**, 105–143 (2006)
7. Balan, R., Casazza, P.G., Heil, C., Landau, Z.: Density, overcompleteness and localization of frames. II. Gabor systems. J. Fourier Anal. Appl. **12**, 309–344 (2006)
8. Berman, K., Halpern, H., Kaftal, V., Weiss, G.: Some C_4 and C_6 norm inequalities related to the paving problem. Proc. Symp. Pure Math. **51**, 29–41 (1970)
9. Berman, K., Halpern, H., Kaftal, V., Weiss, G.: Matrix norm inequalities and the relative Dixmier property. Integral Equ. Oper. Theory **11**, 28–48 (1988)
10. Bodmann, B., Casazza, P.G.: The road to equal-norm Parseval frames. J. Funct. Anal. **258**(2), 397–420 (2010)
11. Bourgain, J., Tzafriri, L.: Invertibility of "large" submatrices and applications to the geometry of Banach spaces and harmonic analysis. Isr. J. Math. **57**, 137–224 (1987)
12. Bourgain, J., Tzafriri, L.: On a problem of Kadison and Singer. J. Reine Angew. Math. **420**, 1–43 (1991)
13. Cahill, J., Casazza, P.G.: The Paulsen problem in operator theory, preprint
14. Casazza, P.G.: Custom building finite frames. In: Wavelets, Frames and Operator Theory, College Park, MD, 2003. Contemporary Mathematics, vol. 345, pp. 61–86. Am. Math. Soc., Providence (2004)
15. Casazza, P.G., Christensen, O., Lindner, A., Vershynin, R.: Frames and the Feichtinger conjecture. Proc. Am. Math. Soc. **133**(4), 1025–1033 (2005)
16. Casazza, P.G., Edidin, D.: Equivalents of the Kadison–Singer problem. Contemp. Math. **435**, 123–142 (2007)
17. Casazza, P.G., Edidin, D., Kalra, D., Paulsen, V.: Projections and the Kadison–Singer problem. Oper. Matrices **1**(3), 391–408 (2007)
18. Casazza, P.G., Fickus, M., Mixon, D.: Auto-tuning unit norm frames. Appl. Comput. Harmon. Anal. **32**, 1–15 (2012)
19. Casazza, P.G., Fickus, M., Mixon, D.G., Tremain, J.C.: The Bourgain–Tzafriri conjecture and concrete constructions of non-pavable projections. Oper. Matrices **5**(2), 351–363 (2011)
20. Casazza, P., Kutyniok, G.: A generalization of Gram–Schmidt orthogonalization generating all Parseval frames. Adv. Comput. Math. **18**, 65–78 (2007)
21. Casazza, P.G., Pfander, G.: An infinite dimensional restricted invertibility theorem, preprint
22. Casazza, P.G., Tremain, J.C.: The Kadison–Singer problem in mathematics and engineering. Proc. Natl. Acad. Sci. **103**(7), 2032–2039 (2006)
23. Casazza, P.G., Fickus, M., Tremain, J.C., Weber, E.: The Kadison–Singer problem in mathematics and engineering—a detailed account. In: Han, D., Jorgensen, P.E.T., Larson, D.R. (eds.) Operator Theory, Operator Algebras and Applications. Contemporary Mathematics, vol. 414, pp. 297–356 (2006)
24. Casazza, P.G., Tremain, J.C.: Revisiting the Bourgain–Tzafriri restricted invertibility theorem. Oper. Matrices **3**(1), 97–110 (2009)
25. Conway, J.H., Hardin, R.H., Sloane, N.J.A.: Packing lines, planes, etc.: packings in Grassmannian spaces. Exp. Math. **5**(2), 139–159 (1996)
26. Dirac, P.A.M.: Quantum Mechanics, 3rd edn. Oxford University Press, London (1947)
27. Gröchenig, K.H.: Localized frames are finite unions of Riesz sequences. Adv. Comput. Math. **18**, 149–157 (2003)
28. Halpern, H., Kaftal, V., Weiss, G.: Matrix pavings and Laurent operators. J. Oper. Theory **16**, 121–140 (1986)

29. Halpern, H., Kaftal, V., Weiss, G.: Matrix pavings in $B(\mathcal{H})$. In: Proc. 10th International Conference on Operator Theory, Increst (1985). Adv. Appl. **24**, 201–214 (1987)
30. Holmes, R.B., Paulsen, V.: Optimal frames for erasures. Linear Algebra Appl. **377**, 31–51 (2004)
31. Janssen, A.J.E.M.: Zak transforms with few zeroes and the tie. In: Feichtinger, H.G., Strohmer, T. (eds.) Advances in Gabor Analysis, pp. 31–70. Birkhäuser, Boston (2002)
32. Kadison, R., Singer, I.: Extensions of pure states. Am. J. Math. **81**, 383–400 (1959)
33. Paulsen, V.: A dynamical systems approach to the Kadison–Singer problem. J. Funct. Anal. **255**, 120–132 (2008)
34. Paulsen, V., Ragupathi, M.: Some new equivalences of Anderson's paving conjecture. Proc. Am. Math. Soc. **136**, 4275–4282 (2008)
35. Spielman, D.A., Srivastava, N.: An elementary proof of the restricted invertibility theorem. Isr. J. Math. **19**(1), 83–91 (2012)
36. Tropp, J.: The random paving property for uniformly bounded matrices. Stud. Math. **185**(1), 67–82 (2008)
37. Vershynin, R.: Remarks on the geometry of coordinate projections in $\mathbb{R}^n$. Isr. J. Math. **140**, 203–220 (2004)
38. Weaver, N.: The Kadison–Singer problem in discrepancy theory. Discrete Math. **278**, 227–239 (2004)

Chapter 12
Probabilistic Frames: An Overview

Martin Ehler and Kasso A. Okoudjou

Abstract Finite frames can be viewed as mass points distributed in N-dimensional Euclidean space. As such they form a subclass of a larger and rich class of probability measures that we call probabilistic frames. We derive the basic properties of probabilistic frames, and we characterize one of their subclasses in terms of minimizers of some appropriate potential function. In addition, we survey a range of areas where probabilistic frames, albeit under different names, appear. These areas include directional statistics, the geometry of convex bodies, and the theory of t-designs.

Keywords Probabilistic frame · POVM · Frame potential · Isotropic measure

12.1 Introduction

Finite frames in $\mathbb{R}^N$ are spanning sets that allow the analysis and synthesis of vectors in a way similar to basis decompositions. However, frames are redundant systems, and as such the reconstruction formula they provide is not unique. This redundancy plays a key role in many applications of frames which appear now in a range of areas that include, but are not limited to, signal processing, quantum computing, coding theory, and sparse representations; cf. [11, 22, 23] for an overview.

By viewing the frame vectors as discrete mass distributions on $\mathbb{R}^N$, one can extend frame concepts to probability measures. This point of view was developed in [16] under the name of probabilistic frames and was further expanded in [18]. The goal of this chapter is to summarize the main properties of probabilistic frames and to bring forth their relationship to other areas of mathematics.

M. Ehler (✉)
Institute of Biomathematics and Biometry, Helmholtz Zentrum München, Ingolstädter Landstr. 1, 85764 Neuherberg, Germany
e-mail: martin.ehler@helmholtz-muenchen.de

K.A. Okoudjou
Department of Mathematics, Norbert Wiener Center, University of Maryland, College Park, MD 20742, USA
e-mail: kasso@math.umd.edu

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_12,

The richness of the set of probability measures together with the availability of analytic and algebraic tools make it straightforward to construct many examples of probabilistic frames. For instance, by convolving probability measures, we have been able to generate new probabilistic frames from existing ones. In addition, the probabilistic framework considered in this chapter allows us to introduce a new distance on frames, namely the Wasserstein distance [35], also known as the Earth Mover's distance [25]. Unlike standard frame distances in the literature such as the ℓ_2-distance, the Wasserstein metric enables us to define a meaningful distance between two frames of different cardinalities.

As we shall see later in Sect. 12.4, probabilistic frames are also tightly related to various notions that appeared in areas such as the theory of t-designs [15], positive operator valued measures (POVM) encountered in quantum computing [1, 13, 14], and isometric measures used in the study of convex bodies [19, 20, 28]. In particular, in 1948, F. John [20] gave a characterization of what is known today as unit norm tight frames in terms of an ellipsoid of maximal volume, called John's ellipsoid. The latter and other ellipsoids in some extremal positions are supports of probability measures that turn out to be probabilistic frames. The connections between frames and convex bodies could offer new insight into the construction of frames, on which we plan to elaborate elsewhere.

Finally, it is worth mentioning the connections between probabilistic frames and statistics. For instance, in directional statistics probabilistic tight frames can be used to measure inconsistencies of certain statistical tests. Moreover, in the setting of M-estimators as discussed in [21, 33, 34], finite tight frames can be derived from maximum likelihood estimators that are used for parameter estimation of probabilistic frames.

This chapter is organized as follows. In Sect. 12.2 we define probabilistic frames, prove some of their main properties, and give a few examples. In Sect. 12.3 we introduce the notion of the probabilistic frame potential and characterize its minima in terms of tight probabilistic frames. In Sect. 12.4 we discuss the relationship between probabilistic frames and other areas such as the geometry of convex bodies, quantum computing, the theory of t-designs, directional statistics, and compressed sensing.

12.2 Probabilistic Frames

12.2.1 Definition and Basic Properties

Let $\mathscr{P} := \mathscr{P}(\mathscr{B}, \mathbb{R}^N)$ denote the collection of probability measures on $\mathbb{R}^N$ with respect to the Borel σ-algebra $\mathscr{B}$. Recall that the support of $\mu \in \mathscr{P}$ denoted by $\operatorname{supp}(\mu)$ is the set of all $x \in \mathbb{R}^N$ such that for all open neighborhoods $U_x \subset \mathbb{R}^N$ of x, we have $\mu(U_x) > 0$. We write $\mathscr{P}(K) := \mathscr{P}(\mathscr{B}, K)$ for those probability measures in $\mathscr{P}$ whose support is contained in $K \subset \mathbb{R}^N$. The linear span of $\operatorname{supp}(\mu)$ in $\mathbb{R}^N$ is denoted by E_μ.

Definition 12.1 A Borel probability measure $\mu \in \mathscr{P}$ is a *probabilistic frame* if there exist $0 < A \le B < \infty$ such that

$$A\|x\|^2 \le \int_{\mathbb{R}^N} \left|\langle x, y\rangle\right|^2 d\mu(y) \le B\|x\|^2, \quad \text{for all } x \in \mathbb{R}^N. \tag{12.1}$$

The constants A and B are called *lower and upper probabilistic frame bounds*, respectively. When $A = B$, μ is called a *tight probabilistic frame*. If only the upper inequality holds, then we call μ a *Bessel probability measure*.

As mentioned in the Introduction, this notion was introduced in [16] and was further developed in [18]. We shall see later in Sect. 12.2.2 that probabilistic frames provide reconstruction formulas similar to those known from finite frames. We begin by giving a complete characterization of probabilistic frames, for which we first need some preliminary definitions.

Let

$$\mathscr{P}_2 := \mathscr{P}_2\big(\mathbb{R}^N\big) = \left\{\mu \in \mathscr{P} : M_2^2(\mu) := \int_{\mathbb{R}^N} \|x\|^2 \, d\mu(x) < \infty\right\} \tag{12.2}$$

be the (convex) set of all probability measures with finite second moments. There exists a natural metric on $\mathscr{P}_2$ called the 2-*Wasserstein metric,* which is given by

$$W_2^2(\mu, \nu) := \min\left\{\int_{\mathbb{R}^N \times \mathbb{R}^N} \|x - y\|^2 \, d\gamma(x, y), \gamma \in \Gamma(\mu, \nu)\right\}, \tag{12.3}$$

where $\Gamma(\mu, \nu)$ is the set of all Borel probability measures γ on $\mathbb{R}^N \times \mathbb{R}^N$ whose marginals are μ and ν, respectively, i.e., $\gamma(A \times \mathbb{R}^N) = \mu(A)$ and $\gamma(\mathbb{R}^N \times B) = \nu(B)$ for all Borel subsets A, B in $\mathbb{R}^N$. The Wasserstein distance represents the "work" that is needed to transfer the mass from μ into ν, and each $\gamma \in \Gamma(\mu, \nu)$ is called a transport plan. We refer to [2, Chap. 7], [35, Chap. 6] for more details on the Wasserstein spaces.

Theorem 12.1 *A Borel probability measure $\mu \in \mathscr{P}$ is a* probabilistic frame *if and only if $\mu \in \mathscr{P}_2$ and $E_\mu = \mathbb{R}^N$. Moreover, if μ is a tight probabilistic frame, then the frame bound A is given by $A = \frac{1}{N} M_2^2(\mu) = \frac{1}{N} \int_{\mathbb{R}^N} \|y\|^2 \, d\mu(y)$.*

Proof Assume first that μ is a probabilistic frame, and let $\{e_i\}_{i=1}^N$ be an orthonormal basis for $\mathbb{R}^N$. By letting $x = e_i$ in (12.1), we have $A \le \int_{\mathbb{R}^N} |\langle e_i, y\rangle|^2 \, d\mu(y) \le B$. Summing these inequalities over i leads to $A \le \frac{1}{N} \int_{\mathbb{R}^N} \|y\|^2 \, d\mu(y) \le B < \infty$, which proves that $\mu \in \mathscr{P}_2$. Note that the latter inequalities also prove the second part of the theorem.

To prove $E_\mu = \mathbb{R}^N$, we assume that $E_\mu^\perp \ne \{0\}$ and choose $0 \ne x \in E_\mu^\perp$. The left-hand side of (12.1) then yields a contradiction.

For the reverse implication, let $M_2(\mu) < \infty$ and $E_\mu = \mathbb{R}^N$. The upper bound in (12.1) is obtained by a simple application of the Cauchy-Schwarz inequality with

$B = \int_{\mathbb{R}^N} \|y\|^2 \, d\mu(y)$. To obtain the lower frame bound, let

$$A := \inf_{x \in \mathbb{R}^N} \left(\frac{\int_{\mathbb{R}^N} |\langle x, y\rangle|^2 \, d\mu(y)}{\|x\|^2} \right) = \inf_{x \in S^{N-1}} \left(\int_{\mathbb{R}^N} \left|\langle x, y\rangle\right|^2 d\mu(y) \right).$$

Due to the dominated convergence theorem, the mapping $x \mapsto \int_{\mathbb{R}^N} |\langle x, y\rangle|^2 \, d\mu(y)$ is continuous and the infimum is in fact a minimum since the unit sphere S^{N-1} is compact. Let x_0 be in S^{N-1} such that

$$A = \int_{\mathbb{R}^N} \left|\langle x_0, y\rangle\right|^2 d\mu(y).$$

We need to verify that $A > 0$: Since $E_\mu = \mathbb{R}^N$, there is $y_0 \in \operatorname{supp}(\mu)$ such that $|\langle x_0, y_0\rangle|^2 > 0$. Therefore, there is $\varepsilon > 0$ and an open subset $U_{y_0} \subset \mathbb{R}^N$ satisfying $y_0 \in U_{y_0}$ and $|\langle x, y\rangle|^2 > \varepsilon$, for all $y \in U_{y_0}$. Since $\mu(U_{y_0}) > 0$, we obtain $A \geq \varepsilon \mu(U_{y_0}) > 0$, which concludes the proof of the first part of the theorem. □

Remark 12.1 A tight probabilistic frame μ with $M_2(\mu) = 1$ will be referred to as a *unit norm tight probabilistic frame*. In this case the frame bound is $A = \frac{1}{N}$, which only depends on the dimension of the ambient space. In fact, any tight probabilistic frame μ whose support is contained in the unit sphere S^{N-1} is a unit norm tight probabilistic frame.

In the sequel, the Dirac measure supported at $\varphi \in \mathbb{R}^N$ is denoted by δ_φ.

Proposition 12.1 *Let $\Phi = (\varphi_i)_{i=1}^M$ be a sequence of nonzero vectors in $\mathbb{R}^N$, and let $\{a_i\}_{i=1}^M$ be a sequence of positive numbers.*

(a) *Φ is a frame with frame bounds $0 < A \leq B < \infty$ if and only if $\mu_\Phi := \frac{1}{M} \sum_{i=1}^M \delta_{\varphi_i}$ is a probabilistic frame with bounds A/M and B/M.*
(b) *Moreover, the following statements are equivalent:*
  (i) *Φ is a (tight) frame.*
  (ii) *$\mu^\Phi := \frac{1}{\sum_{i=1}^M \|\varphi_i\|^2} \sum_{i=1}^M \|\varphi_i\|^2 \delta_{\frac{\varphi_i}{\|\varphi_i\|}}$ is a (tight) unit norm probabilistic frame.*
  (iii) *$\frac{1}{\sum_{i=1}^M a_i^2} \sum_{i=1}^M a_i^2 \delta_{\frac{\varphi_i}{a_i}}$ is a (tight) probabilistic frame.*

Proof Clearly, μ_Φ is a probability measure, and its support is the set $\{\varphi_k\}_{k=1}^M$, which spans $\mathbb{R}^N$. Moreover,

$$\int_{\mathbb{R}^N} \langle x, y\rangle^2 \, d\mu_\Phi(y) = \frac{1}{M} \sum_{i=1}^M \langle x, \varphi_i\rangle^2.$$

Part (a) can be easily derived from the above equality, and direct calculations imply the remaining equivalences. □

Remark 12.2 Though the frame bounds of μ_Φ are smaller than those of Φ, we observe that the ratios of the respective frame bounds remain the same.

Example 12.1 Let dx denote the Lebesgue measure on $\mathbb{R}^N$ and assume that f is a positive Lebesgue integrable function such that $\int_{\mathbb{R}^N} f(x)\,dx = 1$. If $\int_{\mathbb{R}^N} \|x\|^2 f(x)\,dx < \infty$, then the measure μ defined by $d\mu(x) = f(x)\,dx$ is a (Borel) probability measure that is a probabilistic frame. Moreover, if $f(x_1, \ldots, x_N) = f(\pm x_1, \ldots, \pm x_N)$, for all combinations of $\pm$, then μ is a tight probabilistic frame; cf. Proposition 3.13 in [16]. The latter is satisfied, for instance, if f is radially symmetric; i.e., there is a function g such that $f(x) = g(\|x\|)$.

Viewing frames in the probabilistic setting that we have been developing has several advantages. For instance, we can use measure theoretical tools to generate new probabilistic frames from old ones and, in fact, under some mild conditions, the convolution of probability measures leads to probabilistic frames. Recall that the convolution of $\mu, \nu \in \mathscr{P}$ is the probability measure given by $\mu * \nu(A) = \int_{\mathbb{R}^N} \mu(A - x)\,d\nu(x)$ for $A \in \mathscr{B}$. Before we state the result on convolution of probabilistic frames, we need a technical lemma that is related to the support of a probability measure that we consider later. The result is an analog of the fact that adding finitely many vectors to a frame does not change the frame nature, but affects only its bounds. In the case of probabilistic frames, the adjunction of a single point (or finitely many points) to its support does not destroy the frame property, but just changes the frame bounds:

Lemma 12.1 *Let μ be a Bessel probability measure with bound $B > 0$. Given $\varepsilon \in (0, 1)$, set $\mu_\varepsilon = (1 - \varepsilon)\mu + \varepsilon\delta_0$. Then μ_ε is a Bessel measure with bound $B_\varepsilon = (1 - \varepsilon)B$. If in addition μ is a probabilistic frame with bounds $0 < A \le B < \infty$, then μ_ε is also a probabilistic frame with bounds $(1 - \varepsilon)A$ and $(1 - \varepsilon)B$.*

In particular, if μ is a tight probabilistic frame with bound A, then so is μ_ε with bound $(1 - \varepsilon)A$.

Proof μ_ε is clearly a probability measure since it is a convex combination of probability measures. The proof of the lemma follows from the following equations:

$$\begin{aligned}\int_{\mathbb{R}^N} \left|\langle x, y\rangle\right|^2 d\mu_\varepsilon(y) &= (1 - \varepsilon) \int_{\mathbb{R}^N} \left|\langle x, y\rangle\right|^2 d\mu(y) + \varepsilon \int_{\mathbb{R}^N} \left|\langle x, y\rangle\right|^2 d\delta_0(y) \\ &= (1 - \varepsilon) \int_{\mathbb{R}^N} \left|\langle x, y\rangle\right|^2 d\mu(y).\end{aligned}$$

□

We are now ready to understand the action of convolution on probabilistic frames.

Theorem 12.2 *Let $\mu \in \mathscr{P}_2$ be a probabilistic frame and let $\nu \in \mathscr{P}_2$. If* $\operatorname{supp}(\mu)$ *contains at least $N + 1$ distinct vectors, then $\mu * \nu$ is a probabilistic frame.*

Proof

$$\begin{aligned} M_2^2(\mu * \nu) &= \int_{\mathbb{R}^N} \|y\|^2 \, d\mu * \nu(y) \\ &= \iint_{\mathbb{R}^N \times \mathbb{R}^N} \|x+y\|^2 \, d\mu(x)\, d\nu(y) \\ &\leq M_2^2(\mu) + M_2^2(\nu) + 2M_2(\mu)M_2(\nu) \\ &= \big(M_2(\mu) + M_2(\nu)\big)^2 < \infty. \end{aligned}$$

Thus, $\mu * \nu \in \mathscr{P}_2$, and it only remains to verify that the support of $\mu * \nu$ spans $\mathbb{R}^N$; cf. Theorem 12.1. Since $\operatorname{supp}(\mu)$ must span $\mathbb{R}^N$, there are $\{\varphi_i\}_{i=1}^{N+1} \subset \operatorname{supp}(\mu)$ that form a frame for $\mathbb{R}^N$. Due to their linear dependency, for each $x \in \mathbb{R}^N$, we can find $\{c_i\}_{i=1}^{N+1} \subset \mathbb{R}$ such that $x = \sum_{i=1}^{N+1} c_i \varphi_i$ with $\sum_{i=1}^{N+1} c_i = 0$. For $y \in \operatorname{supp}(\nu)$, we then obtain

$$x = x + 0y = \sum_{i=1}^{N+1} c_i \varphi_i + \sum_{i=1}^{N+1} c_i y = \sum_{i=1}^{N+1} c_i (\varphi_i + y) \in \operatorname{span}\big(\operatorname{supp}(\mu) + \operatorname{supp}(\nu)\big).$$

Thus, $\operatorname{supp}(\mu) \subset \operatorname{span}(\operatorname{supp}(\mu) + \operatorname{supp}(\nu))$. Since $\operatorname{supp}(\mu) + \operatorname{supp}(\nu) \subset \operatorname{supp}(\mu * \nu)$, we can conclude the proof. □

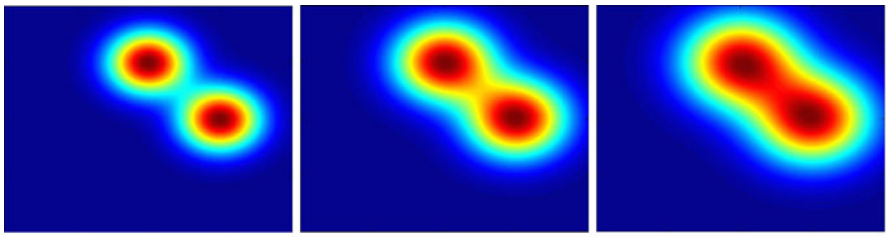
Remark 12.3 By Lemma 12.1 we can assume without loss of generality that $0 \in \operatorname{supp}(\nu)$. In this case, if μ is a probabilistic frame such that $\operatorname{supp}(\mu)$ does not contain $N+1$ distinct vectors, then $\mu * \nu$ is still a probabilistic frame. Indeed, $0 \in \operatorname{supp}(\nu)$ and $E_\mu = \mathbb{R}^N$ together with the fact that $\operatorname{supp}(\mu) + \operatorname{supp}(\nu) \subset \operatorname{supp}(\mu * \nu)$ imply that $\operatorname{supp}(\mu * \nu)$ also spans $\mathbb{R}^N$.

Finally, if μ is a probabilistic frame such that $\operatorname{supp}(\mu)$ does not contain $N+1$ distinct vectors, then $\operatorname{supp}(\mu) = \{\varphi_j\}_{j=1}^N$ forms a basis for $\mathbb{R}^N$. In this case, $\mu * \nu$ is not a probabilistic frame if $\nu = \delta_{-x}$, where x is an affine linear combination of $\{\varphi_j\}_{j=1}^N$. Indeed, $x = \sum_{j=1}^N c_j \varphi_j$ with $\sum_{j=1}^N c_j = 1$ implies $\sum_{j=1}^N c_j(\varphi_j - x) = 0$, although not all c_j can be zero. Therefore, $\operatorname{supp}(\mu * \nu) = \{\varphi_j - x\}_{j=1}^N$ is linearly dependent and, hence, cannot span $\mathbb{R}^N$.

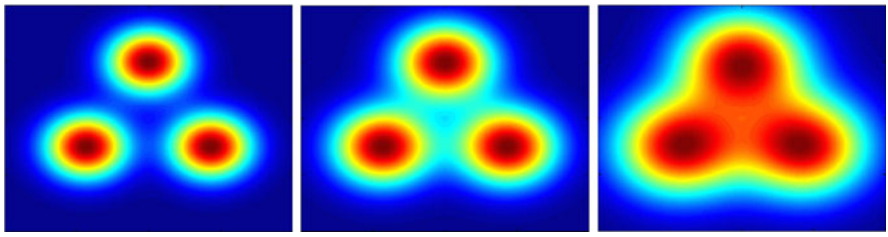
Proposition 12.2 *Let μ and ν be tight probabilistic frames. If ν has zero mean, i.e., $\int_{\mathbb{R}^N} y \, d\nu(y) = 0$, then $\mu * \nu$ is also a tight probabilistic frame.*

Proof Let A_μ and A_ν denote the frame bounds of μ and ν, respectively.

$$\begin{aligned} \int_{\mathbb{R}^N} \big|\langle x, y\rangle\big|^2 \, d\mu * \nu(y) &= \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \big|\langle x, y+z\rangle\big|^2 \, d\mu(y)\, d\nu(z) \\ &= \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \big|\langle x, y\rangle\big|^2 \, d\mu(y)\, d\nu(z) \end{aligned}$$



(a) The orthonormal basis is convolved with a gaussian



(b) The mercedes benz convolved with a gaussian

Fig. 12.1 Heatmaps for the associated probabilistic tight frame, where $\{\varphi_i\}_{i=1}^M \subset \mathbb{R}^2$ is convolved with a Gaussian of increased variance (from *left* to *right*). The origin is at the center, and the axes run from -2 to 2. Each colormap separately scales from zero to the respective density's maximum

$$
\begin{aligned}
&+ \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \big|\langle x, z\rangle\big|^2 \, d\mu(y)\, d\nu(z) \\
&+ 2 \int_{\mathbb{R}^N} \int_{\mathbb{R}^N} \langle x, y\rangle \langle x, z\rangle \, d\mu(y)\, d\nu(z) \\
&= A_\mu \|x\|^2 + A_\nu \|x\|^2 + 2\left\langle \int_{\mathbb{R}^N} \langle x, y\rangle x \, d\mu(y), \int_{\mathbb{R}^N} z \, d\nu(z) \right\rangle \\
&= (A_\mu + A_\nu)\|x\|^2,
\end{aligned}
$$

where the latter equality is due to $\int_{\mathbb{R}^N} z \, d\nu(z) = 0$. □

Example 12.2 Let $\{\varphi_i\}_{i=1}^M \subset \mathbb{R}^N$ be a tight frame, and let ν be a probability measure with $d\nu(x) = g(\|x\|)\, dx$ for some function g. We have already mentioned in Example 12.1 that ν is a tight probabilistic frame, and Proposition 12.2 then implies that $(\frac{1}{M}\sum_{i=1}^M \delta_{-\varphi_i}) * \nu = \frac{1}{M}\sum_{i=1}^M f(x - \varphi_i)\, dx$ is a tight probabilistic frame. See Fig. 12.1 for a visualization.

Proposition 12.3 *Let μ and ν be two probabilistic frames on $\mathbb{R}^{N_1}$ and $\mathbb{R}^{N_2}$ with lower and upper frame bounds A_μ, A_ν and B_μ, B_ν, respectively, such that at least one of them has zero mean. Then the product measure $\gamma = \mu \otimes \nu$ is a probabilistic frame for $\mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$ with lower and upper frame bounds* $\min(A_\mu, A_\nu)$ *and* $\max(B_\mu, B_\nu)$, *respectively.*

If, in addition, μ and ν are tight and $M_2^2(\mu)/N_1 = M_2^2(\nu)/N_2$, then $\gamma = \mu \otimes \nu$ is a tight probabilistic frame.

Proof Let $(z_1, z_2) \in \mathbb{R}^{N_1} \times \mathbb{R}^{N_2}$; then

$$\begin{aligned}
\iint_{\mathbb{R}^{N_1}\times\mathbb{R}^{N_2}} \big|\langle (z_1,z_2),(x,y)\rangle\big|^2 \, d\gamma(x,y) &= \iint_{\mathbb{R}^{N_1}\times\mathbb{R}^{N_2}} \big(\langle z_1,x\rangle + \langle z_2,y\rangle\big)^2 \, d\gamma(x,y) \\
&= \iint_{\mathbb{R}^{N_1}\times\mathbb{R}^{N_2}} \langle z_1,x\rangle^2 \, d\gamma(x,y) \\
&\quad + \iint_{\mathbb{R}^{N_1}\times\mathbb{R}^{N_2}} \langle z_2,y\rangle^2 \, d\gamma(x,y) \\
&\quad + 2\iint_{\mathbb{R}^{N_1}\times\mathbb{R}^{N_2}} \langle z_1,x\rangle\langle z_2,y\rangle \, d\gamma(x,y) \\
&= \int_{\mathbb{R}^{N_1}} \langle z_1,x\rangle^2 \, d\mu(x) + \int_{\mathbb{R}^{N_2}} \langle z_2,y\rangle^2 \, d\nu(y) \\
&\quad + 2\int_{\mathbb{R}^{N_1}}\int_{\mathbb{R}^{N_2}} \langle z_1,x\rangle\langle z_2,y\rangle \, d\mu(x)\, d\nu(y) \\
&= \int_{\mathbb{R}^{N_1}} \langle z_1,x\rangle^2 \, d\mu(x) + \int_{\mathbb{R}^{N_2}} \langle z_2,y\rangle^2 \, d\nu(y)
\end{aligned}$$

where the last equation follows from the fact that one of the two probability measures has zero mean. Consequently,

$$\begin{aligned}
A_\mu\|z_1\|^2 + A_\nu\|z_2\|^2 &\le \iint_{\mathbb{R}^{N_1}\times\mathbb{R}^{N_2}} \big|\langle (z_1,z_2),(x,y)\rangle\big|^2 \, d\gamma(x,y) \\
&\le B_\mu\|z_1\|^2 + B_\nu\|z_2\|^2,
\end{aligned}$$

and the first part of the proposition follows from $\|(z_1,z_2)\|^2 = \|z_1\|^2 + \|z_2\|^2$. The above estimate and Theorem 12.1 imply the second part. □

When $N_1 = N_2 = N$ in Proposition 12.3 and μ and ν are tight probabilistic frames for $\mathbb{R}^N$ such that at least one of them has zero mean, then $\gamma = \mu \otimes \nu$ is a tight probabilistic frame for $\mathbb{R}^N \times \mathbb{R}^N$. It is obvious that the product measure $\gamma = \mu \otimes \nu$ has marginals μ and ν, respectively, and hence is an element in $\Gamma(\mu,\nu)$, where this last set was defined in (12.3). One could ask whether there are any other tight probabilistic frames in $\Gamma(\mu,\nu)$, and if so, how to find them.

The following question is known in frame theory as the Paulsen problem, cf. [7, 9, 10]: Given a frame $\{\varphi_i\}_{i=1}^M \subset \mathbb{R}^N$, how far is the closest tight frame whose elements have equal norm? The distance between two frames $\Phi = \{\varphi_i\}_{i=1}^M$ and $\Psi = \{\psi_i\}_{i=1}^M$ is usually measured by means of the standard ℓ_2-distance $\sum_{i=1}^M \|\varphi_i - \psi_i\|^2$.

The Paulsen problem can be recast in the probabilistic setting we have been considering, and this reformulation seems flexible enough to yield new insights into the

problem. Given any nonzero vectors $\Phi = \{\varphi_i\}_{i=1}^M$, there are two natural embeddings into the space of probability measures, namely

$$\mu_\Phi = \frac{1}{M}\sum_{i=1}^M \delta_{\varphi_i} \quad \text{and} \quad \mu^\Phi := \frac{1}{\sum_{i=1}^M \|\varphi_i\|^2}\sum_{i=1}^M \|\varphi_i\|^2 \delta_{\varphi_i/\|\varphi_i\|}.$$

The 2-Wasserstein distance between μ_Φ and μ_Ψ satisfies

$$M\|\mu_\Phi - \mu_\Psi\|_{W_2}^2 = \inf_{\pi\in\Pi_M}\sum_{i=1}^M \|\varphi_i - \psi_{\pi(i)}\|^2 \le \sum_{i=1}^M \|\varphi_i - \psi_i\|^2, \tag{12.4}$$

where Π_M denotes the set of all permutations of $\{1,\dots,M\}$; cf. [25]. The right-hand side of (12.4) represents the standard distance between frames and is sensitive to the ordering of the frame elements. However, the Wasserstein distance allows us to rearrange elements. More importantly, the ℓ_2-distance requires both frames to have the same cardinalities. On the other hand, the Wasserstein metric enables us to determine how far two frames of different cardinalities are from each other. Therefore, in trying to solve the Paulsen problem, one can seek the closest tight unit norm frame without requiring it to have the same cardinality.

The second embedding μ^Φ can be used to illustrate this point.

Example 12.3 If, for $\varepsilon \ge 0$,

$$\Phi_\varepsilon = \left\{(1,0)^\top, \sqrt{\frac{1}{2}}\big(\sin(\varepsilon),\cos(\varepsilon)\big)^\top, \sqrt{\frac{1}{2}}\big(\sin(-\varepsilon),\cos(-\varepsilon)\big)^\top\right\},$$

then $\mu^{\Phi_\varepsilon} \to \frac{1}{2}(\delta_{e_1} + \delta_{e_2})$ in the 2-Wasserstein metric as $\varepsilon \to 0$, where $\{e_i\}_{i=1}^2$ is the canonical orthonormal basis for $\mathbb{R}^2$. Thus, $\{e_i\}_{i=1}^2$ is close to Φ_ε in the probabilistic setting. Since $\{e_i\}_{i=1}^2$ has only 2 vectors, it is not even under consideration when one looks for any tight frame that is close to Φ_ε in the standard ℓ_2-distance.

We finish this subsection with a list of open problems whose solution can shed new light on frame theory. The first three questions are related to the Paulsen problem, cf. [7, 9, 10], that we have already mentioned above.

Problem 12.1

(a) Given a probabilistic frame $\mu \in \mathscr{P}(S^{N-1})$, how far is the closest probabilistic tight unit norm frame $\nu \in \mathscr{P}(S^{N-1})$ with respect to the 2-Wasserstein metric and how can we find it? Notice that in this case, $\mathscr{P}_2(S^{N-1}) = \mathscr{P}(S^{N-1})$ is a compact set; see, e.g., [29, Theorem 6.4].

(b) Given a unit norm probabilistic frame $\mu \in \mathscr{P}_2$, how far is the closest probabilistic tight unit norm frame $\nu \in \mathscr{P}_2$ with respect to the 2-Wasserstein metric and how can we find it?

(c) Replace the 2-Wasserstein metric in the preceding two problems with different Wasserstein p-metrics $W_p^p(\mu,\nu) = \inf_{\gamma\in\Gamma(\mu,\nu)} \int_{\mathbb{R}^N\times\mathbb{R}^N} \|x-y\|^p \, d\gamma(x,y)$, where $2 \neq p \in (1,\infty)$.

(d) Let μ and ν be two probabilistic tight frames on $\mathbb{R}^N$, such that at least one of them has zero mean. Recall that $\Gamma(\mu,\nu)$ is the set of all probability measures on $\mathbb{R}^N \times \mathbb{R}^N$ whose marginals are μ and ν, respectively. Is the minimizer $\gamma_0 \in \Gamma(\mu,\nu)$ for $W_2^2(\mu,\nu)$ a probabilistic tight frame? Alternatively, are there any other probabilistic tight frames in $\Gamma(\mu,\nu)$ besides the product measure?

12.2.2 The Probabilistic Frame and the Gram Operators

To better understand the notion of probabilistic frames, we consider some related operators that encode all the properties of the measure μ. Let $\mu \in \mathscr{P}$ be a probabilistic frame. The *probabilistic analysis operator* is given by

$$T_\mu : \mathbb{R}^N \to L^2\big(\mathbb{R}^N,\mu\big), \quad x \mapsto \langle x,\cdot\rangle.$$

Its adjoint operator is defined by

$$T_\mu^* : L^2\big(\mathbb{R}^N,\mu\big) \to \mathbb{R}^N, \quad f \mapsto \int_{\mathbb{R}^N} f(x)x \, d\mu(x)$$

and is called the *probabilistic synthesis operator*, where the above integral is vector-valued. The *probabilistic Gram operator*, also called the *probabilistic Gramian* of μ, is $G_\mu = T_\mu T_\mu^*$. The *probabilistic frame operator* of μ is $S_\mu = T_\mu^* T_\mu$, and one easily verifies that

$$S_\mu : \mathbb{R}^N \to \mathbb{R}^N, \quad S_\mu(x) = \int_{\mathbb{R}^N} \langle x,y\rangle y \, d\mu(y).$$

If $\{e_j\}_{j=1}^N$ is the canonical orthonormal basis for $\mathbb{R}^N$, then the vector-valued integral yields

$$\int_{\mathbb{R}^N} y^{(i)} y \, d\mu(y) = \sum_{j=1}^N \int_{\mathbb{R}^N} y^{(i)} y^{(j)} \, d\mu(y) e_j,$$

where $y = (y^{(1)},\dots,y^{(N)})^\top \in \mathbb{R}^N$. If we denote the second moments of μ by $m_{i,j}(\mu)$, i.e.,

$$m_{i,j}(\mu) = \int_{\mathbb{R}^N} x^{(i)} x^{(j)} \, d\mu(x), \quad \text{for } i,j = 1,\dots,N,$$

then we obtain

$$S_\mu e_i = \int_{\mathbb{R}^N} y^{(i)} y \, d\mu(y) = \sum_{j=1}^N \int_{\mathbb{R}^N} y^{(i)} y^{(j)} \, d\mu(y) e_j = \sum_{j=1}^N m_{i,j}(\mu) e_j.$$

Thus, the probabilistic frame operator is the matrix of second moments.

The Gramian of μ is the integral operator defined on $L^2(\mathbb{R}^N, \mu)$ by

$$G_\mu f(x) = T_\mu T_\mu^* f(x) = \int_{\mathbb{R}^N} K(x, y) f(y)\, d\mu(y) = \int_{\mathbb{R}^N} \langle x, y\rangle f(y)\, d\mu(y).$$

It is trivially seen that G_μ is a compact operator on $L^2(\mathbb{R}^N, \mu)$ and in fact it is trace class and Hilbert-Schmidt. Indeed, its kernel is symmetric, continuous, and in $L^2(\mathbb{R}^N \times \mathbb{R}^N, \mu \otimes \mu) \subset L^1(\mathbb{R}^N \times \mathbb{R}^N, \mu \otimes \mu)$. Note that the last inclusion follows from the fact that $\mu \otimes \mu$ is a (finite) probability measure on $\mathbb{R}^N \times \mathbb{R}^N$. Moreover, for any $f \in L^2(\mathbb{R}^N, \mu)$, $G_\mu f$ is a uniformly continuous function on $\mathbb{R}^N$.

Let us collect some properties of S_μ and G_μ.

Proposition 12.4 *If $\mu \in \mathscr{P}$, then the following statements hold*:

(a) *S_μ is well defined (and hence bounded) if and only if*

$$M_2(\mu) < \infty.$$

(b) *μ is a probabilistic frame if and only if S_μ is well defined and positive definite.*
(c) *The nullspace of G_μ consists of all functions in $L^2(\mathbb{R}^N, \mu)$ such that*

$$\int_{\mathbb{R}^N} y f(y)\, d\mu(y) = 0.$$

Moreover, the eigenvalue 0 of G_μ has infinite multiplicity; that is, its eigenspace is infinite dimensional.

For completeness, we give a detailed proof of Proposition 12.4.

Proof Part (a): If S_μ is well defined, then it is bounded as a linear operator on a finite dimensional Hilbert space. If $\|S_\mu\|$ denotes its operator norm and $\{e_i\}_{i=1}^N$ is an orthonormal basis for $\mathbb{R}^N$, then

$$\begin{aligned}\int_{\mathbb{R}^N} \|y\|^2\, d\mu(y) &= \sum_{i=1}^N \int_{\mathbb{R}^N} \langle e_i, y\rangle\langle y, e_i\rangle\, d\mu(y) \\ &= \sum_{i=1}^N \big\langle S_\mu(e_i), e_i\big\rangle \le \sum_{i=1}^N \big\| S_\mu(e_i)\big\| \le N\|S_\mu\|.\end{aligned}$$

On the other hand, if $M_2(\mu) < \infty$, then

$$\int_{\mathbb{R}^N} \big|\langle x, y\rangle\big|^2\, d\mu(y) \le \int_{\mathbb{R}^N} \|x\|^2\|y\|^2\, d\mu(y) = \|x\|^2 M_2^2(\mu),$$

and, therefore, T_μ is well defined and bounded. So is T_μ^*, and hence S_μ is well defined and bounded.

Part (b): If μ is a probabilistic frame, then $M_2(\mu) < \infty$, cf. Theorem 12.1, and hence S_μ is well defined. If $A > 0$ is the lower frame bound of μ, then we obtain

$$\langle x, S_\mu(x)\rangle = \int_{\mathbb{R}^N} \langle x, y\rangle\langle x, y\rangle \, d\mu(y) = \int_{\mathbb{R}^N} \big|\langle x, y\rangle\big|^2 \, d\mu(y) \geq A\|x\|^2,$$
$$\text{for all } x \in \mathbb{R}^N,$$

so that S_μ is positive definite.

Now, let S_μ be well defined and positive definite. According to part (a), $M_2^2(\mu) < \infty$ so that the upper frame bound exists. Since S_μ is positive definite, its eigenvectors $\{v_i\}_{i=1}^N$ are a basis for $\mathbb{R}^N$ and the eigenvalues $\{\lambda_i\}_{i=1}^N$, respectively, are all positive. Each $x \in \mathbb{R}^N$ can be expanded as $x = \sum_{i=1}^N a_i v_i$ such that $\sum_{i=1}^N a_i^2 = \|x\|^2$. If $\lambda > 0$ denotes the smallest eigenvalue, then we obtain

$$\int_{\mathbb{R}^N} \big|\langle x, y\rangle\big|^2 \, d\mu(y) = \langle x, S_\mu(x)\rangle = \sum_{i,j} a_i \langle v_i, \lambda_j a_j v_j\rangle = \sum_{i=1}^N a_i^2 \lambda_i \geq \lambda\|x\|^2,$$

so that λ is the lower frame bound.

For part (c) notice that f is in the nullspace of G_μ if and only if

$$0 = \int_{\mathbb{R}^N} \langle x, y\rangle f(y) \, d\mu(y) = \left\langle x, \int_{\mathbb{R}^N} y f(y) \, d\mu(y)\right\rangle, \quad \text{for each } x \in \mathbb{R}^N.$$

This condition is equivalent to $\int_{\mathbb{R}^N} y f(y) \, d\mu(y) = 0$. The fact that the eigenspace corresponding to the eigenvalue 0 has infinite dimension follows from general principles about compact operators. □

A key property of probabilistic frames is that they give rise to a reconstruction formula similar to the one used in frame theory. Indeed, if $\mu \in \mathscr{P}_2$ is a probabilistic frame, define $\tilde{\mu} = \mu \circ S_\mu$ by

$$\tilde{\mu}(B) = \mu\big(\big(S_\mu^{-1}\big)^{-1} B\big) = \mu(S_\mu B)$$

for each Borel set $B \subset \mathbb{R}^N$. This is equivalent to

$$\int_{\mathbb{R}^N} f\big(S_\mu^{-1}(y)\big) \, d\mu(y) = \int_{\mathbb{R}^N} f(y) \, d\tilde{\mu}(y).$$

We point out that $\tilde{\mu}$ is the *pushforward* of μ through S_μ^{-1}. We refer to [2, Sect. 5.2] for more details on the pushforward of probability measures. Consequently, using the fact that $S_\mu^{-1} S_\mu = S_\mu S_\mu^{-1} = Id$ we have

$$\int_{\mathbb{R}^N} \langle x, y\rangle S_\mu y \, d\tilde{\mu}(y) = \int_{\mathbb{R}^N} \langle x, S_\mu^{-1} y\rangle S_\mu S_\mu^{-1}(y) \, d\mu(y)$$

$$= \int_{\mathbb{R}^N} \langle S_\mu^{-1} x, y \rangle y \, d\mu(y)$$
$$= S_\mu S_\mu^{-1}(x)$$
$$= x$$

for each $x \in \mathbb{R}^N$. Therefore, we have just derived the reconstruction formula:

$$x = \int_{\mathbb{R}^N} \langle x, y \rangle S_\mu y \, d\tilde{\mu}(y) = \int_{\mathbb{R}^N} y \langle S_\mu y, x \rangle \, d\tilde{\mu}(y), \quad \text{for all } x \in \mathbb{R}^N. \tag{12.5}$$

In fact, if μ is a probabilistic frame for $\mathbb{R}^N$, then $\tilde{\mu}$ is a probabilistic frame for $\mathbb{R}^N$. Note that if μ is the counting measure corresponding to a finite unit norm tight frame $\{\varphi_i\}_{i=1}^M$, then $\tilde{\mu}$ is the counting measure associated to the canonical dual frame of $\{\varphi_i\}_{i=1}^M$, and Eq. (12.5) reduces to the known reconstruction formula for finite frames. These observations motivate the following definition:

Definition 12.2 If μ is a probabilistic frame, then $\tilde{\mu} = \mu \circ S_\mu$ is called the *probabilistic canonical dual frame* of μ.

Many properties of finite frames can be carried over. For instance, we can follow the methods in [12] to derive a generalization of the canonical tight frame.

Proposition 12.5 *If μ is a probabilistic frame for $\mathbb{R}^N$, then $\mu \circ S_\mu^{1/2}$ is a tight probabilistic frame for $\mathbb{R}^N$.*

Remark 12.4 The notion of probabilistic frames that we developed thus far in finite dimensional Euclidean spaces can be defined on any infinite dimensional separable real Hilbert space X with norm $\|\cdot\|_X$ and inner product $\langle \cdot, \cdot \rangle_X$. We call a Borel probability measure μ on X a *probabilistic frame for X* if there exist $0 < A \leq B < \infty$ such that

$$A\|x\|^2 \leq \int_X |\langle x, y \rangle|^2 \, d\mu(y) \leq B\|x\|^2, \quad \text{for all } x \in X.$$

If $A = B$, then we call μ a probabilistic tight frame.

12.3 Probabilistic Frame Potential

The frame potential was defined in [6, 16, 31, 36]. In particular, the frame potential of $\Phi = \{\varphi_i\}_{i=1}^M \subset \mathbb{R}^N$ is the function $\mathrm{FP}(\cdot)$ defined on $\mathbb{R}^N \times \mathbb{R}^N \times \cdots \times \mathbb{R}^N$ by

$$\mathrm{FP}(\Phi) := \sum_{i=1}^M \sum_{j=1}^M |\langle \varphi_i, \varphi_j \rangle|^2.$$

Its probabilistic analog is given by the following definition.

Definition 12.3 For $\mu \in \mathscr{P}_2$, the *probabilistic frame potential* is

$$\mathrm{PFP}(\mu) = \iint_{\mathbb{R}^N \times \mathbb{R}^N} \left|\langle x, y\rangle\right|^2 d\mu(x)\, d\mu(y). \tag{12.6}$$

Note that $\mathrm{PFP}(\mu)$ is well defined for each $\mu \in \mathscr{P}_2$ and $\mathrm{PFP}(\mu) \leq M_2^4(\mu)$.

In fact, the probabilistic frame potential is just the Hilbert-Schmidt norm of the operator G_μ, that is,

$$\|G_\mu\|_{\mathrm{HS}}^2 = \iint_{\mathbb{R}^N \times \mathbb{R}^N} \langle x, y\rangle^2\, d\mu(x)\, d\mu(y) = \sum_{\ell=0}^{\infty} \lambda_\ell^2,$$

where $\lambda_k := \lambda_k(\mu)$ is the k-th eigenvalue of G_μ.

If $\Phi = \{\varphi_i\}_{i=1}^M$ $M \geq N$ is a finite unit norm tight frame, and $\mu = \frac{1}{M}\sum_{i=1}^M \delta_{\varphi_i}$ is the corresponding probabilistic tight frame, then

$$\mathrm{PFP}(\mu) = \frac{1}{M^2} \sum_{i,j=1}^{M} \langle \varphi_i, \varphi_j\rangle^2 = \frac{1}{M^2}\frac{M^2}{N} = \frac{1}{N}.$$

According to Theorem 4.2 in [16], we have

$$\mathrm{PFP}(\mu) \geq \frac{1}{N} M_2^4(\mu),$$

and, except for the measure δ_0, equality holds if and only if μ is a probabilistic tight frame.

Theorem 12.3 *If $\mu \in \mathscr{P}_2$ such that $M_2(\mu) = 1$, then*

$$\mathrm{PFP}(\mu) \geq 1/n, \tag{12.7}$$

where n is the number of nonzero eigenvalues of S_μ. Moreover, equality holds if and only if μ is a probabilistic tight frame for E_μ.

Note that we must identify E_μ with the real $\dim(E_\mu)$-dimensional Euclidean space in Theorem 12.3 to speak about probabilistic frames for E_μ. Moreover, Theorem 12.3 yields that if $\mu \in \mathscr{P}_2$ such that $M_2(\mu) = 1$, then $\mathrm{PFP}(\mu) \geq 1/N$, and equality holds if and only if μ is a probabilistic tight frame for $\mathbb{R}^N$.

Proof Recall that $\sigma(G_\mu) = \sigma(S_\mu) \cup \{0\}$, where $\sigma(T)$ denotes the spectrum of the operator T. Moreover, because G_μ is compact, its spectrum consists only of eigenvalues. Moreover, the condition on the support of μ implies that the eigenvalues $\{\lambda_k\}_{k=1}^N$ of S_μ are all positive. Since

$$\sigma(G_\mu) = \sigma(S_\mu) \cup \{0\} = \{\lambda_k\}_{k=1}^N \cup \{0\},$$

the proposition reduces to minimizing $\sum_{k=1}^N \lambda_k^2$ under the constraint $\sum_{k=1}^N \lambda_k = 1$, which concludes the proof. □

12.4 Relations to Other Fields

Probabilistic frames, isotropic measures, and the geometry of convex bodies A finite nonnegative Borel measure μ on S^{N-1} is called *isotropic* in [19, 26] if

$$\int_{S^{N-1}} \left|\langle x, y\rangle\right|^2 d\mu(y) = \frac{\mu(S^{N-1})}{N} \quad \forall x \in S^{N-1}.$$

Thus, every tight probabilistic frame $\mu \in \mathscr{P}(S^{N-1})$ is an isotropic measure. The term isotropic is also used for special subsets in $\mathbb{R}^N$. Recall that a subset $K \subset \mathbb{R}^N$ is called a convex body if K is compact, convex, and has nonempty interior. Denote by $\mathrm{vol}_N(B)$ the N-dimensional volume of $B \subset \mathbb{R}^N$. According to [28, Sect. 1.6] and [19], a convex body K with centroid at the origin and unit volume, i.e., $\int_K x\,dx = 0$ and $\mathrm{vol}_N(K) = \int_K dx = 1$, is said to be in *isotropic position* if there exists a constant L_K such that

$$\int_K \left|\langle x, y\rangle\right|^2 d\sigma_K(y) = L_K \quad \forall x \in S^{N-1}, \tag{12.8}$$

where σ_K denotes the uniform measure on K.

Thus, K is in isotropic position if and only if the uniform probability measure on K (σ_K) is a tight probabilistic frame. The constant L_K must then satisfy $L_K = \frac{1}{N}\int_K \|x\|^2 d\sigma_K(x)$.

In fact, the two concepts, isotropic measures and being in isotropic position, can be combined within probabilistic frames as follows: Given any tight probabilistic frame $\mu \in \mathscr{P}$ on $\mathbb{R}^N$, let K_μ denote the convex hull of $\mathrm{supp}(\mu)$. Then for each $x \in \mathbb{R}^N$ we have

$$\int_{\mathbb{R}^N} \left|\langle x, y\rangle\right|^2 d\mu(y) = \int_{\mathrm{supp}(\mu)} \left|\langle x, y\rangle\right|^2 d\mu(y) = \int_{K_\mu} \left|\langle x, y\rangle\right|^2 d\mu(y).$$

Although K_μ might not be a convex body, we see that the convex hull of the support of every tight probabilistic frame is in "isotropic position" with respect to μ.

In the following, let $\mu \in \mathscr{P}(S^{N-1})$ be a probabilistic unit norm tight frame with zero mean. K_μ is a convex body and

$$\mathrm{vol}_N(K_\mu) \geq \frac{(N+1)^{(N+1)/2}}{N!} N^{-N/2},$$

where equality holds if and only if K_μ is a regular simplex; cf. [3, 26]. Note that the extremal points of the regular simplex form an equiangular tight frame $\{\varphi_i\}_{i=1}^{N+1}$, i.e., a tight frame whose pairwise inner products $|\langle \varphi_i, \varphi_j\rangle|$ do not depend on $i \neq j$. Moreover, the polar body $P_\mu := \{x \in \mathbb{R}^N : \langle x, y\rangle \leq 1, \text{ for all } y \in \mathrm{supp}(\mu)\}$ satisfies

$$\mathrm{vol}_N(P_\mu) \leq \frac{(N+1)^{(N+1)/2}}{N!} N^{N/2},$$

and, again, equality holds if and only if K_μ is a regular simplex; cf. [3, 26].

Probabilistic tight frames are also related to inscribed ellipsoids of convex bodies. Note that each convex body contains a unique ellipsoid of maximal volume, called John's ellipsoid; cf. [20]. Therefore, there is an affine transformation Z such that the ellipsoid of maximal volume of $Z(K)$ is the unit ball. A characterization of such transformed convex bodies was derived in [20]; see also [3].

Theorem 12.4 *The unit ball $B \subset \mathbb{R}^N$ is the ellipsoid of maximal volume in the convex body K if and only if $B \subset K$ and, for some $M \geq N$, there are $\{\varphi_i\}_{i=1}^M \subset S^{N-1} \cap \partial K$ and positive numbers $\{c_i\}_{i=1}^M$ such that*

(a) $\sum_{i=1}^M c_i \varphi_i = 0$ *and*
(b) $\sum_{i=1}^M c_i \varphi_i \varphi_i^\top = I_N$.

Note that the conditions (a) and (b) in Theorem 12.4 are equivalent to saying that $\frac{1}{N}\sum_{i=1}^M c_i \delta_{\varphi_i} \in \mathscr{P}(S^{N-1})$ is a probabilistic unit norm tight frame with zero mean.

Last but not least, we comment on a deep open problem in convex analysis. Bourgain raised in [8] the following question: *Is there a universal constant $c > 0$ such that for any dimension N and any convex body K in $\mathbb{R}^N$ with $\mathrm{vol}_N(K) = 1$, there exists a hyperplane $H \subset \mathbb{R}^N$ for which $\mathrm{vol}_{N-1}(K \cap H) > c$?* The positive answer to this question has become known as the hyperplane conjecture. By applying results in [28], we can rephrase this conjecture by means of probabilistic tight frames: *There is a universal constant C such that for any convex body K, on which the uniform probability measure σ_K forms a probabilistic tight frame, the probabilistic tight frame bound is less than C.* Due to Theorem 12.1, the boundedness condition is equivalent to $M_2^2(\sigma_K) \leq CN$. The hyperplane conjecture is still open, but there are large classes of convex bodies, for instance, Gaussian random polytopes [24], for which an affirmative answer has been established.

Probabilistic frames and positive operator valued measures Let Ω be a locally compact Hausdorff space, $\mathscr{B}(\Omega)$ be the Borel-sigma algebra on Ω, and H be a real separable Hilbert space with norm $\|\cdot\|$ and inner product $\langle\cdot,\cdot\rangle$. We denote by $\mathscr{L}(H)$ the space of bounded linear operators on H.

Definition 12.4 A *positive operator valued measure (POVM)* on Ω with values in $\mathscr{L}(H)$ is a map $F : \mathscr{B}(\Omega) \to \mathscr{L}(H)$ such that:

(i) $F(A)$ is positive semidefinite for each $A \in \mathscr{B}(\Omega)$;
(ii) $F(\Omega)$ is the identity map on H;
(iii) If $\{A_i\}_{i\in I}^\infty$ is a countable family of pairwise disjoint Borel sets in $\mathscr{B}(\Omega)$, then

$$F\left(\bigcup_{i\in I} A_i\right) = \sum_{i\in I} F(A_i),$$

where the series on the right-hand side converges in the weak topology of $\mathscr{L}(H)$, i.e., for all vectors $x, y \in H$, the series $\sum_{i\in I}\langle F(A_i)x, y\rangle$converges.

We refer to [1, 13, 14] for more details on POVMs.

In fact, every probabilistic tight frame on $\mathbb{R}^N$ gives rise to a POVM on $\mathbb{R}^N$ with values in the set of real $N \times N$ matrices:

Proposition 12.6 *Assume that $\mu \in \mathscr{P}_2(\mathbb{R}^N)$ is a probabilistic tight frame. Define the operator F from $\mathscr{B}$ to the set of real $N \times N$ matrices by*

$$F(A) := \frac{N}{M_2^2(\mu)} \left(\int_A y_i y_j \, d\mu(y) \right)_{i,j}. \tag{12.9}$$

Then F is a POVM.

Proof Note that for each Borel measurable set A, the matrix $F(A)$ is positive semidefinite, and we also have $F(\mathbb{R}^N) = Id_N$. Finally, for a countable family of pairwise disjoint Borel measurable sets $\{A_i\}_{i \in I}$, we clearly have, for each $x \in \mathbb{R}^N$,

$$F\left(\bigcup_{i \in I} A_i \right) x = \sum_{k \in I} F(A_k) x.$$

Thus, any probabilistic tight frame in $\mathbb{R}^N$ gives rise to a POVM. □

We have not been able to prove or disprove whether the converse of this proposition holds.

Problem 12.2 Given a POVM $F : \mathscr{B}(\mathbb{R}^N) \to \mathscr{L}(\mathbb{R}^N)$, is there a tight probabilistic frame μ such that F and μ are related through (12.9)?

Probabilistic frames and t-designs Let σ denote the uniform probability measure on S^{N-1}. A *cubature formula of strength t* is a finite collection of points $\{\varphi_i\}_{i=1}^M \subset S^{N-1}$ with weights $\{\omega_i\}_{i=1}^M$ such that

$$\sum_{i=1}^{M} \omega_i h(\varphi_i) = \int_{S^{N-1}} h(x) \, d\sigma(x),$$

for all homogeneous polynomials h of total degree less than or equal to t in N variables [30]. Cubature formulas are used in numerical integration, and the weights are usually required to be positive. A *spherical t-design* is a cubature formula of strength t whose weights are all equal to $1/M$. The parameter t quantifies how well the spherical design samples the sphere. Spherical designs were introduced in [15] as analogs on the sphere of the classical combinatorial designs. One commonly aims to find the strongest spherical design for fixed M, or seeks to minimize M for a fixed strength t. Exact answers are essentially known only for small M and t. Instead, lower and upper bounds, respectively, and some asymptotic statements have been derived; cf. [4, 5, 15, 32]. In general, it is extremely difficult to explicitly construct spherical t-designs for large t.

This notion of t-design can be extended to the probabilistic setting considered in this chapter. In particular, a probability measure $\mu \in \mathscr{P}(S^{N-1})$ is called a *probabilistic spherical t-design* in [16] if

$$\int_{S^{N-1}} h(x)\, d\mu(x) = \int_{S^{N-1}} h(x)\, d\sigma(x), \tag{12.10}$$

for all homogeneous polynomials h with total degree less than or equal to t. Since the weights are hidden in the measure, it would also make sense to call μ a probabilistic cubature formula. The following result has been established in [16].

Theorem 12.5 *If $\mu \in \mathscr{P}(S^{N-1})$, then the following are equivalent*:

(i) *μ is a probabilistic spherical 2-design.*
(ii) *μ minimizes*

$$\frac{\int_{S^{N-1}} \int_{S^{N-1}} |\langle x, y\rangle|^2 \, d\mu(x)\, d\mu(y)}{\int_{S^{N-1}} \int_{S^{N-1}} \|x-y\|^2 \, d\mu(x)\, d\mu(y)} \tag{12.11}$$

among all probability measures $\mathscr{P}(S^{N-1})$.
(iii) *μ is a tight probabilistic unit norm frame with zero mean.*

In particular, if μ is a tight probabilistic unit norm frame, then $\nu(A) := \frac{1}{2}(\mu(A) + \mu(-A))$, for $A \in \mathscr{B}$, defines a probabilistic spherical 2-design.

Note that conditions (a) and (b) of Theorem 12.4 can be rephrased as saying that $\frac{1}{N}\sum_{i=1}^{M} c_i \delta_{\varphi_i} \in \mathscr{P}(S^{N-1})$ is a probabilistic spherical 2-design.

Remark 12.5 By using results in [18], the equivalence between (i) and (ii) in Theorem 12.5 can be generalized to spherical t-designs if t is an even integer. In this case, μ is a probabilistic spherical t-design if and only if μ minimizes

$$\frac{\int_{S^{N-1}} \int_{S^{N-1}} |\langle x, y\rangle|^t \, d\mu(x)\, d\mu(y)}{\int_{S^{N-1}} \int_{S^{N-1}} \|x-y\|^2 \, d\mu(x)\, d\mu(y)}.$$

Probabilistic frames and directional statistics Common tests in directional statistics focus on whether or not a sample on the unit sphere S^{N-1} is uniformly distributed. The *Bingham test* rejects the hypothesis of directional uniformity of a sample $\{\varphi_i\}_{i=1}^{M} \subset S^{N-1}$ if the *scatter matrix*

$$\frac{1}{M}\sum_{i=1}^{M} \varphi_i \varphi_i^{\top}$$

is far from $\frac{1}{N} I_N$; cf. [27]. Note that this scatter matrix is the scaled frame operator of $\{\varphi_i\}_{i=1}^{M}$ and, hence, one measures the sample's deviation from being a tight frame. Probability measures μ that satisfy $S_\mu = \frac{1}{N} I_N$ are called *Bingham alternatives* in [17], and the probabilistic unit norm tight frames on the sphere S^{N-1} are the Bingham alternatives.

Tight frames also occur in relation to M-estimators as discussed in [21, 33, 34]: The family of angular central Gaussian distributions are given by densities f_Γ with respect to the uniform surface measure on the sphere S^{N-1}, where

$$f_\Gamma(x) = \frac{\det(\Gamma)^{-1/2}}{a_N}\left(x^\top \Gamma^{-1} x\right)^{-N/2}, \quad \text{for } x \in S^{N-1}.$$

Note that Γ is only determined up to a scaling factor. According to [34], the maximum likelihood estimate of Γ based on a random sample $\{\varphi_i\}_{i=1}^M \subset S^{N-1}$ is the solution $\hat{\Gamma}$ to

$$\hat{\Gamma} = \frac{M}{N}\sum_{i=1}^{M} \frac{\varphi_i \varphi_i^\top}{\varphi_i^\top \hat{\Gamma}^{-1} \varphi_i},$$

which can be found, under mild assumptions, through the iterative scheme

$$\Gamma_{k+1} = \frac{N}{\sum_{i=1}^M \frac{1}{\varphi_i^\top \Gamma_k^{-1} \varphi_i}} \sum_{i=1}^{M} \frac{\varphi_i \varphi_i^\top}{\varphi_i^\top \Gamma_k^{-1} \varphi_i},$$

where $\Gamma_0 = I_N$, and then $\Gamma_k \to \hat{\Gamma}$ as $k \to \infty$. It is not hard to see that $\{\psi_i\}_{i=1}^M := \{\frac{\hat{\Gamma}^{-1/2}\varphi_i}{\|\hat{\Gamma}^{-1/2}\varphi_i\|}\}_{i=1}^M \subset S^{N-1}$ forms a tight frame. If $\hat{\Gamma}$ is close to the identity matrix, then $\{\psi_i\}_{i=1}^M$ is close to $\{\varphi_i\}_{i=1}^M$ and it is likely that f_Γ represents a probability measure that is close to being tight, in fact, close to the uniform surface measure.

Probabilistic frames and random matrices Random matrices are used in multivariate statistics, physics, compressed sensing, and many other fields. Here, we shall point out some results on random matrices as related to probabilistic tight frames.

For a point cloud $\{\varphi_i\}_{i=1}^M$, the frame operator is a scaled version of the sample covariance matrix up to subtracting the mean and can be related to the population covariance when chosen at random. To properly formulate a result in [16], let us recall some notation. For $\mu \in \mathscr{P}_2$, we define $E(Z) := \int_{\mathbb{R}^N} Z(x)\, d\mu(x)$, where $Z : \mathbb{R}^N \to \mathbb{R}^{p\times q}$ is a random matrix/vector that is distributed according to μ. The following was proven in [16] where $\|\cdot\|_{\mathscr{F}}$ denotes the Frobenius norm on matrices.

Theorem 12.6 *Let $\{X_k\}_{k=1}^M$ be a collection of random vectors, independently distributed according to probabilistic tight frames $\{\mu_k\}_{k=1}^M \subset \mathscr{P}_2$, respectively, whose 4-th moments are finite, i.e., $M_4^4(\mu_k) := \int_{\mathbb{R}^N} \|y\|^4\, d\mu_k(y) < \infty$. If F denotes the random matrix associated to the analysis operator of $\{X_k\}_{k=1}^M$, then we have*

$$E\left(\left\|\frac{1}{M}F^*F - \frac{L_1}{N} I_N\right\|_{\mathscr{F}}^2\right) = \frac{1}{M}\left(L_4 - \frac{L_2}{N}\right), \tag{12.12}$$

where $L_1 := \frac{1}{M}\sum_{k=1}^M M_2(\mu_k)$, $L_2 := \frac{1}{M}\sum_{k=1}^M M_2^2(\mu_k)$, and $L_4 = \frac{1}{M}\sum_{k=1}^M M_4^4(\mu_k)$.

Under the notation of Theorem 12.6, the special case of probabilistic unit norm tight frames was also addressed in [16].

Corollary 12.1 *Let $\{X_k\}_{k=1}^M$ be a collection of random vectors, independently distributed according to probabilistic unit norm tight frames $\{\mu_k\}_{k=1}^M$, respectively, such that $M_4(\mu_k) < \infty$. If F denotes the random matrix associated to the analysis operator of $\{X_k\}_{k=1}^M$, then*

$$E\left(\left\|\frac{1}{M}F^*F - \frac{1}{N}I_N\right\|_{\mathscr{F}}^2\right) = \frac{1}{M}\left(L_4 - \frac{1}{N}\right), \tag{12.13}$$

where $L_4 = \frac{1}{M}\sum_{k=1}^M M_4^4(\mu_k)$.

In compressed sensing, random matrices are used to design measurements and are commonly based on Bernoulli, Gaussian, and sub-Gaussian distributions. Since each row of such a random matrix can be considered as a random vector, it follows that these compressed sensing matrices are induced by probabilistic tight frames, so that we can also apply Theorem 12.1.

Example 12.4 Let $\{X_k\}_{k=1}^M$ be a collection of N-dimensional random vectors such that each vector's entries are independently identically distributed (i.i.d) according to a probability measure with zero mean and finite 4-th moments. This implies that each X_k is distributed with respect to a probabilistic tight frame whose 4-th moments exist. Thus, the assumptions in Theorem 12.6 are satisfied, and we can compute (12.12) for some specific distributions that are related to compressed sensing:

- If the entries of X_k, $k = 1, \ldots, M$, are i.i.d. according to a Bernoulli distribution that takes the values $\pm\frac{1}{\sqrt{N}}$ with probability $\frac{1}{2}$, then X_k is distributed according to a normalized counting measure supported on the vertices of the N-dimensional hypercube. Thus, X_k is distributed according to a probabilistic unit norm tight frame for $\mathbb{R}^N$.
- If the entries of X_k, $k = 1, \ldots, M$, are i.i.d. according to a Gaussian distribution with zero mean and variance $\frac{1}{\sqrt{N}}$, then X_k is distributed according to a multivariate Gaussian probability measure μ whose covariance matrix is $\frac{1}{N}I_N$, and μ forms a probabilistic tight frame for $\mathbb{R}^N$. Since the moments of a multivariate Gaussian random vector are well known, we can explicitly compute $L_4 = 1 + \frac{2}{N}$, $L_1 = 1$, and $L_2 = 1$ in Theorem 12.6. Thus, the right-hand side of (12.12) equals $\frac{1}{M}(1 + \frac{1}{N})$.

Acknowledgements M. Ehler was supported by the NIH/DFG Research Career Transition Awards Program (EH 405/1-1/575910). K.A. Okoudjou was supported by ONR grants N000140910324 and N000140910144, by a RASA from the Graduate School of UMCP, and by the Alexander von Humboldt Foundation. He would also like to express his gratitude to the Institute for Mathematics at the University of Osnabrück.

References

1. Albini, P., De Vito, E., Toigo, A.: Quantum homodyne tomography as an informationally complete positive-operator-valued measure. J. Phys. A **42**(29), 12 (2009)
2. Ambrosio, L., Gigli, N., Savaré, G.: Gradients Flows in Metric Spaces and in the Space of Probability Measures. Lectures in Mathematics ETH Zürich. Birkhäuser, Basel (2005)
3. Ball, K.: Ellipsoids of maximal volume in convex bodies. Geom. Dedic. **41**(2), 241–250 (1992)
4. Bannai, E., Damerell, R.: Tight spherical designs I. J. Math. Soc. Jpn. **31**, 199–207 (1979)
5. Bannai, E., Damerell, R.: Tight spherical designs II. J. Lond. Math. Soc. **21**, 13–30 (1980)
6. Benedetto, J.J., Fickus, M.: Finite normalized tight frames. Adv. Comput. Math. **18**(2–4), 357–385 (2003)
7. Bodmann, B.G., Casazza, P.G.: The road to equal-norm Parseval frames. J. Funct. Anal. **258**(2–4), 397–420 (2010)
8. Bourgain, J.: On high-dimensional maximal functions associated to convex bodies. Am. J. Math. **108**(6), 1467–1476 (1986)
9. Cahill, J., Casazza, P.G.: The Paulsen problem in operator theory. arXiv:1102.2344v2 (2011)
10. Casazza, P.G., Fickus, M., Mixon, D.G.: Auto-tuning unit norm frames. Appl. Comput. Harmon. Anal. **32**(1), 1–15 (2012)
11. Christensen, O.: An Introduction to Frames and Riesz Bases. Birkhäuser, Boston (2003)
12. Christensen, O., Stoeva, D.T.: p-frames in separable Banach spaces. Adv. Comput. Math. **18**, 117–126 (2003)
13. Davies, E.B.: Quantum Theory of Open Systems. Academic Press, London–New York (1976)
14. Davies, E.B., Lewis, J.T.: An operational approach to quantum probability. Commun. Math. Phys. **17**, 239–260 (1970)
15. Delsarte, P., Goethals, J.M., Seidel, J.J.: Spherical codes and designs. Geom. Dedic. **6**, 363–388 (1977)
16. Ehler, M.: Random tight frames. J. Fourier Anal. Appl. **18**(1), 1–20 (2012)
17. Ehler, M., Galanis, J.: Frame theory in directional statistics. Stat. Probab. Lett. **81**(8), 1046–1051 (2011)
18. Ehler, M., Okoudjou, K.A.: Minimization of the probabilistic p-frame potential. J. Stat. Plan. Inference **142**(3), 645–659 (2012)
19. Giannopoulos, A.A., Milman, V.D.: Extremal problems and isotropic positions of convex bodies. Isr. J. Math. **117**, 29–60 (2000)
20. John, F.: Extremum problems with inequalities as subsidiary conditions. In: Courant Anniversary Volume, pp. 187–204. Interscience, New York (1948)
21. Kent, J.T., Tyler, D.E.: Maximum likelihood estimation for the wrapped Cauchy distribution. J. Appl. Stat. **15**(2), 247–254 (1988)
22. Kovačević, J., Chebira, A.: Life beyond bases: the advent of frames (Part I). IEEE Signal Process. Mag. **24**(4), 86–104 (2007)
23. Kovačević, J., Chebira, A.: Life beyond bases: the advent of frames (Part II). IEEE Signal Process. Mag. **24**(5), 115–125 (2007)
24. Klartag, B., Kozma, G.: On the hyperplane conjecture for random convex sets. Isr. J. Math. **170**, 253–268 (2009)
25. Levina, E., Bickel, P.: The Earth Mover's distance is the Mallows distance: some insights from statistics. In: Eighth IEEE International Conference on Computer Vision, vol. 2, pp. 251–256 (2001)
26. Lutwak, E., Yang, D., Zhang, G.: Volume inequalities for isotropic measures. Am. J. Math. **129**(6), 1711–1723 (2007)
27. Mardia, K.V., Peter, E.J.: Directional Statistics. Wiley Series in Probability and Statistics. Wiley, New York (2008)
28. Milman, V., Pajor, A.: Isotropic position and inertia ellipsoids and zonoids of the unit ball of normed n-dimensional space. In: Geometric Aspects of Functional Analysis. Lecture Notes in Math., pp. 64–104. Springer, Berlin (1987–1988)

29. Parthasarathy, K.R.: Probability Measures on Metric Spaces. Probability and Mathematical Statistics, vol. 3. Academic Press, New York–London (1967)
30. Radon, J.: Zur mechanischen Kubatur. Monatshefte Math. **52**, 286–300 (1948)
31. Renes, J.M., et al.: Symmetric informationally complete quantum measurements. J. Math. Phys. **45**, 2171–2180 (2004)
32. Seymour, P., Zaslavsky, T.: Averaging sets: a generalization of mean values and spherical designs. Adv. Math. **52**, 213–240 (1984)
33. Tyler, D.E.: A distribution-free M-estimate of multivariate scatter. Ann. Stat. **15**(1), 234–251 (1987)
34. Tyler, D.E.: Statistical analysis for the angular central Gaussian distribution. Biometrika **74**(3), 579–590 (1987)
35. Villani, C.: Optimal Transport: Old and New. Grundlehren der Mathematischen Wissenschaften, vol. 338. Springer, Berlin (2009)
36. Waldron, S.: Generalised Welch bound equality sequences are tight frames. IEEE Trans. Inf. Theory **49**, 2307–2309 (2003)

Chapter 13
Fusion Frames

Peter G. Casazza and Gitta Kutyniok

Abstract Novel technological advances have significantly increased the demand to model applications requiring distributed processing. Frames are, however, too restrictive for such applications, wherefore it was necessary to go beyond classical frame theory. Fusion frames, which can be regarded as frames of subspaces, satisfy exactly those needs. They analyze signals by projecting them onto multidimensional subspaces, in contrast to frames which consider only one-dimensional projections. This chapter serves as an introduction to and a survey about this exciting area of research as well as a reference for the state of the art of this research field.

Keywords Compressed sensing · Distributed processing · Fusion coherence · Fusion frame · Fusion frame potential · Isoclinic subspaces · Mutually unbiased bases · Sparse fusion frames · Spectral Tetris · Nonorthogonal fusion frames

13.1 Introduction

In the twenty-first century, scientists face massive amounts of data, which can typically no longer be handled with a single processing system. A seemingly unrelated problem arises in sensor networks when communication between any pair of sensors is not possible due to, for instance, low communication bandwidth. Yet another question is the design of erasure-resilient packet-based encoding when data is broken into packets for separate transmission.

All these problems can be regarded as belonging to the field of distributed processing. However, they have an even more special structure in common, since each can be regarded as a special case of the following mathematical framework: Given data and a collection of subspaces, project the data onto the subspaces, then process the data within each subspace, and finally "fuse" the locally computed objects.

P.G. Casazza
Mathematics Department, University of Missouri, Columbia, MO 65211, USA
e-mail: casazzap@missouri.edu

G. Kutyniok (✉)
Institut für Mathematik, Technische Universität Berlin, 10623 Berlin, Germany
e-mail: kutyniok@math.tu-berlin.de

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3_13,

The decomposition of the given data into the subspaces coincides with—relating to the initial three problems—the splitting into different processing systems, the local measurements of groups of close sensors, and the generation of packets. The distributed fusion models the reconstruction procedure, also enabling, for instance, an error analysis of resilience against erasures. This is however only possible if the data is decomposed in a *redundant* way, which forces the subspaces to be redundant.

Fusion frames provide a suitable mathematical framework to design and analyze such applications under distributed processing requirements. Interestingly, fusion frames are also a versatile tool for more theoretically oriented problems in mathematics, and we will see various examples of this throughout the chapter.

13.1.1 The Fusion Frame Framework

Let us now give a half-formal introduction to fusion frames, utilizing another motivation as a guideline. One goal in frame theory is to construct large frames by fusing "smaller" frames, and, in fact, this was the original reasoning for introducing fusion frames by the two authors in [21]. We will come back to the three signal processing applications in Sect. 13.1.3 and show in more detail how they fit into this framework.

Locality of frames can be modeled as frame sequences, i.e., frames for their closed linear span. Now assume we have a collection of frame sequences $(\varphi_{ij})_{j=1}^{J_i}$ in $\mathcal{H}^N$ with $i = 1, \dots, M$, and set $\mathcal{W}_i := \operatorname{span}\{\varphi_{ij} : j = 1, \dots, J_i\}$ for each i. The two key questions are: Does the collection $(\varphi_{ij})_{i=1,j=1}^{M,J_i}$ form a frame for $\mathcal{H}^N$? If yes, which frame properties does it have? The first question is easy to answer, since what is required is the spanning property of the family $(\mathcal{W}_i)_{i=1}^{M}$. The second question requires more thought. But it is intuitively clear that—besides the knowledge of the frame bounds of the frame sequences—it will depend solely on the structural properties of the family of subspaces $(\mathcal{W}_i)_{i=1}^{M}$. In fact, it can be proven that the crucial property is the constants associated with the ℓ_2-stability of the mapping

$$\mathcal{H}^N \ni x \mapsto \big(P_i(x)\big)_{i=1}^{M} \in \mathbb{R}^{NM}, \tag{13.1}$$

where P_i denotes the orthogonal projection onto the subspace $\mathcal{W}_i$. A family of subspaces $(\mathcal{W}_i)_{i=1}^{M}$ satisfying such a stability condition is then called a *fusion frame*.

We would like to emphasize that (13.1) leads to the basic fusion frame definition. It can, for instance, be modified by considering weighted projections to allow flexibility in the significance of each subspace, and hence of each locally constructed frame $(\varphi_{ij})_{j=1}^{J_i}$.

We also stress that in [21] the introduced notion was coined "frames of subspaces" for reasons which become clear in the sequel. Later, to avoid confusion with the term "frames *for* subspaces" and to emphasize the local fusing of information, it was baptized "fusion frames" in [22].

13.1.2 Fusion Frames versus Frames

The main distinction between frames and fusion frames lies in the fact that a frame $(\varphi_i)_{i=1}^M$ for $\mathcal{H}^N$ provides the following measurements of a signal $x \in \mathcal{H}^N$:

$$x \mapsto \big(\langle x, \varphi_i \rangle\big)_{i=1}^M \in \mathbb{R}^M.$$

A fusion frame $(\mathcal{W}_i)_{i=1}^M$ for $\mathcal{H}^N$ on the other hand analyzes the signal x by

$$x \mapsto \big(P_i(x)\big)_{i=1}^M \in \mathbb{R}^{MN}.$$

Thus the *scalar* measurements of a frame are substituted by *vector* measurements, and consequently, the representation space of a frame is $\mathbb{R}^M$, whereas that of a fusion frame is $\mathbb{R}^{MN}$. This latter space can sometimes be reduced, and we refer to the next section for details.

A further natural question is whether the theory of fusion frames includes the theory of frames, which is indeed the case. In fact—and the next section will provide more detailed information—a frame can be regarded as a collection of the one-dimensional subspaces its frame vectors generate. Taking the norms of the frame vectors as the aforementioned weights, it can be shown that this is a fusion frame with similar properties. Conversely, taking a fusion frame, one can fix an orthonormal basis in each subspace and then consider the union of these bases. This will form a frame, which can be regarded as being endowed with a particular substructure.

Even at this point these two viewpoints indicate that fusion frame theory is much more difficult than frame theory. In fact, most results in this chapter will be solely stated for the case of the weights being equal to 1, and even in this situation many questions which are answered for frames remain open in the general situation of fusion frames.

13.1.3 Applications of Fusion Frames

The generality of the framework of fusion frames allows their application to various problems both practical as well as theoretical in nature—which then certainly require additional adaptations in the specific setting considered. We first highlight the three signal processing applications mentioned at the beginning of the chapter.

- *Distributed Sensing.* Given a collection of small and inexpensive sensors scattered over a large area, the measurement each sensor produces of an incoming signal $x \in \mathcal{H}^N$ can be modeled as $\langle x, \varphi_i \rangle$, $\varphi_i \in \mathcal{H}^N$ being the specific characteristics of the sensor. Since due to, for instance, limited bandwidth and transmission power, sensors can only communicate locally, the recovery of the signal x can first only be performed among groups of sensors. Let $(\varphi_{ij})_{j=1}^{J_i}$ in $\mathcal{H}^N$ with $i = 1, \dots, M$ be such a grouping. Then, setting $\mathcal{W}_i := \operatorname{span}\{\varphi_{ij} : j = 1, \dots, J_i\}$ for each i, local

frame reconstruction leads to the collection of vectors $(P_i(x))_{i=1}^M$. This data is then passed on by special transmitters to a central processing station for joint processing. At this point, fusion frame theory kicks in and provides a means for performing and analyzing the reconstruction of the signal x. The modeling of sensor networks through fusion frames was considered in the series of papers [23, 38]. A similar local-global signal processing principle is applicable to modeling of the human visual cortex as discussed in [43].

- *Parallel Processing.* If a frame is too large for efficient processing—from a computational complexity or a numerical stability standpoint—one approach is to divide it into multiple small subsystems for simple and ideally parallel processing. Fusion frames allow a stable splitting into smaller frames and afterwards a stable recombining of the local outputs. Splitting of a large system into smaller subsystems for parallel processing was first considered in [3, 42].
- *Packet Encoding*. Transmission of data over a communication network, for instance the Internet, is often achieved by first encoding it into a number of packets. By introducing redundancy into the encoding scheme, the communication scheme becomes resilient against corruption or even complete loss of transmitted packets. Fusion frames provide a means to achieve and analyze redundant subspace representations, where each packet carries one of the fusion frame projections. The use of fusion frames for packet encoding is considered in [4].

Fusion frames also arise in more theoretical problems, as the next two examples show.

- *Kadison-Singer Problem.* The 1959 Kadison-Singer problem [25] is one of the most famous unsolved problems in analysis today. One of the many equivalent formulations is the following question: Can a bounded frame be partitioned such that the spans of the partitions as a fusion frame lead to a "good" lower fusion frame bound? Therefore, advances in the design of fusion frames will have direct impact in providing new angles for a renewed attack on the Kadison-Singer problem.
- *Optimal Packings.* Fusion frame theory also has close connections with Grassmannian packings. It was shown in [38] that the special class of Parseval fusion frames consisting of equidistance and equidimensional subspaces are in fact optimal Grassmannian packings. Thus, novel methods for constructing such fusion frames simultaneously provide ways to construct optimal packings.

13.1.4 Related Approaches

Several approaches related to fusion frames have appeared in the literature. The concept of a frame-like collection of subspaces was first exploited in relation to domain decomposition techniques in papers by Bjørstad and Mandel [3] and Oswald [42]. In 2003, Fornasier introduced in [33] what he called quasi-orthogonal decompositions. The framework of fusion frames was in fact developed at the same time by

the two authors in [21] and contains those decompositions as a special case. Also note that Sun introduced G-frames in the series of papers [44, 45], which extend the definition of fusion frames by generalizing the utilized orthogonal projections to arbitrary operators. However, the generality of this notion is not suitable for modeling distributed processing.

13.1.5 Outline

In Sect. 13.2, we introduce the basic notions and definitions of fusion frame theory, discuss the relation to frame theory, and present a reconstruction formula. Sect. 13.3 is concerned with the introduction and application of the fusion frame potential as a highly useful method for analyzing fusion frames. The construction of fusion frames is then the focus of Sect. 13.4. In this section, we present the Spectral Tetris algorithm as a versatile means to construct general fusion frames followed by a discussion on the construction of equi-isoclinic fusion frames and the construction of fusion frames for filter banks. Section 13.5 discusses the resilience of fusion frames against the impacts of additive noise, erasures, and perturbations. The relation of fusion frames to the novel paradigm of sparsity—optimally sparse fusion frames and sparse recovery from fusion frame measurements—is the topic of Sect. 13.6. We finish this chapter with the new direction of nonorthogonal fusion frames presented in Sect. 13.7.

13.2 Basics of Fusion Frames

We start by making the intuitive view of fusion frames presented in the introduction mathematically precise. We then state a reconstruction formula for reconstructing signals from fusion frame measurements, which will also require the introduction of the fusion frame operator.

We should mention that fusion frames were initially introduced in the general setting of a Hilbert space. We are restricting ourselves here to the finite dimensional setting, which is of more interest for applications; note that the level of difficulty is not diminished by this restriction.

13.2.1 What Is a Fusion Frame?

Let us start by stating the mathematically precise definition of a fusion frame, which we have already motivated in the introduction.

Definition 13.1 Let $(\mathcal{W}_i)_{i=1}^M$ be a family of subspaces in $\mathcal{H}^N$, and let $(w_i)_{i=1}^M \subseteq \mathbb{R}^+$ be a family of weights. Then $((\mathcal{W}_i, w_i))_{i=1}^M$ is a *fusion frame* for $\mathcal{H}^N$, if there exist constants $0 < A \leq B < \infty$ such that

$$A\|x\|_2^2 \leq \sum_{i=1}^{M} w_i^2 \left\| P_i(x) \right\|_2^2 \leq B\|x\|_2^2 \quad \text{for all } x \in \mathcal{H}^N,$$

where P_i denotes the orthogonal projection onto $\mathcal{W}_i$ for each i. The constants A and B are called the *lower* and *upper fusion frame bound*, respectively. The family $((\mathcal{W}_i, w_i))_{i=1}^M$ is referred to as *tight fusion frame*, if $A = B$ is possible. In this case we also refer to the fusion frame as an *A-tight fusion frame*. Moreover, it is called a *Parseval fusion frame*, if A and B can be chosen as $A = B = 1$. Finally, if $w_i = 1$ for all i, often the notation $(\mathcal{W}_i)_{i=1}^M$ is simply utilized.

To illustrate the notion of a fusion frame, we first present some illuminating examples, which also show the delicateness of constructing fusion frames.

Example 13.1

(a) Let $(e_i)_{i=1}^3$ be an orthonormal basis of $\mathbb{R}^3$, define subspaces $\mathcal{W}_1$ and $\mathcal{W}_2$ by $\mathcal{W}_1 = \text{span}\{e_1, e_2\}$ and $\mathcal{W}_2 = \text{span}\{e_2, e_3\}$, and let w_1 and w_2 be two weights. Then $((\mathcal{W}_i, w_i))_{i=1}^2$ is a fusion frame for $\mathbb{R}^3$ with optimal fusion frame bounds $\min\{w_1^2, w_2^2\}$ and $w_1^2 + w_2^2$. We omit the obvious proof, but mention that this example shows that even changing the weights does not always allow us to turn a fusion frame into a tight fusion frame.

(b) Let now $(\varphi_j)_{j=1}^J$ be a frame for $\mathcal{H}^N$ with bounds A and B. A natural question is whether the set $\{1, \ldots, J\}$ can be partitioned into subsets $J_1, \ldots, J_M$ such that the family of subspaces $\mathcal{W}_i = \text{span}\{\varphi_j : j \in J_i\}$, $i = 1, \ldots, M$, forms a fusion frame with "good" fusion frame bounds — in the sense of their ratio being close to 1—since this ensures a low computational complexity of reconstruction. Remembering the sensor network application, we also seek to choose the partitioning such that $(\varphi_j)_{j \in J_i}$ possesses "good" frame bounds. However, it was shown in [25] that the problem of dividing a frame into a finite number of subsets, each of which has good lower frame bounds, is equivalent to the still-unsolved Kadison-Singer problem; see Sect. 13.1.3. The next subsection will, however, present some computationally possible scenarios for deriving a fusion frame by partitioning a frame into subsets.

13.2.2 Fusion Frames versus Frames

One question when introducing a new notion is its relation to the previously considered classical notion, in this case to *frames*. Our first result shows that fusion frames can be regarded as a generalization of frames in the following sense.

Lemma 13.1 *Let $(\varphi_i)_{i=1}^M$ be a frame for $\mathcal{H}^N$ with frame bounds A and B. Then $(\mathrm{span}\{\varphi_i\}, \|\varphi_i\|_2)_{i=1}^M$ constitutes a fusion frame for $\mathcal{H}^N$ with fusion frame bounds A and B.*

Proof Let P_i be the orthogonal projection onto $\mathrm{span}\{\varphi_i\}$. Then, for all $x \in \mathcal{H}^N$, we have

$$\sum_{i=1}^M \|\varphi_i\|_2^2 \left\| P_i(x) \right\|_2^2 = \sum_{i=1}^M \|\varphi_i\|_2^2 \left\| \left\langle x, \frac{\varphi_i}{\|\varphi_i\|_2} \right\rangle \frac{\varphi_i}{\|\varphi_i\|_2} \right\|_2^2 = \sum_{i=1}^M \left| \langle x, \varphi_i \rangle \right|^2.$$

Applying the definitions of frames and fusion frames finishes the proof. □

On the other hand, if we choose any spanning set inside each subspace of a given fusion frame, the collection of these families of vectors forms a frame for $\mathcal{H}^N$. In this sense, a fusion frame might also be considered as a structured frame. Note though, that this viewpoint depends heavily on the selection of the subspace spanning sets. The next theorem states this local-global interaction in detail.

Theorem 13.1 [21] *Let $(\mathcal{W}_i)_{i=1}^M$ be a family of subspaces in $\mathcal{H}^N$, and let $(w_i)_{i=1}^M \subseteq \mathbb{R}^+$ be a family of weights. Further, let $(\varphi_{ij})_{j=1}^{J_i}$ be a frame for $\mathcal{W}_i$ with frame bounds A_i and B_i for each i, and set $A := \min_i A_i$ and $B := \max_i B_i$. Then the following conditions are equivalent.*

1. *$((\mathcal{W}_i, w_i))_{i=1}^M$ is a fusion frame for $\mathcal{H}^N$.*
2. *$(w_i \varphi_{ij})_{i=1,j=1}^{M,J_i}$ is a frame for $\mathcal{H}^N$.*

In particular, if $((\mathcal{W}_i, w_i))_{i=1}^M$ is a fusion frame with fusion frame bounds C and D, then $(w_i \varphi_{ij})_{i=1,j=1}^{M,J_i}$ is a frame with bounds AC and BD. On the other hand, if $(w_i \varphi_{ij})_{i=1,j=1}^{M,J_i}$ is a frame with bounds C and D, then $((\mathcal{W}_i, w_i))_{i=1}^M$ is a fusion frame with fusion frame bounds $\frac{C}{B}$ and $\frac{D}{A}$.

Proof To prove the theorem, it is sufficient to prove the *in particular* part. For this, first assume that $((\mathcal{W}_i, w_i))_{i=1}^M$ is a fusion frame with fusion frame bounds C and D. Then

$$\begin{aligned} \sum_{i=1}^M w_i^2 \sum_{j=1}^{J_i} \left| \langle x, \varphi_{ij} \rangle \right|^2 &= \sum_{i=1}^M w_i^2 \left[\sum_{j=1}^{J_i} \left| \left\langle P_i(x), \varphi_{ij} \right\rangle \right|^2 \right] \\ &\leq \sum_{i=1}^M w_i^2 B_i \left\| P_i(x) \right\|_2^2 \leq BD \|x\|_2^2. \end{aligned}$$

The lower frame bound AC can be proved similarly.

Secondly, we assume that $(w_i\varphi_{ij})_{i=1,j=1}^{M,J_i}$ is a frame with bounds C and D. In this case, we obtain

$$\sum_{i=1}^{M} w_i^2 \left\| P_i(x) \right\|_2^2 \leq \frac{1}{A} \sum_{i=1}^{M} w_i^2 \left[\sum_{j=1}^{J_i} \left| \left\langle P_i(x), \varphi_{ij} \right\rangle \right|^2 \right] \leq \frac{D}{A} \|x\|_2^2.$$

As before, the lower fusion frame bound $\frac{C}{B}$ can be shown using similar arguments. This finishes the proof. □

The following is an immediate consequence.

Corollary 13.1 *Let $(\mathcal{W}_i)_{i=1}^M$ be a family of subspaces in $\mathcal{H}^N$, and let $(w_i)_{i=1}^M \subseteq \mathbb{R}^+$ be a family of weights. Then $((\mathcal{W}_i, w_i))_{i=1}^M$ is a fusion frame for $\mathcal{H}^N$ if and only if the subspaces $\mathcal{W}_i$ span $\mathcal{H}^N$.*

Since tight fusion frames play a particularly important role due to their advantageous reconstruction properties (see Theorem 13.2), we state the special case of the previous result for tight fusion frames explicitly. It follows immediately from Theorem 13.1.

Corollary 13.2 *Let $(\mathcal{W}_i)_{i=1}^M$ be a family of subspaces in $\mathcal{H}^N$, and let $(w_i)_{i=1}^M \subseteq \mathbb{R}^+$ be a family of weights. Further, let $(\varphi_{ij})_{j=1}^{J_i}$ be an A-tight frame for $\mathcal{W}_i$ for each i. Then the following conditions are equivalent.*

1. *$((\mathcal{W}_i, w_i))_{i=1}^M$ is a C-tight fusion frame for $\mathcal{H}^N$.*
2. *$(w_i\varphi_{ij})_{i=1,j=1}^{M,J_i}$ is an AC-tight frame for $\mathcal{H}^N$.*

This result has an interesting consequence. Since redundancy is the crucial property of a fusion frame and also of a frame, one might be interested in a quantitative way to measure it. In the situation of frames, the rather crude measure of the number of frame vectors divided by the dimension—which is the frame bound in the case of a tight frame with normalized vectors—has recently been replaced by a more appropriate measure, see [5]. In the situation of fusion frames, this is still under investigation. However, as a first notion of redundancy in the situation of a tight fusion frame, we can choose its fusion frame bound as a measure. The following result computes its value.

Proposition 13.1 *Let $((\mathcal{W}_i, w_i))_{i=1}^M$ be an A-tight fusion frame for $\mathcal{H}^N$. Then we have*

$$A = \frac{\sum_{i=1}^M w_i^2 \dim \mathcal{W}_i}{N}.$$

Proof Let $(e_{ij})_{j=1}^{\dim \mathcal{W}_i}$ be an orthonormal basis for $\mathcal{W}_i$ for each $1 \le i \le M$. By Corollary 13.2, the sequence $(w_i e_{ij})_{i=1,j=1}^{M,\dim \mathcal{W}_i}$ is an A-tight frame. Thus, we obtain

$$A = \frac{\sum_{i=1}^{M} \sum_{j=1}^{\dim \mathcal{W}_i} \|w_i e_{ij}\|^2}{N} = \frac{\sum_{i=1}^{M} w_i^2 \dim \mathcal{W}_i}{N}. \qquad \square$$

13.2.3 The Fusion Frame Operator

As discussed before, the fusion frame measurements of a signal $x \in \mathcal{H}^N$ are its (weighted) orthogonal projections onto the given family of subspaces. Consequently, given a fusion frame $\mathcal{W} = ((\mathcal{W}_i, w_i))_{i=1}^{M}$ for $\mathcal{H}^N$, we define the associated *analysis operator* $T_{\mathcal{W}}$ by

$$T_{\mathcal{W}} : \mathcal{H}^N \to \mathbb{R}^{MN}, \qquad x \mapsto \big(w_i P_i(x)\big)_{i=1}^{M}.$$

To reduce the dimension of the representation space $\mathbb{R}^{MN}$, we can select an orthonormal basis in each subspace $\mathcal{W}_i$, which we combine to an $N \times \dim \mathcal{W}_i$-matrix U_i. Then the analysis operator can be modified to $T_{\mathcal{W}}(x) = (w_i U_i^T(x))_{i=1}^{M}$. This approach was undertaken, for instance, in [38].

As is customary in frame theory, the synthesis operator is defined to be the adjoint of the analysis operator. Hence in this situation, the *synthesis operator* $T_{\mathcal{W}}^*$, has the form

$$T_{\mathcal{W}}^* : \mathbb{R}^{MN} \to \mathbb{R}^N, \qquad (y_i)_{i=1}^{M} \mapsto \sum_{i=1}^{M} w_i P_i(y_i).$$

This leads to the following definition of an associated *fusion frame operator* $S_{\mathcal{W}}$:

$$S_{\mathcal{W}} = T_{\mathcal{W}}^* T_{\mathcal{W}} : \mathcal{H}^N \to \mathcal{H}^N, \qquad x \mapsto \sum_{i=1}^{M} w_i^2 P_i(x).$$

13.2.4 Reconstruction Formula

Having introduced a fusion frame operator associated with each fusion frame, we expect it to lead to a reconstruction formula as in the frame theory case. Indeed, a similar result is true, as the following theorem shows.

Theorem 13.2 [21] *Let $\mathcal{W} = ((\mathcal{W}_i, w_i))_{i=1}^{M}$ be a fusion frame for $\mathcal{H}^N$ with fusion frame bounds A and B and associated fusion frame operator $S_{\mathcal{W}}$. Then $S_{\mathcal{W}}$ is a*

positive, self-adjoint, invertible operator on $\mathcal{H}^N$ *with* $A\,Id \leq S_{\mathcal{W}} \leq B\,Id$. *Moreover, we have the reconstruction formula*

$$x = \sum_{i=1}^{M} w_i^2 S_{\mathcal{W}}^{-1}\big(P_i(x)\big) \quad \text{for all } x \in \mathcal{H}^N.$$

Note however, that this reconstruction formula—in contrast to the analogous one for frames—does not automatically lead to a "dual fusion frame." In fact, the appropriate definition of a dual fusion frame is still a topic of research.

Theorem 13.2 immediately implies that a fusion frame is tight if and only if $S_{\mathcal{W}} = A\,Id$, and in this situation the reconstruction formula takes the advantageous form

$$x = A^{-1} \sum_{i=1}^{M} w_i^2 \big(P_i(x)\big) \quad \text{for all } x \in \mathcal{H}^N.$$

This fact makes tight fusion frames particularly attractive for applications.

If practical constraints prevent the utilization or construction of an appropriate tight fusion frame, inverting the fusion frame operator can be still circumvented for reconstruction. Recalling the frame algorithm introduced in Chap. 1, we can generalize it to an iterative algorithm for reconstruction of signals from fusion frame measurements. The proof of the following result follows the arguments of the frame analog very closely; therefore, we omit it.

Proposition 13.2 [22] *Let* $((\mathcal{W}_i, w_i))_{i=1}^M$ *be a fusion frame in* $\mathcal{H}^N$ *with fusion frame operator* $S_{\mathcal{W}}$ *and fusion frame bounds* A *and* B. *Further, let* $x \in \mathcal{H}^N$, *and define the sequence* $(x_n)_{n\in\mathbb{N}_0}$ *by*

$$x_n = \begin{cases} 0, & n = 0, \\ x_{n-1} + \frac{2}{A+B} S_{\mathcal{W}}(x - x_{n-1}), & n \geq 1. \end{cases}$$

Then we have $x = \lim_{n\to\infty} x_n$ *with the error estimate*

$$\|x - x_n\| \leq \left(\frac{B-A}{B+A}\right)^n \|x\|.$$

This algorithm enables reconstruction of a signal x from its fusion frame measurements $(w_i P_i(x))_{i=1}^M$, since $S_{\mathcal{W}}(x)$—necessary for the algorithm—only requires the knowledge of those measurements and of the sequence of weights $(w_i)_{i=1}^M$.

13.3 Fusion Frame Potential

The frame potential, which was introduced in [2] (see also Chap. 1), gives a quantitative estimate of the orthogonality of a system of vectors by measuring the total

potential energy stored in the system under a certain force which encourages orthogonality. It was proven in [16] that, given a complete set of vectors, the minimizers of the associated frame potential are precisely the tight frames. This fact made the frame potential attractive for both theoretical results as well as for deriving large classes of tight frames. However, a slight drawback is the lack of an associated algorithm to actually construct such frames, wherefore these results are mostly used as existence results.

The question of whether a similar quantitative measure exists for fusion frames was answered in [14] by the introduction of a fusion frame potential. These results were significantly generalized and extended in [41]. In this section, we will present a selection of the most fundamental results of this theory.

Let us start by stating the definition of the fusion frame potential. Recalling that in the case of a frame $\Phi = (\varphi_i)_{i=1}^M$ its frame potential is defined by

$$FP(\Phi) = \sum_{i,j=1}^{M} \big|\langle \varphi_i, \varphi_j \rangle\big|^2,$$

it is not initially clear how this can be extended. The following definition from [14] presents a suitable candidate. Note that this includes the classical frame potential by Lemma 13.1.

Definition 13.2 Let $\mathcal{W} = ((\mathcal{W}_i, w_i))_{i=1}^M$ be a fusion frame for $\mathcal{H}^N$ with associated fusion frame operator $S_{\mathcal{W}}$. Then the associated *fusion frame potential* of $\mathcal{W}$ is defined by

$$FFP(\mathcal{W}) = \sum_{i,j=1}^{M} w_i^2 w_j^2 \operatorname{Tr}[P_i P_j] = \operatorname{Tr}\big[S_{\mathcal{W}}^2\big].$$

The following result is immediate.

Lemma 13.2 *Let $\mathcal{W} = ((\mathcal{W}_i, w_i))_{i=1}^M$ be a fusion frame for $\mathcal{H}^N$ with associated fusion frame operator $S_{\mathcal{W}}$, and let $(\lambda_i)_{i=1}^N$ be the eigenvalues of $S_{\mathcal{W}}$. Then*

$$FFP(\mathcal{W}) = \sum_{i=1}^{N} \lambda_i^2.$$

We next define the class of fusion frames over which we seek to minimize the fusion frame potential.

Definition 13.3 Letting $d = (d_i)_{i=1}^M$ be a sequence of positive integers and $w = (w_i)_{i=1}^M$ be a sequence of positive weights, we define the set

$$B_{M,N}(d) = \Big\{\big((\mathcal{W}_i, v_i)\big)_{i=1}^M : \big((\mathcal{W}_i, v_i)\big)_{i=1}^M \text{ is a fusion frame with } \dim \mathcal{W}_i = d_i \text{ for all } i = 1, 2, \ldots, M\Big\}$$

and the two subsets

$$B_{M,N}(d,w) = \left\{ \left((\mathcal{W}_i, v_i)\right)_{i=1}^M \in B_{M,N}(d) \colon v_i = w_i \text{ for all } i = 1, 2, \ldots, M \right\},$$

$$B_{M,N}^1(d) = \left\{ \mathcal{W} = \left((\mathcal{W}_i, v_i)\right)_{i=1}^M \in B_{M,N}(d) : \mathrm{Tr}[S_{\mathcal{W}}] = \sum_{i=1}^M v_i^2 d_i = 1 \right\}.$$

We first focus on the set $B_{M,N}^1(d)$, and start with a crucial property of the fusion frame potential of elements therein. In the following result, by $\|\cdot\|_F$ we denote the Frobenius norm.

Proposition 13.3 [41] *Let* $\mathcal{W} = ((\mathcal{W}_i, w_i))_{i=1}^M \in B_{M,N}^1(d)$; *then*

$$\left\| \frac{1}{N} Id - S_{\mathcal{W}} \right\|_F^2 = FFP(\mathcal{W}) - \frac{1}{N}.$$

Proof Since $\mathrm{Tr}[S_{\mathcal{W}}] = 1$ by the definition of $B_{M,N}^1(d)$, a direct computation shows that

$$\left\| \frac{1}{N} Id - S_{\mathcal{W}} \right\|_F^2 = \mathrm{Tr}\left[\frac{1}{N^2} Id - \frac{2}{N} S_{\mathcal{W}} + S_{\mathcal{W}}^2 \right] = \mathrm{Tr}\left[S_{\mathcal{W}}^2\right] - \frac{1}{N}.$$

The definition of $FFP(\mathcal{W})$ finishes the proof. □

This result implies that minimizing the fusion frame potential over the family of fusion frames of $B_{M,N}^1(d)$ is equivalent to minimizing the Frobenius distance between $S_{\mathcal{W}}$ and a multiple of the identity.

In this spirit the following result does not seem surprising, but it requires a technical proof which we omit here.

Theorem 13.3 [41] *Local minimizers of FFP over* $B_{M,N}^1(d)$ *are global minimizers, and they are tight fusion frames.*

We caution the reader that this theorem does not necessarily imply the existence of local minimizers, only that they are tight fusion frames *if* they exist. Lower bounds of FFP provide a means to show the existence of local minimizers. The following result is a direct consequence of Proposition 13.3.

Corollary 13.3 *Let* $\mathcal{W} \in B_{M,N}^1(d)$. *Then we have* $FFP(\mathcal{W}) \geq \frac{1}{N}$. *Moreover,* $FFP(\mathcal{W}) = \frac{1}{N}$ *if and only if* $\mathcal{W}$ *is a tight fusion frame for* $\mathcal{H}^N$.

We now turn to analyzing the fusion frame potential defined on $B_{M,N}(d, v)$. As a first step, we state a lower bound for FFP, which will also lead to a fundamental equality for tight fusion frames.

Proposition 13.4 [41] *Let $d = (d_i)_{i=1}^M$ be a sequence of positive integers and $w = (w_i)_{i=1}^M$ be a decreasing sequence of positive weights such that $\sum_{i=1}^M w_i^2 d_i = 1$ and $\sum_{i=1}^M d_i \geq N$, and let $\mathcal{W} = ((\mathcal{W}_i, w_i))_{i=1}^M \in B_{M,N}(d, w)$. Further, let $j_0 \in \{1, \ldots, M\}$ be defined by*

$$j_0 = j_0(N, d, v) = \max_{1 \leq j \leq M} \left\{ j : \left(N - \sum_{i=1}^{j} d_i \right) w_j^2 > \sum_{i=j+1}^{M} w_i^2 d_i \right\},$$

and let $j_0 = 0$ if the set is empty. If

$$c := \frac{\sum_{i=j_0+1}^M w_i^2 d_i}{N - \sum_{i=1}^{j_0} d_i} < w_{j_0}^2,$$

then

$$FFP(\mathcal{W}) \geq \sum_{i=1}^{j_0} w_i^4 d_i + \left(N - \sum_{i=j_0+1}^{M} d_i \right) c^2. \tag{13.2}$$

Moreover, we have equality in (13.2) if and only if the following two conditions are satisfied:

(1) $P_i P_j = 0$ *for all* $1 \leq i \neq j \leq j_0$,
(2) $((\mathcal{W}_i, w_i))_{i=j_0+1}^M$ *is a tight fusion frame for* $\operatorname{span}\{\mathcal{W}_i : 1 \leq i \leq j_0\}^{\perp}$.

The main result from [41] is highly technical. Its statement utilizes the notion of admissible $(M+1)$-tuples $(J_0, J_1, \ldots, J_M)$ with

$$J_r = \{1 \leq j_1 < j_2 < \cdots < j_r \leq N\},$$

and an associated partition

$$\lambda(J) = (j_r - r, \ldots, j_1 - 1),$$

where $r \leq N$. Due to lack of space we are not able to go into more detail. We merely mention that an *admissible* $(M+1)$-tuple is defined as one for which the Littlewood-Richardson coefficient of the associated partitions $\lambda(J_0), \ldots, \lambda(J_M)$ is positive [34]. This allows us to phrase the following result.

Theorem 13.4 [41] *Let $d = (d_i)_{i=1}^M$ be a sequence of positive integers satisfying $\sum_i d_i \geq N$, let $w = (w_i)_{i=1}^M$ be a sequence of positive weights, and set $c = \sum_{i=1}^M w_i^2 d_i$. Then the following conditions are equivalent.*

(i) *There exists a $\frac{c}{N}$-tight fusion frame in $B_{M,N}(d, w)$.*
(ii) *For every $1 \leq r \leq N-1$ and every admissible $(M+1)$-tuple $(J_0, \ldots, J_M)$,*

$$\frac{r \cdot c}{N} \leq \sum_{i=1}^{M} w_i^2 \cdot \#\big(J_i \cap \{1, 2, \ldots, d_i\}\big).$$

Finally, we mention that [41] also provides necessary and sufficient conditions for the existence of uniform tight fusion frames by exploiting the Horn-Klyachko inequalities. We refer to [41] for the statement and proof of this deep result.

13.4 Construction of Fusion Frames

Different applications might have different desiderata which a fusion frame is required to satisfy. In this chapter we present three approaches for constructing fusion frames: first, a construction procedure based on a given sequence of eigenvalues of the fusion frame operator; second, a construction which focuses on the angles between subspaces; and third a construction which yields fusion frames with particular filter-bank-like properties.

13.4.1 Spectral Tetris Fusion Frame Constructions

Both from a theoretical standpoint and for applications, we often seek to construct fusion frames with a prescribed sequence of eigenvalues of the fusion frame operator. Examples are the analysis of streaming signals for which a fusion frame must be designed with respect to eigenbases of inverse noise covariance matrices with given associated eigenvalues, similar to water-filling principles for precoder design in wireless communication or face recognition in which significance-weighted bases of eigenfaces might be given.

Let us go back to frame theory for a moment to see how the development of this theory has impacted fusion frame theory. Although unit norm tight frames are the most useful frames in practice, until recently very few techniques for constructing such frames existed. In fact, the main methodology employed was to truncate harmonic frames, and a constructive method for obtaining all equal norm tight frames was available only for $\mathbb{R}^2$ [36]. For years, the field was relying on *existence proofs* given by frame potentials and majorization techniques [24]. A recent significant advance in frame construction occurred with the introduction of Spectral Tetris methods [17] (see Chap. 2). In this paper, Spectral Tetris was used to both classify and construct all tight fusion frames which exist for equal dimensional subspaces and weights equal to one. Quickly afterwards, this was generalized to constructing fusion frames with prescribed fusion frame operators restricted to the case where the eigenvalues are ≥ 2 [11]. It was further generalized in [15] to construct fusion frames $(\mathcal{W}_i)_{i=1}^M$ for $\mathcal{H}^N$ with prescribed eigenvalues for the fusion frame operator and with prescribed dimensions for the subspaces. The results in [15], which include the case of eigenvalues smaller than 2, are achieved by first extending the Spectral Tetris algorithm and changing the basic building blocks from adjusted 2×2 unitary matrices to adjusted $k \times k$ discrete Fourier transform matrices.

13.4.2 Constructing Tight Fusion Frames

We start with a result on the existence and construction of tight fusion frames $((\mathcal{W}_i, w_i))_{i=1}^M$ for $\mathcal{H}^N$ with $M \geq 2N$ for equal dimensional subspaces.

The first result from [17] we present is a slightly technical result which will allow us to immediately construct new tight fusion frames from given ones. The associated procedures are given by the following definitions from [11], which for later use we state for more general non-equal dimensional subspaces.

Definition 13.4 Let $\mathcal{W} = ((\mathcal{W}_i, w_i))_{i=1}^M$ be an A-tight fusion frame for $\mathcal{H}^N$.

(a) If $\dim \mathcal{W}_i < N$ for all $i = 1, \ldots, M$ and $\bigcap_{i=1}^M \mathcal{W}_i = \{0\}$, then the *spatial complement of* $\mathcal{W}$ is defined as the fusion frame

$$\big((\mathcal{W}_i^\perp, w_i)\big)_{i=1}^M.$$

(b) For $i = 1, 2, \ldots, M$, let $(e_{ij})_{j=1}^{m_i}$ be an orthonormal basis for $\mathcal{W}_i$, hence $(\frac{w_i}{\sqrt{A}} e_{ij})_{i=1,\,j=1}^{M,m_i}$ is a Parseval frame for $\mathcal{H}^N$. Set $m = \sum_{i=1}^M m_i$, and let P denote the orthogonal projection which maps an orthonormal basis $(e'_{ij})_{i=1,\,j=1}^{M,m_i}$ for a containing Hilbert space $\mathcal{H}^m$ onto the Parseval frame $(\frac{w_i}{\sqrt{A}} e_{ij})_{i=1,\,j=1}^{M,m_i}$ given by Naimark's theorem (see Chap. 1). Then the fusion frame

$$\Big(\operatorname{span}\big\{(Id - P)e_{ij}\big\}_{j=1}^{m_i}, \sqrt{A - w_i^2}\Big)_{i=1}^M$$

is called the *Naimark complement of* $\mathcal{W}$ *with respect to* $(e_{ij})_{i=1,\,j=1}^{M\,,\,m_i}$.

We should mention that the Naimark complement of a fusion frame depends on the particular choice of initial orthonormal bases for the subspaces. If we do not need to make this dependence explicit, we also speak of *a Naimark complement of* $\mathcal{W}$.

We next quickly check whether in the case of tight fusion frames—our situation in this subsection—this indeed yields tight fusion frames.

Lemma 13.3 *Let $\mathcal{W} = ((\mathcal{W}_i, w_i))_{i=1}^M$ be a tight fusion frame for $\mathcal{H}^N$, not all of whose subspaces equal $\mathcal{H}^N$. Then both the spatial complement and each Naimark complement of $\mathcal{W}$ are tight fusion frames.*

Proof To show the claim for the spatial complement, let $x \in \mathcal{H}^N$ denote the tight frame bound of $\mathcal{W}$ by A, and observe that

$$\sum_{i=1}^M w_i^2 \big\|(Id - P_i)(x)\big\|_2^2 = \sum_{i=1}^M w_i^2 \big(\|x\|_2^2 - \big\|P_i(x)\big\|_2^2\big) = \left(\sum_{i=1}^M w_i^2 - A\right)\|x\|_2^2.$$

Since $\sum_{i=1}^{M} \omega_i^2 - A = 0$ if and only if $\dim \mathcal{W}_i = N$ for all $1 \leq i \leq M$, we have that $((\mathcal{W}_i^{\perp}, w_i))_{i=1}^{M}$ is a tight fusion frame.

Turning to Naimark complements, since

$$\langle Pe_{ij}, Pe_{i\ell} \rangle = -\big\langle (Id - P)e_{ij}, (Id - P)e_{i\ell} \big\rangle,$$

for $j \neq \ell$, it follows that $((Id - P)e_{ij})_{j=1}^{m_i}$ is an orthogonal set. This implies that $(\text{span}\{(Id - P)e_{ij}\}_{j=1}^{m_i}, \sqrt{1 - w_i^2})_{i=1}^{M}$ is a tight fusion frame. □

Armed with these definitions, we can now state and prove our first result from [17].

Proposition 13.5 [17] *Let N, M, and m be positive integers such that $1 < m < N$.*

(i) *There exist tight fusion frames $((\mathcal{W}_i, w_i))_{i=1}^{M}$ for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$ if and only if tight fusion frames $((\mathcal{V}_i, v_i))_{i=1}^{M}$ for $\mathcal{H}^N$ with $\dim \mathcal{V}_i = N - m$ for all $i = 1, \ldots, M$ exist.*

(ii) *There exist tight fusion frames $((\mathcal{W}_i, w_i))_{i=1}^{M}$ for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$ if and only if tight fusion frames $((\mathcal{V}_i, v_i))_{i=1}^{M}$ for $\mathbb{R}^{Mm-N}$ with $\dim \mathcal{V}_i = (M-1)m - N$ for all $i = 1, \ldots, M$ exist.*

Proof Part (i) follows directly by taking the spatial complement and then using Lemma 13.3. Part (ii) follows from repeated spatial complement constructions followed by applications of Naimark complements and again application of Lemma 13.3. □

We now turn to the main theorem of this subsection, which can be used to answer the question: For a given triple (M, m, N) of positive integers, does a tight fusion frame (with weights equal to one) of M subspaces of equal dimension m exist for $\mathcal{H}^N$? The result is not merely an existence result but answers the question by explicitly constructing a fusion frame of the given parameters in most cases where one exists. Therefore, besides our previous construction of fusion frames from given ones through complement methods, we need a construction for fusion frames to begin with. Using Theorem 13.1, one way to construct a tight fusion frame with the parameters (M, m, N) is to construct a tight unit norm frame $(\varphi_{i,j})_{i=1,j=1}^{M,m}$ of Mm elements for $\mathcal{H}^N$, such that $(\varphi_{i,j})_{j=1}^{m}$ is an orthogonal sequence for all $i = 1, \ldots, M$. We can then define the desired tight fusion frame $(\mathcal{W}_i)_{i=1}^{M}$ by letting $\mathcal{W}_i$ be the span of $(\varphi_{i,j})_{j=1}^{m}$ for $i = 1, \ldots, M$.

The tool of choice to construct unit norm tight frames whose elements can be partitioned into sets of orthogonal vectors is the Spectral Tetris construction (see Chap. 2). In general, fusion frame constructions involving Spectral Tetris work due to the fact that frames constructed via Spectral Tetris are sparse (cf. also Sect. 13.6). The sparsity property ensures that the constructed frames can be partitioned into sets of orthonormal vectors, the spans of which are the desired fusion frames.

Theorem 13.5 [17] *Let N, M, and m be positive integers such that $m \leq N$.*

(i) *Suppose that $m|N$. Then there exist tight fusion frames $(\mathcal{W}_i)_{i=1}^M$ for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$ if and only if $M \geq \frac{N}{m}$.*

(ii) *Suppose that $m \not| N$. Then the following is true.*

(a) *If there exists a tight fusion frame $(\mathcal{W}_i)_{i=1}^M$ for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$, then $M \geq \lceil \frac{N}{m} \rceil + 1$.*

(b) *If $M \geq \lceil \frac{N}{m} \rceil + 2$, then tight fusion frames $(\mathcal{W}_i)_{i=1}^M$ for $\mathbb{C}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$ do exist.*

Proof (Sketch of proof) (i) Suppose that there exists a tight fusion frame $(\mathcal{W}_i)_{i=1}^M$ for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$. Then any collection of spanning sets for its subspaces consists of at least Mm vectors which span $\mathcal{H}^N$; thus $M \geq \frac{N}{m}$.

Conversely, assume that $M \geq \frac{N}{m}$ with $K := \frac{N}{m}$ being an integer by assumption. Let $(e_j)_{j=1}^K$ be an orthonormal basis for $\mathcal{H}^K$. There exists a unit norm tight frame $(\varphi_i)_{i=1}^M$ for $\mathcal{H}^K$ (see Chap. 1). Now consider the m sets of orthonormal bases given by $(e_{i+(k-1)m})_{k=1}^K$ for $i = 1, \ldots, m$, and project the tight frame elements onto each of the generated spaces, leading to m unit norm tight frames $(\varphi_{ij})_{i=1}^M$ for $j = 0, \ldots, m-1$. Setting $\mathcal{W}_i = \text{span}\{\varphi_{ij} : j = 0, \ldots, m-1\}$, we obtain the required fusion frame.

(ii)(a) If there exists a tight fusion frame $(\mathcal{W}_i)_{i=1}^M$ for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$, then $M \geq \frac{N}{m}$. Since m does not divide N, it follows that $M > \frac{N}{m}$. Hence, by Lemma 13.3, there exists a tight fusion frame $(\mathcal{V}_i)_{i=1}^M$ for $\mathcal{H}^{Mm-N}$ with $\dim \mathcal{V}_i = m$ for all $i = 1, \ldots, M$. Thus, there exist m orthonormal vectors in $\mathcal{H}^{Mm-N}$ implying that $m \leq Mm - N$. Hence, $M \geq \frac{N}{m} + 1$. The claim follows now from the fact that M is an integer.

(ii)(b) This part of the proof uses the sparsity of frames generated by Spectral Tetris. For the arguments we refer to [17], and just remark that, first since Spectral Tetris can in general only be used to construct frames consisting of at least twice as many vectors as the dimension of the space, spatial complements have to be used. Second, the orthogonality relations of the frames constructed by Spectral Tetris then allow us to stack modulated copies of such frames, resulting in complex *Gabor fusion frames*. □

Theorem 13.5 leaves one case unanswered. Does a tight fusion frame of M subspaces of equal dimension m exist in $\mathbb{C}^N$ in the case that m does not divide N and $M = \lceil \frac{N}{m} \rceil + 1$? As it happens, the answer is *sometimes yes* and *sometimes no*. Which it is can be decided by repeatedly using Theorem 13.5 in conjunction with Proposition 13.5 for at most $m - 1$ times; we again refer to [17] for the details. Also note that this result answers a nontrivial problem in operator theory; i.e., it classifies the triples (N, M, m) so that an N-dimensional Hilbert space has M rank m projections which sum to a multiple of the identity.

13.4.3 Spectral Tetris Constructions of General Fusion Frames

We next discuss a general construction introduced in [15], encompassing different eigenvalues of the fusion frame operator as well as different dimensions of the subspaces, therefore including [11] as a special case.

We start by introducing a *reference fusion frame* for a given sequence of eigenvalues. This carefully constructed fusion frame—while having prescribed eigenvalues for its fusion frame operator—will have the striking property that the dimensions of its subspaces are in a certain sense "maximal," allowing for a given sequence of dimensions to decide whether an associated fusion frame can be constructed using the generalized Spectral Tetris algorithm STC presented in Fig. 13.1 (cf. [11]). This algorithm is a straightforward generalization of the original Spectral Tetris algorithm from the case of tight frames to the case of frames with prescribed spectrum for the frame operator; i.e., now the rows of the synthesis matrix that is being constructed square sum to the respective prescribed eigenvalues. We will say a tight fusion frame is *constructible via STC*, if there is a frame constructed by STC whose vectors can be partitioned in such a way that the vectors in each set of the partition are orthogonal and span the respective subspaces of the fusion frame.

The construction of the reference fusion frame for a prescribed sequence of eigenvalues is achieved by the following algorithm called RFF (Fig. 13.2). We will denote the reference fusion frame constructed for the sequence $(\lambda_j)_{j=1}^N$ via RFF by $RFF((\lambda_j)_{j=1}^N)$. In RFF and the following results of this section we restrict ourselves to the case of eigenvalues ≥ 2 and just want to mention that this restriction is dropped in [15], where the general case is handled by first extending the Spectral Tetris construction.

The main goal will now be to derive necessary and sufficient conditions for the constructibility of a fusion frame with prescribed eigenvalues of the fusion frame operator and prescribed dimensions of its subspaces via STC. This will require us to compare the dimensions of the subspaces of a reference fusion frame constructed by RFF with the prescribed sequence of dimensions.

We first need to recall the notion of majorization. Given a sequence $a = (a_n)_{n=1}^N \in \mathcal{H}^N$, we will denote the sequence obtained by rearranging the coordinates of a in decreasing order by $a^{\downarrow} \in \mathcal{H}^N$. For $(a_n)_{n=1}^N, (b_n)_{n=1}^N \in \mathcal{H}^N$, the sequence $(a_n)_{n=1}^N$ *majorizes* $(b_n)_{n=1}^N$, denoted by $(a_n) \succeq (b_n)$, provided that $\sum_{n=1}^m a_n^{\downarrow} \geq \sum_{n=1}^m b_n^{\downarrow}$ for all $m = 1, \ldots, N-1$ and $\sum_{n=1}^N a_n = \sum_{n=1}^N b_n$.

This notion will be the key ingredient for deriving a characterization of the constructibility via Spectral Tetris of a fusion frame with prescribed eigenvalues and dimensions. We note that we will also use the notion of majorization between sequences of different lengths by agreeing to add zero entries to the shorter sequence in order to have sequences of the same length.

The proof of the following condition is constructive, and we refer to [15] for how to iteratively construct the desired fusion frame starting from the reference fusion frame.

STC: SPECTRAL TETRIS CONSTRUCTION FOR PRESCRIBED EIGENVALUES

Parameters:

- Dimension: N.
- Number of frame vectors: M.
- Eigenvalues: $(\lambda_j)_{j=1}^N \subseteq [2,\infty)$ satisfying $\sum_{j=1}^N \lambda_j = M$.

Algorithm:

Set $k := 1$.
For $j = 1, \ldots, N$ do
  Repeat
    If $\lambda_j < 2$ and $\lambda_j \neq 1$ then
      $\varphi_k := \sqrt{\frac{\lambda_j}{2}} \cdot e_j + \sqrt{1 - \frac{\lambda_j}{2}} \cdot e_{j+1}$.
      $\varphi_{k+1} := \sqrt{\frac{\lambda_j}{2}} \cdot e_j - \sqrt{1 - \frac{\lambda_j}{2}} \cdot e_{j+1}$.
      $k := k + 2$.
      $\lambda_j := 0$.
      $\lambda_{j+1} := \lambda_{j+1} - (2 - \lambda_j)$.
    else
      $\varphi_k := e_j$.
      $k := k + 1$.
      $\lambda_j := \lambda_j - 1$.
    end;
  until $\lambda_j = 0$.
end;

Output:

- Unit norm $(\varphi_i)_{i=1}^M \subset \mathcal{H}^N$ with eigenvalues $(\lambda_j)_{j=1}^N$ for its frame operator.

Fig. 13.1 The STC for constructing a frame with prescribed spectrum of its frame operator

Theorem 13.6 [15] *Let M, N be positive integers with $M \geq 2N$, let $(\lambda_j)_{j=1}^N \subseteq [2,\infty)$, and let $(d_i)_{i=1}^D$ be a sequence of positive integers such that $\sum_{j=1}^N \lambda_j = \sum_{i=1}^D d_i = M$. Further, let $(\mathcal{V}_i)_{i=1}^K = RFF((\lambda_j)_{j=1}^N)$. If $(\dim \mathcal{V}_i) \succeq (d_i)$, then a fusion frame $(\mathcal{W}_i)_{i=1}^D$ for $\mathcal{H}^N$ such that $\dim \mathcal{W}_i = d_i$ for $i = 1, \ldots, D$ and whose fusion frame operator has the eigenvalues $(\lambda_j)_{j=1}^N$ can be constructed via STC.*

In the special case of tight fusion frames the majorization condition is also necessary for constructibility via a partitioning into orthonormal sets of a frame constructed via STC.

Theorem 13.7 [15] *Let M, N be positive integers with $M \geq 2N$, and let $(d_i)_{i=1}^D$ be a sequence of positive integers such that $\sum_{i=1}^D d_i = M$. Further, let $(\mathcal{V}_i)_{i=1}^K = RFF((\lambda_j)_{j=1}^N)$ with $(\lambda_j)_{j=1}^N = (\frac{M}{N}, \ldots, \frac{M}{N})$. Then the following conditions are equivalent.*

RFF (REFERENCE FUSION FRAME)

Parameters:

- Dimension: N.
- Eigenvalues: $(\lambda_j)_{j=1}^N \subseteq [2, \infty)$.

Algorithm:

1) Run STC for $(\lambda_j)_{j=1}^N$ and $M := \sum_{j=1}^N \lambda_j$ to obtain the frame $(\varphi_i)_{i=1}^M$.
2) K :=maximal support size of the rows of the synthesis matrix of $(\varphi_i)_{i=1}^M$.
3) $S_i := \emptyset$ for $i = 1, \dots, K$.
4) $k := 0$.
5) Repeat
6) $\quad k := k + 1$.
7) $\quad j := \min\{1 \le r \le K : \operatorname{supp} \varphi_k \cap \operatorname{supp} \varphi_s = \emptyset \text{ for all } \varphi_s \in S_r\}$.
8) $\quad S_j := S_j \cup \{\varphi_k\}$.
9) until $k = M$.

Output:

- Fusion frame $(\mathcal{V}_i)_{i=1}^K$, where $\mathcal{V}_i = \operatorname{span} S_i$ for $i = 1, \dots, K$.

Fig. 13.2 The RFF algorithm for constructing the reference fusion frame

(i) *A tight fusion frame* $(\mathcal{W}_i)_{i=1}^M$ *for* $\mathcal{H}^N$ *with* $\dim \mathcal{W}_i = d_i$ *for* $i = 1, \dots, M$, *is constructible via STC.*
(ii) $(\dim \mathcal{V}_i) \succeq (d_i)$.

13.4.4 Equi-Isoclinic Fusion Frames

Equal norm equiangular Parseval frames are highly useful for applications, in particular due to their optimal erasure resilience alongside an optimal condition number of the synthesis matrix. Examples include reconstruction without phase [1] and quantum state tomography [46].

The fusion frame analog of this class of Parseval frames is that of fusion frames whose subspaces have equal chordal distances or—as the stricter requirement—whose subspaces are equi-isoclinic [39]. The notion of *chordal distance* was introduced by Conway, Hardin, and Sloane in [27], whereas the notion of equi-isoclinic subspaces was introduced by Lemmens and Seidel in [39], the latter being further studied by Hoggar [37] and others [30–32, 35]. Similarly as in frame theory, this analog class of fusion frames—with equal chordal distances as well as with equi-isoclinic subspaces—is also optimally resilient against noise and erasures. For more details, we refer to the discussion in Sect. 13.5.2. At this point, to provide a first intuitive understanding, let us just mention that this class of fusion frames distributes the incoming energy most evenly to the fusion frame measurements.

As a prerequisite we first require the notion of principal angles.

Definition 13.5 Let $\mathcal{W}_1$ and $\mathcal{W}_2$ be subspaces of $\mathcal{H}^N$ with $m := \dim \mathcal{W}_1 \leq \dim \mathcal{W}_2$. Then the *principal angles* $\theta_1, \theta_2, \ldots, \theta_m$ between $\mathcal{W}_1$ and $\mathcal{W}_2$ are defined as follows.

Let

$$\theta_1 = \min\left\{\arccos\left(\frac{\langle x_1, x_2\rangle}{\|x_1\|_2\|x_2\|_2}\right) : x_i \in \mathcal{W}_i, i = 1, 2\right\}$$

be the first principal angle, and let $x_i^{(1)} \in \mathcal{W}_i$, $i = 1, 2$ be chosen such that

$$\cos\theta_1 = \frac{\langle x_1^{(1)}, x_2^{(1)}\rangle}{\|x_1^{(1)}\|_2\|x_2^{(1)}\|_2}.$$

Then, for any $1 \leq j \leq m$, the principal angle θ_j is defined recursively by

$$\theta_j = \min\left\{\arccos\left(\frac{\langle x_1, x_2\rangle}{\|x_1\|_2\|x_2\|_2}\right) : x_i \in \mathcal{W}_i, x_i \perp x_i^{(\ell)}\ \forall 1 \leq \ell \leq j-1, i = 1, 2\right\},$$

and letting $x_i^{(j)} \in \mathcal{W}_i$ with $x_i \perp x_i^{(\ell)}$ for all $1 \leq \ell \leq j-1$, $i = 1, 2$ be chosen such that

$$\cos\theta_j = \frac{\langle x_1^{(j)}, x_2^{(j)}\rangle}{\|x_1^{(j)}\|_2\|x_2^{(j)}\|_2}.$$

Armed with this notion, we can now introduce the notion of chordal distance and isoclinicness.

Definition 13.6 Let $\mathcal{W}_1$ and $\mathcal{W}_2$ be subspaces of $\mathcal{H}^N$ with $m := \dim \mathcal{W}_1 = \dim \mathcal{W}_2$ and denote by P_i the orthogonal projection onto $\mathcal{W}_i$, $i = 1, 2$. Further, let $(\theta_j)_{j=1}^m$ denote the principal angles for this pair.

(a) The *chordal distance* $d_c(\mathcal{W}_1, \mathcal{W}_2)$ between $\mathcal{W}_1$ and $\mathcal{W}_2$ is given by

$$d_c^2(\mathcal{W}_1, \mathcal{W}_2) = m - \mathrm{Tr}[P_1 P_2] = m - \sum_{j=1}^{m} \cos^2\theta_j.$$

(b) The subspaces $\mathcal{W}_1$ and $\mathcal{W}_2$ are called *isoclinic*, if

$$\theta_{j_1} = \theta_{j_2} \quad \text{for all } 1 \leq j_1, j_2 \leq m.$$

Multiple subspaces are called *equi-isoclinic*, if they are pairwise isoclinic.

Part (b) of Definition 13.6 is an equivalent formulation of the standard definition. The main result of this subsection will be a construction of an equi-isoclinic fusion frame, meaning a fusion frame with equi-isoclinic subspaces. One main ingredient is the method of a Naimark complement (cf. Definition 13.4). As a first

step—and also as an interesting result on its own—we analyze the change of the principal angles under computing a Naimark complement. The proof is a straightforward computation, and we refer to [18] for the details.

Theorem 13.8 [18] *Let $((\mathcal{W}_i, w_i))_{i=1}^M$ be a Parseval fusion frame for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $1 \le i \le M$, and let $((\mathcal{W}_i', \sqrt{1-w_i^2}))_{i=1}^M$ be a Naimark complement of it. For $1 \le i_1 \ne i_2 \le M$, we denote the principal angles for the pair of subspaces $\mathcal{W}_{i_1}, \mathcal{W}_{i_2}$ by $(\theta_j^{(i_1 i_2)})_{j=1}^m$. Then the principal angles for the pair $\mathcal{W}_{i_1}', \mathcal{W}_{i_2}'$ are*

$$\left(\arccos\left(\frac{w_{i_1}}{\sqrt{1-w_{i_1}^2}} \cdot \frac{w_{i_2}}{\sqrt{1-w_{i_2}^2}} \cdot \cos\left(\theta_j^{(i_1 i_2)}\right)\right)\right)_{j=1}^M.$$

Next, we utilize this result to provide a method to construct equi-isoclinic fusion frames, which was developed in [7].

Theorem 13.9 [7] *Let $(e_{ij})_{i=1,j=1}^{M,N}$ be a union of M orthonormal bases for $\mathcal{H}^N$. Then $(\operatorname{span}\{e_{ij} : j = 1, \dots, N\}, \sqrt{1/M})_{i=1}^M$ is a Parseval fusion frame for $\mathcal{H}^N$, and we let $(\mathcal{W}_i', \sqrt{(M-1)/M})_{i=1}^M$ denote the Parseval fusion frame for $\mathbb{R}^{(M-1)N}$ derived as its Naimark complement with respect to $(e_{ij})_{i=1,j=1}^{M,N}$. Then the following hold.*

(i) *For all $i \in \{1, 2, \dots, M\}$, we have*

$$\operatorname{span}\left\{\mathcal{W}_{i'}'\right\}_{i' \ne i} = \mathbb{R}^{(M-1)N}.$$

(ii) *The principal angles for the pair $\mathcal{W}_{i_1}', \mathcal{W}_{i_2}'$ are given by*

$$\theta_j^{(i_1 i_2)} = \arccos\left(\frac{1}{M-1}\right).$$

Thus, $(\mathcal{W}_i', \sqrt{(M-1)/M})_{i=1}^M$ forms an equi-isoclinic Parseval fusion frame.

Proof The fact that $(\operatorname{span}\{e_{ij} : j = 1, \dots, N\}, \sqrt{1/M})_{i=1}^M$ is a Parseval fusion frame for $\mathcal{H}^N$ is immediate. Let now $P : \mathbb{R}^{MN} \to \mathcal{H}^N$ denote the orthogonal projection given by Naimark's theorem, so that $e_{ij} = \sqrt{1/M} \cdot P e_{ij}'$ for some orthonormal basis $(e_{ij}')_{i=1,j=1}^{M,N}$ in $\mathbb{R}^{MN}$.

(i) Since, for a fixed i, the set $(e_{ij})_{j=1}^N$ is linearly independent, [6, Corollary 2.6] implies that

$$\mathcal{W}_i' = \operatorname{span}\left\{(Id - P)e_{i'j'} : i' \ne i\right\} \quad \text{for all } i = 1, \dots, M.$$

This proves (i).

(ii) For this, let $i_1 \neq i_2 \in \{1, \ldots, M\}$. Note that the principal angles for the pair $\mathcal{W}_{i_1}, \mathcal{W}_{i_2}$ are all equal to 0. Hence, by Theorem 13.8, principal angles for the pair $\mathcal{W}'_{i_1}, \mathcal{W}'_{i_2}$ are given by

$$\arccos\left(\frac{\frac{1}{\sqrt{M}}}{\sqrt{1-(\frac{1}{\sqrt{M}})^2}}\frac{\frac{1}{\sqrt{M}}}{\sqrt{1-(\frac{1}{\sqrt{M}})^2}}\cos 0\right) = \arccos\left(\frac{1}{M-1}\right).$$

Thus, (ii) is also proved. □

We now present a particularly interesting special case of this result, namely, when the family $(e_{ij})_{i=1,j=1}^{M,N}$ is chosen to be a family of *mutually unbiased bases*. We first define this notion.

Definition 13.7 A family of orthonormal sequences $\{e_{ij}\}_{i=1}^{M}$, $j = 1, \ldots, L$, in $\mathcal{H}^N$ is called *mutually unbiased* if there exists a constant $c > 0$ such that

$$\left|\langle e_{i_1 j_1}, e_{i_2 j_2}\rangle\right| = c \quad \text{for all } j_1 \neq j_2.$$

If $N = M$, then necessarily $c = \sqrt{1/N}$, and we refer to $\{e_{ij}\}_{i=1,j=1}^{M,L}$ as a *family of mutually unbiased bases*.

Now choosing $(e_{ij})_{i=1,j=1}^{M,N}$ to be a family of mutually unbiased bases leads to the following special case of Theorem 13.9.

Corollary 13.4 *Let $(e_{ij})_{i=1,j=1}^{M,N}$ be a family of mutually unbiased bases for $\mathcal{H}^N$. Then $(\text{span}\{e_{ij} : j = 1, \ldots, N\}, \sqrt{1/M})$ is a Parseval fusion frame for $\mathcal{H}^N$, and we let $(\mathcal{W}'_i, \sqrt{(M-1)/M})_{j=1}^{M}$ denote the Parseval fusion frame for $\mathbb{R}^{(M-1)N}$ derived as its Naimark complement with respect to $(e_{ij})_{i=1,j=1}^{M,N}$. Then $(\mathcal{W}'_i, \sqrt{(M-1)/M})_{j=1}^{M}$ is an equi-isoclinic fusion frame, and, moreover, the subspaces $\mathcal{W}'_i$ are spanned by mutually unbiased sequences.*

Since mutually unbiased bases are known to exist in all prime power dimensions p^r [47], this result implies the existence of Parseval fusion frames with $M \leq p^r + 1$ equi-isoclinic subspaces of dimension p^r, spanned by mutually unbiased basic sequences in $\mathbb{R}^{(M-1)p^r}$. If neither equidistance nor equi-isoclinic Parseval fusion frames are realizable, a weaker version are families of subspaces with at most two different values; see [12].

Finally, we mention that a different class of equi-isoclinic fusion frames was recently introduced in [7] by using multiple copies of orthonormal bases.

13.4.5 Fusion Frame Filter Banks

In [26], the first efficiently implementable construction of fusion frames was derived. The main idea is to use specifically designed oversampled filter banks. A *filter* is a linear operator which computes the inner products of an input signal with all translates of a fixed function. In a *filter bank*, several filters are applied to the input, and each of the resulting signals is then downsampled.

The problem in designing filter bank frames is to ensure that they satisfy the large number of conditions needed on the frame for the typical application. An important tool here is the *polyphase matrix*. The fundamental works on filter bank frames [8, 28] characterize translation-invariant frames in $\ell^2(\mathbb{Z})$ in terms of polyphase matrices. In particular, filter bank frames are characterized in [28], and [8] derives the optimal frame bounds of a filter bank frame in terms of the singular values of its polyphase matrix. In the paper [26], these characterizations are then subsequently utilized to construct filter bank fusion frame versions of discrete wavelet and Gabor transforms.

13.5 Robustness of Fusion Frames

Applications naturally call for robustness, which could mean resilience against noise and erasures or stability under perturbation. In this section we will give an introduction to several types of robustness properties of fusion frames.

13.5.1 Noise

One main advantage of redundancy is its property to provide resilience against noise and erasures. Theoretical guarantees for a given fusion frame are determined only in the situation of random signals; see [38]. Note that we focus on non-weighted fusion frames in this subsection.

13.5.1.1 Stochastic signal model

Let $(\mathcal{W}_i)_{i=1}^M$ be a fusion frame for $\mathbb{R}^N$ with bounds A and B, and for $i = 1, \ldots, M$, let m_i be the dimension of $\mathcal{W}_i$ and let U_i be an $N \times m_i$-matrix whose columns form an orthonormal basis of $\mathcal{W}_i$ for $i = 1, \ldots, M$. Further, let $x \in \mathbb{R}^N$ be a zero-mean random vector with covariance matrix $E[xx^T] = R_{xx} = \sigma_x^2 Id$. The noisy fusion frame measurements can then be modeled as

$$z_i = U_i^T x + n_i, \quad i = 1, \ldots, M,$$

where $n_i \in \mathbb{R}^{m_i}$ is an additive white noise vector with zero mean and covariance matrix $E[n_i n_i^T] = \sigma_n^2 Id$, $i = 1, \ldots, M$. It is assumed that the noise vectors for different subspaces are mutually uncorrelated and that the signal vector x and the noise vectors n_i, $i = 1, \ldots, N$, are uncorrelated.

Setting

$$z = \begin{pmatrix} z_1^T & z_2^T & \cdots & z_M^T \end{pmatrix}^T \quad \text{and} \quad U = (U_1 \ U_2 \ \cdots \ U_M),$$

the composite covariance matrix between x and z can be written as

$$E\left[\begin{pmatrix} x \\ z \end{pmatrix} \begin{pmatrix} x^T & z^T \end{pmatrix}\right] = \begin{pmatrix} R_{xx} & R_{xz} \\ R_{zx} & R_{zz} \end{pmatrix},$$

where

$$R_{xz} = E\left[x z^T\right] = R_{xx} U$$

is the $M \times L$ ($L = \sum_{i=1}^{M} m_i$) cross-covariance matrix between x and z, $R_{zx} = R_{xz}^T$, and

$$R_{zz} = E\left[z z^T\right] = U^T R_{xx} U + \sigma_n^2 Id_L$$

is the $L \times L$ composite measurement covariance matrix. The linear mean squared error (MSE) minimizer for estimating x from z is the Wiener filter or the linear minimum mean squared error (LMMSE) filter $F = R_{xz} R_{zz}^{-1}$, which estimates x by $\hat{x} = Fz$. Then the associated error covariance matrix R_{ee} is given by

$$R_{ee} = E\left[(x - \hat{x})(x - \hat{x})^T\right] = \left(R_{xx}^{-1} + \frac{1}{\sigma_n^2} \sum_{i=1}^{M} P_i\right)^{-1},$$

which is derived using the Sherman-Morrison-Woodbury formula. The MSE is obtained by taking the trace of R_{ee}.

A result from [38] shows that, as in the frame case, a fusion frame is optimally resilient against noise if it is tight.

Theorem 13.10 [38] *Assuming the model previously introduced, the following conditions are equivalent.*

(i) *The MSE is minimized.*
(ii) *The fusion frame is tight.*

In this case, the MSE is given by

$$MSE = \frac{N \sigma_n^2 \sigma_x^2}{\sigma_n^2 + \frac{\sigma_x^2 L}{N}}.$$

Proof Since $R_{xx} = \sigma_x^2 Id$ and denoting the frame bounds by A and B, we obtain

$$\frac{N}{\frac{1}{\sigma_x^2} + \frac{B}{\sigma_n^2}} \leq \big(MSE = \mathrm{Tr}[R_{ee}]\big) \leq \frac{N}{\frac{1}{\sigma_x^2} + \frac{A}{\sigma_n^2}}.$$

This implies that the lower bound will be achieved, provided that the fusion frame is tight. The explicit value of the MSE follows from here. □

13.5.2 Erasures

Similar to resilience against noise, redundancy is also beneficial for resilience against erasures. Again, we can distinguish between a deterministic and a stochastic signal model. The first case was analyzed in [4], whereas the second case was studied in [38]. As before, in this subsection we focus on non-weighted fusion frames.

13.5.2.1 Deterministic signal model

Let $\mathcal{W} = (\mathcal{W}_i)_{i=1}^M$ be a fusion frame for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$. Further, let $T_{\mathcal{W}}$ and $S_{\mathcal{W}}$ be the associated analysis and fusion frame operator, respectively.

The loss of a set of subspaces will be modeled deterministically in the following way. Given $K \subseteq \{1, \ldots, M\}$, the associated operator modeling erasures is defined by

$$E_K : \mathbb{R}^{MN} \to \mathbb{R}^{MN}, \qquad E_K\big((x_i)_{i=1}^M\big)_j = \begin{cases} x_j: & j \notin K, \\ 0: & j \in K. \end{cases}$$

The next ingredient of the model is the measure for the imposed error. In [4], the worst case measure was chosen, which in the case of k lost subspaces is defined by

$$e_k(\mathcal{W}) = \max\big\{\big\|Id - S_{\mathcal{W}}^{-1} T_{\mathcal{W}}^* E_K T_{\mathcal{W}}\big\| : K \subset \{1, \ldots, M\}, |K| = k\big\}.$$

We first state the result from [4] for one subspace erasure.

Theorem 13.11 [4] *Assuming the model previously introduced, the following conditions are equivalent.*

(i) *The worst case error $e_1(\mathcal{W})$ is minimized.*
(ii) *The fusion frame $\mathcal{W}$ is a Parseval fusion frame.*

Proof Setting $D_K := Id - E_K$ for some $K \subset \{1, \ldots, M\}$ with $K = \{i_0\}$, we obtain

$$\big\|Id - S_{\mathcal{W}}^{-1} T_{\mathcal{W}}^* E_K T_{\mathcal{W}}\big\| = \big\|S_{\mathcal{W}}^{-1} T_{\mathcal{W}}^* D_K T_{\mathcal{W}}\big\| = \big\|S_{\mathcal{W}}^{-1} P_{i_0}\big\|.$$

Hence, the quantity

$$e_1(\mathcal{W}) = \max\left\{\left\|S_{\mathcal{W}}^{-1} P_{i_0}\right\| : i_0 \in \{1, \ldots, M\}\right\}$$

needs to be minimized. This is achieved if and only if $S_{\mathcal{W}} = Id$, which is equivalent to $\mathcal{W}$ being a Parseval fusion frame. □

To analyze the situation of two subspace erasures, we now restrict ourselves to the class of fusion frames, already shown to behave optimally under one erasure, and reduce the measure $e_2(\mathcal{W})$ accordingly. Then the following result is true; we refer to [4] for its lengthy proof.

Theorem 13.12 [4] *Assuming the model previously introduced, the following conditions are equivalent.*

(i) *The worst case error $e_2(\mathcal{W})$ is minimized.*
(ii) *The fusion frame $\mathcal{W}$ is an equi-isoclinic fusion frame.*

This shows the need to develop construction methodologies for equi-isoclinic fusion frames, and we refer the reader to Sect. 13.4.4 for details.

13.5.2.2 Stochastic signal model

We assume the model already detailed in Sect. 13.5.1. By Theorem 13.10, tight fusion frames are maximally robust against noise. Hence, from now on we restrict ourselves to tight fusion frames and study within this class which fusion frames are optimally resilient with respect to one, two, and more erasures. Also, we mention that all erasures are considered equally important.

Again, the MSE shall be determined when the LMMSE filter F, as defined before, is applied to a measurement vector now with erasures. To model the erasures, let $K \subset \{1, 2, \ldots, M\}$ be the set of indices corresponding to the erased subspaces. Then, the measurements take the form

$$\tilde{z} = (Id - E)z,$$

where E is an $L \times L$ block diagonal erasure matrix whose ith diagonal block is an $m_i \times m_i$ zero matrix, if $i \notin K$, or an $m_i \times m_i$ identity matrix, if $i \in K$.

The estimate of x is now given by

$$\tilde{x} = F\tilde{z},$$

with associated error covariance matrix

$$\widetilde{R}_{ee} = E\left[(x - \tilde{x})(x - \tilde{x})^T\right] = E\left[\big(x - F(Id - E)z\big)\big(x - F(Id - E)z\big)^T\right].$$

The MSE for this estimate can be written as

$$MSE = \mathrm{Tr}[\widetilde{R}_{ee}] = MSE_0 + \overline{MSE},$$

where $MSE_0 = \mathrm{Tr}[R_{ee}]$ and $\overline{MSE}$ is the extra MSE due to erasures given by

$$\overline{MSE} = \alpha^2 \,\mathrm{Tr}\left[\sigma_x^2 \left(\sum_{i\in\mathbb{S}} P_i\right)^2 + \sigma_n^2 \left(\sum_{i\in\mathbb{S}} P_i\right)\right],$$

where $\alpha = \sigma_x^2/(A\sigma_x^2 + \sigma_n^2)$.

This leads to the following result from [38] for one subspace. We also refer to this paper for its proof.

Theorem 13.13 [38] *Assuming the model previously introduced and letting* $(\mathcal{W}_i)_{i=1}^M$ *be a tight fusion frame, the following conditions are equivalent.*

(i) *The MSE due to the erasure of one subspace is minimized.*
(ii) *All subspaces* $\mathcal{W}_i$ *have the same dimension; i.e.,* $(\mathcal{W}_i)_{i=1}^M$ *is an* equidimensional *fusion frame.*

Recalling the definition of *chordal distance* $d_c(i, j)$ from Sect. 13.4.4, we can state the result for two and more erasures. As before, we now restrict to the class of fusion frames, already shown to behave optimally under noise and one erasure.

Theorem 13.14 [38] *Assuming the model previously introduced and letting* $(\mathcal{W}_i)_{i=1}^M$ *be a tight equidimensional fusion frame, the following conditions are equivalent.*

(i) *The MSE due to the erasure of two subspaces is minimized.*
(ii) *The chordal distance between each pair of subspaces is the same and maximal; i.e.,* $(\mathcal{W}_i)_{i=1}^M$ *is a* maximal equidistance *fusion frame.*

Finally, let $(\mathcal{W}_i)_{i=1}^M$ *be an equidimensional, maximal equidistance tight fusion frame. Then the MSE due to k subspace erasures,* $3 \le k < N$, *is constant.*

As we mentioned in the introduction, we will end this subsection with a brief remark on the relation of the previously discovered optimal family of fusion frames with Grassmannian packings. For this, we first state the following problem, which is typically referred to as the classical packing problem (see also [27]).

Classical Packing Problem: For given m, M, N, find a set of m-dimensional subspaces $(\mathcal{W}_i)_{i=1}^M$ in $\mathcal{H}^N$ such that $\min_{i\neq j} d_c(i, j)$ is as large as possible. In this case we call $(\mathcal{W}_i)_{i=1}^M$ an *optimal packing*.

A lower bound is given by the *simplex bound*

$$\frac{m(N-m)M}{N(M-1)}.$$

Theorem 13.15 [27] *Each packing of m-dimensional subspaces* $(\mathcal{W}_i)_{i=1}^M$ *in* $\mathcal{H}^N$ *satisfies*

$$d_c^2(i,j) \leq \frac{m(N-m)}{N}\frac{M}{M-1}, \quad i,j=1,\ldots,M.$$

Interestingly, there is a close connection between tight fusion frames and optimal packings given by the following theorem.

Theorem 13.16 [38] *Let* $(\mathcal{W}_i)_{i=1}^M$ *be a fusion frame of equidimensional subspaces with pairwise equal chordal distances* d_c*. Then, the fusion frame is tight if and only if* d_c^2 *equals the simplex bound.*

This shows that equidistance tight fusion frames are optimal Grassmannian packings.

13.5.3 Perturbations

Perturbations are another common disturbance with respect to which one might seek resilience of a fusion frame. Several scenarios of perturbations of the subspaces can be envisioned. In [22], the following canonical Paley-Wiener-type definition was employed.

Definition 13.8 Let $(\mathcal{W}_i)_{i=1}^M$ and $(\mathcal{V}_i)_{i=1}^M$ be subspaces of $\mathcal{H}^N$ with associated orthogonal projections denoted by $(P_i)_{i=1}^M$ and $(Q_i)_{i=1}^M$, respectively. Further, let $(w_i)_{i=1}^M$ be positive weights, $0 \leq \lambda_1, \lambda_2 < 1$, and $\epsilon > 0$. If, for all $x \in \mathcal{H}^N$ and $1 \leq i \leq M$, we have

$$\big\|(P_i - Q_i)(x)\big\| \leq \lambda_1 \big\|P_i(x)\big\| + \lambda_2 \big\|Q_i(x)\big\| + \epsilon\|x\|,$$

then $((\mathcal{V}_i, w_i))_{i=1}^M$ is called a $(\lambda_1, \lambda_2, \epsilon)$-*perturbation* of $((\mathcal{W}_i, w_i))_{i=1}^M$.

Employing this definition, we obtain the following result on robustness of fusion frames under small perturbations of the associated subspaces. We wish to mention that a perturbation result using a different definition of perturbation can be derived by restricting [45, Theorem 3.1] to fusion frames, however without weights.

Proposition 13.6 [22] *Let* $((\mathcal{W}_i, w_i))_{i=1}^M$ *be a fusion frame for* $\mathcal{H}^N$ *with fusion frame bounds A and B. Further, let* $\lambda_1 \in [0,1)$ *and* $\epsilon > 0$ *be such that*

$$(1-\lambda_1)\sqrt{A} - \epsilon\left(\sum_{i=1}^M w_i^2\right)^{1/2} > 0.$$

Moreover, let $((\mathcal{V}_i, w_i))_{i=1}^M$ *be a* $(\lambda_1, \lambda_2, \epsilon)$*-perturbation of* $((\mathcal{W}_i, w_i))_{i=1}^M$ *for some* $\lambda_2 \in [0, 1)$*. Then* $((\mathcal{V}_i, w_i))_{i=1}^M$ *is a fusion frame with fusion frame bounds*

$$\left(\frac{(1-\lambda_1)\sqrt{A} - \epsilon(\sum_{i=1}^M w_i^2)^{1/2}}{1+\lambda_2}\right)^2 \quad \text{and} \quad \left(\frac{\sqrt{B}(1+\lambda_1) + \epsilon(\sum_{i=1}^M w_i^2)^{1/2}}{1-\lambda_2}\right)^2.$$

For the proof, we refer to [22].

An even more delicate problem is the perturbation of local frame vectors if we consider the full sensor network problem. The difficulty in this case is the possibility of frame vectors leaving the subspace and hence even changing the dimension of those subspaces. A collection of results in this direction can also be found in [22].

13.6 Fusion Frames and Sparsity

In this section we present two different types of results concerning sparsity properties of fusion frames. The first result concerns the construction of tight fusion frames consisting of optimally sparse vectors for efficient processing [19, 20], and the second analyzes the sparse recovery from underdetermined fusion frame measurements [9]. At this point we also refer to Chap. 9 for the theory of sparse recovery and compressed sensing.

13.6.1 Optimally Sparse Fusion Frames

Typically, data processing applications face low on-board computing power and/or a small bandwidth budget. When the signal dimension is large, the decomposition of the signal into its fusion frame measurements requires a large number of additions and multiplications, which may be infeasible for on-board data processing. Thus it would be a significant improvement if the vectors of each orthonormal basis for the subspaces would contain very few nonzero entries, i.e., if they were sparse in the standard unit vector basis, thereby ensuring low-complexity processing. In [19, 20], an algorithmic construction of optimally sparse tight fusion frames with prescribed fusion frame operators was derived, which we will present and discuss in this subsection.

13.6.1.1 Sparseness measure

As already elaborated, we aim for sparsity of orthonormal bases for the subspaces with respect to the standard unit vector basis, which ensures low-complexity processing. Since we are interested in the performance of the whole fusion frame, the

total number of nonzero entries seems to be a suitable sparsity measure. This viewpoint can also be slightly generalized by assuming that there exists a unitary transformation mapping the fusion frame into one having this "sparsity" property. Taking these considerations into account, we state the following definition for a sparse fusion frame, which then reduces to the notion of a sparse frame.

Definition 13.9 Let $(\mathcal{W}_i)_{i=1}^M$ be a fusion frame for $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m_i$ for all $i = 1, \ldots, M$ and let $(e_j)_{j=1}^N$ be an orthonormal basis for $\mathcal{H}^N$. If for each $i \in \{1, \ldots, M\}$ there exists an orthonormal basis $(\varphi_{i,\ell})_{\ell=1}^{m_i}$ for $\mathcal{W}_i$ with the property that for each $\ell = 1, \ldots, m_i$ there is a subset $J_{i,\ell} \subset \{1, \ldots, N\}$ such that

$$\varphi_{i,\ell} \in \operatorname{span}\{e_j : j \in J_{i,\ell}\} \quad \text{and} \quad \sum_{i=1}^{M} \sum_{\ell=1}^{m_i} |J_{i,\ell}| = k,$$

we refer to $(\varphi_{i,\ell})_{i=1,\ell=1}^{M,m_i}$ as an *associated k-sparse frame*. The fusion frame $(\mathcal{W}_i)_{i=1}^M$ is called *k-sparse* with respect to $(e_j)_{j=1}^N$, if it has an associated k-sparse frame and if, for any associated j-sparse frame, we have $k \leq j$.

13.6.1.2 Optimality and maximally achievable sparsity

We now have the necessary machinery at hand to introduce a notion of an *optimally* sparse fusion frame. Optimality will typically be considered within a particular class of fusion frames, e.g., in the class of tight ones.

Definition 13.10 Let $\mathcal{FF}$ be a class of fusion frames for $\mathcal{H}^N$, let $(\mathcal{W}_i)_{i=1}^M \in \mathcal{FF}$, and let $(e_j)_{j=1}^N$ be an orthonormal basis for $\mathcal{H}^N$. Then $(\mathcal{W}_i)_{i=1}^M$ is called *optimally sparse in $\mathcal{FF}$ with respect to* $(e_j)_{j=1}^N$, if $(\mathcal{W}_i)_{i=1}^M$ is k_1-sparse with respect to $(e_j)_{j=1}^N$ and there does not exist a fusion frame $(\mathcal{V}_i)_{i=1}^K \in \mathcal{FF}$ which is k_2-sparse with respect to $(e_j)_{j=1}^N$ with $k_2 < k_1$.

Let N, M, m be positive integers. Then the class of tight fusion frames $(\mathcal{W}_i)_{i=1}^M$ in $\mathcal{H}^N$ with $\dim \mathcal{W}_i = m$ for all $i = 1, \ldots, M$ will be denoted by $\mathcal{FF}(M, m, N)$.

In the case $\frac{Mm}{N} \geq 2$ and $\lfloor \frac{Mm}{N} \rfloor \leq M - 3$ we know that $\mathcal{FF}(M, m, N)$ is not empty; moreover, we can construct a tight fusion frame in this class using the algorithm STFF introduced in Fig. 13.3 (see [11]). STFF can be used to construct fusion frames of equal dimensional subspaces with certain prescribed eigenvalues for the fusion frame operator. We want to use STFF to construct tight fusion frames; i.e., we apply STFF for the constant sequence of eigenvalues $\lambda_j = \frac{Mm}{N}$ for all $j = 1, \ldots, N$, and will refer to the constructed fusion frame as $\mathrm{STFF}(M, m, N)$. The following result shows that $\mathrm{STFF}(M, m, N)$ is optimally sparse in the class $\mathcal{FF}(M, m, N)$. It is a consequence of [19, Theorem 4.4], the analogous result for frames.

STFF (SPECTRAL TETRIS FOR FUSION FRAMES)

Parameters:

- Dimension: N.
- Number of subspaces: M.
- Dimension of subspaces: m.
- Eigenvalues: $(\lambda_j)_{j=1}^N \subseteq [2,\infty)$ satisfying $\sum_{j=1}^N \lambda_j = Mm$ and $\lfloor \lambda_j \rfloor \leq M-3$ for all $j = 1,\ldots,N$.

Algorithm:

1) Set $k := 1$.
2) For $j = 1,\ldots,N$ do
3) Repeat
4) If $\lambda_j < 2$ and $\lambda_j \neq 1$ then
5) $\varphi_k := \sqrt{\frac{\lambda_j}{2}} \cdot e_j + \sqrt{1-\frac{\lambda_j}{2}} \cdot e_{j+1}$.
6) $\varphi_{k+1} := \sqrt{\frac{\lambda_j}{2}} \cdot e_j - \sqrt{1-\frac{\lambda_j}{2}} \cdot e_{j+1}$.
7) $k := k+2$.
8) $\lambda_j := 0$.
9) $\lambda_{j+1} := \lambda_{j+1} - (2-\lambda_j)$.
10) else
11) $\varphi_k := e_j$.
12) $k := k+1$.
13) $\lambda_j := \lambda_j - 1$.
14) end;
15) until $\lambda_j = 0$.
16) end;

Output:

- Fusion frame $(\mathcal{W}_i)_{i=1}^M$ with $\mathcal{W}_i := \text{span}\{\varphi_{i+kM} : k = 0,\ldots,m-1\}$.

Fig. 13.3 The STFF algorithm for constructing a fusion frame

Theorem 13.17 [20] *Let N, M, and m be positive integers such that $\frac{Mm}{N} \geq 2$ and $\lfloor \frac{Mm}{N} \rfloor \leq M-3$. Then the tight fusion frame $STFF(M,m,N)$ is optimally sparse in the class $\mathcal{FF}(M,m,N)$ with respect to the standard unit vector basis.*

In particular, this tight fusion frame is $mM + 2(N - \gcd(Mm,N))$-sparse with respect to the standard unit vector basis.

13.6.2 Compressed Sensing and Fusion Frames

One possible application of fusion frames is music segmentation, in which each note is not characterized by a single frequency but by the subspace spanned by the fundamental frequency of the instrument and its harmonics. Depending on the type of instrument, certain harmonics might or might not be present in the subspace. The overlapping subspaces from distinct instruments can be appropriately modeled by

fusion frames. A canonical question is whether by receiving linear combinations of a collection of signals, each being in one of the subspaces, these signals can be extracted; preferably from as few linear combinations—the measurements—as possible.

This leads to the fundamental question of sparse recovery from fusion frame measurements, which can also be interpreted as structured sparse measurements. In this subsection, we will discuss the answer to this question given in [9], in which sparse recovery results in terms of coherence and restricted isometry property (RIP)-type conditions as well as an average case analysis are provided. In this subsection, due to lack of space, we only focus on the first two.

13.6.2.1 Sparse recovery from underdetermined fusion frame measurements

The just-described scenario can be modeled in the following way. Let $(\mathcal{W}_i)_{i=1}^M$ be a fusion frame for $\mathcal{H}^N$, and let

$$x^0 = \left(x_i^0\right)_{i=1}^M \in \mathcal{H} := \left\{(x_i)_{i=1}^M : x_i \in \mathcal{W}_i \text{ for all } i = 1, \dots, M\right\} \subseteq \mathbb{R}^{MN}.$$

Now assume that we only observe n linear combinations of those vectors; i.e., there exist some scalars a_{ji} satisfying $\|(a_{ji})_{j=1}^n\|_2 = 1$ for all $i = 1, \dots, M$ such that we observe

$$y = (y_j)_{j=1}^n = \left(\sum_{i=1}^M a_{ji} x_i^0\right)_{j=1}^n.$$

We first notice that this equation can be rewritten as

$$y = A_I x^0, \quad \text{where } A_I = (a_{ji} Id_N)_{1 \le j \le n,\, 1 \le i \le M},$$

i.e., A_I is the matrix consisting of the block $a_{ij} Id_N$.

We now aim to recover x^0 from those measurements. Since typically only a few subspaces will contain a signal, it is instructive to impose sparsity conditions as follows; we encourage the reader to compare this with the classical definition of sparsity in Chap. 9.

Definition 13.11 Let $x \in \mathcal{H}$. Then x is called *k-sparse*, if

$$\|x\|_0 := \sum_{i=1}^M \|x_i\|_0 \le k.$$

The initial minimization problem to consider would hence be

$$\hat{x} = \operatorname{argmin}_{x \in \mathcal{H}} \|x\|_0 \quad \text{subject to } A_I x = y.$$

From the theory of compressed sensing, we know that this minimization is NP-hard. A means to circumvent this problem is to consider the associated ℓ_1 minimization problem. In this case, the suitable ℓ_1 norm on $\mathcal{H}$ is a mixed $\ell_{2,1}$ norm defined by

$$\left\|(x_i)_{i=1}^M\right\|_{2,1} := \sum_{i=1}^M \|x_i\|_2 \quad \text{for any } (x_i)_{i=1}^M \in \mathcal{H}.$$

This leads to the investigation of the following minimization problem:

$$\hat{x} = \text{argmin}_{x\in\mathcal{H}} \|x\|_{2,1} \quad \text{subject to } A_I x = y.$$

Taking the special structure of $x \in \mathcal{H}$ into account, we can rewrite this minimization problem as

$$\hat{x} = \text{argmin}_{x\in\mathcal{H}} \|x\|_{2,1} \quad \text{subject to } A_P x = y,$$

where

$$A_P = (a_{ji} P_i)_{1\le i\le M,\, 1\le j\le n}. \tag{13.3}$$

This problem is still difficult to implement, since minimization runs over $\mathcal{H}$. To come to the final utilizable form, let $m_i = \dim \mathcal{W}_i$ and U_i be an $N \times m_i$-matrix whose columns form an orthonormal basis of $\mathcal{W}_i$. This leads to the following two problems—one being equivalent to the previous ℓ_0 minimization problem, the other being equivalent to the just stated ℓ_1 minimization problem—which now merely use matrix-only notation:

$$(P_0) \quad \hat{c} = \text{argmin}_c \|c\|_0 \text{ subject to } Y = AU(c)$$

and

$$(P_1) \quad \hat{c} = \text{argmin}_c \|c\|_{2,1} \text{ subject to } Y = AU(c),$$

in which $A = (a_{ij}) \in \mathbb{R}^{n\times M}$, $j \in \mathbb{R}^{m_j}$, and $y_i \in \mathbb{R}^N$, and

$$U(c) = \begin{pmatrix} \frac{c_1^T U_1^T}{\vdots} \\ \hline c_M^T U_M^T \end{pmatrix} \in \mathbb{R}^{M\times N}, \qquad Y = \begin{pmatrix} \frac{y_1^T}{\vdots} \\ \hline y_n^T \end{pmatrix} \in \mathbb{R}^{n\times N}.$$

13.6.2.2 Coherence results

A typically exploited measure for the coherence of the measurement matrix is its mutual coherence. In [9], the following notion adapted to fusion frame measurements was introduced.

Definition 13.12 The *fusion coherence* of a matrix $A \in \mathbb{R}^{n\times M}$ with normalized columns $(a_i = a_{\cdot,i})_{i=1}^M$ and a fusion frame $(\mathcal{W}_i)_{i=1}^M$ for $\mathbb{R}^N$ is given by

$$\mu_f\big(A, (\mathcal{W}_i)_{i=1}^M\big) = \max_{j\neq k}\big[\big|\langle a_j, a_k\rangle\big| \cdot \|P_j P_k\|_2\big].$$

The reader should note that $\|P_j P_k\|_2 = |\lambda_{\max}(P_j P_k)|^{1/2}$ equals the largest absolute value of the cosines of the principal angles between $\mathcal{W}_j$ and $\mathcal{W}_k$.

This new notion now enables us to phrase the first main result about sparse recovery. Its proof follows some of the arguments of the proof of the analogous "frame result" in [29] with increased technical difficulty; therefore, we refer the reader to the original paper [9].

Theorem 13.18 [9] *Let $A \in \mathbb{R}^{n\times M}$ have normalized columns $(a_i)_{i=1}^M$, let $(\mathcal{W}_i)_{i=1}^M$ be a fusion frame in $\mathbb{R}^N$, and let $Y \in \mathbb{R}^{n\times N}$. If there exists a solution c^0 of the system $Y = AU(c)$ satisfying*

$$\|c^0\|_0 < \frac{1}{2}\big(1 + \mu_f\big(A, (\mathcal{W}_i)_{i=1}^M\big)^{-1}\big),$$

then this solution is the unique solution of (P_0) as well as of (P_1).

This result generalizes the classical sparse recovery result from [29] by letting $N = 1$, since in this case $P_i = 1$ for all $i = 1, \ldots, M$.

13.6.2.3 RIP results

The RIP property, which complements the mutual coherence conditions, was also adapted to the fusion frame setting in [9] in the following way.

Definition 13.13 Let $A \in \mathbb{R}^{n\times M}$ and $(\mathcal{W}_i)_{i=1}^M$ be a fusion frame for $\mathcal{H}^N$. Then the *fusion restricted isometry constant* δ_k is the smallest constant such that

$$(1-\delta_k)\|z\|_2^2 \leq \|A_P z\|_2^2 \leq (1+\delta_k)\|z\|_2^2$$

for all $z \in \mathbb{R}^{NM}$ of sparsity $\|z\|_0 \leq k$, where A_P is defined as in (13.3).

The definition of the *restricted isometry constant* in [13] is a special case of Definition 13.13 with $N = 1$ and $\dim \mathcal{W}_i = 1$ for $i = 1, \ldots, M$. Again, we refer to [9] for the proof of the following theorem.

Theorem 13.19 [9] *Let $(A, (\mathcal{W}_i)_{i=1}^M)$ have the fusion frame restricted isometry constant $\delta_{2k} < 1/3$. Then (P_1) recovers all k-sparse c from $Y = AU(c)$.*

13.7 Nonorthogonal Fusion Frames

Until recently, fusion frame theory has mainly focused on the construction of fusion frames with specified properties. However, in practice, we might not have the freedom to choose the "best fusion frame," since it is often given by the application. One example is the application to modeling of sensor networks (cf. Sect. 13.1.3), in which each sensor spans a fixed subspace $\mathcal{W}$ of $\mathcal{H}^N$ generated by the spatial reversal and the translates of the sensor's impulse response function [40].

Although in such applications selection or manipulation of the subspaces is not possible, sometimes there is the freedom to choose the measuring procedure, i.e., the operators mapping the signal onto each element from the family of subspaces. Let us consider again the example of distributed sensing. At the first stage, each sensor in a particular area measures the scalar $\langle x, \varphi_i \rangle$ of an incoming signal $x \in \mathcal{H}^N$, where $\varphi_i \in \mathcal{H}^N$ depend on the characteristics of the respective sensor for all $i \in I$, say. Now, assume that $\mathcal{W} = \text{span}\{\varphi_i : i \in I\}$. Instead of combining the scalars $\langle x, \varphi_i \rangle$ to obtain the orthogonal projection of x onto $\mathcal{W}$, also $P(x)$, where P is a nonorthogonal projection onto $\mathcal{W}$, could be computed. In such cases, one objective is sparsity of the fusion frame operator, which ensures, despite the fact that tightness might not be achievable, an efficient reconstruction algorithm. It would be particularly desirable if the fusion frame operator were a multiple of the identity or at least a diagonal operator.

Another problem is the limited availability of tight fusion frames. The effectiveness of fusion frame applications in distributed systems is heavily dependent on the end fusion process. This in turn depends upon the efficiency of the inversion of the fusion frame operator. Tight fusion frames take care of this problem because the frame operator is a multiple of the identity and hence its inverse operator is also a multiple of the identity. But tight fusion frames do not exist in most situations. The idea here is to use nonorthogonal projections which will result in much larger classes of fusion frames with the (nonorthogonal) fusion frame operator equal to a multiple of the identity.

To tackle these problems, the theory of nonorthogonal fusion frames was recently introduced in [10]. The main idea is to replace the orthogonal projections in the definition of a fusion frame with general projections, i.e., with linear operators Q from $\mathcal{H}^N$ onto a subspace $\mathcal{W}$ of $\mathcal{H}^N$ which satisfy $Q = Q^2$. Recall that in this case, the adjoint Q^* is also a non-orthogonal projection onto $\mathcal{N}(Q)^\perp$ with $\mathcal{N}(Q) \oplus \mathcal{W} = \mathcal{H}^N$, where $\mathcal{N}(Q) = \{x \in \mathcal{H}^N : Qx = 0\}$. This yields the following definition, which generalizes the classical notion of a fusion frame.

Definition 13.14 Let $(\mathcal{W}_i)_{i=1}^M$ be a family of subspaces in $\mathcal{H}^N$, and let $(w_i)_{i=1}^M \subseteq \mathbb{R}^+$ be a family of weights. For each $i = 1, 2, \ldots, M$ let Q_i be a (orthogonal or nonorthogonal) projection onto $\mathcal{W}_i$. Then $((Q_i, w_i))_{i=1}^M$ is a *nonorthogonal fusion frame* for $\mathcal{H}^N$, if there exist constants $0 < A \leq B < \infty$ such that

$$A\|x\|_2^2 \leq \sum_{i=1}^{M} w_i^2 \left\| Q_i(x) \right\|_2^2 \leq B\|x\|_2^2 \quad \text{for all } x \in \mathcal{H}^N.$$

The constants A and B are called the *lower* and *upper fusion frame bound*, respectively.

Letting $\mathcal{W} = ((Q_i, w_i))_{i=1}^M$ be a nonorthogonal fusion frame for $\mathcal{H}^N$, the associated *analysis operator* $T_{\mathcal{W}}$ is defined by

$$T_{\mathcal{W}} : \mathcal{H}^N \to \mathbb{R}^{MN}, \qquad x \mapsto \big(w_i Q_i(x)\big)_{i=1}^M,$$

and the *synthesis operator* $T_{\mathcal{W}}^*$ has the form

$$T_{\mathcal{W}}^* : \mathbb{R}^{MN} \to \mathbb{R}^N, \qquad (y_i)_{i=1}^M \mapsto \sum_{i=1}^M w_i Q_i^*(y_i).$$

The *nonorthogonal fusion frame operator* $S_{\mathcal{W}}$ is then given by

$$S_{\mathcal{W}} = T_{\mathcal{W}}^* T_{\mathcal{W}} : \mathcal{H}^N \to \mathcal{H}^N, \qquad x \mapsto \sum_{i=1}^M w_i^2 Q_i^* Q_i(x).$$

Similar to Theorem 13.2, we have the following result.

Theorem 13.20 [10] *Let $\mathcal{W} = ((Q_i, w_i))_{i=1}^M$ be a nonorthogonal fusion frame for $\mathcal{H}^N$ with fusion frame bounds A and B and associated nonorthogonal fusion frame operator $S_{\mathcal{W}}$. Then $S_{\mathcal{W}}$ is a positive, self-adjoint, invertible operator on $\mathcal{H}^N$ with $A\,Id \le S_{\mathcal{W}} \le B\,Id$. Moreover, we have the reconstruction formula*

$$x = \sum_{i=1}^M w_i^2 S_{\mathcal{W}}^{-1}\big(Q_i^* Q_i(x)\big) \quad \text{for all } x \in \mathcal{H}^N.$$

We now focus on the second problem, when we have the freedom to choose the subspaces as well as the projections. Surprisingly, this additional freedom enables the construction of tight (nonorthogonal) fusion frames in almost all situations as the next result shows.

Theorem 13.21 [10] *Let $m_i \le \frac{N}{2}$ for all $i = 1, 2, \ldots, M$ satisfy $\sum_{i=1}^M m_i \ge N$. Then there exists a tight nonorthogonal fusion frame $((Q_i, w_i))_{i=1}^M$ for $\mathbb{R}^N$ such that $\text{rank}(Q_i) = m_i$ for all $i = 1, \ldots, M$.*

This result shows that if the dimensions of the subspaces are less than or equal to half the dimension of the ambient space, there always exists a tight nonorthogonal fusion frame. The proof in fact shows that the weights can even be chosen to be equal to 1. Thus, nonorthogonality allows a much larger class of tight fusion frames.

To prove this result, we first require a particular classification of positive, self-adjoint operators by projections. In order to build up some intuition, let $T : \mathbb{R}^N \to$

$\mathbb{R}^N$ be a positive, self-adjoint operator. The goal is to classify the set

$$\Omega(T) = \{Q : Q^2 = Q,\ Q^*Q = T\}.$$

We first observe that, by the spectral theorem, T can be written as

$$T = \sum_{i=1}^{M} \lambda_i P_i,$$

where the λ_i is the ith eigenvalue of T and P_i is the orthogonal projection onto the space generated by the ith eigenvector of T. Hence $Q \in \Omega(T)$ if and only if the eigenvalues and eigenvectors of Q^*Q coincide with those of T. Noting that $Q \in \Omega(T)$ implies $\ker(Q) = \mathrm{im}(T)^{\perp}$ and recalling that a projection is uniquely determined by its kernel and its image, it suffices to consider the set

$$\tilde{\Omega}(T) = \{\mathrm{im}(Q) : Q \in \Omega(T)\}.$$

Moreover, observe that since the only projection of rank N is the identity, we may assume $\mathrm{rank}(T) < N$.

The next result states the classification of $\tilde{\Omega}(T)$ (and hence $\Omega(T)$) which we require for the proof of Theorem 13.21. Although the proof is fairly elementary, we refer the reader to the complete argument in [10].

Theorem 13.22 *Let $T : \mathbb{R}^n \to \mathbb{R}^n$ be a positive, self-adjoint operator of rank $k \leq \frac{N}{2}$. Let $(\lambda_j)_{j=1}^k$ be the nonzero eigenvalues of T and suppose that $\lambda_j \geq 1$ for $j = 1, \ldots, k$ and $(e_j)_{j=1}^k$ is an orthonormal basis of* $\mathrm{im}(T)$ *consisting of eigenvectors of T associated to the eigenvalues $(\lambda_j)_{j=1}^k$. Then*

$$\tilde{\Omega}(T) = \left\{ span\left\{ \frac{1}{\sqrt{\lambda_j}} e_j + \sqrt{\frac{\lambda_j - 1}{\lambda_j}} e_{j+k} \right\}_{j=1}^{k} : (e_j)_{j=1}^{2k} \text{ is orthonormal} \right\}.$$

Let $T : \mathbb{R}^N \to \mathbb{R}^N$ be a positive, self-adjoint operator. Applying Theorem 13.22 to $\frac{1}{\lambda_k} T$, where λ_k is the smallest nonzero eigenvalue of T and setting $v = \sqrt{\lambda_k}$, yields the following corollary.

Corollary 13.5 *Let $T : \mathbb{R}^N \to \mathbb{R}^N$ be a positive, self-adjoint operator of rank $\leq \frac{N}{2}$. Then there exists a projection Q and a weight v so that $T = v^2 Q^* Q$.*

Having these prerequisites, we can now prove Theorem 13.21.

Proof of Theorem 13.21 Let $(e_j)_{j=1}^N$ be an orthonormal basis of $\mathbb{R}^N$, and let $(\mathcal{W}_i)_{i=1}^M$ be a family of subspaces of $\mathcal{H}^N$ such that

- $\mathcal{W}_i = \mathrm{span}\{e_j\}_{j \in J_i}$ with $|J_i| = m_i$ for each $i = 1, \ldots, M$,

- $\mathcal{W}_1 + \cdots + \mathcal{W}_M = \mathcal{H}^N$.

Also, let P_i denote the orthogonal projection onto $\mathcal{W}_i$, and set $S = \sum_{i=1}^{M} P_i$.

Notice that

$$Id = S^{-1}S = \sum_{i=1}^{M} S^{-1}P_i.$$

Since each projection P_i is diagonal with respect to $(e_j)_{j=1}^{N}$, the operator S^{-1} commutes with P_i for each $i = 1, \ldots, M$. Hence, for all $i = 1, \ldots, M$, $S^{-1}P_i$ is positive and self-adjoint. Now, letting γ denote the smallest nonzero eigenvalue of all $S^{-1}P_i$, $i = 1, \ldots, M$, the operator $\frac{1}{\gamma}S^{-1}P_i$ satisfies the hypotheses of Theorem 13.22. Thus, there exists a projection Q_i so that

$$Q_i^* Q_i = \frac{1}{\gamma} S^{-1} P_i,$$

leading to

$$\sum_{i=1}^{M} Q_i^* Q_i = \frac{1}{\gamma} Id.$$

The theorem is proved. □

If we are willing to extend the framework even further and allow *two* projections onto each subspace, it can be shown that Parseval nonorthogonal fusion frames can be constructed for any sequence of dimensions of the subspaces [10].

Acknowledgements The first author is supported by NSF DMS 1008183, NSF ATD 1042701, and AFOSR FA9550-11-1-0245. The second author acknowledges support by the Einstein Foundation Berlin, by Deutsche Forschungsgemeinschaft (DFG) Grant SPP-1324 KU 1446/13 and DFG Grant KU 1446/14, and by the DFG Research Center MATHEON "Mathematics for key technologies" in Berlin. The authors are indebted to Andreas Heinecke for his careful reading of this chapter and various useful comments and suggestions.

References

1. Balan, R., Bodmann, B.G., Casazza, P.G., Edidin, D.: Painless reconstruction from magnitudes of frame coefficients. J. Fourier Anal. Appl. **15**(4), 488–501 (2009)
2. Benedetto, J.J., Fickus, M.: Finite normalized tight frames. Adv. Comput. Math. **18**(2–4), 357–385 (2003)
3. Bjørstad, P.J., Mandel, J.: On the spectra of sums of orthogonal projections with applications to parallel computing. BIT **1**, 76–88 (1991)
4. Bodmann, B.G.: Optimal linear transmission by loss-insensitive packet encoding. Appl. Comput. Harmon. Anal. **22**(3), 274–285 (2007)
5. Bodmann, B.G., Casazza, P.G., Kutyniok, G.: A quantitative notion of redundancy for finite frames. Appl. Comput. Harmon. Anal. **30**, 348–362 (2011)

6. Bodmann, B.G., Casazza, P.G., Paulsen, V.I., Speegle, D.: Spanning and independence properties of frame partitions. Proc. Am. Math. Soc. **40**(7), 2193–2207 (2012)
7. Bodmann, B.G., Casazza, P.G., Peterson, J., Smalyanu, I., Tremain, J.C.: Equi-isoclinic fusion frames and mutually unbiased basic sequences, preprint
8. Bölcskei, H., Hlawatsch, F., Feichtinger, H.G.: Frame-theoretic analysis of oversampled filter banks. IEEE Trans. Signal Process. **46**, 3256–3269 (1998)
9. Boufounos, B., Kutyniok, G., Rauhut, H.: Sparse recovery from combined fusion frame measurements. IEEE Trans. Inf. Theory **57**, 3864–3876 (2011)
10. Cahill, J., Casazza, P.G., Li, S.: Non-orthogonal fusion frames and the sparsity of fusion frame operators, preprint
11. Calderbank, R., Casazza, P.G., Heinecke, A., Kutyniok, G., Pezeshki, A.: Sparse fusion frames: existence and construction. Adv. Comput. Math. **35**(1), 1–31 (2011)
12. Calderbank, A.R., Hardin, R.H., Rains, E.M., Shore, P.W., Sloane, N.J.A.: A group-theoretic framework for the construction of packings in Grassmannian spaces. J. Algebr. Comb. **9**(2), 129–140 (1999)
13. Candés, E.J., Romberg, J.K., Tao, T.: Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
14. Casazza, P.G., Fickus, M.: Minimizing fusion frame potential. Acta Appl. Math. **107**(103), 7–24 (2009)
15. Casazza, P.G., Fickus, M., Heinecke, A., Wang, Y., Zhou, Z.: Spectral tetris fusion frame constructions, preprint
16. Casazza, P.G., Fickus, M., Kovačević, J., Leon, M., Tremain, J.C.: A physical interpretation of tight frames. In: Heil, C. (ed.) Harmonic Analysis and Applications, pp. 51–76. Birkhäuser, Boston (2006)
17. Casazza, P.G., Fickus, M., Mixon, D., Wang, Y., Zhou, Z.: Constructing tight fusion frames. Appl. Comput. Harmon. Anal. **30**(2), 175–187 (2011)
18. Casazza, P.G., Fickus, M., Mixon, D., Peterson, J., Smalyanau, I.: Every Hilbert space frame has a Naimark complement, preprint
19. Casazza, P.G., Heinecke, A., Krahmer, F., Kutyniok, G.: Optimally sparse frames. IEEE Trans. Inf. Theory **57**, 7279–7287 (2011)
20. Casazza, P.G., Heinecke, A., Kutyniok, G.: Optimally sparse fusion frames: existence and construction. In: Proc. SampTA'11 (Singapore, 2011)
21. Casazza, P.G., Kutyniok, G.: Frames of subspaces. In: Wavelets, Frames and Operator Theory, College Park, MD, 2003. Contemp. Math., vol. 345, pp. 87–113. Am. Math. Soc., Providence (2004)
22. Casazza, P.G., Kutyniok, G., Li, S.: Fusion frames and distributed processing. Appl. Comput. Harmon. Anal. **25**, 114–132 (2008)
23. Casazza, P.G., Kutyniok, G., Li, S., Rozell, C.J.: Modeling sensor networks with fusion frames. In: Wavelets XII, San Diego, 2007. SPIE Proc., vol. 6701, pp. 67011M-1–67011M-11. SPIE, Bellingham (2007)
24. Casazza, P.G., Leon, M.: Existence and construction of finite frames with a given frame operator. Int. J. Pure Appl. Math. **63**(2), 149–158 (2010)
25. Casazza, P.G., Tremain, J.C.: The Kadison-Singer problem in mathematics and engineering. Proc. Natl. Acad. Sci. **103**, 2032–2039 (2006)
26. Chebira, A., Fickus, M., Mixon, D.G.: Filter bank fusion frames. IEEE Trans. Signal Process. **59**, 953–963 (2011)
27. Conway, J.H., Hardin, R.H., Sloane, N.J.A.: Packing lines, planes, etc.: packings in Grassmannian spaces. Exp. Math. **5**(2), 139–159 (1996)
28. Cvetković, Z., Vetterli, M.: Oversampled filter banks. IEEE Trans. Signal Process. **46**, 1245–1255 (1998)
29. Donoho, D.L., Elad, M.: Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. Proc. Natl. Acad. Sci. USA **100**(5), 2197–2202 (2003)
30. Et-Taoui, B., Fruchard, A.: Equi-isoclinic subspaces of Euclidean space. Adv. Geom. **9**(4), 471–515 (2009)

31. Et-Taoui, B.: Equi-isoclinic planes in Euclidean even dimensional spaces. Adv. Geom. **7**(3), 379–384 (2007)
32. Et-Taoui, B.: Equi-isoclinic planes of Euclidean spaces. Indag. Math. (N. S.) **17**(2), 205–219 (2006)
33. Fornasier, M.: Quasi-orthogonal decompositions of structured frames. J. Math. Anal. Appl. **289**, 180–199 (2004)
34. Fulton, W.: Young Tableaux. With Applications to Representation Theory and Geometry. London Math. Society Student Texts, vol. 35. Cambridge University Press, Cambridge (1997)
35. Godsil, C.D., Hensel, A.D.: Distance regular covers of the complete graph. J. Comb. Theory, Ser. B **56**(2), 205–238 (1992)
36. Goyal, V., Kovačević, J., Kelner, J.A.: Quantized frame expansions with erasures. Appl. Comput. Harmon. Anal. **10**(3), 203–233 (2001)
37. Hoggar, S.G.: New sets of equi-isoclinic n-planes from old. Proc. Edinb. Math. Soc. **20**(4), 287–291 (1977)
38. Kutyniok, G., Pezeshki, A., Calderbank, A.R., Liu, T.: Robust dimension reduction, fusion frames, and Grassmannian packings. Appl. Comput. Harmon. Anal. **26**(1), 64–76 (2009)
39. Lemmens, P.W.H., Seidel, J.J.: Equi-isoclinic subspaces of Euclidean spaces. Ned. Akad. Wet. Proc. Ser. A 76, Indag. Math. **35**, 98–107 (1973)
40. Li, S., Yan, D.: Frame fundamental sensor modeling and stability of one-sided frame perturbation. Acta Appl. Math. **107**(1–3), 91–103 (2009)
41. Massey, P.G., Ruiz, M.A., Stojanoff, D.: The structure of minimizers of the frame potential on fusion frames. J. Fourier Anal. Appl. (to appear)
42. Oswald, P.: Frames and space splittings in Hilbert spaces. Lecture Notes, Part 1, Bell Labs, Technical Report, pp. 1–32 (1997)
43. Rozell, C.J., Johnson, D.H.: Analyzing the robustness of redundant population codes in sensory and feature extraction systems. Neurocomputing **69**, 1215–1218 (2006)
44. Sun, W.: G-frames and G-Riesz bases. J. Math. Anal. Appl. **322**, 437–452 (2006)
45. Sun, W.: Stability of G-frames. J. Math. Anal. Appl. **326**, 858–868 (2007)
46. Wootters, W.K.: Quantum mechanics without probability amplitudes. Found. Phys. **16**(4), 391–405 (1986)
47. Wootters, W.K., Fields, B.D.: Optimal state-determination by mutually unbiased measurements. Ann. Phys. **191**(2), 363–381 (1989)

Index

P.G. Casazza, G. Kutyniok (eds.), *Finite Frames*,
Applied and Numerical Harmonic Analysis,
DOI 10.1007/978-0-8176-8373-3, © Springer Science+Business Media New York 2013

Applied and Numerical Harmonic Analysis

J.M. Cooper: *Introduction to Partial Differential Equations with MATLAB* (ISBN 978-0-8176-3967-9)

C.E. D'Attellis and E.M. Fernández-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)

T.M. Peters, J.H.T. Bates, G.B. Pike, P. Munger, and J.C. Williams: *The Fourier Transform in Biomedical Engineering* (ISBN 978-0-8176-3941-9)

A.I. Saichev and W.A. Woyczyński: *Distributions in the Physical and Engineering Sciences* (ISBN 978-0-8176-3924-2)

R. Tolimieri and M. An: *Time-Frequency Representations* (ISBN 978-0-8176-3918-1)

G.T. Herman: *Geometry of Digital Spaces* (ISBN 978-0-8176-3897-9)

A. Procházka, J. Uhlíř, P.J.W. Rayner, and N.G. Kingsbury: *Signal Analysis and Prediction* (ISBN 978-0-8176-4042-2)

J. Ramanathan: *Methods of Applied Fourier Analysis* (ISBN 978-0-8176-3963-1)

A. Teolis: *Computational Signal Processing with Wavelets* (ISBN 978-0-8176-3909-9)

W.O. Bray and C.V. Stanojević: *Analysis of Divergence* (ISBN 978-0-8176-4058-3)

G.T Herman and A. Kuba: *Discrete Tomography* (ISBN 978-0-8176-4101-6)

J.J. Benedetto and P.J.S.G. Ferreira: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)

A. Abbate, C.M. DeCusatis, and P.K. Das: *Wavelets and Subbands* (ISBN 978-0-8176-4136-8)

L. Debnath: *Wavelet Transforms and Time-Frequency Signal Analysis* (ISBN 978-0-8176-4104-7)

K. Gröchenig: *Foundations of Time-Frequency Analysis* (ISBN 978-0-8176-4022-4)

D.F. Walnut: *An Introduction to Wavelet Analysis* (ISBN 978-0-8176-3962-4)

O. Bratteli and P. Jorgensen: *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)

J. Davis: *Methods of Applied Mathematics with a MATLAB Overview* (ISBN 978-0-8176-4331-7)

G. Bi and Y. Zeng: *Transforms and Fast Algorithms for Signal Analysis and Representations* (ISBN 978-0-8176-4279-2)

J.J. Benedetto and A. Zayed: *Sampling, Wavelets, and Tomography* (ISBN 978-0-8176-4304-1)

E. Prestini: *The Evolution of Applied Harmonic Analysis* (ISBN 978-0-8176-4125-2)

O. Christensen and K.L. Christensen: *Approximation Theory* (ISBN 978-0-8176-3600-5)

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)

O. Calin and D.-C. Chang: *Geometric Mechanics on Riemannian Manifolds* (ISBN 978-0-8176-4354-6)

Applied and Numerical Harmonic Analysis (Cont'd)

J.A. Hogan and J.D. Lakey: *Time-Frequency and Time-Scale 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, V. Mang I, 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. J. 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, K.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)

B. Luong: *Fourier Analysis on Finite Abelian Groups* (ISBN 978-0-8176-4915-9)

S.G. Krantz: *Explorations in Harmonic Analysis* (ISBN 978-0-8176-4668-4)

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)

P. Massopust and B. Forster: *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. Chang, 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. 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)

J.A. Hogan and J.D. Lakey: *Duration and Bandwidth Limiting* (ISBN 978-0-8176-8306-1)

G. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 2* (ISBN 978-0-8176-4943-2)

G. Kutyniok and D. Labate: *Shearlets* (ISBN 978-0-8176-8315-3)

P.G. Casazza and G. Kutyniok: *Finite Frames* (ISBN 978-0-8176-8372-6)

For a fully up-to-date list of ANHA titles, visit ***http://www.springer.com/series/4968?detailsPage=titles*** *or* ***http://www.springerlink.com/content/t7k8lm/****.*