Applied and Numerical Harmonic Analysis

Götz E. Pfander
Editor

Sampling Theory, a Renaissance

Compressive Sensing and Other Developments

Birkhäuser

Applied and Numerical Harmonic Analysis

Götz E. Pfander
Editor

Sampling Theory, a Renaissance

Compressive Sensing and Other Developments

Birkhäuser

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

ISSN 2296-5009 ISSN 2296-5017 (electronic)
Applied and Numerical Harmonic Analysis
ISBN 978-3-319-19748-7 ISBN 978-3-319-19749-4 (eBook)
DOI 10.1007/978-3-319-19749-4

Library of Congress Control Number: 2015953322

Mathematics Subject Classification (2010): 94A20, 94A12, 42C15, 41A45, 30H20

Springer Cham Heidelberg New York Dordrecht London

Printed on acid-free paper

Springer International Publishing AG Switzerland is part of Springer Science+Business Media (www.springer.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 richness and relevance of applications and implementation depend fundamentally on the structure and depth of theoretical underpinnings. Thus, from our point of view, the interleaving of theory and applications and their creative symbiotic evolution is axiomatic.

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

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

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

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

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

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

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

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

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

Nowadays, some of the theory of PDEs has given way to the study of Fourier integral operators. Problems in antenna theory are studied in terms of unimodular trigonometric polynomials. Applications of Fourier analysis abound in signal processing, whether with the fast Fourier transform (FFT), or filter design, or the

adaptive modeling inherent in time-frequency-scale methods such as wavelet theory. The coherent states of mathematical physics are translated and modulated Fourier transforms, and these are used, in conjunction with the uncertainty principle, for dealing with signal reconstruction in communications theory. We are back to the raison d'être of the *ANHA* series!

College Park, MD, USA John J. Benedetto

Preface

renaissance [...]

> enthusiastic and vigorous activity along literary, artistic, and cultural lines distinguished by a revival of interest in the past, by an increasing pursuit of learning, and by an imaginative response to broader horizons generally [...]
> a return of youthful vigor, freshness, zest, or productivity a renewal of life or interest in some aspect of it [...]
>
> Webster's Third New International Dictionary

Sampling theory has played a central role in mathematics, science, and engineering for over 75 years now. The original quest of identifying a continuous function on Euclidean space from discrete data is addressed in the classical sampling theorem, commonly attributed to Cauchy, Kotelnikov, Ogura, Raabe, Shannon, and/or Whittaker. It states that a bandlimited function can be recovered in full from values measured on a regular sampling grid whenever the bandlimitation is described by an interval whose length does not exceed the density of the sampling grid. A multitude of variants and extensions of this result have cemented the extensive role of sampling theory in engineering and science during the second half of the 20th century.

Today, the original emphasis on recovery from samples is complemented by the need for efficient digital representations of signals and images by various kinds of available, but at first sight insufficient, measurements. In addition, fast and noise resistant algorithms aimed at recovering from such measurements are of increasing importance. The assumption that a signal is bandlimited in the classical setting is commonly replaced by possibly nonlinear constraints on the objects at hand; and the need to efficiently obtain reliable nonredundant representations of such objects may involve a nonlinear measurement procedure as well.

Such and related considerations have lately reenergized the area of sampling theory and inspired the rapid growth of new interdisciplinary research areas such as compressive sensing and phase retrieval.

Compressive sensing is based on the observation that many practical signals like images, speech, music, radar signals, ultrasound signals, and man-made communication signals are well characterized by a relatively small number of relevant parameters when compared to the dimension of the ambient space. That is, we assume that the signal is contained in – or is well approximated by a signal in – the union of low-dimensional subspaces of a high dimensional space; the signal depends on a *sparse* set of parameters and the difficulty lies in realizing which parameters are active and which ones can be ignored. For example, if a high dimensional signal is known to have few nonzero Fourier coefficients of unknown locations, then compressive sensing algorithms exploit this sparsity assumption and recover the signal from samples far below the Nyquist rate.

In compressive sensing, the nonlinearity of the signal space leads to challenging mathematical problems when attempting to prove performance guarantees for realistic recovery algorithms such as Basis Pursuit or Orthogonal Matching Pursuit. State-of-the-art results control the recovery probability of sparse signals when the number of required measurements grows only linearly in the number of nonzero parameters and logarithmically in the ambient dimension.

The second example of a flourishing research area in sampling theory is motivated by X-ray crystallography where, in essence, only magnitudes of Fourier coefficients of an image are measured. In order to reconstruct the image, some additional insights on the image need to be utilized to recover the phase of each Fourier coefficient and thereby the original image. To achieve this in a provably numerically stable manner remains an open problem to date. This being said, the described problem spearheaded the novel research area of phase retrieval. The question addressed herein is the following: in which settings and for what kind of measurements can we design algorithms that recover images or other signals from magnitudes of those measurements?

Compressed sensing and phase retrieval are just two examples that illustrate the influx of new ideas and paradigms in sampling theory; they form the foundation of the sampling theory renaissance that we enjoy today in mathematics, science, and engineering.

The contributed chapters in this volume are authored by invited speakers and session organizers of the *10th International Conference on Sampling Theory and Applications* (SampTA) which took place on July 1st to 5th, 2013, in Bremen, Germany. The authors' contributions are organized into five parts, "Random Measurements of High Dimensional Data," "Finite and Structured Frames," "Band-limitation and Generalizations," "Sampling and Parametric Partial Differential Equations," and "Data Acquisition," thereby representing a good portion of research areas discussed at SampTA 2013.

The success of the conference was made possible through the enthusiasm and commitment of a number of colleagues working in the vast area of sampling theory. Foremost, I would like to thank my colleagues on the local organization team, Peter Oswald, Werner Henkel, Peter Maaß, Peter Massopust, Anja Müller, and Holger Rauhut, as well as my technical program co-chairs Yonina Eldar, Laurent Fesquet, Gitta Kutyniok, Pina Marziliano, and Bruno Torrésani. The support by the

SampTA steering committee, Akram Aldroubi, John Benedetto, Paul Butzer, Hans Feichtinger, Paulo Ferreira, Karlheinz Gröchenig, Rowland Higgins, Abdul Jerri, Yuri Lyubarskii, Farokh Marvasti, Gerhard Schmeißer, Bruno Torrésani, Michael Unser, and Ahmed Zayed, is greatly appreciated.

Preparation of SampTA as well as of this volume was carried out in part during my sabbatical stay at the Mathematics Department and the Research Laboratory of Electronics at the Massachusetts Institute of Technology in Spring 2012, my visit as John von Neumann Visiting Professor at the Technical University Munich in Spring 2014, and my one semester visit to the Catholic University Eichstätt-Ingolstadt in Fall 2014. I would like to thank the three institutions, in particular, my hosts Laurent Demanet, Vivek Goyal, Massimo Fornasier, and Rene Grothmann for their hospitality and the great working conditions that I enjoyed during my stays. Last but not least, I would like to thank my mathematics mentors, Hermann Pfander and John Benedetto, for their continued support.

Jacobs University, Bremen Götz E. Pfander
15.10.2015

Contents

Part I
Random Measurements of High Dimensional Data

Chapter 1
Estimation in High Dimensions: A Geometric Perspective

Roman Vershynin

Abstract This tutorial provides an exposition of a flexible geometric framework for high-dimensional estimation problems with constraints. The tutorial develops geometric intuition about high-dimensional sets, justifies it with some results of asymptotic convex geometry, and demonstrates connections between geometric results and estimation problems. The theory is illustrated with applications to sparse recovery, matrix completion, quantization, linear and logistic regression, and generalized linear models.

1.1 Introduction

1.1.1 Estimation with constraints

This chapter provides an exposition of an emerging mathematical framework for high-dimensional *estimation problems with constraints*. In these problems, the goal is to estimate a point $\boldsymbol{x}$ which lies in a certain known feasible set $K \subseteq \mathbb{R}^n$, from a small sample $y_1, \dots, y_m$ of independent observations of $\boldsymbol{x}$. The point $\boldsymbol{x}$ may represent a signal in signal processing, a parameter of a distribution in statistics, or an unknown matrix in problems of matrix estimation or completion. The feasible set K is supposed to represent properties that we know or want to impose on $\boldsymbol{x}$.

The geometry of the high-dimensional set K is a key to understanding estimation problems. A powerful intuition about *what high-dimensional sets look like* has been developed in the area known as *asymptotic convex geometry* [6, 32]. The intuition is supported by many rigorous results, some of which can be applied to estimation problems. The main goals of this chapter are:

Partially supported by NSF grant DMS 1265782 and USAF Grant FA9550-14-1-0009.

R. Vershynin (✉)
Department of Mathematics, University of Michigan, 530 Church Street,
Ann Arbor, MI 48109, USA
e-mail: romanv@umich.edu

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical
Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_1

(a) develop geometric intuition about high-dimensional sets;
(b) explain results of asymptotic convex geometry which validate this intuition;
(c) demonstrate connections between high-dimensional geometry and high-dimensional estimation problems.

This chapter is not a comprehensive survey but is rather a tutorial. It does not attempt to chart vast territories of high-dimensional inference that lie on the interface of statistics and signal processing. Instead, this chapter proposes a useful geometric viewpoint, which could help us find a common mathematical ground for many (and often dissimilar) estimation problems.

1.1.2 Quick examples

Before we proceed with a general theory, let us mention some concrete examples of estimation problems that will be covered here. A particular class of estimation problems with constraints is considered in the young field of *compressed sensing* [15, 19, 26, 39]. There K is supposed to enforce *sparsity*; thus K usually consists of vectors that have few nonzero coefficients. Sometimes more restrictive *structured sparsity* assumptions are placed, where only certain arrangements of nonzero coefficients are allowed [5, 61]. The observations y_i in compressed sensing are assumed to be *linear* in $\boldsymbol{x}$, which means that $y_i = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$. Here $\boldsymbol{a}_i$ are typically i.i.d. vectors drawn from some known distribution in $\mathbb{R}^n$ (for example, normal).

Another example of estimation problems with constraints is the *matrix completion problem* [12, 13, 34, 37, 63, 68] where K consists of matrices with *low rank*, and $y_1, \dots, y_m$ is a sample of matrix entries. Such observations are still linear in $\boldsymbol{x}$.

In general, observations do not have to be linear; good examples are *binary observations* $y_i \in \{-1, 1\}$, which satisfy $y_i = \operatorname{sign}(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle)$, see [10, 36, 57, 59], and more generally $\mathbb{E}\, y_i = \theta(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle)$, see [2, 58, 60].

In statistics, these classes of estimation problems can be interpreted as *linear regression* (for linear observations with noise), *logistic regression* (for binary observations), and *generalized linear models* (for more general non-linear observations).

All these examples, and more, will be explored in this chapter. However, our main goal is to advance a general approach, which would not be tied to a particular nature of the feasible set K. Some general estimation problems of this nature were considered in [3, 47] for linear observations and in [2, 58–60] for nonlinear observations.

1.1.3 Plan of the chapter

In Section 1.2.1, we introduce a general class of estimation problems with constraints. We explain how the constraints (given by feasible set K) represent *low-complexity structures*, which could make it possible to estimate $\boldsymbol{x}$ from few observations.

In Section 1.3, we make a short excursion into the field of *asymptotic convex geometry*. We explain intuitively the shape of high-dimensional sets K and state some known results supporting this intuition. In view of estimation problems, we especially emphasize one of these results—the so-called M^* *bound* on the size of high-dimensional sections of K by a random subspace E. It depends on the single geometric parameter of K that quantifies the complexity of K; this quantity is called the *mean width*. We discuss mean width in some detail, pointing out its connections to convex geometry, stochastic processes, and statistical learning theory.

In Section 1.4, we apply the M^* bound to the general estimation problem with linear observations. We formulate an estimator first as a convex feasibility problem (following [47]) and then as a convex optimization problem.

In Section 1.5, we prove a general form of the M^* bound. Our proof borrowed from [59] is quite simple and instructive. Once the M^* bound is stated in the language of stochastic processes, it follows quickly by application of symmetrization, contraction, and rotation invariance.

In Section 1.6, we apply the general M^* bound to estimation problems; observations here are still linear but can be noisy. Examples of such problems include *sparse recovery* problems and *linear regression* with constraints, which we explore in Section 1.7.

In Section 1.8, we extend the theory from Gaussian to sub-Gaussian observations. A sub-Gaussian M^* bound (similar to the one obtained in [47]) is deduced from the previous (Gaussian) argument followed by an application of a deep comparison theorem of X. Fernique and M. Talagrand (see [71]).

In Section 1.9, we pass to *exact recovery* results, where an unknown vector $\boldsymbol{x}$ can be inferred from the observations y_i without any error. We present a simple geometric argument based on Y. Gordon's "escape through a mesh" theorem [33]. This argument was first used in this context for sets of sparse vectors in [66], was further developed in [53, 69], and pushed forward for general feasible sets in [3, 16, 72].

In Section 1.10, we explore matrix estimation problems. We first show how the general theory applies to a *low-rank matrix recovery* problem. Then we address a *matrix completion* problem with a short and self-contained argument from [60].

Finally, we pass to nonlinear observations. In Section 1.11, we consider *single-bit observations* $y_i = \operatorname{sign} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$. Analogously to linear observations, there is a clear geometric interpretation for these as well. Namely, the estimation problem reduces in this case to a *pizza cutting problem* about random hyperplane tessellations of K. We discuss a result from [59] on this problem, and we apply it to estimation by formulating it as a feasibility problem.

Similarly to what we did for linear observations, we replace the feasibility problem by optimization problem in Section 1.12. Unlike before, such replacement is not trivial. We present a simple and self-contained argument from [58] about estimation from single-bit observations via convex optimization.

In Section 1.13, we discuss the estimation problem for general (not only single bit) observations following [60]. The new crucial step of estimation is the metric projection onto the feasible set; this projection was studied recently in [17] and [60].

In Section 1.14, we outline some natural extensions of the results for general distributions and to a localized version of mean width.

1.1.4 Acknowledgements

The author is grateful to Vladimir Koltchinskii, Shahar Mendelson, Renato Negrinho, Robert Nowak, Yaniv Plan, Elizaveta Rebrova, Joel Tropp, and especially the anonymous referees for their helpful discussions, comments, and corrections, which lead to a better presentation of this chapter.

1.2 High-dimensional estimation problems

1.2.1 Estimating vectors from random observations

Suppose we want to estimate an unknown vector $\boldsymbol{x} \in \mathbb{R}^n$. In signal processing, $\boldsymbol{x}$ could be a signal to be reconstructed, while in statistics $\boldsymbol{x}$ may represent a parameter of a distribution. We assume that information about $\boldsymbol{x}$ comes from a sample of independent and identically distributed observations $y_1, \dots, y_m \in \mathbb{R}$, which are drawn from a certain distribution which depends on $\boldsymbol{x}$:

$$y_i \sim \text{distribution}(\boldsymbol{x}), \quad i = 1, \dots, m.$$

So we want to estimate $\boldsymbol{x} \in \mathbb{R}^n$ from the observation vector

$$\boldsymbol{y} = (y_1, \dots, y_m) \in \mathbb{R}^m.$$

One example of this situation is the classical linear regression problem in statistics,

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\nu}, \tag{1.1}$$

in which one wants to estimate the coefficient vector $\boldsymbol{\beta}$ from the observation vector $\boldsymbol{y}$. We will see many more examples later; for now let us continue with setting up the general mathematical framework.

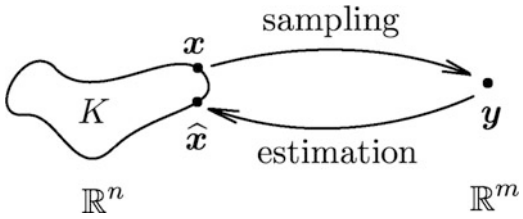


Fig. 1.1 Estimation problem in high dimensions

1.2.2 Low complexity structures

It often happens that we know in advance, believe in, or want to enforce some properties of the vector $\boldsymbol{x}$. We can formalize such extra information as the assumption that

$$\boldsymbol{x} \in K$$

where K is some fixed and known subset of $\mathbb{R}^n$, a *feasible set*. This is a very general and flexible assumption. At this point, we are not stipulating any properties of the feasible set K.

To give a quick example, in regression problem (1.1), one often believes that $\boldsymbol{\beta}$ is a sparse vector, i.e., among its coefficients only few are nonzero. This is important because it means that a few explanatory variables can adequately explain the dependent variable. So one could choose K to be a set of all s-sparse vectors in $\mathbb{R}^n$—those with at most s nonzero coordinates, for a fixed sparsity level $s \leq n$. More examples of natural feasible sets K will be given later.

Figure 1.1 illustrates the estimation problem. Sampling can be thought of as a map taking $\boldsymbol{x} \in K$ to $\boldsymbol{y} \in \mathbb{R}^m$; estimation is a map from $\boldsymbol{y} \in \mathbb{R}^m$ to $\hat{\boldsymbol{x}} \in K$ and is ideally the inverse of sampling.

How can a prior information encoded by K help in high-dimensional estimation? Let us start with a quick and non-rigorous argument based on the number of degrees of freedom. The unknown vector $\boldsymbol{x}$ has n dimensions and the observation vector $\boldsymbol{y}$ has m dimensions. So in principle, it should be possible to estimate $\boldsymbol{x}$ from $\boldsymbol{y}$ with

$$m = O(n)$$

observations. Moreover, this bound should be tight in general.

Now let us add the restriction that $\boldsymbol{x} \in K$. If K happens to be *low dimensional*, with algebraic dimension $\dim(K) = d \ll n$, then $\boldsymbol{x}$ has d degrees of freedom. Therefore, in this case the estimation should be possible with fewer observations,

$$m = O(d) = o(n).$$

It rarely happens that feasible sets of interest literally have small algebraic dimension. For example, the set of all s-sparse vectors in $\mathbb{R}^n$ has full dimension n. Nevertheless, the intuition about low dimensionality remains valid. Natural feasible sets, such as regression coefficient vectors, images, adjacency matrices of networks, do tend to have *low complexity*. Formally K may live in an n-dimensional space where n can be very large, but the actual complexity of K, or "effective dimension" (which we will formally quantify in Section 1.3.5.6), is often much smaller.

This intuition motivates the following three goals, which we will discuss in detail in this chapter:

1. Quantify the complexity of general subsets K of $\mathbb{R}^n$.
2. Demonstrate that estimation can be done with few observations as long as the feasible set K has low complexity.
3. Design estimators that are algorithmically efficient.

We will start by developing intuition about the geometry of sets K in high dimensions. This will take us a short excursion into high-dimensional convex geometry. Although convexity assumption for K will not be imposed in most results of this chapter, it is going to be useful in Section 1.3 for developing a good intuition about geometry in high dimensions.

1.3 An excursion into high-dimensional convex geometry

High-dimensional convex geometry studies *convex bodies* K in $\mathbb{R}^n$ for large n; those are closed, bounded, convex sets with nonempty interior. This area of mathematics is sometimes also called asymptotic convex geometry (referring to n increasing to infinity) and geometric functional analysis. The tutorial [6] could be an excellent first contact with this field; the survey [30] and books [4, 32, 52, 56] cover more material and in more depth.

1.3.1 What do high-dimensional convex bodies look like?

A central problem in high-dimensional convex geometry is—*what do convex bodies look like in high dimensions?* A heuristic answer to this question is—a convex body K usually consists of a *bulk* and *outliers*. The bulk makes up most of the volume of K, but it is usually small in diameter. The outliers contribute little to the volume, but they are large in diameter.

If K is properly scaled, the bulk usually looks like a Euclidean ball. The outliers look like thin, long tentacles. This is best seen Figure 1.2a, which depicts V. Milman's vision of high-dimensional convex sets [51]. This picture does not look convex, and there is a good reason for this. The volume in high dimensions scales differently than in low dimensions—dilating of a set by the factor 2 increases its

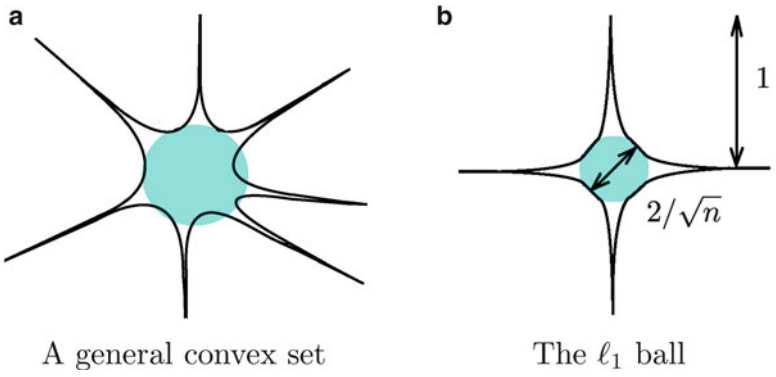


Fig. 1.2 V. Milman's "hyperbolic" drawings of high-dimensional convex sets

volume by the factor 2^n. This is why it is not surprising that the tentacles contain exponentially less volume than the bulk. Such behavior is best seen if a picture looks "hyperbolic." Although not convex, pictures like Figure 1.2 more accurately reflect the distribution of volume in higher dimensions.

Example 3.1 (The ℓ_1 ball). To illustrate this heuristic on a concrete example, consider the set

$$K = B_1^n = \{x \in \mathbb{R}^n : \ \|x\|_1 \leq 1\},$$

i.e., the unit ℓ_1 ball in $\mathbb{R}^n$. The inscribed Euclidean ball in K, which we will denote by B, has diameter $2/\sqrt{n}$. One can then check that volumes of B and of K are comparable:[1]

$$\mathrm{vol}_n(B)^{1/n} \asymp \mathrm{vol}_n(K)^{1/n} \asymp \frac{1}{n}.$$

Therefore, B (perhaps inflated by a constant factor) forms the bulk of K. It is round, makes up most of the volume of K, but has small diameter. The outliers of K are thin and long tentacles protruding quite far in the coordinate directions. This can be best seen in a hyperbolic drawing, see Figure 1.2b.

1.3.2 Concentration of volume

The heuristic representation of convex bodies just described can be supported by some rigorous results about *concentration of volume*.

[1]Here $a_n \asymp b_n$ means that there exists positive absolute constants c and C such that $ca_n \leq b_n \leq Ca_n$ for all n.

These results assume that K is *isotropic*, which means that the random vector X distributed uniformly in K (according to the Lebesgue measure) has zero mean and identity covariance:

$$\mathbb{E}\, X = 0, \quad \mathbb{E}\, XX^{\mathsf{T}} = I_n. \tag{1.2}$$

Isotropy is just an assumption of proper scaling—one can always make a convex body K isotropic by applying a suitable invertible linear transformation.

With this scaling, most of the volume of K is located around the Euclidean sphere of radius $\sqrt{n}$. Indeed, taking traces on both sides of the second equation in (1.2), we obtain

$$\mathbb{E}\, \|X\|_2^2 = n.$$

Therefore, by Markov's inequality, at least 90% of the volume of K is contained in a Euclidean ball of size $O(\sqrt{n})$. Much more powerful concentration results are known—the bulk of K lies very near the sphere of radius $\sqrt{n}$ and the outliers have exponentially small volume. This is the content of the two major results in high-dimensional convex geometry, which we summarize in the following theorem.

Theorem 3.2 (Distribution of volume in high-dimensional convex sets). *Let K be an isotropic convex body in $\mathbb{R}^n$, and let X be a random vector uniformly distributed in K. Then the following is true:*

1. (Concentration of volume) For every $t \geq 1$, one has

$$\mathbb{P}\left\{\|X\|_2 > t\sqrt{n}\right\} \leq \exp(-ct\sqrt{n}).$$

2. (Thin shell) For every $\varepsilon \in (0, 1)$, one has

$$\mathbb{P}\left\{\left|\|X\|_2 - \sqrt{n}\right| > \varepsilon\sqrt{n}\right\} \leq C\exp(-c\varepsilon^3 n^{1/2}).$$

Here and later in this chapter, C, c denote positive absolute constants.

The concentration part of Theorem 3.2 is due to G. Paouris [54]; see [1] for an alternative and shorter proof. The thin shell part is an improved version of a result of B. Klartag [38], which is due to O. Guedon and E. Milman [35].

1.3.3 Low-dimensional random sections

The intuition about bulk and outliers of high-dimensional convex bodies K can help us to understand what *random sections* of K should look like. Suppose E is a random subspace of $\mathbb{R}^n$ with fixed dimension d, i.e., E is drawn at random from the Grassmanian manifold $G_{n,d}$ according to the Haar measure. What does the section $K \cap E$ look like on average?

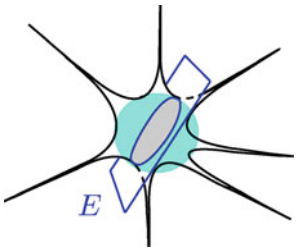


Fig. 1.3 Random section of a high-dimensional convex set

If d is sufficiently small, then we should expect E to pass through the bulk of K and miss the outliers, as those have very small volume. Thus if the bulk of K is a round ball,[2] we should expect the section $K \cap E$ to be a *round ball* as well; see Figure 1.3.

There is a rigorous result which confirms this intuition. It is known as Dvoretzky's theorem [23, 24], which we shall state in the form of V. Milman [48]; expositions of this result can be found, e.g., in [32, 56]. Dvoretzky–Milman's theorem has laid a foundation for the early development of asymptotic convex geometry. Informally, this result says that random sections of K of dimension $d \sim \log n$ are round with high probability.

Theorem 3.3 (Dvoretzky's theorem). *Let K be an origin-symmetric convex body in $\mathbb{R}^n$ such that the ellipsoid of maximal volume contained in K is the unit Euclidean ball B_2^n. Fix $\varepsilon \in (0,1)$. Let E be a random subspace of dimension $d = c\varepsilon^{-2} \log n$ drawn from the Grassmanian $G_{n,d}$ according to the Haar measure. Then there exists $R \geq 0$ such that with high probability (say, 0.99) we have*

$$(1-\varepsilon)\, B(R) \subseteq K \cap E \subseteq (1+\varepsilon)\, B(R).$$

Here $B(R)$ is the centered Euclidean ball of radius R in the subspace E.

Several important aspects of this theorem are not mentioned here—in particular how, for a given convex set K, to compute the radius R and the largest dimension d of round sections of K. These aspects can be found in modern treatments of Dvoretzky theorem such as [32, 56].

[2] This intuition is a good approximation to truth, but it should be corrected. While concentration of volume tells us that the bulk is *contained* in a certain Euclidean ball (and even in a thin spherical shell), it is not always true that the bulk *is* a Euclidean ball (or shell); a counterexample is the unit cube $[-1,1]^n$. In fact, the cube is the worst convex set in the Dvoretzky theorem, which we are about to state.

1.3.4 High-dimensional random sections?

Dvoretzky's Theorem 3.3 describes the shape of *low*-dimensional random sections $K \cap E$, those of dimensions $d \sim \log n$. Can anything be said about *high*-dimensional sections, those with small codimension? In this more difficult regime, we can no longer expect such sections to be round. Instead, as the codimension decreases, the random subspace E becomes larger and it will probably pick more and more of the outliers (tentacles) of K. The shape of such sections $K \cap E$ is difficult to describe.

Nevertheless, it turns out that we can accurately predict the *diameter* of $K \cap E$. A bound on the diameter is known in asymptotic convex geometry as the low M^* estimate, or M^* bound. We will state this result in Section 1.3.6 and prove it in Section 1.5. For now, let us only mention that M^* bound is particularly attractive in applications as it depends only on two parameters—the codimension of E and a single geometric quantity, which informally speaking, measures the size of the bulk of K. This geometric quantity is called the *mean width* of K. We will pause briefly to discuss this important notion.

1.3.5 Mean width

The concept of mean width captures important geometric characteristics of sets in $\mathbb{R}^n$. One can mentally place it in the same category as other classical geometric quantities like volume and surface area.

Consider a bounded subset K in $\mathbb{R}^n$. (The convexity, closedness, and nonempty interior will not be imposed from now on.) The *width* of K in the direction of a given unit vector $\eta \in S^{n-1}$ is defined as the width of the smallest slab between two parallel hyperplanes with normal η that contains K; see Figure 1.4.

Analytically, we can express the width in the direction of η as

$$\sup_{u,v \in K} \langle \eta, u - v \rangle = \sup_{z \in K-K} \langle \eta, z \rangle$$

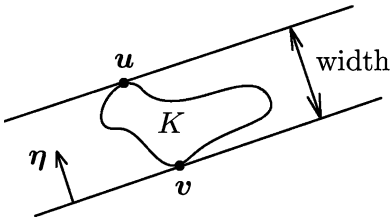


Fig. 1.4 Width of K in the direction of η

where $K-K = \{\boldsymbol{u}-\boldsymbol{v} : \boldsymbol{u}, \boldsymbol{v} \in K\}$ is the Minkowski sum of K and $-K$. Equivalently, we can define the width using the standard notion of *support function* of K, which is $h_K(\boldsymbol{\eta}) = \sup_{\boldsymbol{u}\in K} \langle \boldsymbol{\eta}, \boldsymbol{u}\rangle$, see [64]. The width of K in the direction of $\boldsymbol{\eta}$ can be expressed as $h_K(\boldsymbol{\eta}) + h_K(-\boldsymbol{\eta})$.

Averaging over $\boldsymbol{\eta}$ uniformly distributed on the sphere S^{n-1}, we can define the *spherical mean width* of K:

$$\tilde{w}(K) := \mathbb{E} \sup_{z\in K-K} \langle \boldsymbol{\eta}, z\rangle .$$

This notion is standard in asymptotic geometric analysis.

In other related areas, such as high-dimensional probability and statistical learning theory, it is more convenient to replace the *spherical* random vector $\boldsymbol{\eta} \sim \mathrm{Unif}(S^{n-1})$ by the standard *Gaussian* random vector $\boldsymbol{g} \sim N(0, I_n)$. The advantage is that $\boldsymbol{g}$ has independent coordinates while $\boldsymbol{\eta}$ does not.

Definition 3.4 (Gaussian mean width). The Gaussian *mean width* of a bounded subset K of $\mathbb{R}^n$ is defined as

$$w(K) := \mathbb{E} \sup_{\boldsymbol{u}\in K-K} \langle \boldsymbol{g}, \boldsymbol{u}\rangle , \tag{1.3}$$

where $\boldsymbol{g} \sim N(0, I_n)$ is a standard Gaussian random vector in $\mathbb{R}^n$. We will often refer to Gaussian mean width as simply the *mean width*.

1.3.5.1 Simple properties of mean width

Observe first that the Gaussian mean width is about $\sqrt{n}$ times larger than the spherical mean width. To see this, using rotation invariance we realize $\boldsymbol{\eta}$ as $\boldsymbol{\eta} = \boldsymbol{g}/\|\boldsymbol{g}\|_2$. Next, we recall that the direction and magnitude of a standard Gaussian random vector are independent, so $\boldsymbol{\eta}$ is independent of $\|\boldsymbol{g}\|_2$. It follows that

$$w(K) = \mathbb{E}\,\|\boldsymbol{g}\|_2 \cdot \tilde{w}(K).$$

Further, the factor $\mathbb{E}\,\|\boldsymbol{g}\|_2$ is of order $\sqrt{n}$; this follows, for example, from known bounds on the χ^2 distribution:

$$c\sqrt{n} \le \mathbb{E}\,\|\boldsymbol{g}\|_2 \le \sqrt{n} \tag{1.4}$$

where $c > 0$ is an absolute constant. Therefore, the Gaussian and spherical versions of mean width are equivalent (up to scaling factor $\sqrt{n}$), so it is mostly a matter of personal preference which version to work with. In this chapter, we will mostly work with the Gaussian version.

Let us observe a few standard and useful properties of the mean width, which follow quickly from its definition.

Proposition 3.5. *The mean width is invariant under translations, orthogonal transformations, and taking convex hulls.* □

Especially useful for us will be the last property, which states that

$$w(\mathrm{conv}(K)) = w(K). \tag{1.5}$$

This property will come handy later, when we consider convex relaxations of optimization problems.

1.3.5.2 Computing mean width on examples

Let us illustrate the notion of mean width on some simple examples.

Example 3.6. If K is *the unit Euclidean ball* B_2^n or sphere S^{n-1}, then

$$w(K) = \mathbb{E}\,\|g\|_2 \le \sqrt{n}$$

and also $w(K) \ge c\sqrt{n}$, by (1.4).

Example 3.7. Let K be a subset of B_2^n with *linear algebraic dimension* d. Then K lies in a d-dimensional unit Euclidean ball, so as before we have

$$w(K) \le 2\sqrt{d}.$$

Example 3.8. Let K be *a finite subset* of B_2^n. Then

$$w(K) \le C\sqrt{\log |K|}.$$

This follows from a known and simple computation of the expected maximum of $k = |K|$ Gaussian random variables.

Example 3.9 (Sparsity). Let K consist of all unit s-sparse vectors in $\mathbb{R}^n$—those with at most s nonzero coordinates:

$$K = \{x \in \mathbb{R}^n : \ \|x\|_2 = 1, \ \|x\|_0 \le s\}.$$

Here $\|x\|_0$ denotes the number of nonzero coordinates of x. A simple computation (see, e.g., [58, Lemma 2.3]) shows that

$$c\sqrt{s\log(2n/s)} \le w(K) \le C\sqrt{s\log(2n/s)}.$$

Example 3.10 (Low rank). Let K consist of $d_1 \times d_2$ matrices with unit Frobenius norm and rank at most r:

$$K = \{X \in \mathbb{R}^{d_1 \times d_2} : \ \|X\|_F = 1, \ \mathrm{rank}(X) \le r\}.$$

We will see in Proposition 10.4,

$$w(K) \le C\sqrt{r(d_1 + d_2)}.$$

1.3.5.3 Computing mean width algorithmically

Can we estimate the mean width of a given set K fast and accurately? Gaussian concentration of measure (see [42, 43, 56]) implies that, with high probability, the random variable

$$w(K, \boldsymbol{g}) = \sup_{\boldsymbol{u} \in K-K} \langle \boldsymbol{g}, \boldsymbol{u} \rangle$$

is close to its expectation $w(K)$. Therefore, to estimate $w(K)$, it is enough to generate a single realization of a random vector $\boldsymbol{g} \sim N(0, I_n)$ and compute $w(K, \boldsymbol{g})$; this should produce a good estimator of $w(K)$.

Since we can convexify K without changing the mean width by Proposition 3.5, computing this estimator is a *convex optimization problem* (and often even a linear problem if K is a polytope).

1.3.5.4 Computing mean width theoretically

Finding theoretical estimates on the mean width of a given set K is a nontrivial problem. It has been extensively studied in the areas of probability in Banach spaces and stochastic processes.

Two classical results in the theory of stochastic processes—Sudakov's inequality (see [43, Theorem 3.18]) and Dudley's inequality (see [43, Theorem 11.17])—relate the mean width to the metric entropy of K. Let $N(K, t)$ denote the smallest number of Euclidean balls of radius t whose union covers K. Usually $N(K, t)$ is referred to as a *covering number* of K, and $\log N(K, t)$ is called the *metric entropy* of K.

Theorem 3.11 (Sudakov's and Dudley's inequalities). *For any bounded subset K of $\mathbb{R}^n$, we have*

$$c \sup_{t>0} t\sqrt{\log N(K, t)} \le w(K) \le C \int_0^\infty \sqrt{\log N(K, t)}\, dt.$$

The lower bound is Sudakov's inequality and the upper bound is Dudley's inequality.

Neither Sudakov's nor Dudley's inequality is tight for all sets K. A more advanced method of *generic chaining* produces a tight (but also more complicated) estimate of the mean width in terms of *majorizing measures*; see [71].

Let us only mention some other known ways to control mean width. In some cases, *comparison inequalities* for Gaussian processes can be useful, especially

Slepian's and Gordon's; see [43, Section 3.3]. There is also a combinatorial approach to estimating the mean width and metric entropy, which is based on *VC dimension* and its generalizations; see [44, 65].

1.3.5.5 Mean width and Gaussian processes

The theoretical tools of estimating mean width we just mentioned, including Sudakov's, Dudley's, Slepian's, and Gordon's inequalities, have been developed in the context of stochastic processes. To see the connection, consider the Gaussian random variables $G_{\boldsymbol{u}} = \langle \boldsymbol{g}, \boldsymbol{u} \rangle$ indexed by points $\boldsymbol{u} \in \mathbb{R}^n$. The collection of these random variables $(G_{\boldsymbol{u}})_{\boldsymbol{u} \in K-K}$ forms a *Gaussian process*, and the mean width measures the size of this process:

$$w(K) = \mathbb{E} \sup_{\boldsymbol{u} \in K-K} G_{\boldsymbol{u}}.$$

In some sense, any Gaussian process can be approximated by a process of this form. We will return to the connection between mean width and Gaussian processes in Section 1.5 where we prove the M^* bound.

1.3.5.6 Mean width, complexity, and effective dimension

In the context of stochastic processes, Gaussian mean width (and its non-Gaussian variants) plays an important role in statistical learning theory. There it is more natural to work with classes $\mathcal{F}$ of real-valued functions on $\{1, \dots, n\}$ than with geometric sets $K \subseteq \mathbb{R}^n$. (We identify a vector in $\mathbb{R}^n$ with a function on $\{1, \dots, n\}$.) The Gaussian mean width serves as a measure of complexity of a function class in statistical learning theory, see [45]. It is sometimes called *Gaussian complexity* and is usually denoted $\gamma_2(\mathcal{F})$.

To get a better feeling of mean width as complexity, assume that K lies in the unit Euclidean ball B_2^n. The square of the mean width, $w(K)^2$, may be interpreted as the *effective dimension* of K. By Example 3.7, the effective dimension is always bounded by the linear algebraic dimension. However, unlike algebraic dimension, the effective dimension is *robust*—a small perturbation of K leads to a small change in $w(K)^2$.

1.3.6 Random sections of small codimension: M bound*

Let us return to the problem we posed in Section 1.3.4 – bounding the diameter of random sections $K \cap E$ where E is a high-dimensional subspace. The following important result in asymptotic convex geometry gives a good answer to this question.

Theorem 3.12 (M^* bound). *Let K be a bounded subset of $\mathbb{R}^n$. Let E be a random subspace of $\mathbb{R}^n$ of a fixed codimension m, drawn from the Grassmanian $G_{n,n-m}$ according to the Haar measure. Then*

$$\mathbb{E}\,\mathrm{diam}(K \cap E) \le \frac{Cw(K)}{\sqrt{m}}.$$

We will prove a stronger version of this result in Section 1.5. The first variant of M^* bound was found by V. Milman [49, 50]; its present form is due to A. Pajor and N. Tomczak-Jaegermann [55]; an alternative argument which yields tight constants was given by Y. Gordon [33]; an exposition of M^* bound can be found in [43, 56].

To understand the M^* bound better, it is helpful to recall from Section 1.3.5.1 that $w(K)/\sqrt{n}$ is equivalent to the *spherical* mean width of K. Heuristically, the spherical mean width measures the *size of the bulk* of K.

For subspace E of not very high dimension, where $m = \Omega(n)$, the M^* bound states that the size of the random section $K \cap E$ is bounded by the spherical mean width of K. In other words, subspace E of proportional dimension *passes through the bulk* of K and ignores the outliers ("tentacles"), just as Figure 1.3 illustrates. But when the dimension of the subspace E grows toward n (so the codimension m becomes small), the diameter of $K \cap E$ also grows by a factor of $\sqrt{n/m}$. This gives a precise control of how E in this case *interferes with the outliers* of K.

1.4 From geometry to estimation: linear observations

Having completed the excursion into geometry, we can now return to the high-dimensional estimation problems that we started to discuss in Section 1.2. To recall, our goal is to estimate an unknown vector

$$\boldsymbol{x} \in K \subseteq \mathbb{R}^n$$

that lies in a known feasible set K, from a random observation vector

$$\boldsymbol{y} = (y_1, \dots, y_m) \in \mathbb{R}^m,$$

whose coordinates y_i are random i.i.d. observations of x.

So far, we have not been clear about possible distributions of the observations y_i. In this section, we will study perhaps the simplest model—*Gaussian linear observations*. Consider i.i.d. standard Gaussian vectors

$$\boldsymbol{a}_i \sim N(0, I_n)$$

and define

$$y_i = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle, \quad i = 1, \dots, m.$$

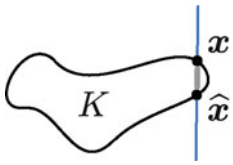


Fig. 1.5 Estimating $\boldsymbol{x}$ by any vector $\hat{\boldsymbol{x}}$ in the intersection of K with the affine subspace $\{\boldsymbol{x}' : A\boldsymbol{x}' = \boldsymbol{y}\}$

Thus the observation vector $\boldsymbol{y}$ depends linearly on $\boldsymbol{x}$. This is best expressed in a matrix form:

$$\boldsymbol{y} = A\boldsymbol{x}.$$

Here A is an $m \times n$ Gaussian random matrix, which means that the entries of A are i.i.d. $N(0, 1)$ random variables; the vectors $\boldsymbol{a}_i$ form the rows of A.

The interesting regime is when the number of observations is smaller than the dimension, i.e., when $m < n$. In this regime, the problem of estimating $\boldsymbol{x} \in \mathbb{R}^n$ from $\boldsymbol{y} \in \mathbb{R}^m$ is ill posed. (In the complementary regime, where $m \geq n$, the linear system $\boldsymbol{y} = A\boldsymbol{x}$ is well posed since A has full rank almost surely, so the solution is trivial.)

1.4.1 Estimation based on M bound*

Recall that we know two pieces of information about $\boldsymbol{x}$:

1. $\boldsymbol{x}$ lies in a known random affine subspace $\{\boldsymbol{x}' : A\boldsymbol{x}' = \boldsymbol{y}\}$;
2. $\boldsymbol{x}$ lies in a known set K.

Therefore, a good estimator of $\boldsymbol{x}$ can be obtained by picking any vector $\hat{\boldsymbol{x}}$ from the *intersection of these two sets*; see Figure 1.5. Moreover, since just these two pieces of information about $\boldsymbol{x}$ are available, such estimator is best possible in some sense.

How good is such estimate? The maximal error is, of course, the distance between two farthest points in the intersection of K with the affine subspace $\{\boldsymbol{x}' : A\boldsymbol{x}' = \boldsymbol{y}\}$. This distance in turn equals the diameter of the section of K by this random subspace. But this diameter is controlled by M^* bound, Theorem 3.12. Let us put together this argument more rigorously.

In the following theorem, the setting is the same as above: $K \subset \mathbb{R}^n$ is a bounded subset, $\boldsymbol{x} \in K$ is an unknown vector, and $\boldsymbol{y} = A\boldsymbol{x}$ is the observation vector, where A is an $m \times n$ Gaussian matrix.

Theorem 4.1 (Estimation from linear observations: feasibility program). *Choose $\hat{\boldsymbol{x}}$ to be any vector satisfying*

$$\hat{\boldsymbol{x}} \in K \quad \text{and} \quad A\hat{\boldsymbol{x}} = \boldsymbol{y}. \tag{1.6}$$

Then

$$\mathbb{E} \sup_{x \in K} \|\hat{x} - x\|_2 \leq \frac{Cw(K)}{\sqrt{m}}.$$

Proof. We apply the M^* bound, Theorem 3.12, for the set $K - K$ and the subspace $E = \ker(A)$. Rotation invariance of Gaussian distribution implies that E is uniformly distributed in the Grassmanian $G_{n,n-m}$, as required by the M^* bound. Moreover, it is straightforward to check that $w(K - K) \leq 2w(K)$. It follows that

$$\mathbb{E} \operatorname{diam}((K - K) \cap E) \leq \frac{Cw(K)}{\sqrt{m}}.$$

It remains to note that since $\hat{x}, x \in K$ and $A\hat{x} = Ax = y$, we have $\hat{x} - x \in (K - K) \cap E$. □

The argument we just described was first suggested by S. Mendelson, A. Pajor, and N. Tomczak-Jaegermann [47].

1.4.2 Estimation as an optimization problem

Let us make one step forward and replace the feasibility program (1.6) by a more flexible *optimization* program.

For this, let us make an additional (but quite mild) assumption that K has nonempty interior and is *star-shaped*. Being star-shaped means that together with each point, the set K contains the segment joining that point to the origin; in other words,

$$tK \subseteq K \quad \text{for all } t \in [0, 1].$$

For such set K, let us revise the feasibility program (1.6). Instead of intersecting a fixed set K with the affine subspace $\{x' : Ax' = y\}$, we may *blow up* K (i.e., consider a dilate tK with increasing $t \geq 0$) until it touches that subspace. Choose $\hat{x}$ to be the touching point, see Figure 1.6. The fact that K is star-shaped implies that $\hat{x}$ still belongs to K and (obviously) the affine subspace; thus $\hat{x}$ satisfies the same error bound as in Theorem 4.1.

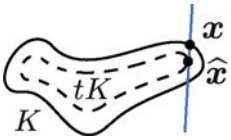


Fig. 1.6 Estimating x by blowing up K until it touches the affine subspace $\{x' : Ax' = y\}$

To express this estimator analytically, it is convenient to use the notion of *Minkowski functional* of K, which associates to each point $\boldsymbol{x} \in \mathbb{R}^n$ a nonnegative number $\|\boldsymbol{x}\|_K$ defined by the rule

$$\|\boldsymbol{x}\|_K = \inf\left\{\lambda > 0 : \ \lambda^{-1}\boldsymbol{x} \in K\right\}.$$

Minkowski functionals, also called *gauges*, are standard notions in geometric functional analysis and convex analysis. Convex analysis textbooks such as [64] offer thorough treatments of this concept. We just mention here a couple of elementary properties. First, the function $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|_K$ is continuous on $\mathbb{R}^n$ and it is positive homogeneous (that is, $\|a\boldsymbol{x}\|_K = a\|\boldsymbol{x}\|_K$ for $a > 0$). Next, a closed set K is the 1-sublevel set of its Minkowski functional, that is,

$$K = \{x : \|\boldsymbol{x}\|_K \le 1\}.$$

A typical situation to think of is when K is a symmetric convex body (i.e. K is closed, bounded, has nonempty interior, and is origin symmetric); then $\|\boldsymbol{x}\|_K$ defines a *norm* on $\mathbb{R}^n$ with K being the unit ball.

Let us now accurately state an optimization version of Theorem 4.1. It is valid for an arbitrary bounded star-shaped set K with nonempty interior.

Theorem 4.2 (Estimation from linear observations: optimization program). *Choose $\hat{\boldsymbol{x}}$ to be a solution of the program*

$$\text{minimize} \|\boldsymbol{x}'\|_K \quad \text{subjectto} \quad A\boldsymbol{x}' = \boldsymbol{y}. \tag{1.7}$$

Then

$$\mathbb{E} \sup_{x \in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \frac{Cw(K)}{\sqrt{m}}.$$

Proof. It suffices to check that $\hat{\boldsymbol{x}} \in K$; the conclusion would then follow from Theorem 4.1. Both $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$ satisfy the linear constraint $A\boldsymbol{x}' = \boldsymbol{y}$. Therefore, by choice of $\hat{\boldsymbol{x}}$, we have

$$\|\hat{\boldsymbol{x}}\|_K \le \|\boldsymbol{x}\|_K \le 1;$$

the last inequality is nothing else than our assumption that $\boldsymbol{x} \in K$. Thus $\hat{\boldsymbol{x}} \in K$ as claimed. □

1.4.3 Algorithmic aspects: convex programming

What does it take to solve the optimization problem (1.7) algorithmically? If the feasible set K is convex, then (1.7) is a *convex program*. In this case, to

solve this problem numerically one may tap into an array of available convex optimization solvers, in particular interior-point methods [8] and proximal-splitting algorithms [7].

Further, if K is a polytope, then (1.7) can be cast as a *linear program*, which widens an array of algorithmic possibilities even further. For a quick preview, let us mention that examples of the latter kind will be discussed in detail in Section 1.7, where we will use K to enforce *sparsity*. We will thus choose K to be a ball of ℓ_1 norm in $\mathbb{R}^n$, so the program (1.7) will minimize $\|\boldsymbol{x}'\|_1$ subject to $A\boldsymbol{x}' = \boldsymbol{y}$. This is a typical linear program in the area of compressed sensing.

If K is *not* convex, then we can convexify it, thereby replacing K with its convex hull $\mathrm{conv}(K)$. Convexification does not change the mean width according to the remarkable property (1.5). Therefore, the generally nonconvex problem (1.7) can be relaxed to the convex program

$$\text{minimize} \|\boldsymbol{x}'\|_{\mathrm{conv}(K)} \quad \text{subject to} \quad A\boldsymbol{x}' = \boldsymbol{y}, \tag{1.8}$$

without compromising the guarantee of estimation stated in Theorem 4.2. The solution $\hat{\boldsymbol{x}}$ of the convex program (1.8) satisfies

$$\mathbb{E} \sup_{\boldsymbol{x} \in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \mathbb{E} \sup_{\boldsymbol{x} \in \mathrm{conv}(K)} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \frac{Cw(\mathrm{conv}(K))}{\sqrt{m}} = \frac{Cw(K)}{\sqrt{m}}.$$

Summarizing, we see that in any case, whether K is convex or not, the estimation problem reduces to solving an algorithmically tractable convex program. Of course, one needs to be able to compute $\|\boldsymbol{z}\|_{\mathrm{conv}(K)}$ algorithmically for a given vector $\boldsymbol{z} \in \mathbb{R}^n$. This is possible for many (but not all) feasible sets K.

1.4.4 Information-theoretic aspects: effective dimension

If we fix a desired error level, for example if we aim for

$$\mathbb{E} \sup_{\boldsymbol{x} \in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le 0.01,$$

then

$$m \sim w(K)^2$$

observations will suffice. The implicit constant factor here is determined by the desired error level.

Notice that this result is *uniform*. By Markov's inequality, with probability, say 0.9 in A (which determines the observation model) the estimation is accurate simultaneously for all vectors $\boldsymbol{x} \in K$. Moreover, as we observed in Section 1.5.2,

the actual probability is much better than 0.9; it converges to 1 exponentially fast in the number of observations m.

The square of the mean width, $w(K)^2$, can be thought of an *effective dimension* of the feasible set K, as we pointed out in Section 1.3.5.6.

We can summarize our findings as follows.

Using convex programming, one can estimate a vector $\boldsymbol{x}$ in a general feasible set K from m random linear observations. A sufficient number of observations m is the same as the effective dimension of K (the mean width squared), up to a constant factor.

1.5 High-dimensional sections: proof of a general M^* bound

Let us give a quick proof of the M^* bound, Theorem 3.12. In fact, without much extra work we will be able to derive a more general result from [59]. First, it would allow us to treat noisy observations of the form $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{v}$. Second, it will be generalizable for non-Gaussian observations.

Theorem 5.1 (General M^* bound). *Let T be a bounded subset of $\mathbb{R}^n$. Let A be an $m \times n$ Gaussian random matrix (with i.i.d. $N(0,1)$ entries). Fix $\varepsilon \geq 0$ and consider the set*

$$T_\varepsilon := \Big\{ \boldsymbol{u} \in T : \ \frac{1}{m} \|A\boldsymbol{u}\|_1 \leq \varepsilon \Big\}. \tag{1.9}$$

Then[3]

$$\mathbb{E} \sup_{\boldsymbol{u} \in T_\varepsilon} \|\boldsymbol{u}\|_2 \leq \sqrt{\frac{8\pi}{m}} \ \mathbb{E} \sup_{\boldsymbol{u} \in T} |\, \langle \boldsymbol{g}, \boldsymbol{u} \rangle \,| + \sqrt{\frac{\pi}{2}}\, \varepsilon, \tag{1.10}$$

where $g \sim N(0, I_n)$ is a standard Gaussian random vector in $\mathbb{R}^n$.

To see that this result contains the classical M^* bound, Theorem 3.12, we can apply it for $T = K - K$, $\varepsilon = 0$, and identify $\ker(A)$ with E. In this case,

$$T_\varepsilon = (K - K) \cap E.$$

It follows that $T_\varepsilon \supseteq (K \cap E) - (K \cap E)$, so the left-hand side of (1.10) is bounded below by $\operatorname{diam}(K \cap E)$. The right-hand side of (1.10) by symmetry equals $\sqrt{8\pi/m}\, w(K)$. Thus we recover Theorem 3.12 with $C = \sqrt{8\pi}$.

Our proof of Theorem 5.1 will be based on two basic tools in the theory of *stochastic processes*—symmetrization and contraction.

[3] Conclusion (1.10) is stated with the convention that $\sup_{\boldsymbol{u} \in T_\varepsilon} \|u\|_2 = 0$ whenever $T_\varepsilon = \emptyset$.

A stochastic process is simply a collection of random variables $(Z(t))_{t\in T}$ on the same probability space. The index space T can be arbitrary; it may be a time interval (such as in Brownian motion) or a subset of $\mathbb{R}^n$ (as will be our case). To avoid measurability issues, we can assume that T is finite by discretizing it if necessary.

Proposition 5.2. *Consider a finite collection of stochastic processes $Z_1(t), \dots, Z_m(t)$ indexed by $t \in T$. Let ε_i be independent Rademacher random variables (that is, ε_i independently take values -1 and 1 with probabilities $1/2$ each). Then we have the following:*

(i) (Symmetrization)

$$\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^m \big[Z_i(t) - \mathbb{E}\,Z_i(t)\big]\Big| \le 2\,\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^m \varepsilon_i Z_i(t)\Big|.$$

(ii) (Contraction)

$$\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^m \varepsilon_i |Z_i(t)|\Big| \le 2\,\mathbb{E}\sup_{t\in T}\Big|\sum_{i=1}^m \varepsilon_i Z_i(t)\Big|.$$

Both statements are relatively easy to prove even in greater generality. For example, taking the absolute values of $Z_i(t)$ in the contraction principle can be replaced by applying general Lipschitz functions. Proofs of symmetrization and contraction principles can be found in [43, Lemma 6.3] and [43, Theorem 4.12], respectively.

1.5.1 Proof of Theorem 3.12

Let $\boldsymbol{a}_i^{\mathsf{T}}$ denote the rows of A; thus $\boldsymbol{a}_i$ are independent $N(0, I_n)$ random vectors. The desired bound (1.10) would follow from the deviation inequality

$$\mathbb{E}\sup_{\boldsymbol{u}\in T}\Big|\frac{1}{m}\sum_{i=1}^m |\,\langle \boldsymbol{a}_i, \boldsymbol{u}\rangle\,| - \sqrt{\frac{2}{\pi}}\,\|\boldsymbol{u}\|_2\Big| \le \frac{4}{\sqrt{m}}\,\mathbb{E}\sup_{\boldsymbol{u}\in T}|\,\langle \boldsymbol{g}, \boldsymbol{u}\rangle\,|. \tag{1.11}$$

Indeed, if this inequality holds, then same is true if we replace T by the smaller set T_ε on the left-hand side of (1.11). But for $\boldsymbol{u} \in T_\varepsilon$, we have $\frac{1}{m}\sum_{i=1}^m |\,\langle \boldsymbol{a}_i, \boldsymbol{u}\rangle\,| = \frac{1}{m}\|A\boldsymbol{u}\|_1 \le \varepsilon$, and the bound (1.10) follows by triangle inequality.

The rotation invariance of Gaussian distribution implies that

$$\mathbb{E}\,|\,\langle \boldsymbol{a}_i, \boldsymbol{u}\rangle\,| = \sqrt{\frac{2}{\pi}}\,\|\boldsymbol{u}\|_2. \tag{1.12}$$

Thus using symmetrization and then contraction inequalities from Proposition 5.2, we can bound the left-hand side of (1.11) by

$$4\,\mathbb{E}\sup_{\boldsymbol{u}\in T}\left|\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\,\langle \boldsymbol{a}_i,\boldsymbol{u}\rangle\,\right| = 4\,\mathbb{E}\sup_{\boldsymbol{u}\in T}\left|\left\langle\frac{1}{m}\sum_{i=1}^{m}\varepsilon_i\boldsymbol{a}_i,\boldsymbol{u}\right\rangle\right|. \tag{1.13}$$

Here ε_i are independent Rademacher variables.

Conditioning on ε_i and using rotation invariance, we see that the random vector

$$\boldsymbol{g} := \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varepsilon_i\boldsymbol{a}_i$$

has distribution $N(0, I_n)$. Thus (1.13) can be written as

$$\frac{4}{\sqrt{m}}\,\mathbb{E}\sup_{\boldsymbol{u}\in T}|\,\langle \boldsymbol{g},\boldsymbol{u}\rangle\,|.$$

This proves (1.11) and completes the proof of Theorem 5.1. □

1.5.2 From expectation to overwhelming probability

The M^* bound that we just proved and in fact all results in this survey are stated in terms of expected value for simplicity of presentation. One can upgrade them to estimates with overwhelming probability using *concentration of measure*, see [42]. We will illustrate this method with a couple of examples; the reader can apply similar reasoning for several other results we have proved.

Let us first obtain a high-probability version of the deviation inequality (1.11) using the *Gaussian concentration inequality*. We will consider the deviation

$$Z(A) := \sup_{\boldsymbol{u}\in T}\left|\frac{1}{m}\sum_{i=1}^{m}|\,\langle \boldsymbol{a}_i,\boldsymbol{u}\rangle\,| - \sqrt{\frac{2}{\pi}}\,\|\boldsymbol{u}\|_2\right|$$

as a function of the matrix $A \in \mathbb{R}^{m\times n}$. Let us show that it is a Lipschitz function on $\mathbb{R}^{m\times n}$ equipped with Frobenius norm $\|\cdot\|_F$ (which is the same as the Euclidean norm on $\mathbb{R}^{mn}$). Indeed, two applications of the triangle inequality followed by two applications of the Cauchy–Schwarz inequality imply that for matrices A and B with rows $\boldsymbol{a}_i^{\mathsf{T}}$ and $\boldsymbol{b}_i^{\mathsf{T}}$, respectively, we have

$$
\begin{aligned}
|Z(A) - Z(B)| &\le \sup_{\boldsymbol{u}\in T} \frac{1}{m} \sum_{i=1}^{m} |\, \langle \boldsymbol{a}_i - \boldsymbol{b}_i, \boldsymbol{u}\rangle \,| \\
&\le \frac{d(T)}{m} \sum_{i=1}^{m} \|\boldsymbol{a}_i - \boldsymbol{b}_i\|_2 \quad (\text{where} d(T) = \max_{\boldsymbol{u}\in T} \|\boldsymbol{u}\|_2) \\
&\le \frac{d(T)}{\sqrt{m}} \|A - B\|_F.
\end{aligned}
$$

Thus the function $A \mapsto Z(A)$ has Lipschitz constant bounded by $d(K)/\sqrt{m}$. We may now bound the deviation probability for Z using the Gaussian concentration inequality (see [43, Equation 1.6]) as follows:

$$
\mathbb{P}\left\{|Z - \mathbb{E}\, Z| \ge t\right\} \le 2\exp\Big(-\frac{mt^2}{2d(T)^2}\Big), \quad t \ge 0.
$$

This is a high-probability version of the deviation inequality (1.11).

Using this inequality, one quickly deduces a corresponding high-probability version of Theorem 5.1. It states that

$$
\sup_{\boldsymbol{u}\in T_\varepsilon} \|\boldsymbol{u}\|_2 \le \sqrt{\frac{8\pi}{m}}\; \mathbb{E} \sup_{\boldsymbol{u}\in T} |\, \langle \boldsymbol{g}, \boldsymbol{u}\rangle \,| + \sqrt{\frac{\pi}{2}}\, (\varepsilon + t)
$$

with probability at least $1 - 2\exp(-mt^2/2d(T)^2)$.

As before, we obtain from this the following *high-probability version of the M^* bound*, Theorem 5.1. It states that

$$
\operatorname{diam}(K \cap E) \le \frac{Cw(K)}{\sqrt{m}} + Ct
$$

with probability at least $1 - 2\exp(-mt^2/2\operatorname{diam}(K)^2)$.

1.6 Consequences: estimation from noisy linear observations

Let us apply the general M^* bound, Theorem 5.1, to estimation problems. This will be even more straightforward than our application of the standard M^* bound in Section 1.4. Moreover, we will now be able to treat *noisy* observations.

Like before, our goal is to estimate an unknown vector $\boldsymbol{x}$ that lies in a known feasible set $K \subset \mathbb{R}^n$, from a random observation vector $\boldsymbol{y} \in \mathbb{R}^m$. This time we assume that, for some known level of noise $\varepsilon \ge 0$, we have

$$
\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{v}, \qquad \frac{1}{m}\|\boldsymbol{v}\|_1 = \frac{1}{m}\sum_{i=1}^{m} |v_i| \le \varepsilon. \tag{1.14}
$$

Here A is an $m \times n$ Gaussian matrix as before. The noise vector $\boldsymbol{\nu}$ may be unknown and have arbitrary structure. In particular, $\boldsymbol{\nu}$ may depend on A, so even adversarial errors are allowed. The ℓ_1 constraint in (1.14) can clearly be replaced by the stronger ℓ_2 constraint

$$\frac{1}{m}\|\boldsymbol{\nu}\|_2^2 = \frac{1}{m}\sum_{i=1}^{m} \nu_i^2 \le \varepsilon^2.$$

The following result is a generalization of Theorem 4.1 for noisy observations (1.14). As before, it is valid for any bounded set $K \subset \mathbb{R}^n$.

Theorem 6.1 (Estimation from noisy linear observations: feasibility program). *Choose $\hat{\boldsymbol{x}}$ to be any vector satisfying*

$$\hat{\boldsymbol{x}} \in K \quad \text{and} \quad \frac{1}{m}\|A\hat{\boldsymbol{x}} - \boldsymbol{y}\|_1 \le \varepsilon. \tag{1.15}$$

Then

$$\mathbb{E} \sup_{\boldsymbol{x}\in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \sqrt{8\pi}\left(\frac{w(K)}{\sqrt{m}} + \varepsilon\right).$$

Proof. We apply the general M^* bound, Theorem 5.1, for the set $T = K - K$, and with 2ε instead of ε. It follows that

$$\mathbb{E} \sup_{\boldsymbol{u}\in T_{2\varepsilon}} \|\boldsymbol{u}\|_2 \le \sqrt{\frac{8\pi}{m}}\, \mathbb{E} \sup_{\boldsymbol{u}\in T} |\langle \boldsymbol{g}, \boldsymbol{u}\rangle| + \sqrt{2\pi}\,\varepsilon \le \sqrt{8\pi}\left(\frac{w(K)}{\sqrt{m}} + \varepsilon\right).$$

The last inequality follows from the definition of mean width and the symmetry of T.

To finish the proof, it remains to check that

$$\hat{\boldsymbol{x}} - \boldsymbol{x} \in T_{2\varepsilon}. \tag{1.16}$$

To prove this, first note that $\hat{\boldsymbol{x}}, \boldsymbol{x} \in K$, so $\hat{\boldsymbol{x}} - \boldsymbol{x} \in K - K = T$. Next, by triangle inequality, we have

$$\frac{1}{m}\|A(\hat{\boldsymbol{x}} - \boldsymbol{x})\|_1 = \frac{1}{m}\|A\hat{\boldsymbol{x}} - \boldsymbol{y} + \boldsymbol{\nu}\|_1 \le \frac{1}{m}\|A\hat{\boldsymbol{x}} - \boldsymbol{y}\|_1 + \frac{1}{m}\|\boldsymbol{\nu}\|_1 \le 2\varepsilon.$$

The last inequality follows from (1.14) and (1.15). We showed that the vector $\boldsymbol{u} = \hat{\boldsymbol{x}} - \boldsymbol{x}$ satisfies both constraints that define $T_{2\varepsilon}$ in (1.9). Hence (1.16) holds, and the proof of the theorem is complete. □

And similarly to Theorem 4.2, we can cast estimation as an optimization (rather than feasibility) program. As before, it is valid for any bounded star-shaped set $K \subset \mathbb{R}^n$ with nonempty interior.

Theorem 6.2 (Estimation from noisy linear observations: optimization program). *Choose $\hat{\boldsymbol{x}}$ to be a solution to the program*

$$\text{minimize} \|\boldsymbol{x}'\|_K \quad \text{subject to} \quad \frac{1}{m}\|A\boldsymbol{x}' - \boldsymbol{y}\|_1 \le \varepsilon. \tag{1.17}$$

Then

$$\mathbb{E} \sup_{\boldsymbol{x} \in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \sqrt{8\pi} \left(\frac{w(K)}{\sqrt{m}} + \varepsilon \right).$$

Proof. It suffices to check that $\hat{\boldsymbol{x}} \in K$; the conclusion would then follow from Theorem 6.1. Note first that by choice of $\hat{\boldsymbol{x}}$ we have $\frac{1}{m}\|A\hat{\boldsymbol{x}} - \boldsymbol{y}\|_1 \le \varepsilon$, and by assumption (1.14) we have $\frac{1}{m}\|A\boldsymbol{x} - \boldsymbol{y}\|_1 = \frac{1}{m}\|\boldsymbol{v}\|_1 \le \varepsilon$. Thus both $\hat{\boldsymbol{x}}$ and $\boldsymbol{x}$ satisfy the constraint in (1.17). Therefore, by choice of $\hat{\boldsymbol{x}}$, we have

$$\|\hat{\boldsymbol{x}}\|_K \le \|\boldsymbol{x}\|_K \le 1;$$

the last inequality is nothing else than our assumption that $\boldsymbol{x} \in K$. It follows $\hat{\boldsymbol{x}} \in K$ as claimed. □

The remarks about algorithmic aspects of estimation made in Sections 1.4.3 and 1.4.4 apply also to the results of this section. In particular, the estimation from noisy linear observations (1.14) can be formulated as a *convex program*.

1.7 Applications to sparse recovery and regression

Remarkable examples of feasible sets K with low complexity come from the notion of *sparsity*. Consider the set K of all unit s-sparse vectors in $\mathbb{R}^n$. As we mentioned in Example 3.9, the mean width of K is

$$w(K) \sim s \log(n/s).$$

According to the interpretation we discussed in Section 1.4.4, this means that the *effective dimension* of K is of order $s \log(n/s)$. Therefore,

$$m \sim s \log(n/s)$$

observations should suffice to estimate any s-sparse vector in $\mathbb{R}^n$. Results of this type form the core of *compressed sensing*, a young area of signal processing, see [15, 19, 26, 39].

In this section, we consider a more general model, where an unknown vector $\boldsymbol{x}$ has a sparse representation *in some dictionary*.

We will specialize Theorem 6.2 to the sparse recovery problem. The convex program will in this case amount to minimizing the ℓ_1 norm of the coefficients. We will note that the notion of sparsity can be relaxed to accommodate approximate, or "effective," sparsity. Finally, we will observe that the estimate $\hat{\boldsymbol{x}}$ is most often unique and m-sparse.

1.7.1 Sparse recovery for general dictionaries

Let us fix a *dictionary* of vectors $\boldsymbol{d}_1, \dots, \boldsymbol{d}_N \in \mathbb{R}^n$, which may be arbitrary (even linearly dependent). The choice of a dictionary depends on the application; common examples include unions of orthogonal bases and more generally tight frames (in particular, Gabor frames). See [18, 20, 21, 62] for an introduction to sparse recovery problems with general dictionaries.

Suppose an unknown vector $\boldsymbol{x} \in \mathbb{R}^n$ is *s-sparse in the dictionary* $\{\boldsymbol{d}_i\}$. This means that $\boldsymbol{x}$ can be represented as a linear combination of at most s dictionary elements, i.e.,

$$\boldsymbol{x} = \sum_{i=1}^{N} \alpha_i \boldsymbol{d}_i \quad \text{with at most } s \text{ nonzero coefficients } \alpha_i \in \mathbb{R}. \tag{1.18}$$

As in Section 1.6, our goal is to recover $\boldsymbol{x}$ from a noisy observation vector $\boldsymbol{y} \in \mathbb{R}^m$ of the form

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\nu}, \qquad \frac{1}{m}\|\boldsymbol{\nu}\|_1 = \frac{1}{m}\sum_{i=1}^{m} |\nu_i| \le \varepsilon.$$

Recall that A is a known $m \times n$ Gaussian matrix, and $\boldsymbol{\nu}$ is an unknown noise vector, which can have arbitrary structure (in particular, correlated with A).

Theorem 6.2 will quickly imply the following recovery result.

Theorem 7.1 (Sparse recovery: general dictionaries). *Assume for normalization that all dictionary vectors satisfy $\|\boldsymbol{d}_i\|_2 \le 1$. Choose $\hat{\boldsymbol{x}}$ to be a solution to the convex program*

$$\text{minimize} \|\boldsymbol{\alpha}'\|_1 \text{ such that } \boldsymbol{x}' = \sum_{i=1}^{N} \alpha_i' \boldsymbol{d}_i \text{ satisfies } \frac{1}{m}\|A\boldsymbol{x}' - \boldsymbol{y}\|_1 \le \varepsilon. \tag{1.19}$$

Then

$$\mathbb{E}\,\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le C\sqrt{\frac{s\log N}{m}} \cdot \|\boldsymbol{\alpha}\|_2 + \sqrt{2\pi}\,\varepsilon.$$

Proof. Consider the sets

$$\bar{K} := \operatorname{conv}\{\pm \boldsymbol{d}_i\}_{i=1}^N, \quad K := \|\boldsymbol{\alpha}\|_1 \cdot \bar{K}.$$

Representation (1.18) implies that $\boldsymbol{x} \in K$, so it makes sense to apply Theorem 6.2 for K.

Let us first argue that the optimization program in Theorem 6.2 can be written in the form (1.19). Observe that we can replace $\|\boldsymbol{x}'\|_K$ by $\|\boldsymbol{x}'\|_{\bar{K}}$ in the optimization problem (1.17) without changing its solution. (This is because $\|\boldsymbol{x}'\|_{\bar{K}} = \|\boldsymbol{\alpha}\|_1 \cdot \|\boldsymbol{x}'\|_K$ and $\|\boldsymbol{\alpha}\|_1$ is a constant value.) Now, by definition of $\bar{K}$, we have

$$\|\boldsymbol{x}'\|_{\bar{K}} = \min\Big\{\|\boldsymbol{\alpha}'\|_1 : \ \boldsymbol{x}' = \sum_{i=1}^N \alpha_i' \boldsymbol{d}_i\Big\}.$$

Therefore, the optimization programs (1.17) and (1.19) are indeed equivalent.

Next, to evaluate the error bound in Theorem 6.2, we need to bound the mean width of K. The convexification property (1.5) and Example 3.8 yield

$$w(K) = \|\boldsymbol{\alpha}\|_1 \cdot w(\bar{K}) \le C\|\boldsymbol{\alpha}\|_1 \cdot \sqrt{\log N}.$$

Putting this into the conclusion of Theorem 6.2, we obtain the error bound

$$\mathbb{E}\sup_{\boldsymbol{x}\in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \sqrt{8\pi}\ C\sqrt{\frac{\log N}{m}} \cdot \|\boldsymbol{\alpha}\|_1 + \sqrt{2\pi}\,\varepsilon.$$

To complete the proof, it remains to note that

$$\|\boldsymbol{\alpha}\|_1 \le \sqrt{s} \cdot \|\boldsymbol{\alpha}\|_2, \tag{1.20}$$

since $\boldsymbol{\alpha}$ is s-sparse, i.e., it has only s nonzero coordinates. □

1.7.2 Remarkable properties of sparse recovery

Let us pause to look more closely at the statement of Theorem 7.1.

1.7.2.1 General dictionaries

Theorem 7.1 is very flexible with respect to the choice of a dictionary $\{\boldsymbol{d}_i\}$. Note that there are essentially *no restrictions on the dictionary*. (The normalization assumption $\|\boldsymbol{d}_i\|_2 \le 1$ can be dispensed of at the cost of increasing the error bound by the factor of $\max_i \|\boldsymbol{d}_i\|_2$.) In particular, the dictionary may be *linearly dependent*.

1.7.2.2 Effective sparsity

The reader may have noticed that the proof of Theorem 7.1 used sparsity in a quite mild way, only through inequality (1.20). So the result is still true for vectors $\boldsymbol{x}$ that are *approximately sparse* in the dictionary. Namely, Theorem 7.1 will hold if we replace the exact notion of sparsity (the number of nonzero coefficients) by the more flexible notion of *effective sparsity*, defined as

$$\text{effective sparsity}(\boldsymbol{\alpha}) := (\|\boldsymbol{\alpha}\|_1/\|\boldsymbol{\alpha}\|_2)^2.$$

It is now clear how to extend sparsity in a dictionary (1.18) to approximate sparsity. We can say that a vector $\boldsymbol{x}$ is *effectively s-sparse in a dictionary* $\{d_i\}$ if it can be represented as $\boldsymbol{x} = \sum_{i=1}^N \alpha_i \boldsymbol{d}_i$ where the coefficient vector $\boldsymbol{a} = (\alpha_1, \dots, \alpha_N)$ is effectively s-sparse.

The effective sparsity is clearly bounded by the exact sparsity, and it is robust with respect to small perturbations.

1.7.2.3 Linear programming

The convex programs (1.19) and (1.22) can be reformulated as linear programs. This can be done by introducing new variables $u_1, \dots, u_N$; instead of minimizing $\|\boldsymbol{\alpha}'\|_1$ in (1.19), we can equivalently minimize the linear function $\sum_{i=1}^N u_i$ subject to the additional linear constraints $-u_i \le \alpha_i' \le u_i$, $i = 1, \dots, N$. In a similar fashion, one can replace the convex constraint $\frac{1}{m}\|A\boldsymbol{x}' - \boldsymbol{y}\|_1 \le \varepsilon$ in (1.19) by n linear constraints.

1.7.2.4 Estimating the coefficients of sparse representation

It is worthwhile to notice that as a result of solving the convex recovery program (1.19), we obtain not only an estimate $\hat{\boldsymbol{x}}$ of the vector $\boldsymbol{x}$ but also an estimate $\hat{\boldsymbol{\alpha}}$ of the *coefficient vector* in the representation $\boldsymbol{x} = \sum \alpha_i \boldsymbol{d}_i$.

1.7.2.5 Sparsity of solution

The solution of the sparse recovery problem (1.19) may not be exact in general, that is, $\hat{\boldsymbol{x}} \neq \boldsymbol{x}$ can happen. This can be due to several factors—the generality of the dictionary, approximate (rather than exact) sparsity of $\boldsymbol{x}$ in the dictionary, and the noise ν in the observations. But even in this general situation, *the solution* $\boldsymbol{x}$ *is still m-sparse*, in all but degenerate cases. We will now state and prove this known fact (see [26]).

Proposition 7.2 (Sparsity of solution). *Assume that a given convex recovery program* (1.19) *has a unique solution* $\hat{\boldsymbol{\alpha}}$ *for the coefficient vector. Then* $\hat{\boldsymbol{\alpha}}$ *is m-sparse, and consequently* $\hat{\boldsymbol{x}}$ *is m-sparse in the dictionary* $\{\boldsymbol{d}_i\}$*. This is true even in presence of noise in observations, and even when no sparsity assumptions on* $\boldsymbol{x}$ *are in place.*

Proof. The result follows by simple dimension considerations. First note that the constraint on $\boldsymbol{\alpha}'$ in the optimization problem (1.19) can be written in the form

$$\frac{1}{m}\|AD\boldsymbol{\alpha}' - \boldsymbol{y}\|_1 \leq \varepsilon, \tag{1.21}$$

where D is the $n \times N$ matrix whose columns are the dictionary vectors $\boldsymbol{d}_i$. Since matrix AD has dimensions $m \times N$, the constraint defines a cylinder in $\mathbb{R}^N$ whose infinite directions are formed by the kernel of AD, which has dimension at least $N - m$. Moreover, this cylinder is a polyhedral set (due to the ℓ_1 norm defining it), so it has no faces of dimension smaller than $N - m$.

On the other hand, the level sets of the objective function $\|\boldsymbol{\alpha}'\|_1$ are also polyhedral sets; they are dilates of the unit ℓ_1 ball. The solution $\hat{\boldsymbol{\alpha}}$ of the optimization problem (1.19) is thus a point in $\mathbb{R}^N$ where the smallest dilate of the ℓ_1 ball touches the cylinder. The uniqueness of solution means that a touching point is unique. This is illustrated in Figure 1.7.

Consider the faces of these two polyhedral sets of smallest dimensions that contain the touching point; we may call these the touching faces. The touching face of the cylinder has dimension at least $N-m$, as all of its faces do. Then the touching

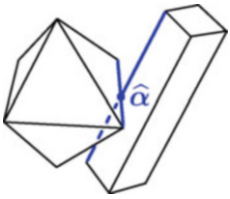


Fig. 1.7 Illustration for the proof of Proposition 7.2. The polytope on the left represents a level set of the ℓ_1 ball. The cylinder on the right represents the vectors $\boldsymbol{\alpha}'$ satisfying constraint (1.21). The two polyhedral sets touch at point $\hat{\boldsymbol{\alpha}}$.

face of the ℓ_1 ball must have dimension at most m, otherwise the two touching faces would intersect by more than one point. This translates into the m-sparsity of the solution $\hat{\boldsymbol{\alpha}}$, as claimed. □

In view of Proposition 7.2, we can ask when the solution $\hat{\boldsymbol{\alpha}}$ of the convex program (1.19) is unique. This does not always happen; for example, this fails if $\boldsymbol{d}_1 = \boldsymbol{d}_2$.

Uniqueness of solutions of optimization problems like (1.19) is extensively studied [26]. Let us mention here a cheap way to obtain uniqueness. This can be achieved by an arbitrarily small generic perturbation of the dictionary elements, such as adding a small independent Gaussian vector to each $\boldsymbol{d}_i$. Then one can see that the solution $\hat{\boldsymbol{\alpha}}$ (and therefore $\hat{\boldsymbol{x}}$ as well) are unique almost surely. Invoking Proposition 7.2 we see that $\hat{\boldsymbol{x}}$ is m-sparse in the perturbed dictionary.

1.7.3 Sparse recovery for the canonical dictionary

Let us illustrate Theorem 7.1 for the simplest example of a dictionary—the canonical basis of $\mathbb{R}^n$:

$$\{\boldsymbol{d}_i\}_{i=1}^n = \{\boldsymbol{e}_i\}_{i=1}^n.$$

In this case, our assumption is that an unknown vector $\boldsymbol{x} \in \mathbb{R}^n$ is *s-sparse* in the usual sense, meaning that $\boldsymbol{x}$ has at most s nonzero coordinates, or *effectively s-sparse* as in Section 1.7.2.2. Theorem 7.1 then reads as follows.

Corollary 7.3 (Sparse recovery). *Choose $\hat{\boldsymbol{x}}$ to be a solution to the convex program*

$$\text{minimize} \|\boldsymbol{x}'\|_1 \text{ subject to } \frac{1}{m}\|A\boldsymbol{x}' - \boldsymbol{y}\|_1 \le \varepsilon. \tag{1.22}$$

Then

$$\mathbb{E}\,\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le C\sqrt{\frac{s\log n}{m}} \cdot \|\boldsymbol{x}\|_2 + \sqrt{2\pi}\,\varepsilon. \qquad \square$$

Sparse recovery results like Corollary 7.3 form the core of the area of *compressed sensing*, see [15, 19, 26, 39].

In the noiseless case ($\varepsilon = 0$) and for sparse (rather than effectively sparse) vectors, one may even hope to recover $\boldsymbol{x}$ *exactly*, meaning that $\hat{\boldsymbol{x}} = \boldsymbol{x}$ with high probability. Conditions for exact recovery are now well understood in compressed sensing. We will discuss some exact recovery problems in Section 1.9.

We can summarize Theorem 7.1 and the discussion around it as follows:

Using linear programming, one can approximately recover a vector $\boldsymbol{x}$ that is s-sparse (or effectively s-sparse) in a general dictionary of size N, from $m \sim s\log N$ random linear observations.

1.7.4 Application: linear regression with constraints

The noisy estimation problem (1.14) is equivalent to *linear regression* with constraints. So in this section, we will translate the story into the statistical language. We present here just one class of examples out of a wide array of statistical problems; we refer the reader to [11, 74] for a recent review of high-dimensional estimation problems from a statistical viewpoint.

Linear regression is a model of linear relationship between one dependent variable and n explanatory variables. It is usually written as

$$\boldsymbol{y} = X\boldsymbol{\beta} + \boldsymbol{\nu}.$$

Here X is an $n \times p$ matrix which contains a sample of n observations of p explanatory variables; $\boldsymbol{y} \in \mathbb{R}^n$ represents a sample of n observations of the dependent variable; $\boldsymbol{\beta} \in \mathbb{R}^p$ is a coefficient vector; $\boldsymbol{\nu} \in \mathbb{R}^n$ is a noise vector. We assume that X and $\boldsymbol{y}$ are known, while $\boldsymbol{\beta}$ and $\boldsymbol{\nu}$ are unknown. Our goal is to estimate $\boldsymbol{\beta}$.

We discussed a classical formulation of linear regression. In addition, we often know, believe, or want to enforce some properties about the coefficient vector $\boldsymbol{\beta}$ (for example, sparsity). We can express such extra information as the assumption that

$$\boldsymbol{\beta} \in K$$

where $K \subset \mathbb{R}^p$ is a known feasible set. Such problem may be called a *linear regression with constraints*.

The high-dimensional estimation results we have seen so far can be translated into the language of regression in a straightforward way. Let us do this for Theorem 6.2; the interested reader can make a similar translation or other results.

We assume that the explanatory variables are independent $N(0, 1)$, so the matrix X has all i.i.d. $N(0, 1)$ entries. This requirement may be too strong in practice; however see Section 1.8 on relaxing this assumption. The noise vector $\boldsymbol{\nu}$ is allowed to have arbitrary structure (in particular, it can be correlated with X). We assume that its magnitude is controlled:

$$\frac{1}{n}\|\boldsymbol{\nu}\|_1 = \frac{1}{n}\sum_{i=1}^{n} |\nu_i| \le \varepsilon$$

for some known noise level ε. Then we can restate Theorem 6.2 in the following way.

Theorem 7.4 (Linear regression with constraints). *Choose $\hat{\boldsymbol{\beta}}$ to be a solution to the program*

$$\text{minimize} \|\boldsymbol{\beta}'\|_K \quad \text{subject to} \quad \frac{1}{n}\|X\boldsymbol{\beta}' - \boldsymbol{y}\|_1 \le \varepsilon.$$

Then

$$\mathbb{E} \sup_{\boldsymbol{\beta} \in K} \|\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}\|_2 \le \sqrt{8\pi} \left(\frac{w(K)}{\sqrt{n}} + \varepsilon \right). \qquad \square$$

1.8 Extensions from Gaussian to sub-Gaussian distributions

So far, all our results were stated for Gaussian distributions. Let us show how to relax this assumption. In this section, we will modify the proof of the M^* bound, Theorem 5.1, for general *sub-Gaussian* distributions, and indicate the consequences for the estimation problem. A result of this type was proved in [47] with a much more complex argument.

1.8.1 Sub-Gaussian random variables and random vectors

A systematic introduction into sub-Gaussian distributions can be found in Sections 5.2.3 and 5.2.5 of [73]; here we briefly mention the basic definitions. According to one of the several equivalent definitions, a random variable X is *sub-Gaussian* if

$$\mathbb{E} \exp(X^2/\psi^2) \le e$$

for some $\psi > 0$. The smallest ψ is called the *sub-Gaussian norm* and is denoted $\|X\|_{\psi_2}$. Normal and all bounded random variables are sub-Gaussian, while exponential random variables are not.

The notion of sub-Gaussian distribution transfers to higher dimensions as follows. A random vector $\boldsymbol{X} \in \mathbb{R}^n$ is called sub-Gaussian if all one-dimensional marginals $\langle \boldsymbol{X}, \boldsymbol{u} \rangle$, $\boldsymbol{u} \in \mathbb{R}^n$, are sub-Gaussian random variables. The sub-Gaussian norm of $\boldsymbol{X}$ is defined as

$$\|\boldsymbol{X}\|_{\psi_2} := \sup_{\boldsymbol{u} \in S^{n-1}} \| \langle \boldsymbol{X}, \boldsymbol{u} \rangle \|_{\psi_2} \tag{1.23}$$

where, as before, S^{n-1} denotes the Euclidean sphere in $\mathbb{R}^n$. Recall also that the random vector $\boldsymbol{X}$ is called *isotropic* if

$$\mathbb{E} \boldsymbol{X}\boldsymbol{X}^{\mathsf{T}} = I_n.$$

Isotropy is a scaling condition; any distribution in $\mathbb{R}^n$ which is not supported in a low-dimensional subspace can be made isotropic by an appropriate linear transformation. To illustrate this notion with a couple of quick examples, one can check that $N(0, I_n)$ and the uniform distribution on the discrete cube $\{-1, 1\}^n$ are isotropic and sub-Gaussian distributions.

1.8.2 M* bound for sub-Gaussian distributions

Now we state and prove a version of M^* bound, Theorem 5.1, for general *sub-Gaussian* distributions. It is a variant of a result from [47].

Theorem 8.1 (General M^* bound for sub-Gaussian distributions). *Let T be a bounded subset of $\mathbb{R}^n$. Let A be an $m \times n$ matrix whose rows $\boldsymbol{a}_i$ are i.i.d., mean zero, isotropic, and sub-Gaussian random vectors in $\mathbb{R}^n$. Choose $\psi \geq 1$ so that*

$$\|\boldsymbol{a}_i\|_{\psi_2} \leq \psi, \quad i = 1, \dots, m. \tag{1.24}$$

Fix $\varepsilon \geq 0$ and consider the set

$$T_\varepsilon := \Big\{ \boldsymbol{u} \in T : \ \frac{1}{m} \|A\boldsymbol{u}\|_1 \leq \varepsilon \Big\}.$$

Then

$$\mathbb{E} \sup_{\boldsymbol{u} \in T_\varepsilon} \|\boldsymbol{u}\|_2 \leq C\psi^4 \Big(\frac{1}{\sqrt{m}} \ \mathbb{E} \sup_{\boldsymbol{u} \in T} |\ \langle \boldsymbol{g}, \boldsymbol{u} \rangle\ | + \varepsilon \Big),$$

where $g \sim N(0, I_n)$ is a standard Gaussian random vector in $\mathbb{R}^n$.

A proof of this result is an extension of the proof of the Gaussian M^* bound, Theorem 5.1. Most of that argument generalizes to sub-Gaussian distributions in a standard way. The only nontrivial new step will be based on the deep *comparison theorem for sub-Gaussian processes* due to X. Fernique and M. Talagrand, see [71, Section 2.1]. Informally, the result states that any sub-Gaussian process is dominated by a Gaussian process with the same (or larger) increments.

Theorem 8.2 (Fernique–Talagrand's comparison theorem). *Let T be an arbitrary set.*[4] *Consider a Gaussian random process $(G(\boldsymbol{t}))_{\boldsymbol{t} \in T}$ and a sub-Gaussian random process $(H(\boldsymbol{t}))_{\boldsymbol{t} \in T}$. Assume that $\mathbb{E}\, G(\boldsymbol{t}) = \mathbb{E}\, H(\boldsymbol{t}) = 0$ for all $\boldsymbol{t} \in T$. Assume also that for some $M > 0$, the following increment comparison holds:*[5]

$$\|H(\boldsymbol{s}) - H(\boldsymbol{t})\|_{\psi_2} \leq M \, (\mathbb{E}\, \|G(\boldsymbol{s}) - G(\boldsymbol{t})\|_2^2)^{1/2} \quad \text{for all } \boldsymbol{s}, \boldsymbol{t} \in T.$$

Then

$$\mathbb{E} \sup_{\boldsymbol{t} \in T} H(\boldsymbol{t}) \leq CM \ \mathbb{E} \sup_{\boldsymbol{t} \in T} G(\boldsymbol{t}).$$

[4] We can assume T to be finite to avoid measurability complications and then proceed by approximation; see, e.g., [43, Section 2.2].

[5] The increment comparison may look better if we replace the L_2 norm on the right-hand side by ψ_2 norm. Indeed, it is easy to see that $\|G(\boldsymbol{s}) - G(\boldsymbol{t})\|_{\psi_2} \asymp (\mathbb{E}\, \|G(\boldsymbol{s}) - G(\boldsymbol{t})\|_2^2)^{1/2}$.

This theorem is a combination of a result of X. Fernique [25] that bounds $\mathbb{E}\sup_{t\in T} H(t)$ above by the so-called *majorizing measure* of T, and a result of M. Talagrand [70] that bounds $\mathbb{E}\sup_{t\in T} G(t)$ below by the same majorizing measure of T.

Proof of Theorem 8.1. Let us examine the proof of the Gaussian M^* bound, Theorem 5.1, check where we used Gaussian assumptions, and try to accommodate sub-Gaussian assumptions instead.

The first such place is identity (1.12). We claim that a version of it still holds for the sub-Gaussian random vector $\boldsymbol{a}$, namely

$$\|\boldsymbol{u}\|_2 \le C_0\psi^3\ \mathbb{E}_{\boldsymbol{a}}\,|\,\langle \boldsymbol{a},\boldsymbol{u}\rangle\,| \tag{1.25}$$

where C_0 is an absolute constant.[6]

To check (1.25), we can assume that $\|\boldsymbol{u}\|_2 = 1$ by dividing both sides by $\|\boldsymbol{u}\|_2$ if necessary. Then $Z := \langle \boldsymbol{a},\boldsymbol{u}\rangle$ is sub-Gaussian random variable, since according to (1.23) and (1.24), we have $\|Z\|_{\psi_2} \le \|\boldsymbol{a}\|_{\psi_2} \le \psi$. Then, since sub-Gaussian distributions have moments of all orders (see [73, Lemma 5.5]), we have $(\mathbb{E}\,Z^3)^{1/3} \le C_1\|Z\|_{\psi_2} \le C_1\psi$, where C_1 is an absolute constant. Using this together with isotropy and Cauchy–Schwarz inequality, we obtain

$$1 = \mathbb{E}\,Z^2 = \mathbb{E}\,Z^{1/2}Z^{3/2} \le (\mathbb{E}\,Z)^{1/2}(\mathbb{E}\,Z^3)^{1/2} \le (\mathbb{E}\,Z)^{1/2}(C_1\psi)^{3/2}.$$

Squaring both sides implies (1.25), since we assumed that $\|\boldsymbol{u}\|_2 = 1$.

The next steps in the proof of Theorem 5.1—symmetrization and contraction—go through for sub-Gaussian distributions without change. So (1.13) is still valid in our case.

Next, the random vector

$$\boldsymbol{h} := \frac{1}{\sqrt{m}}\sum_{i=1}^{m}\varepsilon_i\boldsymbol{a}_i$$

is no longer Gaussian as in the proof of Theorem 5.1. Still, $\boldsymbol{h}$ is sub-Gaussian with

$$\|\boldsymbol{h}\|_{\psi_2} \le C_2\psi \tag{1.26}$$

due to the approximate rotation invariance of sub-Gaussian distributions, see [73, Lemma 5.9].

In the last step of the argument, we need to replace the sub-Gaussian random vector $\boldsymbol{h}$ by the Gaussian random vector $\boldsymbol{g} \sim N(0, I_n)$, i.e., prove an inequality of the form

[6] We should mention that a reverse inequality also holds: by isotropy, one has $\mathbb{E}_{\boldsymbol{a}}\,|\,\langle \boldsymbol{a},\boldsymbol{u}\rangle\,| \le (\mathbb{E}_{\boldsymbol{a}}\,\langle \boldsymbol{a},\boldsymbol{u}\rangle^2)^{1/2} = \|\boldsymbol{u}\|_2$. However, this inequality will not be used in the proof.

$$\mathbb{E} \sup_{\boldsymbol{u} \in T} |\langle \boldsymbol{h}, \boldsymbol{u} \rangle| \lesssim \mathbb{E} \sup_{\boldsymbol{u} \in T} |\langle \boldsymbol{g}, \boldsymbol{u} \rangle|.$$

This can be done by applying the comparison inequality of Theorem 8.2 for the processes

$$H(\boldsymbol{u}) = \langle \boldsymbol{h}, \boldsymbol{u} \rangle \quad \text{and} \quad G(\boldsymbol{u}) = \langle \boldsymbol{g}, \boldsymbol{u} \rangle, \quad \boldsymbol{u} \in T \cup (-T).$$

To check the increment inequality, we can use (1.26), which yields

$$\|H(\boldsymbol{u}) - H(\boldsymbol{v})\|_{\psi_2} = \| \langle \boldsymbol{h}, \boldsymbol{u} - \boldsymbol{v} \rangle \|_{\psi_2} \le \|\boldsymbol{h}\|_{\psi_2} \|\boldsymbol{u} - \boldsymbol{v}\|_2 \le C_2 \psi \|\boldsymbol{u} - \boldsymbol{v}\|_2.$$

On the other hand,

$$(\mathbb{E} \|G(\boldsymbol{u}) - G(\boldsymbol{v})\|_2^2)^{1/2} = \|\boldsymbol{u} - \boldsymbol{v}\|_2.$$

Therefore, the increment inequality in Theorem 8.2 holds with $M = C_2 \psi$. It follows that

$$\mathbb{E} \sup_{\boldsymbol{u} \in T \cup (-T)} \langle \boldsymbol{h}, \boldsymbol{u} \rangle \le C_3 \psi \ \mathbb{E} \sup_{\boldsymbol{u} \in T \cup (-T)} \langle \boldsymbol{g}, \boldsymbol{u} \rangle.$$

This means that

$$\mathbb{E} \sup_{\boldsymbol{u} \in T} |\langle \boldsymbol{h}, \boldsymbol{u} \rangle| \le C_3 \psi \ \mathbb{E} \sup_{\boldsymbol{u} \in T} |\langle \boldsymbol{g}, \boldsymbol{u} \rangle|$$

as claimed.

Replacing all Gaussian inequalities by their sub-Gaussian counterparts discussed above, we complete the proof just like in Theorem 5.1. □

1.8.3 *Estimation from sub-Gaussian linear observations*

It is now straightforward to generalize all recovery results we developed before from Gaussian to sub-Gaussian observations. So our observations are now

$$y_i = \langle \boldsymbol{a}_i, x \rangle + \boldsymbol{v}_i, \quad i = 1, \dots, m$$

where $\boldsymbol{a}_i$ are i.i.d., mean zero, isotropic, and *sub-Gaussian* random vectors in $\mathbb{R}^n$. As in Theorem 8.1, we control the sub-Gaussian norm with the parameter $\psi > 1$, choosing it so that

$$\|\boldsymbol{a}_i\|_{\psi_2} \le \psi, \quad i = 1, \dots, m.$$

We can write observations in the matrix form as in (1.14), i.e.,

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\nu},$$

where A is the $m \times n$ matrix with rows $\boldsymbol{a}_i$. As before, we assume some control on the error:

$$\frac{1}{m}\|\boldsymbol{\nu}\|_1 = \frac{1}{m}\sum_{i=1}^{m} |\nu_i| \le \varepsilon.$$

Let us state a version of Theorem 6.1 for sub-Gaussian observations. Its proof is the same, except we use the sub-Gaussian M^* bound, Theorem 8.1, where previously a Gaussian M^* bound was used.

Theorem 8.3 (Estimation from sub-Gaussian observations). *Choose $\hat{\boldsymbol{x}}$ to be any vector satisfying*

$$\hat{\boldsymbol{x}} \in K \quad \text{and} \quad \frac{1}{m}\|A\hat{\boldsymbol{x}} - \boldsymbol{y}\|_1 \le \varepsilon.$$

Then

$$\mathbb{E} \sup_{\boldsymbol{x} \in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le C\psi^4 \left(\frac{w(K)}{\sqrt{m}} + \varepsilon \right). \square$$

In a similar fashion, one can generalize all other estimation results established before to sub-Gaussian observations. We leave this to the interested reader.

1.9 Exact recovery

In some situations, one can hope to estimate vector $\boldsymbol{x} \in K$ from $\boldsymbol{y}$ *exactly*, without any error. Such results form the core of the area of *compressed sensing* [19, 26, 39]. Here we will present an approach to exact recovery based on Y. Gordon's "escape through a mesh" theorem [33]. This argument goes back to [66] for the set of sparse vectors, it was further developed in [53, 69] and was pushed forward for general feasible sets in [2, 16, 72].

In this tutorial we will present the most basic result; the reader will find a more complete picture and many more examples in the papers just cited.

We will work here with Gaussian observations

$$\boldsymbol{y} = A\boldsymbol{x},$$

where A is an $m \times n$ Gaussian random matrix. This is the same model as we considered in Section 1.4.

1.9.1 Exact recovery condition and the descent cone

When can $\boldsymbol{x}$ be inferred from $\boldsymbol{y}$ exactly? Recall that we only know two things about $\boldsymbol{x}$—that it lies in the feasible set K and in the affine subspace

$$E_{\boldsymbol{x}} := \{\boldsymbol{x}' : A\boldsymbol{x}' = \boldsymbol{y}\}.$$

This two pieces of information determine $\boldsymbol{x}$ uniquely if and only if these two sets intersect at the single point $\boldsymbol{x}$:

$$K \cap E_{\boldsymbol{x}} = \{x\}. \tag{1.27}$$

Notice that this situation would go far beyond the M^* bound on the diameter of $K \cap E$ (see Theorem 3.12)—indeed, in this case the diameter would equal zero!

How can this be possible? Geometrically, the exact recovery condition (1.27) states that *the affine subspace $E_{\boldsymbol{x}}$ is tangent to the set K at the point $\boldsymbol{x}$*; see Figure 1.8a for illustration.

This condition is local. Assuming that K is convex for better understanding, we see that the tangency condition depends on the shape of K in an infinitesimal neighborhood of $\boldsymbol{x}$, while the global geometry of K is irrelevant. So we would not lose anything if we replace K by the *descent cone* at point x, see Figure 1.8b. This set is formed by the rays emanating from $\boldsymbol{x}$ into directions of points from K:

$$D(K, \boldsymbol{x}) := \{t(z - \boldsymbol{x}) : z \in K,\ t \geq 0\}.$$

Translating by $-\boldsymbol{x}$, can we rewrite the exact recovery condition (1.27) as

$$(K - \boldsymbol{x}) \cap (E_{\boldsymbol{x}} - \boldsymbol{x}) = \{0\}.$$

Replacing $K - \boldsymbol{x}$ by the descent cone (a bigger set) and noting that $E_{\boldsymbol{x}} - \boldsymbol{x} = \ker(A)$, we rewrite this again as

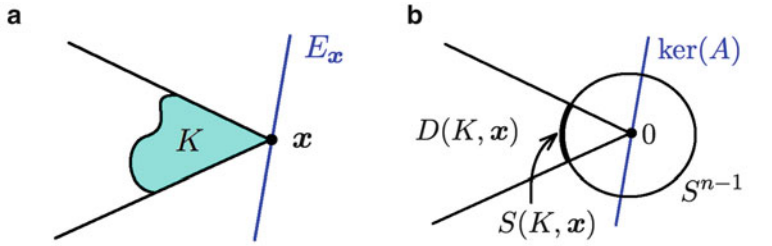
$$D(K, \boldsymbol{x}) \cap \ker(A) = \{0\}.$$

The descent cone can be determined by its intersection with the unit sphere, i.e., by[7]

$$S(K, x) := D(K, \boldsymbol{x}) \cap S^{n-1} = \left\{ \frac{z - \boldsymbol{x}}{\|z - \boldsymbol{x}\|_2} : z \in K \right\}. \tag{1.28}$$

Thus we arrive at the following equivalent form of the exact recovery condition (1.27):

[7] In definition (1.28), we adopt the convention that $0/0 = 0$.



Exact recovery condition (9.1): affine subspace E_x is tangent to K at x

Picture translated by $-x$: subspace $\ker(A)$ is tangent to descent cone $D(K,x)$ at 0

Fig. 1.8 Illustration of the exact recovery condition (1.27)

$$S(K,x) \cap \ker(A) = \varnothing;$$

see Figure 1.8b for an illustration.

1.9.2 Escape through a mesh and implications for exact recovery

It remains to understand under what conditions the random subspace $\ker A$ misses a given subset $S = S(K,x)$ of the unit sphere. There is a remarkably sharp result in asymptotic convex geometry that answers this question for general subsets S. This is the theorem on *escape through a mesh*, which is due to Y. Gordon [33]. Similarly to the other results we saw before, this theorem depends on the *mean width* of S, defined as[8]

$$\bar{w}(S) = \mathbb{E} \sup_{u \in S} \langle g, u \rangle, \quad \text{where} \quad g \sim N(0, I_n).$$

Theorem 9.1 (Escape through a mesh). *Let S be a fixed subset of S^{n-1}. Let E be a random subspace of $\mathbb{R}^n$ of a fixed codimension m, drawn from the Grassmanian $G_{n,n-m}$ according to the Haar measure. Assume that*

$$\bar{w}(S) < \sqrt{m}.$$

[8]The only (minor) difference with our former definition (1.3) of the mean width is that we take supremum over S instead of $S - S$, so $\bar{w}(S)$ is a smaller quantity. The reason we do not need to consider $S - S$ because we already subtracted x in the definition of the descent cone.

Then

$$S \cap E = \varnothing$$

with high probability, namely $1 - 2.5 \exp\big[- (m/\sqrt{m+1} - \bar{w}(S))^2/18\big]$.

Before applying this result to high-dimensional estimation, let us see how a slightly weaker result follows from the general M^* bound, Theorem 5.1. Indeed, applying the latter theorem for $T = S$, $E = \ker(A)$, and $\varepsilon = 0$, we obtain

$$\mathbb{E} \sup_{\boldsymbol{u} \in S \cap E} \|\boldsymbol{u}\|_2 \le \sqrt{\frac{8\pi}{m}}\, \mathbb{E} \sup_{\boldsymbol{u} \in S} |\langle \boldsymbol{g}, \boldsymbol{u} \rangle| \le \sqrt{\frac{8\pi}{m}}\, \bar{w}(S). \tag{1.29}$$

Since $S \subset S^{n-1}$, the supremum on the left-hand side equals 1 when $S \cap E \neq \emptyset$ and zero otherwise. Thus the expectation in (1.29) equals $\mathbb{P}\big\{S \cap E \neq \emptyset\big\}$. Further, one can easily check that $\mathbb{E} \sup_{\boldsymbol{u} \in S} |\langle \boldsymbol{g}, \boldsymbol{u} \rangle| \le \bar{w}(S) + \sqrt{2/\pi}$, see [57, Proposition 2.1]. Thus we obtain

$$\mathbb{P}\big\{S \cap E \neq \emptyset\big\} \le \sqrt{\frac{8\pi}{m}}\Big(\bar{w}(S) + \sqrt{\frac{2}{\pi}}\Big).$$

In other words, $S \cap E = \emptyset$ with high probability if the codimension m is sufficiently large so that $\bar{w}(S) \ll \sqrt{m}$. Thus we obtain a somewhat weaker form of Escape Theorem 9.1.

Now let us apply Theorem 9.1 for the descent $S = S(K, x)$ and $E = \ker(A)$. We conclude by the argument above that the exact recovery condition (1.27) holds with high probability if

$$m > \bar{w}(S)^2.$$

How can we *algorithmically* recover $\boldsymbol{x}$ in these circumstances? We can do the same as in Section 1.4.1, either using the feasibility program (1.6) or, better yet, the optimization program (1.7). The only difference is that the diameter of the intersection is now zero, so the recovery is exact. The following is an exact version of Theorem 4.2.

Theorem 9.2 (Exact recovery from linear observations). *Choose* $\hat{\boldsymbol{x}}$ *to be a solution of the program*

$$\text{minimize} \|\boldsymbol{x}'\|_K \quad \text{subject to} \quad A\boldsymbol{x}' = \boldsymbol{y}.$$

Assume that the number of observations satisfies

$$m > \bar{w}(S)^2 \tag{1.30}$$

where $S = S(K, x)$ is the spherical part of the descent cone of K, defined in (1.28). Then

$$\hat{\boldsymbol{x}} = \boldsymbol{x}$$

with high probability (the same as in Theorem 9.1). □

Note the familiar condition (1.30) on m which we have seen before, see, e.g., Section 1.4.3. Informally, it states the following:

> *Exact recovery is possible when the number of measurements exceeds the effective dimension of the descent cone.*

Remarkably, condition (1.30) does not have absolute constant factors which we had in results before.

1.9.3 Application: exact sparse recovery

Let us illustrate how Theorem 9.2 works for *exact sparse recovery*. Assume that $\boldsymbol{x}$ is s-sparse, i.e. it has at most s nonzero coefficients. For the feasible set, we can choose $K := \|\boldsymbol{x}\|_1 B_1^n = \{\boldsymbol{x}' : \|\boldsymbol{x}'\|_1 \le \|\boldsymbol{x}\|_1\}$. One can write down accurately an expression for the descent cone and derive a familiar bound on the mean width of $S = S(K, x)$:

$$\bar{w}(S) \le C\sqrt{s \log(2n/s)}.$$

This computation goes back to [66]; see that paper and also [3, 16, 69] for estimates with explicit absolute constants.

We plug this into Theorem 9.2, where we replace $\|\boldsymbol{x}'\|_K$ in the optimization problem by the proportional quantity $\|\boldsymbol{x}'\|_1$. This leads to the following exact version of Corollary 7.3:

Theorem 9.3 (Exact sparse recovery). *Assume that an unknown vector $\boldsymbol{x} \in \mathbb{R}^n$ is s-sparse. Choose $\hat{\boldsymbol{x}}$ to be a solution to the convex program*

$$\text{minimize } \|\boldsymbol{x}'\|_1 \quad \text{subject to} \quad A\boldsymbol{x}' = \boldsymbol{y}.$$

Assume that the number of observations satisfies $m > Cs \log n$. Then

$$\hat{\boldsymbol{x}} = \boldsymbol{x}$$

with high probability, namely $1 - 3e^{-m}$. □

Due to the remarkable sharpness of Gordon's theorem, one may hope to obtain *sharp* conditions on the number of observations m, without any losses in absolute constants. This was done in [22] for the sparse recovery problem (using geometry

of polytopes rather than Gordon's theorem) and more recently in [3] for general feasible cones. The latter paper proposes a notion of statistical dimension, which is a close relative of mean width, and establishes a variant of Gordon's theorem for statistical dimension.

1.10 Low-rank matrix recovery and matrix completion

1.10.1 Background: matrix norms

The theory we developed so far concerns estimation of *vectors* in $\mathbb{R}^n$. It should not be surprising that this theory can also be applied for *matrices*. Matrix estimation problems were studied recently, in particular in [12–14, 37, 63].

Let us recall some basic facts about matrices and their norms. We can identify $d_1 \times d_2$ matrices with vectors in $\mathbb{R}^{d_1 \times d_2}$. The ℓ_2 norm in $\mathbb{R}^{d_1 \times d_2}$ is then nothing else than *Frobenius* (or Hilbert–Schmidt) norm of matrices:

$$\|X\|_F = \Big(\sum_{i=1}^{d_1} \sum_{j=1}^{d_2} |X_{ij}|^2 \Big)^{1/2}.$$

The inner product in $\mathbb{R}^{d_1 \times d_2}$ can be written in matrix form as follows:

$$\langle X, Y \rangle = \operatorname{tr}(X^\mathsf{T} Y).$$

Denote $d = \min(d_1, d_2)$. Let

$$s_1(X) \geq s_2(X) \geq \cdots \geq s_d(X) \geq 0$$

denote the *singular values* of X. Then Frobenius norm has the following spectral representation:

$$\|X\|_F = \Big(\sum_{i=1}^{d} s_i(X)^2 \Big)^{1/2}.$$

Recall also the *operator norm* of X, which is

$$\|X\| = \max_{\boldsymbol{u} \in \mathbb{R}^n \setminus \{0\}} \frac{\|X\boldsymbol{u}\|_2}{\|\boldsymbol{u}\|_2} = \max_{i=1,\ldots,d} s_i(X).$$

Finally, the *nuclear norm* of X is defined as

$$\|X\|_* = \sum_{i=1}^{d} s_i(X).$$

Spectrally, i.e., on the level of singular values, the nuclear norm is a version of ℓ_1 norm for matrices, the Frobenius norm is a version of ℓ_2 norm for matrices, and the operator norm is a version of ℓ_∞ norm for matrices. In particular, the following inequality holds:

$$\|X\| \le \|X\|_F \le \|X\|_*.$$

The reader should be able to derive many other useful inequalities in a similar way, for example,

$$\|X\|_* \le \sqrt{\operatorname{rank}(X)} \cdot \|X\|_F, \quad \|X\|_F \le \sqrt{\operatorname{rank}(X)} \cdot \|X\| \tag{1.31}$$

and

$$\langle X, Y \rangle \le \|X\| \cdot \|Y\|_*. \tag{1.32}$$

1.10.2 Low-rank matrix recovery

We are ready to formulate a matrix version of the sparse recovery problem from Section 1.7. Our goal is to estimate an unknown $d_1 \times d_2$ matrix X from m linear observations given by

$$y_i = \langle A_i, X \rangle, \quad i = 1, \dots, m. \tag{1.33}$$

Here A_i are independent $d_1 \times d_2$ Gaussian matrices with all i.i.d. $N(0, 1)$ entries.

There are two natural matrix versions of sparsity. The first version is the sparsity of entries. We will be concerned with the other, spectral, type of sparsity, where there are only a few nonzero singular values. This simply means that the matrix has *low rank*. So let us assume that the unknown matrix X satisfies

$$\operatorname{rank}(X) \le r \tag{1.34}$$

for some fixed (and possibly unknown) $r \le n$.

The following is a matrix version of Corollary 7.3; for simplicity we are stating it in a noise-free setting ($\varepsilon = 0$).

Theorem 10.1 (Low-rank matrix recovery). *Choose $\hat{X}$ to be a solution to the convex program*

$$\text{minimize} \|X'\|_* \text{subject to} \langle A_i, X' \rangle = y_i, \quad i = 1, \dots, m. \tag{1.35}$$

Then

$$\mathbb{E} \sup_X \|\hat{X} - X\|_F \le 4\sqrt{\pi} \sqrt{\frac{r(d_1 + d_2)}{m}} \cdot \|X\|_F.$$

Here the supremum is taken over all $d_1 \times d_2$ matrices X of rank at most r.

The proof of Theorem 10.1 will closely follow its vector prototype, that of Theorem 7.1; we will just need to replace the ℓ_1 norm by the nuclear norm. The only real difference will be in the computation of the *mean width of the unit ball of the nuclear norm*. This computation will be based on Y. Gordon's bound on the operator norm of Gaussian random matrices, see Theorem 5.32 in [73].

Theorem 10.2 (Gordon's bound for Gaussian random matrices). *Let G be a $d_1 \times d_2$ matrix whose entries are i.i.d., mean zero random variables. Then*

$$\mathbb{E} \|G\| \le \sqrt{d_1} + \sqrt{d_2}.$$

Proposition 10.3 (Mean width of the unit ball of nuclear norm). *Consider the unit ball in the space of $d_1 \times d_2$ matrices corresponding to the nuclear norm:*

$$B_* := \{X \in \mathbb{R}^{d_1 \times d_2} : \|X\|_* \le 1\}.$$

Then

$$w(B_*) \le 2(\sqrt{d_1} + \sqrt{d_2}).$$

Proof. By definition and symmetry of B, we have

$$w(B) = \mathbb{E} \sup_{X \in B_* - B_*} \langle G, X \rangle = 2\, \mathbb{E} \sup_{X \in B_*} \langle G, X \rangle ,$$

where G is a $d_1 \times d_2$ Gaussian random matrix with $N(0, 1)$ entries. Using inequality (1.32) and definition of B_*, we obtain

$$w(B_*) \le 2\, \mathbb{E} \sup_{X \in B_*} \|G\| \cdot \|X\|_* \le 2\, \mathbb{E} \|G\|.$$

(The reader may notice that both these inequalities are in fact equalities, although we do not need this in the proof.) To complete the proof, it remains to apply Theorem 10.2. □

Let us mention an immediate consequence of Proposition 10.3, although it will not be used in the proof of Theorem 10.1.

Proposition 10.4 (Mean width of the set of low-rank matrices). *Let*

$$D = \{X \in \mathbb{R}^{d_1 \times d_2} : \ \|X\|_F = 1, \ \operatorname{rank}(X) \le r\}.$$

Then

$$w(D) \le 2\sqrt{2r(d_1 + d_2)}.$$

Proof of Proposition 10.4. The bound follows immediately from Proposition 10.3 and the first inequality in (1.31), which implies that $D \subset \sqrt{r} \cdot B_*$. □

Proof of Theorem 10.1. The argument is a matrix version of the proof of Theorem 7.1. We consider the following subsets of $d_1 \times d_2$ matrices:

$$\bar{K} := \{X' : \ \|X'\|_* \le 1\}, \quad K := \|X\|_* \cdot \bar{K}.$$

Then obviously $X \in K$, so it makes sense to apply Theorem 6.2 (with $\varepsilon = 0$) for K. It should also be clear that the optimization program in Theorem 6.2 can be written in the form (1.35).

Applying Theorem 6.2, we obtain

$$\mathbb{E} \sup_X \|\hat{X} - X\|_F \le \sqrt{2\pi} \cdot \frac{w(K)}{\sqrt{m}}.$$

Recalling the definition of K and using Proposition 10.3 to bound its mean width, we have

$$w(K) = w(\bar{K}) \cdot \|X\|_* \le 2\sqrt{2}\sqrt{d_1 + d_2} \cdot \|X\|_*.$$

It follows that

$$\mathbb{E} \sup_X \|\hat{X} - X\|_F \le 4\sqrt{\pi}\sqrt{\frac{d_1 + d_2}{m}} \cdot \|X\|_*.$$

It remains to use the low-rank assumption (1.34). According to the first inequality in (1.31), we have

$$\|X\|_* \le \sqrt{r}\|X\|_F.$$

This completes the proof of Theorem 10.1. □

1.10.3 Low-rank matrix recovery: some extensions

1.10.3.1 From exact to effective low rank

The exact low-rank assumption (1.34) can be replaced by approximate low-rank assumption. This is a matrix version of a similar observation about sparsity which we made in Section 1.7.2.2. Indeed, our argument shows that Theorem 10.1 will hold if we replace the rank by the more flexible *effective rank*, defined for a matrix X as

$$r(X) = (\|X\|_*/\|X\|_F)^2.$$

The effective rank is clearly bounded by the algebraic rank, and it is robust with respect to small perturbations.

1.10.3.2 Noisy and sub-Gaussian observations

Our argument makes it easy to allow noise in the observations (1.33), i.e., consider observations of the form $y_i = \langle A_i, X\rangle + \nu_i$. We leave details to the interested reader.

Further, just like in Section 1.8, we can relax the requirement that A_i be Gaussian random matrices, replacing it with a *sub-Gaussian* assumption. Namely, it is enough to assume that the columns of A_i are i.i.d., mean zero, isotropic, and sub-Gaussian random vectors in $\mathbb{R}^{d_1}$, with a common bound on the sub-Gaussian norm. We again leave details to the interested reader.

We can summarize the results about low-rank matrix recovery as follows.

Using convex programming, one can approximately recover a $d_1 \times d_2$ matrix which has rank (or effective rank) r, from $m \sim r(d_1 + d_2)$ random linear observations.

To understand this number of observations better, note that it is of the same order as the number of degrees of freedom in the set of $d_1 \times d_2$ matrices or rank r.

1.10.4 Matrix completion

Let us now consider a different, and perhaps more natural, model of observations of matrices. Assume that we are given a *small random sample of entries* of an unknown matrix X. Our goal is to estimate X from this sample. As before, we assume that X has low rank. This is called a *matrix completion problem*, and it was extensively studied recently [12, 13, 37, 63].

The theory we discussed earlier in this chapter does not apply here. While sampling of entries is a linear operation, such observations are not Gaussian or sub-Gaussian (more accurately, we should say that the sub-Gaussian norm of such observations is too large). Nevertheless, it is possible to derive a matrix completion

result in this setting. Our exposition will be based on a direct and simple argument from [60]. The reader interested in deeper understanding of the matrix completion problem (and in particular exact completion) is referred to the papers cited above.

Let us formalize the process of sampling the entries of X. First, we fix the average size m of the sample. Then we generate selectors $\delta_{ij} \in \{0, 1\}$ for each entry of X. Those are i.i.d. random variables with

$$\mathbb{E}\,\delta_{ij} = \frac{m}{d_1 d_2} =: p.$$

Our observations are given as the $d_1 \times d_2$ matrix Y whose entries are

$$Y_{ij} = \delta_{ij} X_{ij}.$$

Therefore, the observations are randomly and independently sampled entries of X along with the indices of these entries; the average sample size is fixed and equals m. We will require that

$$m \geq d_1 \log d_1, \quad m \geq d_2 \log d_2. \tag{1.36}$$

These restrictions ensure that, with high probability, the sample contains at least one entry from each row and each column of X (recall the classical coupon collector's problem).

As before, we assume that

$$\operatorname{rank}(X) \leq r.$$

The next result shows that X can be estimated from Y using low-rank approximation.

Theorem 10.5 (Matrix completion). *Choose $\hat{X}$ to be best rank-r approximation*[9] *of $p^{-1}Y$. Then*

$$\mathbb{E}\,\frac{1}{\sqrt{d_1 d_2}}\,\|\hat{X} - X\|_F \leq C\sqrt{\frac{r(d_1 + d_2)}{m}}\,\|X\|_\infty, \tag{1.37}$$

where $\|X\|_\infty = \max_{i,j} |X_{ij}|$.

To understand the form of this estimate, note that the left-hand side of (1.37) measures the *average error per entry* of X:

[9] Formally, consider the singular value decomposition $p^{-1}Y = \sum_i s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}$ with nonincreasing singular values s_i. We define $\hat{X}$ by retaining the r leading terms of this decomposition, i.e., $\hat{X} = \sum_{i=1}^{r} s_i \boldsymbol{u}_i \boldsymbol{v}_i^{\mathsf{T}}$.

$$\frac{1}{\sqrt{d_1 d_2}}\|\hat{X} - X\|_F = \Big(\frac{1}{d_1 d_2}\sum_{i=1}^{d_1}\sum_{j=1}^{d_2}|\hat{X}_{ij} - X_{ij}|^2\Big)^{1/2}.$$

So Theorem 10.5 allows to make the average error per entry arbitrarily smaller than the maximal entry of the matrix. Such estimation succeeds with a sample of $m \sim r(d_1 + d_2)$ entries of X.

The proof of Theorem 10.5 will be based on a known bound on the operator norm of random matrices, which is more general than Y. Gordon's Theorem 10.2. There are several ways to obtain general bounds; see [73] for a systematic treatment of this topic. We will use one such result due to Y. Seginer [67].

Theorem 10.6 (Seginer's bound for general random matrices). *Let G be a $d_1 \times d_2$ matrix whose entries are i.i.d., mean zero random variables. Then*

$$\mathbb{E}\,\|G\| \le C\Big(\mathbb{E}\max_i \|G_i\|_2 + \mathbb{E}\max_j \|G^j\|_2\Big)$$

where the maxima are taken over all rows G_i and over all columns G^j of G, respectively.

Proof of Theorem 10.5. We shall first control the error in the operator norm. By triangle inequality,

$$\|\hat{X} - X\| \le \|\hat{X} - p^{-1}Y\| + \|p^{-1}Y - X\|. \tag{1.38}$$

Since $\hat{X}$ is the best rank-r approximation to $p^{-1}Y$, and both X and $\hat{X}$ are rank-r matrices, the first term in (1.38) is bounded by the second term. Thus

$$\|\hat{X} - X\| \le 2\|p^{-1}Y - X\| = \frac{2}{p}\|Y - pX\|. \tag{1.39}$$

The matrix $Y - pX$ has independent mean zero entries, namely

$$(Y - pX)_{ij} = (\delta_{ij} - p)X_{ij}.$$

So we can apply Y. Seginer's Theorem 10.6, which yields

$$\mathbb{E}\,\|Y - pX\| \le C\Big(\mathbb{E}\max_{i\le d_1}\|(Y - pX)_i\|_2 + \mathbb{E}\max_{j\le d_2}\|(Y - pX)^j\|_2\Big). \tag{1.40}$$

It remains to bound the ℓ_2 norms of rows and columns of $Y - pX$. Let us do this for rows; a similar argument would control the columns. Note that

$$\|(Y - pX)_i\|_2^2 = \sum_{j=1}^{d_2}(\delta_{ij} - p)^2|X_{ij}|^2 \le \sum_{j=1}^{d_2}(\delta_{ij} - p)^2 \cdot \|X\|_\infty^2, \tag{1.41}$$

where $\|X\|_\infty = \max_{i,j} |X_{ij}|$ is the ℓ_∞ norm of X considered as a vector in $\mathbb{R}^{d_1 \times d_2}$. To further bound the quantity in (1.41) we can use concentration inequalities for sums of independent random variables. In particular, we can use Bernstein's inequality (see [9]), which yields

$$\mathbb{P}\left\{ \sum_{j=1}^{d_2} (\delta_{ij} - p)^2 > p d_2 t \right\} \le \exp(-c p d_2 t), \quad t \ge 2.$$

The first restriction in (1.36) guarantees that $p d_2 \ge \log d_1$. This enables us to use the union bound over $i \le d_1$, which yields

$$\mathbb{E} \max_{i \le d_1} \Big[\sum_{j=1}^{d_2} (\delta_{ij} - p)^2 \Big]^{1/2} \le C_1 \sqrt{p d_2}.$$

This translates into the following bound for the rows of $Y - pX$:

$$\mathbb{E} \max_{i \le d_1} \|(Y - pX)_i\|_2 \le C_1 \sqrt{p d_2}\, \|X\|_\infty.$$

Repeating this argument for columns and putting the two bounds into (1.40), we obtain

$$\mathbb{E}\, \|Y - pX\| \le C_2 \sqrt{p(d_1 + d_2)}\, \|X\|_\infty.$$

Substituting into (1.39), we conclude that

$$\mathbb{E}\, \|\hat{X} - X\| \le C_3 \sqrt{\frac{d_1 + d_2}{p}}\, \|X\|_\infty. \tag{1.42}$$

It remains to pass to the Frobenius norm. This is where we use the low-rank assumption on X. Since both X and $\hat{X}$ have ranks bounded by r, we have $\operatorname{rank}(\hat{X} - X) \le 2r$. Then, according to the second inequality in (1.31),

$$\|\hat{X} - X\|_F \le \sqrt{2r}\, \|\hat{X} - X\|.$$

Combining this with (1.42) and recalling that $p = m/(d_1 d_2)$ by definition, we arrive at the desired bound (1.37). □

Remark 10.7 (Noisy observations). One can easily extend Theorem 10.5 for noisy sampling, where every observed entry of X is independently corrupted by a mean zero noise. Formally, we assume that the entries of the observation matrix Y are

$$Y_{ij} = \delta_{ij}(X_{ij} + \nu_{ij})$$

where ν_{ij} are independent and mean zero random variables. Let us further assume that $|\nu_{ij}| \le M$ almost surely. Then a slight modification of the proof of Theorem 10.5 yields the following error bound:

$$\mathbb{E}\,\frac{1}{\sqrt{d_1 d_2}}\,\|\hat{X} - X\|_F \le C\sqrt{\frac{r(d_1+d_2)}{m}}\,\big(\|X\|_\infty + M\big).$$

We leave details to the interested reader.

1.11 Single-bit observations via hyperplane tessellations

It may perhaps be surprising that a theory of similar strength can be developed for estimation problems with *nonlinear* observations, in which the observation vector $\boldsymbol{y} \in \mathbb{R}^m$ depends nonlinearly on the unknown vector $\boldsymbol{x} \in \mathbb{R}^n$.

In this and next sections, we explore an example of extreme non-linearity—the one given by the sign function. In Section 1.13, we will extend the theory to completely general nonlinearities.

1.11.1 Single-bit observations

As before, our goal is to estimate an unknown vector $\boldsymbol{x}$ that lies in a known feasible set $K \subset \mathbb{R}^n$, from a random observation vector $\boldsymbol{y} = (y_1, \dots, y_m) \in \mathbb{R}^m$. This time, we will work with *single-bit observations* $y_i \in \{-1, 1\}$. So we assume that

$$y_i = \operatorname{sign} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle, \quad i = 1, \dots, m, \tag{1.43}$$

where $\boldsymbol{a}_i$ are standard Gaussian random vectors, i.e., $\boldsymbol{a}_i \sim N(0, I_n)$. We can represent the model in a matrix form:

$$\boldsymbol{y} = \operatorname{sign}(A\boldsymbol{x}),$$

where A is an $m \times n$ Gaussian random matrix with rows $\boldsymbol{a}_i$, and where our convention is that the sign function is applied to each coordinate of the vector $A\boldsymbol{x}$.

The single-bit model represents an extreme *quantization* of the linear model we explored before, where $\boldsymbol{y} = A\boldsymbol{x}$. Only one bit is retained from each linear observation y_i. Yet we hope to estimate $\boldsymbol{x}$ as accurately as if all bits were available.

The model of single-bit observations was first studied in this context in [10]. Our discussion will follow [59].

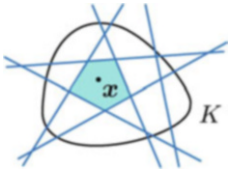


Fig. 1.9 A tessellation of the feasible set K by hyperplanes. The cell containing $\boldsymbol{x}$ is highlighted.

1.11.2 Hyperplane tessellations

Let us try to understand single-bit observations y_i from a geometric perspective. Each $y_i \in \{-1, 1\}$ represents the orientation of the vector $\boldsymbol{x}$ with respect to the hyperplane with normal $\boldsymbol{a}_i$. There are m such hyperplanes. The observation vector $\boldsymbol{y} = (y_1, \dots, y_m)$ represents orientation of $\boldsymbol{x}$ with respect to all these hyperplanes.

Geometrically, the m hyperplanes induce a *tessellation* of $\mathbb{R}^n$ by *cells*. A cell is a set of points that have the same orientation with respect to all hyperplanes; see Figure 1.9. Knowing $\boldsymbol{y}$ is the same as knowing the cell where $\boldsymbol{x}$ lies.

How can we estimate $\boldsymbol{x}$? Recall that we know two pieces of information about $\boldsymbol{x}$:

1. $\boldsymbol{x}$ lies in a known cell of the hyperplane tessellation;
2. $\boldsymbol{x}$ lies in a known set K.

Therefore, a good estimator of $\boldsymbol{x}$ can be obtained by picking any vector $\hat{\boldsymbol{x}}$ from the *intersection of these two sets*. Moreover, since just these two pieces of information about $\boldsymbol{x}$ are available, such an estimator is best possible in some sense.

1.11.3 M^* bound for random tessellations

How good is such an estimate? The maximal error is of course the diameter of the intersection of the cell with K. So in order to bound the error, we need to prove that this diameter is small.

Note that our strategy is parallel to what we have done for linear observations in Section 1.4.1. The only piece we are missing is a version of M^* bound for random tessellations instead of random subspaces. Informally, we need a result about the following question:

Question 11.1 (Pizza cutting). How many random hyperplanes would cut a given set K into pieces that are at most ε in size?

A result about this problem was proved in [59].

Theorem 11.2 (M^* bound for random tessellations). *Consider a set $K \subseteq S^{n-1}$ and m independent random hyperplanes drawn uniformly from the Grassmanian $G_{n,n-1}$. Then*

$$\mathbb{E} \max_{\mathcal{C}} \operatorname{diam}(K \cap \mathcal{C}) \le \left[\frac{Cw(K)}{\sqrt{m}} \right]^{1/3}, \tag{1.44}$$

where the maximum is taken over all cells $\mathcal{C}$ of the hyperplane tessellation.[10]

Apart from the exponent $1/2$ which is unlikely to be optimal, this result is indeed a version of the M^* bound, Theorem 3.12. To further highlight the similarity, note that when $m < n$, the intersection of the m random hyperplanes is a random linear subspace E of codimension m. This subspace lies in each cell of the tessellation. So in particular, Theorem 11.2 controls the quantity $\mathbb{E} \operatorname{diam}(K \cap E)$ appearing in the standard M^* bound, Theorem 3.12.

1.11.4 Estimation based on M^* bound for random tessellations

Now we can apply Theorem 11.2 for the estimation problem. Based on our discussion in Section 1.11.2, this result immediately implies the following.

Theorem 11.3 (Estimation from single-bit observations: feasibility program). *Assume the unknown vector $\boldsymbol{x}$ lies in some known set $K \subseteq S^{n-1}$, and the single-bit observation vector $\boldsymbol{y}$ is given by (1.43). Choose $\hat{\boldsymbol{x}}$ to be any vector satisfying*

$$\hat{\boldsymbol{x}} \in K \quad \text{and} \quad \operatorname{sign}(A\hat{\boldsymbol{x}}) = \boldsymbol{y}. \tag{1.45}$$

Then

$$\mathbb{E} \sup_{\boldsymbol{x} \in K} \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \left[\frac{Cw(K)}{\sqrt{m}} \right]^{1/3}. \quad \square$$

We assumed in this result that feasible set K lies on the unit sphere. This is because the magnitude $\|\boldsymbol{x}\|_2$ is obviously lost in the single-bit observations. So we can only hope to estimate the direction of $\boldsymbol{x}$, which is the vector $\boldsymbol{x}/\|\boldsymbol{x}\|_2$ on the unit sphere.

A good news is that estimation can be made from $m \sim w(K)^2$ single-bit observations, the same as for linear observations. So perhaps surprisingly, the essential information about $\boldsymbol{x}$ is contained in a single bit of each observation.

Bad news is that the feasibility program (1.45) is *not convex*. When K is restricted to lie on the sphere, it can never be convex or be convexified. One can get around this issue, for example, by lifting the restriction; see [59] for pizza cutting of general sets in $\mathbb{R}^n$.

[10] A high-probability version of Theorem 11.2 was proved in [59]. Namely, denoting by δ the right-hand side of (1.44), we have $\max_{\mathcal{C}} \operatorname{diam}(K \cap \mathcal{C}) \le \delta$ with probability at least $1 - 2\exp(-c\delta^2 m)$, as long as $m \ge C\delta^{-6} w(K)^2$. The reader will easily deduce the statement of Theorem 11.2 from this.

But a better idea will be to replace the feasibility problem (1.45) by an optimization problem—just like we did in Section 1.4.2—which will work for general sets K in the unit *ball* B_2^n rather than the unit sphere. Such sets can be convexified. We will do this in the next section.

1.12 Single-bit observations via optimization and applications to logistic regression

Our goal remains the same as we described in Section 1.11.1. We would like to estimate a vector $\boldsymbol{x}$ that lies in a known feasible set $K \subset \mathbb{R}^n$, from single-bit observations given as

$$\boldsymbol{y} = \operatorname{sign}(A\boldsymbol{x}) \in \{-1, 1\}^m.$$

Instead of formulating estimation as a feasibility problem (1.45), we will now state it as an *optimization* problem, as follows:

$$\text{maximize } \langle A\boldsymbol{x}', \boldsymbol{y} \rangle \text{ subject to } \boldsymbol{x}' \in K. \tag{1.46}$$

This program tries to fit linear observations $A\boldsymbol{x}'$ to the single-bit observations $\boldsymbol{y}$. It does so by maximizing the correlation between linear and single-bit observations while searching inside the feasible set K.

If K is a convex set, (1.46) is a convex program. Otherwise one can convexify K as we did several times before.

The following result from [58] provides a guarantee for such estimator.

Theorem 12.1 (Estimation from single-bit observations: optimization program). *Assume the unknown vector $\boldsymbol{x} \in \mathbb{R}^n$ satisfies $\|\boldsymbol{x}\|_2 = 1$ and $\boldsymbol{x}$ lies in some known set $K \subseteq B_2^n$. Choose $\hat{\boldsymbol{x}}$ to be a solution to the program (1.46). Then*

$$\mathbb{E}\, \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 \le \frac{Cw(K)}{\sqrt{m}}.$$

Here $C = \sqrt{8\pi} \approx 5.01$.

Our proof of Theorem 12.1 will be based on properties of the *loss function*, which we define as

$$L_{\boldsymbol{x}}(\boldsymbol{x}') = -\frac{1}{m} \langle A\boldsymbol{x}', \boldsymbol{y} \rangle = -\frac{1}{m} \sum_{i=1}^m y_i \langle \boldsymbol{a}_i, \boldsymbol{x}' \rangle .$$

The index $\boldsymbol{x}$ indicates that the loss function depends on $\boldsymbol{x}$ through $\boldsymbol{y}$. The negative sign is chosen so that program (1.46) minimizes the loss function over K.

We will now compute the expected value and the deviation of the loss function for fixed $\boldsymbol{x}$ and $\boldsymbol{x}'$.

Lemma 12.2 (Expectation of loss function). *Let $\boldsymbol{x} \in S^{n-1}$ and $\boldsymbol{x}' \in \mathbb{R}^n$. Then*

$$\mathbb{E}\, L_{\boldsymbol{x}}(\boldsymbol{x}') = -\sqrt{\frac{2}{\pi}}\, \langle \boldsymbol{x}, \boldsymbol{x}' \rangle .$$

Proof. We have

$$\mathbb{E}\, L_{\boldsymbol{x}}(\boldsymbol{x}') = -\mathbb{E}\, y_1 \langle \boldsymbol{a}_1, \boldsymbol{x}' \rangle = -\mathbb{E}\, \operatorname{sign}(\langle \boldsymbol{a}_1, \boldsymbol{x} \rangle) \langle \boldsymbol{a}_1, \boldsymbol{x}' \rangle .$$

It remains to note that $\langle \boldsymbol{a}_1, \boldsymbol{x} \rangle$ and $\langle \boldsymbol{a}_1, \boldsymbol{x}' \rangle$ are normal random variables with zero mean, variances $\|\boldsymbol{x}\|_2^2 = 1$ and $\|\boldsymbol{x}'\|_2^2$, respectively, and covariance $\langle \boldsymbol{x}, \boldsymbol{x}' \rangle$. A simple calculation renders the expectation above as $-\langle \boldsymbol{x}, \boldsymbol{x}' \rangle \cdot \mathbb{E}\, \operatorname{sign}(g) g$ where $g \sim N(0,1)$. It remains to recall that $\mathbb{E}\, \operatorname{sign}(g) g = \mathbb{E}\, |g| = \sqrt{2/\pi}$. □

Lemma 12.3 (Uniform deviation of loss function). *We have*

$$\mathbb{E} \sup_{\boldsymbol{u} \in K - K} |L_{\boldsymbol{x}}(\boldsymbol{u}) - \mathbb{E}\, L_{\boldsymbol{x}}(\boldsymbol{u})| \le \frac{2w(K)}{\sqrt{m}}. \tag{1.47}$$

Proof. Due to the form of loss function, we can apply the symmetrization inequality of Proposition 5.2, which bounds the left-hand side of (1.47) by

$$\frac{2}{m}\, \mathbb{E} \sup_{\boldsymbol{u} \in K - K} \Big| \sum_{i=1}^m \varepsilon_i y_i \langle \boldsymbol{a}_i, \boldsymbol{u} \rangle \Big| = \frac{2}{m}\, \mathbb{E} \sup_{\boldsymbol{u} \in K - K} \Big| \Big\langle \sum_{i=1}^m \varepsilon_i y_i \boldsymbol{a}_i, \boldsymbol{u} \Big\rangle \Big|. \tag{1.48}$$

By symmetry and since $y_i \in \{-1, 1\}$, the random vectors $\{\varepsilon_i y_i \boldsymbol{a}_i\}$ are distributed identically with $\{\boldsymbol{a}_i\}$. In other words, we can remove $\varepsilon_i y_i$ from (1.48) without changing the value of the expectation.

Next, by rotation invariance, $\sum_{i=1}^m \boldsymbol{a}_i$ is distributed identically with $\sqrt{m}\, \boldsymbol{g}$, where $\boldsymbol{g} \sim N(0, I_n)$. Therefore, the quantity in (1.48) equals

$$\frac{2}{\sqrt{m}}\, \mathbb{E} \sup_{\boldsymbol{u} \in K - K} |\langle \boldsymbol{g}, \boldsymbol{u} \rangle| = \frac{2w(K)}{\sqrt{m}}.$$

This completes the proof. □

Proof of Theorem 12.1. Fix $\boldsymbol{x}' \in K$. Let us try to bound $\|\boldsymbol{x} - \boldsymbol{x}'\|_2$ in terms of $L_{\boldsymbol{x}}(\boldsymbol{x}) - L_{\boldsymbol{x}}(\boldsymbol{x}')$. By linearity of the loss function, we have

$$L_{\boldsymbol{x}}(\boldsymbol{x}) - L_{\boldsymbol{x}}(\boldsymbol{x}') = L_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}') = \mathbb{E}\, L_{\boldsymbol{x}}(\boldsymbol{x} - \boldsymbol{x}') + D_{\boldsymbol{x}} \tag{1.49}$$

where the deviation

$$D_x := \sup_{u \in K-K} |L_x(u) - \mathbb{E}\, L_x(u)|$$

will be controlled using Lemma 12.3 a bit later.

To compute the expected value in (1.49), we can use Lemma 12.2 along with the conditions $\|x\|_2 = 1$, $\|x'\|_2 \le 1$ (the latter holds since $x' \in K \subseteq B_2^n$). This way we obtain

$$\mathbb{E}\, L_x(x - x') = -\sqrt{\frac{2}{\pi}}\, \langle x, x - x' \rangle \le -\frac{1}{2}\sqrt{\frac{2}{\pi}}\, \|x - x'\|_2^2.$$

Putting this into (1.49), we conclude that

$$L_x(x) - L_x(x') \le -\frac{1}{\sqrt{2\pi}}\, \|x - x'\|_2^2 + D_x. \tag{1.50}$$

This bound holds for any fixed $x' \in K$ and for any point in the probability space (i.e., for any realization of the random variables appearing in this bound). Therefore, (1.50) must hold for the random vector $x' = \hat{x}$, again for any point in the probability space.

The solution $\hat{x}$ was chosen to minimize the loss function; thus $L_x(\hat{x}) \le L_x(x)$. This means that for $x' = \hat{x}$, the left-hand side of (1.50) is non-negative. Rearranging the terms, we obtain

$$\|x - \hat{x}\|_2^2 \le \sqrt{2\pi}\, D_x.$$

It remains to take expectation on both sides and use Lemma 12.3. This yields

$$\mathbb{E}\, \|x - \hat{x}\|_2^2 \le \sqrt{2\pi}\, \frac{2w(K)}{\sqrt{m}}.$$

This completes the proof of Theorem 12.1. □

1.12.1 Single-bit observations with general nonlinearities

The specific nonlinearity of observations that we considered so far—the one given by sign function—did not play a big role in our argument in the last section. The same argument, and surprisingly the same optimization program (1.46), can serve any nonlinearity in the observations.

So let us consider a general model of single-bit observations $y = (y_1, \dots, y_m) \in \{-1, 1\}^m$, which satisfy

$$\mathbb{E}\, y_i = \theta(\langle a_i, x \rangle), \quad i = 1, \dots, m \tag{1.51}$$

Here $\theta : \mathbb{R} \to \mathbb{R}$ is some *link function*, which describes nonlinearity of observations. We assume that y_i are independent given $\boldsymbol{a}_i$, which are standard Gaussian random vectors as before. The matrix form of this model can be written as

$$\mathbb{E}\boldsymbol{y} = \theta(A\boldsymbol{x}),$$

where A is an $m \times n$ Gaussian random matrix with rows $\boldsymbol{a}_i$, and where our convention is that θ is applied to each coordinate of the vector $A\boldsymbol{x}$.

To estimate $\boldsymbol{x}$, an unknown vector in a known feasible set K, we will try to use the same optimization program (1.46) in the last section. This may be surprising since *the program does not even need to know the nonlinearity* θ, nor does it attempt to estimate θ. Yet, this idea works in general as nicely as for the specific sign function. The following result from [58] is a general version of Theorem 12.1.

Theorem 12.4 (Estimation from single-bit observations with general nonlinearity). *Assume the unknown vector* $\boldsymbol{x} \in \mathbb{R}^n$ *satisfies* $\|\boldsymbol{x}\|_2 = 1$ *and* $\boldsymbol{x}$ *lies in some known set* $K \subseteq B_2^n$. *Choose* $\hat{\boldsymbol{x}}$ *to be a solution to the program* (1.46). *Then*

$$\mathbb{E}\,\|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2^2 \le \frac{4w(K)}{\lambda\sqrt{m}}.$$

Here we assume that

$$\lambda := \mathbb{E}\,\theta(g)g > 0 \quad \text{for} g \sim N(0,1). \tag{1.52}$$

Proof. The argument follows very closely the proof of Theorem 12.1. The only different place is the computation of expected loss function in Lemma 12.2. When the sign function is replaced by a general nonlinearity θ, one easily checks that the expected value becomes

$$\mathbb{E}\,L_{\boldsymbol{x}}(\boldsymbol{x}') = -\lambda\left\langle \boldsymbol{x}, \boldsymbol{x}'\right\rangle.$$

The rest of the argument is the same. □

For $\theta(z) = \operatorname{sign}(z)$, Theorem 12.4 is identical with Theorem 12.1. However, the new result is much more general. *Virtually no restrictions are imposed on the nonlinearity* θ. In particular, θ need not be continuous or one to one.

The parameter λ simply measures the information content retained through the nonlinearity. It might be useful to express λ as

$$\lambda = \mathbb{E}\,\theta(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle)\,\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle\,,$$

so λ measures how much the nonlinear observations $\theta(\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle)$ are correlated with linear observations $\langle \boldsymbol{a}_i, \boldsymbol{x}\rangle$.

The assumption that $\lambda > 0$ is made for convenience; if $\lambda < 0$ we can switch the sign of θ. However, if $\lambda = 0$, the nonlinear and linear measurements are

uncorrelated, and often no estimation is possible. An extreme example of the latter situation occurs when θ is a constant function, which clearly carries no information about $\boldsymbol{x}$.

1.12.2 Logistic regression and beyond

For the link function $\theta(z) = \tanh(z/2)$, the estimation problem (1.51) is equivalent to *logistic regression with constraints*. In the usual statistical notation explained in Section 1.7.4, logistic regression takes the form

$$\mathbb{E}\boldsymbol{y} = \tanh(X\boldsymbol{\beta}/2).$$

The coefficient vector β is constrained to lie in some known feasible set K. We will leave it to the interested reader to translate Theorem 12.4 into the language of logistic regression, just like we did in Section 1.7.4 for linear regression.

The fact that Theorem 12.4 applies for general and unknown link function should be important in statistics. It means that one *does not need to know the non-linearity of the model (the link function) to make inference*. Be it the tanh function specific to logistic regression or (virtually) any other non-linearity, the estimator $\hat{\boldsymbol{\beta}}$ is the same.

1.13 General nonlinear observations via metric projection

Finally, we pass to the most general model of observations $\boldsymbol{y} = (y_1, \dots, y_m)$, which are not necessarily linear or single bit. In fact, we will not even specify a dependence of y_i on $\boldsymbol{x}$. Instead, we only require that y_i be i.i.d. random variables, and

$$\text{each observation } y_i \text{ may depend on } \boldsymbol{a}_i \text{ only through } \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle. \tag{1.53}$$

Technically, the latter requirement means that, given $\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle$, the observation y_i is independent from $\boldsymbol{a}_i$. This type of observation models are called *single-index models* in statistics.

How can we estimate $\boldsymbol{x} \in K$ from such general observation vector $\boldsymbol{y}$? Let us look again at the optimization problem (1.46), writing it as follows:

$$\text{maximize } \langle \boldsymbol{x}', A^{\mathsf{T}}\boldsymbol{y} \rangle \text{ subject to } \boldsymbol{x}' \in K.$$

It might be useful to imagine solving this program as a sequence of two steps: (a) compute a *linear estimate* of $\boldsymbol{x}$, which is

$$\hat{\boldsymbol{x}}_{\text{lin}} = \frac{1}{m} A^{\mathsf{T}} \boldsymbol{y} = \frac{1}{m} \sum_{i=1}^{m} y_i \boldsymbol{a}_i, \tag{1.54}$$

and then (b) *fitting* $\hat{\boldsymbol{x}}_{\text{lin}}$ to the feasible set K, which is done by choosing a point in K that is most correlated with $\hat{\boldsymbol{x}}_{\text{lin}}$.

Surprisingly, almost the same estimation procedure succeeds for the general single-index model (1.53). We just need to adjust the second, fitting, step. Instead of maximizing the correlation, let us metrically *project* $\hat{\boldsymbol{x}}_{\text{lin}}$ *onto the feasible set* K, thus choosing $\hat{\boldsymbol{x}}$ to be a solution of the program

$$\text{minimize} \|\boldsymbol{x}' - \hat{\boldsymbol{x}}_{\text{lin}}\|_2 \text{ subject to } \boldsymbol{x}' \in K. \tag{1.55}$$

Just like in the previous section, it may be surprising that this estimator does not need to know the nature of the nonlinearity in observations $\boldsymbol{y}$. To get a heuristic evidence of why this knowledge may not be needed, one can quickly check (using rotation invariance) that

$$\mathbb{E}\, \hat{\boldsymbol{x}}_{\text{lin}} = \mathbb{E}\, y_1 \boldsymbol{a}_1 = \lambda \bar{\boldsymbol{x}}, \quad \text{where} \quad \bar{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|_2, \quad \lambda = \mathbb{E}\, y_1 \langle \boldsymbol{a}_1, \bar{\boldsymbol{x}} \rangle .$$

So despite not knowing the nonlinearity, $\hat{\boldsymbol{x}}_{\text{lin}}$ already provides an *unbiased estimate* of $\boldsymbol{x}$, up to scaling.

A result from [60] provides a guarantee for the two-step estimator (1.54), (1.55). Let us state this result in a special case where K is a *cone*, i.e., $tK = K$ for all $t \geq 0$. A version for general sets K is not much more difficult, see [60] for details.

Since cones are unbounded sets, the standard mean width (as defined in (1.3)) would be infinite. To get around this issue, we should consider a *local* version of mean width, which we can define as

$$w_1(K) = \mathbb{E} \sup_{\boldsymbol{u} \in (K-K) \cap B_2^n} \langle \boldsymbol{g}, \boldsymbol{u} \rangle , \quad \boldsymbol{g} \sim N(0, I_n).$$

Theorem 13.1 (Estimation from nonlinear observations). *Assume the unknown vector* $\boldsymbol{x}$ *lies in a known closed cone* K *in* $\mathbb{R}^n$. *Choose* $\hat{\boldsymbol{x}}$ *to be a solution to the program* (1.55). *Let* $\bar{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|_2$. *Then*

$$\mathbb{E}\, \hat{\boldsymbol{x}} = \lambda \bar{\boldsymbol{x}} \quad \text{and} \quad \mathbb{E}\, \|\hat{\boldsymbol{x}} - \lambda \bar{\boldsymbol{x}}\|_2 \leq \frac{M w_1(K)}{\sqrt{m}}.$$

Here we assume that

$$\lambda = \mathbb{E}\, y_1 \langle \boldsymbol{a}_1, \bar{\boldsymbol{x}} \rangle > 0 \quad \text{and} \quad M = \sqrt{2\pi} \big[\mathbb{E}\, y_1^2 + \text{Var}\left(y_1 \langle \boldsymbol{a}_1, \bar{\boldsymbol{x}} \rangle \right) \big]^{1/2}.$$

The proof of Theorem 13.1 is given in [60, Theorem 2.1]. It is not difficult, and is close in spirit to the arguments we saw here; we will not reproduce it.

The role of parameters λ and M is to determine the correct magnitude and deviation of the estimator; one can think of them as *constants* that are usually easy to compute or estimate. By rotation invariance, λ and M depend on the *magnitude* $\|\boldsymbol{x}\|_2$ (through y_1) but not on the direction $\bar{\boldsymbol{x}} = \boldsymbol{x}/\|\boldsymbol{x}\|_2$ of the unknown vector $\boldsymbol{x}$.

We can summarize results of this and previous section as follows.

One can estimate a vector $\boldsymbol{x}$ in a general feasible set K from $m \sim w(K)^2$ random nonlinear observations, even if the nonlinearity is not known. If K is convex, estimation can be done using convex programming.

1.13.1 Examples of observations

To give a couple of concrete examples, consider *noisy linear observations*

$$y_i = \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle + \nu_i.$$

We already explored this model in Section 1.6, where ν_i were arbitrary numbers representing noise. This time, let us assume ν_i are independent random variables with zero mean and variance σ^2. A quick computation gives

$$\lambda = \|\boldsymbol{x}\|_2, \quad M = C(\|\boldsymbol{x}\|_2 + \sigma).$$

Theorem 13.1 then yields the following error bound:

$$\mathbb{E}\, \|\hat{\boldsymbol{x}} - \boldsymbol{x}\|_2 \le \frac{C w_1(K)}{\sqrt{m}} \, (\|\boldsymbol{x}\|_2 + \sigma).$$

Let us give one more example, for the *single-bit observations*

$$y_i = \operatorname{sign} \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle .$$

We explored this model in Sections 1.11 and 1.12. A quick computation gives

$$\lambda = \sqrt{\frac{2}{\pi}}, \quad M = C.$$

Theorem 13.1 then yields the following error bound:

$$\mathbb{E}\, \Big\| \hat{\boldsymbol{x}} - \sqrt{\frac{2}{\pi}}\, \boldsymbol{x} \Big\|_2 \le \frac{C w_1(K)}{\sqrt{m}}.$$

1.13.2 Examples of feasible cones

To give a couple of concrete examples of feasible cones, consider the set K of *s-sparse vectors* in $\mathbb{R}^n$, those with at most s nonzero coordinates. As we already noted in Example 3.9,

$$w_1(K) \sim \sqrt{s \log(2n/s)}.$$

Further, solving the program (1.55) (i.e., computing the metric projection of $\hat{\boldsymbol{x}}_{\text{lin}}$ onto K) amounts to *hard thresholding* of $\boldsymbol{x}'$. The solution $\hat{\boldsymbol{x}}$ is obtained from $\hat{\boldsymbol{x}}_{\text{lin}}$ by keeping the s largest coefficients (in absolute value) and zeroing out all other coefficients.

So Theorem 13.1 in this case can be stated informally as follows:

One can estimate an s-sparse vector $\boldsymbol{x}$ in $\mathbb{R}^n$ from $m \sim s \log n$ nonlinear observations $\boldsymbol{y}$, even if the nonlinearity is not known. The estimation is given by the hard thresholding of $\hat{\boldsymbol{x}}_{\text{lin}} = m^{-1} A^{\mathsf{T}} \boldsymbol{y}$.

Another popular example of a feasible cone is a set of *low-rank matrices*. Let K be the set of $d_1 \times d_2$ matrices with rank at most r. Proposition 10.4 implies that

$$w_1(K) \le C\sqrt{r(d_1 + d_2)}.$$

Further, solving the program (1.55) (i.e., computing the metric projection of $\boldsymbol{x}'$ onto K) amounts to computing the best rank-r approximation of $\hat{\boldsymbol{x}}_{\text{lin}}$. This amounts to *hard thresholding of singular values* of $\hat{\boldsymbol{x}}_{\text{lin}}$, i.e., keeping the leading s terms of the singular value decomposition. Recall that we already came across this thresholding in the matrix completion problem, Theorem 10.5.

So Theorem 13.1 in this case can be stated informally as follows:

One can estimate a $d_1 \times d_2$ matrix with rank r from $m \sim r(d_1 + d_2)$ nonlinear observations, even if the nonlinearity is not known. The estimation is given by the hard thresholding of singular values of $\hat{\boldsymbol{x}}_{\text{lin}}$.

1.14 Some extensions

1.14.1 From global to local mean width

As we have seen, the concept of Gaussian mean width captures the complexity of a feasible set K quite accurately. Still, it is not exactly the optimal quantity in geometric and estimation results. An optimal quantity is the *local mean width*, which is a function of radius $r > 0$, defined as

$$w_r(K) = \mathbb{E} \sup_{\boldsymbol{u} \in (K-K) \cap rB_2^n} \langle \boldsymbol{g}, \boldsymbol{u} \rangle, \quad \boldsymbol{g} \sim N(0, I_n).$$

Comparing with Definition 3.4 of the usual mean width, we see that

$$w_r(K) \le w(K) \quad \text{forall} r.$$

The usefulness of local mean width was noted in asymptotic convex geometry by A. Giannopoulos and V. Milman [27–29, 31]. They showed that the function $w_r(K)$ completely describes the diameter of high dimensional sections $K \cap E$, thus proving *two-sided* versions of the M^* bound (Theorem 3.12). An observation of a similar nature was made recently by S. Chatterjee [17] in the context of high-dimensional estimation. He noted that a variant of local mean width provides optimal error rates for the *metric projection* onto a feasible set considered in Section 1.13.

For most results discussed in this survey, one can replace the usual mean width by a local mean width, thus making them stronger. Let us briefly indicate how this can be done for the M^* bound (Theorem 3.12); see [28, 29, 31, 47] for a more detailed discussion.

Such localization is in a sense automatic; it can be done as a "post-processing" of the M^* estimate. The conclusion of the general M^* bound, Theorem 5.1, for $T \cap rB_2^n$, is that

$$\sup_{\boldsymbol{u} \in T_\varepsilon \cap rB_2^n} \|\boldsymbol{u}\|_2 \le C\Big(\frac{1}{\sqrt{m}} \mathbb{E} \sup_{\boldsymbol{u} \in T \cap rB_2^n} |\langle \boldsymbol{g}, \boldsymbol{u} \rangle| + \varepsilon \Big) \tag{1.56}$$

with high probability (see also Section 1.5.2). Let us show that the intersection with the ball rB_2^n can be automatically removed from the left side. Since

$$\sup_{\boldsymbol{u} \in T_\varepsilon \cap rB_2^n} \|\boldsymbol{u}\|_2 = \min\Big(\sup_{\boldsymbol{u} \in T_\varepsilon} \|\boldsymbol{u}\|_2, r \Big),$$

it follows that if $\sup_{\boldsymbol{u} \in T_\varepsilon \cap rB_2^n} \|\boldsymbol{u}\|_2 < r$ then $\sup_{\boldsymbol{u} \in T_\varepsilon} \|\boldsymbol{u}\|_2 \le r$. Thus, if the right-hand side of (1.56) is smaller than r, then $\sup_{\boldsymbol{u} \in T_\varepsilon} \|\boldsymbol{u}\|_2 \le r$.

When applied to the classical M^* bound, Theorem 3.12, this argument localizes it as follows:

$$\frac{w_r(K)}{r} \le c\sqrt{m} \quad \text{implies} \quad \operatorname{diam}(K \cap E) \le r$$

with high probability.

1.14.2 More general distributions

For simplicity of exposition, the estimation results in this survey were stated for isotropic Gaussian vectors $\boldsymbol{a}_i$. We showed in Section 1.8 how to extend the M^*

bound and the corresponding linear estimation results for line for *sub-Gaussian* distributions. For more heavy-tailed distributions, a version of M^* bound was proved recently in [46]; compressed sensing for such distributions was examined in [40, 41].

For single-bit observations of Section 1.12, a generalization for sub-Gaussian distributions is discussed in [2]. Some results can be formulated for *anisotropic* Gaussian distributions, where $\boldsymbol{a}_i \sim N(0, \Sigma)$ with $\Sigma \neq I_n$, see, e.g., [58, Section 3.4].

Results for extremely heavy-tailed distributions, such as samples of entries and random Fourier measurements, exist currently only for special cases of feasible sets K. When K consists of sparse vectors, reconstruction of $\boldsymbol{x}$ from Fourier measurements (random frequencies of $\boldsymbol{x}$) was extensively studied in compressed sensing [15, 19, 26, 39]. Reconstruction of a matrix from a random sample of entries was discussed in Section 1.10.4 in the context of matrix completion problem.

There are currently no results, for instance, about reconstruction of $\boldsymbol{x} \in K$ from random Fourier measurements, where K is a general feasible set. It is clear that K needs to be *incoherent* with the Fourier basis of exponentials, but this is yet to be quantified. In the special case where K is a set of sparse vectors, basic results of compressed sensing quantify this incoherence via a *restricted isometry property* [15, 19, 26, 39].

References

1. R. Adamczak, R. Latala, A. Litvak, K. Oleszkiewicz, A. Pajor, N. Tomczak-Jaegermann, A short proof of Paouris' inequality. Can. Math. Bull. **57**, 3–8 (2014)
2. A. Ai, A. Lapanowski, Y. Plan, R. Vershynin, One-bit compressed sensing with non-Gaussian measurements. Linear Algebra Appl. **441**, 222–239 (2014)
3. D. Amelunxen, M. Lotz, M. McCoy, J.A. Tropp, Living on the edge: a geometric theory of phase transitions in convex optimization. Inf. Inference **3**, 224–294 (2014)
4. S. Artstein-Avidan, A. Giannopoulos, V. Milman, *Asymptotic Geometric Analysis, Part I. AMS Mathematical Surveys and Monographs* (2015)
5. F. Bach, R. Jenatton, J. Mairal, G. Obozinski, Structured sparsity through convex optimization. Stat. Sci. **27**, 450–468 (2012)
6. K. Ball, An elementary introduction to modern convex geometry, in *Flavors of Geometry*. Mathematical Sciences Research Institute Publications, vol. 31 (Cambridge University Press, Cambridge, 1997), pp. 1–58
7. H. Bauschke, P. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC (Springer, New York, 2011)
8. A. Ben-Tal, A. Nemirovski, *Lectures on Modern Convex Optimization. Analysis, Algorithms, and Engineering Applications.* MPS/SIAM Series on Optimization (Society for Industrial and Applied Mathematics (SIAM)/Mathematical Programming Society (MPS), Philadelphia, 2001)
9. S. Boucheron, O. Bousquet, G. Lugosi, Concentration inequalities, in *Advanced Lectures in Machine Learning*, ed. by O. Bousquet, U. Luxburg, G. Rätsch (Springer, Berlin, 2004), pp. 208–240
10. P. Boufounos, R. Baraniuk, 1-bit compressive sensing, in *Conference on Information Sciences and Systems (CISS)*, March 2008 (Princeton, New Jersey, 2008)

11. P. Bühlmann, S. van de Geer, *Statistics for High-Dimensional Data. Methods, Theory and Applications.* Springer Series in Statistics (Springer, Heidelberg, 2011)
12. E. Candès, B. Recht, Exact matrix completion via convex optimization. Found. Comput. Math. **9**, 717–772 (2009)
13. E. Candès, T. Tao, The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory **56**, 2053–2080 (2010)
14. E. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM **58**(3), Art. 11, 37 pp. (2011)
15. D. Chafaï, O. Guédon, G. Lecué, A. Pajor, *Interactions Between Compressed Sensing Random Matrices and High Dimensional Geometry.* Panoramas et Synthéses, vol. 37 (Société Mathématique de France, Paris, 2012)
16. V. Chandrasekaran, B. Recht, P. Parrilo, A. Willsky, The convex geometry of linear inverse problems. Found. Comput. Math. **12**, 805–849 (2012)
17. S. Chatterjee, A new perspective on least squares under convex constraint. Ann. Stat. **42**, 2340–2381 (2014)
18. S. Chen, D. Donoho, M. Saunders, Atomic decomposition by Basis Pursuit. SIAM J. Sci. Comput. **20**, 33–61 (1998)
19. M. Davenport, M. Duarte, Y. Eldar, G. Kutyniok, Introduction to compressed sensing, in *Compressed Sensing* (Cambridge University Press, Cambridge, 2012), pp. 1–64
20. M. Davenport, D. Needell, M. Wakin, Signal space CoSaMP for sparse recovery with redundant dictionaries. IEEE Trans. Inf. Theory **59**, 6820–6829 (2013)
21. D. Donoho, M. Elad, Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization. Proc. Natl. Acad. Sci. USA **100**, 2197–2202 (2003)
22. D. Donoho, J. Tanner, Counting faces of randomly projected polytopes when the projection radically lowers dimension. J. Am. Math. Soc. **22**, 1–53 (2009)
23. A. Dvoretzky, A theorem on convex bodies and applications to Banach spaces. Proc. Natl. Acad. Sci. USA **45**, 223–226 (1959)
24. A. Dvoretzky, Some results on convex bodies and Banach spaces, in *Proceedings of the International Symposium on Linear Spaces* (Jerusalem, 1961), pp. 123–161
25. X. Fernique, Regularité des trajectoires des fonctions aléatoires gaussiennes, in *École d'Été de Probabilités de Saint-Flour, IV-1974*. Lecture Notes in Mathematics, vol. 480 (Springer, Berlin, 1975), pp. 1–96
26. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing.* Applied and Numerical Harmonic Analysis (Birkhäuser/Springer, New York, 2013)
27. A. Giannopoulos, V. Milman, How small can the intersection of a few rotations of a symmetric convex body be? C. R. Acad. Sci. Paris Ser. I Math. **325**, 389–394 (1997)
28. A. Giannopoulos, V. Milman, On the diameter of proportional sections of a symmetric convex body. Int. Math. Res. Not. **1**, 5–19 (1997)
29. A. Giannopoulos, V.D. Milman, Mean width and diameter of proportional sections of a symmetric convex body. J. Reine Angew. Math. **497**, 113–139 (1998)
30. A. Giannopoulos, V. Milman, Asymptotic convex geometry: short overview, in *Different Faces of Geometry*. International Mathematical Series (NY), vol. 3 (Kluwer/Plenum, New York, 2004), pp. 87–162
31. A. Giannopoulos, V.D. Milman, Asymptotic formulas for the diameter of sections of symmetric convex bodies. J. Funct. Anal. **223**, 86–108 (2005)
32. A. Giannopoulos, S. Brazitikos, P. Valettas, B.-H. Vritsiou, *Geometry of Isotropic Convex Bodies.* Mathematical Surveys and Monographs, vol. 196 (American Mathematical Society, Providence, 2014)
33. Y. Gordon, On Milman's inequality and random subspaces which escape through a mesh in $\mathbb{R}^n$, in *Geometric Aspects of Functional Analysis. Israel Seminar 1986–1987*. Lecture Notes in Mathematics, vol. 1317 (Springer, Berlin, 1988), pp. 84–106
34. D. Gross, Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**, 1548–1566 (2011)

35. O. Guédon, E. Milman, Interpolating thin-shell and sharp large-deviation estimates for isotropic log-concave measures. Geom. Funct. Anal. **21**, 1043–1068 (2011)
36. L. Jacques, J. Laska, P. Boufounos, R. Baraniuk, Robust 1-bit compressive sensing via binary stable embeddings of sparse vectors. IEEE Trans. Inf. Theory **59**(4), 2082–2102 (2013)
37. R. Keshavan, A. Montanari, S. Oh, Matrix completion from a few entries. IEEE Trans. Inf. Theory **56**, 2980–2998 (2010)
38. B. Klartag, Power-law estimates for the central limit theorem for convex sets. J. Funct. Anal. **245**, 284–310 (2007)
39. G. Kutyniok, Theory and applications of compressed sensing. GAMM-Mitt. **36**, 79–101 (2013)
40. G. Lecué, S. Mendelson, Sparse recovery under weak moment assumptions. J. Eur. Math. Soc. (2015, to appear)
41. G. Lecué, S. Mendelson, Necessary moment conditions for exact reconstruction via Basis Pursuit (submitted)
42. M. Ledoux, *The Concentration of Measure Phenomenon.* Mathematical Surveys and Monographs, vol. 89 (American Mathematical Society, Providence, 2001)
43. M. Ledoux, M. Talagrand, *Probability in Banach Spaces. Isoperimetry and Processes* [Reprint of the 1991 Edition]. Classics in Mathematics (Springer, Berlin, 2011)
44. S. Mendelson, A few notes on statistical learning theory, in *Advanced Lectures in Machine Learning*, ed. by S. Mendelson, A.J. Smola. Lecture Notes in Computer Science, vol. 2600 (Springer, Berlin, 2003), pp. 1–40
45. S. Mendelson, Geometric parameters in learning theory, in *Geometric Aspects of Functional Analysis*. Lecture Notes in Mathematics, vol. 1850 (Springer, Berlin, 2004), pp. 193–235
46. S. Mendelson, A remark on the diameter of random sections of convex bodies, in *Geometric Aspects of Functional Analysis (GAFA Seminar Notes)*. Lecture Notes in Mathematics, vol. 2116 (2014), pp. 395–404
47. S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, Reconstruction and subgaussian operators in asymptotic geometric analysis. Geom. Funct. Anal. **17**, 1248–1282 (2007)
48. V. Milman, New proof of the theorem of Dvoretzky on sections of convex bodies. Funct. Anal. Appl. **5**, 28–37 (1971)
49. V. Milman, Geometrical inequalities and mixed volumes in the local theory of Banach spaces. Astérisque **131**, 373–400 (1985)
50. V. Milman, *Random Subspaces of Proportional Dimension of Finite Dimensional Normed Spaces: Approach Through the Isoperimetric Inequality*. Lecture Notes in Mathematics, vol. 1166 (1985, Springer), pp. 106–115
51. V. Milman, Surprising geometric phenomena in high-dimensional convexity theory, in *European Congress of Mathematics*, vol. II (Budapest, 1996). Progress in Mathematics, vol. 169 (Birkhäuser, Basel, 1998), pp. 73–91
52. V. Milman, G. Schechtman, *Asymptotic Theory of Finite-Dimensional Normed Spaces. With an Appendix by M. Gromov*. Lecture Notes in Mathematics, vol. 1200 (Springer, Berlin, 1986)
53. S. Oymak, B. Hassibi, New null space results and recovery thresholds for matrix rank minimization. Available at arxiv.org/abs/1011.6326 (2010)
54. G. Paouris, Concentration of mass on convex bodies. Geom. Funct. Anal. **16**, 1021–1049 (2006)
55. A. Pajor, N. Tomczak-Jaegermann, Subspaces of small codimension of finite dimen- sional Banach spaces. Proc. Am. Math. Soc. **97**, 637–642 (1986)
56. G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry.* Cambridge Tracts in Mathematics, vol. 94 (Cambridge University Press, Cambridge, 1989)
57. Y. Plan, R. Vershynin, One-bit compressed sensing by linear programming. Commun. Pure Appl. Math. **66**, 1275–1297 (2013)
58. Y. Plan, R. Vershynin, Robust 1-bit compressed sensing and sparse logistic regression: a convex programming approach. IEEE Trans. Inf. Theory **59**, 482–494 (2013)
59. Y. Plan, R. Vershynin, Dimension reduction by random hyperplane tessellations. Discret. Comput. Geom. **51**, 438–461 (2014)

60. Y. Plan, R. Vershynin, E. Yudovina, High-dimensional estimation with geometric constraints [Arxiv: 1404.3749] (submitted)
61. N. Rao, B. Recht, R. Nowak, Tight measurement bounds for exact recovery of structured sparse signals, in *Proceedings of AISTATS* (2012)
62. H. Rauhut, K. Schnass, P. Vandergheynst, Compressed sensing and redundant dictionaries. IEEE Trans. Inf. Theory **54**, 2210–2219 (2008)
63. B. Recht, A simpler approach to matrix completion. J. Mach. Learn. Res. **12**, 3413–3430 (2011)
64. R. Rockafellar, *Convex Analysis*. Princeton Mathematical Series, vol. 28 (Princeton University Press, Princeton, 1970)
65. M. Rudelson, R. Vershynin, Combinatorics of random processes and sections of convex bodies. Ann. Math. **164**, 603–648 (2006)
66. M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**, 1025–1045 (2008)
67. Y. Seginer, The expected norm of random matrices. Comb. Probab. Comput. **9**, 149–166 (2000)
68. N. Srebro, N. Alon, T. Jaakkola, Generalization error bounds for collaborative prediction with low-rank matrices, in *Proceedings of Advances in Neural Information Processing Systems (NIPS)*, vol. 17 (2005)
69. M. Stojnic, Various thresholds for ℓ_1-optimization in compressed sensing (2009) [Arxiv: 0907.3666]
70. M. Talagrand, Regularity of Gaussian processes. Acta Math. **159**, 99–149 (1987)
71. M. Talagrand, *The Generic Chaining. Upper and Lower Bounds of Stochastic Processes*. Springer Monographs in Mathematics (Springer, Berlin, 2005)
72. J. Tropp, Convex recovery of a structured signal from independent random linear measurements, in *Sampling Theory, a Renaissance* (to appear)
73. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, in *Compressed Sensing* (Cambridge University Press, Cambridge, 2012), pp. 210–268
74. M. Wainwright, Structured regularizers for high-dimensional problems: statistical and computational issues. Ann. Rev. Stat. Appl. **1**, 233–253 (2014)

Chapter 2
Convex Recovery of a Structured Signal from Independent Random Linear Measurements

Joel A. Tropp

Abstract This chapter develops a theoretical analysis of the convex programming method for recovering a structured signal from independent random linear measurements. This technique delivers bounds for the sampling complexity that are similar to recent results for standard Gaussian measurements, but the argument applies to a much wider class of measurement ensembles. To demonstrate the power of this approach, the chapter presents a short analysis of phase retrieval by trace-norm minimization. The key technical tool is a framework, due to Mendelson and coauthors, for bounding a nonnegative empirical process.

2.1 Motivation

Signal reconstruction from random measurements is a central preoccupation in contemporary signal processing. In this problem, we acquire linear measurements of an unknown, structured signal through a random sampling process. Given these random measurements, a standard method for recovering the unknown signal is to solve a convex optimization problem that enforces our prior knowledge about the structure. The basic question is how many measurements suffice to resolve a particular type of structure.

Recent research has led to a comprehensive answer when the measurement operator follows the standard Gaussian distribution [1, 6, 10, 22, 24–26, 29, 31, 33]. The literature also contains satisfying answers for sub-Gaussian measurements [22] and subexponential measurements [18]. Other types of measurement systems are quite common, but we are not aware of a simple approach that allows us to analyze general measurements in a unified way.

This chapter describes an approach that addresses a wide class of convex signal reconstruction problems involving random sampling. To understand these questions, the core challenge is to produce a lower bound on a nonnegative empirical process.

J.A. Tropp (✉)
Department of Computing and Mathematical Sciences, California Institute of Technology, 1200 E. California Blvd. Pasadena, CA, 91125-5000

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_2

For this purpose, we rely on a powerful framework, called the *Small Ball Method*, that was developed by Shahar Mendelson and coauthors in a sequence of papers, including [14, 16, 19–21].

To complete the estimates required by Mendelson's Small Ball Method, we propose a technique based on conic duality. One advantage of this approach is that we can exploit the same insights and calculations that have served so well in the Gaussian setting. We refer to this little argument as the *bowling scheme* in honor of David Gross's *golfing scheme* [13]. We anticipate that it will offer researchers an effective way to analyze many signal recovery problems with random measurements.

2.1.1 Roadmap

The first half of the chapter summarizes the established analysis of convex signal reconstruction with a Gaussian sampling model. In Section 2.2, we introduce a convex optimization framework for solving structured signal recovery problems with linear measurements, and we present a geometric formulation of the optimality conditions. Section 2.3 specializes to the case where the measurements come from a Gaussian model, and we explain how classical results for Gaussian processes lead to a sharp bound for the number of Gaussian measurements that suffice. These results are framed in terms of a geometric parameter, the conic Gaussian width, associated with the convex optimization problem. Section 2.4 explains how to use duality to obtain a numerically sharp bound for the conic Gaussian width, and it develops two important examples in detail.

In the second half of the chapter, we consider more general sampling models. Section 2.5 introduces Mendelson's Small Ball Method and the technical arguments that support it. As a first application, in Section 2.6, we use this strategy to analyze signal reconstruction from sub-Gaussian measurements. Section 2.7 presents the bowling scheme, which merges the conic duality estimates with Mendelson's Small Ball Method. This technique allows us to study more general types of random measurements. Finally, in Section 2.8, we demonstrate the vigor of these ideas by applying them to the phase retrieval problem.

2.2 Signal reconstruction from linear measurements

We begin with a framework that describes many convex optimization methods for recovering a structured signal from linear measurements. Examples include the ℓ_1 minimization approach for identifying a sparse vector and the Schatten 1-norm minimization approach for identifying a low-rank matrix. We develop a simple error bound for convex signal reconstruction by exploiting the geometric formulation of the optimality conditions. This analysis leads us to study the minimum conic singular value of a matrix.

2.2.1 Linear acquisition of data

Let $\boldsymbol{x}^\natural \in \mathbb{R}^d$ be an unknown but "structured" signal. Suppose that we observe a vector $\boldsymbol{y}$ in $\mathbb{R}^m$ that consists of m linear measurements of the unknown:

$$\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^\natural + \boldsymbol{e}. \tag{2.1}$$

We assume that $\boldsymbol{\Phi}$ is a known $m \times d$ sampling matrix, and $\boldsymbol{e} \in \mathbb{R}^m$ is a vector of unknown errors. Expression (2.1) offers a model for data acquisition that describes a wide range of problems in signal processing, statistics, and machine learning. Our goal is to compute an approximation of the unknown $\boldsymbol{x}^\natural$ by exploiting our prior knowledge about its structure.

2.2.2 Reconstruction via convex optimization

Convex optimization is a popular approach for recovering a structured vector from linear measurements. Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper convex function[1] that reflects the "complexity" of a signal. Then we can frame the convex program

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad \|\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y}\| \le \eta \tag{2.2}$$

where $\|\cdot\|$ denotes the Euclidean norm and η is a specified bound on the norm of the error $\boldsymbol{e}$. In words, the optimization problem (2.2) searches for the most structured signal $\boldsymbol{x}$ that is consistent with the observed data $\boldsymbol{y}$. In practice, it is common to consider the Lagrangian formulation of (2.2) or to consider a problem where the objective and constraint are interchanged. We can often solve (2.2) and its variants efficiently using standard algorithms.

Remark 2.1 (Alternative programs). The optimization problem (2.2) is not the only type of convex method for signal reconstruction. Suppose that $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ is a gauge, i.e., a function that is nonnegative, positively homogeneous, and convex. Then we may consider the convex program

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad f(\boldsymbol{x}) \quad \text{subject to} \quad f^\circ\big(\boldsymbol{\Phi}^{\mathsf{t}}(\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y})\big) \le \eta,$$

where f° denotes the polar of the gauge [28, Chap. 15] and $^{\mathsf{t}}$ denotes transposition. This reconstruction method submits to an analysis similar to the approach in this note. For example, see [4, Thm. 1].

[1] The extended real numbers $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$. A *proper* convex function takes at least one finite value but never the value $-\infty$.

2.2.3 Examples

Before we continue, let us mention a few structures that arise in applications and the complexity measures that are typically associated with these structures.

Example 2.2 (Sparse vectors). A vector $\boldsymbol{x}^\natural \in \mathbb{R}^d$ is *sparse* when many or most of its entries are equal to zero. We can promote sparsity by minimizing the ℓ_1 norm $\|\cdot\|_{\ell_1}$. This heuristic leads to a problem of the form

$$\underset{\boldsymbol{x}\in\mathbb{R}^d}{\text{minimize}} \quad \|\boldsymbol{x}\|_{\ell_1} \quad \text{subject to} \quad \|\boldsymbol{\Phi}\boldsymbol{x} - \boldsymbol{y}\| \le \eta. \tag{2.3}$$

Sparsity has become a dominant modeling tool in statistics, machine learning, and signal processing.

Example 2.3 (Low-rank matrices). We say that a matrix $\boldsymbol{X}^\natural \in \mathbb{R}^{d_1\times d_2}$ has *low rank* when its rank is small compared with minimum of d_1 and d_2. Suppose that we have acquired noisy measurements

$$\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{X}^\natural) + \boldsymbol{e}, \tag{2.4}$$

where $\boldsymbol{\Phi}$ is a linear operator that maps a matrix in $\mathbb{R}^{d_1\times d_2}$ to a vector in $\mathbb{R}^m$. To reconstruct the unknown low-rank matrix $\boldsymbol{X}^\natural$, we can minimize the Schatten 1-norm $\|\cdot\|_{S_1}$, which returns the sum of the singular values of a matrix. This heuristic suggests that we consider an optimization problem of the form

$$\underset{\boldsymbol{X}\in\mathbb{R}^{d_1\times d_2}}{\text{minimize}} \quad \|\boldsymbol{X}\|_{S_1} \quad \text{subject to} \quad \|\boldsymbol{\Phi}(\boldsymbol{X}) - \boldsymbol{y}\| \le \eta. \tag{2.5}$$

In recent years, this approach to fitting low-rank matrices has become common.

It is possible to consider many other types of structures. For instance, see [6, 10].

2.2.4 A deterministic error bound for convex recovery

We can obtain a deterministic error bound for the convex reconstruction method (2.2) using a standard geometric analysis. Recall that a *cone* is a set $K \subset \mathbb{R}^d$ that is positively homogeneous: $K = \tau K$ for all $\tau > 0$. A *convex cone* is a cone that is also a convex set. Let us introduce the cone of descent directions of a convex function.

Definition 2.4 (Descent cone). Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper convex function. The *descent cone* $\mathscr{D}(f, \boldsymbol{x})$ of the function f at a point $\boldsymbol{x} \in \mathbb{R}^d$ is defined as

$$\mathscr{D}(f, \boldsymbol{x}) := \bigcup_{\tau>0} \{\boldsymbol{u} \in \mathbb{R}^d : f(\boldsymbol{x} + \tau\boldsymbol{u}) \le f(\boldsymbol{x})\}.$$

The descent cone of a convex function is always a convex cone, but it may not be closed.

We are interested in the behavior of the measurement matrix $\mathbf{\Phi}$ when it is restricted to a descent cone.

Definition 2.5 (Minimum conic singular value). Let $\mathbf{\Phi}$ be an $m \times d$ matrix, and let K be a cone in $\mathbb{R}^d$. The minimum singular value of $\mathbf{\Phi}$ with respect to the cone K is defined as

$$\lambda_{\min}(\mathbf{\Phi}; K) := \inf\left\{ \|\mathbf{\Phi}\boldsymbol{u}\| : \boldsymbol{u} \in K \cap \mathsf{S}^{d-1} \right\}$$

where S^{d-1} is the Euclidean unit sphere in $\mathbb{R}^d$.

The terminology originates in the fact that $\lambda_{\min}(\mathbf{\Phi}; \mathbb{R}^d)$ coincides with the usual minimum singular value.

With these definitions at hand, we reach the following basic result.

Proposition 2.6 (A deterministic error bound for convex recovery). *Let $\boldsymbol{x}^\natural$ be a signal in $\mathbb{R}^d$, let $\mathbf{\Phi}$ be an $m \times d$ measurement matrix, and let $\boldsymbol{y} = \mathbf{\Phi}\boldsymbol{x}^\natural + \boldsymbol{e}$ be a vector of measurements in $\mathbb{R}^m$. Assume that $\|\boldsymbol{e}\| \leq \eta$, and let $\hat{\boldsymbol{x}}_\eta$ be any solution to the optimization problem (2.2). Then*

$$\left\|\hat{\boldsymbol{x}}_\eta - \boldsymbol{x}^\natural\right\| \leq \frac{2\eta}{\lambda_{\min}\big(\mathbf{\Phi};\ \mathscr{D}(f, \boldsymbol{x}^\natural)\big)}.$$

This statement is adapted from [6]. For completeness, we include the short proof.

Proof. It is natural to write the decision variable $\boldsymbol{x}$ in the convex program (2.2) relative to the true unknown: $\boldsymbol{u} := \boldsymbol{x} - \boldsymbol{x}^\natural$. Using expression (2.1) for the measurement vector $\boldsymbol{y}$, we obtain the equivalent problem

$$\underset{\boldsymbol{u} \in \mathbb{R}^d}{\text{minimize}} \quad f(\boldsymbol{x}^\natural + \boldsymbol{u}) \quad \text{subject to} \quad \|\mathbf{\Phi}\boldsymbol{u} - \boldsymbol{e}\| \leq \eta. \tag{2.6}$$

Owing to the bound $\|\boldsymbol{e}\| \leq \eta$, the point $\boldsymbol{u} = \mathbf{0}$ is feasible for (2.6). Therefore, each optimal point $\hat{\boldsymbol{u}}$ verifies $f(\boldsymbol{x}^\natural + \hat{\boldsymbol{u}}) \leq f(\boldsymbol{x}^\natural)$. In summary, any optimal point of (2.6) satisfies two conditions:

$$\hat{\boldsymbol{u}} \in \mathscr{D}(f, \boldsymbol{x}^\natural) \quad \text{and} \quad \|\mathbf{\Phi}\hat{\boldsymbol{u}} - \boldsymbol{e}\| \leq \eta.$$

As a consequence, we simply need to determine how far we can travel in a descent direction before we violate the bound constraint. See Figure 2.1 for an illustration of the geometry.

To complete the argument, assume that $\boldsymbol{u}$ is a nonzero point in $\mathscr{D}(f, \boldsymbol{x}^\natural)$ that is feasible for (2.6). Then

$$\lambda_{\min}\big(\mathbf{\Phi};\ \mathscr{D}(f, \boldsymbol{x}^\natural)\big) \leq \frac{\|\mathbf{\Phi}\boldsymbol{u}\|}{\|\boldsymbol{u}\|} \leq \frac{\|\mathbf{\Phi}\boldsymbol{u} - \boldsymbol{e}\| + \|\boldsymbol{e}\|}{\|\boldsymbol{u}\|} \leq \frac{2\eta}{\|\boldsymbol{u}\|}.$$

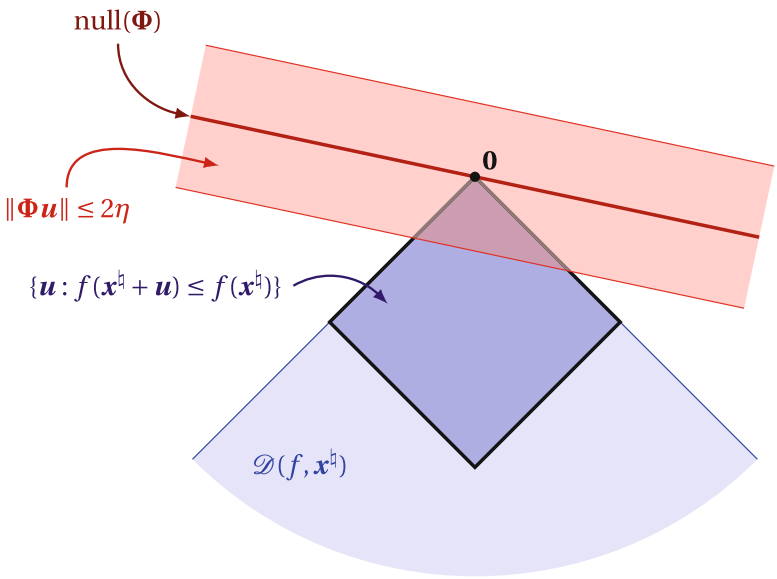


Fig. 2.1 [Geometry of convex recovery] This diagram illustrates the geometry of the optimization problem (2.6). The cone $\mathcal{D}(f, \boldsymbol{x}^\natural)$ contains the directions $\boldsymbol{u}$ in which f is decreasing at $\boldsymbol{x}^\natural$. Assuming that $\|\boldsymbol{e}\| \le \eta$, the diagonal tube contains every point $\boldsymbol{u}$ that satisfies the bound constraint $\|\boldsymbol{\Phi u} + \boldsymbol{e}\| \le \eta$. Each optimal point $\hat{\boldsymbol{u}}$ for (2.6) lies in the intersection of the tube and the cone.

The first inequality follows from Definition 2.5 of the conic singular value. The second relation is the triangle inequality. The last bound holds because $\boldsymbol{u}$ satisfies the constraint in (2.6), and we have assumed that $\|\boldsymbol{e}\| \le \eta$. Finally, rearrange the display, and rewrite $\boldsymbol{u}$ in terms of the original decision variable $\boldsymbol{x}$. □

Although Proposition 2.6 is elegant, it can be difficult to apply because we must calculate the minimum conic singular value of a matrix $\boldsymbol{\Phi}$ with respect to a descent cone. This challenge becomes less severe, however, when the matrix $\boldsymbol{\Phi}$ is drawn at random.

2.3 A universal error bound for Gaussian measurements

We will study the prospects for convex recovery when the sampling matrix $\boldsymbol{\Phi}$ is chosen at random. This modeling assumption arises in signal processing applications where the matrix describes a data-acquisition system that can extract random measurements. This kind of model also appears in statistics and machine learning when each row of the matrix tabulates measured variables for an individual subject in an experiment.

2.3.1 Standard Gaussian measurements

In this section, we treat one of the simplest mathematical models for the $m \times d$ random measurement matrix $\mathbf{\Phi}$. We assume that each of the m rows of $\mathbf{\Phi}$ is drawn independently from the standard Gaussian distribution $\text{NORMAL}(\mathbf{0}, \mathbf{I}_d)$, where the covariance $\mathbf{I}_d$ is the d-dimensional identity matrix. For this special case, we can obtain a sharp estimate for the minimum conic singular value $\lambda_{\min}(\mathbf{\Phi}; K)$ for any convex cone K.

2.3.2 The conic Gaussian width

The analysis of Gaussian sampling depends on a geometric summary parameter for cones.

Definition 3.1 (Conic Gaussian width). Let $K \subset \mathbb{R}^d$ be a cone, not necessarily convex. The *conic Gaussian width* $w(K)$ is defined as

$$w(K) := \mathbb{E} \sup_{\boldsymbol{u} \in K \cap \mathsf{S}^{d-1}} \langle \boldsymbol{g}, \boldsymbol{u} \rangle$$

where $\boldsymbol{g} \sim \text{NORMAL}(\mathbf{0}, \mathbf{I}_d)$ is a standard Gaussian vector in $\mathbb{R}^d$.

The Gaussian width plays a central role in asymptotic convex geometry [17, 23, 27]. Most of the classical techniques for bounding widths are only accurate up to constant factors (or worse). In contrast, ideas from the contemporary signal processing literature frequently allow us to produce numerically sharp estimates for the Gaussian width of a cone. These techniques were developed in the papers [1, 6, 10, 24, 31]. We will outline one of the methods in Section 2.4.

Remark 3.2 (Statistical dimension). The conic Gaussian width $w(K)$ is a convenient functional because it arises from the probabilistic tools that we use. The theory of conic integral geometry, however, delivers a better summary parameter [1]. The *statistical dimension* $\delta(K)$ of a convex cone K can be defined as

$$\delta(K) := \mathbb{E}\left[\left(\sup_{\boldsymbol{u} \in K \cap \mathsf{B}^d} \langle \boldsymbol{g}, \boldsymbol{u} \rangle\right)^2\right],$$

where B^d is the Euclidean unit ball in $\mathbb{R}^d$ and $\boldsymbol{g} \sim \text{NORMAL}(\mathbf{0}, \mathbf{I}_d)$. The statistical dimension canonically extends the dimension of a subspace to the class of convex cones, and it satisfies many elegant identities [1, Prop. 3.1]. For some purposes, the two parameters are interchangeable because of the following comparison [1, Prop. 10.2]:

$$w^2(K) \le \delta(K) \le w^2(K) + 1.$$

As a consequence, we can interpret $w^2(K)$ as a rough measure of the "dimension" of a cone.

2.3.3 Conic singular values and conic Gaussian widths

As it turns out, the conic Gaussian width $w(K)$ controls the minimum conic singular value $\lambda_{\min}(\boldsymbol{\Phi}; K)$ when $\boldsymbol{\Phi}$ follows the standard normal distribution.

Proposition 3.3 (Minimum conic singular value of a Gaussian matrix). *Let $K \subset \mathbb{R}^d$ be a cone, not necessarily convex, and let $\boldsymbol{\Phi}$ be an $m \times d$ matrix whose rows are independent vectors drawn from the standard Gaussian distribution* $\text{NORMAL}(\mathbf{0}, \mathbf{I}_d)$. *Then*

$$\lambda_{\min}(\boldsymbol{\Phi}; K) \geq \sqrt{m-1} - w(K) - t$$

with probability at least $1 - \mathrm{e}^{-t^2/2}$.

In essence, this result dates to the work of Gordon [11, 12]. We have drawn the proof from the survey [8, Sec. 3.2] of Davidson & Szarek; see also [6, 22, 29, 31]. Note that the argument relies on special results for Gaussian processes that do not extend to other distributions.

Proof sketch. We can express the minimum conic singular value as

$$\lambda_{\min}(\boldsymbol{\Phi}; K) = \inf_{\boldsymbol{u} \in K \cap \mathbb{S}^{d-1}} \sup_{\boldsymbol{v} \in \mathbb{S}^{m-1}} \langle \boldsymbol{v},\ \boldsymbol{\Phi}\boldsymbol{u} \rangle$$

It is a consequence of Gordon's comparison inequality [11, Thm. 1.4] that

$$\mathbb{E} \inf_{\boldsymbol{u} \in K \cap \mathbb{S}^{d-1}} \sup_{\boldsymbol{v} \in \mathbb{S}^{m-1}} \langle \boldsymbol{v},\ \boldsymbol{\Phi}\boldsymbol{u} \rangle \geq \mathbb{E} \sup_{\boldsymbol{v} \in \mathbb{S}^{m-1}} \langle \boldsymbol{g}',\ \boldsymbol{v} \rangle - \mathbb{E} \sup_{\boldsymbol{u} \in K \cap \mathbb{S}^{d-1}} \langle \boldsymbol{g},\ \boldsymbol{u} \rangle = \mathbb{E}\ \|\boldsymbol{g}'\| - w(K),$$

where $\boldsymbol{g}' \sim \text{NORMAL}(\mathbf{0}, \mathbf{I}_m)$ and $\boldsymbol{g} \sim \text{NORMAL}(\mathbf{0}, \mathbf{I}_d)$. It is well known that $\mathbb{E}\ \|\boldsymbol{g}'\| \geq \sqrt{m-1}$, and therefore

$$\mathbb{E}\, \lambda_{\min}(\boldsymbol{\Phi}; K) \geq \sqrt{m-1} - w(K). \tag{2.7}$$

To complete the argument, note that the map

$$\lambda_{\min}(\cdot; K) : \boldsymbol{A} \mapsto \inf_{\boldsymbol{u} \in K \cap \mathbb{S}^{d-1}} \|\boldsymbol{A}\boldsymbol{u}\|$$

is 1-Lipschitz with respect to the Frobenius norm. The usual Gaussian concentration inequality [3, Sec. 5.4] implies that

$$\mathbb{P}\big\{\lambda_{\min}(\boldsymbol{\Phi}; K) \leq \mathbb{E}\, \lambda_{\min}(\boldsymbol{\Phi}; K) - t\big\} \leq \mathrm{e}^{-t^2/2}. \tag{2.8}$$

Introduce the lower bound (2.7) for the expectation of the minimum conic singular value into (2.8) to reach the advertised result. □

Remark 3.4 (Sharpness for convex cones). It is a remarkable fact that the bound in Proposition 3.3 is essentially sharp. For any cone K, we can reinterpret the statement as saying that

$$\lambda_{\min}(\boldsymbol{\Phi}; K) > 0 \quad \text{with high probability when} \quad m \geq w^2(K) + Cw(K).$$

(The letter C always denotes a positive absolute constant, but its value may change from place to place.) Conversely, for a convex cone K, it holds that

$$\lambda_{\min}(\boldsymbol{\Phi}; K) = 0 \quad \text{with high probability when} \quad m \leq w^2(K) - Cw(K). \tag{2.9}$$

Result (2.9) follows from research of Amelunxen et al. [1, Thm. I and Prop. 10.2]. This claim can also be derived by supplementing the proof of Proposition 3.3 with a short polarity argument. It is productive to interpret the pair of estimates in this remark as a *phase transition* for convex signal recovery; see [1] for more information.

2.3.4 An error bound for Gaussian measurements

Combining Proposition 2.6 and Proposition 3.3, we obtain a general error bound for convex recovery from Gaussian measurements.

Corollary 3.5 (Signal recovery from Gaussian measurements). *Let $\boldsymbol{x}^\natural$ be a signal in $\mathbb{R}^d$. Let $\boldsymbol{\Phi}$ be an $m \times d$ matrix whose rows are independent random vectors drawn from the standard Gaussian distribution* $\text{NORMAL}(\mathbf{0}, \mathbf{I}_d)$*, and let $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^\natural + \boldsymbol{e}$ be a vector of measurements in $\mathbb{R}^m$. With probability at least $1 - \mathrm{e}^{-t^2/2}$, the following statement holds. Assume that $\|\boldsymbol{e}\| \leq \eta$, and let $\hat{\boldsymbol{x}}_\eta$ be any solution to the optimization problem (2.2). Then*

$$\left\|\hat{\boldsymbol{x}}_\eta - \boldsymbol{x}^\natural\right\| \leq \frac{2\eta}{\left[\sqrt{m-1} - w\big(\mathscr{D}(f, \boldsymbol{x}^\natural)\big) - t\right]_+}.$$

The operation $[a]_+ := \max\{a, 0\}$ returns the positive part of a number.

The overall argument that leads to this result was proposed by Rudelson & Vershynin [29, Sec. 4]; the statement here is adapted from [6].

Corollary 3.5 provides for stable recovery of the unknown $\boldsymbol{x}^\natural$ when the number m of measurements satisfies

$$m \geq w^2\big(\mathscr{D}(f, \boldsymbol{x}^\natural)\big) + Cw\big(\mathscr{D}(f, \boldsymbol{x}^\natural)\big).$$

In view of Remark 3.4, Corollary 3.5 provides a refined estimate for the amount of information that suffices to identify a structured vector from Gaussian measurements via convex optimization.

Remark 3.6 (The normal error model). It is possible to improve the error bound in Corollary 3.5 if we instate a Gaussian model for the error vector $\boldsymbol{e}$. See the papers [25, 26, 33] for an analysis of this case.

2.4 Controlling the width of a descent cone via polarity

As soon as we know the conic Gaussian width of the descent cone, Corollary 6.4 yields error bounds for convex recovery of a structured signal from Gaussian measurements. To make use of this result, we need technology for calculating these widths. This section describes a mechanism, based on polarity, that leads to extremely accurate estimates. We can trace this method to the papers [24, 31], where it is couched in the language of duality for cone programs. The subsequent papers [1, 6] rephrase these ideas in a more geometric fashion. It can be shown that the approach in this section gives sharp results for many natural examples; see [1, Thm. 4.3] or [10, Prop. 1]. Although polar bounds for widths are classic in asymptotic convex geometry [17, 23, 27], the refined arguments here are just a few years old.

2.4.1 Polarity and weak duality for cones

We begin with some classical facts about conic geometry.

Fact 4.1 (Polarity). *Let K be a general cone in $\mathbb{R}^d$. The* polar cone K° *is the closed convex cone*

$$K^\circ := \big\{ \boldsymbol{v} \in \mathbb{R}^d : \langle \boldsymbol{v}, \boldsymbol{x} \rangle \le 0 \text{ for all } \boldsymbol{x} \in K \big\}.$$

It is easy to verify that $K \subset (K^\circ)^\circ$.

Recall that the *distance* from a point $\boldsymbol{x} \in \mathbb{R}^d$ to a set $E \subset \mathbb{R}^d$ is defined by the relation

$$\operatorname{dist}(\boldsymbol{x}, E) := \inf_{\boldsymbol{u} \in E} \|\boldsymbol{x} - \boldsymbol{u}\| .$$

With these definitions, we reach the following weak duality result.

Proposition 4.2 (Weak duality for cones). *Let K be a general cone in $\mathbb{R}^d$. For $\boldsymbol{x} \in \mathbb{R}^d$,*

$$\sup_{\boldsymbol{u} \in K \cap \mathbb{S}^{d-1}} \langle \boldsymbol{x}, \boldsymbol{u} \rangle \le \operatorname{dist}(\boldsymbol{x}, K^\circ).$$

Proof. The argument is based on a simple duality trick. First, write

$$\operatorname{dist}(\boldsymbol{x}, K^\circ) = \inf_{\boldsymbol{v} \in K^\circ} \|\boldsymbol{x} - \boldsymbol{v}\| = \inf_{\boldsymbol{v} \in K^\circ} \sup_{\boldsymbol{u} \in \mathsf{S}^{d-1}} \langle \boldsymbol{x} - \boldsymbol{v},\, \boldsymbol{u} \rangle .$$

Apply the inf–sup inequality:

$$\operatorname{dist}(\boldsymbol{x}, K^\circ) \geq \sup_{\boldsymbol{u} \in \mathsf{S}^{d-1}} \inf_{\boldsymbol{v} \in K^\circ} \langle \boldsymbol{x} - \boldsymbol{v},\, \boldsymbol{u} \rangle = \sup_{\boldsymbol{u} \in \mathsf{S}^{d-1}} \left[\langle \boldsymbol{x},\, \boldsymbol{u} \rangle - \sup_{\boldsymbol{v} \in K^\circ} \langle \boldsymbol{v},\, \boldsymbol{u} \rangle \right].$$

By definition of polarity, the inner supremum takes the value $+\infty$ unless $\boldsymbol{u} \in (K^\circ)^\circ$. We determine that

$$\operatorname{dist}(\boldsymbol{x}, K^\circ) \geq \sup_{\boldsymbol{u} \in (K^\circ)^\circ \cap \mathsf{S}^{d-1}} \langle \boldsymbol{x},\, \boldsymbol{u} \rangle \geq \sup_{\boldsymbol{u} \in K \cap \mathsf{S}^{d-1}} \langle \boldsymbol{x},\, \boldsymbol{u} \rangle .$$

The last inequality holds because $K \subset (K^\circ)^\circ$. □

Remark 4.3 (Strong duality for cones). If K is a convex cone and we replace the sphere with a ball, then we have strong duality instead:

$$\sup_{\boldsymbol{u} \in K \cap \mathsf{B}^d} \langle \boldsymbol{x},\, \boldsymbol{u} \rangle = \operatorname{dist}(\boldsymbol{x}, K^\circ).$$

The proof uses Sion's minimax theorem [30] and the bipolar theorem [28, Thm. 14.1].

2.4.2 The conic Gaussian width of a descent cone

We can use Proposition 4.2 to obtain an effective bound for the width of a descent cone. This approach is based on a classical polarity correspondence [28, Thm. 23.7].

Fact 4.4 (Polarity for descent cones). *The* subdifferential *of a proper convex function $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ at a point $\boldsymbol{x} \in \mathbb{R}^d$ is the closed convex set*

$$\partial f(\boldsymbol{x}) := \left\{ \boldsymbol{v} \in \mathbb{R}^d : f(\boldsymbol{y}) \geq f(\boldsymbol{x}) + \langle \boldsymbol{v},\, \boldsymbol{y} - \boldsymbol{x} \rangle \text{ for all } \boldsymbol{y} \in \mathbb{R}^d \right\}.$$

Assume that the subdifferential $\partial f(\boldsymbol{x})$ is nonempty and does not contain the origin. Then

$$\mathscr{D}(f, \boldsymbol{x})^\circ = \overline{\operatorname{cone}}(\partial f(\boldsymbol{x})) := \operatorname{closure}\left(\bigcup_{\tau \geq 0} \tau \cdot \partial f(\boldsymbol{x}) \right). \tag{2.10}$$

Combining Proposition 4.2 and Fact 4.4, we reach a bound for the conic Gaussian width of a descent cone.

Proposition 4.5 (The width of a descent cone). *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper convex function, and fix a point $\boldsymbol{x} \in \mathbb{R}^d$. Assume that the subdifferential $\partial f(\boldsymbol{x})$ is nonempty and does not contain the origin. Then*

$$w^2\big(\mathscr{D}(f,\boldsymbol{x})\big) \le \mathbb{E} \inf_{\tau \ge 0} \operatorname{dist}^2\big(\boldsymbol{g},\ \tau \cdot \partial f(\boldsymbol{x})\big)$$

Several specific instances of Proposition 4.5 appear in [6, App. C], while the general statement here is adapted from [1, Sec. 4.1]. Sections 2.4.3 and 2.4.4 exhibit how Proposition 4.5 works.

Proof. Proposition 4.2 implies that

$$w\big(\mathscr{D}(f,\boldsymbol{x})\big) = \mathbb{E} \sup_{\boldsymbol{u} \in \mathscr{D}(f,\boldsymbol{x}) \cap \mathbb{S}^{d-1}} \langle \boldsymbol{g},\ \boldsymbol{u} \rangle \le \mathbb{E} \operatorname{dist}\big(\boldsymbol{g},\ \mathscr{D}(f,\boldsymbol{x})^\circ\big).$$

Expression (2.10) for the polar of a descent cone implies that

$$w\big(\mathscr{D}(f,\boldsymbol{x})\big) \le \mathbb{E} \operatorname{dist}\left(\boldsymbol{g},\ \operatorname{closure}\left(\bigcup_{\tau \ge 0} \tau \cdot \partial f(\boldsymbol{x})\right)\right) = \mathbb{E} \inf_{\tau \ge 0} \operatorname{dist}\big(\boldsymbol{g},\ \tau \cdot \partial f(\boldsymbol{x})\big).$$

Indeed, the distance to a set is the same as the distance to its closure, and the distance to a union is the infimal distance to one of its members. Square the latter display and apply Jensen's inequality to complete the argument. □

2.4.3 Example: Sparse vectors

Suppose that $\boldsymbol{x}^\natural$ is a vector in $\mathbb{R}^d$ with s nonzero entries. Let $\boldsymbol{\Phi}$ be an $m \times d$ matrix whose rows are independent random vectors distributed as $\textsc{normal}(\boldsymbol{0}, \mathbf{I}_d)$, and suppose that we acquire a vector $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^\natural + \boldsymbol{e}$ consisting of m noisy measurements. We can solve the ℓ_1-minimization problem (2.3) in an attempt to reconstruct $\boldsymbol{x}^\natural$.

How many measurements are sufficient to ensure that this approach succeeds? We will demonstrate that

$$w^2\big(\mathscr{D}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^\natural)\big) \le 2s \log(d/s) + 2s. \tag{2.11}$$

Therefore, Corollary 3.5 implies that $m \gtrsim 2s \log(d/s)$ measurements are enough for us to recover $\boldsymbol{x}^\natural$ approximately. When $s \ll d$, the first term in (2.11) is numerically sharp because of [10, Prop. 1].

2.4.3.1 The width calculation

Let us establish the width bound (2.11). This analysis is adapted from [6, App. C] and [1, App. D.2]; see also [10, App. B]. The result [1, Prop. 4.5] contains a more complicated formula for the width that is sharp for all choices of the sparsity s.

When estimating widths, a useful strategy is to change coordinates so that the calculations are more transparent. The ℓ_1 norm is invariant under signed permutation, so

$$\mathscr{D}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^\natural) = \boldsymbol{P}^{\mathsf{t}}\, \mathscr{D}(\|\cdot\|_{\ell_1}, \boldsymbol{P}\boldsymbol{x}^\natural) \quad \text{where } \boldsymbol{P} \text{ is a signed permutation.}$$

The distribution of a standard Gaussian random variable is invariant under signed permutation, so the conic Gaussian width has the same invariance. Therefore,

$$w\big(\mathscr{D}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^\natural)\big) = w\big(\boldsymbol{P}^{\mathsf{t}}\, \mathscr{D}(\|\cdot\|_{\ell_1}, \boldsymbol{P}\boldsymbol{x}^\natural)\big) = w\big(\mathscr{D}(\|\cdot\|_{\ell_1}, \boldsymbol{P}\boldsymbol{x}^\natural)\big).$$

We will use this type of transformation several times without detailed justification.

As a consequence of the argument in the last paragraph, we may assume that $\boldsymbol{x}^\natural$ takes the form

$$\boldsymbol{x}^\natural = (x_1, \dots, x_s, 0, \dots, 0)^{\mathsf{t}} \in \mathbb{R}^d \quad \text{where} \quad x_1 \geq \cdots \geq x_s > 0.$$

Proposition 4.5 ensures that

$$w^2\big(\mathscr{D}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^\natural)\big) \leq \mathbb{E}\, \operatorname{dist}^2\big(\boldsymbol{g},\ \tau \cdot \partial\, \|\boldsymbol{x}^\natural\|_{\ell_1}\big) \quad \text{for each } \tau \geq 0 \tag{2.12}$$

where $\boldsymbol{g} \sim \textsc{normal}(\boldsymbol{0}, \mathbf{I}_d)$. The subdifferential of the ℓ_1 norm at $\boldsymbol{x}^\natural$ satisfies

$$\partial\, \|\boldsymbol{x}^\natural\|_{\ell_1} = \left\{ \begin{bmatrix} \mathbf{1}_s \\ \boldsymbol{y} \end{bmatrix} \in \mathbb{R}^d : \|\boldsymbol{y}\|_{\ell_\infty} \leq 1 \right\} \quad \text{where} \quad \mathbf{1}_s := (1, \dots, 1)^{\mathsf{t}} \in \mathbb{R}^s.$$

Therefore,

$$\mathbb{E}\, \operatorname{dist}^2(\boldsymbol{g},\ \tau \cdot \partial\, \|\boldsymbol{x}^\natural\|_{\ell_1}) = \sum_{j=1}^{s} \mathbb{E}\, \big(g_j - \tau\big)^2 + \sum_{j=s+1}^{d} \mathbb{E}\, \big[\, |g_j| - \tau \big]_+^2. \tag{2.13}$$

As usual, $[a]_+ := \max\{a, 0\}$. For $1 \leq j \leq s$, a direct calculation gives

$$\mathbb{E}\, \big(g_j - \tau\big)^2 = 1 + \tau^2. \tag{2.14}$$

For $s < j \leq d$, we apply a familiar tail bound for the standard normal variable to obtain

$$\begin{aligned} \mathbb{E}\, \big[\, |g_j|^2 - \tau \big]_+^2 &= \int_\tau^\infty 2(a - \tau)^2\, \mathbb{P}\big\{\, |g_j| \geq a \big\}\, \mathrm{d}a \\ &\leq \int_\tau^\infty 2a \left(\sqrt{\frac{2}{\pi}}\, a^{-1}\, \mathrm{e}^{-a^2/2} \right) \mathrm{d}a \\ &= 2\mathbb{P}\, \{|g_j\| \geq \tau\} \leq 2\mathrm{e}^{-\tau^2/2} \end{aligned} \tag{2.15}$$

Combine (2.12), (2.13), (2.14), and (2.15) to obtain

$$w^2\big(\mathscr{D}(\|\cdot\|_{\ell_1}, \boldsymbol{x}^\natural)\big) \le \mathbb{E}\,\operatorname{dist}^2(\boldsymbol{g},\ \tau \cdot \partial\,\|\boldsymbol{x}^\natural\|_{\ell_1}) = s \cdot \big(1+\tau^2\big) + (d-s) \cdot 2\mathrm{e}^{-\tau^2/2}.$$

Choose $\tau^2 = 2\log(d/s)$ and simplify to reach (2.11).

2.4.4 Example: Low-rank matrices

Let $\boldsymbol{X}^\natural$ be a matrix in $\mathbb{R}^{d_1 \times d_2}$ with rank r. Let $\boldsymbol{\Phi} : \mathbb{R}^{d_1 \times d_2} \to \mathbb{R}^m$ be a linear operator whose matrix has independent standard Gaussian entries. Suppose we acquire m noisy measurements of the form $\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{X}^\natural) + \boldsymbol{e}$. We can solve the S_1-minimization problem (2.5) to reconstruct $\boldsymbol{X}^\natural$.

How many measurements are enough to guarantee that this approach works? We will prove that

$$w^2\big(\mathscr{D}(\|\cdot\|_{S_1}, \boldsymbol{X}^\natural)\big) \le 3r \cdot (d_1 + d_2 - r). \tag{2.16}$$

As a consequence, Corollary 3.5 implies that $m \gtrsim 3r \cdot (d_1 + d_2 - r)$ measurements allow us to identify $\boldsymbol{X}^\natural$ approximately.

2.4.4.1 The width calculation

Let us establish the width bound (2.16). This analysis is adapted from [6, App. C] and [1, App. D.3]; see also [10, App. E]. The result [1, Prop. 4.6] contains a more complicated formula for the width that is sharp whenever the rank r is proportional to the dimension $\min\{d_1, d_2\}$.

The Schatten 1-norm is unitarily invariant, so we may also select a coordinate system where

$$\boldsymbol{X}^\natural = \begin{bmatrix} \boldsymbol{\Sigma} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix} \quad \text{where} \quad \boldsymbol{\Sigma} = \operatorname{diag}(\sigma_1, \dots, \sigma_r) \quad \text{and} \quad \sigma_j > 0 \text{ for } j = 1, \dots, r.$$

Let $\boldsymbol{G}$ be a $d_1 \times d_2$ matrix with independent standard normal entries, partitioned as

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{G}_{11} & \boldsymbol{G}_{12} \\ \boldsymbol{G}_{21} & \boldsymbol{G}_{22} \end{bmatrix} \quad \text{where} \quad \boldsymbol{G}_{11} \text{ is } r \times r \quad \text{and} \quad \boldsymbol{G}_{22} \text{ is } (d_1 - r) \times (d_2 - r).$$

Define a random parameter $\tau = \|\boldsymbol{G}_{22}\|$, where $\|\cdot\|$ denotes the spectral norm. Proposition 4.5 ensures that

$$w^2\big(\mathscr{D}(\|\cdot\|_{S_1}, \boldsymbol{X}^\natural)\big) \le \mathbb{E}\,\operatorname{dist}_{\mathrm{F}}^2\big(\boldsymbol{G},\ \tau \cdot \partial\,\|\boldsymbol{X}^\natural\|_{S_1}\big). \tag{2.17}$$

Note that we must calculate distance with respect to the Frobenius norm $\|\cdot\|_{\mathrm{F}}$. According to [37, Ex. 2], the subdifferential of the Schatten 1-norm takes the form

$$\partial \|\boldsymbol{X}^\natural\|_{S_1} = \left\{ \begin{bmatrix} \mathbf{I}_r & \mathbf{0} \\ \mathbf{0} & \boldsymbol{Y} \end{bmatrix} \in \mathbb{R}^{d_1 \times d_2} : \|\boldsymbol{Y}\| \le 1 \right\} \quad \text{where} \quad \mathbf{I}_r \text{ is the } r \times r \text{ identity matrix.}$$

We may calculate that

$$\begin{aligned} \mathbb{E}\, \operatorname{dist}_{\mathrm{F}}^2 \left(\boldsymbol{G},\ \tau \cdot \|\boldsymbol{X}^\natural\|_{S_1} \right) = \mathbb{E}\, \|\boldsymbol{G}_{11} - \tau \cdot \mathbf{I}_r\|_{\mathrm{F}}^2 + \mathbb{E}\, \|\boldsymbol{G}_{12}\|_{\mathrm{F}}^2 \\ + \mathbb{E}\, \|\boldsymbol{G}_{21}\|_{\mathrm{F}}^2 + \mathbb{E} \inf_{\|\boldsymbol{Y}\| \le 1} \|\boldsymbol{G}_{22} - \tau \cdot \boldsymbol{Y}\|_{\mathrm{F}}^2 . \end{aligned} \tag{2.18}$$

Our selection of τ ensures that the last term on the right-hand side of (2.18) vanishes. By direct calculation,

$$\mathbb{E}\, \|\boldsymbol{G}_{12}\|_{\mathrm{F}}^2 + \mathbb{E}\, \|\boldsymbol{G}_{21}\|_{\mathrm{F}}^2 = r \cdot (d_1 + d_2 - 2r). \tag{2.19}$$

To bound the first term on right-hand side of (2.18), observe that

$$\mathbb{E}\, \|\boldsymbol{G}_{11} - \tau \cdot \mathbf{I}_r\|_{\mathrm{F}}^2 = r^2 + r \cdot \mathbb{E}\, \tau^2 \tag{2.20}$$

because the random variable τ is independent of $\boldsymbol{G}_{11}$. We need to compute $\mathbb{E}\, \tau^2 = \mathbb{E}\, \|\boldsymbol{G}_{22}\|_{\mathrm{F}}^2$. A short argument [8, Sec. 2.3] based on the Slepian comparison inequality shows that

$$\mathbb{E}\, \|\boldsymbol{G}_{22}\| \le \sqrt{d_1 - r} + \sqrt{d_2 - r} \le \sqrt{2(d_1 + d_2 - 2r)}.$$

The spectral norm is 1-Lipschitz, so the Gaussian Poincaré inequality [3, Thm. 3.20] implies

$$\mathbb{E}\, \|\boldsymbol{G}_{22}\|^2 - \left(\mathbb{E}\, \|\boldsymbol{G}_{22}\| \right)^2 = \operatorname{Var}\left(\|\boldsymbol{G}_{22}\| \right) \le 1.$$

Combining the last two displays,

$$\mathbb{E}\, \tau^2 = \mathbb{E}\, \|\boldsymbol{G}_{22}\|^2 \le \left(\mathbb{E}\, \|\boldsymbol{G}_{22}\| \right)^2 + 1 \le 2\,(d_1 + d_2 - 2r) + 1. \tag{2.21}$$

Finally, we incorporate (2.18), (2.19), (2.20), (2.21) into the width bound (2.17) to reach

$$w^2\big(\mathscr{D}(\|\cdot\|_{S_1}, \boldsymbol{X}^\natural)\big) \le 3r \cdot (d_1 + d_2 - 2r) + r^2 + r.$$

Simplify this expression to obtain result (2.16).

2.5 Mendelson's Small Ball Method

In Sections 2.2–2.4, we analyzed a convex programming method for recovering structured signals from standard Gaussian measurements. The main result, Corollary 3.5, is appealing because it applies to any convex complexity measure f. Proposition 4.5 allows us to instantiate this result because it provides a mechanism for controlling the Gaussian width of a descent cone. On the other hand, this approach only works when the sampling matrix $\mathbf{\Phi}$ follows the standard Gaussian distribution.

For other sampling models, researchers use a variety of ad hoc techniques to study the recovery problem. It is common to see a separate and intricate argument for each new complexity measure f and each new distribution for $\mathbf{\Phi}$. It is natural to wonder whether there is a single approach that can address a broad class of complexity measures and sampling matrices.

The primary goal of this chapter is to analyze convex signal reconstruction with more general random measurements. Our argument is based on Mendelson's *Small Ball Method*, a powerful strategy for establishing a lower bound on a nonnegative empirical process [14, 16, 19–21]. This section contains an overview of Mendelson's Small Ball Method. Section 2.6 uses this technique to study sub-Gaussian measurement models. In Section 2.7, we extend these ideas to a larger class of sampling distributions. In Section 2.8, we conclude with an application to the problem of phase retrieval.

2.5.1 *The minimum conic singular value as a nonnegative empirical process*

Suppose that $\boldsymbol{\varphi}$ is a random vector on $\mathbb{R}^d$, and draw independent copies $\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_m$ of the random vector $\boldsymbol{\varphi}$. Form an $m \times d$ sampling matrix $\mathbf{\Phi}$ whose rows are these random vectors:

$$\mathbf{\Phi} = \begin{bmatrix} \boldsymbol{\varphi}_1^{\mathsf{t}} \\ \vdots \\ \boldsymbol{\varphi}_m^{\mathsf{t}} \end{bmatrix}. \tag{2.22}$$

Fix a cone $K \in \mathbb{R}^d$, not necessarily convex, and define the set $E := K \cap \mathsf{S}^{d-1}$. Then we can express the minimum conic singular value $\lambda_{\min}(\mathbf{\Phi}; K)$ of the sampling matrix as a nonnegative empirical process:

$$\lambda_{\min}(\mathbf{\Phi}; K) = \inf_{\boldsymbol{u} \in E} \left(\sum_{i=1}^{m} |\langle \boldsymbol{\varphi}_i, \boldsymbol{u} \rangle|^2 \right)^{1/2}. \tag{2.23}$$

When the sampling matrix is Gaussian, we can use Gordon's theorem [11, Thm. 1.4] to obtain a lower bound for expression (2.23), as in Proposition 3.3. The challenge is to find an alternative method for producing a lower bound in a more general setting.

2.5.2 A lower bound for nonnegative empirical processes

The main technical component in Mendelson's Small Ball Method is a remarkable estimate that was developed in the paper [20]. This result delivers an effective lower bound for a nonnegative empirical process.

Proposition 5.1 (Lower bound for a nonnegative empirical process [20, Thm. 5.4]). *Fix a set $E \subset \mathbb{R}^d$. Let $\boldsymbol{\varphi}$ be a random vector on $\mathbb{R}^d$, and let $\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_m$ be independent copies of $\boldsymbol{\varphi}$. Define the $m \times d$ matrix $\boldsymbol{\Phi}$ as on (2.22). Introduce the marginal tail function*

$$Q_\xi(E; \boldsymbol{\varphi}) := \inf_{\boldsymbol{u} \in E} \mathbb{P}\{ |\langle \boldsymbol{\varphi}, \boldsymbol{u} \rangle| \geq \xi \} \quad \text{where } \xi \geq 0.$$

Let $\varepsilon_1, \dots, \varepsilon_m$ be independent Rademacher random variables,[2] independent of everything else, and define the mean empirical width of the set:

$$W_m(E; \boldsymbol{\varphi}) := \mathbb{E} \sup_{\boldsymbol{u} \in E} \langle \boldsymbol{h}, \boldsymbol{u} \rangle \quad \text{where} \quad \boldsymbol{h} := \frac{1}{\sqrt{m}} \sum_{i=1}^m \varepsilon_i \boldsymbol{\varphi}_i. \tag{2.24}$$

Then, for any $\xi > 0$ and $t > 0$,

$$\inf_{\boldsymbol{u} \in E} \left(\sum_{i=1}^m |\langle \boldsymbol{\varphi}_i, \boldsymbol{u} \rangle|^2 \right)^{1/2} \geq \xi \sqrt{m}\, Q_{2\xi}(E; \boldsymbol{\varphi}) - 2W_m(E; \boldsymbol{\varphi}) - \xi t$$

with probability at least $1 - \mathrm{e}^{-t^2/2}$.

The proof appears below in Section 2.5.5. In the sequel, we usually lighten our notation for Q_ξ and W_m by suppressing the dependence on $\boldsymbol{\varphi}$.

Before we continue, it may be helpful to remark on this result. The marginal tail function $Q_\xi(E)$ reflects the probability that the random variable $|\langle \boldsymbol{\varphi}, \boldsymbol{u} \rangle|$ is close to zero for any fixed vector $\boldsymbol{u} \in E$. When $Q_\xi(E)$ is bounded away from zero for some ξ, the nonnegative empirical process is likely to be large. Koltchinskii & Mendelson [14] point out that the marginal tail function reflects the absolute continuity of the distribution of $\boldsymbol{\varphi}$, so Q_ξ may be quite small when $\boldsymbol{\varphi}$ is "spiky."

[2] A *Rademacher* random variable takes the two values ± 1 with equal probability.

The mean empirical width $W_m(E)$ is a distribution-dependent measure of the size of the set E. When $\boldsymbol{\varphi}$ follows a standard Gaussian distribution, $W_m(E)$ reduces to the usual Gaussian width $W(E) := \mathbb{E} \sup_{\boldsymbol{u} \in E} \langle \boldsymbol{g}, \boldsymbol{u} \rangle$. As the number m tends to infinity, the distribution of the random vector $\boldsymbol{h}$ converges in distribution to a centered Gaussian variable with covariance $\mathbb{E}[\boldsymbol{\varphi}\boldsymbol{\varphi}^*]$. Therefore, $W_m(E) \to W(E)$ when $\boldsymbol{\varphi}$ is centered and isotropic.

2.5.3 Mendelson's Small Ball Method

Proposition 5.1 shows that we can obtain a lower bound for (2.23) by performing two simpler estimates. To achieve this goal, Mendelson has developed a general strategy, which consists of three steps:

MENDELSON'S SMALL BALL METHOD

(1) Apply Proposition 5.1 to bound the minimum conic singular value $\lambda_{\min}(\boldsymbol{\Phi};\ K)$ below in terms of the marginal tail function $Q_{2\xi}(E; \boldsymbol{\varphi})$ and the mean empirical width $W_m(E; \boldsymbol{\varphi})$. The index set $E := K \cap \mathsf{S}^{d-1}$.
(2) Bound the marginal tail function $Q_{2\xi}(E; \boldsymbol{\varphi})$ below using a Paley–Zygmund inequality.
(3) Bound the mean empirical width $W_m(E; \boldsymbol{\varphi})$ above by imitating techniques for controlling the Gaussian width of E.

This presentation is distilled from the corpus [14, 16, 19–21]. A more sophisticated variant of this method appears in [20, Thm. 5.3]. Later in this chapter, we will encounter several concrete applications of this strategy.

2.5.4 Expected Scope

Mendelson's Small Ball Method provides lower bounds for (2.23) in many situations, but it does not offer a universal prescription. Let us try to delineate the circumstances where this approach is likely to be useful for signal recovery problems.

- Mendelson's Small Ball Method assumes that the sampling matrix $\boldsymbol{\Phi}$ has independent, identically distributed rows. Although this model describes many of the sampling strategies in the literature, there are some examples, such as random filtering [34], that do not conform to this assumption.

- A major advantage of Mendelson's Small Ball Method is that it applies to sampling distributions with heavy tails. On the other hand, the random vector $\boldsymbol{\varphi}$ cannot be too "spiky," or else it may not be possible to produce a good lower bound for the marginal tail function $Q_{2\xi}(E)$. This requirement indicates that the approach may require significant improvements before it applies to problems like matrix completion.

There are a number of possible extensions of Mendelson's Small Ball Method that could expand its bailiwick. For example, it is easy to extend Proposition 5.1 to address the case where the random vector $\boldsymbol{\varphi}$ is complex valued. A more difficult, but very useful, modification would allow us to block the measurements into groups. This revision could reduce the difficulties associated with spiky distributions, but it seems to demand some additional ideas.

2.5.5 *Proof of Proposition 5.1*

Let us establish the Mendelson bound for a nonnegative empirical process. First, we introduce a directional version of the marginal tail function:

$$Q_\xi(\boldsymbol{u}) := \mathbb{P}\big\{ |\langle \boldsymbol{\varphi}, \boldsymbol{u} \rangle| \geq \xi \big\} \quad \text{for } \boldsymbol{u} \in E \text{ and } \xi > 0.$$

Lyapunov's inequality and Markov's inequality give the numerical bounds

$$\left(\frac{1}{m} \sum_{i=1}^m |\langle \boldsymbol{\varphi}_i, \boldsymbol{u} \rangle|^2 \right)^{1/2} \geq \frac{1}{m} \sum_{i=1}^m |\langle \boldsymbol{\varphi}_i, \boldsymbol{u} \rangle| \geq \frac{\xi}{m} \sum_{i=1}^m \mathbb{1}\big\{ |\langle \boldsymbol{\varphi}_i, \boldsymbol{u} \rangle| \geq \xi \big\}.$$

We write $\mathbb{1}A$ for the 0–1 random variable that takes the value one when the event A takes place. Add and subtract $Q_{2\xi}(\boldsymbol{u})$ inside the sum and then take the infimum over $\boldsymbol{u} \in E$ to reach the inequality

$$\begin{aligned} \inf_{\boldsymbol{u} \in E} \left(\frac{1}{m} \sum_{i=1}^m |\langle \boldsymbol{\varphi}_i, \boldsymbol{u} \rangle|^2 \right)^{1/2} &\geq \xi \inf_{\boldsymbol{u} \in E} Q_{2\xi}(\boldsymbol{u}) \\ &\quad - \frac{\xi}{m} \sup_{\boldsymbol{u} \in E} \sum_{i=1}^m \big[Q_{2\xi}(\boldsymbol{u}) - \mathbb{1}\big\{ |\langle \boldsymbol{\varphi}_i, \boldsymbol{u} \rangle| \geq \xi \big\} \big]. \end{aligned} \tag{2.25}$$

To control the supremum in probability, we can invoke the bounded difference inequality [3, Sec. 6.1]. Observe that each summand is independent and bounded in magnitude by one. Therefore,

$$\sup_{\boldsymbol{u}\in E} \sum_{i=1}^{m} \left[Q_{2\xi}(\boldsymbol{u}) - \mathbb{1}\{ |\langle \boldsymbol{\varphi}_i, \boldsymbol{u}\rangle| \geq \xi \} \right]$$

$$\leq \mathbb{E} \sup_{\boldsymbol{u}\in E} \sum_{i=1}^{m} \left[Q_{2\xi}(\boldsymbol{u}) - \mathbb{1}\{ |\langle \boldsymbol{\varphi}_i, \boldsymbol{x}\rangle| \geq \xi \} \right] + t\sqrt{m} \tag{2.26}$$

with probability at least $1 - \mathrm{e}^{-t^2/2}$.

Next, we simplify the expected supremum. Introduce a soft indicator function:

$$\psi_\xi : \mathbb{R} \to [0,1] \quad \text{where} \quad \psi_\xi(s) := \begin{cases} 0, & |s| \leq \xi \\ (|s| - \xi)/\xi, & \xi < |s| \leq 2\xi \\ 1, & 2\xi < |s| \,. \end{cases}$$

We need two properties of the soft indicator. First, the soft indicator is bracketed by two hard indicators: $\mathbb{1}\{|s| \geq 2\xi\} \leq \psi_\xi(s) \leq \mathbb{1}\{|s| \geq \xi\}$ for all $s \in \mathbb{R}$. Second, $\xi\psi_\xi$ is a *contraction*, i.e., a 1-Lipschitz function on $\mathbb{R}$ that fixes the origin. Therefore, we can make the following calculation:

$$\begin{aligned} &\mathbb{E} \sup_{\boldsymbol{u}\in E} \sum_{i=1}^{m} \left[Q_{2\xi}(\boldsymbol{u}) - \mathbb{1}\{ |\langle \boldsymbol{\varphi}_i, \boldsymbol{u}\rangle| \geq \xi \} \right] \\ &= \mathbb{E} \sup_{\boldsymbol{u}\in E} \sum_{i=1}^{m} \left[\mathbb{E}\, \mathbb{1}\{ |\langle \boldsymbol{\varphi}, \boldsymbol{u}\rangle| \geq 2\xi \} - \mathbb{1}\{ |\langle \boldsymbol{\varphi}_i, \boldsymbol{u}\rangle| \geq \xi \} \right] \\ &\leq \mathbb{E} \sup_{\boldsymbol{u}\in E} \sum_{i=1}^{m} \left[\mathbb{E}\, \psi_\xi(\langle \boldsymbol{\varphi}, \boldsymbol{u}\rangle) - \psi_\xi(\langle \boldsymbol{\varphi}_i, \boldsymbol{u}\rangle) \right] \\ &\leq 2\, \mathbb{E} \sup_{\boldsymbol{u}\in E} \sum_{i=1}^{m} \varepsilon_i \psi_\xi(\langle \boldsymbol{\varphi}_i, \boldsymbol{u}\rangle) \\ &\leq \frac{2}{\xi}\, \mathbb{E} \sup_{\boldsymbol{u}\in E} \sum_{i=1}^{m} \varepsilon_i \langle \boldsymbol{\varphi}_i, \boldsymbol{u}\rangle \,. \end{aligned} \tag{2.27}$$

In the first line, we write the marginal tail function as an expectation and then we bound the two indicators using the soft indicator function. The next inequality is the Giné–Zinn symmetrization [35, Lem. 2.3.1]. The last line follows from the Rademacher comparison principle [17, Eqn. (4.20)] because $\xi\psi_\xi$ is a contraction.

Combine the inequalities (2.25), (2.26), and (2.27) to reach

$$\inf_{\boldsymbol{u}\in E} \left(\frac{1}{m} \sum_{i=1}^{m} |\langle \boldsymbol{\varphi}_i, \boldsymbol{u}\rangle|^2 \right)^{1/2} \geq \xi \inf_{\boldsymbol{u}\in E} Q_{2\xi}(\boldsymbol{u}) - \frac{\xi}{m} \left[\frac{2}{\xi}\, \mathbb{E} \sup_{\boldsymbol{u}\in E} \sum_{i=1}^{m} \varepsilon_i \langle \boldsymbol{\varphi}_i, \boldsymbol{u}\rangle + t\sqrt{m} \right].$$

Define $\boldsymbol{h} := m^{-1/2} \sum_{i=1}^{m} \varepsilon_i \boldsymbol{\varphi}_i$ and clear the factor $\sqrt{m}$ to conclude that

$$\inf_{\boldsymbol{u} \in E} \left(\sum_{i=1}^{m} |\langle \boldsymbol{\varphi}_i, \boldsymbol{u} \rangle|^2 \right)^{1/2} \geq \xi \sqrt{m} \inf_{\boldsymbol{u} \in E} Q_{2\xi}(\boldsymbol{u}) - 2\,\mathbb{E} \sup_{\boldsymbol{u} \in E} \langle \boldsymbol{h}, \boldsymbol{u} \rangle - \xi t.$$

with probability at least $1 - \mathrm{e}^{-t^2/2}$. Identify the marginal tail function $Q_{2\xi}(E)$ and the empirical width $W_m(E)$ to establish Proposition 5.1.

2.6 A universal error bound for sub-Gaussian measurements

In this section, we invoke Mendelson's Small Ball Method to study convex signal recovery from independent sub-Gaussian measurements. This class of examples provides a wide generalization of standard Gaussian measurements. We will establish a variant of the Gaussian recovery result, Corollary 3.5, in this setting.

2.6.1 Sub-Gaussian measurements

Let us set out the conditions we require for the sampling matrix. Suppose that $\boldsymbol{\varphi}$ is a random vector in $\mathbb{R}^d$ that has the following properties:

- **[Centering]** The vector has zero mean: $\mathbb{E}\,\boldsymbol{\varphi} = \mathbf{0}$.
- **[Nondegeneracy]** There is a positive constant α for which

$$\alpha \leq \mathbb{E}\,|\langle \boldsymbol{\varphi}, \boldsymbol{u} \rangle| \quad \text{for each } \boldsymbol{u} \in \mathsf{S}^{d-1}.$$

- **[Sub-Gaussian marginals]** There is a positive constant σ for which

$$\mathbb{P}\big\{ |\langle \boldsymbol{\varphi}, \boldsymbol{u} \rangle| \geq t \big\} \leq 2\mathrm{e}^{-t^2/(2\sigma^2)} \quad \text{for each } \boldsymbol{u} \in \mathsf{S}^{d-1}.$$

- **[Low eccentricity]** The eccentricity $\rho := \sigma/\alpha$ of the distribution should be small.

Finally, we construct a random $m \times d$ sampling matrix $\boldsymbol{\Phi}$ whose rows are independent copies of $\boldsymbol{\varphi}^{\mathsf{t}}$, as in expression (2.22).

A few examples of sub-Gaussian distributions may be helpful.

Example 6.1 (Nonstandard Gaussian matrices). Suppose that $\boldsymbol{\varphi} \in \mathbb{R}^d$ follows the $\textsc{normal}(\mathbf{0}, \boldsymbol{\Sigma})$ distribution where the covariance $\boldsymbol{\Sigma}$ satisfies $\frac{\pi}{2}\alpha^2 \leq \boldsymbol{u}^{\mathsf{t}} \boldsymbol{\Sigma} \boldsymbol{u} \leq \sigma^2$ for each vector $\boldsymbol{u} \in \mathsf{S}^{d-1}$. Then the required conditions follow from basic facts about a normal distribution.

Example 6.2 (Independent bounded entries). Let X be a symmetric random variable whose magnitude is bounded by σ. Suppose that each entry of $\boldsymbol{\varphi}$ is an independent copy of X.

The vector $\boldsymbol{\varphi}$ inherits centering from X. Next, $\boldsymbol{\varphi}$ is nondegenerate with $\alpha \geq 2^{-1/2}\,\mathbb{E}\,|X|$ because of the Khintchine inequality [15] and a convexity argument. Finally, $\boldsymbol{\varphi}$ has sub-Gaussian marginals with the parameter σ because of Hoeffding's inequality [3, Sec. 2.6].

2.6.2 The minimum conic singular value of a sub-Gaussian matrix

The main result of this section gives a lower bound for the minimum conic singular value of a matrix $\boldsymbol{\Phi}$ that satisfies the conditions in Section 2.6.1.

Theorem 6.3 (Minimum conic singular value of a sub-Gaussian matrix). *Suppose $\boldsymbol{\Phi}$ is an $m \times d$ random matrix that satisfies the conditions in Section 2.6.1. Let $K \subset \mathbb{R}^d$ be a cone, not necessarily convex. Then*

$$\lambda_{\min}(\boldsymbol{\Phi}; K) \geq c\alpha\rho^{-2} \cdot \sqrt{m} - C\sigma \cdot w(K) - \alpha t$$

with probability at least $1 - \mathrm{e}^{-ct^2}$. The quantities c and C are positive absolute constants.

Observe that, when the eccentricity ρ has constant order, the bound in Theorem 6.3 matches the result for Gaussian matrices in Proposition 3.3. A similar result appears in the paper [22], so we do not claim any novelty. We establish Theorem 6.3 below in Section 2.6.4.

2.6.3 An error bound for sub-Gaussian measurements

Combining Proposition 2.6 and Theorem 6.3, we reach an immediate consequence for signal recovery from sub-Gaussian measurements.

Corollary 6.4 (Signal recovery from sub-Gaussian measurements). *Let $\boldsymbol{x}^\natural$ be a signal in $\mathbb{R}^d$. Let $\boldsymbol{\Phi}$ be an $m \times d$ random matrix that satisfies the conditions in Section 2.6.1, and let $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^\natural + \boldsymbol{e}$ be a vector of measurements in $\mathbb{R}^m$. With probability at least $1 - \mathrm{e}^{-ct^2}$, the following statement holds. Assume that $\|\boldsymbol{e}\| \leq \eta$, and let $\hat{\boldsymbol{x}}_\eta$ be any solution to the optimization problem (2.2). Then*

$$\left\|\hat{\boldsymbol{x}}_\eta - \boldsymbol{x}^\natural\right\| \leq \frac{2\eta}{\left[c\alpha\rho^{-2} \cdot \sqrt{m} - C\sigma \cdot w\big(\mathscr{D}(f, \boldsymbol{x}^\natural)\big) - \alpha t\right]_+}.$$

The quantities c and C are positive absolute constants. The operation $[a]_+ := \max\{a, 0\}$ returns the positive part of a number.

Corollary 6.4 provides for stable recovery of $\boldsymbol{x}^\natural$ as soon as the number m of sub-Gaussian measurements satisfies

$$m \geq C' \rho^6 \cdot w^2\big(\mathscr{D}(f, \boldsymbol{x}^\natural)\big).$$

How accurate is this result? Note that standard Gaussian measurements satisfy the assumptions of the corollary with ρ constant, and we need at least $w^2\big(\mathscr{D}(f, \boldsymbol{x}^\natural)\big)$ standard normal measurements to recover the structured signal $\boldsymbol{x}^\natural$ with the complexity measure f. Therefore, the bound is correct up to the constant factor C' and the precise dependence on the eccentricity ρ.

2.6.4 Proof of Theorem 6.3: Setup and Step 1

To establish Theorem 6.3, we rely on Mendelson's Small Ball Method. The argument also depends on some deep ideas from the theory of generic chaining [32], but we only use these results in a naïve way.

Fix a cone K in $\mathbb{R}^d$ and define the set $E := K \cap \mathsf{S}^{d-1}$. Suppose that $\boldsymbol{\varphi}$ is a random vector in $\mathbb{R}^d$ that satisfies the conditions set out in Section 2.6.1 and construct an $m \times d$ random matrix $\boldsymbol{\Phi}$ whose rows are independent copies of $\boldsymbol{\varphi}$. Proposition 5.1 implies that

$$\lambda_{\min}(\boldsymbol{\Phi}; K) \geq \xi \sqrt{m}\, Q_{2\xi}(E) - 2W_m(E) - \xi t \quad \text{with probability} \geq 1 - \mathrm{e}^{-t^2/2}. \tag{2.28}$$

This result holds for all $\xi > 0$ and $t > 0$. To establish Theorem 6.3, we must develop a constant lower bound for the marginal tail function $Q_{2\xi}(E)$, and we also need to compare the mean empirical width $W_m(E)$ with the conic Gaussian width $w(K)$.

2.6.5 Step 2: The marginal tail function

We begin with the lower bound for the marginal tail function $Q_{2\xi}$. This result is an easy consequence of the second moment method, also known as the Paley–Zygmund inequality. Let $\boldsymbol{u}$ be any vector in E. One version of the second moment method states that

$$\mathbb{P}\big\{ |\langle \boldsymbol{\varphi}, \boldsymbol{u} \rangle| \geq 2\xi \big\} \geq \frac{\big[\, \mathbb{E}\, |\langle \boldsymbol{\varphi}, \boldsymbol{u} \rangle| - 2\xi \big]_+^2}{\mathbb{E}\, |\langle \boldsymbol{\varphi}, \boldsymbol{u} \rangle|^2}. \tag{2.29}$$

To control the denominator on the right-hand side of (2.29), we use the sub-Gaussian marginal condition to estimate that

$$\mathbb{E}\,|\langle \boldsymbol{\varphi},\, \boldsymbol{u}\rangle|^2 = \int_0^\infty 2s \cdot \mathbb{P}\big\{\,|\langle \boldsymbol{\varphi},\, \boldsymbol{u}\rangle| \geq s\big\}\,\mathrm{d}s \leq 4\sigma^2.$$

To bound the numerator on the right-hand side of (2.29), we use the nondegeneracy assumption: $\mathbb{E}\,|\langle \boldsymbol{\varphi},\, \boldsymbol{u}\rangle| \geq \alpha$. Combining these results and taking the infimum over $\boldsymbol{u} \in E$, we reach

$$Q_{2\xi}(E) = \inf_{\boldsymbol{u}\in E} \mathbb{P}\big\{\,|\langle \boldsymbol{\varphi},\, \boldsymbol{u}\rangle| \geq 2\xi\big\} \geq \frac{(\alpha - 2\xi)^2}{4\sigma^2} \tag{2.30}$$

for any ξ that satisfies $2\xi < \alpha$.

2.6.6 Step 3: The mean empirical width

Next, we demonstrate that the empirical width $W_m(E)$ is controlled by the conic Gaussian width $w(K)$. This argument requires sophisticated results from the theory of generic chaining [32]. First, observe that the vector $\boldsymbol{h} = m^{-1/2} \sum_{i=1}^m \varepsilon_i \boldsymbol{\varphi}_i$ inherits sub-Gaussian marginals from the centered sub-Gaussian distribution $\boldsymbol{\varphi}$. Indeed,

$$\mathbb{P}\big\{\,|\langle \boldsymbol{h},\, \boldsymbol{u}\rangle| \geq t\big\} \leq C_1 \mathrm{e}^{-c_1 t^2/\sigma^2} \quad \text{for each } \boldsymbol{u} \in \mathsf{S}^{d-1}.$$

See [36, Sec. 5.2.3] for an introduction to sub-Gaussian random variables. In particular, we have the bound

$$\mathbb{P}\big\{\,|\langle \boldsymbol{h},\, \boldsymbol{u} - \boldsymbol{v}\rangle| \geq t\big\} \leq C_1 \mathrm{e}^{-c_1 t^2/(\sigma^2 \|\boldsymbol{u}-\boldsymbol{v}\|^2)} \quad \text{for all } \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^d.$$

Under the latter condition, the generic chaining theorem [32, Thm. 1.2.6] asserts that

$$W_m(E) = \mathbb{E} \sup_{\boldsymbol{u}\in E} \langle \boldsymbol{h},\, \boldsymbol{u}\rangle \leq C_2\sigma \cdot \gamma_2(E, \ell_2)$$

where γ_2 is a geometric functional. The precise definition of γ_2 is not important for our purposes because the majorizing measure theorem [32, Thm. 2.1.1] states that

$$\gamma_2(E, \ell_2) \leq C_3 \cdot \mathbb{E} \sup_{\boldsymbol{u}\in E} \langle \boldsymbol{g},\, \boldsymbol{u}\rangle$$

where $\boldsymbol{g} \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_d)$. It follows that

$$W_m(E) \leq C_4\sigma \cdot \mathbb{E} \sup_{\boldsymbol{u}\in E} \langle \boldsymbol{g},\, \boldsymbol{u}\rangle = C_4\sigma \cdot w(K). \tag{2.31}$$

We have recalled that $E = K \cap \mathsf{S}^{d-1}$ to identify the conic Gaussian width $w(K)$.

2.6.7 Combining the bounds

Combine the bounds (2.28), (2.30), and (2.31) to discover that

$$\lambda_{\min}(\boldsymbol{\Phi}; K) \geq \xi\sqrt{m} \cdot \frac{(\alpha - 2\xi)^2}{4\sigma^2} - 2C_4\sigma\, w(K) - \xi t \quad \text{with probability} \geq 1 - \mathrm{e}^{-t^2/2},$$

provided that $2\xi < \alpha$. Select $\xi = \alpha/6$ to see that

$$\lambda_{\min}(\boldsymbol{\Phi}; K) \geq \frac{1}{54} \cdot \frac{\alpha^3}{\sigma^2}\sqrt{m} - C_5\sigma\, w(K) - \frac{\alpha}{6}t \quad \text{with probability} \geq 1 - \mathrm{e}^{-t^2/2}. \tag{2.32}$$

Using the eccentricity $\rho = \sigma/\alpha$, we simplify expression (2.32) to reach a bound for the minimum conic singular value of a sub-Gaussian random matrix $\boldsymbol{\Phi}$ that satisfies the conditions set out in Section 2.6.1. This completes the proof of Theorem 6.3.

2.7 The bowling scheme

As we have seen in Theorem 6.3, sub-Gaussian sampling models exhibit behavior similar to the standard Gaussian measurement model. Yet there are many interesting problems where the random sampling matrix does not conform to the sub-Gaussian assumption. In this section, we explain how to adapt Mendelson's Small Ball Method to a range of other sampling ensembles. The key idea is to use the conic duality arguments from Section 2.4 to complete the estimate for the mean empirical width.

2.7.1 The mean empirical width of a descent cone

Let us state a simple duality result for the mean empirical width of a descent cone. This bound is based on the same principles as Proposition 4.5.

Proposition 7.1 (The mean empirical width of a descent cone). *Let $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ be a proper convex function, and fix a point $\boldsymbol{x} \in \mathbb{R}^d$. Assume that the subdifferential $\partial f(\boldsymbol{x})$ is nonempty and does not contain the origin. For any random vector $\boldsymbol{\varphi} \in \mathbb{R}^d$,*

$$W_m\big(\mathscr{D}(f, \boldsymbol{x}) \cap \mathsf{S}^{d-1}; \boldsymbol{\varphi}\big) \leq \mathbb{E} \inf_{\tau \geq 0} \operatorname{dist}^2\big(\boldsymbol{h},\ \tau \cdot \partial f(\boldsymbol{x})\big) \quad \text{where} \quad \boldsymbol{h} := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \varepsilon_i \boldsymbol{\varphi}_i.$$

The mean empirical width W_m is defined in (2.24). The random vectors $\boldsymbol{\varphi}_1, \dots, \boldsymbol{\varphi}_m$ are independent copies of $\boldsymbol{\varphi}$, and $\varepsilon_1, \dots, \varepsilon_m$ are independent Rademacher random variables.

Proof. The argument is identical with the proof of Proposition 4.5 once we replace the Gaussian vector $\boldsymbol{g}$ with the random vector $\boldsymbol{h}$. □

2.7.2 The bowling scheme

We are now prepared to describe a general approach for convex signal recovery from independent random measurements.

The setup is similar to previous sections. Consider an unknown structured signal $\boldsymbol{x}^\natural \in \mathbb{R}^d$ and a complexity measure $f : \mathbb{R}^d \to \overline{\mathbb{R}}$ that is proper and convex. Let $\boldsymbol{\Phi}$ be a known $m \times d$ sampling matrix, and suppose that we acquire m noisy linear measurements of the form $\boldsymbol{y} = \boldsymbol{\Phi}\boldsymbol{x}^\natural + \boldsymbol{e}$. We wish to analyze the performance of the convex recovery method (2.2). Proposition 2.6 shows that we can accomplish this goal by finding a lower bound for the minimum conic singular value of the descent cone:

$$\lambda_{\min}\big(\boldsymbol{\Phi};\ \mathscr{D}(f, \boldsymbol{x}^\natural)\big) \geq \underline{\qquad ??? \qquad}. \tag{2.33}$$

We want to produce a bound of the form (2.33) when the rows of the measurement matrix $\boldsymbol{\Phi}$ are independent copies of a random vector $\boldsymbol{\varphi}$. This problem falls within the scope of Mendelson's Small Ball Method. Introduce the index set $E := \mathscr{D}(f, \boldsymbol{x}^\natural) \cap \mathsf{S}^{d-1}$. In light of (2.23),

$$\lambda_{\min}\big(\boldsymbol{\Phi};\ \mathscr{D}(f, \boldsymbol{x}^\natural)\big) = \inf_{\boldsymbol{u} \in E} \left(\sum_{i=1}^{m} |\langle \boldsymbol{\varphi}_i,\ \boldsymbol{u} \rangle|^2 \right)^{1/2}.$$

We follow Mendelson's general strategy to control the minimum conic singular value, but we propose a specific technique for bounding the mean empirical width that exploits the structure of the index set E.

THE BOWLING SCHEME

(1) Apply Proposition 5.1 to bound the minimum conic singular value $\lambda_{\min}\big(\boldsymbol{\Phi};\ \mathscr{D}(f, \boldsymbol{x}^\natural)\big)$ below in terms of the marginal tail function $Q_{2\xi}(E; \boldsymbol{\varphi})$ and the mean empirical width $W_m(E; \boldsymbol{\varphi})$. The index set $E := \mathscr{D}(f; \boldsymbol{x}^\natural) \cap \mathsf{S}^{d-1}$.

(2) Bound the marginal tail function $Q_{2\xi}(E; \boldsymbol{\varphi})$ below using a Paley–Zygmund inequality.

(3′) Apply Proposition 7.1 to control the mean empirical width $W_m(E; \boldsymbol{\varphi})$.

In other words, Step (3) of Mendelson's framework has been specialized to Step (3′).

We refer to this instance of Mendelson's Small Ball Method as *the bowling scheme*. The name is chosen as a salute to David Gross's *golfing scheme*. Whereas the golfing scheme is based on dual optimality conditions for the signal recovery problem (2.2), the bowling scheme is based on the primal optimality condition through Proposition 2.6. In the bowling scheme, duality enters only when we are ready to estimate the mean empirical width.

In our experience, this idea has been successful whenever we understand how to bound the conic Gaussian width of the descent cone. The main distinction is that the random vector $\boldsymbol{\varphi}$ may not share the rotational invariance of the standard Gaussian distribution.

2.8 Example: Phase retrieval

To demonstrate how the bowling scheme works, we consider the question of phase retrieval. In this problem, we collect linear samples of an unknown signal, but we are only able to observe their magnitudes. To reconstruct the original signal, we must resolve the uncertainty about the phases (or signs) of the measurements. There is a natural convex program that can achieve this goal, and the bowling scheme offers an easy way to analyze the number of measurements that are required.

2.8.1 Phase retrieval by convex optimization

In the phase retrieval problem, we wish to recover a signal $\boldsymbol{x}^\natural \in \mathbb{R}^d$ from a family of measurements of the form

$$y_i = |\langle \boldsymbol{\psi}_i, \boldsymbol{x}^\natural \rangle|^2 \quad \text{for } i = 1, 2, 3, \dots, m. \tag{2.34}$$

The sampling ensemble $\boldsymbol{\psi}_1, \dots, \boldsymbol{\psi}_m$ consists of known vectors in $\mathbb{R}^d$. For clarity of presentation, we do not consider the case where the samples are noisy or complex-valued.

Although the samples do not initially appear linear, we can apply a lifting method proposed by Balan et al. [2]. Observe that

$$|\langle \boldsymbol{\psi}, \boldsymbol{x} \rangle|^2 = \boldsymbol{\psi}^\mathsf{t} \boldsymbol{x} \cdot \boldsymbol{x}^\mathsf{t} \boldsymbol{\psi} = \operatorname{trace}\left(\boldsymbol{x}\boldsymbol{x}^\mathsf{t} \cdot \boldsymbol{\psi}\boldsymbol{\psi}^\mathsf{t}\right).$$

In view of this expression, it is appropriate to introduce the rank-one positive-semidefinite matrices

$$\boldsymbol{X}^\natural = (\boldsymbol{x}^\natural)(\boldsymbol{x}^\natural)^\mathsf{t} \in \mathbb{R}^{d \times d} \quad \text{and} \quad \boldsymbol{\Psi}_i = \boldsymbol{\psi}_i \boldsymbol{\psi}_i^\mathsf{t} \in \mathbb{R}^{d \times d} \quad \text{for } i = 1, 2, 3, \dots, m. \tag{2.35}$$

Then we can express the samples y_i as *linear* functions of the matrix $\boldsymbol{X}^\natural$:

$$y_i = \operatorname{trace}\left(\boldsymbol{X}^\natural \cdot \boldsymbol{\Psi}_i\right) \quad \text{for } i = 1, 2, 3, \dots, m. \tag{2.36}$$

Expression (2.36) coincides with the measurement model (2.1) we have been considering.

We can use convex optimization to reconstruct the unknown matrix $\boldsymbol{X}^\natural$. It is natural to minimize the Schatten 1-norm to promote low rank, but we also want to enforce the fact that $\boldsymbol{X}^\natural$ is positive semidefinite [9]. To that end, we consider the convex program

$$\begin{aligned} &\underset{\boldsymbol{X} \in \mathbb{R}^{d \times d}}{\text{minimize}} \quad \operatorname{trace}(\boldsymbol{X}) \quad \text{subject to} \quad \boldsymbol{X} \succcurlyeq \mathbf{0} \quad \text{and} \quad y_i = \operatorname{trace}\left(\boldsymbol{X}\boldsymbol{\Psi}_i\right) \\ &\quad \text{for each } i = 1, 2, 3, \dots, m. \end{aligned} \tag{2.37}$$

This formulation involves the lifted variables (2.35). We say that the optimization problem (2.37) *recovers* $\boldsymbol{x}^\natural$ if the matrix $\boldsymbol{X}^\natural$ is the unique minimizer. Indeed, in this case, we can reconstruct the original signal by factorizing the solution to the optimization problem.

Remark 8.1 (Citation for convex phase retrieval). Formulation (2.37) was developed by a working group at the meeting "Frames for the finite world: Sampling, coding and quantization," which took place at the American Institute of Mathematics in Palo Alto in August 2008. Most of the recent literature attributes this idea incorrectly.

2.8.2 Phase retrieval from Gaussian measurements

Recently, researchers have started to consider phase retrieval problems with random data; see [5] for example. In the simplest instance, we choose each sampling vector $\boldsymbol{\psi}_i$ independently from the standard normal distribution on $\mathbb{R}^d$:

$$\boldsymbol{\psi}_i \sim \text{NORMAL}(\mathbf{0}, \mathbf{I}_d).$$

Then each sampling matrix $\boldsymbol{\Psi}_i = \boldsymbol{\psi}_i \boldsymbol{\psi}_i^*$ follows a Wishart distribution. These random matrices do not have sub-Gaussian marginals, so we cannot apply Corollary 6.4 to study the performance of the optimization problem (2.37). Nevertheless, we can make short work of the analysis by using the bowling scheme.

Theorem 8.2 (Phase retrieval from Gaussian measurements). *Let $\boldsymbol{x}^\natural$ be a signal in $\mathbb{R}^d$. Let $\boldsymbol{\psi}_i \sim \text{NORMAL}(\mathbf{0}, \mathbf{I}_d)$ be independent standard Gaussian vectors, and consider random measurements $y_i = |\langle \boldsymbol{\psi}_i, \boldsymbol{x}^\natural \rangle|^2$ for $i = 1, 2, 3, \dots, m$. Assuming that $m \geq Cd$, the convex phase retrieval problem (2.37) recovers $\boldsymbol{x}^\natural$ with probability at least $1 - \mathrm{e}^{-cm}$. The numbers c and C are positive absolute constants.*

The sampling complexity $m \geq Cd$ established in Theorem 8.2 is qualitatively optimal. Indeed, a dimension-counting argument shows that we need at least $m \geq d$ nonadaptive linear measurements to reconstruct a general vector in $\boldsymbol{x}^\natural \in \mathbb{R}^d$.

Remark 8.3 (Extensions). There are a number of obvious improvements to Theorem 8.2 that follow with a little more effort. For example, it is clear that the convex phase retrieval method is stable. The exceedingly high success probability also allows us to establish uniform results for all d-dimensional vectors by means of net arguments and union bounds. Furthermore, the Gaussian assumption is inessential; it is possible to establish similar theorems for other sampling distributions. We leave these refinements for the avid reader.

2.8.3 *Proof of Theorem 8.2: Setup*

Let us rewrite the optimization problem (2.37) in a form that is more conducive to our methods of analysis. First, introduce the inner product space $\mathbb{R}^{d\times d}_{\text{sym}}$ of $d \times d$ symmetric matrices, equipped with the trace inner product $\langle \boldsymbol{A}, \boldsymbol{B} \rangle := \operatorname{trace}(\boldsymbol{AB})$ and the Frobenius norm $\|\cdot\|_{\text{F}}$. Define the linear operator

$$\boldsymbol{\Phi} : \mathbb{R}^{d\times d}_{\text{sym}} \to \mathbb{R}^m \quad \text{where} \quad [\boldsymbol{\Phi}(\boldsymbol{X})]_i = \langle \boldsymbol{\Psi}_i, \boldsymbol{X} \rangle \quad \text{for } i = 1, 2, 3, \dots, m.$$

Collect the measurements into a vector $\boldsymbol{y} = (y_1, \dots, y_m)^{\mathsf{t}} \in \mathbb{R}^m$ and observe that $\boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{X}^\natural)$ because of expression (2.36). Next, define the convex indicator function of the positive-semidefinite cone:

$$\iota : \mathbb{R}^{d\times d}_{\text{sym}} \to \overline{\mathbb{R}} \quad \text{where} \quad \iota(\boldsymbol{X}) = \begin{cases} 0, & \boldsymbol{X} \text{ is positive semidefinite} \\ +\infty, & \text{otherwise.} \end{cases}$$

Introduce the convex regularizer

$$f : \mathbb{R}^{d\times d}_{\text{sym}} \to \overline{\mathbb{R}} \quad \text{where} \quad f(\boldsymbol{X}) = \operatorname{trace}(\boldsymbol{X}) + \iota(\boldsymbol{X}).$$

With this notation, we can write (2.37) in the form

$$\underset{\boldsymbol{X}\in\mathbb{R}^{d\times d}_{\text{sym}}}{\text{minimize}} \quad f(\boldsymbol{X}) \quad \text{subject to} \quad \boldsymbol{y} = \boldsymbol{\Phi}(\boldsymbol{X}). \tag{2.38}$$

Formulation (2.38) matches our core problem (2.2) with the error vector $\boldsymbol{e} = \boldsymbol{0}$ and error tolerance $\eta = 0$.

Proposition 2.6 demonstrates that $\boldsymbol{X}^\natural$ is the unique solution of (2.38) whenever

$$\lambda_{\min}\big(\boldsymbol{\Phi};\ \mathscr{D}(f, \boldsymbol{X}^\natural)\big) > 0.$$

We must determine how many measurements m suffice for this event to hold with high probability.

2.8.4 Step 1: The nonnegative empirical process bound

Define the set

$$E := \{\boldsymbol{U} \in \mathscr{D}(f, \boldsymbol{X}^\natural) : \|\boldsymbol{U}\|_{\mathrm{F}} = 1\} \subset \mathbb{R}^{d\times d}_{\mathrm{sym}}.$$

Proposition 5.1 demonstrates that

$$\lambda_{\min}\big(\boldsymbol{\Phi};\ \mathscr{D}(f, \boldsymbol{X}^\natural)\big) = \inf_{\boldsymbol{U}\in E} \left(\sum_{i=1}^{m} |\langle \boldsymbol{\Psi}_i,\ \boldsymbol{U}\rangle|^2\right)^{1/2} \geq \xi\sqrt{m}\, Q_{2\xi}(E) - 2\, W_m(E) - \xi t \tag{2.39}$$

with probability at least $1 - \mathrm{e}^{-t^2/2}$. In this setting, the marginal tail function is defined as

$$Q_{2\xi}(E) := \inf_{\boldsymbol{U}\in E} \mathbb{P}\big\{\, |\langle \boldsymbol{\Psi}_1,\ \boldsymbol{U}\rangle| \geq 2\xi \big\}.$$

The mean empirical width is defined as

$$W_m(E) := \mathbb{E} \sup_{\boldsymbol{U}\in E} \langle \boldsymbol{H},\ \boldsymbol{U}\rangle \quad \text{where} \quad \boldsymbol{H} := \frac{1}{\sqrt{m}} \sum_{i=1}^{m} \varepsilon_i \boldsymbol{\Psi}_i.$$

Here, $\{\varepsilon_i\}$ is an independent family of Rademacher random variables, independent of everything else.

2.8.5 Step 2: The marginal tail function

We can use the Paley–Zygmund inequality to show that

$$Q_1(E) = \inf_{\boldsymbol{U}\in E} \mathbb{P}\big\{\, |\langle \boldsymbol{\Psi}_1,\ \boldsymbol{U}\rangle| \geq 1 \big\} \geq c_0. \tag{2.40}$$

We have implicitly chosen $\xi = \frac{1}{2}$, and c_0 is a positive absolute constant.

2.8.5.1 The tail bound

To perform this estimate, we apply the Paley–Zygmund inequality in the form

$$\mathbb{P}\Big\{ |\langle \boldsymbol{\Psi}_1,\ \boldsymbol{U}\rangle|^2 \geq \tfrac{1}{2}\big(\mathbb{E}\, |\langle \boldsymbol{\Psi}_1,\ \boldsymbol{U}\rangle|^2\big) \Big\} \geq \frac{1}{4} \cdot \frac{\big(\mathbb{E}\, |\langle \boldsymbol{\Psi}_1,\ \boldsymbol{U}\rangle|^2\big)^2}{\mathbb{E}\, |\langle \boldsymbol{\Psi}_1,\ \boldsymbol{U}\rangle|^4}.$$

The easiest way to treat the expectation in the denominator is to invoke Gaussian hypercontractivity [17, Sec. 3.2]. Indeed,

$$\left(\mathbb{E}\,|\langle \boldsymbol{\Psi}_1,\, \boldsymbol{U}\rangle|^4\right)^{1/4} \le C_0 \left(\mathbb{E}\,|\langle \boldsymbol{\Psi}_1,\, \boldsymbol{U}\rangle|^2\right)^{1/2}$$

because $\langle \boldsymbol{\Psi}_1,\, \boldsymbol{U}\rangle$ is a second-order polynomial in the entries of $\boldsymbol{\psi}_1$. Combine the last two displays to obtain

$$\mathbb{P}\left\{|\langle \boldsymbol{\Psi}_1,\, \boldsymbol{U}\rangle|^2 \ge \tfrac{1}{2}\left(\mathbb{E}\,|\langle \boldsymbol{\Psi}_1,\, \boldsymbol{U}\rangle|^2\right)\right\} \ge \frac{1}{4 \cdot C_0^4} = c_0.$$

We can bound the remaining expectation by means of an explicit calculation. Assuming that $\boldsymbol{U} \in E$,

$$\mathbb{E}\,|\langle \boldsymbol{\Psi}_1,\, \boldsymbol{U}\rangle|^2 = 3\sum_{i=1}^{m} |u_{ii}|^2 + 2\sum_{i,j=1}^{m} |u_{ij}|^2 + \left|\sum_{i=1}^{m} u_{ii}\right|^2 \ge 2.$$

We have used the fact that $\boldsymbol{U}$ is a symmetric matrix with unit Frobenius norm. In conclusion,

$$\mathbb{P}\left\{|\langle \boldsymbol{\Psi}_1,\, \boldsymbol{U}\rangle|^2 \ge 1\right\} \ge c_0 \quad \text{for each } \boldsymbol{U} \in E.$$

This inequality implies (2.40).

2.8.6 *Step 3′: The mean empirical width of the descent cone*

We can apply Proposition 7.1 to demonstrate that the mean empirical width satisfies

$$W_m(E) \le C_1\sqrt{d} \quad \text{for } m \ge C_2 d. \tag{2.41}$$

The numbers C_1 and C_2 are positive, absolute constants.

2.8.6.1 The width bound

The bound holds trivially when $\boldsymbol{X}^\natural = \mathbf{0}$, so we may assume that the unknown matrix is nonzero. Select a coordinate system where

$$\boldsymbol{X}^\natural = \begin{bmatrix} a & \mathbf{0}^{\mathsf{t}} \\ \mathbf{0} & \mathbf{0} \end{bmatrix} \in \mathbb{R}^{d\times d}_{\mathrm{sym}} \quad \text{where } a > 0.$$

Recall that the matrix $\boldsymbol{H} = m^{-1/2} \sum_{i=1}^{m} \varepsilon_i \boldsymbol{\Psi}_i$, where $\boldsymbol{\Psi}_i = \boldsymbol{\psi}_i \boldsymbol{\psi}_i^{\mathsf{t}}$ and $\boldsymbol{\psi}_i \sim \textsc{normal}(\mathbf{0}, \mathbf{I}_d)$. Partition $\boldsymbol{H}$ conformally with $\boldsymbol{X}^\natural$:

$$\boldsymbol{H} = \begin{bmatrix} h_{11} & \boldsymbol{h}_{21}^{\mathsf{t}} \\ \boldsymbol{h}_{21} & \boldsymbol{H}_{22} \end{bmatrix}.$$

Define the random parameter $\tau = \|H_{22}\|$.

$$W_m(E) = \mathbb{E} \sup_{\boldsymbol{U} \in E} \langle \boldsymbol{H}, \boldsymbol{U} \rangle \le \left(\mathbb{E} \operatorname{dist}_{\mathrm{F}}^2 \left(\boldsymbol{H}, \tau \cdot \partial f(\boldsymbol{X}^\natural) \right) \right)^{1/2}. \tag{2.42}$$

Using standard calculus rules for subdifferentials [28, Chap. 23], we determine that

$$\partial f(\boldsymbol{X}^\natural) = \left\{ \begin{bmatrix} 1 & \mathbf{0}^{\mathsf{t}} \\ \mathbf{0} & \boldsymbol{Y} \end{bmatrix} \in \mathbb{R}^{d \times d}_{\mathrm{sym}} : \lambda_{\max}(\boldsymbol{Y}) \le 1 \right\}.$$

We write $\lambda_{\max}$ denotes the maximum eigenvalue of a symmetric matrix. Proposition 7.1 delivers the width bound,

$$\mathbb{E} \operatorname{dist}_{\mathrm{F}}^2 \left(\boldsymbol{H}, \partial f(\boldsymbol{X}^\natural) \right) = \mathbb{E} \, (h_{11} - \tau)^2 + 2 \, \mathbb{E} \, \|\boldsymbol{h}_{21}\|^2 + \mathbb{E} \inf_{\lambda_{\max}(\boldsymbol{Y}) \le 1} \|\boldsymbol{H}_{22} - \tau \cdot \boldsymbol{Y}\|_{\mathrm{F}}^2. \tag{2.43}$$

By construction, the third term on the right-hand side of (2.43) is zero. By direct calculation, the second term on the right-hand side of (2.43) satisfies

$$\mathbb{E} \, \|\boldsymbol{h}_{21}\|^2 = d - 1. \tag{2.44}$$

Finally, we turn to the first term on the right-hand side of (2.43). Relatively crude bounds suffice here. By interlacing of singular values,

$$\tau = \|H_{22}\| \le \|H\| = \frac{1}{\sqrt{m}} \left\| \sum_{i=1}^{m} \varepsilon_i \boldsymbol{\psi}_i \boldsymbol{\psi}_i^{\mathsf{t}} \right\|.$$

Standard net arguments, such as those in [36, Sec. 5.4.1], demonstrate that

$$\mathbb{P}\left\{ \|\boldsymbol{H}\| \ge C_3 \sqrt{d} \right\} \le \mathrm{e}^{-c_1 d}, \quad \text{provided that } m \ge C_2 d.$$

Together, the last two displays imply that $\mathbb{E} \, \tau^2 \le C_4 d$. Therefore,

$$\mathbb{E} \, (h_{11} - \tau)^2 \le C_5 d. \tag{2.45}$$

Introducing (2.43), (2.44), and (2.45) into (2.42), we arrive at the required bound (2.41).

Remark 8.4 (Other sampling distributions). The only challenging part of the calculation is the bound on $\|\boldsymbol{H}\|$. For more general sampling distributions, we can easily obtain the required estimate from the matrix moment inequality [7, Thm. A.1].

2.8.7 Combining the bounds

Assume that $m \geq C_2 d$. Combine the estimates (2.39), (2.40), and (2.41) to reach

$$\lambda_{\min}\big(\boldsymbol{\Phi};\ \mathscr{D}(f, \boldsymbol{X}^\natural)\big) \geq c_2\sqrt{m} - C_6\sqrt{d} - \tfrac{1}{2}t$$

with probability at least $1 - \mathrm{e}^{-t^2/2}$. Choosing $t = c_3\sqrt{m}$, we find that the minimum conic singular value is positive with probability at least $1 - \mathrm{e}^{-c_4 m}$. In this event, Proposition 2.6 implies that $\boldsymbol{X}^\natural$ is the unique solution to the phase retrieval problem (2.37). This observation completes the proof of Theorem 8.2.

Acknowledgements JAT gratefully acknowledges support from ONR award N00014-11-1002, AFOSR award FA9550-09-1-0643, and a Sloan Research Fellowship. Thanks are also due to the Moore Foundation.

References

1. D. Amelunxen, M. Lotz, M.B. McCoy, J.A. Tropp, Living on the edge: phase transitions in convex programs with random data. Inf. Inference **3**(3), 224–294 (2014). Available at http://arXiv.org/abs/1303.6672
2. R. Balan, B.G. Bodmann, P.G. Casazza, D. Edidin, Painless reconstruction from magnitudes of frame coefficients. J. Fourier Anal. Appl. **15**(4), 488–501 (2009)
3. S. Boucheron, G. Lugosi, P. Massart, *Concentration Inequalities: A Nonasymptotic Theory of Independence* (Oxford University Press, Oxford, 2013)
4. T.T. Cai, T. Liang, A. Rakhlin, Geometrizing local rates of convergence for linear inverse problems (April 2014). Available at http://arXiv.org/abs/1404.4408
5. E.J. Candès, T. Strohmer, V. Voroninski, PhaseLift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**(8), 1241–1274 (2013)
6. V. Chandrasekaran, B. Recht, P.A. Parrilo, A.S. Willsky, The convex geometry of linear inverse problems. Found. Comput. Math. **12**(6), 805–849 (2012)
7. R.Y. Chen, A. Gittens, J.A. Tropp, The masked sample covariance estimator: an analysis via the matrix Laplace transform method. Inf. Inference **1**, 2–20 (2012)
8. K.R. Davidson, S.J. Szarek, Local operator theory, random matrices and Banach spaces, in *Handbook of the Geometry of Banach Spaces*, vol. I (North-Holland, Amsterdam, 2001), pp. 317–366
9. M. Fazel. *Matrix Rank Minimization with Applications*. Ph.D. thesis, Stanford University, 2002.

10. R. Foygel, L. Mackey, Corrupted sensing: novel guarantees for separating structured signals. Trans. Inf. Theory **60**(2), 1223–1247 (2014)
11. Y. Gordon, Some inequalities for Gaussian processes and applications. Isr. J. Math. **50**(4), 265–289 (1985)
12. Y. Gordon. On Milman's inequality and random subspaces which escape through a mesh in $\mathbf{R}^n$, in *Geometric Aspects of Functional Analysis (1986/1987)*. Lecture Notes in Mathematics, vol. 1317 (Springer, Berlin, 1988), pp. 84–106
13. D. Gross, Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011)
14. V. Koltchinskii, S. Mendelson, Bounding the smallest singular value of a random matrix without concentration (December 2013). Available at http://arXiv.org/abs/1312.3580
15. R. Latała, K. Oleszkiewicz, On the best constant in the Khinchin-Kahane inequality. Studia Math. **109**(1), 101–104 (1994)
16. G. Lecué, S. Mendelson, Compressed sensing under weak moment assumptions (January 2014). Available at http://arXiv.org/abs/1401.2188
17. M. Ledoux, M. Talagrand, *Probability in Banach Spaces: Isoperimetry and Processes* (Springer, Berlin, 1991)
18. S. Mendelson. Empirical processes with a bounded ψ_1 diameter. Geom. Funct. Anal. **20**(4), 988–1027 (2010)
19. S. Mendelson, A remark on the diameter of random sections of convex bodies (December 2013). Available at http://arXiv.org/abs/1312.3608
20. S. Mendelson, Learning without concentration. J. Assoc. Comput. Mach. (2014, to appear). **62**(3), (2015). Available at http://arXiv.org/abs/1401.0304
21. S. Mendelson, Learning without concentration for general loss functions (October 2014). Available at http://arXiv.org/abs/1410.3192
22. S. Mendelson, A. Pajor, N. Tomczak-Jaegermann, Reconstruction and subgaussian operators in asymptotic geometric analysis. Geom. Funct. Anal. **17**(4), 1248–1282 (2007)
23. V. Milman, G. Schechtman, *Asymptotic Theory of Finite-Dimensional Normed Linear Spaces*. Number 1200 in LNM (Springer, New York, 1986)
24. S. Oymak, B. Hassibi, New null space results and recovery thresholds for matrix rank minimization. Partial results presented at ISIT 2011 (2010). Available at http://arXiv.org/abs/1011.6326
25. S. Oymak, B. Hassibi, Sharp MSE bounds for proximal denoising. Partial results presented at Allerton 2012 (March 2013). Available at http://arxiv.org/abs/1305.2714
26. S. Oymak, C. Thrampoulides, B. Hassibi, Simple bounds for noisy linear inverse problems with exact side information (December 2013). Available at http://arXiv.org/abs/1312.0641
27. G. Pisier, *The Volume of Convex Bodies and Banach Space Geometry* (Cambridge University Press, Cambridge, 1989)
28. R.T. Rockafellar, *Convex Analysis* (Princeton University Press, Princeton, 1970)
29. M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**(8), 1025–1045 (2008)
30. M. Sion, On general minimax theorems. Pac. J. Math. **8**, 171–176 (1958)
31. M. Stojnic, Various thresholds for ℓ_1-optimization in compressed sensing (2009). Available at http://arXiv.org/abs/0907.3666
32. M. Talagrand, *The Generic Chaining. Upper and Lower Bounds of Stochastic Processes*. Springer Monographs in Mathematics (Springer, Berlin, 2005)
33. C. Thrampoulides, S. Oymak, B. Hassibi, Simple error bounds for regularized noisy linear inverse problems. Appeared at ISIT 2014 (January 2014). Available at http://arXiv.org/abs/1401.6578
34. J. Tropp, M. Wakin, M. Duarte, D. Baron, R. Baraniuk, Random filters for compressive sampling and reconstruction, in *Proceedings of the 2006 IEEE International Conference on Acoustics, Speech and Signal Processing, 2006 (ICASSP 2006)*, vol. 3, May 2006, pp. III–875

35. A.W. van der Vaart, J.A. Wellner, *Weak Convergence and Empirical Processes*. Springer Series in Statistics (Springer, New York, 1996). With applications to statistics.
36. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, in *Compressed Sensing* (Cambridge University Press, Cambridge, 2012), pp. 210–268
37. G.A. Watson, Characterization of the subdifferential of some matrix norms. Linear Algebra Appl. **170**, 33–45 (1992)

Chapter 3
Low Complexity Regularization of Linear Inverse Problems

Samuel Vaiter, Gabriel Peyré, and Jalal Fadili

Abstract Inverse problems and regularization theory is a central theme in imaging sciences, statistics, and machine learning. The goal is to reconstruct an unknown vector from partial indirect, and possibly noisy, measurements of it. A now standard method for recovering the unknown vector is to solve a convex optimization problem that enforces some prior knowledge about its structure. This chapter delivers a review of recent advances in the field where the regularization prior promotes solutions conforming to some notion of simplicity/low complexity. These priors encompass as popular examples sparsity and group sparsity (to capture the compressibility of natural signals and images), total variation and analysis sparsity (to promote piecewise regularity), and low rank (as natural extension of sparsity to matrix-valued data). Our aim is to provide a unified treatment of all these regularizations under a single umbrella, namely the theory of partial smoothness. This framework is very general and accommodates all low complexity regularizers just mentioned, as well as many others. Partial smoothness turns out to be the canonical way to encode low-dimensional models that can be linear spaces or more general smooth manifolds. This review is intended to serve as a one stop shop toward the understanding of the theoretical properties of the so-regularized solutions. It covers a large spectrum including (i) recovery guarantees and stability to noise, both in terms of ℓ^2-stability and model (manifold) identification; (ii) sensitivity analysis to perturbations of the parameters involved (in particular the observations), with applications to unbiased risk estimation; (iii) convergence properties of the forward-backward proximal splitting scheme that is particularly well suited to solve the corresponding large-scale regularized optimization problem.

S. Vaiter (✉)
Ceremade, Université Paris-Dauphine, Paris, France
e-mail: samuel.vaiter@ceremade.dauphine.fr

G. Peyré
CNRS and Ceremade, Université Paris-Dauphine, Paris, France
e-mail: gabriel.peyre@ceremade.dauphine.fr

J. Fadili
GREYC, CNRS-ENSICAEN-Université de Caen, Caen, France
e-mail: Jalal.Fadili@greyc.ensicaen.fr

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_3

3.1 Inverse Problems and Regularization

In this chapter, we deal with finite-dimensional linear inverse problems.

3.1.1 Forward Model

Let $x_0 \in \mathbb{R}^N$ be the unknown vector of interest. Suppose that we observe a vector $y \in \mathbb{R}^P$ of P linear measurements according to

$$y = \Phi x_0 + w, \tag{3.1}$$

where $w \in \mathbb{R}^P$ is a vector of unknown errors contaminating the observations. The forward model (3.1) offers a model for data acquisition that describes a wide range of problems in data processing, including signal and image processing, statistics, and machine learning. The linear operator $\Phi : \mathbb{R}^N \to \mathbb{R}^P$, assumed to be known, is typically an idealization of the acquisition hardware in imaging science applications, or the design matrix in a parametric statistical regression problem. The noise w can be either deterministic (in this case, one typically assumes to know some bound on its ℓ^2 norm $\|w\|$) or random (in which case its distribution is assumed to be known). Except in Sections 3.4.4 and 3.5.3 where the noise is explicitly assumed random, w is deterministic throughout the rest of the chapter. We refer to [189] and [22] for a comprehensive account on noise models in imaging systems.

Solving an inverse problem amounts to recovering x_0, to a good approximation, knowing y and Φ according to (3.1). Unfortunately, the number of measurements P can be much smaller than the ambient dimension N of the signal. Even when $P = N$, the mapping Φ is in general ill conditioned or even singular. This entails that the inverse problem is in general ill posed. In signal or image processing, one might for instance think of Φ as a convolution with the camera point-spread function, or a subsampling accounting for low-resolution or damaged sensors. In medical imaging, typical operators represent a (possibly subsampled) Radon transform (for computerized tomography), a partial Fourier transform (for magnetic resonance imaging), a propagation of the voltage/magnetic field from the dipoles to the sensors (for electro- or magnetoencephalography). In seismic imaging, the action of Φ amounts to a convolution with a wavelet-like impulse response that approximates the solution of a wave propagation equation in media with discontinuities. For regression problems in statistics and machine learning, Φ is the design matrix whose columns are P covariate vectors.

3.1.2 Variational Regularization

As argued above, solving an inverse problem from observations (3.1) is in general ill posed. In order to reach the land of well-posedness, it is necessary to restrict the inversion process to a well-chosen subset of $\mathbb{R}^N$ containing the plausible solutions including x_0, e.g., a linear space or a union of subspaces. A closely related procedure, that we describe next, amounts to adopting a variational framework where the sought-after solutions are those where a prior penalty/regularization function is the smallest. Though this approach may have a maximum a posteriori Bayesian interpretation, where a random prior is placed on x_0, this is not the only interpretation. In fact, we put no randomness whatsoever on the class of signals we look for. We will not elaborate more on these differences in this chapter, but the reader may refer to [119] for an insightful discussion.

The foundations of regularization theory can be traced back to the pioneering work of the Russian school, and in particular of Tikhonov in 1943 when he proposed the notion of conditional well-posedness. In 1963, Tikhonov [216, 217] introduced what is now commonly referred to as Tikhonov (or also Tikhonov-Phillips) regularization, see also the book [218]. This corresponds, for $\lambda > 0$, to solving an optimization problem of the form

$$x^\star \in \operatorname*{Argmin}_{x\in\mathbb{R}^N} \frac{1}{2\lambda}\|\Phi x - y\|^2 + J(x). \qquad (\mathscr{P}_{y,\lambda})$$

3.1.2.1 Data fidelity

In $(\mathscr{P}_{y,\lambda})$, $\|\Phi x - y\|^2$ stands for the data fidelity term. If the noise happens to be random, then using a likelihood argument, an appropriate fidelity term conforming to the noise distribution can be used instead of the quadratic data fidelity. Clearly, it is sufficient then to replace the latter by the negative log-likelihood of the distribution underlying the noise. Think for instance of the Csiszár's I-divergence for Poisson noise. We would also like to stress that many of the results provided in this chapter extend readily when the quadratic loss in the fidelity term, i.e., $\mu \mapsto \|y-\mu\|^2$, is replaced by any smooth and strongly convex function, see in particular Remark 13. To make our exposition concrete and digestible, we focus in the sequel on the quadratic loss.

3.1.2.2 Regularization

The function $J : \mathbb{R}^N \to \mathbb{R}$ is the regularization term which is intended to promote some prior on the vector to recover. We will consider throughout this chapter that J is a convex finite-valued function. Convexity plays an important role at many locations, both on the recovery guarantees and the algorithmic part. See for instance

Section 3.6 which gives a brief overview of recent algorithms that are able to tackle this class of convex optimization problems. It is however important to realize that non-convex regularizing penalties, as well as non-variational methods (e.g., greedy algorithms), are routinely used for many problems such as sparse or low-rank recovery. They may even outperform in practice their convex counterparts/relaxation. It is however beyond the scope of this chapter to describe these algorithms and the associated theoretical performance guarantees. We refer to Section 3.2.1 for a brief account on non-convex model selection approaches.

The scalar $\lambda > 0$ is the regularization parameter. It balances the trade-off between fidelity and regularization. Intuitively, and anticipating on our theoretical results hereafter, this parameter should be adapted to the noise level $\|w\|$ and the known properties of the vector x_0 to recover. Selecting optimally and automatically λ for a given problem is however difficult in general. This is at the heart of Section 3.5, where unbiased risk estimation strategies are shown to offer a versatile solution.

Note that since Φ is generally not injective and J is not coercive, the objective function of $(\mathscr{P}_{y,\lambda})$ is neither coercive nor strictly convex. In turn, there might be existence (of minimizers) issues, and even if minimizers exist, they are not unique in general.

Under mild assumptions, problem $(\mathscr{P}_{y,\lambda})$ is formally equivalent to the constrained formulations

$$\min \{J(x) \;;\; \|y - \Phi x\| \leqslant \varepsilon\}\,, \tag{$\mathscr{P}^1_{y,\varepsilon}$}$$

$$\min \{\|y - \Phi x\| \;;\; J(x) \leqslant \gamma\}\,, \tag{$\mathscr{P}^2_{y,\gamma}$}$$

in the sense that there exists a bijection between each pair of parameters among $(\lambda, \varepsilon, \gamma)$ so that the corresponding problems share the same set of solutions. However, this bijection is not explicit and depends on y, so that both from an algorithmic point of view and a theoretical one, each problem may need to be addressed separately. See the recent paper [60] and references therein for a detailed discussion, and [154, Theorem 2.3] valid also in the non-convex case. We focus in this chapter on the penalized/Tikhonov formulation $(\mathscr{P}_{y,\lambda})$, though most of the results stated can be extended to deal with the constrained ones $(\mathscr{P}^1_{y,\varepsilon})$ and $(\mathscr{P}^2_{y,\gamma})$ (the former is known as the residual method or Mozorov regularization and the latter as Ivanov regularization in the inverse problems literature).

The value of λ should typically be an increasing function of $\|w\|$. In the special case where there is no noise, i.e., $w = 0$, the fidelity to data should be perfect, which corresponds to considering the limit of $(\mathscr{P}_{y,\lambda})$ as $\lambda \to 0^+$. Thus, assuming that $y \in \mathrm{Im}(\Phi)$, as is the case when $w = 0$, it can be proved that the solutions of $(\mathscr{P}_{y,\lambda})$ converge to the solutions of the following constrained problem [196, 216]

$$x^\star \in \operatorname*{Argmin}_{x \in \mathbb{R}^N} J(x) \quad \text{subject to} \quad \Phi x = y. \tag{$\mathscr{P}_{y,0}$}$$

3.1.3 Notations

For any subspace T of $\mathbb{R}^N$, we denote P_T the orthogonal projection onto T, $x_T = \mathrm{P}_T(x)$ and $\Phi_T = \Phi\,\mathrm{P}_T$. For a matrix A, we denote A^* its transpose, and A^+ its Moore-Penrose pseudoinverse. For a convex set E, $\mathrm{aff}(E)$ denotes its affine hull (i.e., the smallest affine space containing it) and $\mathrm{lin}(E)$ its linear hull (i.e., the linear space parallel to $\mathrm{aff}(E)$). Its relative interior $\mathrm{ri}(E)$ is the interior for the topology of $\mathrm{aff}(E)$ and $\mathrm{rbd}(E)$ is its relative boundary. For a manifold $\mathscr{M}$, we denote $\mathscr{T}_{\mathscr{M}}(x)$ the tangent space of $\mathscr{M}$ at $x \in \mathscr{M}$. A good source on smooth manifold theory is [143].

A function $J : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is said to be proper if it is not identically $+\infty$. It is said to be finite valued if $J(x) \in \mathbb{R}$ for all $x \in \mathbb{R}^N$. We denote $\mathrm{dom}(J)$ the set of points x where $J(x) \in \mathbb{R}$ is finite. J is said to be closed if its epigraph $\{(x, y)\ ;\ J(x) \leqslant y\}$ is closed. For a set $C \subset \mathbb{R}^N$, the indicator function ι_C is defined as $\iota_C(x) = 0$ if $x \in C$ and $\iota_C(x) = +\infty$ otherwise.

We recall that the subdifferential at x of a proper and closed convex function $J : \mathbb{R}^N \to \mathbb{R} \cup \{+\infty\}$ is the set

$$\partial J(x) = \left\{\eta \in \mathbb{R}^N\ ;\ \forall\, \delta \in \mathbb{R}^N,\ J(x+\delta) \geqslant J(x) + \langle \eta, \delta \rangle\right\}.$$

Geometrically, when J is finite at x, $\partial J(x)$ is the set of normals to the hyperplanes supporting the graph of J and tangent to it at x. Thus, $\partial J(x)$ is a closed convex set. It is moreover bounded, hence compact, if and only if $x \in \mathrm{int}(\mathrm{dom}(J))$. The size of the subdifferential at $x \in \mathrm{dom}(J)$ reflects in some sense the degree of non-smoothness of J at x. The larger the subdifferential at x, the larger the "kink" of the graph of J at x. In particular, if J is differentiable at x, then $\partial J(x)$ is a singleton and $\partial J(x) = \{\nabla J(x)\}$.

As an illustrative example, the subdifferential of the absolute value is

$$\forall\, x \in \mathbb{R}, \quad \partial|\cdot|(x) = \begin{cases} \mathrm{sign}(x) & \text{if} \quad x \neq 0, \\ [-1, 1] & \text{otherwise.} \end{cases} \tag{3.2}$$

The ℓ^1 norm

$$\forall\, x \in \mathbb{R}^N, \quad \|x\|_1 = \sum_{i=1}^{N} |x_i|$$

is a popular low complexity prior (see Section 3.2.3.1 for more details). Formula (3.2) is extended by separability to obtain the subdifferential of the ℓ^1 norm

$$\partial\, \|\cdot\|_1\,(x) = \left\{\eta \in \mathbb{R}^N\ ;\ \|\eta\|_\infty \leqslant 1 \quad \text{and} \quad \forall\, i \in I,\ \mathrm{sign}(\eta_i) = \mathrm{sign}(x_i)\right\} \tag{3.3}$$

where $I = \mathrm{supp}(x) = \{i\ ;\ x_i \neq 0\}$. Note that at a point $x \in \mathbb{R}^N$ such that $x_i \neq 0$ for all i, $\|\cdot\|_1$ is differentiable and $\partial\, \|\cdot\|_1\,(x) = \{\mathrm{sign}(x)\}$.

3.2 Low Complexity Priors

A recent trend in signal and image processing, statistics, and machine learning is to make use of large collections of the so-called models to account for the complicated structures of the data to handle. Generally speaking, these are manifolds $\mathscr{M}$ (most of the time linear subspaces), and hopefully of low complexity (to be detailed later), that capture the properties of the sought after signal, image, or higher dimensional data. In order to tractably manipulate these collections, the key idea underlying this approach is to encode these manifolds in the nonsmooth parts of the regularizer J. As we detail here, the theory of partial smoothness turns out to be natural to provide a mathematically grounded and unified description of these regularizing functions.

3.2.1 Model Selection

The general idea is thus to describe the data to recover using a large collection of models $\mathbb{M} = \{\mathscr{M}\}_{\mathscr{M}\in\mathbb{M}}$, which are manifolds. The "complexity" of elements in such a manifold $\mathscr{M}$ is measured through a penalty $\mathrm{pen}(\mathscr{M})$. A typical example is simply the dimensionality of $\mathscr{M}$, and it should reflect the intuitive notion of the number of parameters underlying the description of the vector $x_0 \in \mathscr{M}$ that one aims at recovering from the noisy measurements of the form (3.1). As popular examples of such low complexity, one thinks of sparsity, piecewise regularity, or low rank. Penalizing in accordance to some notion of complexity is a key idea, whose roots can be traced back to the statistical and information theory literature, see for instance [2, 161].

Within this setting, the inverse problem associated to measurements (3.1) is solved by restricting the inversion to an optimal manifold as selected by $\mathrm{pen}(\mathscr{M})$. Formally, this would correspond to solving ($\mathscr{P}_{y,\lambda}$) with the combinatorial regularizer

$$J(x) = \inf\{\mathrm{pen}(\mathscr{M}) \;;\; \mathscr{M} \in \mathbb{M} \quad \text{and} \quad x \in \mathscr{M}\}. \tag{3.4}$$

A typical example of such a model selection framework is that with sparse signals, where the collection $\mathbb{M}$ corresponds to a union of subspaces, each of the form

$$\mathscr{M} = \left\{x \in \mathbb{R}^N \;;\; \mathrm{supp}(x) \subseteq I\right\}.$$

Here $I \subseteq \{1,\ldots,N\}$ indexes the supports of signals in $\mathscr{M}$ and can be arbitrary. In this case, one uses $\mathrm{pen}(\mathscr{M}) = \dim(\mathscr{M}) = |I|$, so that the associated combinatorial penalty is the so-called ℓ^0 pseudonorm

$$J(x) = \|x\|_0 = |\,\mathrm{supp}(x)| = |\,\{i \in \{1,\ldots,N\} \;;\; x_i \neq 0\}\,|. \tag{3.5}$$

Thus, solving ($\mathscr{P}_{y,\lambda}$) is intended to select a few active variables (corresponding to nonzero coefficients) in the recovered vector.

These sparse models can be extended in many ways. For instance, piecewise regular signals or images can be modeled using manifolds $\mathscr{M}$ that are parameterized by the locations of the singularities and some low-order polynomial between these singularities. The dimension of $\mathscr{M}$ thus grows with the number of singularities, hence the complexity of the model.

Literature review. The model selection literature [11, 17, 18] proposes many theoretical results to quantify the performance of these approaches. However, a major bottleneck of this class of methods is that the corresponding J function defined in (3.4) is non-convex, and even not necessarily closed, thus typically leading to highly intractable combinatorial optimization problems. For instance, in the case of ℓ^0 penalty (3.5) and for an arbitrary operator Φ, ($\mathscr{P}_{y,\lambda}$) is known to be NP-hard, see, e.g., [167].

It then appears crucial to propose alternative strategies which allow us to deploy fast computational algorithms. A first line of work consists in finding stationary points of ($\mathscr{P}_{y,\lambda}$) using descent-like schemes. For instance, in the case of ℓ^0 pseudo-norm, this can be achieved using iterative hard thresholding [20, 210], or iterative reweighting schemes which consist of solving a sequence of weighted ℓ^1- or ℓ^2-minimization problems where the weights used for the next iteration are computed from the values of the current solution, see for instance [45, 72, 187] and references therein. Another class of approaches is that of greedy algorithms. These are algorithms which explore the set of possible manifolds $\mathscr{M}$ by progressively, actually in a greedy fashion, increasing the value of pen($\mathscr{M}$). The most popular schemes are matching pursuit [160] and its orthogonal variant [73, 179], see also the comprehensive review [168] and references therein. The last line of research, which is the backbone of this chapter, consists in considering convex regularizers which are built in such a way that they promote the same set of low complexity manifolds $\mathbb{M}$. In some cases, the convex regularizer proves to be the convex hull of the initial (restricted) non-convex combinatorial penalty (3.4). But these convex penalties can also be designed without being necessarily convexified surrogates of the original non-convex ones.

In the remainder of this section, we describe in detail a general framework that allows model selection through the general class of convex partly smooth functions.

3.2.2 *Encoding Models into Partly Smooth Functions*

Before giving the precise definition of our class of convex priors, we define formally the subspace T_x.

Definition 1 (Model tangent subspace). For any vector $x \in \mathbb{R}^N$, we define the *model tangent subspace* of x associated to J

$$T_x = \mathrm{lin}(\partial J(x))^{\perp}.$$

In fact, the terminology "tangent" originates from the sharpness property of Definition 2(ii) below, when x belongs to the manifold $\mathscr{M}$.

When J is differentiable at x, i.e., $\partial J(x) = \{\nabla J(x)\}$, one has $T_x = \mathbb{R}^N$. On the contrary, when J is not smooth at x, the dimension of T_x is of a strictly smaller dimension, and J essentially promotes elements living on or close to the affine space $x + T_x$.

We can illustrate this using the ℓ^1 norm $J = \|\cdot\|_1$ defined in (3.2). Using formula (3.3) for the subdifferential, one obtains that

$$T_x = \left\{u \in \mathbb{R}^N \;;\; \mathrm{supp}(u) \subseteq \mathrm{supp}(x)\right\},$$

which is the set of vector having the same sparsity pattern as x.

Toward the goal of studying the recovery guarantees of problem (3.4), our central assumption is that J is a partly smooth function relative to some manifold $\mathscr{M}$. Partial smoothness of functions was originally defined [145]. Loosely speaking, a partly smooth function behaves smoothly as we move on the manifold $\mathscr{M}$, and sharply if we move normal to it. Our definition hereafter specializes that of [145] to the case of finite-valued convex functions.

Definition 2. Let J be a finite-valued convex function. J is *partly smooth at x relative to a set $\mathscr{M}$* containing x if

(i) (Smoothness) $\mathscr{M}$ is a C^2-manifold around x and J restricted to $\mathscr{M}$ is C^2 around x.
(ii) (Sharpness) The tangent space $\mathscr{T}_{\mathscr{M}}(x)$ is T_x.
(iii) (Continuity) The set-valued mapping ∂J is continuous at x relative to $\mathscr{M}$.

J is said to be *partly smooth relative to a set $\mathscr{M}$* if $\mathscr{M}$ is a manifold and J is partly smooth at each point $x \in \mathscr{M}$ relative to $\mathscr{M}$. J is said to be *locally partly smooth at x relative to a set $\mathscr{M}$* if $\mathscr{M}$ is a manifold and there exists a neighborhood U of x such that J is partly smooth at each point of $\mathscr{M} \cap U$ relative to $\mathscr{M}$.

Remark 1 (Uniqueness of $\mathscr{M}$). In the previous definition, $\mathscr{M}$ needs only to be defined locally around x, and it can be shown to be locally unique, see [131, Corollary 4.2]. In the following we will thus often denote $\mathscr{M}_x$ any such a manifold for which J is partly smooth at x.

Taking once again the example of $J = \|\cdot\|_1$, one sees that in this case, $\mathscr{M}_x = T_x$ because this function is polyhedral. Section 3.2.3.6 below defines functions J for which $\mathscr{M}_x$ differs in general from T_x.

3.2.3 Examples of Partly Smooth Regularizers

We describe below some popular examples of partly smooth regularizers that are widely used in signal and image processing, statistics, and machine learning. We first expose basic building blocks (sparsity, group sparsity, anti-sparsity) and then show how the machinery of partial smoothness enables a powerful calculus to create new priors (using pre- and post-composition, spectral lifting, and positive linear combinations).

3.2.3.1 ℓ^1 Sparsity

One of the most popular nonquadratic convex regularization is the ℓ^1 norm

$$J(x) = \|x\|_1 = \sum_{i=1}^{N} |x_i|,$$

which promotes sparsity. Indeed, it is easy to check that J is partly smooth at x relative to the subspace

$$\mathscr{M}_x = T_x = \left\{u \in \mathbb{R}^N \;;\; \mathrm{supp}(u) \subseteq \mathrm{supp}(x)\right\}.$$

Another equivalent way to interpret this ℓ^1 prior is that it is the convex envelope (restricted to the ℓ^2-ball) of the ℓ^0 pseudonorm (3.5), in the sense that the ℓ^1-unit ball is the convex hull of the restriction of the unit ball of the ℓ^0 pseudonorm to the ℓ^2-unit ball.

Literature review. The use of the ℓ^1 norm as a sparsity-promoting regularizer traces back several decades. An early application was deconvolution in seismology [61, 195, 211]. Rigorous recovery results began to appear in the late 1980s [80, 81]. In the mid-1990s, ℓ^1 regularization of least-square problems has been popularized in the signal processing literature under the name Basis Pursuit [58] and in the statistics literature under the name Lasso [212]. Since then, the applications and understanding of ℓ^1 minimization have continued to increase dramatically.

3.2.3.2 $\ell^1 - \ell^2$ Group Sparsity

To better capture the sparsity pattern of natural signals and images, it is useful to structure the sparsity into nonoverlapping groups $\mathscr{B}$ such that $\bigcup_{b\in\mathscr{B}} b = \{1, \ldots, N\}$. This group structure is enforced by using typically the mixed $\ell^1 - \ell^2$ norm

$$J(x) = \|x\|_{1,\mathscr{B}} = \sum_{b\in\mathscr{B}} \|x_b\|\,, \tag{3.6}$$

where $x_b = (x_i)_{i\in b} \in \mathbb{R}^{|b|}$. Unlike the ℓ^1 norm, and except the case $|b| = 1$ for all $b \in \mathscr{B}$, the $\ell^1 - \ell^2$ norm is not polyhedral, but is still partly smooth at x relative to the linear manifold

$$\mathscr{M}_x = T_x = \{u\,;\ \mathrm{supp}_{\mathscr{B}}(u) \subseteq \mathrm{supp}_{\mathscr{B}}(x)\} \quad \text{where} \quad \mathrm{supp}_{\mathscr{B}}(x) = \bigcup \{b\,;\ x_b \neq 0\}\,.$$

Literature review. The idea of group/block sparsity has been first proposed by [31, 125, 126] for wavelet block shrinkage, i.e., when $\Phi = \mathrm{Id}$. For overdetermined regression problems of the form (3.1), it has been introduced by [9, 242]. Group sparsity has also been extensively used in machine learning in, e.g., [7] (regression and multiple kernel learning) and [174] (for multitask learning). The wavelet coefficients of a natural image typical exhibit some group structure, see [159] and references therein on natural image modeling. Indeed, edges and textures induce strong dependencies between coefficients. In audio processing, it has proved useful to structure sparsity in multi-channel data [122]. Group sparsity is also at the heart of the so-called multiple measurements vector (MMV) model, see for instance [57, 69]. It is possible to replace the ℓ^2 norm with more general functionals, such as ℓ^p norms for $p > 1$, see for instance [169, 224, 236].

3.2.3.3 ℓ^∞ Anti-sparsity

In some cases, the vector to be reconstructed is expected to be flat. Such a prior can be captured using the ℓ^∞ norm

$$J(x) = \|x\|_\infty = \max_{i\in\{1,\ldots,n\}} |x_i|.$$

It can be readily checked that this regularizer is partly smooth (in fact polyhedral) relative to the subspace

$$\mathscr{M}_x = T_x = \{u\,;\ u_I = \rho x_I \text{ for some } \rho \in \mathbb{R}\}\,, \quad \text{where} \quad I = \{i\,;\ x_i = \|x\|_\infty\}\,.$$

Literature review. The ℓ^∞ regularization has found applications in computer vision, such as for database image retrieval [136]. For this application, it is indeed useful to have a compact signature of a signal x, ideally with only two values $\pm\|x\|_\infty$ (thus achieving optimal anti-sparsity since $\dim(T_x) = 1$ in such a case). An approach proposed in [137] for realizing this binary quantification is to compute these vectors as solutions of $(\mathscr{P}_{y,\lambda})$ for $J = \|\cdot\|_\infty$ and a random Φ. A study of this regularization is done in [108], where an homotopy-like algorithm is provided. The use of this ℓ^∞ regularization is also connected to Kashin's representation [156],

which is known to be useful in stabilizing the quantization error for instance. Other applications such as wireless network optimization [209] also rely on the ℓ^∞ prior.

3.2.3.4 Synthesis Regularizers

Sparsity or more general low complexity regularizations are often used to model coefficients $\alpha \in \mathbb{R}^Q$ describing the data $x = D\alpha$ in a dictionary $D \in \mathbb{R}^{N\times Q}$ of Q atoms in $\mathbb{R}^N$. Given a partly smooth function $J_0 : \mathbb{R}^Q \to \mathbb{R}$, we define the following synthesis-type prior $J : \mathbb{R}^N \to \mathbb{R}$ as the pre-image of J_0 under the linear mapping D

$$J(x) = \min_{\alpha\in\mathbb{R}^Q} J_0(\alpha) \quad \text{s.t.} \quad D\alpha = x$$

Since J_0 is bounded below and convex, J is convex. If D is surjective (as in most cases with redundant dictionaries), then J is also finite valued. The initial optimization $(\mathscr{P}_{y,\lambda})$ can equivalently been solved directly over the coefficients domain to obtain $x^\star = D\alpha^\star$ where

$$\alpha^\star \in \operatorname*{Argmin}_{\alpha\in\mathbb{R}^Q} \frac{1}{2\lambda}\|y - \Phi D\alpha\|^2 + J_0(\alpha) \tag{3.7}$$

which can be interpreted as a regularized inversion of the operator ΦD using the prior J_0.

It is possible to study directly the properties of the solutions $\alpha^\star$ to (3.7), which involves directly partial smoothness of J_0. A slightly different question is to understand the behavior of the solutions $x^\star = D\alpha^\star$ of $(\mathscr{P}_{y,\lambda})$, which requires to study partial smoothness of J itself. In the case where D is invertible, both problems are completely equivalent.

Literature review. Sparse synthesis regularization using $J_0 = \|\cdot\|_1$ is popular in signal and image processing to model natural signals and images, see for instance [159, 205] for a comprehensive account. The key problem to achieve good performance in these applications is to design a dictionary to capture sparse representations of the data to process. Multiscale dictionaries built from wavelet pyramids are popular to sparsely represent transient signals with isolated singularities and natural images [158]. The curvelet transform is known to provide nonadaptive near-optimal sparse representation of piecewise smooth images away from smooth edges (the so-called cartoon images) [34]. Gabor dictionaries (made of localized and translated Fourier atoms) are popular to capture locally stationary oscillating signals for audio processing [3]. To cope with richer and diverse contents, researchers have advocated to concatenate several dictionaries to solve difficult problems in signal and image processing, such as component separation or inpainting, see for instance [98]. A line of current active research is to learn and optimize the dictionary from exemplars or even from the available data themselves. We refer to [97, Chapter 12] for a recent overview of the relevant literature.

3.2.3.5 Analysis Regularizers

Analysis-type regularizers (following the terminology introduced in [99]) are of the form

$$J(x) = J_0(D^* x) \ ,$$

where $D \in \mathbb{R}^{N \times Q}$ is a linear operator. Such a prior controls the low complexity (as measured by J_0) of the correlations between the columns of D and the signal x. If J_0 is partly smooth at $z = D^* x$ for the manifold $\mathscr{M}_z^0$, then it is shown in [145, Theorem 4.2] that J is partly smooth at x relative to the manifold

$$\mathscr{M}_x = \left\{ u \in \mathbb{R}^N \ ; \ D^* u \in \mathscr{M}_z^0 \right\}$$

provided that the following transversality condition holds [143, Theorem 6.30(a)]

$$\mathrm{Ker}(D) \cap \mathscr{T}_{\mathscr{M}_z^0}(z)^\perp = \{0\} \iff \mathrm{Im}(D^*) + \mathscr{T}_{\mathscr{M}_z^0}(z) = \mathbb{R}^N \ .$$

Literature review. A popular example is when $J_0 = \|\cdot\|_1$ and D^* is a finite-difference discretization of the derivative of a 1-D signal or a 2-D image. This defines the anisotropic total variation semi-norm, which promotes piecewise constant signals or images [194]. The 2-D isotropic total variation semi-norm can be interpreted as taking $J_0 = \|\cdot\|_{1,2}$ with blocks of size two. A comprehensive review of total variation regularization can be found in [53]. TV regularization has been extended in several ways to model piecewise polynomial functions, see in particular the Total Generalized Variation prior [28].

One can also use a wavelet dictionary D which is shift invariant, such that the corresponding regularization J can be seen as a kind of multiscale total variation. This is typically the case of the Haar wavelet dictionary [206]. When using higher order wavelets, the corresponding priors favor models $\mathscr{M}$ composed of discrete piecewise polynomials.

The Fused Lasso [215] corresponds to J_0 being the ℓ^1 norm and D is the concatenation of the identity and the adjoint of a finite-difference operator. The corresponding models $\mathscr{M}$ are composed of disjoint blocks over which the signals are constant.

Defining a block extracting operator $D^* x = (x_b)_{b \in \mathscr{B}}$ allows to rewrite the group ℓ^1–ℓ^2 norm (3.6), even with overlapping blocks (i.e., $\exists (b, b') \in \mathscr{B}^2$ with $b \cap b' \neq \emptyset$), as $J = J_0 \circ D^*$ where $J_0 = \|\cdot\|_{1,2}$ without overlap, see [32, 138, 182, 244]. To cope with correlated covariates in linear regression, analysis-type sparsity-enforcing priors were proposed in [118, 191] using $J_0 = \|\cdot\|_*$ the nuclear norm (as defined in Section 3.2.3.6).

For unitary D, the solutions of ($\mathscr{P}_{y,\lambda}$) with synthesis and analysis regularizations are obviously the same. In the general case (e.g., D overcomplete), however, these two regularizations are different. Some authors have reported results comparing these two priors for the case where J_0 is the ℓ^1 norm [99, 197]. A first discussion on

the relation and distinction between analysis and synthesis ℓ^1-sparse regularizations can be found in [99]. But only very recently, some theoretical recovery results and algorithmic developments on ℓ^1- analysis regularization (so-called cosparse model) have begaun to be developed, see, e.g., [166, 229].

3.2.3.6 Spectral Functions

The natural extension of low complexity priors to matrix-valued data $x \in \mathbb{R}^{N_0 \times N_0}$ (where $N = N_0^2$) is to impose the low complexity on the singular values of the matrix. We denote $x = U_x \operatorname{diag}(\Lambda_x) V_x^*$ an SVD decomposition of x, where $\Lambda_x \in \mathbb{R}_+^{N_0}$. If $j : \mathbb{R}^{N_0} \to \mathbb{R}$ is a permutation-invariant closed convex function, then one can consider the function

$$J(x) = j(\Lambda_x)$$

which can be shown to be a convex function as well [146]. When restricted to the linear space of symmetric matrices, j is partly smooth at Λ_x for a manifold m_{Λ_x}, if and only if J is partly smooth at x relative to the manifold

$$\mathscr{M}_x = \{U \operatorname{diag}(\Lambda) U^* \;;\; \Lambda \in m_{\Lambda_x}, U \in \mathscr{O}_{N_0}\},$$

where $\mathscr{O}_{N_0} \subset \mathbb{R}^{N_0 \times N_0}$ is the orthogonal group. The proof of this assertion can be found in [70, Theorem 3.19], which builds upon the work of [71] on manifold smoothness transfer under spectral lifting. This result can be extended to non-symmetric matrices by requiring that j is an absolutely permutation-invariant closed convex function, see [70, Theorem 5.3].

Literature review. The most popular spectral prior is obtained for $j = \|\cdot\|_1$. This defines the nuclear norm, or 1-Schatten norm, as

$$J(x) = \|x\|_* = \|\Lambda_x\|_1 \;. \tag{3.8}$$

It can be shown that the nuclear norm is the convex hull of the rank function with respect to the spectral norm ball, see [102, 132]. It then corresponds to promoting a low-rank prior. Moreover, the nuclear norm can be shown to be partly smooth at x relative to the set [147, Example 2]

$$\mathscr{M}_x = \{u \;;\; \operatorname{rank}(u) = \operatorname{rank}(x)\}$$

which is a manifold around x.

The nuclear norm has been used in signal and image processing, statistics, and machine learning for various applications, including low-rank matrix completion [38, 188, 203], principal component pursuit [47], model reduction [103], and phase retrieval [49]. It is also used for some imaging applications, see for instance [151].

3.2.3.7 Mixed Regularizations

Starting from a collection of convex functions $\{J_\ell\}_{\ell\in\mathscr{L}}$, $\mathscr{L} = \{1,\dots,L\}$, it is possible to design a convex function as

$$J_{\ell(x)} = \sum_{\ell\in\mathscr{L}} \rho_\ell J_\ell(x),$$

where $\rho_\ell > 0$ are weights. If each J_ℓ is partly smooth at x relative to a manifold $\mathscr{M}_x^\ell$, then it is shown in [145, Corollary 4.8] that J is also partly smooth at x for

$$\mathscr{M}_x = \bigcap_{\ell\in\mathscr{L}} \mathscr{M}_x^\ell \ ,$$

with the proviso that the manifolds $\mathscr{M}_x^\ell$ intersect transversally [143, Theorem 6.30(b)], i.e. the sum of their respective tangent spaces $\mathscr{T}_{\mathscr{M}_x^\ell}(x)$ spans the whole ambient space $\mathbb{R}^N$.

Literature review. A popular example is to impose both sparsity and low rank of a matrix, when using $J_1 = \|\cdot\|_1$ and $J_2 = \|\cdot\|_*$, see for instance [114, 176].

3.2.3.8 Separable Regularization

Let $\{J_\ell\}_{\ell\in\mathscr{L}}$, $\mathscr{L} = \{1,\dots,L\}$, be a family of convex functions. If J_ℓ is partly smooth at x_ℓ relative to a manifold $\mathscr{M}_{x_\ell}^\ell$, then the separable function

$$J\left(\{x_\ell\}_{\ell\in\mathscr{L}}\right) = \sum_{\ell\in\mathscr{L}} J_\ell(x_\ell)$$

is partly smooth at $(x_1,\dots,x_L)$ relative to $\mathscr{M}_{x_1}^1 \times \cdots \times \mathscr{M}_{x_L}^L$ [145, Proposition 4.5].

Literature review. One fundamental problem that has attracted a lot of interest in the recent years in data processing involves decomposing an observed object into a linear combination of components/constituents x_ℓ, $\ell \in \mathscr{L} = \{1,\dots,L\}$. One instance of such a problem is image decomposition into texture and piecewise-smooth (cartoon) parts. The corresponding forward model can be cast in the form (3.1), where $x_0 = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}$, x_1 and x_2 are the texture and cartoon components, and $\Phi = [\mathrm{Id} \quad \mathrm{Id}]$. The decomposition is then achieved by solving the variational problem ($\mathscr{P}_{y,\lambda}$), where J_1 is designed to promote the discontinuities in the image and J_2 to favor textures; see [6, 181, 204] and references therein. Another example of decomposition is principal component pursuit, proposed in [47], to decompose a matrix which is the superposition of a low-rank component and a sparse component. In this case $J_1 = \|\cdot\|_1$ and $J_2 = \|\cdot\|_*$.

3.3 ℓ^2 Stability

In this section, we assume that J is a finite-valued convex function, but it is not assumed to be partly smooth.

The observations y are in general contaminated by noise, as described by the forward model (3.1). It is thus important to study the ability of $(\mathscr{P}_{y,\lambda})$ to recover x_0 to a good approximation in presence of such a noise w and to assess how the reconstruction error decays as a function of the noise level. In this section, we present a generic result ensuring a so-called linear convergence rate in terms of ℓ^2-error between a recovered vector and x_0 (see Theorem 1), which encompasses a large body of literature from the inverse problems community.

3.3.1 Dual Certificates

It is intuitively expected that if $(\mathscr{P}_{y,\lambda})$ is good at recovering an approximation of x_0 in presence of noise, then $(\mathscr{P}_{y,0})$ should be able to identify x_0 uniquely when the noise vanishes, i.e., $y = \Phi x_0$. For this to happen, the solution to $(\mathscr{P}_{y,0})$ has to satisfy some nondegeneracy condition. To formalize this, we first introduce the notion of dual certificate.

Definition 3 (Dual certificates). For any vector $x \in \mathbb{R}^N$, the set of *dual certificates* at x is defined as

$$\mathscr{D}(x) = \mathrm{Im}(\Phi^*) \cap \partial J(x) \ .$$

The terminology "dual certificate" was introduced in [38]. One can show that the image by Φ^* of the set of solutions of the Fenchel-Rockafellar dual to $(\mathscr{P}_{y,0})$ is precisely $\mathscr{D}(x)$.

It is also worth noting that x_0 being a solution of $(\mathscr{P}_{y,0})$ for $y = \Phi x_0$ is equivalent to $\mathscr{D}(x_0) \neq \emptyset$. Indeed, this is simply a convenient rewriting of the first-order optimality condition for $(\mathscr{P}_{y,0})$.

To ensure stability of the set of minimizers $(\mathscr{P}_{y,\lambda})$ to noise perturbing the observations Φx_0, one needs to introduce the additional requirement that the dual certificates should be strictly inside the subdifferential of J at x_0. This is precisely the nondegeneracy condition mentioned previously.

Definition 4 (Nondegenerate dual certificates). For any vector $x \in \mathbb{R}^N$, we define the set of *nondegenerate dual certificates* of x

$$\tilde{\mathscr{D}}(x) = \mathrm{Im}(\Phi^*) \cap \mathrm{ri}(\partial J(x)) \ .$$

3.3.2 Stability in ℓ^2 Norm

The following theorem, proved in [101], establishes a linear convergence rate valid for any regularizer J, without any particular assumption beside being a proper closed convex function. In particular, it does not assume partial smoothness of J. This generic result encompasses many previous works, as discussed in Section 3.3.3.

Theorem 1. *Assume that*

$$\mathrm{Ker}(\Phi) \cap T_{x_0} = \{0\} \quad \text{and} \quad \tilde{\mathscr{D}}(x_0) \neq \emptyset \tag{3.9}$$

and consider the choice $\lambda = c\,\|w\|$, for some $c > 0$. Then we have for all minimizers $x^\star$ of $(\mathscr{P}_{y,\lambda})$

$$\|x^\star - x_0\|_2 \leqslant C\,\|w\| \ , \tag{3.10}$$

where $C > 0$ is a constant (see Remark 4 for details).

In plain words, this bound tells us that the distance of x_0 to the set of minimizers of $(\mathscr{P}_{y,\lambda})$ is within a factor of the noise level, which justifies the terminology "linear convergence rate."

Remark 2 (The role of nonsmoothness). The injectivity of Φ when restricted to T_{x_0} is intimately related to the fact that J is nonsmooth at x_0. The higher the degree of nonsmoothness, the lower the dimension of the subspace T_{x_0}, and hence the more likely the restricted injectivity. If J is smooth around x_0 (e.g., quadratic regularizers), however, the restricted injectivity condition cannot be fulfilled, unless Φ is itself injective. The reason is that T_{x_0} is the whole $\mathbb{R}^N$ at the smoothness points. For smooth regularizations, it can be shown that the convergence rate is slower than linear, we refer to [196] for more details.

Remark 3 (Uniqueness). One can show that condition (3.9) implies that x_0 is the unique solution of $(\mathscr{P}_{y,0})$ for $y = \Phi x_0$. This condition however does not imply in general that $(\mathscr{P}_{y,\lambda})$ has a unique minimizer for $\lambda > 0$.

Remark 4 (Stability constant). Result (3.10) ensures that the mapping $y \mapsto x^\star$ (that might be set valued) is C-Lipschitz-continuous at $y = \Phi x_0$. Condition $\tilde{\mathscr{D}}(x_0) \neq \emptyset$ is equivalent to the existence of some $\eta \in \tilde{\mathscr{D}}(x_0)$. The value of C (in fact an upper bound) can be found in [101]. It depends on Φ, T_{x_0}, c and the chosen nondegenerate dual certificate η. In particular, the constant degrades critically as η gets closer to the relative boundary of $\tilde{\mathscr{D}}(x_0)$, which reflects the intuition of how far is η from being a nondegenerate certificate.

Remark 5 (Source condition). The condition $\mathscr{D}(x_0) \neq \emptyset$ is often called "source condition" or "range condition" in the literature of inverse problems. We refer to the monograph [196] for a general overview of this condition and its implications. It is an abstract condition, which is not easy to check in practice, since exhibiting a

valid nondegenerate certificate is not trivial. We give in Section 3.4.1 further insights about this in the context of compressed sensing. Section 3.4.1 describes a particular construction of a good candidate (the so-called linearized pre-certificate) for being such an $\eta \in \tilde{\mathscr{D}}(x_0)$, and it is shown to govern stability of the manifold $\mathscr{M}_{x_0}$ for partly smooth regularizers.

Remark 6 (Infinite dimension). It is important to remind that, in its full general form, Theorem 1 only holds in finite dimension. The constant C indeed may depend on the ambient dimension N, in which case the constant can blow up as the discretization grid of the underlying continuous problem is made finer (i.e., as N grows). We detail below some relevant literature where similar results are shown in infinite dimension.

3.3.3 Related Works

3.3.3.1 Convergence Rates

For quadratic regularizations of the form $J = \|D^*\cdot\|^2$ for some linear operator D^*, the ℓ^2-error decay can be proved to be $O(\sqrt{\|w\|})$, which is not linear, see [196, Chapter 3] for more details and extensions to infinite-dimensional Hilbert spaces. For nonsmooth priors, in [30], the authors show the Bregman distance between $x^\star$ and x_0 exhibits a linear convergence rate for both the Lagrangian ($\mathscr{P}_{y,\lambda}$) and the constrained ($\mathscr{P}^1_{y,\varepsilon}$) problems under the source condition $\mathscr{D}(x_0) \neq 0$. These results hold more generally over infinite-dimensional Banach spaces. They have been subsequently generalized to ill-posed nonlinear inverse problems by [190] and [133]. It is important to observe that in order to prove convergence rates in terms of ℓ^2-error, as done in (3.10), it is necessary to strengthen the source condition to its nondegenerate version, i.e., $\tilde{\mathscr{D}}(x_0) \neq 0$.

In [153], the authors consider the case where J is a ℓ^p norm with $1 \leqslant p \leqslant 2$ and establish convergence rates of $\|\Phi x_0 - \Phi x^\star\|$ in $O(\|w\|)$ and of $\|x^\star - x_0\|$ in $O(\sqrt{\|w\|})$. [117] prove Theorem 1 for $J = \|\cdot\|_1$. They show that the nondegeneracy condition is also necessary for linear convergence and draw some connections with the restricted isometry property (RIP), see below. Under a condition that bears similarities with (3.9), linear convergence with respect to J, i.e., $J(x^\star - x_0) = O(\|w\|)$, is proved in [116] for positively homogeneous regularizers. This result is equivalent to Theorem 1 but only when J is coercive, which precludes many important regularizers, such as for instance analysis-type regularizers including total variation.

3.3.3.2 RIP-based Compressed Sensing

The recovery performance of compressed sensing (i.e., when Φ is drawn from suitable random ensembles) for $J = \|\cdot\|_1$ has been widely analyzed under the so-called restricted isometry property (RIP) introduced in [41, 43, 44]. For any integer $k \geqslant 0$, the kth order restricted isometry constant of a matrix Φ is defined as the smallest $\delta_k \geqslant 0$ such that

$$(1 - \delta_k) \|x\|^2 \leqslant \|\Phi x\|^2 \leqslant (1 + \delta_k) \|x\|^2 ,$$

for all vectors x such that $\|x\|_0 \leqslant k$. It is shown [43] that if $\delta_{2k} + \delta_{3k} < 1$, then for every vector x_0 with $\|x_0\|_0 \leqslant k$, there exists a nondegenerate certificate [40, Lemma 2.2], see also the discussion in [117]. In turn, this implies linear convergence rate and is applied in [44] to show ℓ^2-stability to noise of compressed sensing. This was generalized in [46] to analysis sparsity $J = \|D^* \cdot\|_1$, where D is assumed to be a tight frame, structured sparsity in [46], and matrix completion in [37, 188] using $J = \|\cdot\|_*$. The goal is then to design RIP matrices Φ with constants such that $\delta_{2k} + \delta_{3k}$ (or a related quantity) is small enough. This is possible if Φ is drawn from an appropriate random ensemble for some (hopefully optimal) scaling of (N, P, k). For instance, if Φ is drawn from the standard Gaussian ensemble (i.e., with i.i.d. zero-mean standard Gaussian entries), there exists a constant C such that the RIP constants of $\Phi/\sqrt{P}$ obey $\delta_{2k} + \delta_{3k} < 1$ with overwhelming probability provided that

$$P \geqslant Ck \log(N/k) , \tag{3.11}$$

see for instance [41]. This result remains true when the entries of Φ are drawn independently from a subgaussian distribution. When Φ is a structured random matrix, e.g., random partial Fourier matrix, the RIP constants of $\Phi/\sqrt{P}$ can also satisfy the desired bound, but at the expense of polylog terms in the scaling (3.11), see [105] for a comprehensive treatment. Note that in general, computing the RIP constants for a given matrix is an NP-hard problem [10, 219].

3.3.3.3 RIP-less Compressed Sensing

RIP-based guarantees are uniform, in the sense that the recovery holds with high probability for *all* sparse signals. There is a recent wave of work in RIP-less analysis of the recovery guarantees for compressed sensing. The claims are nonuniform, meaning that they hold for a fixed signal with high probability on the random matrix Φ. This line of approaches improves on RIP-based bounds providing typically sharper constants. When Φ is drawn from the Gaussian ensemble, it is proved in [193] for $J = \|\cdot\|_1$ that if the number of measurements P obeys $P \geqslant Ck \log(N/k)$ for some constant $C > 0$, where $k = \|x_0\|_0$, then condition (3.9) holds with high probability on Φ. This result is based on Gordon's comparison principle for

Gaussian processes and depends on a summary parameter for convex cones called the Gaussian width. Equivalent lower bounds on the number of measurements for matrix completion from random measurements by minimizing the nuclear norm were provided in [42] to ensure that (3.9) holds with high probability. This was used to prove ℓ^2-stable matrix completion in [35].

The authors in [54] have recently showed that the Gaussian width-based approach leads to sharp lower bounds on P required to solve regularized inverse problems from Gaussian random measurements. For instance, they showed for $J = \|\cdot\|_1$ that

$$P > 2k \log(N/k) \tag{3.12}$$

guarantees exact recovery from noiseless measurements by solving ($\mathscr{P}_{y,0}$). An overhead in the number of measurements is necessary to get linear convergence of the ℓ^2-error in presence of noise by solving ($\mathscr{P}^1_{y,\varepsilon}$) with $\varepsilon = \|w\|$, i.e., x_0 is feasible. Their results handle for instance the case of group sparsity (3.6) and the nuclear norm (3.8). In the polyhedral case, it can be shown that (3.12) implies the existence of a non-degenerate dual certificate, i.e., (3.9), with overwhelming probability. The Gaussian width is closely related to another geometric quantity called the statistical dimension in conic integral geometry. The statistical dimension canonically extends the linear dimension to convex cones, and has been proposed in [4] to deliver reliable predictions about the quantitative aspects of the phase transition for exact noiseless recovery from Gaussian measurements.

To deal with non-Gaussian matrix measurements (such as for instance partial Fourier matrices), [123] introduced the "golfing scheme" for noiseless low-rank matrix recovery guarantees using $J = \|\cdot\|_*$. The golfing scheme is an iterative procedure to construct an (approximate) nondegenerate certificate. This construction is also studied in [36] for noiseless and noisy sparse recovery with $J = \|\cdot\|_1$. In another chapter of this volume [220], the author develops a technique, called the "bowling scheme," which is able to deliver bounds on the number of measurements that are similar to the Gaussian width-based bounds for standard Gaussian measurements, but the argument applies to a much wider class of measurement ensembles.

3.4 Model Stability

In the remainder of this chapter, we assume that J is finite-valued convex and locally partly smooth around x_0, as defined in Section 3.2.2. This means in particular that the prior J promotes locally solution which belongs to the manifold $\mathscr{M} = \mathscr{M}_{x_0}$. In the previous section, we were only concerned with ℓ^2-stability guarantees and partial smoothness was not necessary then. Owing to the additional structure conveyed by partial smoothness, we will be able to provide guarantees on the identification of the correct $\mathscr{M} = \mathscr{M}_{x_0}$ by solving ($\mathscr{P}_{y,\lambda}$), i.e., whether the (unique) solution $x^\star$ of ($\mathscr{P}_{y,\lambda}$) satisfies $x^\star \in \mathscr{M}$. Such guarantees are of paramount importance for many applications. For instance, consider the case where ℓ^1 regularization is used

to localize some (sparse) sources. Then $x^\star \in \mathscr{M}$ means that one perfectly identifies the correct source locations. Another example is that of the nuclear norm for low-rank matrix recovery. The correct model identification implies that $x^\star$ has the correct rank, and consequently that the eigenspaces of $x^\star$ have the correct dimensions and are close to those of x_0.

3.4.1 Linearized Pre-certificate

We saw in Section 3.3.2 that ℓ^2-stability of the solutions to $(\mathscr{P}_{y,\lambda})$ is governed by the existence of a nondegenerate dual certificate $p \in \tilde{\mathscr{D}}(x_0)$. It turns out that not all dual certificates are equally good for stable model identification, and toward the latter, one actually needs to focus on a particular dual certificate, which we call "minimal norm" certificate.

Definition 5 (Minimal norm certificate). Assume that x_0 is a solution of $(\mathscr{P}_{y,0})$. We define the "minimal-norm certificate" as

$$\eta_0 = \Phi^* \operatorname*{argmin}_{\Phi^* p \in \partial J(x_0)} \|p\| \ . \tag{3.13}$$

A remarkable property, stated in Proposition 1 below, is that, as long as one is concerned with checking whether η_0 is nondegenerate, i.e., $\eta_0 \in \mathrm{ri}(\partial J(x_0))$, one can instead use the vector η_F defined below, which can be computed in closed form.

Definition 6 (Linearized pre-certificate). Assume that

$$\mathrm{Ker}(\Phi) \cap T_{x_0} = \{0\}. \tag{3.14}$$

We define the "linearized pre-certificate" as

$$\eta_F = \Phi^* \operatorname*{argmin}_{\Phi^* p \in \mathrm{aff}(\partial J(x_0))} \|p\| \, . \tag{3.15}$$

Remark 7 (Well-posedness of the definitions). Note that the hypothesis that x_0 is a solution of $(\mathscr{P}_{y,0})$ is equivalent to saying that $\mathscr{D}(x_0)$ is a nonempty convex compact set. Hence in (3.13), the optimal p is the orthogonal projection of 0 on a nonempty closed convex set, and thus η_0 is uniquely defined. Similarly, the hypothesis (3.14) implies that the constraint set involved in (3.15) is a nonempty affine space, and thus η_F is also uniquely defined.

Remark 8 (Certificate vs. pre-certificate). Note that the only difference between (3.13) and (3.15) is that the convex constraint set $\partial J(x_0)$ is replaced by a simpler affine constraint. This means that η_F does not always qualify as a valid certificate, i.e., $\eta_F \in \partial J(x_0)$, hence the terminology "pre-certificate" is used. This condition is actually at the heart of the model identification result exposed in Theorem 2.

From now on, let us remark that η_F is actually simple to compute, since it amounts to solving a linear system in the least-squares sense.

Proposition 1. *Under condition* (3.14)*, one has*

$$\eta_F = \Phi^* \Phi_{T_{x_0}}^{+,*} e_{x_0} \quad \text{where} \quad e_{x_0} = \mathrm{P}_{T_{x_0}}(\partial J(x_0)) \in \mathbb{R}^N. \tag{3.16}$$

Remark 9 (Computating e_x). The vector e_x appearing in (3.16) can be computed in closed form for most of the regularizers discussed in Section 3.2.2. For instance, for $J = \|\cdot\|_1$, $e_x = \mathrm{sign}(x)$. For $J = \|\cdot\|_{1,\mathscr{B}}$, it reads $e_x = (e_b)_{b\in\mathscr{B}}$, where $e_b = x_b / \|x_b\|$ if $x_b \neq 0$, and $e_b = 0$ otherwise. For $J = \|\cdot\|_*$ and a SVD decomposition $x = U_x \,\mathrm{diag}(\Lambda_x) V_x^*$, one has $e_x = U_x V_x^*$.

The following proposition, whose proof can be found in [232], exhibits a precise relationship between η_0 and η_F. In particular, it implies that η_F can be used in place of η_0 to check whether η_0 is nondegenerate, i.e., $\eta_0 \in \mathrm{ri}(\partial J(x_0))$.

Proposition 2. *Under condition* (3.14)*, one has*

$$\begin{aligned} \eta_F \in \mathrm{ri}(\partial J(x_0)) \quad &\Longrightarrow \quad \eta_F = \eta_0, \\ \eta_0 \in \mathrm{ri}(\partial J(x_0)) \quad &\Longrightarrow \quad \eta_F = \eta_0. \end{aligned}$$

3.4.2 Model Identification

The following theorem provides a sharp sufficient condition to establish model selection. It is proved in [232]. It encompasses as special cases many previous works in the signal processing, statistics, and machine learning literatures, as we discuss in Section 3.4.5.1.

Theorem 2. *Let J be locally partly smooth at x_0 relative to $\mathscr{M} = \mathscr{M}_{x_0}$. Assume that*

$$\mathrm{Ker}(\Phi) \cap T_{x_0} = \{0\} \quad \text{and} \quad \eta_F \in \mathrm{ri}(\partial J(x_0)). \tag{3.17}$$

Then there exists C such that if

$$\max(\lambda, \|w\| / \lambda) \leqslant C, \tag{3.18}$$

the solution $x^\star$ of ($\mathscr{P}_{y,\lambda}$) from measurements (3.1) *is unique and satisfies*

$$x^\star \in \mathscr{M} \quad \text{and} \quad \|x_0 - x^\star\| = O(\max(\lambda, \|w\|)). \tag{3.19}$$

Remark 10 (Linear convergence rate vs. model identification). Obviously, assumptions (3.17) of Theorem 2 imply those of Theorem 1. They are of course stronger,

but imply a stronger result, since uniqueness of $x^\star$ and model identification (i.e., $x^\star \in \mathscr{M}$) are not guaranteed by Theorem 1 (which does not even need J to be partly smooth). A chief advantage of Theorem 2 is that its hypotheses can be easily checked and analyzed for a particular operator Φ. Indeed, computing η_F only requires solving a linear system, as clearly seen from formula (3.16).

Remark 11 (Minimal signal-to-noise ratio). Another important distinction between Theorems 1 and 2 is the second assumption (3.18). In plain words, it requires that the noise level is small enough and that the regularization parameter is wisely chosen. Such an assumption is not needed in Theorem 2 to ensure linear convergence of the ℓ^2-error. In fact, this condition is quite natural. To see this, consider for instance the case of sparse recovery where $J = \|\cdot\|_1$. If the minimal signal-to-noise ratio is low, the noise will clearly dominate the amplitude of the smallest entries, so that one cannot hope to recover the exact support, but it is still possible to achieve a low ℓ^2-error by forcing those small entries to zero.

Remark 12 (Identification of the manifold). For all the regularizations considered in Section 3.2.3, the conclusion of Theorem 2 is even stronger as it guarantees that $\mathscr{M}_{x^\star} = \mathscr{M}$. The reason is that for any x and nearby points x' with $x' \in \mathscr{M}_x$, one has $\mathscr{M}_{x'} = \mathscr{M}_x$.

Remark 13 (General loss/data fidelity). It is possible to extend Theorem 2 to account for general loss/data fidelity terms beyond the quadratic one, i.e., $\frac{1}{2}\|y - \Phi x\|^2$. More precisely, this result holds true for loss functions of the form $F(\Phi x, y)$, where $F : \mathbb{R}^P \times \mathbb{R}^P \to \mathbb{R}$ is a C^2 strictly convex function in its first argument, ∇F is C^1 in the second argument, with $\nabla F(y, y) = 0$, where ∇F is the gradient with respect to the first variable. In this case, expression (3.16) of η_F becomes simply

$$\eta_F = \Gamma(\mathrm{P}_T\, \Gamma\, \mathrm{P}_T)^+ e_{x_0} \quad \text{where} \quad \begin{cases} T = T_{x_0} \\ \Gamma = \Phi^* \partial^2 F(\Phi x_0, \Phi x_0) \Phi \end{cases},$$

and where $\partial^2 F$ is the Hessian with respect to the first variable (which is a positive definite operator). We refer to [232] for more details.

3.4.3 Sharpness of the Model Identification Criterion

The following proposition, proved in [232], shows that Theorem 2 is in some sense sharp, since the hypothesis $\eta_F \in \mathrm{ri}(\partial J(x_0))$ (almost) characterizes the stability of $\mathscr{M}$.

Proposition 3. *We suppose that x_0 is the unique solution of $(\mathscr{P}_{y,0})$ for $y = \Phi x_0$ and that*

$$\mathrm{Ker}(\Phi) \cap T_{x_0} = \{0\}, \quad \text{and} \quad \eta_F \notin \partial J(x_0). \tag{3.20}$$

Then there exists $C > 0$ such that if (3.18) holds, then any solution $x^\star$ of $(\mathscr{P}_{y,\lambda})$ for $\lambda > 0$ obeys $x^\star \notin \mathscr{M}$.

In the particular case where $w = 0$ (no noise), this result shows that the manifold $\mathscr{M}$ is not correctly identified when solving $(\mathscr{P}_{y,\lambda})$ for $y = \Phi x_0$ and for any $\lambda > 0$ small enough.

Remark 14 (Critical case). The only case not covered by neither Theorem 2 nor Proposition 3 is when $\eta_F \in \mathrm{rbd}(\partial J(x_0))$, where rbd stands for the boundary relative to the affine hull. In this case, one cannot conclude, since depending on the noise w, one can have either stability or non-stability of $\mathscr{M}$. We refer to [229] where an example illustrates this situation for the 1-D total variation $J = \left\|D^*_{\mathrm{DIF}} \cdot\right\|_1$, where D^*_{DIF} is a finite-difference discretization of the 1-D derivative operator.

3.4.4 Probabilistic Model Consistency

Theorem 2 assumes a deterministic noise w, and the operator Φ is fixed. For applications in statistics and machine learning, it makes sense to rather assume a random model for both Φ and w. The natural question is then to assert that the estimator defined by solving $(\mathscr{P}_{y,\lambda})$ is consistent in the sense that it correctly estimates x_0 and possibly the model $\mathscr{M}_{x_0}$ as the number of observations $P \to +\infty$. This requires to handle operators Φ with an increasing number of rows, and thus to also assess sensitivity of the optimization problem $(\mathscr{P}_{y,\lambda})$ to perturbations of Φ (and not only to (w, λ) as done previously).

To be more concrete, in this section, we work under the classical setting where N an x_0 are fixed as the number of observations $P \to +\infty$. The data (φ_i, w_i) are assumed to be random vectors in $\mathbb{R}^N \times \mathbb{R}$, where φ_i is the ith row of Φ for $i = 1, \ldots, P$. These vectors are supposed independent and identically distributed (i.i.d.) samples from a joint probability distribution such that $\mathbb{E}(w_i|\varphi_i) = 0$, finite fourth-order moments, i.e., $\mathbb{E}\left(w_i^4\right) < +\infty$ and $\mathbb{E}\left(\|\varphi_i\|^4\right) < +\infty$. Note that in general, w_i and φ_i are not necessarily independent. It is possible to consider other distribution models by weakening some of the assumptions and strengthening others, see, e.g., [7, 142, 243]. Let us denote $\Gamma = \mathbb{E}(\varphi_i^* \phi_i) \in \mathbb{R}^{N \times N}$, where ϕ_i is any row of Φ. We do not make any assumption on the invertibility of Γ.

In this setting, a natural extension of η_F defined by (3.16) in the deterministic case is

$$\tilde{\eta}_F = \Gamma \Gamma^+_{T_{x_0}} e_{x_0}$$

where $\Gamma_{T_{x_0}} = \mathrm{P}_{T_{x_0}} \Gamma \mathrm{P}_{T_{x_0}}$, and we use the fact that $\Gamma_{T_{x_0}}$ is symmetric and $\mathrm{Im}(\Gamma^{+}_{T_{x_0}}) \subset T_{x_0}$. It is also implicitly assumed that $\mathrm{Ker}(\Gamma) \cap T_{x_0} = \{0\}$ which is the equivalent adaptation of the restricted injectivity condition in (3.17) to this setting.

To make the discussion clearer, the parameters $(\lambda = \lambda_P, \Phi = \Phi_P, w = w_P)$ are now indexed by P. The estimator $x^\star_P$ obtained by solving $(\mathscr{P}_{\lambda_P, y_P})$ for $y_P = \Phi_P x_0 + w_P$ is said to be consistent for x_0 if

$$\lim_{P \to +\infty} \Pr\left(x^\star_P \text{ is unique}\right) = 1$$

and $x^\star_P \to x_0$ in probability. The estimator is said to be model consistent if

$$\lim_{P \to +\infty} \Pr\left(x^\star_P \in \mathscr{M}\right) = 1,$$

where $\mathscr{M} = \mathscr{M}_{x_0}$ is the manifold associated to x_0.

The following result, whose proof can be found in [232], guarantees model consistency for an appropriate scaling of μ_P. It generalizes several previous works in the statistical and machine learning literature as we review in Section 3.4.5.1.

Theorem 3. *If*

$$\mathrm{Ker}(\Gamma) \cap T_{x_0} = \{0\} \quad \text{and} \quad \tilde{\eta}_F \in \mathrm{ri}(\partial J(x_0)), \tag{3.21}$$

and

$$\lambda_P = o(P) \quad \text{and} \quad \lambda_P^{-1} = o(P^{-1/2}), \tag{3.22}$$

then the estimator $x^\star_P$ of x_0 is model consistent.

3.4.5 Related Works

3.4.5.1 Model Consistency

Theorem 2 is a generalization of a large body of results in the literature. For the Lasso, i.e. $J = \|\cdot\|_1$, to the best of our knowledge, this result was initially stated in [107]. In this setting, result (3.19) corresponds to the correct identification of the support, i.e., $\mathrm{supp}(x^\star) = \mathrm{supp}(x_0)$. Condition (3.21) for $J = \|\cdot\|_1$ is known in the statistics literature under the name "irrepresentable condition" (generally stated in a nongeometrical form), see, e.g., [243]. [142] have shown estimation consistency for Lasso for fixed N and x_0 and asymptotic normality of the estimates. The authors in [243] prove Theorem 3 for $J = \|\cdot\|_1$, though under slightly different assumptions on the covariance and noise distribution. A similar result is established in [140] for the elastic net, i.e., $J = \|\cdot\|_1 + \rho\|\cdot\|_2^2$ for $\rho > 0$. In [7] and [8], the author

proves Theorem 3 for two special cases, namely the group Lasso and nuclear norm minimization. Note that these previous works assume that the asymptotic covariance Γ is invertible. We do not impose such an assumption and only require the weaker restricted injectivity condition $\mathrm{Ker}(\Gamma) \cap T = \{0\}$. In a previous work [229], we have proved an instance of Theorem 2 when $J(x) = \|D^* x\|_1$, where $D \in \mathbb{R}^{N \times Q}$ is an arbitrary linear operator. This covers as special cases the discrete anisotropic total variation or the fused Lasso. This result was further generalized in [228] when J belongs to the class of partly smooth functions relative to linear manifolds $\mathscr{M}$, i.e., $\mathscr{M} = T_x$. Typical instances encompassed in this class are the $\ell^1 - \ell^2$ norm, or its analysis version, as well as polyhedral gauges including the ℓ^∞ norms. Note that the nuclear norm (and composition of it with linear operators as proposed for instance in [118, 191]), whose manifold is not linear, does not fit into the framework of [228], while it is covered by Theorem 2. Lastly, a similar result is proved in [89] for a continuous (infinite-dimensional) sparse recovery problem over the space of Radon measures normed by the total variation of a measure J (not to be confused with the total variation of functions). In this continuous setting, an interesting finding is that, when $\eta_0 \in \mathrm{ri}(\partial J(x_0))$, η_0 is not equal to η_F but to a different certificate (called "vanishing derivative" certificate in [89]) that can also be computed by solving a linear system.

3.4.5.2 Stronger Criteria for ℓ^1

Many sufficient conditions have been proposed in the literature to ensure that η_F is a nondegenerate certificate, and hence to guarantee stable identification of the support (i.e., model). We illustrate this here for $J = \|\cdot\|_1$, but similar reasoning can be carried out for $\|\cdot\|_{1,\mathscr{B}}$ or $\|\cdot\|_*$.

The strongest criterion makes use of mutual coherence, first considered in [78]

$$\mu(\Phi) = \max_{i \neq j} |\langle \varphi_i, \varphi_j \rangle|$$

where each column φ_i of Φ is assumed normalized to a unit ℓ^2 norm. Mutual coherence measures the degree of ill conditioning of Φ through the correlation of its columns $(\varphi_i)_{1 \leqslant i \leqslant N}$. Mutual coherence is always lower bounded by $\sqrt{\frac{N-P}{P(N-1)}}$, and equality holds if and only if $(\varphi_i)_{1 \leqslant i \leqslant N}$ is an equiangular tight frame, see [208]. Finer variants based on cumulative coherences have been proposed in [24, 120]. To take into account the influence of the support $I = \mathrm{supp}(x_0)$ of the vector x_0 to recover, Tropp introduced in [221] the Exact Recovery Condition (ERC), defined as

$$\mathrm{ERC}(I) = \left\| \Phi_{I^c}^* \Phi_I^{+,*} \right\|_{\infty,\infty} = \max_{j \notin I} \left\| \Phi_I^+ \varphi_j \right\|_1$$

where $\|\cdot\|_{\infty,\infty}$ is the matrix operator norm induced by the ℓ^∞ vector norm, $\Phi_I = (\varphi_i)_{i \in I}$, and I^c is the complement of the set I. Φ_I is assumed injective which, in view

of Section 3.2.3.1, is nothing but a specialization to ℓ^1 of the restricted injectivity condition in (3.17). A weak ERC criterion, which does not involve matrix inversion, is derived in [83]

$$\mathrm{wERC}(I) = \frac{\max_{j \in I^c} \sum_{i \in I} |\langle \varphi_i, \varphi_j \rangle|}{1 - \max_{j \in I} \sum_{i \neq j \in I} |\langle \varphi_i, \varphi_j \rangle|}.$$

Given the structure of the subdifferential of the ℓ^1 norm, it is easy to check that

$$\eta_F \in \mathrm{ri}(\partial J(x_0)) \quad \Longleftrightarrow \quad \mathrm{IC}(x_0) = \left\| \Phi_{I^c}^* \Phi_I^{+,*} \operatorname{sign}(x_{0,I}) \right\|_\infty < 1.$$

The right-hand side in the equivalence is precisely what is called the irrepresentable condition in statistics and machine learning. Clearly, $\mathrm{IC}(x_0)$ involves both the sign vector and the support of x_0. The following proposition gives ordered upper bounds of $\mathrm{IC}(x_0)$ in terms of the cruder criteria ERC, wERC, and mutual coherence. A more elaborate discussion of them can be found in [159].

Proposition 4. *Assume that Φ_I is injective and denote $k = |I| = \|x_0\|_0$. Then,*

$$\mathrm{IC}(x_0) \leqslant \mathrm{ERC}(I) \leqslant \mathrm{wERC}(I) \leqslant \frac{k\mu(\Phi)}{1 - (k-1)\mu(\Phi)}.$$

3.4.5.3 Linearized Pre-certificate for Compressed Sensing Recovery

Stable support identification has been established in [84, 239] for the Lasso problem when Φ is drawn from the Gaussian ensemble. These works show that for $k = \|x_0\|_0$, if

$$P > 2k \log(N)$$

then indeed $\eta_F \in \mathrm{ri}(\partial J(x_0))$, and this scaling can be shown to be sharp. This scaling should be compared with (3.12) ensuring that there exists a nondegenerate certificate. The gap in the log term indicates that there exists vectors that can be stably recovered by ℓ^1 minimization in ℓ^2-error sense, but whose support cannot be stably identified. Equivalently, for these vectors, there exists a nondegenerate certificate but it is not η_F.

The pre-certificate η_F is also used to ensure exact recovery of a low-rank matrix from incomplete noiseless measurements by minimizing the nuclear norm [38, 42]. This idea is further generalized by [39] for a family of decomposable norms (including in particular ℓ^1-ℓ^2 norm and the nuclear norm), which turns to be a subset of partly smooth regularizers. In these works, lower bounds on the number of random measurements needed for η_F to be a nondegenerate certificate are developed. In fact, these measurement lower bounds combined with Theorem 2

allow us to conclude that matrix completion by solving ($\mathscr{P}_{y,\lambda}$) with $J = \|\cdot\|_*$ identifies the correct rank at high signal-to-noise levels.

3.4.5.4 Sensitivity Analysis

Sensitivity analysis is a central theme in variational analysis. Comprehensive monographs on the subject are [23, 165]. The function to be analyzed underlying problems ($\mathscr{P}_{y,\lambda}$) and ($\mathscr{P}_{y,0}$) is

$$f(x,\theta) = \begin{cases} \frac{1}{2\lambda}\|y-\Phi x\|^2 + J(x) & \text{if } \lambda > 0 \\ \iota_{\mathscr{H}_y}(x) + J(x) & \text{if } \lambda = 0 \end{cases}, \tag{3.23}$$

where $\mathscr{H}_y = \{y\,;\, \Phi x = y\}$ and where the parameters are $\theta = (\lambda, y, \Phi)$ for $\lambda \geqslant 0$. Theorems 2 and 3 can be understood as a sensitivity analysis of the minimizers of f at a point $(x = x_0, \theta = \theta_0 = (0, \Phi x_0, \Phi))$.

Classical sensitivity analysis of nonsmooth optimization problems seeks conditions to ensure smoothness of the mapping $\theta \mapsto x_\theta$ where x_θ is a minimizer of $f(\cdot,\theta)$, see for instance [23, 192]. This is usually guaranteed by the nondegenerate source condition and restricted injectivity condition (3.9), which, as already exposed in Section 3.3.2, ensure linear convergence rate, and hence Lipschitz behavior of this mapping. The analysis proposed by Theorem 2 goes one step further, by assessing that $\mathscr{M}_{x_0}$ is a stable manifold (in the sense of [240]), since the minimizer x_θ is unique and remains in $\mathscr{M}_{x_0}$ for θ close to θ_0. Our starting point for establishing Theorem 2 is the inspiring work of Lewis [145] who first introduced the notion of partial smoothness and showed that this broad class of functions enjoys a powerful calculus and sensitivity theory. For convex functions (which is the setting considered in our work), partial smoothness is closely related to $\mathscr{U} - \mathscr{V}$-decompositions developed in [144]. In fact, the behavior of a partly smooth function and of its minimizers (or critical points) depend essentially on its restriction to the manifold, hence offering a powerful framework for sensitivity analysis theory. In particular, critical points of partly smooth functions move stably on the manifold as the function undergoes small perturbations [148]. An important and distinctive feature of Theorem 2 is that partial smoothness of J at x_0 relative to $\mathscr{M}$ transfers to $f(\cdot,\theta)$ for $\lambda > 0$, but not when $\lambda = 0$ in general. In particular, [145, Theorem 5.7] does not apply to prove our claim.

3.5 Sensitivity Analysis and Parameter Selection

In this section, we study local variations of the solutions of ($\mathscr{P}_{y,\lambda}$) considered as functions of the observations y. In a variational-analytic language, this corresponds to analyzing the sensitivity of the optimal values of ($\mathscr{P}_{y,\lambda}$) to small perturbations

of y seen as a parameter. This analysis will have important implications, and we exemplify one of them by constructing unbiased estimators of the quadratic risk, which in turn will allow us to have an objectively guided way to select the optimal value of the regularization parameter λ.

As argued in Section 3.4.5.4, assessing the recovery performance by solving ($\mathscr{P}_{y,\lambda}$) for w and λ small amounts to a sensitivity analysis of the minimizers of f in (3.23) at $(x = x_0, \theta = \theta_0 = (0, \Phi x_0, \Phi))$. This section involves again sensitivity analysis of (3.23) to perturbations of y but for $\lambda > 0$. Though we focus our attention on sensitivity to y, our arguments extend to any parameters, for instance λ or Φ.

Similarly to the previous section, we suppose here that J is a finite-valued convex and partly smooth function. For technical reasons, we furthermore assume that the partial smoothness manifold is linear, i.e., $\mathscr{M}_x = T_x$. We additionally suppose that the set of all possible models $\mathscr{T} = \{T_x\}_{x \in \mathbb{R}^N}$ is finite. All these assumptions hold true for the regularizers considered in Section 3.2.3, with the notable exception of the nuclear norm, whose manifolds of partial smoothness are nonlinear.

3.5.1 *Differentiability of Minimizers*

Let us denote $x^\star(y)$ a minimizer of ($\mathscr{P}_{y,\lambda}$) for a fixed value of $\lambda > 0$. Our main goal is to study differentiability of $x^\star(y)$ and find a closed-form formula of the derivative of $x^\star(y)$ with respect to the observations y. Since $x^\star(y)$ is not necessarily a unique minimizer, such a result means actually that we have to single out one solution $x^\star(y)$, which hopefully should be a locally smooth function of y. However, as J is non-smooth, one cannot hope for such a result to hold for any observation $y \in \mathbb{R}^P$. For applications to risk estimation (see Section 3.5.3), it is important to characterize precisely the smallest set $\mathscr{H}$ outside of which $x^\star(y)$ is indeed locally smooth. It turns out that one can actually write down an analytical expression of such a set $\mathscr{H}$, containing points where one cannot find locally a smooth parameterization of the minimizers. This motivates our definition of what we coin a "transition space."

Definition 7 (Transition space). We define the *transition space* $\mathscr{H}$ as

$$\mathscr{H} = \bigcup_{T \in \mathscr{T}} \mathrm{bd}(\mathscr{H}_T),$$

where $\mathrm{bd}(C)$ is the boundary of a set C, and

$$\mathscr{H}_T = \left\{ y \in \mathbb{R}^P \;;\; \exists x \in \tilde{T},\ \lambda^{-1} \Phi_T^*(\Phi x - y) \in \mathrm{rbd}(\partial J(x)) \right\},$$

where $\tilde{T} = \left\{ x \in \mathbb{R}^N \;;\; T_x = T \right\}$.

The set $\mathscr{H}$ contains the observations $y \in \mathbb{R}^P$ such that the model subspace $T_{\tilde{x}(y)}$ associated to a well-chosen solution $\tilde{x}(y)$ of ($\mathscr{P}_{y,\lambda}$) is not stable with respect to small perturbations of y. In particular, when $J = \|\cdot\|_1$, it can be checked that $\mathscr{H}$

is a finite union of hyperplanes and when $J = \|\cdot\|_{1,2}$ it is a semi-algebraic set (see Definition 8). This stability is not only crucial to prove smoothness of $\tilde{x}(y)$, it is also important to be able to write down an explicit formula for the derivative, as detailed in the following theorem whose proof is given in [226].

Theorem 4. *Let $y \notin \mathscr{H}$ and $x^\star$ a solution of $(\mathscr{P}_{y,\lambda})$ such that*

$$\operatorname{Ker} \Phi_T \cap \operatorname{Ker} \mathrm{D}^2 J_T(x^\star) = \{0\} \qquad (\mathscr{I}_{x^\star})$$

where $T = T_{x^\star}$. Then, there exists an open neighborhood $\mathscr{V} \subset \mathbb{R}^N$ of y, and a mapping $\tilde{x} : \mathscr{V} \to T$ such that

1. *for every $\bar{y} \in \mathscr{V}$, $\tilde{x}(\bar{y})$ is a solution of $(\mathscr{P}_{\lambda,\bar{y}})$, and $\tilde{x}(y) = x^\star$;*
2. *the mapping $\tilde{x}$ is $\mathrm{C}^1(\mathscr{V})$ and*

$$\forall\, \bar{y} \in \mathscr{V}, \quad \mathrm{D}\tilde{x}(\bar{y}) = (\Phi_T^* \Phi_T + \lambda \mathrm{D}^2 J_T(x^\star))^{-1} \Phi_T.$$

Here $\mathrm{D}^2 J_T$ is the Hessian (second order derivative) of J restricted to T. This Hessian is surely well defined owing to partial smoothness, see Definition 2(i).

3.5.2 Semi-algebraic Geometry

Our goal now is to show that the set $\mathscr{H}$ is in some sense "small" (in particular to show that it has zero Lebesgue measure), which will entail differentiability of $y \mapsto x^\star$ Lebesgue almost everywhere. For this, additional geometrical structure on J is needed. Such a rich class of functions is provided by the notion of a semi-algebraic subset of $\mathbb{R}^{\mathbb{N}}$ to be defined shortly. Semi-algebraic sets and functions have been broadly applied to various areas of optimization. The wide applicability of semi-algebraic functions follows largely from their stability under many mathematical operations. In particular, the celebrated Tarski-Seidenberg theorem states, loosely, that the projection of a semi-algebraic set is semi-algebraic. These stability properties are crucial to obtain the following result, proved in [226].

Definition 8 (Semi-algebraic set and function). A set E is semi-algebraic if it is a finite union of sets defined by polynomial equations and (possibly strict) inequalities. A function $f : E \to F$ is semi-algebraic if E and its graph $\{(u, f(u)) \; ; \; u \in E\}$ are semi-algebraic sets.

Remark 15 (From semi-algebraic to o-minimal geometry). The class of semi-algebraic functions is large, and subsumes, for instance, all the regularizers J described in Section 3.2.3. The qualitative properties of semi-algebraic functions are shared by a much bigger class called functions definable in an o-minimal structure over $\mathbb{R}$, or simply definable functions. O-minimal structures over $\mathbb{R}$ correspond in some sense to an axiomatization of some of the prominent geometrical properties of semi-algebraic geometry [68] and particularly of the stability under projection. For

example, the function $J(x) = \sum_i |x_i|^s$, for an arbitrary $s \geqslant 0$, is semi-algebraic only for rational $s \in \mathbb{Q}$, while it is always definable in an o-minimal structure [235]. Due to the variety of regularizations J that can be formulated within the framework of o-minimal structures, all our results stated in this section apply to definable functions, see [226] for a detailed treatment.

Semi-algebraic functions are stable for instance under (sub)differentiation and projection. These stability properties are crucial to obtain the following result, proved in [226].

Proposition 5. *If J is semi-algebraic, the transition space $\mathscr{H}$ is semi-algebraic and has zero Lebesgue measure.*

3.5.3 Unbiased Risk Estimation

A problem of fundamental practical importance is to automatically adjust the parameter λ to reach the best recovery performance when solving ($\mathscr{P}_{y,\lambda}$). Parameter selection is a central theme in statistics, and is intimately related to the question of model selection, as introduced in Section 3.2.1.

We then adopt a statistical framework in which the observation model (3.1) becomes

$$Y = \Phi x_0 + W \tag{3.24}$$

where W is random noise having an everywhere strictly positive probability density function, assumed to be known. Though the forthcoming results can be stated for a large family of distributions, for the sake of concreteness, we only consider the white Gaussian model where $W \sim \mathscr{N}(0, \sigma^2 \mathrm{Id}_{P\times P})$, with known variance σ^2.

Under the observation model (3.24), the ideal choice of λ should be the one which minimizes the quadratic estimation risk $\mathbb{E}_W(\|x^\star(Y) - x_0\|^2)$. This is obviously not realistic as x_0 is not available, and in practice, only one realization of Y is observed. To overcome these obstacles, the traditional approach is to replace the quadratic risk with some estimator that solely depends on Y. The risk estimator is also expected to enjoy nice statistical properties among which unbiasedness is highly desirable.

However, it can be shown, see, e.g., [100, Section IV], that the quadratic risk $\mathbb{E}_W(\|x^\star(Y) - x_0\|^2)$ cannot be reliably estimated on $\mathrm{Ker}(\Phi)$. Nonetheless, we may still obtain a reliable assessment of the part that lies in $\mathrm{Im}(\Phi^*) = \mathrm{Ker}(\Phi)^\perp$ or any linear image of it. For instance, the most straightforward surrogate of the above risk is the so-called prediction risk $\mathbb{E}_W(\|\mu(Y) - \mu_0\|^2)$, where

$$\mu_0 = \Phi x_0 \quad \text{and} \quad \mu(y) = \Phi x^\star(y),$$

where $x^\star(y)$ is any solution of $(\mathscr{P}_{y,\lambda})$. One can easily show that $\mu(y) \in \mathbb{R}^P$ is well defined as a single-valued mapping and thus does not depend on the particular choice of $x^\star(y)$, see [226]. Consequently, Theorem 4 shows that $y \mapsto \mu(y)$ is a C^1 mapping on $\mathbb{R}^P \setminus \mathscr{H}$.

3.5.4 Degrees of Freedom

The degrees of freedom (DOF) quantifies the model "complexity" of a statistical modeling procedure [95]. It is at the heart of several risk estimation procedures. Therefore, in order to design estimators of the prediction risk, an important step is to get an estimator of the corresponding DOF.

Definition 9 (Empirical DOF). Suppose that $y \mapsto \mu(y)$ is differentiable Lebesgue almost everywhere, as is the case when it is Lipschitz-continuous (Rademacher's theorem). The empirical number of degrees of freedom is defined as

$$\mathrm{df}(y) = \mathrm{div}(\mu)(y) = \mathrm{tr}(\mathrm{D}\mu(y)),$$

where the derivative is to be understood in the weak sense, i.e., to hold Lebesgue almost everywhere (a.e.).

An instructive example to get the gist of this formula is the case where μ is the orthogonal projection onto some linear subspace V. We then get easily that $\mathrm{df}(y) = \dim(V)$, which is in agreement with the intuitive notion of the number of DOF.

The following result delivers the closed-form expression of $\mathrm{df}(y)$, valid on a full Lebesgue measure set, for $\mu(y) = \Phi x^\star(y)$ and $x^\star(y)$ an appropriate solution of $(\mathscr{P}_{y,\lambda})$. At this stage, it is important to realize that the main difficulty does not lie in showing almost everywhere differentiability of $\mu(y)$; this mapping is in fact Lipschitz-continuous by classical arguments of sensitivity analysis applied to $(\mathscr{P}_{y,\lambda})$. Rather, it is the existence of such a formula and its validity Lebesgue a. e. that requires more subtle arguments obtained owing to partial smoothness of J. For this, we need also to rule out the points y where $(\mathscr{I}_{x^\star})$ does not hold. This is the rationale behind the following set.

Definition 10 (Non-injectivity set). We define the *Non-injectivity set* $\mathscr{G}$ as

$$\mathscr{G} = \left\{ y \notin \mathscr{H} \;;\; (\mathscr{I}_{x^\star}) \text{ does not hold for any minimizer } x^\star \text{ of} (\mathscr{P}_{y,\lambda}) \right\}.$$

Theorem 5. *For every $y \notin \mathscr{H} \cup \mathscr{G}$, there is $x^\star$ such that $(\mathscr{P}_{y,\lambda})$ holds and*

$$\mathrm{df}(y) = \mathrm{tr}(\Delta_{x^\star}(y)) \quad \textit{where} \quad \Delta_{x^\star}(y) = \Phi_T \circ (\Phi_T{}^* \Phi_T + \lambda \mathrm{D}^2 J_T(x^\star))^{-1} \circ \Phi_T{}^*, \tag{3.25}$$

where $T = T_{x^\star}$.

Remark 16 (Non-injectivity set). It turns out that $\mathscr{G}$ is in fact empty for many regularizers. This is typically the case for $J = \|\cdot\|_1$ [85], $J = \|D^*\cdot\|_1$ [227], and the underlying reasoning can be more generally extended to polyhedral regularizers. The same result was also shown for $J = \|\cdot\|_{1,2}$ in [230]. More precisely, in all these works, it was shown that for each $y \notin \mathscr{H}$, there exists a solution $x^\star$ of $(\mathscr{P}_{y,\lambda})$ that fulfills $(\mathscr{I}_{x^\star})$. The proof is moreover constructive allowing to build such a solution starting from any other one.

3.5.5 Stein Unbiased Risk Estimator (SURE)

We now have all necessary ingredients at hand to design an estimator of the prediction risk.

Definition 11. Suppose that $y \mapsto \mu(y)$ is differentiable Lebesgue almost everywhere, as is the case when it is Lipschitz-continuous. The SURE is defined as

$$\mathrm{SURE}(y) = \|y - \mu(y)\|^2 + 2\sigma^2 \,\mathrm{df}(y) - P\sigma^2. \tag{3.26}$$

In this definition, we have anticipated on unbiasedness of this estimator. In fact, this turns out to be a fundamental property owing to the celebrated lemma of Stein [207], which indeed asserts that the SURE (3.26) is an unbiased estimator of the prediction risk. Therefore, putting together Theorem 5, Proposition 5, and Stein's lemma, we get the following.

Theorem 6. *Suppose that J is semi-algebraic and $\mathscr{G}$ is of zero Lebesgue measure. Then,*

$$\mathbb{E}_W(\mathrm{SURE}(Y)) = \mathbb{E}_W(\|\mu(Y) - \mu_0\|^2)$$

where (3.25) *is plugged into* (3.26)*, and* $\mu(Y) = \Phi x^\star(Y)$.

Remark 17 (Parameter selection). A practical usefulness of the SURE is its ability to provide an objectively guided way to select a good λ from a single observation y by minimizing $\mathrm{SURE}(y)$. While unbiasedness of the SURE is guaranteed, it is hard to control its variance and hence its consistency. This is an open problem in general, and thus little can be said about the actual theoretical efficiency of such an empirical parameter selection method. It works however remarkably well in practice, see the discussion in Section 3.5.6.5 and references therein.

Remark 18 (Projection risk). The SURE can be extended to unbiasedly estimate other risks than the predicted one. For instance, as argued in Section 3.5.3, one can estimate the so-called projection risk defined as $\mathbb{E}_W(\|\,\mathrm{P}_{\mathrm{Ker}(\Phi)^\perp}(x^\star(Y) - x_0)\|^2)$. This is obviously than the prediction risk as a surrogate for the estimation risk.

3.5.6 Related Works

3.5.6.1 Sensitivity Analysis

In Section 3.4.5.4, we reviewed the relevant literature pertaining to sensitivity analysis for partly smooth functions, which is obviously very connected to Theorem 4. See also [21] for the case of linear optimization over a convex semi-algebraic partly smooth feasible set, where the authors prove a sensitivity result with a zero-measure transition space. A distinctive feature of our analysis toward proving unbiasedness of the SURE is the need to ensure that sensitivity analysis can be carried out on a full Lebesgue measure set. In particular, it necessitates local stability of the manifold $\mathscr{M}_{x^\star}$ associated to an appropriate solution $x^\star$, and this has to hold Lebesgue almost everywhere. Thus the combination of partial smoothness and semi-algebraicity is the key.

3.5.6.2 Risk Estimators

In this section, we put emphasis on the SURE as an unbiased estimator of the prediction risk. There are other alternatives in the literature which similarly rely on estimator of the DOF. One can think for instance of the generalized cross-validation (GCV) [115]. Thus our results apply equally well to such risk estimators. Extensions of the SURE to independent variables from a continuous exponential family are considered in [134]. [100] generalizes the SURE principle to continuous multivariate exponential families, see also [180, 227] for the multivariate Gaussian case. The results described here can be extended to these setting as well, see [226].

3.5.6.3 Applications of SURE in Statistics and Imaging

Applications of SURE emerged for choosing the parameters of linear estimators such ridge regression or smoothing splines [149]. After its introduction in the wavelet community through the SURE-Shrink estimator [79], it has been extensively used for various image restoration problems, e.g., with sparse regularization [19, 33, 55, 155, 180, 184–186, 237] or with nonlocal means [75, 90, 233, 234].

3.5.6.4 Closed-form Expressions for SURE

For the Lasso problem, i.e., $J = \|\cdot\|_1$, the divergence formula (3.25) reads

$$\mathrm{df}(y) = |\operatorname{supp}(x^\star)|,$$

where $x^\star$ is a solution of $(\mathscr{P}_{y,\lambda})$ such that $(\mathscr{I}_{x^\star})$ holds, i.e., $\Phi_{\operatorname{supp}(x^\star)}$ has full rank. This result is proved in [245] for injective Φ and in [85] for arbitrary Φ. This result

is extended to analysis ℓ^1-sparsity, i.e., $J = \|D^* \cdot\|_1$, in [214, 227]. A formula for the DOF in the case where $x^\star(y)$ is the orthogonal projection onto a partly smooth convex set C is proved in [141]. This work extends that of [163] which treats the case where C is a convex polyhedral cone. These two works allow one to compute the degrees of freedom of estimators defined by solving $(\mathscr{P}^2_{y,\gamma})$ in the case where Φ is injective. [127] studied the DOF of the metric projection onto a closed set (nonnecessarily convex), and gave a precise representation of the bias when the projection is not sufficiently differentiable.

A formula of an estimate of the DOF for the group Lasso, i.e., $J = \|\cdot\|_{1,2}$ when Φ is orthogonal within each group was conjectured in [242]. An estimate is also given by [200] using heuristic derivations that are valid only when Φ is injective, though its unbiasedness is not proved. [225] derived an estimator of the DOF of the group Lasso and proved its unbiasedness when Φ is injective. Closed-form expression of the DOF estimate for denoising with the nuclear norm, i.e., $\Phi = \mathrm{Id}$ and $J = \|\cdot\|_*$, was concurrently provided in [48, 77].

3.5.6.5 Numerical Methods for SURE

Deriving the closed-form expression of the DOF is in general challenging and has to be addressed on a case-by-case basis. The implementation of the divergence formula such as (3.25) can be computationally expensive in high dimension. But since only the trace of the Jacobian is needed, it is possible to speed up these computations through Monte Carlo sampling, but at the price of mild approximations. If the Jacobian is not known in closed-form or prohibitive to compute, one may appeal to finite-difference approximations along Monte Carlo sampled directions [199, 241], see [111, 184] for applications to imaging problems.

In practice, the analytical formula (3.25) might be subject to serious numerical instabilities, and thus cannot always be applied safely when the solution $x^\star$ is only known approximately. Think for instance of the case where $x^\star$ is approximated by an iterate computed after finitely many iterations of an algorithm as detailed in Section 3.6. A better practice is then to directly compute the DOF, hence the SURE, recursively from the iterates themselves, as proposed by [76, 112, 237].

3.6 Proximal Splitting for Structured Optimization

Though problems $(\mathscr{P}_{y,\lambda})$, $(\mathscr{P}_{y,0})$, $(\mathscr{P}^1_{y,\varepsilon})$, and $(\mathscr{P}^2_{y,\gamma})$ are nonsmooth, they enjoy enough structure to be solved by efficient algorithms. The type of algorithm to be used depends in particular on the properties of J. We first briefly mention some popular nonsmooth optimization schemes in Section 3.6.1 and focus our attention on proximal splitting schemes afterward.

3.6.1 Convex Optimization for Regularized Inverse Problems

3.6.1.1 (Sub)gradient Descent

Consider for example problem ($\mathscr{P}_{y,\lambda}$). This is a convex composite optimization problem where one of the functions is smooth with a Lipschitz-continuous gradient. If J were smooth enough, then a simple gradient (or possibly (quasi-)Newton) descent method could be used. However, as detailed in Section 3.2.2, low complexity regularizers J are intended to be nonsmooth in order to promote models $\mathscr{M}$ of low intrinsic dimension, and J is precisely nonsmooth transverse to $\mathscr{M}$. One can think of replacing gradients by subgradients (elements of the subdifferential), since J is assumed finite-valued (hence closed) convex, which are bounded. This results in a subgradient descent algorithm which is guaranteed to converge but under stringent assumptions on the descent step sizes, which in turn makes their global convergence rate quite slow, see [171].

3.6.1.2 Interior Point Methods

Clearly, the key to getting efficient algorithms is to exploit the structure of the optimization problems at hand while handling nonsmoothness properly. For a large class of regularizers J, such as those introduced in Section 3.2.3, the corresponding optimization problems can be cast as conic programs. The cone constraint can be enforced using a self-concordant barrier function, and the optimization problem can hence be solved using interior point methods, as pioneered by [173], see also the monograph [25]. This class of methods enjoys fast convergence rate. Each iteration however is typically quite costly and can become prohibitive as the dimension increases.

3.6.1.3 Conditional Gradient

This algorithm is historically one of the first methods for smooth constrained convex optimization (a typical example being ($\mathscr{P}^2_{y,\gamma}$)) and was extensively studied in the 70s. It is also known as Frank-Wolfe algorithm, since it was introduced by [106] for quadratic programming and extended in [88]. The conditional gradient algorithm is premised on being able to easily solve (at each iteration) linear optimization problems over the feasible region of interest. This is in contrast to other first-order methods, such as forward-backward splitting and its variants (see Section 3.6.3), which are premised on being able to easily solve (at each iteration) a projection problem. Moreover, in many applications the solutions to the linear optimization subproblem are highly structured and exhibit particular sparsity and/or low-rank properties. These properties have renewed interest in the conditional gradient method to solve sparse recovery (ℓ_1 and total variation), low-rank matrix recovery

(nuclear norm minimization), anti-sparsity recovery, and various other problems in signal processing and machine learning; see, e.g., [62, 87, 128, 135, 198].

3.6.1.4 Homotopy/Path-following

Homotopy and path-following-type methods have been introduced in the case of ℓ^1-minimization to solve $(\mathscr{P}_{y,\lambda})$ by [175]. They were then adapted to analysis ℓ^1, i.e., $J = \|D^*\cdot\|_1$, in [213], and ℓ^∞ regularization, $\|\cdot\|_\infty$, in [108]. One can in fact show that these methods can be applied to any polyhedral regularization (see [231]) because these methods only rely on the crucial fact that the solution path $\lambda \mapsto x^\star_\lambda$, where $x^\star_\lambda$ is a solution of $(\mathscr{P}_{y,\lambda})$, is piecewise affine. The LARS algorithm [96] is an accelerated version of homotopy which computes an approximate homotopy path for $J = \|\cdot\|_1$ along which the support increases monotonically along the course of iterations. In the noiseless compressed sensing case, with Φ drawn from the Gaussian ensemble, it is shown in [82] that if x_0 is k-sparse with $P > 2k\log(N)$, the homotopy method reaches x_0 in only k iterations. This k-solution property was empirically observed for other random matrix ensembles, but at different thresholds for P. In [157], the authors proved that in the worst case, the number of segments in the solution path is exponential in the number of variables, and thus the homotopy method can then take as many iterations to converge.

As for interior points, the cost per iteration of homotopy-like methods, without particular ad hoc optimization, scales badly with the dimension, thus preventing them to be used for large-scale problems such as those encountered in imaging. This class of solvers is thus a wise choice for problems of medium size, and when high accuracy (or even exact computation up to machine precision for the homotopy algorithm) is needed. Extensions of these homotopy methods can deal with progressive changes in the operator Φ or the observations y, and are thus efficient for these settings, see [5].

3.6.1.5 Approximate Message Passing

In the last five years, ideas from graphical models and message passing and approximate message passing algorithms have been proposed to solve large-scale problems of the form $(\mathscr{P}_{y,\lambda})$ for various regularizers J, in particular ℓ^1, $\ell^1 - \ell^2$, and the nuclear norm. A comprehensive review is given in [164]. However, rigorous convergence results have been proved so far only in the case in which Φ is standard Gaussian, though numerical results show that the same behavior should apply for broader random matrix ensembles.

3.6.2 Proximal Splitting Algorithms

Proximal splitting methods are first-order iterative algorithms that are tailored to solve structured nonsmooth (essentially convex) optimization problems. The first operator splitting method has been developed from the 70s. Since then, the class of splitting methods has been regularly enriched with increasingly sophisticated algorithms, as the structure of problems to handle becomes more complex.

To make our discussion more concrete, consider the general problem of minimizing the proper closed convex function

$$f = h + \sum_{k=1}^{K} g_k \circ A_k$$

where $h : \mathbb{R}^N \to \mathbb{R}$ is convex and smooth, the $A_k : \mathbb{R}^N \to \mathbb{R}^{N_k}$ are linear operators, and $g_k : \mathbb{R}^{N_k} \to \mathbb{R}$ are proper closed convex functions for which the so-called proximity operator (to be defined shortly) can be computed easily (typically in closed form). We call such a function g_k "simple."

Definition 12. The proximity operator of a proper closed convex function g is defined as, for $\gamma > 0$,

$$\mathrm{prox}_{\gamma g}(x) = \operatorname*{argmin}_{u \in \mathbb{R}^N} \frac{1}{2} \|x - u\|^2 + \gamma g(u).$$

The proximal operator generalizes the notion of orthogonal projection onto a nonempty closed convex set C that one recovers by taking $g = \iota_C$.

Proximal splitting algorithms may evaluate (possibly approximately) the individual operators (e.g., gradient of h), the proximity operators of the g_ks, the linear operators A_k, all separately at various points in the course of iteration, but never those of sums of functions nor composition by a linear operator. Therefore, each iteration is cheap to compute for large-scale problems. They also enjoy rigorous convergence guarantees, stability to errors, with possibly quantified convergence rates and iteration complexity bounds on various quantities. This justifies their popularity in contemporary signal and image processing or machine learning, despite that their convergence is either sublinear or at best linear.

It is beyond the scope of this chapter to describe thoroughly the huge literature on proximal splitting schemes, as it is a large and extremely active research field in optimization theory. Good resources and reviews on the subject are [13, 16, 64, 177]. We instead give a brief classification of the most popular algorithms according to the class of structured objective functions they are able to handle:

- *Forward-Backward (FB)* algorithm [66, 162, 178]. It is designed to minimize (3.6.2) when h has a Lipschitz-continuous gradient, $K = 1$, $A_1 = \mathrm{Id}$, and g_1 is simple. There are accelerated (optimal) variants of FB, such as the popular Nesterov [172] and Fista [15], but the convergence of the iterates is no

longer guaranteed for these schemes. FB and its variants are good candidates to solve $(\mathscr{P}_{y,\lambda})$. We will further elaborate on FB in Section 3.6.3.

- *Douglas-Rachford (DR)* algorithm [86, 152]. It is designed to minimize (3.6.2) for $h = 0$, $K = 2$, $A_k = \mathrm{Id}$, and g_k is simple for $k = 1, 2$. It can be easily extended to the case of $K > 2$ by either lifting to a product space, see, e.g., [63], or through projective splitting [94]. DR can be used to solve $(\mathscr{P}_{y,0})$, $(\mathscr{P}^1_{y,\varepsilon})$, or $(\mathscr{P}^2_{y,\gamma})$ for certain operators Φ.
- *Generalized Forward-Backward (GFB)* algorithm [183]. It can handle the case of an arbitrary K with $A_k = \mathrm{Id}$, g_k simple and h has a Lipschitz-continuous gradient. It can be interpreted as hybridization of FB scheme and the DR scheme on a product space.
- *Alternate Direction Method of Multipliers (ADMM)* algorithm [104, 109, 110, 113]. It is adapted to minimize (3.6.2) for $h = 0$, $K = 2$ with $A_1 = \mathrm{Id}$ and A_2 is injective. It can be shown [93, 110] that ADMM is equivalent to DR applied to the Fenchel-Rockafellar dual problem $\min_u g_1^* \circ -A_2^*(u) + g_2^*(u)$, where g_k^* is the Legendre-Fenchel conjugate of g_k. While DR applies when g_1 and $g_2 \circ A_2$ are simple, ADMM is a better alternative whereas both $g_1 \circ -A_2^*$ and g_2^* are simple. Extension to the case $K > 2$ was proposed for instance in [92].
- *Dykstra* algorithm [91]. It is able to solve the case where $h(x) = \|x - y\|^2$, $A_k = \mathrm{Id}$, and the g_k are simple functions. It was initially introduced by [91] in the case where the g_k are indicator functions of closed convex sets, and is generalized in [12] to arbitrary convex functions. It is also extended in [14, 51] to the case where h is a Bregman divergence.
- *Primal-Dual schemes*. Recently, primal-dual splitting algorithms have been proposed to minimize (3.6.2) in its full generality, and even more complex objectives, see for instance [29, 52, 56, 65, 67, 201, 222, 238]. Primal-dual schemes can be used to solve $(\mathscr{P}_{y,\lambda})$, $(\mathscr{P}_{y,0})$, $(\mathscr{P}^1_{y,\varepsilon})$, or $(\mathscr{P}^2_{y,\gamma})$.

3.6.3 Finite Model Identification with Forward Backward

The FB algorithm is a good candidate to solve $(\mathscr{P}_{y,\lambda})$ when J is simple. Starting from some $x^{(0)} \in \mathbb{R}^N$, the FB iteration applied to $(\mathscr{P}_{y,\lambda})$ reads

$$x^{(n+1)} = \mathrm{Prox}_{\tau_n \lambda J}\left(x^{(n)} + \tau_n \Phi^*(y - \Phi x^{(n)})\right),$$

where the step-size sequence should satisfy $0 < \underline{\tau} \leqslant \tau_n \leqslant \overline{\tau} < 2/\|\Phi\|^2$ to ensure convergence of the sequence $x^{(n)}$ to a minimizer of $(\mathscr{P}_{y,\lambda})$.

In fact, owing to partial smoothness of J, much more can be said about the iterates of the FB algorithm. More precisely, after a finite number of iterations, Forward-Backward algorithm correctly identifies the manifold $\mathscr{M}$. This is made formal in the following theorem whose proof can be found in [150].

Theorem 7. *Under the assumptions of Theorem 2, $x^{(n)} \in \mathscr{M}$ for n large enough.*

This result sheds some light on the convergence behavior of this algorithm in the favorable case where condition (3.14) holds and $(\|w\|/\lambda, \lambda)$ are sufficiently small. In fact, it is shown in [150] that FB identifies in finite time the manifold of any nondegenerate minimizer $x^\star$. As a corollary, if condition (3.14) holds at x_0 and $(\|w\|/\lambda, \lambda)$ are sufficiently small, then we recover Theorem 7. These results shed light on the typical convergence behavior of FB observed in such circumstances (e.g., in compressed sensing problems).

Remark 19 (Local linear convergence). The FB generally exhibits a global sublinear $O(1/n)$ convergence rate in terms of the objective function. However, under partial smoothness of J, it is shown in [150] that once the active manifold is identified, the FB algorithm enters a local linear convergence regime (Q-linear in general and R-linear if $\mathscr{M}$ is a linear manifold), whose rate can be characterized precisely in terms of the condition number of $\Phi_{T_{x_0}}$.

3.6.4 Related Works

Finite support identification and local R-linear convergence of FB to solve $(\mathscr{P}_{y,\lambda})$ is established in [26] under either a very restrictive injectivity assumption or a nondegeneracy assumption that is a specialization of ours to the ℓ_1 norm. A similar result is proved in [124]. The ℓ_1 norm is a partly smooth function and is therefore covered by Theorem 7. [170] proved Q-linear convergence of FB to solve $(\mathscr{P}_{y,\lambda})$ with a data fidelity satisfying restricted smoothness and strong convexity assumptions, and J a so-called convex decomposable regularizer. Again, the latter falls within the class of partly smooth functions, and their result is then subsumed by our analysis.

For general programs, a variety of algorithms, such as proximal and projected-gradient schemes, were observed to have the finite identification property of the active manifold. In [130, 131], the authors have shown finite identification of manifolds associated to partly smooth functions via the (sub)gradient projection method, Newton-like methods, and the proximal point algorithm. Their work extends that of e.g., [240] on identifiable surfaces from the smooth constrained convex case to a general nonsmooth setting. Using these results, [129] considered the algorithm [223] to solve (3.6.2) when h is C^2, $K = 1, A_1 = \mathrm{Id}$, and g_1 is simple and partly smooth, but not necessarily convex, and proved finite identification of the active manifold. However, the convergence rates remain an open problem in all these works.

3.7 Summary and Perspectives

In this chapter, we have reviewed work covering a large body of literature on the regularization of linear inverse problems. We also showed how these previous works can all be seen as particular instances of a unified framework, namely sensitivity analysis for minimization of convex partly smooth functions. We believe this general framework is the one that should be adopted as long as one is interested in studying fine properties and guarantees of these regularizers, and in particular when the stability of the low complexity manifold associated to the data to recover is at stake.

This analysis is however only the tip of the iceberg, and there is actually a flurry of open problems to go beyond the theoretical results presented in this chapter. We list here a few ones that we believe are important avenues for future works:

- *Non-convexity and/or nonfiniteness:* in this chapter, for the sake of simplicity, we focused on smooth convex fidelity terms and finite-valued convex regularizers. All the results stated in this chapter extend readily to proper lower semicontinuous convex regularizers, since any such a function is subdifferentially regular. Generalizations of some of the results to non-convex regularizers is possible as well, though some regularity assumptions are needed. This is of practical importance to deal with settings where Φ is not a linear operator, or to impose more aggressive regularization (for instance when using ℓ^p functional with $0 \leqslant p < 1$ instead of the ℓ^1 norm). There are however many difficulties to tackle in this case. For instance, regularity properties that hold automatically for the convex case have to be either imposed or proved. Another major bottleneck is that some of the results presented here, if extended verbatim, will only assess the recovery of a stationary/critical point. The latter is not a local minimum in general, and even less global.
- *Dictionary learning:* a related non-convex sensitivity analysis problem is to understand the recovery of the dictionary D in synthesis regularization (as defined in Section 3.2.3.4) when solving problems of the form

$$\min_{\{\alpha_k\}_k, D \in \mathscr{D}} \sum_k \frac{1}{2} \|y - \Phi D \alpha_k\|^2 + \lambda J_0(\alpha_k)$$

  where the $(y_k)_k$ are a set of input exemplars and $\mathscr{D}$ stands for the set of constraints imposed on the dictionary to avoid trivial solutions. Such a non-convex variational problem is popular to compute adapted dictionaries, in particular when $J_0 = \|\cdot\|_1$, see [97] and references therein. Although the dictionary learning problem has been extensively studied when $J_0 = \|\cdot\|_1$, most of the methods lack theoretical guarantees. The theory of dictionary learning is only beginning to develop, see, e.g., [1, 121, 139, 202]. Tackling other regularizers, including analysis ℓ^1 of the form $J = J_0 \circ D^*$ is even more difficult, see, e.g., [59] for some computational schemes.
- *Infinite-dimensional problems:* we dealt in this chapter with finite-dimensional vector spaces. It is not straightforward to extend these results to infinite-dimensional cases. As far as ℓ_2-stability is concerned, the constants involved in

the upper bounds depend on the dimension N, and the scaling might diverge as $N \to +\infty$. We refer to Section 3.3.3 for previous works on convergence rates of Tikhonov regularization in infinite-dimensional Hilbert or Banach spaces. Extending Theorem 2 for possibly nonreflexive Banach spaces is however still out of reach (nonreflexivity is a typical degeneracy when considering low complexity regularization). There exists however some extensions of classical stability results over spaces of measures, such as weak convergence [27], exact recovery [50, 74], and stable support recovery [89].

- *Compressed sensing:* as highlighted in Sections 3.3.3.3 and 3.4.5.3, the general machinery of partly smooth regularizers (and the associated dual certificates) is well adapted to derive optimal recovery bounds for compressed sensing. Unfortunately, this analysis has been for now only applied to norms ($\|\cdot\|_1$, $\|\cdot\|_{1,\mathscr{B}}$, $\|\cdot\|_*$, and $\|\cdot\|_\infty$). Extending this framework for synthesis and analysis regularizers (see Sections 3.2.3.4 and 3.2.3.5) is a difficult open problem.
- *Convergence and acceleration of the optimization schemes:* Section 3.6.3 showed how partial smoothness can be used to achieve exact manifold identification after a finite number of iterations using the FB algorithm. This in turn implies a local linear convergence of the iterates and raises the hope of acceleration using either first-order or second-order information for the function along the identified manifold (in which we recall it is C^2). Studying such accelerations and their guarantees as well as extending this idea to other proximal splitting schemes is thus of practical importance to tackle more complicated problems such as, e.g., ($\mathscr{P}_{y,0}$), ($\mathscr{P}^1_{y,\varepsilon}$), or ($\mathscr{P}^2_{y,\gamma}$).

Acknowledgements This work was supported by the European Research Council (ERC project SIGMA-Vision). We would like to thank our collaborators Charles Deledalle, Charles Dossal, Mohammad Golbabaee, and Vincent Duval who have helped to build this unified view of the field.

References

1. A. Agarwal, A. Anandkumar, P. Netrapalli, Exact recovery of sparsely used overcomplete dictionaries (2013) [arxiv]
2. H. Akaike, Information theory and an extension of the maximum likelihood principle, in *Second International Symposium on Information Theory* (Springer, New York, 1973), pp. 267–281
3. J. Allen, Short-term spectral analysis, and modification by discrete Fourier transform. IEEE Trans. Acoust. Speech Signal Process. **25**(3), 235–238 (1977)
4. D. Amelunxen, M. Lotz, M.B. McCoy, J.A. Tropp, Living on the edge: a geometric theory of phase transitions in convex optimization. CoRR, abs/1303.6672 (2013)
5. M.S. Asif, J. Romberg, Sparse recovery of streaming signals using L1-homotopy. Technical report, Preprint (2013) [arxiv 1306.3331]
6. J.-F. Aujol, G. Aubert, L. Blanc-Féraud, A. Chambolle, Image decomposition into a bounded variation component and an oscillating component. J. Math. Imaging Vis. **22**, 71–88 (2005)

7. F. Bach, Consistency of the group Lasso and multiple kernel learning. J. Mach. Learn. Res. **9**, 1179–1225 (2008)
8. F. Bach, Consistency of trace norm minimization. J. Mach. Learn. Res. **9**, 1019–1048 (2008)
9. S. Bakin, Adaptive regression and model selection in data mining problems. Ph.D. thesis, Australian National University, 1999
10. A.S. Bandeira, E. Dobriban, D.G. Mixon, W.F. Sawin, Certifying the restricted isometry property is hard. IEEE Trans. Inf. Theory **59**(6), 3448–3450 (2013)
11. A. Barron, L. Birgé, P. Massart, Risk bounds for model selection via penalization. Probab. Theory Relat. Fields **113**(3), 301–413 (1999)
12. H.H. Bauschke, P.L. Combettes, A dykstra-like algorithm for two monotone operators. Pac. J. Optim. **4**(3), 383–391 (2008)
13. H.H. Bauschke, P.L. Combettes, *Convex Analysis and Monotone Operator Theory in Hilbert Spaces* (Springer, New York, 2011)
14. H.H. Bauschke, A.S Lewis, Dykstras algorithm with bregman projections: a convergence proof. Optimization **48**(4), 409–427 (2000)
15. A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm for linear inverse problems. SIAM J. Imaging Sci. **2**(1), 183–202 (2009)
16. A. Beck, M. Teboulle, Gradient-based algorithms with applications to signal recovery, in *Convex Optimization in Signal Processing and Communications* (Cambridge University Press, Cambridge, 2009)
17. L. Birgé, P. Massart, From model selection to adaptive estimation, Chapter 4, in *Festschrift for Lucien Le Cam*, ed. by D. Pollard, E. Torgersen, L.Y. Grace (Springer, New York, 1997), pp. 55–87
18. L. Birgé, P. Massart, Minimal penalties for Gaussian model selection. Probab. Theory Relat. Fields **138**(1–2), 33–73 (2007)
19. T. Blu, F. Luisier, The SURE-LET approach to image denoising. IEEE Trans. Image Process. **16**(11), 2778–2786 (2007)
20. T. Blumensath, M.E. Davies, Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. **27**(3), 265–274 (2009)
21. J. Bolte, A. Daniilidis, A.S. Lewis, Generic optimality conditions for semialgebraic convex programs. Math. Oper. Res. **36**(1), 55–70 (2011)
22. C. Boncelet, Image noise models. *Handbook of Image and Video Processing* (Academic, New York, 2005)
23. J.F. Bonnans, A. Shapiro, *Perturbation Analysis of Optimization Problems*. Springer Series in Operations Research and Financial Engineering (Springer, New York, 2000)
24. L. Borup, R. Gribonval, M. Nielsen, Beyond coherence: recovering structured time-frequency representations. Appl. Comput. Harmon. Anal. **24**(1), 120–128 (2008)
25. S.P. Boyd, L. Vandenberghe, *Convex Optimization* (Cambridge University Press, Cambridge, 2004)
26. K. Bredies, D.A. Lorenz, Linear convergence of iterative soft-thresholding. J. Four. Anal. Appl. **14**(5–6), 813–837 (2008)
27. K. Bredies, H.K. Pikkarainen, Inverse problems in spaces of measures. ESAIM Control Optim. Calc. Var. **19**, 190–218 (2013)
28. K. Bredies, K. Kunisch, T. Pock, Total generalized variation. SIAM J. Imaging Sci. **3**(3), 492–526 (2010)
29. L.M. Briceño Arias, P.L. Combettes, A monotone+skew splitting model for composite monotone inclusions in duality. SIAM J. Optim. **21**(4), 1230–1250 (2011)
30. M. Burger, S. Osher, Convergence rates of convex variational regularization. Inverse Prob. **20**(5), 1411 (2004)
31. T.T. Cai, Adaptive wavelet estimation: a block thresholding and oracle inequality approach. Ann. stat. **27**(3), 898–924 (1999)
32. T.T. Cai, B.W. Silverman, Incorporating information on neighbouring coefficients into wavelet estimation. Sankhya Indian J. Stat. Ser. B **63**, 127–148 (2001)

33. T.T. Cai, H.H. Zhou, A data-driven block thresholding approach to wavelet estimation. Ann. Stat. **37**(2), 569–595 (2009)
34. E.J. Candès, D.L. Donoho, Curvelets: a surprisingly effective nonadaptive representation for objects with edges. Technical report, DTIC Document (2000)
35. E.J. Candès, Y. Plan, Matrix completion with noise. Proc. IEEE **98**(6), 925–936 (2010)
36. E.J. Candès, Y. Plan, A probabilistic and RIPless theory of compressed sensing. IEEE Trans. Inf. Theory **57**(11), 7235–7254 (2011)
37. E.J. Candès, Y. Plan, Tight oracle inequalities for low-rank matrix recovery from a minimal number of noisy random measurements. IEEE Trans. Inf. Theory **57**(4), 2342–2359 (2011)
38. E.J. Candès, B. Recht, Exact matrix completion via convex optimization. Found. Comput. Math. **9**(6), 717–772 (2009)
39. E.J. Candès, B. Recht, Simple bounds for recovering low-complexity models. Math. Program. **141**(1–2), 577–589 (2013)
40. E. J. Candès, T. Tao, Decoding by linear programming. IEEE Trans. Inf. Theory **51**(12), 4203–4215 (2005)
41. E.J. Candès, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies? IEEE Trans. Inf. Theory **52**(12), 5406–5425 (2006)
42. E.J. Candès, T. Tao, The power of convex relaxation: near-optimal matrix completion. IEEE Trans. Inf. Theory **56**(5), 2053–2080 (2010)
43. E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
44. E.J. Candès, J. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
45. E.J. Candès, M. Wakin, S. Boyd, Enhancing sparsity by reweighted ℓ_1 minimization. J. Four. Anal. Appl. **14**, 877–905 (2007)
46. E.J. Candès, Y.C. Eldar, D. Needell, P. Randall, Compressed sensing with coherent and redundant dictionaries. Appl. Comput. Harmon. Anal. **31**(1), 59–73 (2011)
47. E.J. Candès, X. Li, Y. Ma, J. Wright, Robust principal component analysis? J. ACM **58**(3), 11:1–11:37 (2011)
48. E.J. Candès, C.A. Sing-Long, J.D. Trzasko, Unbiased risk estimates for singular value thresholding and spectral estimators. IEEE Trans. Signal Process. **61**(19), 4643–4657 (2012)
49. E.J. Candès, T. Strohmer, V. Voroninski, Phaselift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**(8), 1241–1274 (2013)
50. E.J. Candès, C. Fernandez-Granda, Towards a mathematical theory of super-resolution. Commun. Pure Appl. Math. **67**(6), 906–956 (2014)
51. Y. Censor, S. Reich, The dykstra algorithm with bregman projections. Commun. Appl. Anal. **2**, 407–419 (1998)
52. A. Chambolle, T. Pock, A first-order primal-dual algorithm for convex problems with applications to imaging. J. Math. Imaging Vis. **40**(1), 120–145 (2011)
53. A. Chambolle, V. Caselles, D. Cremers, M. Novaga, T. Pock, An introduction to total variation for image analysis, in *Theoretical Foundations and Numerical Methods for Sparse Recovery* (De Gruyter, Berlin, 2010)
54. V. Chandrasekaran, B. Recht, P.A. Parrilo, A. Willsky, The convex geometry of linear inverse problems. Found. Comput. Math. **12**(6), 805–849 (2012)
55. C. Chaux, L. Duval, A. Benazza-Benyahia, J.-C. Pesquet, A nonlinear stein-based estimator for multichannel image denoising. IEEE Trans. Signal Process. **56**(8), 3855–3870 (2008)
56. G. Chen, M. Teboulle, A proximal-based decomposition method for convex minimization problems. Math. Program. **64**(1–3), 81–101 (1994)
57. J. Chen, X. Huo, Theoretical results on sparse representations of multiple-measurement vectors. IEEE Trans. Signal Process. **54**(12), 4634–4643 (2006)
58. S.S. Chen, D.L. Donoho, M.A. Saunders, Atomic decomposition by Basis Pursuit. SIAM J. Sci. Comput. **20**(1), 33–61 (1999)

59. Y. Chen, T. Pock, H. Bischof, Learning ℓ_1-based analysis and synthesis sparsity priors using bi-level optimization, in *NIPS* (2012)
60. R. Ciak, B. Shafei, G. Steidl, Homogeneous penalizers and constraints in convex image restoration. J. Math. Imaging Vis. **47**, 210–230 (2013)
61. J.F. Claerbout, F. Muir, Robust modeling with erratic data. Geophysics **38**(5), 826–844 (1973)
62. K.L. Clarkson, Coresets, sparse greedy approximation, and the frank-wolfe algorithm, in *19th ACM-SIAM Symposium on Discrete Algorithms* (2008), pp. 922–931
63. P.L. Combettes, J.-C. Pesquet, A proximal decomposition method for solving convex variational inverse problems. Inverse Prob. **24**(6), 065014 (2008)
64. P.L. Combettes, J.-C. Pesquet, Proximal splitting methods in signal processing, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, ed. by H.H. Bauschke, R.S. Burachik, P.L. Combettes, V. Elser, D.R. Luke, H. Wolkowicz (Springer, New York, 2011), pp. 185–212
65. P.L. Combettes, J.C. Pesquet, Primal–dual splitting algorithm for solving inclusions with mixtures of composite, lipschitzian, and parallel-sum type monotone operators. Set-Valued Var. Anal. **20**(2), 307–330 (2012)
66. P.L. Combettes, V.R. Wajs, Signal recovery by proximal forward-backward splitting. Multiscale Model. Simul. **4**(4), 1168–1200 (2005)
67. L. Condat, A primal–dual splitting method for convex optimization involving lipschitzian, proximable and linear composite terms. J. Optim. Theory Appl. **158**, 1–20 (2012)
68. M Coste, *An introduction to o-minimal geometry*. Pisa: Istituti editoriali e poligrafici internazionali (2000)
69. S.F. Cotter, B.D. Rao, J. Engan, K. Kreutz-Delgado, Sparse solutions to linear inverse problems with multiple measurement vectors. IEEE Trans. Signal Process. **53**(7), 2477–2488 (2005)
70. A. Daniilidis, D. Drusvyatskiy, A.S. Lewis, Orthogonal invariance and identifiability. Technical report (2013) [arXiv 1304.1198]
71. A. Daniilidis, J. Malick, H.S. Sendov, Spectral (isotropic) manifolds and their dimension, to appear in Journal d'Analyse Mathematique, 25 (2014)
72. I. Daubechies, R. DeVore, M. Fornasier, C.S. Gunturk, Iteratively reweighted least squares minimization for sparse recovery. Commun. Pure Appl. Math. **63**(1), 1–38 (2010)
73. G. Davis, S.G. Mallat, Z. Zhang, Adaptive time-frequency approximations with matching pursuits. Technical report, Courant Institute of Mathematical Sciences (1994)
74. Y. de Castro, F. Gamboa, Exact reconstruction using beurling minimal extrapolation. J. Math. Anal. Appl. **395**(1), 336–354 (2012)
75. C-A. Deledalle, V. Duval, J. Salmon, Non-local Methods with Shape-Adaptive Patches (NLM-SAP). J. Math. Imaging Vis. **43**, 1–18 (2011)
76. C. Deledalle, S. Vaiter, G. Peyré, M.J. Fadili, C. Dossal, Proximal splitting derivatives for risk estimation, in *2nd International Workshop on New Computational Methods for Inverse Problems (NCMIP)*, Paris (2012)
77. C.-A. Deledalle, S. Vaiter, G. Peyré, M.J. Fadili, C. Dossal, Risk estimation for matrix recovery with spectral regularization, in *ICML'12 Workshops* (2012) [arXiv:1205.1482v1]
78. D.L. Donoho, X. Huo, Uncertainty principles and ideal atomic decomposition. IEEE Trans. Inf. Theory **47**(7), 2845–2862 (2001)
79. D.L. Donoho, Johnstone, I.M.: Adapting to unknown smoothness via wavelet shrinkage. J. Am. Stat. Assoc. **90**(432), 1200–1224 (1995)
80. D.L. Donoho, B.F. Logan, Signal recovery and the large sieve. SIAM J. Appl. Math. **52**(2), 577–591 (1992)
81. D.L. Donoho, P.B. Stark, Uncertainty principles and signal recovery. SIAM J. Appl. Math. **49**(3), 906–931 (1989)
82. D.L. Donoho, Y. Tsaig, Fast solution of ℓ^1-norm minimization problems when the solution may be sparse. IEEE Trans. Inf. Theory **54**(11), 4789–4812 (2008)
83. C. Dossal, S. Mallat, Sparse spike deconvolution with minimum scale, in *Proc. SPARS 2005* (2005)

84. C. Dossal, M.-L. Chabanol, G. Peyré, J.M. Fadili, Sharp support recovery from noisy random measurements by ℓ^1-minimization. Appl. Comput. Harmon. Anal. **33**(1), 24–43 (2012)
85. C. Dossal, M. Kachour, M.J. Fadili, G. Peyré, C. Chesneau, The degrees of freedom of the Lasso for general design matrix. Stat. Sin. **23**, 809–828 (2013)
86. J. Douglas, H.H. Rachford, On the numerical solution of heat conduction problems in two and three space variables. Trans. Am. Math. Soc. **82**(2), 421–439 (1956)
87. M. Dudík, Z. Harchaoui, J. Malick, Lifted coordinate descent for learning with trace-norm regularization, in *Proc. AISTATS*, ed. by N.D. Lawrence, M. Girolami. JMLR Proceedings, vol. 22, JMLR.org (2012), pp. 327–336
88. J.C. Dunn, S. Harshbarger, Conditional gradient algorithms with open loop step size rules. J. Math. Anal. Appl. **62**(2), 432–444 (1978)
89. V. Duval, G. Peyré, Exact support recovery for sparse spikes deconvolution. Technical report, Preprint hal-00839635 (2013)
90. V. Duval, J.-F. Aujol, Y. Gousseau, A bias-variance approach for the non-local means. SIAM J. Imaging Sci. **4**(2), 760–788 (2011)
91. R.L. Dykstra, An algorithm for restricted least squares regression. J. Am. Stat. **78**, 839–842 (1983)
92. J. Eckstein, Parallel alternating direction multiplier decomposition of convex programs. J. Optim. Theory Appl. **80**(1), 39–62 (1994)
93. J. Eckstein, D.P. Bertsekas, On the douglas–rachford splitting method and the proximal point algorithm for maximal monotone operators. Math. Program. **55**(1–3), 293–318 (1992)
94. J. Eckstein, B.F. Svaiter, General projective splitting methods for sums of maximal monotone operators. SIAM J. Control Optim. **48**(2), 787–811 (2009)
95. B. Efron, How biased is the apparent error rate of a prediction rule? J. Am. Stat. Assoc. **81**(394), 461–470 (1986)
96. B. Efron, T. Hastie, I. Johnstone, R. Tibshirani, Least angle regression. Ann. Stat. **32**(2), 407–451 (2004)
97. M. Elad, *Sparse and Redundant Representations: From Theory to Applications in Signal and Iimage Processing* (Springer, New York, 2010)
98. M. Elad, J.-L. Starck, P. Querre, D.L. Donoho, Simultaneous cartoon and texture image inpainting using morphological component analysis (MCA). Appl. Comput. Harmon. Anal. **19**(3), 340–358 (2005)
99. M. Elad, P. Milanfar, R. Rubinstein, Analysis versus synthesis in signal priors. Inverse Prob. **23**(3), 947 (2007)
100. Y.C. Eldar, Generalized SURE for exponential families: applications to regularization. IEEE Trans. Signal Process. **57**(2), 471–481 (2009)
101. M.J. Fadili, G. Peyré, S. Vaiter, C.-A. Deledalle, J. Salmon, Stable recovery with analysis decomposable priors, in *Proc. SampTA* (2013)
102. M. Fazel, Matrix rank minimization with applications. Ph.D. thesis, Stanford University, 2002
103. M. Fazel, H. Hindi, S.P. Boyd, A rank minimization heuristic with application to minimum order system approximation, in *Proceedings of the 2001 American Control Conference*, vol. 6 (IEEE, Arlington, 2001), pp. 4734–4739
104. M. Fortin, R. Glowinski, *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-Value Problems* (Elsevier, Amsterdam, 2000)
105. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Birkhäuser Series in Applied and Numerical Harmonic Analysis (Birkhäuser, Basel, 2013)
106. M. Frank, P. Wolfe, An algorithm for quadratic programming. Nav. Res. Logist. Q. **3**(1–2), 95–110 (1956)
107. J.-J. Fuchs, On sparse representations in arbitrary redundant bases. IEEE Trans. Inf. Theory **50**(6), 1341–1344 (2004)
108. J.-J. Fuchs, Spread representations, in *Signals, Systems and Computers (ASILOMAR)* (IEEE, Pacific Grove, 2011), pp. 814–817

109. D. Gabay, Applications of the method of multipliers to variational inequalities, in *Augmented Lagrangian Methods: Applications to the Numerical Solution of Boundary-value Problems*, ed. by M. Fortin, R. Glowinski (North-Holland, Amsterdam, 1983)
110. D. Gabay, B. Mercier, A dual algorithm for the solution of nonlinear variational problems via finite element approximation. Comput. Math. Appl. **2**(1), 17–40 (1976)
111. A Girard, A fast Monte-Carlo cross-validation procedure for large least squares problems with noisy data. Numer. Math. **56**(1), 1–23 (1989)
112. R. Giryes, M. Elad, Y.C. Eldar, The projected GSURE for automatic parameter tuning in iterative shrinkage methods. Appl. Comput. Harmon. Anal. **30**(3), 407–422 (2011)
113. R. Glowinski, P. Le Tallec, *Augmented Lagrangian and Operator-Splitting Methods in Nonlinear Mechanics*, vol. 9 (SIAM, Philadelphia, 1989)
114. M. Golbabaee, P. Vandergheynst, Hyperspectral image compressed sensing via low-rank and joint-sparse matrix recovery, in *2012 IEEE International Conference On Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Kyoto, 2012), pp. 2741–2744
115. G.H. Golub, M. Heath, G. Wahba, Generalized cross-validation as a method for choosing a good ridge parameter. Technometrics **21**(2), 215–223 (1979)
116. M. Grasmair, Linear convergence rates for Tikhonov regularization with positively homogeneous functionals. Inverse Prob. **27**(7), 075014 (2011)
117. M. Grasmair, O. Scherzer, M. Haltmeier, Necessary and sufficient conditions for linear convergence of l1-regularization. Commun. Pure Appl. Math. **64**(2), 161–182 (2011)
118. E. Grave, G. Obozinski, F. Bach, Trace Lasso: a trace norm regularization for correlated designs, in *Neural Information Processing Systems (NIPS)*, Spain (2012)
119. R. Gribonval, Should penalized least squares regression be interpreted as maximum a posteriori estimation? IEEE Trans. Signal Process. **59**(5), 2405–2410 (2011)
120. R. Gribonval, M. Nielsen, Beyond sparsity: recovering structured representations by ℓ^1-minimization and greedy algorithms. Adv. Comput. Math. **28**(1), 23–41 (2008)
121. R. Gribonval, K. Schnass, Dictionary identification - sparse matrix factorization via ℓ_1-minimization. IEEE Trans. Inf. Theory **56**(7), 3523–3539 (2010)
122. R. Gribonval, H. Rauhut, K. Schnass, P. Vandergheynst, Atoms of all channels, unite! average case analysis of multi-channel sparse recovery using greedy algorithms. J. Four. Anal. Appl. **14**(5–6), 655–687 (2008)
123. D. Gross, Recovering low-rank matrices from few coefficients in any basis. IEEE Trans. Inf. Theory **57**(3), 1548–1566 (2011)
124. E. Hale, W. Yin, Y. Zhang, Fixed-point continuation for ℓ_1-minimization: methodology and convergence. SIAM J. Optim. **19**(3), 1107–1130 (2008)
125. P. Hall, S. Penev, G. Kerkyacharian, D. Picard, Numerical performance of block thresholded wavelet estimators. Stat. Comput. **7**(2), 115–124 (1997)
126. P. Hall, G. Kerkyacharian, D. Picard, On the minimax optimality of block thresholded wavelet estimators. Stat. Sin. **9**(1), 33–49 (1999)
127. N.R. Hansen, A. Sokol, Degrees of freedom for nonlinear least squares estimation. Technical report (2014) [arXiv 1402.2997]
128. E. Harchaoui, A. Juditsky, A. Nemirovski, Conditional gradient algorithms for norm-regularized smooth convex optimization. Math. Program. **152**(1–2), 75–112 (2014)
129. W.L. Hare, Identifying active manifolds in regularization problems, Chapter 13, in *Fixed-Point Algorithms for Inverse Problems in Science and Engineering*, ed. by H.H. Bauschke, R.S., Burachik, P.L. Combettes, V. Elser, D.R. Luke, H. Wolkowicz. Springer Optimization and Its Applications, vol. 49 (Springer, New York, 2011)
130. W.L. Hare, A.S. Lewis, Identifying active constraints via partial smoothness and prox-regularity. J. Convex Anal. **11**(2), 251–266 (2004)
131. W. Hare, A.S. Lewis, Identifying active manifolds. Algorithmic Oper. Res. **2**(2) (2007)
132. J.-B. Hiriart-Urruty, H.Y. Le, Convexifying the set of matrices of bounded rank: applications to the quasiconvexification and convexification of the rank function. Optim. Lett. **6**(5), 841–849 (2012)

133. B. Hofmann, B. Kaltenbacher, C. Poeschl, O. Scherzer, A convergence rates result for Tikhonov regularization in Banach spaces with non-smooth operators. Inverse Prob. **23**(3), 987 (2007)
134. H.M. Hudson, A natural identity for exponential families with applications in multiparameter estimation. Ann. Stat. **6**(3), 473–484 (1978)
135. M. Jaggi, M. Sulovsky, A simple algorithm for nuclear norm regularized problems, in *ICML* (2010)
136. H. Jégou, M. Douze, C. Schmid, Improving bag-of-features for large scale image search. Int. J. Comput. Vis. **87**(3), 316–336 (2010)
137. H. Jégou, T. Furon, J.-J. Fuchs, Anti-sparse coding for approximate nearest neighbor search, in *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)* (IEEE, Kyoto, 2012), pp. 2029–2032
138. R. Jenatton, J.Y. Audibert, F. Bach, Structured variable selection with sparsity-inducing norms. J. Mach. Learn. Res. **12**, 2777–2824 (2011)
139. R. Jenatton, R. Gribonval, F. Bach, Local stability and robustness of sparse dictionary learning in the presence of noise (2012) [arxiv:1210.0685]
140. J. Jia, B. Yu, On model selection consistency of the elastic net when $p \gg n$. Stat. Sin. **20**, 595–611 (2010)
141. K. Kato, On the degrees of freedom in shrinkage estimation. J. Multivariate Anal. **100**(7), 1338–1352 (2009)
142. K. Knight, W. Fu, Asymptotics for Lasso-Type Estimators. Ann. Stat. **28**(5), 1356–1378 (2000)
143. J.M. Lee, *Smooth Manifolds* (Springer, New York, 2003)
144. C. Lemaréchal, F. Oustry, C. Sagastizábal, The $\mathscr{U}$-lagrangian of a convex function. Trans. Am. Math. Soc. **352**(2), 711–729 (2000)
145. A.S. Lewis, Active sets, nonsmoothness, and sensitivity. SIAM J. Optim. **13**(3), 702–725 (2002)
146. A.S. Lewis, The mathematics of eigenvalue optimization. Math. Program. **97**(1–2), 155–176 (2003)
147. A.S. Lewis, J. Malick, Alternating projections on manifolds. Math. Oper. Res. **33**(1), 216–234 (2008)
148. A.S. Lewis, S. Zhang, Partial smoothness, tilt stability, and generalized hessians. SIAM J. Optim. **23**(1), 74–94 (2013)
149. K.-C. Li. From Stein's unbiased risk estimates to the method of generalized cross validation. Ann. Stat. **13**(4), 1352–1377 (1985)
150. J. Liang, M.J Fadili, G. Peyré, Local linear convergence of forward–backward under partial smoothness. Technical report (2014) [arxiv preprint arXiv:1407.5611]
151. S.G. Lingala, Y. Hu, E.V.R. Di Bella, M. Jacob, Accelerated dynamic MRI exploiting sparsity and low-rank structure: k-t SLR. IEEE Trans. Med. Imaging **30**(5), 1042–1054 (2011)
152. P.L. Lions, B. Mercier, Splitting algorithms for the sum of two nonlinear operators. SIAM J. Numer. Anal. **16**(6), 964–979 (1979)
153. D.A. Lorenz, Convergence rates and source conditions for Tikhonov regularization with sparsity constraints. J. Inverse Ill-Posed Prob. **16**(5), 463–478 (2008)
154. D. Lorenz, N. Worliczek, Necessary conditions for variational regularization schemes. Inverse Prob. **29**(7), 075016 (2013)
155. F. Luisier, T. Blu, M. Unser, Sure-let for orthonormal wavelet-domain video denoising. IEEE Trans. Circuits Syst. Video Technol. **20**(6), 913–919 (2010)
156. Y. Lyubarskii, R. Vershynin, Uncertainty principles and vector quantization. IEEE Trans. Inf. Theory **56**(7), 3491–3501 (2010)
157. J. Mairal, B. Yu, Complexity analysis of the lasso regularization path, in *ICML'12* (2012)
158. S.G. Mallat, A theory for multiresolution signal decomposition: the wavelet representation. IEEE Trans. Pattern Anal. Mach. Intell. **11**(7), 674–693 (1989)
159. S.G. Mallat, *A Wavelet Tour of Signal Processing*, 3rd edn. (Elsevier/Academic, Amsterdam, 2009)

160. S.G. Mallat, Z. Zhang, Matching pursuits with time-frequency dictionaries. IEEE Trans. Signal Process. **41**(12), 3397–3415 (1993)
161. C.L. Mallows, Some comments on C_p. Technometrics **15**(4), 661–675 (1973)
162. B. Mercier, Topics in finite element solution of elliptic problems. Lect. Math. **63** (1979)
163. M. Meyer, M. Woodroofe, On the degrees of freedom in shape-restricted regression. Ann. Stat. **28**(4), 1083–1104 (2000)
164. A. Montanari, Graphical models concepts in compressed sensing, in *Compressed Sensing*, ed. by Y. Eldar, G. Kutyniok (Cambridge University Press, Cambridge, 2012)
165. B.S. Mordukhovich, Sensitivity analysis in nonsmooth optimization, in *Theoretical Aspects of Industrial Design*, vol. 58, ed. by D.A. Field, V. Komkov (SIAM, Philadelphia, 1992), pp. 32–46
166. S. Nam, M.E. Davies, M. Elad, R. Gribonval, The cosparse analysis model and algorithms. Appl. Comput. Harmon. Anal. **34**(1), 30–56 (2013)
167. B.K. Natarajan, Sparse approximate solutions to linear systems. SIAM J. Comput. **24**(2), 227–234 (1995)
168. D. Needell, J. Tropp, R. Vershynin, Greedy signal recovery review, in *Conference on Signals, Systems and Computers* (IEEE, Pacific Grove, 2008), pp. 1048–1050
169. S.N. Negahban, M.J. Wainwright, Simultaneous support recovery in high dimensions: Benefits and perils of block-regularization. IEEE Trans. Inf. Theory **57**(6), 3841–3863 (2011)
170. S. Negahban, P. Ravikumar, M.J. Wainwright, B. Yu, A unified framework for high-dimensional analysis of M-estimators with decomposable regularizers. Stat. Sci. **27**(4), 538–557 (2012)
171. Y. Nesterov, Smooth minimization of non-smooth functions. Math. Program. **103**(1), 127–152 (2005)
172. Y. Nesterov, Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE) (2007)
173. Y. Nesterov, A. Nemirovskii, Y. Ye, *Interior-Point Polynomial Algorithms in Convex Programming*, vol. 13 (SIAM, Philadelphia, 1994)
174. G. Obozinski, B. Taskar, M.I. Jordan, Joint covariate selection and joint subspace selection for multiple classification problems. Stat. Comput. **20**(2), 231–252 (2010)
175. M.R. Osborne, B. Presnell, B.A. Turlach, A new approach to variable selection in least squares problems. IMA J. Numer. Anal. **20**(3), 389–403 (2000)
176. S. Oymak, A. Jalali, M. Fazel, Y.C. Eldar, B. Hassibi, Simultaneously structured models with application to sparse and low-rank matrices. (2012) [arXiv preprint arXiv:1212.3753]
177. N. Parikh, S.P. Boyd, Proximal algorithms. Found. Trends Optim. **1**(3), 123–231 (2013)
178. G.B. Passty, Ergodic convergence to a zero of the sum of monotone operators in hilbert space. J. Math. Anal. Appl. **72**(2), 383–390 (1979)
179. Y.C. Pati, R. Rezaiifar, P.S. Krishnaprasad, Orthogonal Matching Pursuit: recursive function approximation with applications to wavelet decomposition, in *Conference on Signals, Systems and Computers* (IEEE, Pacific Grove, 1993), pp. 40–44
180. J-C. Pesquet, A. Benazza-Benyahia, C. Chaux, A SURE approach for digital signal/image deconvolution problems. IEEE Trans. Signal Process. **57**(12), 4616–4632 (2009)
181. G. Peyré, M.J. Fadili, J.-L. Starck, Learning the morphological diversity. SIAM J. Imaging Sci. **3**(3), 646–669 (2010)
182. G. Peyré, J. Fadili, C. Chesneau, Adaptive structured block sparsity via dyadic partitioning, in *Proc. EUSIPCO 2011* (2011), pp. 1455–1459
183. H. Raguet, J. Fadili, G. Peyré, Generalized forward–backward splitting. SIAM J. Imaging Sci. **6**(3), 1199–1226 (2013)
184. S. Ramani, T. Blu, M. Unser, Monte-Carlo SURE: a black-box optimization of regularization parameters for general denoising algorithms. IEEE Trans. Image Process. **17**(9), 1540–1554 (2008)

185. S. Ramani, Z. Liu, J. Rosen, J.-F. Nielsen, J.A. Fessler, Regularization parameter selection for nonlinear iterative image restoration and mri reconstruction using GCV and SURE-based methods. IEEE Trans. Image Process. **21**(8), 3659–3672 (2012)
186. S. Ramani, J. Rosen, Z. Liu, J.A. Fessler, Iterative weighted risk estimation for nonlinear image restoration with analysis priors, in *Computational Imaging X*, vol. 8296 (2012), pp. 82960N–82960N–12
187. B.D. Rao, K. Kreutz-Delgado, An affine scaling methodology for best basis selection. IEEE Trans. Signal Process. **47**(1), 187–200 (1999)
188. B. Recht, M. Fazel, P.A. Parrilo, Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. SIAM Rev. **52**(3), 471–501 (2010)
189. R. Refregier, F. Goudail, *Statistical Image Processing Techniques for Noisy Images - An Application Oriented Approach* (Kluwer, New York, 2004)
190. E. Resmerita, Regularization of ill-posed problems in Banach spaces: convergence rates. Inverse Prob. **21**(4), 1303 (2005)
191. E. Richard, F. Bach, J.-P. Vert, Intersecting singularities for multi-structured estimation, in *International Conference on Machine Learning*, Atlanta, États-Unis (2013)
192. R.T. Rockafellar, R. Wets, *Variational Analysis*, vol. 317 (Springer, Berlin, 1998)
193. M. Rudelson, R. Vershynin, On sparse reconstruction from Fourier and Gaussian measurements. Commun. Pure Appl. Math. **61**(8), 1025–1045 (2008)
194. L.I. Rudin, S. Osher, E. Fatemi, Nonlinear total variation based noise removal algorithms. Phys. D Nonlinear Phenom. **60**(1), 259–268 (1992)
195. F. Santosa, W.W. Symes, Linear inversion of band-limited reflection seismograms. SIAM J. Sci. Stat. Comput. **7**(4), 1307–1330 (1986)
196. O. Scherzer, M. Grasmair, H. Grossauer, M. Haltmeier, F. Lenzen, *Variational Methods in Imaging*, vol. 167 (Springer, New York, 2009)
197. I.W. Selesnick, M.A.T. Figueiredo, Signal restoration with overcomplete wavelet transforms: comparison of analysis and synthesis priors, in *Proceedings of SPIE*, vol. 7446 (2009), p. 74460D
198. S. Shalev-Shwartz, A. Gonen, O. Shamir, Large-scale convex minimization with a low-rank constraint, in *ICML* (2011)
199. X. Shen, J. Ye, Adaptive model selection. J. Am. Stat. Assoc. **97**(457), 210–221 (2002)
200. Solo, V., Ulfarsson, M.: Threshold selection for group sparsity, in *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)* (IEEE, Dallas, 2010), pp. 3754–3757
201. M.V. Solodov, A class of decomposition methods for convex optimization and monotone variational inclusions via the hybrid inexact proximal point framework. Optim. Methods Softw. **19**(5), 557–575 (2004)
202. D.A. Spielman, H. Wang, J. Wright, Exact recovery of sparsely-used dictionaries. J. Mach. Learn. Res. **23**, 1–35 (2012)
203. N. Srebro, Learning with matrix factorizations. Ph.D. thesis, MIT, 2004
204. J.-L. Starck, M. Elad, D.L. Donoho, Image decomposition via the combination of sparse representatntions and variational approach. IEEE Trans. Image Process. **14**(10), 1570–1582 (2005)
205. J.-L. Starck, F. Murtagh, J.M. Fadili, *Sparse Image and Signal Processing: Wavelets, Curvelets, Morphological Diversity* (Cambridge University Press, Cambridge, 2010)
206. G. Steidl, J. Weickert, T. Brox, P. Mrázek, M. Welk, On the equivalence of soft wavelet shrinkage, total variation diffusion, total variation regularization, and sides. SIAM J. Numer. Anal. **42**(2), 686–713 (2004)
207. C.M. Stein, Estimation of the mean of a multivariate normal distribution. Ann. Stat. **9**(6), 1135–1151 (1981)
208. T. Strohmer, R.W. Heath Jr., Grassmannian frames with applications to coding and communication. Appl. Comput. Harmon. Anal. **14**(3), 257–275 (2003)
209. C. Studer, W. Yin, R.G. Baraniuk, Signal representations with minimum ℓ_∞-norm, in *Proc. 50th Ann. Allerton Conf. on Communication, Control, and Computing* (2012)

210. B.F. Svaiter, H. Attouch, J. Bolte, Convergence of descent methods for semi-algebraic and tame problems: proximal algorithms, forward-backward splitting, and regularized gauss-seidel methods. Math. Program. Ser. A **137**(1–2), 91–129 (2013)
211. H.L. Taylor, S.C. Banks, J.F. McCoy, Deconvolution with the ℓ_1 norm. Geophysics **44**(1), 39–52 (1979)
212. R. Tibshirani, Regression shrinkage and selection via the Lasso. J. R. Stat. Soc. Ser. B. Methodol. **58**(1), 267–288 (1996)
213. R.J. Tibshirani, J. Taylor, The solution path of the generalized Lasso. Ann. Stat. **39**(3), 1335–1371 (2011)
214. R.J. Tibshirani, J. Taylor, Degrees of freedom in Lasso problems. Ann. Stat. **40**(2), 1198–1232 (2012)
215. R. Tibshirani, M. Saunders, S. Rosset, J. Zhu, K. Knight, Sparsity and smoothness via the fused Lasso. J. R. Stat. Soc. Ser. B Stat. Methodol. **67**(1), 91–108 (2005)
216. A.N. Tikhonov, Regularization of incorrectly posed problems. Soviet Math. Dokl. **4**, 1624–1627 (1963)
217. A.N. Tikhonov, Solution of incorrectly formulated problems and the regularization methods. Soviet Math. Dokl. **4**, 1035–1038 (1963)
218. A.N. Tikhonov, V. Arsenin, *Solutions of Ill-Posed Problems*. (V. H. Winston and Sons, Washington, 1977)
219. A.M. Tillman, M.E. Pfetsh., The computational complexity of the restricted isometry property, the nullspace property, and related concepts in compressed sensing. IEEE Trans. Inf. Theory **60**(2), 1248–1259 (2014)
220. J. Tropp, Convex recovery of a structured signal from independent random linear measurements, in *Sampling Theory, a Renaissance* (Birkhäuser, Basel, 2014)
221. J.A. Tropp, Just relax: convex programming methods for identifying sparse signals in noise. IEEE Trans. Inf. Theory **52**(3), 1030–1051 (2006)
222. P. Tseng, Alternating projection-proximal methods for convex programming and variational inequalities. SIAM J. Optim. **7**(4), 951–965 (1997)
223. P. Tseng, S. Yun, A coordinate gradient descent method for nonsmooth separable minimization. Math. Prog. Ser. B **117** (2009)
224. B.A. Turlach, W.N. Venables, S.J. Wright, Simultaneous variable selection. Technometrics **47**(3), 349–363 (2005)
225. S. Vaiter, C. Deledalle, G. Peyré, J. Fadili, C. Dossal, Degrees of freedom of the group Lasso, in *ICML'12 Workshops* (2012), pp. 89–92
226. S. Vaiter, C. Deledalle, G. Peyré, J. Fadili, C. Dossal, The degrees of freedom of partly smooth regularizers. Technical report, Preprint Hal-00768896 (2013)
227. S. Vaiter, C.-A. Deledalle, G. Peyré, C. Dossal, J. Fadili, Local behavior of sparse analysis regularization: applications to risk estimation. Appl. Comput. Harmon. Anal. **35**(3), 433–451 (2013)
228. S. Vaiter, M. Golbabaee, M.J. Fadili, G. Peyré, Model selection with low complexity priors. Technical report (2013) [arXiv preprint arXiv:1307.2342]
229. S. Vaiter, G. Peyré, C. Dossal, M.J. Fadili, Robust sparse analysis regularization. IEEE Trans. Inf. Theory **59**(4), 2001–2016 (2013)
230. S. Vaiter, G. Peyré, J.M. Fadili, C.-A. Deledalle, C. Dossal, The degrees of freedom of the group Lasso for a general design, in *Proc. SPARS'13* (2013)
231. S. Vaiter, G. Peyré, M.J. Fadili, Robust polyhedral regularization, in *Proc. SampTA* (2013)
232. S. Vaiter, G. Peyré, J. Fadili, Model consistency of partly smooth regularizers. Technical report, Preprint Hal-00987293 (2014)
233. D. Van De Ville, M. Kocher, SURE-based Non-Local Means. IEEE Signal Process. Lett. **16**(11), 973–976 (2009)
234. D. Van De Ville, M. Kocher, Non-local means with dimensionality reduction and SURE-based parameter selection. IEEE Trans. Image Process. **9**(20), 2683–2690 (2011)
235. L. van den Dries, C. Miller, Geometric categories and o-minimal structures. Duke Math. J. **84**(2), 497–540, 08 (1996)

236. J.E. Vogt, V. Roth, A complete analysis of the $\ell_{1,p}$ group-Lasso, in *International Conference on Machine Learning* (2012)
237. C. Vonesch, S. Ramani, M. Unser, Recursive risk estimation for non-linear image deconvolution with a wavelet-domain sparsity constraint, in *International Conference on Image Processing* (IEEE, San Diego, 2008), pp. 665–668
238. B.C. Vũ. A splitting algorithm for dual monotone inclusions involving cocoercive operators. Adv. Comput. Math. **38**, 1–15 (2011)
239. M.J. Wainwright, Sharp thresholds for noisy and high-dimensional recovery of sparsity using ℓ^1-constrained quadratic programming (Lasso). IEEE Trans. Inf. Theory **55**(5), 2183–2202 (2009)
240. S.J. Wright, Identifiable surfaces in constrained optimization. SIAM J. Control Optim. **31**(4), 1063–1079 (1993)
241. J. Ye, On measuring and correcting the effects of data mining and model selection. J. Am. Stat. Assoc. **93**, 120–131 (1998)
242. M. Yuan, Y. Lin, Model selection and estimation in regression with grouped variables. J. R. Stat. Soc. Ser. B Stat. Methodol. **68**(1), 49–67 (2005)
243. P. Zhao, B. Yu, On model selection consistency of Lasso. J. Mach. Learn. Res. **7**, 2541–2563 (2006)
244. P. Zhao, G. Rocha, B. Yu, The composite absolute penalties family for grouped and hierarchical variable selection. Ann. Stat. **37**(6A), 3468–3497 (2009)
245. H. Zou, T. Hastie, R. Tibshirani, On the "degrees of freedom" of the Lasso. Ann. Stat. **35**(5), 2173–2192 (2007)

Part II
Finite and Structured Frames

Chapter 4
Noise-Shaping Quantization Methods for Frame-Based and Compressive Sampling Systems

Evan Chou, C. Sinan Güntürk, Felix Krahmer, Rayan Saab, and Özgür Yılmaz

Abstract Noise shaping refers to an analog-to-digital conversion methodology in which quantization error is arranged to lie mostly outside the signal spectrum by means of oversampling and feedback. Recently it has been successfully applied to more general redundant linear sampling and reconstruction systems associated with frames as well as non-linear systems associated with compressive sampling. This chapter reviews some of the recent progress in this subject.

4.1 Introduction

Source coding via quantized linear representations, also known as transform coding, is a classical and well-studied subject. Yet it is poorly understood outside the simple

E. Chou (✉)
Courant Institute, 251 Mercer Street, New York, NY 10012, USA

Google, New York, NY 10011, USA
e-mail: chou@cims.nyu.edu

C.S. Güntürk
Courant Institute, 251 Mercer Street, New York, NY 10012, USA
e-mail: gunturk@cims.nyu.edu

F. Krahmer
Department of Mathematics, Technische Universität München,
Boltzmannstr. 3, 85748 Garching/Munich, Germany
e-mail: felix.krahmer@tum.de

R. Saab
University of California, San Diego, 9500 Gilman Drive,
Dept. 0112, La Jolla, CA 92093, USA
e-mail: rsaab@ucsd.edu

Ö. Yılmaz
University of British Columbia, 1984 Mathematics Road, Vancouver, BC, Canada V6T 1Z2
e-mail: oyilmaz@math.ubc.ca

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_4

setting of orthogonal transforms, namely, for frame-based representations. The same can also be said for partially nonlinear representations such as those based on compressive sampling. The basic reason for the difficulty in solving the quantization problem for these more general sampling and reconstruction systems is the lack of an analog of Parseval's identity which, more or less, dictates the best quantization strategy for orthogonal systems. While some kind of basic reconstruction stability can be ensured relatively easily, these results do not offer correct rate-distortion trade-offs because of their inefficiency in utilizing redundancy, especially under constraints that do not allow for high-resolution quantization.

Redundancy is a key concept of frame-based as well as compressive sampling systems. It can be understood in terms of the sampling process (e.g., what part of the coefficient space is taken up with the actual measurements) or in terms of the reconstruction process (e.g., which perturbations of the measurements have the smallest effect on the reconstruction). Efficient encoding via the first approach is generally not practical because codewords cannot be easily placed arbitrarily in the coefficient space. Indeed, quantized measurements are typically required to lie on a finite rectangular grid. An alternative approach is then to seek ways of arranging the quantization error in the coefficient space to lie in directions that are away from the actual measurements, typically by means of some feedback process. *Noise shaping* is the generic name of this quantization methodology. It has its roots in sigma-delta modulation, which is used for oversampled analog-to-digital (A/D) conversion [9, 25, 34, 41].

Let us explain the philosophy of noise shaping in more concrete terms. In both frame-based and compressive sampling systems, we have a linear sampling operator Φ that can be inverted on a given space $\mathcal{X}$ of signals using some (possibly nonlinear) reconstruction operator Ψ. Given a signal $x \in \mathcal{X}$ and its sampled version $y = \Phi x$, ordinarily we recover x exactly (or approximately, as in compressive sampling) as $\Psi(y)$. In the context of this paper, quantization of y will mean replacing it with a vector q which is of the same dimensionality as y and whose entries are chosen from some given alphabet $\mathcal{A}$. The goal is to choose q so that the approximate reconstruction $x^{\#} := \Psi(q)$ is as close to x as possible as x varies over $\mathcal{X}$.

In the context of finite frames, Φ is a full-rank $m \times k$ matrix where $m > k$, and Ψ is any left inverse of Φ. The rows of Φ form the *analysis frame* and the columns of Ψ form a *synthesis frame* dual to this frame. With $y = \Phi x$ and $x = \Psi y$ as above, when y is replaced by a quantized vector q, the reconstruction error $e := x - x^{\#}$ is equal to $\Psi(y - q)$. Therefore the correct strategy to reduce the size of e is not to minimize the Euclidean norm $\|y - q\|$ as memoryless scalar quantization (MSQ) does, but to minimize the semi-norm $|y - q|_{\Psi} := \|\Psi(y - q)\|$. In other words, we seek $q \in \mathcal{A}^m$ so that the quantization "noise" $y - q$ is close to $\ker(\Psi)$ in the above sense. This is the basic principle of noise shaping. How this goal can be achieved (approximately), i.e., the actual process of noise shaping, as well as what noise shaping can offer for source coding are nontrivial questions that will be addressed in this article.

While the basic principle of noise shaping is formulated above for linear sampling and reconstruction systems, its philosophy extends to compressive sampling systems where the reconstruction operator is generally nonlinear. The simplest

connection is made by considering strictly sparse signals. Let Σ_k^N denote the nonlinear space of N-dimensional vectors which have no more than k nonzero entries. In the context of compressive sampling, Φ is an $m \times N$ matrix where $m \ll N$, which means that the sampling process is lossy for the whole of $\mathbb{R}^N$. However, note that Σ_k^N is the union of (a large number of) k-dimensional linear subspaces on each of which Φ acts like a frame once $m > k$. This observation opens up the possibility of noise shaping. Indeed, fixing any one of these subspaces V, we can envision a noise shaping process associated with any of the linear inverses (duals) of Φ on V. However, it is not clear how one might organize all of these individual noise shaping processes, especially given that these subspaces are not directly available to the quantizer. What comes to the rescue is the notion of an *alternative dual*. While we formulated noise shaping above as matching the quantization operator to a given dual frame, it is also possible to consider matching the dual frame to a given quantization operator. This results in the possibility of "universal" quantization processes (i.e., independent of the signal subspace) which become noise-shaping processes for suitable alternative duals. Even though finding these suitable alternative duals may require extracting information about the signal subspace, this duty purely belongs to the decoder and not the quantizer.

This article is organized as follows. In Section 4.2, we review the basics of classical noise shaping in the setting of sigma-delta ($\Sigma\Delta$) modulation. In Section 4.3, we extend the formulation of noise shaping and introduce various notions of alternative duals for noise shaping in the setting of frames, followed by their performance analysis for random frames in Section 4.4. We then discuss noise-shaping quantization methods for compressive sampling in Section 4.5.

4.2 Classical noise shaping: Sigma-Delta Modulation

The Shannon-Nyquist sampling theorem for bandlimited functions provides the natural framework of conventional A/D conversion systems. With the Fourier transform normalized according to the "ordinary-frequency" convention

$$\hat{x}(\xi) := \int_{-\infty}^{\infty} x(t) e^{-2\pi i \xi t} \, \mathrm{d}t,$$

let us define the space $\mathcal{B}_\Omega$ of bandlimited functions to be all x in $L^2(\mathbb{R})$ such that $\hat{x}$ is supported in $[-\Omega, \Omega]$. The classical sampling theorem says that any $x \in \mathcal{B}_\Omega$ can be reconstructed perfectly from its time samples $(x(n\tau))_{n\in\mathbb{Z}}$ according to the formula

$$x(t) = \tau \sum_{n\in\mathbb{Z}} x(n\tau)\psi(t - n\tau), \tag{4.1}$$

where $\tau \leq \tau_{\text{crit}} := \frac{1}{2\Omega}$, and ψ is any function in $L^2(\mathbb{R})$such that

$$\hat{\psi}(\xi) = \begin{cases} 1, & |\xi| \leq \Omega, \\ 0, & |\xi| > \frac{1}{2\tau}. \end{cases} \tag{4.2}$$

Hence, if we define the sampling operator $(\Phi x)_n := x(n\tau)$ and the reconstruction operator $\Psi(u) := \tau \sum u_n \psi(\cdot - n\tau)$ (on any space it makes sense), then Ψ is a left inverse of Φ on $\mathcal{B}_\Omega$ when τ and ψ satisfy the conditions stated above.

The value $\rho := 1/\tau$ is called the sampling rate, and $\rho_{\text{crit}} := 1/\tau_{\text{crit}} = 2\Omega$ is called the critical (or Nyquist) sampling rate. Their ratio given by

$$\lambda := \frac{\rho}{\rho_{\text{crit}}} \tag{4.3}$$

is called the *oversampling ratio*. According to the value of λ, A/D converters are broadly classified as Nyquist-rate converters ($\lambda \approx 1$) or oversampling converters ($\lambda \gg 1$).

Nyquist-rate converters set their sampling rate ρ slightly above the critical frequency 2Ω so that ψ may be chosen to decay rapidly enough to ensure absolute summability of (4.1). Given any quantization alphabet $\mathcal{A}$, the (nearly) optimal quantization strategy in this (nearly) orthogonal setting is memoryless scalar quantization (MSQ). This means that each sample $y_n := x(n\tau)$ is rounded to the nearest quantization level $q_n \in \mathcal{A}$. This process is also referred to as pulse-code modulation (PCM). If each sample is quantized with error no more than δ, i.e., $\|y - q\|_\infty \leq \delta$, then the error signal

$$e(t) := x(t) - (\Psi q)(t) = \tau \sum_{n \in \mathbb{Z}} \big(y_n - q_n\big)\psi(t - n\tau) \tag{4.4}$$

obeys the bound $\|e\|_{L^\infty} \leq C\delta$ where C is independent of δ. This is essentially the best error bound one can expect for Nyquist-rate converters. Because setting δ very small is costly, Nyquist-rate converters are not very suitable for signals that require high-fidelity such as audio signals.

Oversampling converters are designed to take advantage of the redundancy in the representation (4.1) when $\tau < \tau_{\text{crit}}$. In this case, the interpolation operator Ψ has a kernel which gets bigger as $\tau \to 0$. Indeed, let $\hat{\psi}(\xi) = 0$ for $|\xi| > \Omega_0$. It is easily seen that $\Psi u = 0$ if

$$\sum_{n \in \mathbb{Z}} u_n e^{2\pi i n \xi} = 0 \text{ for } |\xi| < \tau\Omega_0. \tag{4.5}$$

This means that even though $y - q$ may be large everywhere, $e = \Psi(y - q)$ can be very small if $y - q$ can be arranged to be spectrally disjoint from the (discretized) reconstruction kernel ψ. This is the concrete form of noise shaping that we briefly discussed in the Introduction.

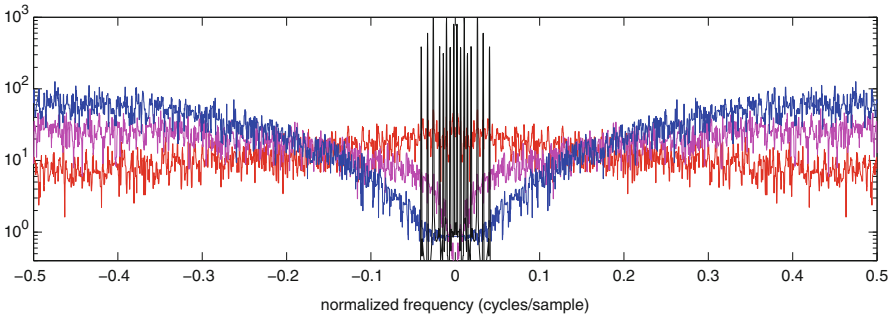


Fig. 4.1 Illustration of classical noise shaping via $\Sigma\Delta$ modulation: The superimposed Fourier spectra of a bandlimited signal (in black), and the quantization error signals using MSQ (in red), 1st order $\Sigma\Delta$ modulation (in magenta), and 2nd order $\Sigma\Delta$ modulation (in blue).

The main focus of an oversampling A/D converter is on its quantization algorithm, which has to be non-local to be useful, but also causal so that it can be implemented in real time. The assignment of each q_n will therefore depend on y_n as well as a set of values (the states) that can be kept in an analog circuit memory, while meeting the spectral constraints on $y - q$ as described in the previous section. $\Sigma\Delta$ modulators operate according to these principles.

As can be seen in (4.5), the kernel of Ψ consists of high-pass sequences. Hence the primary objective of $\Sigma\Delta$ modulation is to arrange the quantization error $y-q$ to be an approximate high-pass sequence (see Fig. 4.1). This objective can be realized by setting up a difference equation, the so-called *canonical* $\Sigma\Delta$ equation, of the form

$$y - q = \Delta^r u, \tag{4.6}$$

where Δ denotes the finite difference operator defined by

$$(\Delta w)_n := w_n - w_{n-1}, \tag{4.7}$$

r denotes the "order" of the scheme, and u is an appropriate auxiliary sequence called the *state sequence*. This equation does not imply anything about q without any constraint on u. The most useful constraint turns out to be boundedness.

In practice, the boundedness of u in (4.6) has to be attained through a recursive algorithm. This means that given any input sequence (y_n), the q_n are found by a given "quantization rule" of the form

$$q_n = F(u_{n-1}, u_{n-2}, \dots, y_n, y_{n-1}, \dots), \tag{4.8}$$

and the u_n are updated via

$$u_n = \sum_{k=1}^{r} (-1)^{k-1} \binom{r}{k} u_{n-k} + y_n - q_n, \tag{4.9}$$

which is a restatement of (4.6). In electrical engineering, such a recursive procedure for quantization is called "feedback quantization" due to the role q_n plays as a feedback control term in (4.9). The role of the quantization rule F is to keep the system *stable*, i.e., u bounded.

Stability is a crucial property. Indeed, it was shown in [13] that a stable rth order scheme results in the error bound

$$\|e\|_{L^\infty} \leq \|u\|_{\ell^\infty} \|\psi^{(r)}\|_{L^1} \tau^r, \tag{4.10}$$

where $\psi^{(r)}$ denotes the rth order derivative of ψ. The implicit Ω- and the explicit τ-dependence of this estimate can be replaced with a single λ-dependence by setting $\psi(t) := \Omega \psi_0(\Omega t)$ where the prototype $\hat{\psi}_0(\xi)$ equals 1 on $[-1, 1]$ and vanishes for $|\xi| \geq 1 + \epsilon_0$, with $\epsilon_0 > 0$ fixed. Let $C_0 := \|\psi_0\|_{L^1}$. Bernstein's inequality applied to ψ yields

$$\|e\|_{L^\infty} \leq C_0 \|u\|_{\ell^\infty} \pi^r (1 + \epsilon_0)^r \lambda^{-r}, \text{ for all } \lambda > 1 + \epsilon_0. \tag{4.11}$$

With this error bound, there are two goals in progression. The first is to keep u bounded and the second is to keep the bound small. Ultimately, the best strategy is to have, for each r, a quantization rule yielding a stable rth order scheme, and then for any given λ, to choose the best one (i.e., the one with the least error bound). This task is significantly complicated by the fact that the bound on u has a strong dependence on r, especially for small quantization alphabets $\mathcal{A}$. In general it is not possible to expect this dependence to be less than $(cr)^r$ for some constant c that depends on the given amplitude range μ for x. This growth order is also what is needed to ensure that the reconstruction error decays exponentially, i.e., as $2^{-p\lambda}$, as a function of λ, which is the best possible due to Kolmogorov entropy estimates for bandlimited functions [21]. The rate p of exponential decay that is achievable by the resulting family of schemes is inversely proportional to c, and gets worse as μ is increased. The question of best achievable accuracy for oversampling converters in this setting remains open. Currently, the best result in the one-bit case with $\mathcal{A} = \{-1, 1\}$ yields $\|e\|_{L^\infty} = O(2^{-p\lambda})$ where $p = \pi/(6e^2 \log 2) \approx 0.1$, and $\mu \approx 0.06$. Higher values of p can be achieved with more levels in $\mathcal{A}$. For example, if $\mathcal{A} = \{-1, 0, 1\}$, then p rises to 0.15 and μ to 0.25 [15]. These are rigorously proven bounds and the actual behavior of the error based on numerical experiments appears to be better. For the details of the quantization rules which result in these exponentially accurate $\Sigma\Delta$ modulators, see [15, 21]. It has also been shown that no matter how the bits are assigned the rate of the exponential decay cannot match that of Nyquist-rate conversion [28].

4.3 Generalized Noise-shaping Operators and Alternative Duals of Frames for Noise Shaping

In this section, we will generalize the classical theory of $\Sigma\Delta$ modulation to more general noise-shaping quantizers as well as sampling and reconstruction systems. For conceptual clarity, we will separate the process of noise shaping from the processes of sampling and reconstruction. While we will present these generalizations in a finite-dimensional setting, extensions to infinite-dimensional settings are usually possible. We will also discuss the notion of alternative duals of frames which are associated with noise-shaping quantizers.

4.3.1 A general framework of noise shaping

The canonical $\Sigma\Delta$ equation we saw in (4.6) is a special case of a more general framework of noise shaping. Let $\mathcal{A}$ be a finite quantization alphabet and J be a compact interval in $\mathbb{R}$. Let $h = (h_j)_{j\geq 0}$ be a given sequence, finite or infinite, where $h_0 = 1$. By a noise-shaping quantizer with the transfer filter h, we mean any sequence $Q = (Q_m)_1^\infty$ of maps $Q_m : J^m \to \mathcal{A}^m$, $m \in \mathbb{N}$, where for each $y \in J^m$, the output $q := Q_m(y)$ satisfies

$$y - q = h * u \tag{4.12}$$

where $u \in \mathbb{R}^m$ and $\|u\|_\infty \leq C$ for some constant C which is independent of m. Here $h * u$ refers to the (finite) convolution of h and u defined by

$$(h * u)_n := \sum_{j\geq 0} h_j u_{n-j}, \quad 1 \leq n \leq m,$$

where it is assumed that $u_n := 0$ for $n \leq 0$. Without any reference to a sampling or a reconstruction operator, noise shaping in this setting refers to the fact that the "quantization noise" $y - q$ is spectrally aligned with h. Note that the operator $H : u \mapsto h * u$ is invertible on $\mathbb{R}^m$ for any m, and therefore given any y and q, there exists $u \in \mathbb{R}^m$ which satisfies (4.12); this is trivial. However, the requirement that $\|u\|_\infty$ must be controlled uniformly in m imposes restrictions on what q can be for a given y; these solutions are certainly non-trivial to find and may not always exist.

The operator H above (defined as convolution by h) is a lower triangular Toeplitz matrix with unit diagonal. With this view, let us relax the notion of a noise-shaping quantizer and assume that H is any lower triangular $m \times m$ matrix with unit diagonal. We will refer to H as a noise-shaping transfer operator where the associated noise-shaping relation is given by

$$y - q = Hu. \tag{4.13}$$

Suppose we are given a sequence $(H_m)_1^\infty$ of $m \times m$ noise-shaping transfer operators. In this general setting, we say that an associated sequence $(Q_m)_1^\infty$ of quantizer maps (for which $q := Q_m(y)$ and u is determined by (4.13)) achieves noise shaping for (H_m), J, and $\mathcal{A}$, if $\|u\|_\infty \leq C$ for some constant C independent of m. A slightly weaker assumption is to only require that $\|u\|_\infty = o(\|H_m^{-1}\|_{\infty\to\infty})$, though we shall not need to work in this generality in this paper.

In many applications, one works with $(H_m)_1^\infty$ which are "progressive" (also called "nested") in the sense that

$$P_m \circ H_{m+1} \circ P_{m+1} = H_m \circ P_m,$$

where P_m is the restriction of a vector to its first m coordinates. Convolution is a standard example. In this case, it may be natural to require that the $(Q_m)_1^\infty$ are progressive as well. The classical $\Sigma\Delta$ modulation we saw in Section 4.2 is of this type. However, our general formulation does not impose progressiveness.

As indicated earlier, noise-shaping quantizers provide non-trivial solutions to (4.13) and therefore do not exist unconditionally, though under certain suitable assumptions on H, J, and $\mathcal{A}$, they exist and can be implemented via recursive algorithms. The simplest is the (non-overloading) *greedy quantizer* whose general formulation is given below:

Proposition 1. *Let $\mathcal{A} := \mathcal{A}_{L,\delta}$ denote the arithmetic progression in $\mathbb{R}$ which is of length L, spacing 2δ, and symmetric about 0. Assume that $H = I - \tilde{H}$, where $\tilde{H}$ is strictly lower triangular, and $\mu \geq 0$ such that $\|\tilde{H}\|_{\infty\to\infty} + \mu/\delta \leq L$. Suppose $\|y\|_\infty \leq \mu$. For each $n \geq 1$, let*

$$q_n := \text{round}_{\mathcal{A}}\left(y_n + \sum_{j=1}^{n-1} \tilde{H}_{n,n-j} u_{n-j}\right)$$

and

$$u_n := y_n + \sum_{j=1}^{n-1} \tilde{H}_{n,n-j} u_{n-j} - q_n.$$

Then the resulting q satisfies (4.13) with $\|u\|_\infty \leq \delta$.

This quantizer is called greedy because for all n, the selection of q_n over $\mathcal{A}$ is made so as to minimize $|u_n|$. The proof of this basic result follows easily by induction once we note that for any $w \in [-L\delta, L\delta]$, we have $|w - \text{round}_{\mathcal{A}}(w)| \leq \delta$, hence the scalar quantizer $\text{round}_{\mathcal{A}}$ is not overloaded. For details, see [11]. Note that the greedy quantizer is progressive if $(H_m)_1^\infty$ is a progressive sequence of noise-shaping transfer operators. In the special case $Hu = h * u$ where $h_0 = 1$, we simply have $\|\tilde{H}\|_{\infty\to\infty} = \|h\|_1 - 1$. This special case is well-known and widely utilized (e.g. [9, 21, 34, 41]).

4.3.2 Canonical duals of frames for noise shaping

The earliest works on noise-shaping quantization in the context of finite frames used $\Sigma\Delta$ quantization and focused on canonical duals for reconstruction. Before we begin our discussion of these contributions we remind the reader of our convention: we identify an analysis frame with (the rows of) its analysis operator and a synthesis frame with (the columns of) its synthesis operator.

Let Φ be a finite frame and $y = \Phi x$ be the frame measurements of a given signal x. Assume that we quantize y using a noise-shaping quantizer with transfer operator H. Any left-inverse (dual) Ψ of Φ gives

$$x - \Psi q = \Psi(y - q) = \Psi H u. \tag{4.14}$$

Using this expression, and specializing to the case of first order $\Sigma\Delta$ quantization, i.e., $H = D$ where D is the lower bidiagonal matrix whose diagonal entries are 1 and subdiagonal entries are -1, [3] observed that the reconstruction error can be bounded as

$$\|x - \Psi q\|_2 \leq \|u\|_\infty \sum_{j=1}^{m} \|(\Psi D)_j\|_2 \tag{4.15}$$

where $(\Psi D)_j$ denotes the jth column of ΨD. This led [3] to introduce the notion of frame variation

$$\mathrm{Var}(\Psi) := \sum_{j=1}^{m} \|\psi_j - \psi_{j+1}\|_2 \tag{4.16}$$

with ψ_j denoting the jth column of Ψ and ψ_{m+1} defined to be zero. Using normalized tight-frames, i.e., frames Φ for which $\Phi^*\Phi = (m/k)I$, this resulted in the error bound

$$\|x - \Phi^\dagger q\|_2 \leq \frac{k}{m}\|u\|_\infty \mathrm{Var}(\Phi^*), \tag{4.17}$$

where $\Psi = \Phi^\dagger$ denotes the *canonical dual* of Φ defined (for an arbitrary frame Φ) by

$$\Phi^\dagger := (\Phi^*\Phi)^{-1}\Phi^*. \tag{4.18}$$

Subsequently, similarly defined higher-order frame variations were used to study the behavior of higher-order $\Sigma\Delta$ schemes (e.g., in [2] and [6]) with corresponding generalizations of (4.17) and the conclusion that frames with lower variations lead to better error bounds. This motivated considering frames obtained via uniform sampling of smooth curves in $\mathbb{R}^k$ (called *frame paths*). As it turned out, however,

this type of analysis based on frame-variation bounds does not provide higher-order reconstruction accuracy unless the frame path terminates smoothly. Smooth termination of the frame path is not available for most of the commonly encountered frames, and finding frames with this property can be challenging. Indeed, designing such frames was a main contribution of [6] which showed a reconstruction error bound decaying as m^r for rth order $\Sigma\Delta$ quantization of measurements using these frames.

In practice, however, one must often work with a given frame rather than design a frame of their choosing. In such cases there are frames, sampled from smooth curves, for which reconstructing with the canonical dual yields reconstruction error that is *lower bounded* by a term behaving like m^{-1}, regardless of the $\Sigma\Delta$ scheme's order $r \geq 3$ (see, [31] for the details). Consequently, to achieve better error decay rates one must seek either different quantization or different reconstruction schemes. We will consider both routes to improving the error bounds in what follows.

4.3.3 Alternative duals of frames for noise shaping

The discussion in Section 4.3.2 was based on canonical duals and it involved a particular method to bound the 2-norm of the reconstruction error $x - \Psi q$, assuming u is bounded in the ∞-norm. It is possible to significantly improve the reconstruction accuracy by allowing for more general duals, here called *alternative duals*. To explain this route, we return to the general noise-shaping quantization relation (4.14). We assume again that u is known to be bounded in the ∞-norm, which is essentially the only type of bound available. Hence, the most natural reconstruction error bound is given by

$$\|x - \Psi q\|_2 \leq \|\Psi H\|_{\infty\to 2}\|u\|_\infty. \tag{4.19}$$

With this bound, the natural objective would be to employ an alternative dual Ψ of Φ which minimizes $\|\Psi H\|_{\infty\to 2}$. An explicit solution for this problem is not readily available mainly because there is no easily computable expression for $\|A\|_{\infty\to 2}$ for a general $k \times m$ matrix A, so we replace it by a simpler upper bound. In fact, this was already done in (4.15) because we have

$$\|A\|_{\infty\to 2} \leq \sum_{j=1}^{m} \|A_j\|_2 \tag{4.20}$$

where again A_j denotes the jth column of A. (This upper bound is also known to be the $L_{2,1}$-norm of A.) Another such bound which is often (but not always) better is given by

$$\|A\|_{\infty\to 2} \leq \sqrt{m}\|A\|_{2\to 2}. \tag{4.21}$$

(Indeed, for a large random matrix with standard Gaussian entries, the upper bound in (4.21) behaves as $m+\sqrt{mk}$ whereas that of (4.20) behaves as $m\sqrt{k}$. Both of these upper bounds are easily seen to be less than $\sqrt{m}\|A\|_{\mathrm{Fr}}$, however.)

With this upper bound, we minimize $\|\Psi H\|_{2\to 2}$ over all alternative duals Ψ of Φ. Then an explicit solution is available and is given by

$$\Psi_{H^{-1}} := (H^{-1}\Phi)^{\dagger}H^{-1}. \tag{4.22}$$

This idea was initially introduced specifically for $\Sigma\Delta$ quantization [4, 31] with the choice $H = D^r$. The resulting alternative duals were called *Sobolev duals* and will be discussed in the next subsection. The above generalized version was stated in [23] where the notation Ψ_H and the term "H-dual" were introduced for the right hand side of (4.22), but because of a further generalization we will discuss in Section 4.3.3.3, we find it more appropriate to use the label H^{-1}.

Note that the no noise-shaping case of $H = I$ yields the canonical dual. In general, we have

$$\|\Psi_{H^{-1}}H\|_{2\to 2} = \|(H^{-1}\Phi)^{\dagger}\|_{2\to 2} = \frac{1}{\sigma_{\min}(H^{-1}\Phi)}$$

so that (4.19) and (4.21) yield the error bound

$$\|x - \Psi_{H^{-1}}q\|_2 \le \frac{\sqrt{m}}{\sigma_{\min}(H^{-1}\Phi)}\|u\|_{\infty}. \tag{4.23}$$

4.3.3.1 Sobolev Duals

In the case of $\Sigma\Delta$ modulation, H is defined by (4.6), and given in matrix form by D^r where the diagonal entries of the lower bidiagonal matrix D are 1 and the subdiagonal entries are -1. Because $\|\Psi D^r\|_{2\to 2}$ resembles a Sobolev norm on Ψ, the corresponding alternative dual was called the (rth order) Sobolev dual of Φ in [4]. In this work, Sobolev duals of certain deterministic frames, such as the harmonic frames, were studied. More precisely, [4] considered frames obtained using a sufficiently dense sampling of vector-valued functions on $[0, 1]$, which had the additional property that their component functions were piecewise C^1 and linearly independent. For such frames, it was shown that

$$\sigma_{\min}(D^{-r}\Phi) \ge c_r m^{r+\frac{1}{2}}, \tag{4.24}$$

hence with (4.23), the reconstruction error using the rth order Sobolev dual satisfies

$$\|x - \Psi_{D^{-r}}q\|_2 \le \frac{C_r}{m^r}\|u\|_{\infty} \tag{4.25}$$

with $C_r := 1/c_r$. Here, for a fixed stable $\Sigma\Delta$ scheme, the constant C_r depends only on the order r and the vector-valued function from which the frame was sampled. The main technique used in [4] to control the operator norm $\|\Psi_{D^{-r}}D^r\|_{2\to 2}$ is a Riemann sum argument. The argument leverages the smoothness of the vector-valued functions from which the frames are sampled to obtain a lower bound on $\|D^{-r}\Phi x\|_2$ over unit norm vectors $x \in \mathbb{R}^d$ and produces the stated lower bound (4.24).

As mentioned before, error bounds similar to (4.25) had also been obtained in [6], albeit for specific tight frames. Nevertheless, in both [4] and [6], the decay of the error associated with $\Sigma\Delta$ quantization is a polynomial function of the number of measurements. The significance of this polynomial error decay stems from the fact that for any frame, a lower bound on the reconstruction error associated with MSQ is known to decay only linearly in m [20].

4.3.3.2 Refined Bounds Using Sobolev Duals

The analysis of [4] was refined in [29] in two special cases: harmonic frames, and the so-called Sobolev self-dual frames. For these frames, [29] established an upper bound on the reconstruction error that decays as a root-exponential function of the number of measurements. More specifically, for harmonic frames, [29] explicitly bounds the constant C_r in (4.25) and, as in [21] and [15], optimizes the $\Sigma\Delta$ scheme's order r as a function of the number of measurements. Quantizing with a $\Sigma\Delta$ scheme of the optimal order $r_{\text{opt}}(m)$ and reconstructing with the associated Sobolev dual results in a root-exponential error bound

$$\|x - \Psi_{D^{-r_{\text{opt}}}} q\|_2 \leq c_1 e^{-c_2\sqrt{m/k}} \tag{4.26}$$

where the constants c_1 and c_2 depend on the quantization alphabet $\mathcal{A}_{L,\delta}$ and possibly on k as well. This possible dependence on k is absent in the similar bound for Sobolev self-dual frames. Sobolev self-dual frames are defined using the singular value decomposition $D^r = U\Sigma V^*$. Here, the $m \times k$ matrix corresponding to a Sobolev self-dual frame consists of the k columns of U associated with the smallest singular values of D^r. This construction implies that the frame admits itself as both a canonical dual and Sobolev dual of order r, hence the name. More importantly, this construction also allows one to bound C_r in (4.25) explicitly and optimize the $\Sigma\Delta$ scheme's order r to obtain the error bound (4.26), without any dependence of the constants on k.

While we have so far discussed deterministic constructions of frames, Gaussian random frames were studied in [23], and later, sub-Gaussian random frames in [30]. We will discuss these random frames extensively in Section 4.4.1, though at this point we note that, like the harmonic and Sobolev self-dual frames, these frames also allow for root-exponential error decay when the order of the $\Sigma\Delta$ scheme is optimized.

In the context of $\Sigma\Delta$ quantization of frame coefficients using a fixed alphabet $\mathcal{A}$, the number of measurements is proportional to the total number of bits. Hence, the error bounds (4.25) and (4.26) can be interpreted as polynomially and root-exponentially decaying in the total number of bits. While these bounds are certainly a big improvement over the linearly decaying lower bound associated with MSQ, they are still sub-optimal. To see this, one observes that the problem of quantizing vectors in the unit ball of $\mathbb{R}^k$ with a maximum reconstruction error of ϵ is analogous to covering the unit-ball with balls of radius ϵ. A simple volume argument shows that to quantize the unit ball of $\mathbb{R}^k$ with an error of ε, one needs at least $k \log_2 \left(\frac{1}{\epsilon}\right)$ bits. Thus, the reconstruction error can at best decay exponentially in the number of bits used. Moreover, since there exists a covering of the unit-ball with no more than $\left(\frac{3}{\epsilon}\right)^k$ elements (see, e.g., [32]), in principle an exponential decay in the error as a function of the number of bits used is possible. This exponential error decay is predicated on a quantization scheme that has direct access to x and, more importantly, the ability to compare x to each of the approximately ϵ^{-k} elements of the covering, to assign it an appropriate binary label. The reconstruction scheme for this quantization would then simply replace the binary label by the center of the element of the covering associated with it. Of course, this setting is markedly different from the noise-shaping quantization of frame coefficients considered in this chapter, but it establishes exponential error decay in the number of bits as optimal.

To achieve exponential error decay in the number of bits, [26] proposed an encoding scheme to follow rth order $\Sigma\Delta$ quantization. The encoding scheme consists of using an $\ell \times m$ Bernoulli random matrix B, with ℓ slightly larger than k, to embed the vector $D^{-r}q$ into a lower dimensional subspace. Since B serves as a distance-preserving Johnson-Lindenstrauss embedding (see, [1, 27]), the vector $BD^{-r}q$ effectively contains all the information needed for accurate reconstruction of x, and it is the only quantity retained. Moreover, the number of bits required to store $BD^{-r}q$ scales only logarithmically in m. Using $(BD^{-r}\Phi)^\dagger$ as a reconstruction operator (acting on $BD^{-r}q$) and employing the properties of Johnson-Lindenstrauss embeddings, [26] shows that the reconstruction error still decays as it would have if no embedding had been employed. In particular, this means an error decay of m^{-r} for the frames discussed in this section. Combining these two observations, i.e., logarithmic scaling of the number of bits with m, and polynomial decay of the error, [26] obtains reconstruction error bounds that decay *exponentially*, i.e., near optimally, in the number of bits.

It turns out that exponential decay of the reconstruction error (in the bit rate or in the oversampling ratio m/k) can also be achieved by means of the "plain route" of noise-shaping quantization and alternative dual reconstruction only, but with noise-shaping unlike $\Sigma\Delta$ quantization and more like the conventional beta encoding [10, 11]. This method, called beta duals, is explained next for general frames, and later in Section 4.4.2 for random frames.

4.3.3.3 Further generalizations: V-duals

Given any $m \times k$ matrix Φ whose rows are a frame for $\mathbb{R}^k$, consider any $p \times m$ matrix V (i.e., not necessarily square) such that $V\Phi$ is also a frame for $\mathbb{R}^k$. We will call

$$\Psi_V := (V\Phi)^\dagger V \tag{4.27}$$

the *V-dual* of Φ. (The square and invertible case of $V = H^{-1}$ was already discussed at the beginning of this subsection.) When $p < m$, we call $V\Phi$ the *V-condensation* of Φ.

With a V-dual, we have $\Psi_V H = (V\Phi)^\dagger VH$ so that

$$\|\Psi_V H\|_{\infty\to 2} \le \frac{\|VH\|_{\infty\to 2}}{\sigma_{\min}(V\Phi)} \le \frac{\sqrt{p}\|VH\|_{\infty\to\infty}}{\sigma_{\min}(V\Phi)}. \tag{4.28}$$

For $V = H^{-1}$ (and therefore, $p = m$), combination of (4.19) with (4.28) agrees with (4.23). However, as shown in [11], optimization of (4.28) over V can produce a strictly smaller reconstruction error upper bound. A highly effective special case is discussed next.

Beta duals

Beta duals have been recently proposed and studied in [10, 11]. They constitute a special case of V-duals, while they relate strongly to classical beta expansions. (See [12, 35] for the classical theory of beta expansions, and [14] for the use of beta expansions in A/D conversion as a robust alternative to successive approximation.) In order to illustrate the main construction of beta duals without technical details, our presentation in this article will be restricted to certain dimensional constraints as described below.

Let $m \ge p \ge k$ and assume that $\lambda' := m/p$ is an integer. For any $\beta > 1$, let h^β be the (length-2) sequence given by $h_0^\beta = 1$ and $h_1^\beta = -\beta$. Define H^β to be the $\lambda' \times \lambda'$ noise-shaping transfer operator corresponding to h^β, and

$$v^\beta := [\beta^{-1} \;\; \beta^{-2} \;\cdots\; \beta^{-\lambda'}].$$

We set

$$H := \begin{bmatrix} H^\beta & & \\ & \ddots & \\ & & H^\beta \end{bmatrix}_{m\times m} \quad \text{and} \quad V := \begin{bmatrix} v^\beta & & \\ & \ddots & \\ & & v^\beta \end{bmatrix}_{p\times m}. \tag{4.29}$$

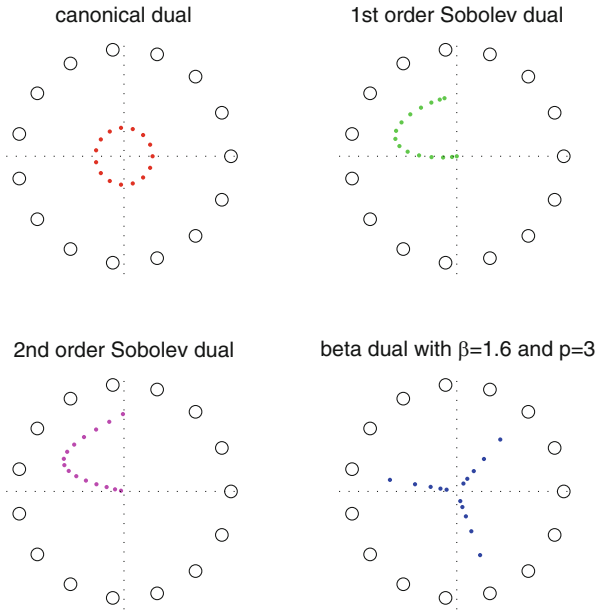


Fig. 4.2 Comparative illustration of the various alternative duals described in this paper: Each plot depicts the original frame in $\mathbb{R}^2$ consisting of the $15th$ roots-of-unity along with one of its duals (scaled up by a factor of two for visual clarity). For the computation of the alternative duals, the analysis frame was ordered counter-clockwise starting from $(1, 0)$.

In other words, $H = I_p \otimes H^\beta$ and $V = I_p \otimes v^\beta$ where $\otimes$ denotes the Kronecker product. It follows that $VH = I_p \otimes (v^\beta H^\beta)$. Since $v^\beta H^\beta = [0 \cdots 0 \ \beta^{-\lambda'}]$, we have $\|VH\|_{\infty\to\infty} = \beta^{-\lambda'}$ which, together with (4.19) and (4.28), yields

$$\|x - \Psi_V q\|_2 \leq \frac{\sqrt{p}\|u\|_\infty}{\sigma_{\min}(V\Phi)}\beta^{-\lambda'}. \tag{4.30}$$

For certain special frames, such as the harmonic semi-circle frames, it is possible to set p as low as k and turn the above bound into a near-optimal one in terms of its bit-rate [11]. The case of random frames will be discussed in the next section.

In Fig. 4.2, we illustrate a beta dual of a certain "roots-of-unity" frame along with the Sobolev duals of order 0 (the canonical dual), 1, and 2.

4.4 Analysis of Alternative Duals for Random Frames

In this section, we consider random frames, that is, frames whose analysis (or synthesis) operator is a random matrix. Certain classes of random matrices have become of considerable importance in high dimensional signal processing, particularly with the advent of compressed sensing. One main reason for this is that

their inherent independence entails good conditioning of not only the matrix, but also its submatrices. Because of the fast growing number of such submatrices with dimension, the latter is very difficult to achieve with deterministic constructions. This also means, however, that any two frame vectors are approximately orthogonal, so frame path conditions that would imply recovery guarantees using canonical dual frames will almost never hold. For this reason, it is crucial to work with alternative duals. We separately consider the two main examples discussed above, Sobolev duals and beta duals.

4.4.1 Sobolev duals of random frames

As noted above, the Sobolev dual of a frame is the dual frame Ψ that minimizes the expression $\|\Psi D^r\|_{2\to 2}$, and the explicit minimizer is given by (4.22) with $H = D^r$. By (4.23), a bound for the error that arises when using this alternative dual to reconstruct is governed by $\sigma_{\min}(D^{-r}\Phi)$. Thus a main goal of this subsection is to discuss the behavior of this minimum singular value.

The matrix $D^{-r}\Phi$ is the product of a deterministic matrix D^{-r}, whose singular values are known to a sufficient approximation, and a random matrix Φ, whose singular values are known to be well concentrated. Nevertheless, using a product bound does not yield good results, mainly because the singular values of D^{-r} differ tremendously, so any worst case bound will not be good enough. One approach to provide a refined bound is to first provide lower bounds for the action of $D^{-r}\Phi$ on a single vector and then proceed via a covering argument. That is, one combines these lower bounds for all of the vectors forming an ϵ-net, obtaining a uniform bound for the net. An approximation argument then allows to pass from the net to all vectors in the sphere. In this way, [23] obtains the following result for Gaussian random frames:

Theorem 1 ([23]). *Let Φ be an $m \times k$ random matrix whose entries are i.i.d. standard Gaussian variables. Given $r \in \mathbb{N}$ and $\alpha \in (0,1)$, there exist constants strictly positive r-dependent constants c_1, c_2, and c_3 such that if $\lambda := m/k \geq (c_1 \log m)^{1/(1-\alpha)}$, then with probability at least $1 - \exp(-c_2 m\lambda^{-\alpha})$,*

$$\sigma_{\min}(D^{-r}\Phi) \geq c_3(r)\lambda^{\alpha(r-\frac{1}{2})}\sqrt{m}. \tag{4.31}$$

In this approach, one explicitly uses the density of the Gaussian distribution. Thus, as soon as the matrix entries fail to be exactly Gaussian, a completely different approach is needed. In what follows, we will present the main idea of the method used in [30] to tackle the case of random matrices with independent sub-Gaussian entries as introduced in the following definition (for alternative characterizations of sub-Gaussian random variables see, for example, [42]). This approach is also related to the RIP-based analysis for quantized compressive sampling presented in [18] (cf. Section 4.5 below).

Definition 1. A random variable ξ is sub-Gaussian with parameter $c > 0$ if it satisfies $\mathbb{P}(|\xi| > t) \leq e^{1-ct^2}$ for all $t \geq 0$.

As in the Gaussian case presented in [23], we employ the singular value decomposition $D^{-r} = U\Sigma V^*$ where U and V are unitary and $\Sigma \in \mathbb{R}^{m\times m}$ is a diagonal matrix with entries $s_1 \geq \cdots \geq s_m \geq 0$. Then

$$\sigma_{\min}(D^{-r}\Phi) = \sigma_{\min}(U\Sigma V^*\Phi) = \sigma_{\min}(\Sigma V^*\Phi),$$

as U is unitary. Furthermore, for $P_\ell : \mathbb{R}^m \to \mathbb{R}^\ell$ the projection onto the first ℓ entries, $\ell \leq m$, one has in the positive semidefinite partial ordering $\succeq$

$$\Sigma \succeq P_\ell \Sigma = P_\ell \Sigma P_\ell^* P_\ell \succeq s_\ell P_\ell.$$

Here the first inequality uses that P_ℓ is a projection, the following equality uses that Σ is diagonal, and the last inequality uses that the diagonal entries of Σ are ordered.

As a consequence, we find that $\sigma_{\min}(D^{-r}\Phi) \geq s_\ell \sigma_{\min}(V^*\Phi)$. For Gaussian matrix entries, this immediately yields Theorem 1, as standard Gaussian vectors are rotation invariant, so $P_\ell V^*\Phi$ is just a standard Gaussian matrix, whose singular value distributions are well understood (see for example [42]). Applying the bound for different values of ℓ yield the theorem for different choices of α.

For independent, zero mean, unit variance sub-Gaussian (rather than Gaussian) matrix entries, one no longer has such a strong version of rotation invariance; while the columns of $V^*\Phi$ will still be sub-Gaussian random vectors, its entries will, in general, no longer be independent. There are also singular value estimates that require only independent sub-Gaussian matrix columns rather than independent entries (see again [42]), but such bounds require that the matrix columns are of constant norm. Even if Φ and hence also $V^*\Phi$ has constant norm columns (such as for example for Bernoulli matrices, $\Phi_{ij} \in \pm 1$), the projection P_ℓ will typically map them to vectors of different length.

In order to nevertheless bound the singular values, we again use a union bound argument, first considering the action on one fixed vector x of unit norm. Then we write

$$\|V^*\Phi x\|_2^2 = \sum_{i,i'=1}^{k} \sum_{j,j'=1}^{m} x_i \Phi_{ji} (VP_\ell^* P_\ell V^*)_{jj'} \Phi_{j'i'} x_{i'}.$$

Thus $\|V^*\Phi x\|_2^2$ is a so-called chaos process, that is, a random quadratic form of the form $\langle \xi, M\xi \rangle$, where ξ is a random vector with independent entries (in this case, the vectorization of Φ). Its expectation is given by

$$\mathbb{E}\|V^*\Phi x\|_2^2 = \sum_{i=1}^{k} \sum_{j=1}^{m} x_i^2 \mathbb{E}\Phi_{ji}^2 (VP_\ell^* P_\ell V^*)_{jj} = \|x\|_2^2 \mathrm{tr}(VP_\ell^* P_\ell V^*) = \ell,$$

where the last equality uses the cyclicity of the trace. Its deviation from the expectation can be estimated using the following refined version of the Hanson-Wright inequality, which has been provided in [37] (see [24] for the original version).

Theorem 2. *Let $\xi = (\xi_1, \ldots, \xi_n) \in \mathbb{R}^n$ be a random vector with independent components ξ_i which are sub-Gaussian with parameter c and satisfy $\mathbb{E}\xi_i = 0$. Let A be an $n \times n$ matrix. Then for every $t \geq 0$,*

$$\mathbb{P}\big\{|\langle \xi, M\xi\rangle - \mathbb{E}\langle \xi, M\xi\rangle| > t\big\} \leq 2\exp\Big(-C_4 \min\big(\frac{t^2}{c^4\|M\|_F^2}, \frac{t}{c^2\|M\|_{2\to 2}}\big)\Big),$$

where C_4 is an absolute constant.

To obtain a deviation bound for the above setup, we thus need to estimate the Frobenius norm $\|M\|_F^2 := \mathrm{tr}M^*M = \sum_{i,i',j,j'} M^2_{(i,i'),(j,j')}$ and the operator norm $\|M\|_{2\to 2} := \sup_{\|y\|_2=1} \|My\|_2$ of the doubly-indexed matrix M given by $M_{(i,j),(i',j')} = x_i x_{i'} (VP_\ell^* P_\ell V^*)_{jj'}$. For the Frobenius norm, we write

$$\|M\|_F^2 = \sum_{i,i',j,j'} x_i^2 x_{i'}^2 (VP_\ell^* P_\ell V^*)^2_{jj'} = \|VP_\ell^* P_\ell V^*\|_F^2 = \mathrm{tr}(VP_\ell^* P_\ell V^* VP_\ell^* P_\ell V^*) = \ell,$$

where in the last equality, we used again the cyclicity of the trace, that V is unitary, and that $P_\ell^* P_\ell$ is a projection. For the operator norm, we note that

$$M = P_\ell V^* \begin{pmatrix} x^T & 0 & \cdots & 0 \\ 0 & x^T & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots \\ 0 & \cdots & 0 & x^T \end{pmatrix},$$

so as all these three factors have operator norm 1, the norm of their product is bounded above by 1. On the other hand, applying M to the unit norm vector y given by $y_{(i,j)} = x_i V_{1j}$ yields $My = e_1$, where e_1 is the first standard basis vector, showing that the norm is also lower bounded by 1. So one indeed has $\|M\|_{2\to 2} = 1$. Combining these bounds with Theorem 2 yields the following generalization of Theorem 1 for sub-Gaussian frames.

Theorem 3 ([30]). *Let Φ be an $m \times k$ random matrix whose entries are zero mean, unit variance, sub-Gaussian random variables with parameter c. Given $r \in \mathbb{N}$ and $\alpha \in (0,1)$, there exist constants $c = c(r) > 0$ and $c' = c'(r) > 0$ such that if $\lambda := \frac{m}{k} \geq c^{\frac{1}{1-\alpha}}$ then one has with probability at least $1 - \exp(-c' m\lambda^{-\alpha})$*

$$\sigma_{\min}(D^{-r}\Phi) \geq \lambda^{\alpha(r-\frac{1}{2})}\sqrt{m}. \tag{4.32}$$

Combining (4.23) for $H = D^r$ with the lower bound of (4.31) or (4.32), the Sobolev dual reconstruction $\Psi_{D^{-r}} q$ from $\Sigma\Delta$ quantized frame coefficients $y = \Phi x$ results in the error bound

$$\|x - \Psi_{D^{-r}} q\|_2 \leq C(r)\lambda^{-\alpha(r-\frac{1}{2})}\|u\|_\infty. \tag{4.33}$$

Thus the error decays polynomially in the oversampling rate λ as long as the underlying $\Sigma\Delta$ scheme is stable. For the greedy quantization rule, stability follows from Proposition 1, as long as $\|y\|_\infty \leq \mu$ for a suitable μ whose range is constrained by the quantization alphabet $\mathcal{A}_{L,\delta}$ and r. (It can be easily computed that for $H = D^r$, we have $\|\tilde{H}\|_{\infty\to\infty} = 2^r - 1$. Hence we require $L > 2^r - 1$, with the value of δ assumed to be adjustable.) If we assume that $\|x\|_2 \leq 1$, then controlling $\|y\|_\infty$ amounts to bounding $\|\Phi\|_{2\to\infty} \leq \|\Phi\|_{2\to 2}$ and thus to bounding the maximum singular value of a rectangular matrix with independent sub-Gaussian entries. This is a well-understood setup, it is known that the singular values of such a matrix are well concentrated and one has $\|\Phi\|_{2\to\infty} \leq \|\Phi\|_{2\to 2} = O(\sqrt{m})$ with high probability (see again [42]). As a consequence, the $\Sigma\Delta$ scheme is stable provided L is chosen large enough and the quantizer level is adjusted accordingly. We conclude that sub-Gaussian frame expansions quantized using a greedy r-th order $\Sigma\Delta$ scheme allow for reconstruction error bounds decaying polynomially in the oversampling rate, where the decay order can be made arbitrarily large by choosing r large enough.

4.4.2 Beta duals of random frames

We return to the Gaussian distribution for the analysis of beta duals for random frames. Based on the error bound (4.30) derived in Section 4.3.3.3, it now suffices to give a probabilistic lower bound for $\sigma_{\min}(V\Phi)$. Note that the entries of the $p \times k$ matrix $V\Phi$ are i.i.d. Gaussian with variance

$$\sigma_{\lambda'}^2 := \beta^{-2} + \cdots + \beta^{-2\lambda'}. \tag{4.34}$$

At this point, a choice for the parameter p needs to be made. In [11], both choices of $p = k$ and $p > k$ were studied in detail. The analysis of the former choice is somewhat cleaner, but the strongest probabilistic estimates follow by choosing p greater than k.

We will primarily be interested in the smallest singular value of $V\Phi$ being near zero. For $p = k$, the following well-known result suffices:

Theorem 4 ([36, Theorem 3.1], [17]). *Let Ω be a $k\times k$ random matrix with entries drawn independently from $\mathcal{N}(0, \sigma^2)$. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\left\{\sigma_{\min}(\Omega) \leq \varepsilon\sigma/\sqrt{k}\right\}\right) \leq \varepsilon.$$

Meanwhile, the stability of the greedy quantizer with alphabet $\mathcal{A}_{L,\delta}$ can be ensured in a way similar to the case of Sobolev duals, noting that $\|\tilde{H}\|_{\infty\to\infty} = \beta$. Hence, we know that if $\beta + \mu/\delta \leq L$, then $\|u\|_\infty \leq \delta$. By standard Gaussian concentration results, $\mu \leq 4\sqrt{m}$ is guaranteed with probability at least $1 - e^{-2m}$. Therefore, with (4.30) and Theorem 4 in which we set $\Omega = V\Phi$, we obtain

$$\|x - \Psi_V q\|_2 \leq kL\varepsilon^{-1}\delta\beta^{-m/k} \tag{4.35}$$

with probability at least $1 - \varepsilon - e^{-2m}$, where we have also used the simple chain of inequalities $1/\sigma_{\lambda'} \leq \beta \leq L$. The value of β can be chosen arbitrarily close to L with sufficiently large values of δ. However, the optimal choice would result from minimizing $\delta\beta^{-m/k}$ subject to $\beta + \mu/\delta = L$. For details, see [11].

For $p > k$, we have the following result:

Theorem 5 ([11, Theorem 4.3]). *Let $p > k$ and Ω be a $p \times k$ random matrix whose entries are drawn independently from $\mathcal{N}(0, \sigma^2)$. Then for any $0 < \varepsilon < 1$,*

$$\mathbb{P}\left(\left\{\sigma_{\min}(\Omega) \leq \varepsilon\sigma\sqrt{p}/2\right\}\right) \leq \left(10 + 8\sqrt{\log \varepsilon^{-1}}\right)^k e^{p/2}\varepsilon^{p-k}.$$

The corresponding error bound

$$\|x - \Psi_V q\|_2 \leq 2L\varepsilon^{-1}\delta\beta^{-m/p} \tag{4.36}$$

now holds with higher probability. The choices $\varepsilon \approx \beta^{-\eta m/p}$ for small η and $p \approx (1 + \eta)k$ turn out to be good ones. For details, again see [11].

4.5 Noise-shaping Quantization for Compressive Sampling

Compressive sampling (also called compressed sensing) has emerged over the last decade as a novel sampling paradigm. It is based on the empirical observation that various important classes of signals encountered in practice, such as audio and images, admit (nearly) sparse approximations when expanded with respect to an appropriate basis or frame, such as a wavelet basis or a Gabor frame. Seminal papers by Candès, Romberg, and Tao [8], and by Donoho [16] established the fundamental theory, specifying how to collect the samples (or measurements), and the relation between the approximation accuracy and the number of samples acquired ("sampling rate") vis-a-vis the sparsity level of the signal. Since then the literature has matured considerably, again focusing on the same issues, i.e., how to construct effective measurement schemes and how one can control the approximation error as a function of the sampling rate, e.g., see [19].

By now compressive sampling is well-established as an effective sampling theory. From the perspective of practicability, however, it also needs to be accompanied by a quantization theory. Here, as in the case of frames, MSQ is highly limited as

a quantization strategy in terms of its rate-distortion performance. Thus, efficient quantization methods are needed for compressive sampling to live up to its name, i.e., to provide compressed representations in the sense of source coding.

In this section, we will discuss how noise-shaping methods can be employed to quantize compressive samples of sparse and compressible signals to vastly improve the reconstruction accuracy compared to the default method of MSQ. We start with the basic framework of compressive sampling as needed for our discussion.

4.5.1 Basics of Compressive Sampling

In the basic theory of compressive sampling, the signals of interest are finite (but potentially high) dimensional vectors that are exactly or approximately *sparse*. More precisely, we say that a signal x in $\mathbb{R}^N$ is k-sparse if it is in $\Sigma_k^N := \{x \in \mathbb{R}^N : \|x\|_0 \leq k\}$. Here $\|x\|_0$ denotes the number of non-zero entries of x. The signals we encounter in practice are typically not sparse, but they can be well-approximated by sparse signals. Such signals are referred to as compressible signals and roughly identified as signals x with small $\sigma_k(x)_{\ell_p}$, *the best k-term approximation error of x in* ℓ_p, defined by

$$\sigma_k(x)_{\ell_p} := \min_{z \in \Sigma_k^N} \|x - z\|_p.$$

Compressive sampling consists of acquiring linear, non-adaptive measurements of sparse or compressible signals, possibly corrupted by noise, and recovering (an approximation to) the original signal from the compressive samples via a computationally tractable algorithm. In other words, the compressive samples are obtained by multiplying the signal of interest by a *compressive sampling (measurement) matrix*. The success of recovery algorithms relies heavily on certain properties of this matrix. To state this dependence precisely, we next define the restricted isometry constants of a matrix.

Definition 2. The restricted isometry constant (see, e.g., [8]) $\gamma_k := \gamma_k(\Phi)$ of a matrix $\Phi \in \mathbb{R}^{m \times N}$ is the smallest constant for which

$$(1 - \gamma_k)\|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \gamma_k)\|x\|_2^2$$

for all $x \in \Sigma_k^N$.

Suppose that $\Phi \in \mathbb{R}^{m \times N}$ is used as a compressive sampling matrix. Here, m denotes the number of measurements and is significantly smaller than N, the ambient dimension of the signal. Let $\tilde{y} := \Phi x + w$ denote the (possibly) perturbed measurements of a signal $x \in \mathbb{R}^N$, where the unknown perturbation w satisfies $\|w\|_2 \leq \epsilon$. A crucial result in the theory of compressive sampling states that if the restricted isometry constants of Φ are suitably controlled (e.g. as originally stated

in [8], or more recently as in [7] which only assumes $\gamma_{ak} \leq \sqrt{(a-1)/a}$ for some $a \geq 4/3$), then there is an approximate recovery $\Delta_1^\epsilon(\Phi, \tilde{y})$ of x which satisfies

$$\|x - \Delta_1^\epsilon(\Phi, \tilde{y})\|_2 \leq C\epsilon + D\sigma_k(x)_{\ell_1}/\sqrt{k}. \tag{4.37}$$

Here, $\Delta_1^\epsilon(\Phi, \tilde{y})$ is found by mapping $\tilde{y}$ to a minimizer of a tractable, convex optimization problem—which is often called the "Basis Pursuit Denoise" algorithm—given by

$$\Delta_1^\epsilon(\Phi, \tilde{y}) := \arg\min_z \|z\|_1 \ \text{ subject to } \ \|\Phi z - \tilde{y}\|_2 \leq \epsilon.$$

C and D are constants that depend on Φ, but can be made absolute by slightly stronger assumptions on Φ.

Note that in the noiseless case, it follows from (4.37) that any k-sparse signal can be exactly recovered from its compressive samples as $\Delta_1^0(\Phi, \Phi x)$. In the general case, the approximation error remains within the noise level and within the best k-term approximation error of x in ℓ_1. Hence the recovery is robust with respect to the amount of noise and stable with respect to violation of the exact sparsity assumption. The decoder Δ_1^ϵ is a *robust compressive sampling decoder* as defined next.

Definition 3 ([30, Definition 4.9]). Let $\varepsilon > 0$, let m, N be positive integers such that $m < N$ and suppose that $\Phi \in \mathbb{R}^{m \times N}$. We say that $\Delta : \mathbb{R}^{m \times N} \times \mathbb{R}^m \to \mathbb{R}^N$ is a robust compressive sampling decoder with parameters (k, a, γ), $k < m$, and constant C if

$$\|x - \Delta(\Phi, \Phi x + e)\| \leq C\varepsilon, \tag{4.38}$$

for all $x \in \Sigma_k^N$, $\|e\|_2 \leq \varepsilon$, and all matrices Φ with a restricted isometry constant $\gamma_{ak} < \gamma$.

Examples of robust decoders include Δ_1^ϵ and its p-norm generalization Δ_p^ϵ with $0 < p \leq 1$ [8, 38], compressive sampling matching pursuit (CoSaMP) [33], Orthogonal Matching Pursuit (OMP) [43], and iterative hard thresholding (IHT) [5]. See also [19] for detailed estimates of the relevant parameters.

4.5.2 Noise-shaping Quantization of Compressive Samples

Even though noise shaping methods are tailored mainly for quantizing redundant representations, perhaps surprisingly, they also provide efficient strategies for quantizing compressive samples [18, 22, 23, 30]. The approach, originally developed in [23] specifically for $\Sigma\Delta$ quantization, relies on the observation that when the original signal is exactly sparse, compressed measurements are in fact redundant

frame coefficients of the sparse signal restricted to its support. Since then it has been extended for beta encoding and applied to compressible signals as well [10]. We start with the case of sparse signals.

4.5.2.1 Sparse signals

Let $x \in \Sigma_k^N$ with $\text{supp}(x) = T$ and $\Phi \in \mathbb{R}^{m\times N}$ be a compressive sampling matrix. Then, we have

$$y = \Phi x \quad \Longrightarrow \quad y = \Phi_T x_T,$$

where Φ_T is the submatrix of Φ consisting of its columns indexed by T and x_T is the restriction of x to T. Accordingly, *any* quantization technique designed for frames could be adopted to compressive sampling as follows:

Quantization: Since the compressive samples are in fact frame coefficients, apply the noise-shaping quantization algorithm directly to the compressive samples y to obtain the quantized samples, say, q. Note that the quantization process is blind to the support of the sparse signal as well as to the sampling operator.

Reconstruction: Reconstruct via the following *two-stage reconstruction algorithm*. To obtain an estimate $x^\#$ of x from q:

1. **Coarse Recovery:** Solve

$$\tilde{x} = \Delta_1^{\epsilon_Q}(\Phi, q) \tag{4.39}$$

   where ϵ_Q is an upper bound on $\|y - q\|_2$, which depends on the quantization scheme and is known explicitly. Note that the decoder $\Delta_1^{\epsilon_Q}$ above can be replaced with any robust compressive sampling decoder Δ. Clearly, by (4.38) $\|x - \tilde{x}\|$ will be small if ϵ_Q is small.
2. **Fine Recovery:** Obtain a support estimate, $\tilde{T}$, of x from $\tilde{x}$. A finer approximation for x is then given by reconstructing with an appropriate alternative dual of the underlying frame $\Phi_{\tilde{T}}$ based on the noise-shaping operator that was employed for quantization.

The success of the two-stage reconstruction algorithm relies on the accurate recovery of the support of x. In turn, this can be guaranteed by a size condition on the smallest-in-magnitude non-zero entry of x. To see this, note that for all $i \in T$, the robustness guarantee (4.38) yields $|\tilde{x}_i - x| \leq C\epsilon_Q$, which, together with the size condition $\min_{i\in T} |x_i| > 2C\epsilon_Q$, gives $|\tilde{x}_i| > C\epsilon_Q$. Moreover, by (4.38) we have $|\tilde{x}_i| \leq C\epsilon_Q$ for all $i \in T^c$. Consequently, the largest-in-magnitude k coefficients of $\tilde{x}$ are supported on T. Thus, we have the following proposition.

Proposition 2. *Suppose that $x \in \Sigma_k^N$ with $supp(x) = T$, and let $\Phi \in \mathbb{R}^{m\times N}$ be a compressive sampling matrix so that* (4.38) *holds for $\Delta = \Delta_1^{\epsilon_Q}$ with robustness*

constant C. Let $\tilde{x}$ be as in (4.39) where $\|\Phi x - q\|_2 \leq \epsilon_Q$. If $\min_{i\in T} |x_i| > 2C\epsilon_Q$, then the k largest-in-magnitude coefficients of $\tilde{x}$ are supported on T.

By this observation, the coarse recovery stage not only yields an estimate $\tilde{x}$ that satisfies $\|x-\tilde{x}\|_2 \leq C\epsilon_Q$, but it also gives an accurate estimate of the support of x (via the support of the k-largest coefficients of $\tilde{x}$). It remains to show that reconstruction techniques associated with noise shaping quantization for frames can be used in the fine recovery stage to produce an estimate $x^{\#}$ that is more accurate than $\tilde{x}$ of the coarse stage.

When q results from a noise-shaping quantization scheme, accurate recovery based on alternative duals can be guaranteed via (4.19). In particular, suppose that H is the noise transfer operator of the quantizer. Conditioned on recovering T, let $\Psi_{H^{-1}}$ be the left inverse of Φ_T as defined in (4.22) and set $x^{\#} := \Psi_{H^{-1}} q$. We then have, as before,

$$\|x - x^{\#}\|_2 \leq \frac{\sqrt{m}}{\sigma_{\min}(H^{-1}\Phi_T)} \|u\|_\infty \tag{4.40}$$

where u is as in (4.13) .

Predominantly, compressed sensing matrices Φ (hence their submatrices Φ_T) are random matrices. Thus, to uniformly control the reconstruction error via (4.40) one needs lower bounds on the smallest singular values of the random matrices $H^{-1}\Phi_T$ for all $T \subset [N] := \{1, \ldots, N\}$, $|T| = k$, as well as a uniform upper bound on $\|u\|_\infty$.

We concentrate again on random matrices Φ with independent and identically distributed Gaussian or sub-Gaussian entries. In these cases, for each fixed support T, Φ_T is a random frame of the type considered in Section 4.4 and a probabilistic lower bound on $\sigma_{\min}(H^{-1}\Phi_T)$ follows from Theorem 1 (for Gaussian entries) and Theorem 3 (for sub-Gaussian entries).

A uniform lower bound on $\sigma_{\min}(H^{-1}\Phi_T)$ over all support sets T of size k can now be deduced via a union bound over the $\binom{N}{k}$ support sets. Note that to obtain a uniform bound over this rather large set of supports, one requires a relatively small bound for the probability of failure on each potential support, and, consequently, a larger embedding dimension m as compared to the case of a single frame. An alternative approach based on the restricted isometry constant, essentially yielding the same result, can be found in [18].

The approaches just outlined are general and can be applied in the case of any noise shaping quantizer that allows exact recovery of the support of sparse vectors via Proposition 2. In the following, however, we focus on the special case of rth-order $\Sigma\Delta$ quantization, where $H = D^{-r}$ and we obtain the following theorem.

Theorem 6 ([23, 30]). *Let $r \in \mathbb{Z}^+$, fix $a \in \mathbb{N}$, $\gamma < 1$, and $c, C > 0$. Then there exist constants C_1, C_2, C_3, C_4 depending only on these parameters such that the following holds.*

Fix $0 < \alpha < 1$. Let Φ be an $m \times N$ matrix with independent sub-Gaussian entries that have zero mean, unit variance, and parameter c, let Δ be a robust compressive sampling decoder and $k \in \mathbb{N}$ is such that

$$\lambda := \frac{m}{k} \geq \left(C_1 \log(eN/k)\right)^{\frac{1}{1-\alpha}}.$$

Suppose that q is obtained by quantizing Φz, $z \in \mathbb{R}^N$, via the rth order greedy $\Sigma\Delta$ scheme with the alphabet $\mathcal{A}_{L,\delta}$, and with $L \geq \lceil \frac{K\lambda^{-1/2}}{\delta} \rceil + 2^r + 1$. Denote by q the quantization output resulting from Φz where $z \in \mathbb{R}^N$. Then with probability exceeding $1 - 4e^{-C_2 m^{1-\alpha} k^\alpha}$ for all $x \in \Sigma_k^N$ having $\min_{j \in \mathrm{supp}(x)} |x_j| > C_3\delta$:

(i) The support of x, T, coincides with the support of the best k-term approximation of $\Delta(\frac{1}{\sqrt{m}}\Phi, \frac{1}{\sqrt{m}}q)$.

(ii) Denoting by Φ_T and F the sub-matrix of Φ corresponding to the support of z and its rth order Sobolev dual respectively, and by $x_T \in \mathbb{R}^k$ the restriction of x to its support, we have

$$\|x_T - Fq\|_2 \leq C_4 \lambda^{-\alpha(r-1/2)}\delta.$$

We remark that in Theorem 6, the requirement that $L \geq \lceil \frac{K\lambda^{-1/2}}{\delta} \rceil + 2^r + 1$ ensures stability of the $\Sigma\Delta$ scheme while $\min_{j \in \mathrm{supp}(x)} |x_j| > C_3\delta$ implies accurate support recovery.

4.5.2.2 Compressible signals

The two-stage reconstruction algorithm for sparse signals presented above applies equally well to noise-shaping quantization based on beta encoding as discussed in Section 4.3.3.3. However, it turns out that for beta encoding there is a more powerful reconstruction algorithm which works for compressible signals as well.

Let Φ now be an $m \times N$ compressive sampling matrix, and let H be the $m \times m$ noise transfer operator and V be the $p \times m$ condensation operator as in (4.29), where again, for simplicity, we have assumed that m/p is an integer. Note that the associated noise-shaping quantization relation

$$\Phi x - q = Hu$$

implies

$$V\Phi x - Vq = VHu,$$

hence we may consider $V\Phi$ as a new condensed measurement matrix and $Vq = V\Phi x + VHu$ as the corresponding perturbed measurement. As before,

$$\|VHu\|_2 \leq \|VH\|_{\infty \to 2} \|u\|_\infty \leq \sqrt{p}\beta^{-m/p}\|u\|_\infty,$$

so that if the greedy quantization rule is stable (i.e., $\|u\|_\infty \leq \delta$), then we can set $\epsilon := \sqrt{p}\beta^{-m/p}\delta$ and consider the decoder

$$(q \mapsto \Delta_1^\epsilon(V\Phi, Vq)).$$

As it follows from the discussion of (4.37), if for some $\alpha > 0$, $\gamma_{2k} := \gamma_{2k}(\alpha V\Phi)$ is sufficiently small (say less than $1/3$), then we have the estimate

$$\|x - \Delta_1^\epsilon(V\Phi, Vq)\|_2 \leq C\alpha\epsilon + D\frac{\sigma_k(x)_1}{\sqrt{k}}, \tag{4.41}$$

where C and D are now absolute constants.

For the random (Gaussian) case, the following result is implied by our discussion above and other tools presented earlier in this paper (for a more detailed derivation of a similar result, see [10]):

Theorem 7. *Let Φ be an $m \times N$ random matrix whose entries are i.i.d. standard Gaussian variables. Let $x \in \mathbb{R}^N$, $\|x\|_2 \leq 1$, and let q be the result of quantizing the measurements Φx with the noise transfer operator H from (4.29) and the alphabet $\mathcal{A}_{L,\delta}$ where $\beta + 2\sqrt{N}/\delta \leq L$. Assume $m \geq p \geq k$ are such that $\lambda' := m/p$ is an integer and*

$$\lambda := \frac{m}{k} \geq C_1\lambda' \log N/k$$

for some numerical constant C_1. Let V be the $p \times m$ condensation matrix as in (4.29) and $\epsilon := \sqrt{p}\beta^{-m/p}\delta$. Then with probability exceeding $1 - e^{-p/C_1'}$ for another numerical constant C_1', we have

$$\|x - \Delta_1^\epsilon(V\Phi, Vq)\|_2 \leq CL\delta\sqrt{p/m}\,\beta^{-m/p} + D\,\frac{\sigma_k(x)_1}{\sqrt{k}}.$$

We note that the optimal choice of the auxiliary parameters p and k in the above theorem depends on the success probability as well as further information on the amount of compressibility of x. A rule of thumb would be to balance the two error terms above corresponding to quantization error and approximation error. Similarly, the choice of β, L, and δ can be optimized. For example, if $L \geq 2$ is given and fixed, but δ is variable, then one would minimize the error bound (over p, k, β and δ) within a given probabilistic guarantee objective and a priori knowledge on x.

Finally, we end with the following remark: a recent work [39, 40] shows that it is in fact possible to obtain an approximation from $\Sigma\Delta$ quantized compressive samples that is robust to additive noise and is stable for compressible signals. This approximation is obtained via a *one-stage reconstruction method* based on solving a simple convex optimization problem. Furthermore, by encoding the quantized measurements via a Johnson- Lindenstrauss dimensionality reducing embedding as in [26], one obtains near-optimal rate-distortion guarantees in the case of sparse signals. For details, see [39, 40].

Acknowledgements FK and RS acknowledge support by the German Science Foundation (DFG) in the context of the Emmy-Noether Junior Research Group KR 4512/1-1 "RaSenQuaSI". ÖY was funded in part by a Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery Grant (22R82411), an NSERC Accelerator Award (22R68054) and an NSERC Collaborative Research and Development Grant DNOISE II (22R07504).

References

1. D. Achlioptas, Database-friendly random projections, in *Proceedings of the Twentieth ACM SIGMOD-SIGACT-SIGART Symposium on Principles of Database Systems* (ACM, New York, 2001), pp. 274–281
2. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Second-order sigma–delta ($\Sigma\Delta$) quantization of finite frame expansions. Appl. Comput. Harmon. Anal. **20**(1), 126–148 (2006)
3. J.J. Benedetto, A.M. Powell, Ö. Yılmaz, Sigma-delta ($\Sigma\Delta$) quantization and finite frames. IEEE Trans. Inf. Theory **52**(5), 1990–2005 (2006)
4. J. Blum, M. Lammers, A.M. Powell, Ö. Yılmaz, Sobolev duals in frame theory and sigma-delta quantization. J. Fourier Anal. Appl. **16**(3), 365–381 (2010)
5. T. Blumensath, M.E. Davies, Iterative hard thresholding for compressed sensing. Appl. Comput. Harmon. Anal. **27**(3), 265–274 (2009)
6. B.G. Bodmann, V.I. Paulsen, S.A. Abdulbaki, Smooth frame-path termination for higher order sigma-delta quantization. J. Fourier Anal. Appl. **13**(3), 285–307 (2007)
7. T.T. Cai, A. Zhang. Sparse representation of a polytope and recovery of sparse signals and low-rank matrices. IEEE Trans. Inf. Theory **60**(1), 122–132 (2014)
8. E.J. Candès, J.K. Romberg, T. Tao, Stable signal recovery from incomplete and inaccurate measurements. Commun. Pure Appl. Math. **59**(8), 1207–1223 (2006)
9. J.C. Candy, G.C. Temes (eds.), *Oversampling Delta-Sigma Data Converters: Theory, Design and Simulation* (Wiley/IEEE, New York, 1991)
10. E. Chou, *Beta-Duals of Frames and Applications to Problems in Quantization*. Ph.D. Thesis, New York University, 2013
11. E. Chou, C.S. Güntürk, Distributed noise-shaping quantization: I. Beta duals of finite frames and near-optimal quantization of random measurements. Preprint (2014) [arXiv:1405.4628]
12. K. Dajani, C. Kraaikamp, From greedy to lazy expansions and their driving dynamics. Expo. Math. **20**(4), 315–327 (2002)
13. I. Daubechies, R. DeVore, Approximating a bandlimited function using very coarsely quantized data: a family of stable sigma-delta modulators of arbitrary order. Ann. Math. (2) **158**(2), 679–710 (2003)
14. I. Daubechies, R.A. DeVore, C.S. Güntürk, V.A. Vaishampayan, A/D conversion with imperfect quantizers. IEEE Trans. Inf. Theory **52**(3), 874–885 (2006)
15. P. Deift, C.S. Güntürk, F. Krahmer, An optimal family of exponentially accurate one-bit sigma-delta quantization schemes. Commun. Pure Appl. Math. **64**(7), 883–919 (2011)
16. D.L. Donoho, Compressed sensing. IEEE Trans. Inf. Theory **52**(4), 1289–1306 (2006)
17. A. Edelman, Eigenvalues and condition numbers of random matrices. SIAM J. Matrix Anal. Appl. **9**(4), 543–560 (1988)
18. J. Feng, F. Krahmer, An RIP approach to Sigma-Delta quantization for compressed sensing. IEEE Signal Process. Lett. **21**(11), 1351–1355 (2014)
19. S. Foucart, H. Rauhut, *A Mathematical Introduction to Compressive Sensing* (Birkhäuser, Basel, 2013)
20. V.K. Goyal, M. Vetterli, N.T. Thao, Quantized overcomplete expansions in $\mathbb{R}^N$: analysis, synthesis, and algorithms. IEEE Trans. Inf. Theory **44**(1), 16–31 (1998)
21. C.S. Güntürk, One-bit sigma-delta quantization with exponential accuracy. Commun. Pure Appl. Math. **56**(11), 1608–1630 (2003)

22. C.S. Güntürk, M. Lammers, A. Powell, R. Saab, Ö. Yılmaz, Sigma delta quantization for compressed sensing, in *2010 44th Annual Conference on Information Sciences and Systems (CISS)*, March 2010, pp. 1–6
23. C.S. Güntürk, M. Lammers, A.M. Powell, R. Saab, Ö. Yılmaz, Sobolev duals for random frames and $\Sigma\Delta$ quantization of compressed sensing measurements. Found. Comput. Math. **13**(1), 1–36 (2013)
24. D.L. Hanson, F.T. Wright, A bound on tail probabilities for quadratic forms in independent random variables. Ann. Math. Stat. **42**(3), 1079–1083 (1971)
25. H. Inose, Y. Yasuda, A unity bit coding method by negative feedback. Proc. IEEE **51**(11), 1524–1535 (1963)
26. M. Iwen, R. Saab, Near-optimal encoding for sigma-delta quantization of finite frame expansions. J. Fourier Anal. Appl. **16**, 1–19 (2013)
27. W.B. Johnson, J. Lindenstrauss, Extensions of Lipschitz mappings into a Hilbert space. Contemp. Math. **26**, 189–206 (1984)
28. F. Krahmer, R. Ward, Lower bounds for the error decay incurred by coarse quantization schemes. Appl. Comput. Harmon. Anal. **32**(1), 131–138 (2012)
29. F. Krahmer, R. Saab, R. Ward, Root-exponential accuracy for coarse quantization of finite frame expansions. IEEE Trans. Inf. Theory **58**(2), 1069–1079 (2012)
30. F. Krahmer, R. Saab, Ö. Yılmaz, Sigma-delta quantization of sub-Gaussian frame expansions and its application to compressed sensing. Inf. Inference **3**(1), 40–58 (2014)
31. M. Lammers, A.M. Powell, Ö. Yılmaz, Alternative dual frames for digital-to-analog conversion in sigma–delta quantization. Adv. Comput. Math. **32**(1), 73–102 (2010)
32. G.G. Lorentz, M. von Golitschek, Y. Makovoz, *Constructive Approximation: Advanced Problems*. Grundlehren der mathematischen Wissenschaften (Springer, New York, 1996)
33. D. Needell, J.A. Tropp, Cosamp: iterative signal recovery from incomplete and inaccurate samples. Appl. Comput. Harmon. Anal. **26**(3), 301–321 (2009)
34. S.R. Norsworthy, R. Schreier, G.C. Temes, (eds.), *Delta-Sigma-Converters: Theory, Design and Simulation* (Wiley/IEEE, New York, 1996)
35. W. Parry, On the β-expansions of real numbers. Acta Math. Acad. Sci. Hungar. **11**, 401–416 (1960)
36. M. Rudelson, R. Vershynin, Non-asymptotic theory of random matrices: extreme singular values, in *Proceedings of the International Congress of Mathematicians*, vol. III (Hindustan Book Agency, New Delhi, 2010), pp. 1576–1602
37. M. Rudelson, R. Vershynin, Hanson-Wright inequality and sub-gaussian concentration. Electron. Commun. Probab. **18**, 1–9 (2013)
38. R. Saab, Ö. Yılmaz, Sparse recovery by non-convex optimization–instance optimality. Appl. Comput. Harmon. Anal. **29**(1), 30–48 (2010)
39. R. Saab, R. Wang, Ö. Yılmaz, Near-optimal compression for compressed sensing, in *Data Compression Conference* (*DCC* 2015) April 2015, pp. 113–122
40. R. Saab, R. Wang, Ö. Yılmaz, Quantization of compressive samples with stable and robust recovery. CoRR, (2015), http://arxiv.org/abs/1504.00087
41. R. Schreier, G.C. Temes, *Understanding Delta-Sigma Data Converters* (Wiley/IEEE Press, New York, 2004)
42. R. Vershynin, Introduction to the non-asymptotic analysis of random matrices, in *Compressed Sensing: Theory and Applications*, ed. by Y.C. Eldar, G. Kutyniok. (Cambridge University Press, Cambridge, 2012), pp. xii+544
43. T. Zhang, Sparse recovery with Orthogonal Matching Pursuit under RIP. IEEE Trans. Inf. Theory **57**(9), 6215–6221 (2011)

Chapter 5
Fourier Operators in Applied Harmonic Analysis

John J. Benedetto and Matthew J. Begué

Abstract We present a panorama describing the pervasiveness of the short-time Fourier transform (STFT) in a host of topics including the following: waveform design and optimal ambiguity function behavior for radar and communications applications; vector-valued ambiguity function theory for multi-sensor environments; finite Gabor frames for deterministic compressive sensing and as a background for the HRT conjecture; generalizations of Fourier frames and non-uniform sampling; and pseudo-differential operator frame inequalities.

5.1 Introduction

5.1.1 The Short Time Fourier Transform (STFT)

Let $\mathbb{Z}$ denote the ring of integers and let $\mathbb{C}$, respectively $\mathbb{R}$, denote the field of complex, respectively real, numbers. Given an integer N, let $\mathbb{Z}/N\mathbb{Z}$ denote the ring of integers modulo N. (We have chosen this well-defined notation, $\mathbb{Z}/N\mathbb{Z}$, and not $\mathbb{Z}_N$, to denote the ring of integers mod N, since we shall deal with primes, p, and $\mathbb{Z}_p$ is universally used to denote the ring of p-adic integers.) Unless otherwise noted, all of the vector spaces herein are complex vector spaces. Let $L^2(\mathbb{R}^d)$ be the space of square-integrable functions defined on the d-dimensional Euclidean space $\mathbb{R}^d$. We let $\hat{\mathbb{R}}^d$ denote $\mathbb{R}^d$ considered as the Fourier, or spectral, domain. We define the Fourier transform of a Schwartz class function, $f \in \mathscr{S}(\mathbb{R}^d)$, as

$$\forall \gamma \in \hat{\mathbb{R}}^d, \quad \hat{f}(\gamma) = \int_{\mathbb{R}^d} f(x) e^{-2\pi i x \cdot \gamma}\, dx.$$

J.J. Benedetto (✉) • M.J. Begué
Department of Mathematics, Norbert Wiener Center, University of Maryland, College Park, MD 20742, USA
e-mail: jjb@math.umd.edu; begue@math.umd.edu

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_5

The Fourier transform can be extended to the space $\mathscr{S}'(\mathbb{R}^d)$ of tempered distributions. In particular, the Fourier transform is well-defined on the Banach algebra $L^1(\mathbb{R}^d)$ and, more generally, on the Banach algebra $M_b(\mathbb{R}^d)$ of bounded Radon measures. Some references on harmonic analysis are [13, 102, 103].

Let $f, g \in L^2(\mathbb{R}^d)$. The *short-time Fourier transform* (STFT) of f with respect to g is the function $V_g f$ defined on $\mathbb{R}^{2d}$ as

$$V_g f(x, \omega) = \int_{\mathbb{R}^d} f(t)\overline{g(t-x)}e^{-2\pi i t\cdot\omega}\, dt,$$

see [51, 52]. The STFT is uniformly continuous on $\mathbb{R}^{2d}$. Furthermore, if $f, g \in L^2(\mathbb{R}^d)$, and $F = \hat{f}$ and $G = \hat{g}$, then the *fundamental identity of time-frequency analysis* is

$$V_g f(x, \omega) = e^{-2\pi i x\cdot\omega} V_G F(\omega, -x).$$

If $f, g \in L^2(\mathbb{R}^d)$, then it can be proved that

$$\left\| V_g f \right\|_{L^2(\mathbb{R}^{2d})} = \|f\|_{L^2(\mathbb{R}^d)} \|g\|_{L^2(\mathbb{R}^d)}. \tag{5.1}$$

Thus, if $\|g\|_{L^2(\mathbb{R}^d)} = 1$, then (5.1) allows us to assert that f is completely determined by $V_g f$. Furthermore, for a fixed "window" function $g \in L^2(\mathbb{R}^d)$ with $\|g\|_{L^2(\mathbb{R}^d)} = 1$, we can recover $f \in L^2(\mathbb{R}^d)$ from its STFT, $V_g f$, by means of the vector-valued integral inversion formula,

$$f = \int_{\mathbb{R}^d} \int_{\hat{\mathbb{R}}^d} V_g f(x, \omega) e_\omega \tau_x g\, d\omega\, dx,$$

where $(e_\omega h)(t) = e^{2\pi i t\cdot\omega} h(t)$ and $\tau_x h(t) = h(t - x)$ represent modulation and translation, respectively, see [51, p. 43].

Remark 1. *a.* Equation (5.1) is *Moyal's formula*. This is a special case of a formulation in 1949 due to José Enrique Moyal in the context of quantum mechanics as a statistical theory. When written in terms of the Wigner distribution from quantum mechanics (1932), this formulation is analogous to the orthogonality relations, that give rise to (5.1) for the STFT. It should also be pointed out that the Ville distribution for signal analysis also appeared in the late 1940s. These ideas are closely related, e.g., see [36, Chapter 8] and [50].

b. Closely related to the STFT and the Wigner and Ville distributions is the *narrow band cross-correlation ambiguity function* of $v, w \in L^2(\mathbb{R})$, defined as

$$\forall (t, \gamma) \in \mathbb{R} \times \hat{\mathbb{R}}, \quad A(v, w)(t, \gamma) = \int_{\mathbb{R}} v(s + t)\overline{w(s)}e^{-2\pi i s\gamma}\, ds.$$

Note that $A(v,w)(t,\gamma) = e^{2\pi it\gamma} V_w u(t,\gamma)$. The *narrow band radar ambiguity function*, $A(v)$, of $v \in L^2(\mathbb{R})$ is defined as

$$\forall (t,\gamma) \in \mathbb{R} \times \hat{\mathbb{R}}, \quad A(v)(t,\gamma) = \int_{\mathbb{R}} v(s+t)\overline{v(s)} e^{-2\pi is\gamma}\, ds$$
$$= e^{\pi it\gamma} \int_{\mathbb{R}} v(s+\tfrac{t}{2})\overline{v(s-\tfrac{t}{2})} e^{-2\pi is\gamma}\, ds.$$

P. M. Woodward (1953) introduced the function, $A(v)$, to describe the effect of range and Doppler on matched filter receivers in radar. Underlying the function itself was his idea of using information theory to optimize resolution in terms of radar waveforms. By comparison with Shannon, Woodward dealt with the problem of mapping information into lower dimensions, prescient of current dimension reduction methodologies. This leads to ambiguities whence, the term, *ambiguity function*. Technical examples of such ambiguity abound in the radar literature, e.g., [80, 100]. In Sections 5.3 and 5.4, we concentrate on discrete versions of $A(v)$.

Whereas the narrow band ambiguity function is essentially time-frequency analysis, the wide band ambiguity function is essentially a wavelet transform, e.g., [8, 64, 106].

c. The STFT can also be formulated in terms of so-called (X, μ) or continuous frames, e.g., see [1, 2, 6, 45, 48].

5.1.2 Outline and theme

Our theme is to interleave and compare various related decompositions whose coefficients are associated with sampled values of a given function. The tentacles of this process are labyrinthine and diverse.

In Section 5.2 we give the necessary background from harmonic analysis. We define balayage, sets of spectral synthesis and strict multiplicity, and provide material from the theory of frames.

Motivated by radar and communications applications of waveform design, Section 5.3 defines and discusses CAZAC sequences and optimal ambiguity function behavior on $\mathbb{Z}/N\mathbb{Z}$, and states a basic result. Because of the importance of dealing effectively with multi-sensor environments, Section 5.4 is devoted to the development of the vector-valued Discrete Fourier Transform (DFT) and proper definitions of vector-valued ambiguity functions. Perhaps surprisingly, this material requires more than using bold-faced notation.

Section 5.5 treats two topics dealing with finite Gabor systems: deterministic compressive sensing in terms of Gabor matrices and conditions to assert the linear independence of finite Gabor sums. The former gives elementary results embedded

in advanced material developed by others. The latter addresses the HRT (Heil, Ramanathan, Topiwala) conjecture, and solves several special cases.

Sections 5.6 and 5.7 use the material on balayage, spectral synthesis, and strict multiplicity to formulate frame inequalities for the STFT and pseudo-differential operators, respectively. It builds on deep work of Beurling and Landau, and it is developed in the spirit of Fourier frames and non-uniform sampling formulas.

We close with a brief appendix showing how the DFT can be used in practice to *compute* Fourier transforms on $\mathbb{R}$. We omit the required error estimates and generalizations. On the other hand, we include the Appendix since this computation requires the Classical Sampling Theorem (Theorem 17), thereby fitting naturally into our theme.

All of the aforementioned topics are unified by the STFT. Further, most of these topics have a long history with contributions by some of the most profound harmonic analysts. Our presentation has to be viewed in that context. Furthermore, our presentation is meant to integrate [5, 6, 15, 18–20, 22]. These references do have a common author, who wants to record the relationships between these topics, but who does not want to give the wrong impression about relative importance by having so many of his papers listed in the references.

5.2 Background from harmonic analysis

5.2.1 Balayage, spectral synthesis, and multiplicity

Let $M_b(G)$ be the algebra of bounded Radon measures on the locally compact abelian group (LCAG), G, with dual group denoted by $\hat{G}$. The space, $M_b(E)$, designates those elements of $M_b(G)$ for which $\text{supp}(\mu) \subseteq E$, see [16]. We use Beurling's definition of balayage from his 1959–60 lectures.

Definition 1. Let $E \subseteq G$, and $\Lambda \subseteq \hat{G}$ be closed sets. *Balayage* is possible for $(E, \Lambda) \subseteq G \times \hat{G}$ if

$$\forall \mu \in M_b(G),\ \exists \nu \in M_b(E) \text{ such that } \hat{\mu} = \hat{\nu} \text{ on } \Lambda.$$

The notion of balayage originated in potential theory by Christoffel in the early 1870s, see [32], and then by Poincaré in 1890, who used the idea of balayage as a method to solve the Dirichlet problem, see [6] for historical background. The set, Λ, of group characters is the analogue of the original role of Λ in balayage as a collection of potential theoretic kernels. Kahane formulated balayage for the harmonic analysis of restriction algebras, see [66].

We shall also require the definition of spectral synthesis due to Wiener and Beurling.

Definition 2. Let $C_b(G)$ be the set of bounded continuous functions on the LCAG G. A closed set $\Lambda \subseteq \hat{G}$ is a *set of spectral synthesis*, or *S-set*, if

$$\forall \mu \in M_b(G) \text{ and } \forall f \in C_b(G), \ \ \text{supp}(\hat{f}) \subseteq \Lambda \text{ and } \hat{\mu} = 0 \text{ on } \Lambda \implies \int_G f \, d\mu = 0, \tag{5.2}$$

see [12].

Remark 2. *a*. Let $A(\hat{G})$ denote the Banach algebra of absolutely convergent Fourier transforms on $\hat{G}$, taken with the transported topology from $L^1(G)$; and let $A'(\hat{G})$ be its dual space. Equivalent to Definition 2, a closed set $\Lambda \subseteq \hat{G}$ is a set of spectral synthesis if for all $T \in A'(\hat{G})$ and for all $\phi \in A(\hat{G})$, if $\text{supp}(T) \subseteq \Lambda$ and $\phi = 0$ on Λ, then $T(\phi) = 0$. This equivalence follows from an elementary functional analysis argument.

b. To determine whether or not $\Lambda \subseteq \hat{G}$ is a set of spectral synthesis is closely related to the problem of determining the ideal structure of the convolution algebra $L^1(G)$, and so a fundamental theorem about sets of spectral synthesis can be thought of in the context of a Nullstellensatz of harmonic analysis. The problem of characterizing S-sets emanated from Wiener's Tauberian theorems and was developed by Beurling in the 1940s. It is "synthesis" in that one wishes to approximate $f \in L^\infty(G)$ in the $\sigma(L^\infty(G), L^1(G))$ (weak-$*$) topology by finite sums of characters, $\gamma : L^\infty(G) \to \mathbb{C}$, that is, each γ is a continuous homomorphism $G \to \{z \in \mathbb{C} : |z| = 1\}$ under multiplication. Further, γ can be considered an element of Λ with $\text{supp}(\delta_\gamma) \subseteq \text{supp}(\hat{f})$, where $\text{supp}(\hat{f})$ is the so-called *spectrum* of f. Such an approximation is elementary to achieve with convolutions of the measures δ_γ, but in this case we lose the essential property that the spectra of the approximants be contained in the spectrum of f.

c. The annihilation property of (5.2) holds when f and μ have balancing smoothness and irregularity. For example, if $\hat{f} \in \mathscr{S}'(\hat{\mathbb{R}}^d)$, $\hat{\mu} = \phi \in \mathscr{S}(\hat{\mathbb{R}}^d)$, and $\phi = 0$ on $\text{supp}(\hat{f})$, then $\hat{f}(\phi) = 0$. Similarly, the same annihilation holds for the pairing of $M_b(\hat{\mathbb{R}}^d)$ and $C_0(\hat{\mathbb{R}}^d)$.

d. The sphere $S^2 \subseteq \hat{\mathbb{R}}^3$ is not an S-set (proven by Schwartz in 1947). Also, every non-discrete $\hat{G}$ has non-S-sets (proven by Malliavin in 1959). Polyhedra are S-sets while the 1/3-Cantor set is an S-set with non-S-subsets, see [12].

Definition 3. A closed set $\Gamma \subseteq \hat{\mathbb{R}}^d$ is a set of *strict multiplicity* if

$$\exists \mu \in M_b(\Gamma) \setminus \{0\} \text{ such that } \lim_{\|x\| \to \infty} |\check{\mu}(x)| = 0,$$

where $\check{\mu}$ is the inverse Fourier transform of μ and $\|x\|$ denotes the standard Euclidean norm of $x \in \mathbb{R}^d$. This is also well-defined for G and $\hat{G}$.

The notion of strict multiplicity was motivated by Riemann's study of sets of uniqueness for trigonometric series. In 1916 Menchov showed that there exist a closed $\Gamma \subseteq \hat{\mathbb{R}}/\mathbb{Z}$ and $\mu \in M(\Gamma) \setminus \{0\}$ such that $|\Gamma| = 0$ and $\check{\mu}(n) = O((\log |n|)^{-1/2})$ as $|n| \to \infty$ ($|\Gamma|$ is the Lebesgue measure of Γ). There have been

intricate refinements of Menchov's result by Bary (1927), Littlewood (1936), Salem [97, 98], Ivašev-Mucatov (1957), and Beurling, et al. see [12].

The above concepts are used in the deep proof of the following theorem.

Theorem 1. *Assume that* $\Lambda \subseteq \hat{\mathbb{R}}^d$ *is an S-set of strict multiplicity, and that balayage is possible for* $(E, \Lambda) \subseteq \mathbb{R}^d \times \hat{\mathbb{R}}^d$. *Let* $\Lambda_\epsilon = \{\gamma \in \hat{\mathbb{R}}^d : dist(\gamma, \Lambda) \leq \epsilon\}$. *There is* $\epsilon_0 > 0$ *such that if* $0 < \epsilon < \epsilon_0$, *then balayage is possible for* (E, Λ_ϵ).

5.2.2 Frames

Definition 4. Let H be a separable Hilbert space, e.g., $H = L^2(\mathbb{R}^d), \mathbb{R}^d$, or $\mathbb{C}^d$. A sequence $F = \{x_i\}_{i\in I} \subseteq H$ is a *frame* for H if there exist constants $A, B > 0$ such that

$$\forall x \in H, \quad A\,\|x\|^2 \leq \sum_{i\in I} |\langle x, x_i\rangle|^2 \leq B\,\|x\|^2\,. \tag{5.3}$$

The constants A and B are *lower* and *upper frame bounds*, respectively. In this paper we shall assume that A is the largest of the lower frame bounds and B is the smallest of the upper frame bounds. In this case, we refer to A and B as *the* lower and upper frame bounds, respectively. If $A = B$, we say that F is a *tight frame* for H. If all the elements of F are of equal norm, we refer to F as an *equal-norm tight frame*. In the case that the tight frame, F, consists of a finite number of elements all with norm equal to 1, then F is a *finite unit-norm tight frame* or *FUNTF*.

Frames are a natural tool for dealing with numerical stability, over-completeness, noise reduction, and robust representation problems. Frames were first defined by Duffin and Schaeffer [39] in 1952 but appeared even earlier in Paley and Wiener's book [86] in 1934. Since then, significant contributions have been made by Beurling [23, 24], Beurling and Malliavin [25, 26], Kahane [65], Landau [79], Jaffard [63], and Seip [85, 99]. Recent expositions on the theory and applications of frames include [34, 75, 76].

Theorem 2. *If* $F = \{x_i\}_{i\in I} \subseteq H$ *is a frame for* H, *then*

$$\forall x \in H, \quad x = \sum_{i\in I} \langle x, S^{-1}x_i\rangle x_i = \sum_{i\in I} \langle x, x_i\rangle S^{-1}x_i,$$

where the map, $S : H \to H$, $x \mapsto \sum_{i\in I} \langle x, x_i\rangle x_i$, *is a well-defined topological isomorphism.*

Theorem 2 illustrates the natural role that frames play in non-uniform sampling formulas, see Example 1.

Let $\Lambda \subseteq \hat{\mathbb{R}}^d$ be a closed set. The *Paley-Wiener space*, PW_Λ, is defined as

$$PW_\Lambda = \{f \in L^2(\mathbb{R}^d) : \operatorname{supp}(\hat{f}) \subseteq \Lambda\}.$$

Definition 5. Let $\Lambda \subseteq \hat{\mathbb{R}}^d$ be a compact set and let $E = \{x_i\}_{i\in I} \subseteq \mathbb{R}^d$ be a sequence. For each $x \in E$, define $f_x = (e_{-x}\mathbb{1}_\Lambda)^\vee \in PW_\Lambda$, where $\mathbb{1}_\Lambda$ denotes the characteristic function of the set Λ. The sequence $\{f_x : x \in E\}$ is a *Fourier frame* for PW_Λ if there exist constants $A, B > 0$ such that

$$\forall f \in PW_\Lambda, \quad A\,\|f\|^2_{L^2(\mathbb{R}^d)} \leq \sum_{x\in E} |f(x)|^2 \leq B\,\|f\|^2_{L^2(\mathbb{R}^d)}\,. \tag{5.4}$$

In fact, (5.4) is a special case of (5.3) since $f(x)$ is an inner product by the Fourier inversion formula.

Definition 6. A sequence $E \subseteq \mathbb{R}^d$ is *separated* if

$$\exists r > 0 \text{ such that } \inf\{\|x - y\| : x, y \in E \text{ and } x \neq y\} \geq r.$$

The following theorem due to Beurling gives a sufficient condition for the existence of Fourier frames in terms of balayage. The proof uses Theorem 1, and its history and structure are analyzed in [6] as part of a more general program.

Theorem 3 (Beurling). *Assume that $\Lambda \subseteq \hat{\mathbb{R}}^d$ is an S-set of strict multiplicity and that $E \subseteq \mathbb{R}^d$ is a separated sequence. Further assume that for every $\gamma \in \Lambda$ and for every compact neighborhood $N(\gamma)$, $\Lambda \cap N(\gamma)$ is a set of strict multiplicity. If balayage is possible for (E, Λ), then $\{(e_{-x}\mathbb{1}_\Lambda)^\vee : x \in E\}$ is a Fourier frame for PW_Λ.*

A host of examples can be deduced satisfying the hypotheses of Theorem 3 as well as Theorem 14 (ahead) from the constructions in [23, Section II].

Example 1. The conclusion of Theorem 3 is the assertion

$$\forall f \in PW_\Lambda, \quad f = \sum_{x\in E} f(x) S^{-1}(f_x) = \sum_{x\in E} \langle f, S^{-1}(f_x)\rangle f_x,$$

where $S(f) = \sum_{x\in E} f(x)\,(e_{-x}\mathbb{1}_\Lambda)^\vee$.

5.3 Optimal ambiguity function behavior on $\mathbb{Z}/N\mathbb{Z}$

Definition 7. A function, $u : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$, is *Constant Amplitude Zero Autocorrelation* (CAZAC) if

$$\forall m \in \mathbb{Z}/N\mathbb{Z}, \quad |u[m]| = 1, \quad \text{(CA)}$$

and

$$\forall m \in \mathbb{Z}/N\mathbb{Z} \setminus \{0\}, \quad \frac{1}{N}\sum_{k=0}^{N-1} u[m+k]\overline{u[k]} = 0. \quad \text{(ZAC)}.$$

Equation (CA) is the condition that u has constant amplitude 1. Equation (ZAC) is the condition that u has zero autocorrelation for $m \in (\mathbb{Z}/N\mathbb{Z}) \setminus \{0\}$, i.e., off the DC-component.

The study of CAZAC sequences and other sequences related to optimal autocorrelation behavior is deeply rooted in several important applications. One of the most prominent applications is the area of waveform design associated with radar and communications. See, e.g., [7, 21, 35, 47, 49, 53, 59, 72, 73, 80, 84, 91, 93, 100, 108, 109]. There has been a striking recent application of low correlation sequences to radar in terms of compressed sensing [60].

There are also purely mathematical roots for the construction of CAZAC sequences. One example, that inspired the role of probability theory in the subject, is due to Wiener, see [17]. Another originated in a question by Per Enflo in 1983 asking about specific Gaussian sequences to deal with the estimation of certain exponential sums, see [96] by Saffari for the role played by Björck, cf. [28, 29].

Do there exist only finitely many non-equivalent CAZAC sequences in $\mathbb{Z}/N\mathbb{Z}$? The answer to this question is "yes" for N prime and "no" for $N = MK^2$, see, e.g., [18, 96].

Definition 8. Let p be a prime number, and so $\mathbb{Z}/p\mathbb{Z}$ is a field. A *Björck CAZAC sequence*, b_p, of length p is defined as

$$\forall k = 0, 1, \ldots, p-1, \quad b_p[k] = e^{i\theta_p(k)},$$

where, for $p \equiv 1 \pmod 4$,

$$\theta_p(k) = \arccos\left(\frac{1}{1+\sqrt{p}}\right)\left(\frac{k}{p}\right)$$

and, for $p \equiv 3 \pmod 4$,

$$\theta_p(k) = \frac{1}{2}\arccos\left(\frac{1-p}{1+p}\right)[(1-\delta_k)\left(\frac{k}{p}\right) + \delta_k].$$

Here, δ_k is the Kronecker delta and $\left(\frac{k}{p}\right)$ is the Legendre symbol defined by

$$\left(\frac{k}{p}\right) = \begin{cases} 0, & \text{if } k \equiv 0 \pmod p, \\ 1, & \text{if } k \equiv n^2 \pmod p \text{ for some } n \in \mathbb{Z}, \\ -1, & \text{if } k \not\equiv n^2 \pmod p \text{ for all } n \in \mathbb{Z}. \end{cases}$$

In [27] Björck proved that Björck sequences are CAZAC sequences, and there is a longstanding collaboration of Björck and Saffari in the general area, see [29] for references.

Definition 9. Let $u : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$. The *discrete narrow band ambiguity function*, $A_N(u) : \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$, of u is defined as

$$\forall (m,n) \in \mathbb{Z}/N\mathbb{Z} \times \widehat{\mathbb{Z}/N\mathbb{Z}}, \quad A_N(u)[m,n] = \frac{1}{N}\sum_{k=0}^{N-1} u[m+k]\overline{u[k]}e^{-2\pi ikn/N}. \quad (5.5)$$

The *discrete autocorrelation* of u is the function, $A_N(u)[\cdot, 0] : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$.

The following estimate is proved in [22]. Notwithstanding the difficulty of proof, its formulation was the result of observations by two of the authors of [22] based on extensive computational work by one of them, viz., Woodworth.

Theorem 4. *Let b_p denote the Björck CAZAC sequence of prime length p, and let $A_p(b_p)$ be the discrete narrow band ambiguity function defined on $\mathbb{Z}/p\mathbb{Z} \times \widehat{\mathbb{Z}/p\mathbb{Z}}$. Then,*

$$\forall (m,n) \in (\mathbb{Z}/p\mathbb{Z} \times \widehat{\mathbb{Z}/p\mathbb{Z}}) \setminus (0,0),$$

$$|A_p(b_p)[m,n]| < \frac{2}{\sqrt{p}} + \frac{4}{p}, \quad \text{if } p \equiv 1 \pmod 4,$$

and

$$|A_p(b_p)[m,n]| < \frac{2}{\sqrt{p}} + \frac{4}{p^{3/2}}, \quad \text{if } p \equiv 3 \pmod 4.$$

The proof of Theorem 4 requires Weil's exponential sum bound [112], which is a consequence of his proof of the Riemann Hypothesis for curves over finite fields [113].

Theorem 4 establishes essentially optimal ambiguity function behavior for b_p, cf. Example 2 and Section 5.5.1. In this regard, and by comparison, if u is any CAZAC sequence of length p, then

$$\frac{1}{\sqrt{p-1}} \le \max\{|A_p(u)[m,n]| : (m,n) \in (\mathbb{Z}/p\mathbb{Z} \times \widehat{\mathbb{Z}/p\mathbb{Z}}) \setminus \{(0,0)\}\}.$$

Example 2. *a.* Let p be a prime number. Alltop [3] defined the sequence, a_p, of length p as

$$\forall k = 0, 1, \ldots, p-1, \quad a_p[k] = e^{2\pi ik^3/p}.$$

Clearly, a_p is of constant amplitude (CA). Alltop proved that

$$\forall m \in (\mathbb{Z}/p\mathbb{Z}) \setminus \{0\} \text{ and } \forall n \in \mathbb{Z}/p\mathbb{Z}, \quad |A_p(a_p)[m,n]| = \frac{1}{\sqrt{p}},$$

which is an excellent bound, cf. Theorem 4 and Section 5.5.1, but also establishes that a_p is *not* a CAZAC sequence in contrast to b_p.

b. The structure of $A_p(b_p)$ is also more complex than that of $A_p(a_p)$ in that $|A_p(b_p)|$ takes values smaller than $1/\sqrt{p}$, a feature that can be used in radar and communications. This goes back to [22] with continuing work by one of those authors and Nava-Tudela.

5.4 The vector-valued DFT and ambiguity functions

5.4.1 The vector-valued DFT

Let $N \geq d$. Form an $N \times d$ matrix using any d columns of the $N \times N$ DFT matrix $(e^{2\pi ijk/N})_{j,k=0}^{N-1}$. The rows of this matrix, up to multiplication by $1/\sqrt{d}$, form a FUNTF for $\mathbb{C}^d$.

Definition 10. Let $N \geq d$ and let $s : \mathbb{Z}/d\mathbb{Z} \to \mathbb{Z}/N\mathbb{Z}$ be injective. The rows $\{E_m\}_{m=0}^{N-1}$ of the $N \times d$ matrix,

$$\left(e^{2\pi i m s(n)/N}\right)_{m,n},$$

form an equal-norm tight frame for $\mathbb{C}^d$, that we call a *DFT frame*.

Definition 11. Let $\{E_k\}_{k=0}^{N-1}$ be a DFT frame for $\mathbb{C}^d$. Given $u : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}^d$, we define the *vector-valued discrete Fourier transform* of u by

$$\forall n \in \mathbb{Z}_N, \quad F(u)(n) = \hat{u}(n) = \sum_{m=0}^{N-1} u(m) * E_{-mn},$$

where $*$ is pointwise (coordinatewise) multiplication. We have that

$$F : \ell^2(\mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/d\mathbb{Z}) \to \ell^2(\mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/d\mathbb{Z})$$

is a linear operator.

The following inversion formula for the vector-valued DFT is proved in [5].

Theorem 5. *The vector-valued Fourier transform is invertible if and only if s, the function defining the DFT frame, has the property that*

$$\forall n \in \mathbb{Z}/d\mathbb{Z}, \quad (s(n), N) = 1.$$

The inverse is given by

$$\forall m \in \mathbb{Z}/N\mathbb{Z}, \quad u(m) = F^{-1}\hat{u}(m) = \frac{1}{N}\sum_{n=0}^{N-1} \hat{u}(n) * E_{mn}.$$

In this case we also have that $F^*F = FF^* = NI$, where I is the identity operator.

In particular, the inversion formula is valid for N prime.

We also note here that vector-valued DFT uncertainty principle inequalities are valid, similar to the results [33] in compressive sensing.

5.4.2 *Vector-valued ambiguity functions and frame multiplication*

5.4.2.1 An ambiguity function for vector-valued functions

Given $u : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}^d$. If $d = 1$, then we can write the discrete ambiguity function, $A_N(u)$, as

$$A_N(u)[m,n] = \frac{1}{N}\sum_{k=0}^{N-1} \langle u(m+k), u(k)e_{nk}\rangle, \tag{5.6}$$

where recall $e_n = e^{2\pi in/N}$. For $d > 1$, the problem of defining a discrete periodic ambiguity function has two natural settings: either it is $\mathbb{C}$-valued or $\mathbb{C}^d$-valued, i.e., $A_N^1(u)[m,n] \in \mathbb{C}$ or $A_N^d(u)[m,n] \in \mathbb{C}^d$. The problem and its solutions were first outlined in [19] (2008).

Let us consider the case $A_N^1(u)[m,n] \in \mathbb{C}$. Motivated by (5.6), we must find a sequence $\{E_k\} \subseteq \mathbb{C}^d$ and an operator, $* : \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}^d$, so that

$$A_N^1(u)[m,n] = \frac{1}{N}\sum_{k=0}^{N-1} \langle u(m+k), u(k) * E_{nk}\rangle \in \mathbb{C} \tag{5.7}$$

defines a meaningful ambiguity function.

To effect this definition, we shall make the following three *ambiguity function assumptions*. First, we assume that there is a sequence $\{E_k\}_{k=0}^{N-1} \subseteq \mathbb{C}^d$ and an operation, $*$, with the property that $E_m * E_n = E_{m+n}$ for $m, n \in \mathbb{Z}/N\mathbb{Z}$. Second, to deal with $u(k) * E_{nk}$ in (5.7), where $u(k) \in \mathbb{C}^d$, we also assume that $\{E_k\}_{k=0}^{N-1} \subseteq \mathbb{C}^d$ is a tight frame for $\mathbb{C}^d$. The multiplication nk is modular multiplication in $\mathbb{Z}/N\mathbb{Z}$. Third, we assume that $* : \mathbb{C}^d \times \mathbb{C}^d \to \mathbb{C}^d$ is bilinear, in particular,

$$\left(\sum_{j=0}^{N-1} c_j E_j\right) * \left(\sum_{k=0}^{N-1} d_k E_k\right) = \sum_{j=0}^{N-1}\sum_{k=0}^{N-1} c_j d_k E_j * E_k.$$

Example 3. Let $\{E_j\}_{j=0}^{N-1} \subseteq \mathbb{C}^d$ satisfy the three ambiguity function assumptions. Then,

$$E_m * E_n = \frac{d^2}{N^2} \sum_{j=0}^{N-1}\sum_{k=0}^{N-1} \langle E_m, E_j\rangle \langle E_n, E_k\rangle E_{j+k}. \tag{5.8}$$

Further, let $\{E_j\}_{j=0}^{N-1}$ be a DFT frame, and let r designate a fixed column. Assume, without loss of generality, that the $N \times d$ matrix for the frame consists of the first d columns of the $N \times N$ DFT matrix. Then (5.8) gives

$$(E_m * E_n)(r) = \frac{e^{2\pi i(m+n)r/N}}{\sqrt{d}} = E_{m+n}(r).$$

Consequently, for DFT frames, $*$ is componentwise multiplication in $\mathbb{C}^d$ with a factor of $\sqrt{d}$. In particular, we have shown that if $u : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}^d$, then $A_N^1(u)$ *is well-defined and can be written explicitly for the case of* DFT *frames and component-wise multiplication,* $*$*, in* $\mathbb{C}^d$.

The definition of $*$ is intrinsically related to the "addition" defined on the indices of the frame elements. In fact, it is not pre-ordained that this "addition" must be modular addition on $\mathbb{Z}/N\mathbb{Z}$, as was the case in Example 3. Formally, we could have $E_m * E_n = E_{m\bullet n}$ for some function $\bullet : \mathbb{Z}/N\mathbb{Z} \times \mathbb{Z}/N\mathbb{Z} \to \mathbb{Z}/N\mathbb{Z}$. The following example exhibits this phenomenon for the familiar case of cross products from the calculus, see [19].

Example 4 ($A_N^1(u)$ for cross product frames). Define $* : \mathbb{C}^3 \times \mathbb{C}^3 \to \mathbb{C}^3$ to be the cross product on $\mathbb{C}^3$. Let $\{i, j, k\}$ be the standard basis for $\mathbb{C}^3$, e.g., $i = (1, 0, 0) \in \mathbb{C}^3$. We have that $i * j = k$, $j * i = -k$, $k * i = j$, $i * k = -j$, $j * k = i$, $k * j = -i$, $i * i = j * j = k * k = 0$. The union of tight frames and the zero vector is a tight frame. In fact, $\{0, i, j, k, -i, -j, -k\}$ is a tight frame for $\mathbb{C}^3$ with frame constant 2. Let $E_0 = 0$, $E_1 = i$, $E_2 = j$, $E_3 = k$, $E_4 = -i$, $E_5 = -j$, and $E_6 = -k$. The index operation corresponding to the frame multiplication is the non-abelian operation $\bullet : \mathbb{Z}/7\mathbb{Z} \times \mathbb{Z}/7\mathbb{Z} \to \mathbb{Z}/7\mathbb{Z}$, where we compute

$$\begin{array}{ccccc} 1\bullet 2 = 3, & 1\bullet 3 = 5, & 1\bullet 4 = 0, & 1\bullet 5 = 6, & 1\bullet 6 = 2, \\ 2\bullet 1 = 6, & 2\bullet 3 = 1, & 2\bullet 4 = 3, & 2\bullet 5 = 0, & 2\bullet 6 = 4, \\ 3\bullet 1 = 2, & 3\bullet 2 = 4, & 3\bullet 4 = 5, & 3\bullet 5 = 1, & 3\bullet 6 = 0, \\ n\bullet n = 0, & & n\bullet 0 = 0\bullet n = 0, & & \text{etc.} \end{array}$$

Thus, the ambiguity function assumptions are valid, with the verification of bilinearity from the definition of the cross product being a tedious calculation. In any case, we can now obtain the following formula:

$$\forall u, v \in \mathbb{C}^3, \quad u * v = \frac{1}{4} \sum_{j=1}^{6} \sum_{k=1}^{6} \langle u, E_j \rangle \langle v, E_k \rangle E_{j \bullet k}.$$

Consequently, $A_N^1(u)$ *is well-defined for the case of this cross product frame and associated bilinear operator*, $*$.

5.4.2.2 Frame multiplication

The essential idea and requirement to define ambiguity functions for $u : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}^d$ is to formulate an effective notion of *frame multiplication*. This was the purpose of the exposition in Section 5.4.2.1 and of [19], where we further noted the substantive role of group theory in this process.

In fact, the set $\{0, \pm i, \pm j, \pm k\}$ of Example 4 is a quasi-group, and the quaternion group of order 8, viz., $\{\pm 1, \pm i, \pm j, \pm k\}$, fits into our theory, see Andrews [4] who develops frame multiplication theory for non-abelian finite groups.

We begin this subsection by defining frame multiplication along the lines motivated in Section 5.4.2.1. Then we shall define frame multiplication associated with a group. Our theory characterizes the groups for which frame multiplication is possible; and, in this case, ambiguity functions can be defined for $\mathbb{C}^d$-valued functions. We shall state some results when the underlying group is abelian, see [5] for the full theory.

Definition 12. *a.* Let $F = \{x_i\}_{i \in I}$ be a frame for a finite dimensional Hilbert space, H, and let $\bullet : I \times I \to I$ be a binary operation. We say $\bullet$ is a *frame multiplication* for F if there is a bilinear map, $* : H \times H \to H$, such that

$$\forall i, j \in I, \quad x_i * x_j = x_{i \bullet j}.$$

Thus, $\bullet$ defines a frame multiplication for F if and only if, for every $x = \sum_{i \in I} a_i x_i$ and $y = \sum_{i \in I} b_i x_i$ in H,

$$x * y = \sum_{i,j \in I} a_i b_j x_{i \bullet j}$$

is well-defined, independent of the frame representations of x and y.

b. Let $(G, \bullet)$ be a finite abelian group, and let $F = \{x_g\}_{g \in G}$ be a frame for a finite dimensional Hilbert space. We say $(G, \bullet)$ defines a *frame multiplication* for F if there is a bilinear map, $* : H \times H \to H$, such that

$$\forall g, h \in G, \quad x_g * x_h = x_{g \bullet h}.$$

Definition 13. Let $(G, \bullet)$ be a finite group. A finite tight frame $F = \{x_g\}_{g \in G}$ for a Hilbert space H is a *G-frame* if there exists $\pi : G \to \mathscr{U}(H)$, a unitary representation of G, such that

$$\forall g, h \in G, \quad \pi(g)x_h = x_{g \bullet h}.$$

Here, $\mathscr{U}(H)$ is the group of unitary operators on H.

Remark 3. The notion of G-frames [105] is a natural one with slightly varying definitions. Definition 13 has been used extensively by Vale and Waldron [111]. Closely related, there are *geometrically uniform frames*, see Bölcskei and Eldar [30], Forney [46], Heath and Strohmer [104], and Slepian [101], as well as a more general formulation due to Han and Larson [54, 55].

The following theorem is proved in [5].

Theorem 6. *Let $(G, \bullet)$ be a finite abelian group and let $F = \{x_g\}_{g \in G}$ be a tight frame for a finite dimensional Hilbert space H. Then G defines a frame multiplication for F if and only if F is a G-frame.*

Definition 14. *a.* Let $(G, \bullet)$ be a finite abelian group of order N. Thus, G has exactly N characters, i.e., N group homomorphisms, $\gamma_j : G \to \mathbb{C}^\times$, where $\mathbb{C}^\times$ is the multiplicative group, $\mathbb{C} \setminus \{0\}$. For each i and j, $\gamma_j(x_i)$ is an Nth root of unity; and the set $\{(\gamma_j(x_i))_{i=1}^N : j = 1, \ldots, N\} \subseteq \mathbb{C}^N$ is an orthonormal basis for $\mathbb{C}^N$.

b. Let $I \subseteq \{1, \ldots, N\}$ have cardinality d. Then, for any $U \in \mathscr{U}(\mathbb{C}^d)$,

$$F = \{U(\gamma_j(x_i))_{j \in I} : i = 1, \ldots, N\} \subseteq \mathbb{C}^d$$

is a frame for $\mathbb{C}^d$, and this is the definition of a *harmonic frame*, see [61, 110].

c. If $(G, \bullet)$ is $\mathbb{Z}/N\mathbb{Z}$ with modular addition, and U is the identity, then F is a DFT-FUNTF.

d. Tight frames $F = \{x_g\}_{g \in G}$ and $H = \{y_g\}_{g \in G}$ for $\mathbb{C}^d$ are said to be *unitarily equivalent* if there exist a unitary map $U \in \mathscr{U}(\mathbb{C}^d)$ and constant $c > 0$ such that

$$\forall g \in G, \quad x_g = cU(y_g).$$

Using Schur's lemma and Maschke's theorem [107], we see the relationship between frame multiplication and harmonic frames in the following result.

Theorem 7. *Let $(G, \bullet)$ be a finite abelian group and let $F = \{x_g\}_{g \in G}$ be a tight frame for $\mathbb{C}^d$. If $(G, \bullet)$ defines a frame multiplication for F, then F is unitarily equivalent to a harmonic frame, and there exist $U \in \mathscr{U}(\mathbb{C}^d)$ and $c > 0$ such that*

$$\forall g, h \in G, \quad \frac{1}{c}U(x_g * x_h) = \frac{1}{c}U(x_g)\frac{1}{c}U(x_h),$$

where the product on the right side is vector pointwise multiplication.

Corollary 1. *Let $F = \{x_k\}_{k \in \mathbb{Z}/N\mathbb{Z}} \subseteq \mathbb{C}^d$ be a tight frame for $\mathbb{C}^d$. If $\mathbb{Z}/N\mathbb{Z}$ defines a frame multiplication for F, then F is unitarily equivalent to a DFT frame.*

5.5 Finite Gabor systems

5.5.1 Gabor matrices

Definition 15. Let $F = \{x_i\}_{i=0}^{N-1} \subseteq \mathbb{C}^d$, $N \geq d$. The *coherence* of F, denoted by $\mu(F)$, is defined as

$$\mu(F) = \max_{j \neq k} \frac{|\langle x_j, x_k \rangle|}{\|x_j\| \, \|x_k\|}.$$

It is well-known that

$$\left(\frac{N-d}{d(N-1)} \right)^{1/2} \leq \mu(F) \leq 1, \tag{5.9}$$

see [92, 114]. The expression on the left side of (5.9) is the *Welch bound* for F. If $\mu(F) = 1$, then there are two elements $x_j, x_k \in F$ that are aligned, and we have maximal coherence. If $\mu(F)$ is the Welch bound, then all of the $x_i \in F$ are spread out in $\mathbb{C}^d$, and we say that we have *maximal incoherence* or *minimal coherence*.

Remark 4. In the case that F is a FUNTF, then $\mu(F)$ is the cosine of the smallest angle between the lines spanned by the elements of the frame. This is not the same as asserting that the coherence is the cosine of the smallest angle between the elements of the frame. For example, in the frame for $\mathbb{R}^2$ that consists of the vectors $(1, 0), (0, 1), (-1, 0)$, and $(0, -1)$, the smallest angle between any two elements is 90 degrees but the smallest angle between any two of the lines spanned by the frame is 0 degrees. Thus, taking the smallest cosine between elements of the frame yields a coherence of 0, whereas taking the smallest cosine between the lines spanned by the frame gives the correct coherence of 1.

A FUNTF, $F = \{x_i\}_{i \in I}$, with $|\langle x_j, x_k \rangle|$ constant for all $j \neq k$ is called an *equiangular frame*. It can be shown that among all FUNTFs of N frame elements in $\mathbb{C}^d$, the equiangular frames are those with minimal coherence. In fact, $\mu(F)$ is the Welch bound if the FUNTF is equiangular. Note that (5.9) implies that an equiangular frame must satisfy $N \leq d^2$, see [104].

Gabor analysis is centered on the interplay of the Fourier transform, translation operators, and modulation operators. Recall that for a given a function $g : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$, we let $\tau_j g(l) = g(l-j)$ and $e_k(l) = e^{2\pi i k l/N}$, for $l = 0, 1, \ldots, N-1$, denote translation and modulation on g, respectively. Let $\top$ denote the transpose operator. The $N \times N^2$ *Gabor matrix*, G, generated by g, is defined as

$$G(g) = [G_0 | G_1 | \cdots | G_{N-1}], \tag{5.10}$$

where each G_j is the $N \times N$ matrix,

$$G_j = \left[e_0 \tau_{j-N} g | e_1 \tau_{j-N} g | \cdots | e_{N-1} \tau_{j-N} g \right],$$

and where each $(e_k\tau_{j-N})^\top$ is the $N \times 1$ column vector, $k = 0, 1, \ldots, N-1$.

Next, we introduce the notation,

$$(g)_k^j = e_k\tau_{j-N}g = \left(e_k(0)\tau_{j-N}g(0), e_k(1)\tau_{j-N}g(1), \ldots, e_k(N-1)\tau_{j-N}g(N-1)\right)^\top.$$

We identify the Gabor matrix $G(g)$ with the set of all these vectors, and so we write

$$G_g = \{(g)_k^j\}_{k,j=0}^{N-1}.$$

This set, G_g, of vectors is referred to as the *Gabor system* generated by g, with corresponding Gabor matrix $G(g)$. Clearly, if $g : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$, then G_g consists of N^2 vectors each of length N, corresponding to all N^2 time-frequency shifts in $\mathbb{Z}/N\mathbb{Z} \times \widehat{\mathbb{Z}/N\mathbb{Z}}$.

The following is elementary to prove, see [87].

Theorem 8. *Given $g : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$, not identically zero. Then, G_g is a tight frame for $\mathbb{C}^N$.*

In this case of Theorem 8, the Gabor system, G_g, is called a *Gabor frame* for $\mathbb{C}^N$, see [87].

Given $g : \mathbb{Z}/N\mathbb{Z} \to \mathbb{C}$, not identically zero. Then, for G_g, the Gabor frame for $\mathbb{C}^N$, (5.9) becomes

$$\sqrt{\frac{N^2 - N}{N(N^2-1)}} = \frac{1}{\sqrt{N+1}} \leq \mu(G_g).$$

The notion of coherence is useful in obtaining sparse solutions to systems of equations. It is well known that for a full rank matrix $A \in \mathbb{C}^{n\times m}$ with $n < m$, there is an infinite number of solutions, $x \in \mathbb{C}^m$, to the system $Ax = b$. One is interested, especially in the context of signal processing and image compression, in finding the sparsest such solution, x, to the linear system. One measure of sparsity of x is by counting the number of nonzero elements, denoted by the ℓ_0 "norm",

$$\|x\|_0 = \#\{i : x(i) \neq 0\}.$$

The sparsest solution to the system $Ax = b$ depends on the coherence of the set of column vectors corresponding to A. A basic theorem is the following.

Theorem 9. *If x is a solution to $Ax = b$, and*

$$\|x\|_0 < \frac{1}{2}\left(1 + \frac{1}{\mu(A)}\right), \tag{5.11}$$

then x is the unique sparsest solution to $Ax = b$, e.g., [31].

Furthermore, the Orthogonal Matching Pursuit (OMP) algorithm constructs x, see [20, 31].

Example 5. We combine Gabor frames, Theorem 9, discrete ambiguity functions, and properties of Alltop and Björck sequences in the following way. The coherence of a Gabor frame, G_g, has an elementary formulaic identity to the discrete ambiguity function of g. Thus, $\mu(G_{a_p})$ and $\mu(G_{b_p})$ are of order $1/\sqrt{p}$ by the comments in Section 5.3. Hence, these values are optimally small because of Welch's theorem, see (5.9).

Consequently, if we let $n = p$ and $m = p^2$ in the setup of Theorem 9, we see that the right side of (5.11) is essentially as large as possible. Thus, the right side of (5.11) is of order $1/2(1 + \sqrt{p})$. Therefore, a large domain is established with regard to unique sparse solutions of $Ax = b$.

5.5.2 The HRT conjecture

Let $g \in L^2(\mathbb{R})$ and let $\Lambda = \{(\alpha_k, \beta_k)\}_{k=1}^N \subseteq \mathbb{R}^2$ be a collection of N distinct points. The *Gabor system* generated by g and Λ is the set,

$$\mathcal{G}(g, \Lambda) = \{e^{2\pi i \beta_k x} g(x - \alpha_k)\}_{k=1}^N.$$

In [57, 58], the Heil, Ramanathan, and Topiwala (HRT) *conjecture* is stated as follows: *Given $g \in L^2(\mathbb{R}) \setminus \{0\}$ and $\Lambda = \{(\alpha_k, \beta_k)\}_{k=1}^N$ as above; then $\mathcal{G}(g, \Lambda)$ is a linearly independent set of functions in $L^2(\mathbb{R})$.* In this case, we shall say that the HRT conjecture holds for $\mathcal{G}(g, \Lambda)$.

Despite its simple statement, the HRT conjecture remains an open problem. On the other hand, some special cases for its validity are known, see [9, 15, 37, 38, 58, 78, 82].

Among the results in [15], the authors prove that the HRT conjecture holds in the setting of ultimately positive functions.

Definition 16. We say that a function $f : \mathbb{R} \to \mathbb{C}$ is *ultimately positive* if

$$\exists x_0 > 0 \text{ such that } \forall x > x_0, \quad f(x) > 0.$$

The HRT results for such functions rely on Kronecker's theorem in Diophantine approximations.

Theorem 10 (Kronecker's theorem). *Let $\{\beta_1, \ldots, \beta_N\} \subseteq \mathbb{R}$ be a linearly independent set over $\mathbb{Q}$, and let $\theta_1, \ldots, \theta_N \in \mathbb{R}$. If $U, \epsilon > 0$, then there exist $p_1, \ldots, p_N \in \mathbb{Z}$ and $u > U$ such that*

$$\forall k = 1, \ldots, N, \quad |\beta_k - p_k - \theta_k| < \epsilon,$$

and, therefore,

$$\forall k = 1, \ldots, N, \quad |e^{2\pi i \beta_k u} - e^{2\pi i \theta_k}| < 4\pi\epsilon.$$

One proof of Kronecker's theorem relies on the Bohr compactification, [12, Theorem 3.2.7]; see [56, Chapter 23], [71, 74] for different proofs.

We shall use the following lemmas in the proof of Theorem 11.

Lemma 1. *Let P be a property that holds for almost every $x \in \mathbb{R}$. For every sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}$, there exists a measurable set $E \subseteq \mathbb{R}$ such that $|\mathbb{R} \setminus E| = 0$ and P holds for $x + u_n$ for each $(n, x) \in \mathbb{N} \times E$.*

Proof. If $E = \bigcap_{n\in\mathbb{N}} \{x : P(x + u_n) \text{ holds}\}$, then P holds for $x + u_n$ for each $(n, x) \in \mathbb{N} \times E$. We know that $|\{x : P(x + u_n) \text{ fails}\}| = 0$ for each $n \in \mathbb{N}$, and so $\left|\bigcup_{n\in\mathbb{N}} \{x : P(x + u_n) \text{ fails}\}\right| = 0$, i.e., $|\mathbb{R} \setminus E| = 0$. □

Lemma 2. *If $A : \mathbb{R}^2 \to \mathbb{R}^2$ is a surjective linear transformation with $\det A = 1$, then there exists a unitary transformation $U_A : L^2(\mathbb{R}) \to L^2(\mathbb{R})$ such that*

$$U_A e_b \tau_a = c_A(a, b) e_v \tau_u U_A,$$

where $(u, v) = A(a, b)$ and $c_A(a, b)$ has the property that $|c_A(a, b)| = 1$.

The operators U_A are *metaplectic transforms*, and form a group of linear transformations of $L^2(\mathbb{R})$ onto itself. Translations, modulations, dilations, and the Fourier transform are all examples of metaplectic transforms on $L^2(\mathbb{R})$.

Theorem 11 (HRT for ultimately positive functions). *Let $g \in L^2(\mathbb{R})$ be ultimately positive and let $\Lambda = \{(\alpha_k, \beta_k)\}_{k=0}^N \subseteq \mathbb{R}^2$ be a set of distinct points with the property that $\{\beta_0, \ldots, \beta_N\}$ is linearly independent over $\mathbb{Q}$. Then, the HRT conjecture holds for $\mathcal{G}(g, \Lambda)$.*

Proof. *i.* We begin by simplifying the setting. First notice that if $\{\beta_0, \ldots, \beta_N\}$ is linearly independent over $\mathbb{Q}$ then $\{\beta_1 - \beta_0, \ldots, \beta_N - \beta_0\}$ is also linearly independent over $\mathbb{Q}$. Furthermore, there exists a linear transformation, A, on $\mathbb{R}^2$ sending (α_0, β_0) to $(0, 0)$ and associated metaplectic transform, $U = U_A$. Then by [15, Lemma 1.3], $\mathcal{G}(g, \Lambda)$ is linearly independent in $L^2(\mathbb{R})$ if and only if $\mathcal{G}(Ug, A(\Lambda))$ is linearly independent in $L^2(\mathbb{R})$. Consequently, without loss of generality, we may assume $(\alpha_0, \beta_0) = (0, 0)$ and $\{\beta_1, \ldots, \beta_N\}$ is linearly independent over $\mathbb{Q}$.

We suppose that $\mathcal{G}(g, \Lambda)$ is linearly dependent in $L^2(\mathbb{R})$ and obtain a contradiction.

ii. Since $\mathcal{G}(g, \Lambda)$ is linearly dependent, there exist constants $c_1, \ldots, c_N \in \mathbb{C}$ not all zero such that

$$g(x) = \sum_{k=1}^{N} c_k e^{2\pi i \beta_k x} g(x - \alpha_k) \quad \text{a.e.} \tag{5.12}$$

In fact, we can take each $c_k \in \mathbb{C} \setminus \{0\}$.
By Kronecker's theorem (Theorem 10) and the linear independence of $\{\beta_1, \ldots, \beta_N\}$ over $\mathbb{Q}$, there exists a sequence $\{u_n\}_{n\in\mathbb{N}} \subseteq \mathbb{R}$ such that $\lim_{n\to\infty} u_n = \infty$, and

$$\forall k = 1, \ldots, N, \quad \lim_{n\to\infty} e^{2\pi i \beta_k u_n} = e^{2\pi i \theta_k}, \tag{5.13}$$

where each

$$\theta_k = \phi_k + 1/4 \quad \text{and} \quad \frac{c_k}{|c_k|} = e^{-2\pi i \phi_k},$$

i.e., we have chosen each θ_k in the application of Theorem 10 so that $e^{2\pi i \theta_k} = |c_k| i / c_k$. Therefore, from (5.13), we compute

$$\forall k = 1, \ldots, N, \quad \lim_{n\to\infty} c_k e^{2\pi i \beta_k u_n} = |c_k| i. \tag{5.14}$$

Then, by Lemma 1, there exists a set $X \subseteq \mathbb{R}$ with $|\mathbb{R} \setminus X| = 0$ such that

$$\forall (n, x) \in \mathbb{N} \times X, \quad g(x + u_n) = \sum_{k=1}^{N} c_k e^{2\pi i \beta_k (x+u_n)} g(x + u_n - \alpha_k). \tag{5.15}$$

iii. Without loss of generality, we may assume that $0 \in X$, for if not, then we can replace g with a translated version of g. Since g is ultimately positive and $u_n \to \infty$, then we may also assume without loss of generality that

$$\forall n \in \mathbb{N} \text{ and } \forall k = 0, 1, \ldots, N, \quad g(u_n - \alpha_k) > 0$$

by simply replacing $\{u_n\}_{n\in\mathbb{N}}$ with a subsequence for which this property *does* hold. Then, by the positivity of g, we divide both sides of (5.15) by $g(x + u_n)$ and evaluate at $x = 0$ to obtain

$$1 = \sum_{k=1}^{N} c_k e^{2\pi i \beta_k u_n} \frac{g(u_n - \alpha_k)}{g(u_n)} = \sum_{k=1}^{N} \left(|c_k| i + c_k e^{2\pi i \beta_k u_n} - |c_k| i \right) \frac{g(u_n - \alpha_k)}{g(u_n)} \tag{5.16}$$

$$\geq \left| \sum_{k=1}^{N} |c_k| i \frac{g(u_n - \alpha_k)}{g(u_n)} \right| - \left| \sum_{k=1}^{N} (c_k e^{2\pi i \beta_k u_n} - |c_k| i) \frac{g(u_n - \alpha_k)}{g(u_n)} \right|$$

$$\geq \sum_{k=1}^{N} |c_k| \frac{g(u_n - \alpha_k)}{g(u_n)} - \sum_{k=1}^{N} |c_k| \left| e^{2\pi i \beta_k u_n} - \frac{|c_k|}{c_k} i \right| \frac{g(u_n - \alpha_k)}{g(u_n)},$$

since $|cd - |c|i| = |c|\,|d - |c|i/c|$ for $c \in \mathbb{C} \setminus \{0\}$ and $d \in \mathbb{C}$.

Now set $\epsilon = 1/(8\pi)$ and apply Theorem 10 to assert that

$$\exists U > 0 \text{ such that } \forall u_n > U \text{ and } \forall k = 1, .., N, \quad \left| e^{2\pi i \beta_k u_n} - \frac{|c_k|}{c_k} i \right| < \frac{1}{2}.$$

This, combined with (5.16), gives

$$\forall u_n > U, \quad 2 \geq \sum_{k=1}^{N} |c_k| \frac{g(u_n - \alpha_k)}{g(u_n)}.$$

Hence, $\{g(u_n - \alpha_k)/g(u_n)\}_{n\in\mathbb{N}}$ is a bounded sequence for each $k = 1, \ldots, N$. Therefore there exists a subsequence $\{v_n\}_{n\in\mathbb{N}}$ of $\{u_n\}_{n\in\mathbb{N}}$ and $r_k \in \mathbb{R}$, $k = 1, \ldots, N$, such that

$$\forall k = 1, \ldots, N, \quad \lim_{n\to\infty} \frac{g(v_n - \alpha_k)}{g(v_n)} = r_k.$$

Then, by the equality of (5.16) and by (5.14), we have

$$1 = \lim_{n\to\infty} \sum_{k=1}^{N} c_k e^{2\pi i \beta_k v_n} \frac{g(v_n - \alpha_k)}{g(v_n)} = \sum_{k=1}^{N} |c_k| r_k i.$$

The left side is real and the right side is imaginary, giving the desired contradiction. □

Definition 17. We say that a function $f : \mathbb{R} \to \mathbb{R}$ is *ultimately decreasing* if

$$\exists x_0 > 0 \text{ such that } \forall y > x > x_0, \quad f(y) \leq f(x).$$

See Figure 5.1 for an illustration of an ultimately positive and ultimately decreasing function on $\mathbb{R}$.

Kronecker's theorem can also be used to prove that the HRT conjecture holds for a four-element Gabor system generated by an ultimately positive function if $g(x)$ and $g(-x)$ are also ultimately decreasing.

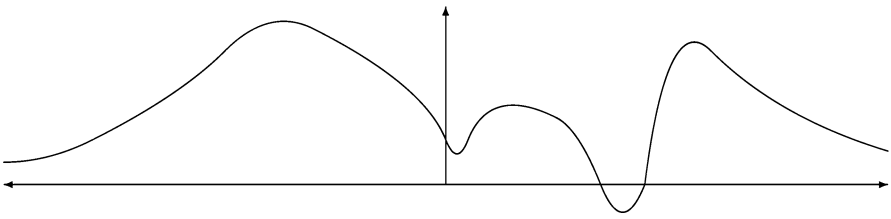


Fig. 5.1 An illustration of an ultimately positive and ultimately decreasing function.

Theorem 12. *Let $g \in L^2(\mathbb{R})$ have the properties that $g(x)$ and $g(-x)$ are ultimately positive and ultimately decreasing, and let $\Lambda = \{(\alpha_k, \beta_k)\}_{k=0}^{3} \subseteq \mathbb{R}^2$ be a set of distinct points. Then, the HRT conjecture holds for $\mathscr{G}(g, \Lambda)$.*

The proof in [15] is omitted here.

Because of the importance of various independent sets in harmonic analysis, Kronecker's theorem motivates the following definition.

Definition 18. Let $E \subseteq \hat{R}$ be compact, and let $C(E)$ be the space of complex-valued continuous functions on E. The set E is a *Kronecker set* if

$$\forall \epsilon > 0 \text{ and } \forall \varphi \in C(E) \text{ for which } \forall x \in E, |\varphi(x)| = 1, \exists x \in \mathbb{R} \text{ such that } \forall \gamma \in E,$$
$$|\varphi(\gamma) - e^{2\pi i x \gamma}| < \epsilon.$$

The resemblance to Kronecker's theorem is apparent. Further, it is clear from Definition 18 how to define Kronecker sets for $E \subseteq \Gamma$, a LCAG. In this setting and going back to the definition of strict multiplicity in Definition 3, we say that a closed set $E \subseteq \Gamma$ is a *Riemann set of uniqueness*, or *U-set*, if it is not a set of multiplicity, i.e., it is not a closed set, F, for which $A_0'(\Gamma) \cap A'(F) \neq \{0\}$. Here, $A_0'(\Gamma) = \{T \in A'(\Gamma) : \hat{T} \in L^\infty(G) \text{ vanishes at infinity}\}$, and $A'(F) = \{T \in A'(\Gamma) : \text{supp}(T) \subseteq F\}$. This definition of a U-set is correct but not highly motivated; however, see [12] for history, motivation, open problems, and important references.

From the point of view of Kronecker's theorem, it is interesting to note that Kronecker sets are sets of strong spectral resolution, i.e., $A'(E) = M_b(E)$, and these, in turn, are U-sets (Malliavin, 1962). There are many other intricacies and open problems in this area combining harmonic analysis, in particular, spectral synthesis, with number theory, see [11, 11, 12, 65, 67, 71, 81, 83, 95].

5.6 Short-time Fourier transform frame inequalities on $\mathbb{R}^d$

Let $g_0(x) = 2^{d/4} e^{-\pi \|x\|^2}$. Then $G_0(\gamma) = \hat{g}_0(\gamma) = 2^{d/4} e^{-\pi \|\gamma\|^2}$ and $\|g_0\|_{L^2(\mathbb{R}^d)} = 1$, see [16] for properties of g_0.

Definition 19. The *Feichtinger algebra*, $\mathscr{S}_0(\mathbb{R}^d)$, is defined as

$$\mathscr{S}_0(\mathbb{R}^d) = \{f \in L^2(\mathbb{R}^d) : \|f\|_{\mathscr{S}_0(\mathbb{R}^d)} = \left\|V_{g_0} f\right\|_{L^1(\mathbb{R}^{2d})} < \infty\}.$$

The Fourier transform of $\mathscr{S}_0(\mathbb{R}^d)$ is an isometric isomorphism onto itself, and, in particular, $f \in \mathscr{S}_0(\mathbb{R}^d)$ if and only if $F \in \mathscr{S}_0(\hat{\mathbb{R}}^d)$, see, e.g., [40–42, 44, 51].

The Feichtinger algebra provides a natural setting for proving non-uniform sampling theorems for the STFT analogous to Beurling's non-uniform sampling theorem, Theorem 3, for Fourier frames. This setting extends to more general modulation spaces. The theories for the STFT and modulation spaces are given in [51].

The following is Gröchenig's non-uniform Gabor frame theorem, and it was also influenced by earlier work with Feichtinger, see [50, Theorem S] and [51], cf. [43, 44] for a precursor of this result.

Theorem 13. *Given any $g \in \mathscr{S}_0(\mathbb{R}^d)$. There is $r = r(g) > 0$ such that if $E = \{(s_n, \sigma_n)\}_{n=1}^{\infty} \subseteq \mathbb{R}^d \times \hat{\mathbb{R}}^d$ is a separated sequence with the property that*

$$\bigcup_{n=1}^{\infty} \overline{B((s_n, \sigma_n), r(g))} = \mathbb{R}^d \times \hat{\mathbb{R}}^d,$$

then the frame operator $S = S_{g,E}$ defined by

$$S_{g,E} f = \sum_{n=1}^{\infty} \langle f, \tau_{s_n} e_{\sigma_n} g \rangle \tau_{s_n} e_{\sigma_n} g,$$

is invertible on $\mathscr{S}_0(\mathbb{R}^d)$. Further, every $f \in \mathscr{S}_0(\mathbb{R}^d)$ has a non-uniform Gabor expansion,

$$f = \sum_{n=1}^{\infty} \langle f, \tau_{s_n} e_{\sigma_n} g \rangle S_{g,E}^{-1}(\tau_{s_n} e_{\sigma_n} g),$$

where the series converges unconditionally in $\mathscr{S}_0(\mathbb{R}^d)$.

It should be noted that the set E depends on g.

The following is proved in [6] and can be compared with Theorem 13.

Theorem 14. *Let $E = \{(s_n, \sigma_n)\}_{n=1}^{\infty} \subseteq \mathbb{R}^d \times \hat{\mathbb{R}}^d$ be a separated sequence; and let $\Lambda \subseteq \hat{\mathbb{R}}^d \times \mathbb{R}^d$ be an S-set of strict multiplicity that is compact, convex, and symmetric about $0 \in \hat{\mathbb{R}}^d \times \mathbb{R}^d$. Assume balayage is possible for (E, Λ). Further, let $g \in L^2(\mathbb{R}^d)$, $\hat{g} = G$, have the property that $\|g\|_{L^2(\mathbb{R}^d)} = 1$. We have that*

$$\exists A, B > 0, \text{ such that } \forall f \in \mathscr{S}_0(\mathbb{R}^d), \text{ for which } \operatorname{supp}(\widehat{V_g f}) \subseteq \Lambda,$$

$$A \|f\|_{L^2(\mathbb{R}^d)}^2 \leq \sum_{n=1}^{\infty} |V_g f(s_n, \sigma_n)|^2 \leq B \|f\|_{L^2(\mathbb{R}^d)}^2 .$$

Consequently, the frame operator $S = S_{g,E}$ is invertible in $L^2(\mathbb{R}^d)$-norm on the subspace of $\mathscr{S}_0(\mathbb{R}^d)$, whose elements f have the property that $\operatorname{supp}(\widehat{V_g f}) \subseteq \Lambda$.

Moreover, every $f \in \mathscr{S}_0(\mathbb{R}^d)$ satisfying the support condition, $\operatorname{supp}(\widehat{V_g f}) \subseteq \Lambda$, has a non-uniform Gabor expansion,

$$f = \sum_{n=1}^{\infty} \langle f, \tau_{s_n} e_{\sigma_n} g \rangle S_{g,E}^{-1}(\tau_{s_n} e_{\sigma_n} g),$$

where the series converges unconditionally in $L^2(\mathbb{R}^d)$.

It should be noted that here the set E does not depend on g.

Example 6. In comparing Theorem 14 with Theorem 13, a possible weakness of the former is the dependence of the set E on g, whereas a possible weakness of the latter is the hypothesis that $\mathrm{supp}(\widehat{V_g f}) \subseteq \Lambda$. We now show that this latter constraint is of no major consequence.

Let $f, g \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$. We know that $V_g f \in L^2(\mathbb{R}^d \times \hat{\mathbb{R}}^d)$, and

$$\widehat{V_g f}(\zeta, z) = \int_{\hat{\mathbb{R}}^d} \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d} f(t) g(t-x) e^{-2\pi i t\cdot\omega}\, dt \right) e^{-2\pi i (x\cdot\zeta + z\cdot\omega)}\, dx\, d\omega.$$

The right side is

$$\int_{\hat{\mathbb{R}}^d} \int_{\mathbb{R}^d} f(t) \left(\int_{\mathbb{R}^d} g(t-x) e^{-2\pi i x\cdot\zeta}\, dx \right) e^{-2\pi i t\cdot\omega} e^{-2\pi i z\cdot\omega}\, dt\, d\omega,$$

where the interchange in integration follows from the Fubini-Tonelli theorem and the hypothesis that $f, g \in L^1(\mathbb{R}^d)$. This, in turn, is

$$\begin{aligned} &\hat{g}(-\zeta) \textstyle\int_{\hat{\mathbb{R}}^d} \left(\int_{\mathbb{R}^d} f(t) e^{-2\pi i t\cdot\zeta} e^{-2\pi i t\cdot\omega}\, dt \right) e^{-2\pi i z\cdot\omega}\, d\omega \\ &= \hat{g}(-\zeta) \textstyle\int_{\hat{\mathbb{R}}^d} \hat{f}(\zeta+\omega) e^{-2\pi i z\cdot\omega}\, d\omega = e^{-2\pi i z\cdot\zeta} f(-z) \hat{g}(-\zeta). \end{aligned}$$

Consequently, we have shown that

$$\forall f, g \in L^1(\mathbb{R}^d) \cap L^2(\mathbb{R}^d), \quad \widehat{V_g f}(\zeta, z) = e^{-2\pi i z\cdot\zeta} f(-z) \hat{g}(-\zeta). \tag{5.17}$$

Let $d = 1$ and let $\Lambda = [-\Omega, \Omega] \times [-T, T] \subseteq \hat{\mathbb{R}}^d \times \mathbb{R}^d$. We can choose $g \in PW_{[-\Omega,\Omega]}$, where $\hat{g}$ is even and smooth enough so that $g \in L^1(\mathbb{R})$. For this window, g, we can take any even $f \in L^2(\mathbb{R})$ which is supported in $[-T, T]$. Equation (5.17) applies. Consequently, in this case, $\mathrm{supp}(\widehat{V_g f}) \subseteq \Lambda$. Clearly, this particular example can be extended significantly, whence our assertion that the hypothesis, $\mathrm{supp}(\widehat{V_g f}) \subseteq \Lambda$, in Theorem 14 is of no major consequence.

5.7 Pseudo-differential operator frame inequalities on $\mathbb{R}^d$

Let $\sigma \in \mathscr{S}'(\mathbb{R}^d \times \hat{\mathbb{R}}^d)$. The operator, K_σ, formally defined as

$$(K_\sigma f)(x) = \int_{\hat{\mathbb{R}}^d} \sigma(x, \gamma) \hat{f}(\gamma) e^{2\pi i x\cdot\gamma}\, d\gamma,$$

is the *pseudo-differential operator* with Kohn-Nirenberg symbol, σ, see [51] Chapter 14, [52] Chapter 8, [62], and [102]. For consistency with the notation and setting of the previous sections, we shall define pseudo-differential operators, K_s, with tempered distributional Kohn-Nirenberg symbols, $s \in \mathscr{S}'(\mathbb{R}^d \times \hat{\mathbb{R}}^d)$, as

$$(K_s\hat{f})(\gamma) = \int_{\mathbb{R}} s(y,\gamma) f(y) e^{-2\pi i y\cdot\gamma}\, dy.$$

Furthermore, we shall deal with Hilbert-Schmidt operators, $K : L^2(\hat{\mathbb{R}}^d) \to L^2(\hat{\mathbb{R}}^d)$; and these, in turn, can be represented as $K = K_s$, where $s \in L^2(\mathbb{R}^d \times \hat{\mathbb{R}}^d)$. Recall that $K : L^2(\hat{\mathbb{R}}^d) \to L^2(\hat{\mathbb{R}}^d)$ is a *Hilbert-Schmidt operator* if

$$\sum_{n=1}^{\infty} \|Ke_n\|^2_{L^2(\hat{\mathbb{R}}^d)} < \infty$$

for some orthonormal basis, $\{e_n\}_{n=1}^{\infty}$, for $L^2(\hat{\mathbb{R}}^d)$, in which case the *Hilbert-Schmidt norm* of K is defined as

$$\|K\|_{HS} = \left(\sum_{n=1}^{\infty} \|Ke_n\|^2_{L^2(\hat{\mathbb{R}}^d)}\right)^{1/2},$$

and $\|K\|_{HS}$ is independent of the choice of orthonormal basis.

The following theorem on Hilbert-Schmidt operators can be found in [94].

Theorem 15. *If $K : L^2(\hat{\mathbb{R}}^d) \to L^2(\hat{\mathbb{R}}^d)$ is a bounded linear mapping and $(K\hat{f})(\gamma) = \int_{\hat{\mathbb{R}}^d} m(\gamma,\lambda)\hat{f}(\lambda)\, d\lambda$, for some measurable function m, then K is a Hilbert-Schmidt operator if and only if $m \in L^2(\hat{\mathbb{R}}^{2d})$ and, in this case, $\|K\|_{HS} = \|m\|_{L^2(\mathbb{R}^{2d})}$.*

The following theorem about pseudo-differential operator frame inequalities is proved in [6].

Theorem 16. *Let $E = \{x_n\} \subseteq \mathbb{R}^d$ be a separated sequence, that is symmetric about $0 \in \mathbb{R}^d$; and let $\Lambda \subseteq \hat{\mathbb{R}}^d$ be an S-set of strict multiplicity, that is compact, convex, and symmetric about $0 \in \hat{\mathbb{R}}^d$. Assume balayage is possible for (E,Λ). Furthermore, let K be a Hilbert-Schmidt operator on $L^2(\hat{\mathbb{R}}^d)$ with pseudo-differential operator representation,*

$$(K\hat{f})(\gamma) = (K_s\hat{f})(\gamma) = \int_{\mathbb{R}^d} s(y,\gamma) f(y) e^{-2\pi i y\cdot\gamma}\, dy,$$

where $s_\gamma(y) = s(y,\gamma) \in L^2(\mathbb{R}^d \times \hat{\mathbb{R}}^d)$ is the Kohn-Nirenberg symbol and where we furthermore assume that

$$\forall \gamma \in \hat{\mathbb{R}}^d, \quad s_\gamma \in C_b(\mathbb{R}^d) \text{ and } \operatorname{supp}(s_\gamma e_{-\gamma})\hat{} \subseteq \Lambda.$$

Then,

$$\exists A, B > 0 \text{ such that } \forall f \in L^2(\mathbb{R}^d) \setminus \{0\},$$

$$A \frac{\left\| K_s \hat{f} \right\|^4_{L^2(\hat{\mathbb{R}}^d)}}{\|f\|^2_{L^2(\mathbb{R}^d)}} \leq \sum_{x \in E} |\langle (K_s \hat{f})(\cdot), \overline{s(x,\cdot)} e_x(\cdot) \rangle|^2 \leq B \|s\|^2_{L^2(\mathbb{R}^d \times \hat{\mathbb{R}}^d)} \left\| K_s \hat{f} \right\|^2_{L^2(\hat{\mathbb{R}}^d)}.$$

Example 7. We shall define a Kohn-Nirenberg symbol class whose elements, s, satisfy the hypotheses of Theorem 16.

Choose $\{\lambda_j\} \subseteq \text{int}(\Lambda)$, where Λ is as described in Theorem 16. Choose $\{a_j\} \subseteq C_b(\mathbb{R}^d) \cap L^2(\mathbb{R}^d)$ and $\{b_j\} \subseteq C_b(\hat{\mathbb{R}}^d) \cap L^2(\hat{\mathbb{R}}^d)$ with the following properties:

i. $\sum_{j=1}^{\infty} |a_j(y) b_j(y)|$ is uniformly bounded and converges uniformly on $\mathbb{R}^d \times \hat{\mathbb{R}}^d$;
ii. $\sum_{j=1}^{\infty} \|a_j\|_{L^2(\mathbb{R}^d)} \|b_j\|_{L^2(\hat{\mathbb{R}}^d)} < \infty$;
iii. $\forall j \in \mathbb{N}, \exists \epsilon_j > 0$ such that $\overline{B(\lambda_j, \epsilon_j)} \subseteq \Lambda$ and $\text{supp}(\hat{a}_j) \subseteq \overline{B(0, \epsilon_j)}$.

These conditions are satisfied for a large class of functions a_j and b_j.

The Kohn-Nirenberg symbol class consisting of functions, s, defined by

$$s(y, \gamma) = \sum_{j=1}^{\infty} a_j(y) b_j(\gamma) e^{-2\pi i y \cdot \lambda_j}$$

satisfy the hypotheses of Theorem 16. To see this, first note that condition *i* guarantees that if we set $s_\gamma(y) = s(y, \gamma)$, then

$$\forall \gamma \in \hat{\mathbb{R}}^d, \quad s_\gamma \in C_b(\mathbb{R}^d).$$

Condition *ii* allows us to assert that $s \in L^2(\mathbb{R}^d \times \hat{\mathbb{R}}^d)$ since Minkowski's inequality can be used to make the following estimate:

$$\begin{aligned} \|s\|_{L^2(\mathbb{R}^d \times \hat{\mathbb{R}}^d)} &\leq \sum_{j=1}^{\infty} \left(\int_{\hat{\mathbb{R}}^d} \int_{\mathbb{R}^d} \left| b_j(\gamma) a_j(y) e^{-2\pi i y \cdot (\lambda_j - \gamma)} \right|^2 dy\, d\gamma \right)^{1/2} \\ &= \sum_{j=1}^{\infty} \|a_j\|_{L^2(\mathbb{R}^d)} \|b_j\|_{L^2(\hat{\mathbb{R}}^d)}. \end{aligned}$$

Finally, using condition *iii*, we obtain the support hypothesis, $\text{supp}(s_\gamma e_{-\gamma})\hat{} \subseteq \Lambda$, of Theorem 16 for each $\gamma \in \hat{\mathbb{R}}^d$, because of the following calculations:

$$\text{supp}(s_\gamma e_{-\gamma})\hat{}(\omega) = \sum_{j=1}^{\infty} b_j(\gamma) (\hat{a}_j * \delta_{-\lambda_j})(\omega)$$

and, for each j,

$$\text{supp}(\hat{a}_j * \delta_{-\lambda_j}) \subseteq \overline{B(0,\epsilon_j)} + \{\lambda_j\} \subseteq \overline{B(\lambda_j,\epsilon_j)} \subseteq \Lambda.$$

Remark 5. Pfander and collaborators combine the theory of Gabor frames and Hilbert-Schmidt operators to obtain results in *operator sampling*. The goal of operator sampling is to determine an operator completely from its action on a single input function or distribution. The question of determining which operators can be identified in this way was addressed in basic work of Kailath [68–70] and Bello [10], who found that the identifiability of a communication channel is characterized by the area of the support of its so-called *spreading function*. The spreading function, $\eta_H(t,\nu)$, of the Hilbert-Schmidt operator, H, on $L^2(\mathbb{R})$ is the *symplectic Fourier transform* of its Kohn-Nirenberg symbol, σ, viz,

$$\eta_H(t,\nu) = \int_{\hat{\mathbb{R}}}\int_{\mathbb{R}} \sigma(x,\gamma)\hat{f}(\gamma)e^{-2\pi i(\nu x-\gamma t)}\,dx\,d\gamma;$$

and we have the representation,

$$Hf(x) = \int_{\mathbb{R}}\int_{\hat{\mathbb{R}}} \eta_H(t,\nu)\tau_t e_\nu f(x)\,d\nu\,dt.$$

In this sense, an operator H whose spreading function has compact support can be said to have a *bandlimited symbol*. This motivates the definition of an *operator Paley-Wiener space* and an associated sampling theorem [89]. The aforementioned communications application was put on a firm mathematical footing, first proving Kailath's conjectures in [77] (Kozek and Pfander) and then proving Bello's assertions in [89] (Pfander and Walnut). Results and an overview of the subject are given in [88, 90].

Appendix

The Classical Sampling Theorem goes back to papers by Cauchy (1840s), see [13, Theorem 3.10.10] for proofs of Theorem 17. It has had a significant impact on various topics in mathematics, including number theory and interpolation theory, long before Shannon's application of it in communications.

Theorem 17 (Classical Sampling Theorem). *Let $T, \Omega > 0$ satisfy the condition that $0 < 2T\Omega \leq 1$, and let s be an element of the Paley-Wiener space $PW_{1/(2T)}$ satisfying the condition that $\hat{s} = S = 1$ on $[-\Omega,\Omega]$ and $S \in L^\infty(\hat{\mathbb{R}})$. Then*

$$\forall f \in PW_\Omega, \quad f = T\sum_{n\in\mathbb{Z}} f(nT)\tau_{nT}s, \tag{5.18}$$

where the convergence of (5.18) *is in the* $L^2(\mathbb{R})$ *norm and uniformly in* $\mathbb{R}$. *One possible sampling function s is*

$$s(t) = \frac{\sin(2\pi\Omega t)}{\pi t}.$$

We can compute Fourier transforms numerically using the following result, whose proof, see [14], requires Theorem 17.

Theorem 18. *Let* $T, \Omega > 0$ *satisfy the property that* $2T\Omega = 1$, *let* $N \geq 2$ *be an even integer, and let* $f \in PW_\Omega \cap L^1(\mathbb{R})$. *Consider the dilation* $f_T(t) = Tf(Tt)$ *as a continuous function on* $\mathbb{R}$, *as well as a function on* $\mathbb{Z}$ *defined by* $m \mapsto f_T[m]$, *where* $f_T[m] = f_T(m)$. *Assume that* $f_T \in \ell^1(\mathbb{Z})$. *Then for every integer* $n \in [-\frac{N}{2}, \frac{N}{2}]$, *we have*

$$\hat{f}\left(\frac{2\Omega n}{N}\right) = \hat{f}\left(\frac{n}{NT}\right) = \sum_{m=0}^{N-1} (f_T)_N^\circ[m] W_N^{mn}, \tag{5.19}$$

where $W_N = e^{-2\pi i/N}$ *and* $(f_T)_N^\circ[m] = \sum_{k\in\mathbb{Z}} f_T[m - kN]$.

In practice, the computation (5.19) requires natural error estimates and the FFT.

Acknowledgements The first named author gratefully acknowledges the support of MURI-ARO Grant W911NF- 09-1-0383, NGA Grant 1582-08-1-0009, and DTRA Grant HDTRA1-13-1-0015. The second named author gratefully acknowledges the support of the Norbert Wiener Center at the University of Maryland, College Park.

References

1. S.T. Ali, J.-P. Antoine, J.-P. Gazeau, Continuous frames in Hilbert space. Ann. Phys. **222**, 1–37 (1993)
2. S.T. Ali, J.-P. Antoine, J.-P. Gazeau, *Coherent States, Wavelets and Their Generalizations*. Graduate Text in Contemporary Physics (Springer, New York, 2000)
3. W.O. Alltop, Complex sequences with low periodic correlations. IEEE Trans. Inf. Theory **26**(3), 350–354 (1980)
4. T. Andrews, Representations of finite groups for frame representation theory (submitted)
5. T. Andrews, J.J. Benedetto, J.J. Donatelli, Frame multiplication theory and vector-valued ambiguity functions (submitted)
6. E. Au-Yeung, J.J. Benedetto, Generalized Fourier frames in terms of balayage (submitted)
7. L. Auslander, P.E. Barbano, Communication codes and Bernoulli transformations. Appl. Comput. Harmon. Anal. **5**(2), 109–128 (1998)
8. L. Auslander, I. Gertner, Wide-band ambiguity function and ax+ b group. Inst. Math. Its Appl. **22**, 1 (1990)
9. R. Balan, A noncommutative Wiener lemma and a faithful tracheal state on Banach algebra of time-frequency operators. Trans. Am. Math. Soc. **360**, 3921–3941 (2008)

10. P.A. Bello, Measurement of random time-variant linear channels. IEEE Trans. Inf. Theory **15**(4), 469–475 (1969)
11. J.J. Benedetto, *Harmonic Analysis on Totally Disconnected Sets*. Springer Lecture Notes, vol. 202 (Springer, New York, 1971)
12. J.J. Benedetto, *Spectral Synthesis* (Academic, New York, 1975)
13. J.J. Benedetto, *Harmonic Analysis and Applications* (CRC Press, Boca Raton 1997)
14. J.J. Benedetto, *Sampling, Sparsity, and CAZAC Sequences*. Lecture Notes in Applied and Numerical Harmonic Analysis (Springer/Birkhäuser, New York, 2015)
15. J.J. Benedetto, A. Bourouihiya, Lineaer independence of finite Gabor systems determined by behavior at infinity. J. Geom. Anal. (2014)
16. J.J. Benedetto, W. Czaja, *Integration and Modern Analysis*. Birkhäuser Advanced Texts (Springer/Birkhäuser, New York, 2009)
17. J.J. Benedetto, S. Datta, Construction of infinite unimodular sequences with zero autocorrelation. Adv. Comput. Math. **32**, 191–207 (2010)
18. J.J. Benedetto, J.J. Donatelli, Ambiguity function and frame theoretic properties of periodic zero autocorrelation waveforms. IEEE J. Spec. Top Signal Process. **1**, 6–20 (2007)
19. J.J. Benedetto, J.J. Donatelli, Frames and a vector-valued ambiguity function, in *Asilomar Conference on Signals, Systems, and Computers* (Oct 2008)
20. J.J. Benedetto, A. Nava-Tudela, Sampling in image representation and compression, in *Sampling Theory in Honor of Paul L. Bützer's 85th Birthday* (Springer/Birkhäuser, New York, 2014)
21. J.J. Benedetto, I. Konstantinidis, M. Rangaswamy, Phase-coded waveforms and their design. IEEE Signal Process. Mag. **26**(1), 22–31 (2009)
22. J.J. Benedetto, R.L. Benedetto, J.T. Woodworth, Optimal ambiguity functions and Weil's exponential sum bound. J. Fourier Anal. Appl. **18**(3), 471–487 (2012)
23. A. Beurling, Local harmonic analysis with some applications to differential operators, in *Some Recent Advances in the Basic Sciences*, vol. 1 (Proceedings Annual Science Conference, Belfer Graduate School of Science, Yeshiva University, New York, 1962–1964), pp. 109–125
24. A. Beurling, *The Collected Works of Arne Beurling. Vol. 2. Harmonic Analysis* (Springer/Birkhäuser, New York, 1989)
25. A. Beurling, P. Malliavin, On Fourier transforms of measures with compact support. Acta Math. **107**, 291–309 (1962)
26. A. Beurling, P. Malliavin, On the closure of characters and the zeros of entire functions. Acta Math. **118**, 79–93 (1967)
27. G. Björck, Functions of modulus one on $\mathbf{Z}_p$ whose Fourier transforms have constant modulus, in *Proceedings of the A. Haar Memorial Conference*, vol. I, II (Budapest, 1985). Colloquim Mathematical Society János Bolyai, vol. 49 (North-Holland, Amsterdam, 1987), pp. 193–197
28. G. Björck, Functions of modulus one on $\mathbb{Z}_n$ whose Fourier transforms have constant modulus, and cyclic n-roots, in *Proceedings of 1989 NATO Advanced Study Institute on Recent Advances in Fourier Analysis and its Applications* (1990), pp. 131–140
29. G. Björck, B. Saffari, New classes of finite unimodular sequences with unimodular Fourier transforms. Circulant Hadamard matrices with complex entries. C. R. Acad. Sci. Paris **320**, 319–324 (1995)
30. H. Bölcskei , Y.C. Eldar, Geometrically uniform frames. IEEE Trans. Inf. Theory **49**(4), 993–1006 (2003)
31. A.M Bruckstein, D.L Donoho, M. Elad, From sparse solutions of systems of equations to sparse modeling of signals and images. SIAM Rev. **51**(1), 34–81 (2009)
32. P.L. Butzer, F. Fehër, *E. B. Christoffel - The Influence of his Work on Mathematics and the Physical Sciences* (Springer/Birkhäuser, New York, 1981)
33. E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. IEEE Trans. Inf. Theory **52**(2), 489–509 (2006)
34. O. Christensen, *An Introduction to Frames and Riesz Bases* (Springer/Birkhäuser, New York, 2003)

35. D.C. Chu, Polyphase codes with good periodic correlation properties. IEEE Trans. Inf. Theory **18**, 531–532 (1972)
36. L. Cohen, *Time-Frequency Analysis: Theory and Applications* (Prentice-Hall Inc, Upper Saddle River, 1995)
37. C. Demeter, Linear independence of time frequency translates for special configurations. Math. Res. Lett. **17**, 761–799 (2010)
38. C. Demeter, A. Zaharescu, Proof of the HRT conjecture for (2, 2) configurations. J. Math. Anal. Appl. **388**(1), 151–159 (2012)
39. R.J. Duffin, A.C. Schaeffer, A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**, 341–366 (1952)
40. H.G. Feichtinger, On a new Segal algebra. Monatsh. Math. **92**, 269–289 (1981)
41. H.G. Feichtinger, Modulation spaces on locally compact abelian groups, in *Proceedings of International Conference on Wavelets and Applications* (1983), pp. 1–56
42. H.G. Feichtinger, Wiener amalgams over Euclidean spaces and some of their applications, in *Proceedings Conference of Function Spaces* (Edwardsville, IL, 1990). Lecture Notes in Pure and Applied Mathematics, vol. 136 (1990), pp. 123–137
43. H.G. Feichtinger, K.H. Gröchenig, A unified approach to atomic decompositions via integrable group representations, in *Function Spaces and Applications* (Lund 1986). Lecture Notes in Mathematics, vol. 1302 (Springer, Berlin, 1988), pp. 52–73
44. H.G. Feichtinger, K.H. Gröchenig, Banach spaces related to integrable group representations and their atomic decompositions. J. Funct. Anal. **86**, 307–340 (1989)
45. M. Fornasier, H. Rauhut, Continuous frames, function spaces, and the discretization problem. J. Fourier Anal. Appl. **11**, 245–287 (2005)
46. G.D. Forney, Geometrically uniform codes. IEEE Trans. Inf. Theory **37**(5), 1241–1260 (1991)
47. R.L. Frank, S.A. Zadoff, Phase shift pulse codes with good periodic correlation properties. IRE Trans. Inf. Theory **8**, 381–382 (1962)
48. J.-P. Gabardo, D. Han, Frames associated with measurable spaces. frames. Adv. Comput. Math. **18**, 127–147 (2003)
49. S.W. Golomb, G. Gong, *Signal Design for Good Correlation* (Cambridge University Press, Cambridge, 2005)
50. K.H. Gröchenig, Describing functions: atomic decompositions versus frames. Monatsh. Math. **112**, 1–42 (1991)
51. K.H. Gröchenig, *Foundations of Time-Frequency Analysis*. Applied and Numerical Harmonic Analysis (Springer/Birkhäuser, New York, 2001)
52. K.H. Gröchenig, A pedestrian's approach to pseudodifferential operators, in *Harmonic Analysis and Applications*, ed. by C. Heil (Springer/Birkhäuser, New York, 2006), pp. 139–169
53. J.-C. Guey, M.R. Bell, Diversity waveform sets for delay-doppler imaging. IEEE Trans. Inf. Theory **44**(4), 1504–1522 (1998)
54. D. Han, Classification of finite group-frames and super-frames. Can. Math. Bull. **50**(1), 85–96 (2007)
55. D. Han, D. Larson, Frames, bases and group representations. Mem. Am. Math. Soc. **147**(697) (2000)
56. G.H. Hardy, E.M. Wright, *An Introduction to the Theory of Numbers*, 4th edn. (Oxford University, Oxford, 1965)
57. C. Heil, Linear independence of finite Gabor systems, in *Harmonic Analysis and Applications: A Volume in Honor of John J. Benedetto* (Springer/Birkhäuser, New York, 2006)
58. C. Heil, J. Ramanathan, P. Topiwala, Linear independence of time-frequency translates. Proc. Am. Math. Soc. **124**(9), 2787–2795 (1996)
59. T. Helleseth, P. Vijay Kumar, Sequences with low correlation, in *Handbook of Coding Theory*, vol. I, II, ed. by V.S. Pless, W.C. Huffman (North-Holland, Amsterdam, 1998), pp. 1765–1853
60. M.A. Herman, T. Strohmer, High-resolution radar via compressed sensing. IEEE Trans. Signal Process. **57**(6), 2275–2284 (2009)
61. M.J. Hirn, The number of harmonic frames of prime order. Linear Algebra Appl. **432**(5), 1105–1125 (2010)

62. L.Hörmander, The Weyl calculus of pseudo differential operators. Commun. Pure Appl. Math. **32**, 360–444 (1979)
63. S. Jaffard, A density criterion for frames of complex exponentials. Mich. Math. J. **38**, 339–348 (1991)
64. P. Jaming, Phase retrieval techniques for radar ambiguity problems. J. Fourier Anal. Appl. **5**(4), 309–329 (1999) MR 1700086 (2000g:94007)
65. J.-P. Kahane, Sur certaines classes de séries de Fourier absolument convergentes. J. Math. Pures Appl. (9) **35**, 249–259 (1956)
66. J.-P. Kahane, *Séries de Fourier Absolument Convergentes*, (Springer, New York, 1970)
67. J.-P. Kahane, R. Salem, *Ensembles Parfaits et Séries Trigonométriques* (Paris, 1963)
68. T. Kailath, Sampling models for linear time-variant filters. Technical Report 352, Massachusetts Institute of Technology, Research Laboratory of Electronics (1959)
69. T. Kailath, Measurements on time-variant communication channels. IRE Trans. Inf. Theory **8**(5), 229–236 (1962)
70. T. Kailath, Time-variant communication channels. IEEE Trans. Inf. Theory **9**(4), 233–237 (1963)
71. Y. Katznelson, *An Introduction to Harmonic Analysis*, 3rd Original 1968 edn., Cambridge Mathematical Library (Cambridge University Press, Cambridge, 2004)
72. J.R. Klauder, The design of radar signals having both high range resolution and high velocity resolution. Bell Syst. Tech. J. **39**, 809–820 (1960)
73. J.R. Klauder, A.C. Price, S. Darlington, W.J. Albersheim, The theory and design of chirp radars. Bell Syst. Tech. J. **39**, 745–808 (1960)
74. J.F. Koksma, The theory of asymptotic distribution modulo one. Compos. Math. **16**, 1–22 (1964)
75. J. Kovačević, A. Chebira, Life beyond bases: the advent of frames (part I). IEEE Signal Process. Mag. **24**(4), 86–104 (2007)
76. J. Kovačević, A. Chebira, Life beyond bases: the advent of frames (part II). IEEE Signal Process. Mag. **24**, 115–125 (2007)
77. W. Kozek, G.E. Pfander, Identification of operators with bandlimited symbols. SIAM J. Math. Anal. **37**(3), 867–888 (2005)
78. G. Kutyniok, Linear independence of time-frequency shifts under a generalized Schrödinger representation. Arch. Math. **78**, 135–144 (2002)
79. H.J. Landau, Necessary density conditions for sampling and interpolation of certain entire functions. Acta Math. **117**, 37–52 (1967)
80. N. Levanon, E. Mozeson, *Radar Signals* (Wiley Interscience/IEEE Press, Hoboken, New Jersey, 2004)
81. L.Ê. Lindahl, F. Poulsen, *Thin Sets in Harmonic Analysis: Seminars Held at Institute Mittag-Leffler, 1969/70* (M. Dekker, New York, 1971)
82. P.A. Linnell, Von Neumann algebras and linear independence of translates. Proc. Am. Math. Soc. **127**, 3269–3277 (1999)
83. Y. Meyer, *Algebraic Numbers and Harmonic Analysis* (Elsevier, New York, 1972)
84. W.H. Mow, A new unified construction of perfect root-of-unity sequences, in *Proceedings of IEEE 4th International Symposium on Spread Spectrum Techniques and Applications* (Germany) (September 1996), pp. 955–959
85. J. Ortega-Cerdà, K. Seip, Fourier frames. Ann. Math. **155**(3), 789–806 (2002)
86. R.E.A.C. Paley, N. Wiener, *Fourier Transforms in the Complex Domain*. American Mathematical Society Colloquium Publications, vol. XIX (American Mathematical Society, Providence, RI, 1934)
87. G.E. Pfander, Gabor frames in finite dimensions, in *Finite Frames, Theory and Applications*, ed. by P.G. Casazza, G. Kutyniok (Springer/Birkhäuser, New York, 2013), pp. 193–239
88. G.E. Pfander, Sampling of operators. J. Fourier Anal. Appl. **19**(3), 612–650 (2013)
89. G.E. Pfander, D.F. Walnut, Measurement of time-variant linear channels. IEEE Trans. Inf. Theory **52**(11), 4808–4820 (2006)

90. G.E. Pfander, D.F. Walnut, *Operator identification and Feichtinger's algebra.* Sampling Theory Signal Image Process. **5**(2), 183–200 (2006)
91. B.M. Popovic, Generalized chirp-like polyphase sequences with optimum correlation properties. IEEE Trans. Inf. Theory **38**(4), 1406–1409 (1992)
92. R.A. Rankin, The closest packing of spherical caps in *n* dimensions, in *Proceedings of the Glasgow Mathematical Association*, vol. 2 (Cambridge University Press, Cambridge, 1955), pp. 139–144
93. M.A. Richards, J.A. Scheer, W.A. Holm (eds.), *Principles of Modern Radar* (SciTech Publishing Inc., Raleigh, 2010)
94. F. Riesz, B. Sz.-Nagy, *Functional Analysis* (Frederick Ungar Publishing Co., New York, 1955)
95. W. Rudin, *Fourier Analysis on Groups*. Interscience Tracts in Pure and Applied Mathematics (Interscience Publishers, New York, 1962)
96. B. Saffari, Some polynomial extremal problems which emerged in the twentieth century, in *Twentieth Century Harmonic Analysis—A Celebration*. NATO Science Series II Mathematics, Physics and Chemistry. vol. 33, (Kluwer Academic Publishers, Dordrecht, 2001) pp. 201–233.
97. R. Salem, On singular monotonic functions of the Cantor type. J. Math. Phys. **21**, 69–82 (1942)
98. R. Salem, On singular monotonic functions whose spectrum has a given Hausdorff dimension. Ark. Mat. **1**(4), 353–365 (1951)
99. K. Seip, On the connection between exponential bases and certain related sequences in $L^2(-\pi, \pi)$. J. Funct. Anal. **130**, 131–160 (1995)
100. M.I. Skolnik, *Introduction to Radar Systems* (McGraw-Hill Book Company, New York, 1980)
101. D. Slepian, Group codes for the gaussian channel. Bell Syst. Tech. J. **47**(4), 575–602 (1968)
102. E.M. Stein, *Harmonic Analysis* (Princeton University Press, Princeton, 1993)
103. E.M. Stein, G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces* (Princeton University Press, Princeton, 1971)
104. T. Strohmer, R.W. Heath Jr., Grassmannian frames with applications to coding and communications. Appl. Comput. Harmon. Anal. **14**, 257–275 (2003)
105. W. Sun, G-frames and g-Riesz bases. J. Math. Anal. Appl. **322**(1), 437–452 (2006)
106. D.A. Swick, A review of wideband ambiguity functions. Technical Report, DTIC Document (1969)
107. A. Terras, *Fourier Analysis on Finite Groups and Applications*, vol. 43 (Cambridge University Press, Cambridge , 1999)
108. R.J. Turyn, Sequences with small correlation, in *Error Correcting Codes*, (Wiley, New York, 1968), pp. 195–228
109. D.E. Vakman, *Sophisticated Signals and the Uncertainty Principle in Radar* (Springer, New York, 1969)
110. R. Vale, S. Waldron, Tight frames and their symmetries. Constr. Approx. **21**(1), 83–112 (2004)
111. R. Vale, S. Waldron, Tight frames generated by finite nonabelian groups. Numer. Algorithms **48**(1–3), 11–27 (2008)
112. A. Weil, On some exponential sums. Proc. Natl. Acad. Sci. USA **34**, 204–207 (1948)
113. A. Weil, *Sur les courbes algébriques et les variétés qui s'en déduisent*, Actualités Sci. et Ind. no. 1041 (Hermann, Paris, 1948)
114. L. Welch, Lower bounds on the maximum cross correlation of signals. IEEE Trans. Inf. Theory **20**(3), 397–399 (1974)

Chapter 6
The Fundamentals of Spectral Tetris Frame Constructions

Peter G. Casazza and Lindsey M. Woodland

Abstract In Casazza et al. (Appl. Comput. Harmon. Anal. **30**(2), 175–187, 2011), Casazza, Fickus, Mixon, Wang and Zhou introduced a fundamental method for constructing unit norm tight frames, which they called *Spectral Tetris*. This was a significant advancement for finite frame theory - especially constructions of finite frames. This paper then generated a vast amount of literature as Spectral Tetris was steadily developed, refined, and generalized until today we have a complete picture of what are the broad applications as well as the limitations of Spectral Tetris. In this paper, we will put this vast body of literature into a coherent theory.

2010 *Mathematics Subject Classification.* Primary 42C15

6.1 Introduction

Hilbert space frames were introduced by Duffin and Schaeffer in [16] while studying deep questions in non-harmonic Fourier series. Today they have broad application to problems in pure mathematics, applied mathematics, engineering, medicine and much more. Due to the redundancy, flexibility and stability of a frame, frame theory has proven to be a powerful area of research with applications to a wide array of fields, including signal processing, noise and erasure reduction, compressed sensing, sampling theory, data quantization, quantum measurements, coding, image processing, wireless communications, time-frequency analysis, speech recognition, bio-imaging, and much more. The reader is referred to [3] and references therein for further information regarding these topics. A fundamental problem for applications

The authors were supported by: NSF DMS 1008183; NSF ATD 1042701; AFOSR DGE51: FA9550-11-1-0245.

P.G. Casazza (✉) • L.M. Woodland
Department of Mathematics, University of Missouri, Columbia, MO 65211, USA
e-mail: casazzap@missouri.edu; lmwvh4@mail.missouri.edu

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_6

of frames is to construct frames with the necessary properties for the application. This can often be very difficult if not impossible in practice.

In [9], Casazza, Fickus, Mixon, Wang and Zhou introduced a construction technique for unit norm tight frames, which they called *Spectral Tetris*. Prior to this technique only ad-hoc methods were used to construct desired frames and in many cases the theory relied on *existence proofs*. Since [9], there has been a flurry of activity around Spectral Tetris as it was steadily developed, refined, and generalized until today we have necessary and sufficient conditions for the frames and fusion frames for which Spectral Tetris can construct. In this paper we will put this vast quantity of literature into a coherent theory so that researchers will be able to quickly tell if these methods will work for their problems.

The present paper is divided into two main parts: the first half introduces finite frame theory and then develops the progression of Spectral Tetris frame constructions while the second half introduces and analyzes the algorithms used to construct Spectral Tetris fusion frames.

6.2 Spectral Tetris Frame Constructions

Before Spectral Tetris is introduced, we will briefly discuss the basics of finite frame theory. After this, we then illustrate Spectral Tetris through an example as it is applied to unit norm tight frames. From here, the Spectral Tetris algorithm is further generalized in each concurrent subsection until finally in Subsection 6.2.7 Spectral Tetris frames are completely characterized. We have included the progression of Spectral Tetris as many of the specialized cases throughout are easier to implement than the general form and hence could be of particular interest to some researchers.

6.2.1 Hilbert Space Frames

We now introduce the basics of finite frame theory.

Definition 2.1. A family of vectors $\{f_n\}_{n=1}^N$ in an M-dimensional Hilbert space $\mathcal{H}_M$ is a frame if there are constants $0 < A \leq B < \infty$ so that for all $x \in \mathcal{H}_M$,

$$A\|x\|^2 \leq \sum_{n=1}^{N} |\langle x, f_n \rangle|^2 \leq B\|x\|^2,$$

where A and B are the lower and upper frame bounds, respectively.

(1) In the finite dimensional setting, a frame is simply a spanning set of vectors in the Hilbert space.
(2) The *optimal lower frame bound and optimal upper frame bound*, denoted A_{op} and B_{op}, are the largest lower frame bound and the smallest upper frame bound, respectively.

(3) If $A = B$ is possible, then $\{f_n\}_{n=1}^N$ is a *tight frame*. Moreover, if $A = B = 1$ is possible, then $\{f_n\}_{n=1}^N$ is a *Parseval frame*.
(4) If there is a constant c so that $\|f_n\| = c$ for all $n = 1, \ldots, N$ then $\{f_n\}_{n=1}^N$ is an *equal norm frame*. Moreover, if $c = 1$ then $\{f_n\}_{n=1}^N$ is a *unit norm frame*.
(5) $\{\langle x, f_n\rangle\}_{n=1}^N$ are called the *frame coefficients* of the vector $x \in \mathcal{H}_M$ with respect to frame $\{f_n\}_{n=1}^N$.
(6) We will refer to a unit norm, tight frame as a UNTF.

If $\{f_n\}_{n=1}^N$ is a frame for $\mathcal{H}_M$, then the *analysis operator* of the frame is the operator $T : \mathcal{H}_M \to \ell_2(N)$ given by

$$T(x) = \{\langle x, f_n\rangle\}_{n=1}^N$$

and the *synthesis operator* is the adjoint operator, T^*, which satisfies

$$T^*\left(\{a_n\}_{n=1}^N\right) = \sum_{n=1}^N a_n f_n.$$

The *frame operator* is the positive, self-adjoint, invertible operator $S = T^*T$ on $\mathcal{H}_M$ and satisfies

$$S(x) = T^*T(x) = \sum_{n=1}^N \langle x, f_n\rangle f_n.$$

That is, $\{f_n\}_{n=1}^N$ is a frame if and only if there are constants $0 < A \leq B < \infty$ such that its frame operator S satisfies $AI \leq S \leq BI$ where I is the identity on $\mathcal{H}_M$.

A frame has a certain spectrum or certain eigenvalues if its frame operator S has this spectrum or respectively these eigenvalues. Note that the spectrum of a frame operator S is positive and real. Also, the smallest and largest eigenvalues of a frame operator S coincide with the optimal lower and upper frame bounds. For any frame with spectrum $\{\lambda_m\}_{m=1}^M$, the sum of its eigenvalues counting multiplicities, equals the sum of the squares of the norms of its vectors:

$$\sum_{m=1}^M \lambda_m = \sum_{n=1}^N ||f_n||^2.$$

This quantity will be exactly the number of vectors N when we work with unit norm frames.

Definition 2.2. Let $N \geq M > 0$ and let the real values $\lambda_1, \ldots, \lambda_M \geq 2$ satisfy $\sum_{m=1}^M \lambda_m = N$ (unit norm). Then the class of unit norm frames $\{f_n\}_{n=1}^N$ in $\mathcal{H}_M$ whose frame operator has eigenvalues $\lambda_1, \ldots, \lambda_M$ will be denoted by $\mathcal{F}\left(N, \{\lambda_m\}_{m=1}^M\right)$.

Remark 2.3. Later in Corollary 2.10, we will see that the sets $\mathcal{F}\left(N, \{\lambda_m\}_{m=1}^M\right)$ are non-empty.

Proposition 2.4. *If $\{f_n\}_{n=1}^N$ is a UNTF for $\mathcal{H}_M$ then the frame bound will be $c = \frac{N}{M}$.*

To each frame we can associate the matrix of its synthesis operator, where the columns correspond to the frame vectors represented against an orthonormal basis for $\mathcal{H}_M$. Note, any rank M, $M \times N$ matrix with $N \geq M$ represents the synthesis matrix of some frame; however, this arbitrary matrix representation may not have many useful properties in applications. If instead the frame vectors are represented against the eigenbasis of its frame operator S then the frame operator can be represented via a diagonal matrix for which its eigenvalues are the diagonal entries.

Theorem 2.5 ([3]). *Let $T : \mathcal{H}_M \to \ell_2(N)$ be a linear operator, let $\{e_m\}_{m=1}^M$ be an orthonormal basis for $\mathcal{H}_M$, and let $\{\lambda_m\}_{m=1}^M$ be a sequence of positive numbers. Let B^* denote the $M \times N$ matrix representation of T^* with respect to $\{e_m\}_{m=1}^M$ and the standard basis $\{\hat{e}_n\}_{n=1}^N$ of $\ell_2(N)$. Then the following conditions are equivalent.*

(1) *$\{B^*\hat{e}_n\}_{n=1}^N$ forms a frame for $\mathcal{H}_M$ whose frame operator has eigenvectors $\{e_m\}_{m=1}^M$ and associated eigenvalues $\{\lambda_m\}_{m=1}^M$.*
(2) *The rows of B^* are orthogonal and the m-th row square sums to λ_m.*
(3) *The columns of B^* form a frame for $\ell_2(N)$ and $B^*B = \text{diag}(\lambda_1, \ldots, \lambda_M)$, where B^*B represents the frame operator and "diag" is the diagonal operator with diagonal values $\{\lambda_m\}_{m=1}^M$.*

Applying Theorem 2.5 to a UNTF yields the following characteristics: the frame operator is a scalar multiple of the identity, the rows of the synthesis matrix all square sum to the same constant and the columns must all square sum to one. Because of these characteristics, in the present paper we only consider frames represented with respect to the eigenbasis of the frame operator. One can relax this condition but little information is then available from the representation of the frame vectors. Also, sparsity is very sensitive to the basis with respect to which we represent the frame vectors.

Theorem 2.5 also justifies calling such a matrix a *frame matrix* or just a *frame* and hence we will use the term *frame* interchangeably to mean a frame or a frame matrix.

Definition 2.6. Given an $M \times N$ frame matrix $T^* = [f_1 \cdots f_N]$ representing an N-element frame in $\mathcal{H}_M$ against the eigenbasis of its frame operator, we have the following:

(1) The *support size of a row* is the number of nonzero entries in that row.
(2) The *support of a frame vector f_n*, denoted $\text{supp}\, f_n$, is the index set of the nonzero entries.

With the necessary definitions from finite frame theory needed for the present paper complete, we refer the interested reader to [3, 15] for a more in-depth study on the topic.

6.2.2 Before Spectral Tetris

Before introducing the Spectral Tetris construction technique, it is important to mention previous frame construction methods. In particular, one well known theorem used for frame construction illustrates how to construct a Parseval frame from the knowledge of an existing Parseval frame. Consider the following construction: for $N \geq M$, given an $N \times N$ unitary matrix, if any M rows are selected from this matrix then the column vectors from these rows form a Parseval frame for $\mathcal{H}_M$. Moreover, the leftover set of $N - M$ rows, also has the property that its N columns form a Parseval frame for $\mathcal{H}_{N-M}$. The next theorem, known as Naimark's Theorem, utilizes this type of operation and is one of the most fundamental results in frame theory.

Theorem 2.7 (Naimark's Theorem [3]). *Let $F = \{f_n\}_{n=1}^N$ be a frame for $\mathcal{H}_M$ with analysis operator T, let $\{e_n\}_{n=1}^N$ be the standard basis of $\ell_2(N)$, and let $P : \ell_2(N) \to \ell_2(N)$ be the orthogonal projection onto range (T). Then the following conditions are equivalent:*

(1) *$\{f_n\}_{n=1}^N$ is a Parseval frame for $\mathcal{H}_M$.*
(2) *For all $n = 1, \ldots, N$, we have $Pe_n = Tf_n$.*
(3) *There exist $\psi_1, \ldots, \psi_N \in \mathcal{H}_{N-M}$ such that $\{f_n \oplus \psi_n\}_{n=1}^N$ is an orthogonal basis of $\mathcal{H}_N$.*

Moreover, if (3) holds, then $\{\psi_n\}_{n=1}^N$ is a Parseval frame for $\mathcal{H}_{N-M}$. If $\{\psi_n'\}_{n=1}^N$ is another Parseval frame as in (3), then there exists a unique linear operator L on $\mathcal{H}_{N-M}$ such that $L\psi_n = \psi_n'$, for all $n = 1, \ldots, N$, and L is unitary.

Explicitly, we call $\{\psi_n\}_{n=1}^N$ the *Naimark Complement* of F.

Naimark's Theorem has its limitations as it requires the use of a previously known Parseval frame and only constructs Parseval frames. However, prior to Spectral Tetris, more general construction methods did not exist and instead the field relied on *existence theorems*, which fail to give precise details about desired frames. Existence results, such as the *Schur-Horn Theorem*, can be found in [4, 5, 7]. Preceding these papers, variations of the Schur-Horn theorem had appeared in the literature although they were in forms which were indistinguishable at the time. For a frame theoretic based proof of the Schur-Horn theorem, see [4, 5].

Definition 2.8. After arranging both sequences, $\{a_n\}_{n=1}^N$ and $\{\lambda_m\}_{m=1}^M$, in non-increasing order, if $\sum_{i=1}^n a_i^2 \leq \sum_{i=1}^n \lambda_i$ for every $n = 1, \ldots, M$ and $\sum_{i=1}^N a_i^2 = \sum_{i=1}^M \lambda_i$, then $\{\lambda_m\}_{m=1}^M$ *majorizes* $\{a_n^2\}_{n=1}^N$. We denote this by $\{\lambda_m\}_{m=1}^M \succeq \{a_n^2\}_{n=1}^N$. Moreover, if $M \neq N$, pad the shorter sequence with zeroes until the lengths are the same.

Theorem 2.9 (Schur-Horn Theorem [5]). *Let S be a positive, self-adjoint operator on $\mathcal{H}_M$, and let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_M > 0$ be the eigenvalues of S. Further, let $N \geq M$, and let $a_1 \geq a_2 \geq \cdots \geq a_N$ be positive real numbers. The following are equivalent:*

(1) There exists a frame $\{f_n\}_{n=1}^N$ *for* $\mathcal{H}_M$ *having frame operator* S *and satisfying* $\|f_n\| = a_n$ *for all* $n = 1, 2, \ldots, N$.
(2) $\{\lambda_m\}_{m=1}^M \succeq \{a_n^2\}_{n=1}^N$.

The Schur-Horn theorem provides a straightforward method for determining when frames exist. As a consequence, N-element equal norm frames exist in $\mathcal{H}_M$ for every $N \geq M$.

Corollary 2.10 ([7]). *For every* $N \geq M$ *and every invertible, positive, self-adjoint operator* S *on* $\mathcal{H}_M$ *there exists an equal norm frame for* $\mathcal{H}_M$ *with* N*-elements and frame operator* S*. In particular, there exists an equal norm Parseval frame with* N*-elements in* $\mathcal{H}_M$ *for every* $N \geq M$.

Both Theorem 2.9 and Corollary 2.10 guarantee existence of certain frames; however these theorems provide minimal insight into the construction of such frames.

6.2.3 *Spectral Tetris Frame Constructions: The Basics of Spectral Tetris*

Spectral Tetris was introduced in [9] as a method for constructing *sparse*, unit norm, tight frames and *sparse*, unit weighted, tight fusion frames via a quick and easy to use algorithm. We start with an example which illustrates the basics of Spectral Tetris for UNTFs. Note that we will call any frame constructed via Spectral Tetris, a *Spectral Tetris frame*.

Definition 2.11. The N-element Spectral Tetris frame in $\mathcal{H}_M$ with eigenvalues $\lambda_1, \ldots, \lambda_M \geq 2$ will be denoted by $STF(N; \lambda_1, \ldots, \lambda_M)$.

Before we begin the example, recall a few necessary facts. The N-element UNTF in $\mathcal{H}_M$, represented by an $M \times N$ matrix, must have the following properties:

(1) The columns square sum to one, to obtain unit norm vectors.
(2) The rows are orthogonal, which is equivalent to the frame operator, S, being a diagonal $M \times M$ matrix.
(3) The rows have constant norm, to obtain tightness, meaning that $S = cI$ for some constant c, where I is the $M \times M$ identity matrix.

One drawback of Spectral Tetris, in its original form, is that it can only construct frames with redundancy of at least 2, that is $N \geq 2M$, where N is the number of frame elements and M is the dimension of the Hilbert space. For a UNTF, the unique eigenvalue is $\frac{N}{M}$ and hence the restriction on the redundancy of the frame equates to the requirement that the unique eigenvalue is at least 2.

The main idea of Spectral Tetris is to iteratively construct a synthesis matrix, T^*, for a UNTF one to two vectors at a time, which satisfies properties (1) and (2) at each step and gets closer to and eventually satisfies property (3) when complete. When it

is necessary to build two vectors at a time throughout the Spectral Tetris process, we will utilize the following 2×2 matrix as a building block for our construction:

$$A(x) = \begin{bmatrix} \sqrt{\frac{x}{2}} & \sqrt{\frac{x}{2}} \\ \sqrt{1-\frac{x}{2}} & -\sqrt{1-\frac{x}{2}} \end{bmatrix},$$

where $0 \leq x \leq 2$.

Notice that $A(x)$ satisfies the following properties:

(1) the columns of $A(x)$ square sum to 1,
(2) $A(x)$ has orthogonal rows,
(3) the square sum of the first row is x.

These properties combined are equivalent to

$$A(x)A^*(x) = \begin{bmatrix} x & 0 \\ 0 & 2-x \end{bmatrix}.$$

We start with an example of how the Spectral Tetris algorithm works.

Example 2.12. Use Spectral Tetris to construct a *sparse*, unit norm, tight frame with 11 elements in $\mathcal{H}_4$, so the tight frame bound will be $\frac{11}{4}$. Note, by Corollary 2.10, such a frame exists.

To do this, create a 4×11 matrix T^*, which satisfies the following conditions:

(1) The columns square sum to 1.
(2) T^* has orthogonal rows.
(3) The rows square sum to $\frac{11}{4}$.
(4) $S = T^*T = \frac{11}{4}I$.

Note that (4) follows if (1), (2) and (3) are all satisfied.

Define $t_{i,j}$ to be the entry in the i^{th} row and j^{th} column of T^*. With an empty 4×11 matrix, start at $t_{1,1}$ and work left to right to fill out the matrix. By requirement (1), the square sum of column one needs to be 1 and by requirement (2) the square sum of row one needs to be $\frac{11}{4} \geq 1$. Hence, start by being greedy and put the maximum weight of 1 in $t_{1,1}$. This forces the rest of the entries in column 1 to be zero, from requirement (1). This yields:

$$T^* = \begin{bmatrix} 1 \cdots\cdots\cdots\cdots \\ 0 \cdots\cdots\cdots\cdots \\ 0 \cdots\cdots\cdots\cdots \\ 0 \cdots\cdots\cdots\cdots \end{bmatrix}.$$

Next, since row one needs to square sum to $\frac{11}{4}$, by (3), and so far row one only has a total weight of 1, we need to add $\frac{11}{4} - 1 = \frac{7}{4} = 1 + \frac{3}{4} \geq 1$ more weight to row one. Again be greedy and add another 1 in $t_{1,2}$. This forces the rest of the entries in

column 2 to be zero, by (1). Also note that row one now has a total square sum of 2. This yields:

$$T^* = \begin{bmatrix} 1\ 1 & \cdots & \cdots \\ 0\ 0 & \cdots & \cdots \\ 0\ 0 & \cdots & \cdots \\ 0\ 0 & \cdots & \cdots \end{bmatrix}.$$

In order to have a total square sum of $\frac{11}{4}$ in the first row, we need to add a total of $\frac{11}{4} - 2 = \frac{3}{4} < 1$ more weight. If the remaining unknown entries are chosen so that T^* has orthogonal rows, then S will be a diagonal matrix. Currently, the diagonal entries of S are mostly unknowns, having the form $\{2+?, \cdot, \cdot, \cdot\}$. Therefore we need to add $\frac{3}{4}$ more weight in the first row without compromising the orthogonality of the rows of T^* nor the normality of its columns. That is, if we get "greedy" and try to add $\sqrt{\frac{3}{4}}$ to position $t_{1,3}$ then the rest of row one must be zero, yielding:

$$T^* = \begin{bmatrix} 1\ 1 & \sqrt{\frac{3}{4}} & 0\ 0\ 0\ 0\ 0\ 0\ 0\ 0 \\ 0\ 0 & \cdot & \cdots\cdots\cdots \\ 0\ 0 & \cdot & \cdots\cdots\cdots \\ 0\ 0 & \cdot & \cdots\cdots\cdots \end{bmatrix}.$$

In order for column three to square sum to one, at least one of the entries $t_{2,3}, t_{3,3}$ or $t_{4,3}$ is non-zero. But then, it is impossible for the rows to be orthogonal and thus we cannot proceed. Hence, instead add two columns of information in attempts to satisfy these conditions. The key idea is to utilize our 2×2 building block, $A(x)$.

Define the third and fourth columns of T^* according to such a matrix $A(x)$, where $x = \frac{11}{4} - 2 = \frac{3}{4}$. Notice that by doing this, column three and column four each square sum to one within the first two rows, hence the rest of the unknown entries in these two columns will be zero. This yields:

$$T^* = \begin{bmatrix} 1\ 1 & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & \cdots\cdots\cdots \\ 0\ 0 & \sqrt{\frac{5}{8}} & -\sqrt{\frac{5}{8}} & \cdots\cdots\cdots \\ 0\ 0 & 0 & 0 & \cdots\cdots\cdots \\ 0\ 0 & 0 & 0 & \cdots\cdots\cdots \end{bmatrix}.$$

The diagonal entries of T^* are now $\{\frac{11}{4}, \frac{5}{4}+?, \cdot, \cdot\}$. The first row of T^*, and equivalently the first diagonal entry of S, now have sufficient weight and so its remaining entries are set to zero. The second row, however, is currently falling short by $\frac{11}{4} - \left(\left(\sqrt{\frac{5}{8}}\right)^2 + \left(-\sqrt{\frac{5}{8}}\right)^2\right) = \frac{6}{4} = 1 + \frac{2}{4}$. Since $1 + \frac{2}{4} \geq 1$, again be greedy and add a weight of 1 in $t_{2,5}$. Hence, column five becomes e_2. Next, with a weight

of $\frac{2}{4} < 1$ left to add to row two, utilize the 2×2 building block $A(x)$, with $x = \frac{2}{4}$. Adding this 2×2 block in columns six and seven yields sufficient weight in these columns and hence we finish these two columns with zeros. This yields:

$$T^* = \begin{bmatrix} 1 & 1 & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\frac{5}{8}} & -\sqrt{\frac{5}{8}} & 1 & \sqrt{\frac{2}{8}} & \sqrt{\frac{2}{8}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{6}{8}} & -\sqrt{\frac{6}{8}} & \cdot & \cdot & \cdot & \cdot \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \cdot & \cdot & \cdot & \cdot \end{bmatrix}.$$

The diagonal entries of T^* are now $\{\frac{11}{4}, \frac{11}{4}, \frac{6}{4}+?, \cdot\}$, where the third diagonal entry, and equivalently the third row, are falling short by $\frac{11}{4} - \frac{6}{4} = \frac{5}{4} = 1 + \frac{1}{4}$. Since $1 + \frac{1}{4} \geq 1$, then we take the eighth column of T^* to be e_3. Complete the matrix following these same strategies by letting the ninth and tenth columns arise from $A\left(\frac{1}{4}\right)$, and making the final column e_4, yielding the desired UNTF:

$$T^* = \begin{bmatrix} 1 & 1 & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\frac{5}{8}} & -\sqrt{\frac{5}{8}} & 1 & \sqrt{\frac{2}{8}} & \sqrt{\frac{2}{8}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{6}{8}} & -\sqrt{\frac{6}{8}} & 1 & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{7}{8}} & -\sqrt{\frac{7}{8}} & 1 \end{bmatrix}.$$

In this construction, column vectors are either introduced one at a time, such as columns $1, 2, 5, 8$, and 11, or in pairs, such as columns $\{3, 4\}, \{6, 7\}$, and $\{9, 10\}$. Each singleton contributes a value of 1 to a particular diagonal entry of T^*, 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 one or two. This construction is reminiscent of the game Tetris, as we fill in blocks of mixed area to obtain a flat spectrum.

6.2.4 Sparsity

Example 2.12 illustrates an important property of the frames that Spectral Tetris constructs, namely sparsity.

Definition 2.13.

(1) Given a fixed orthonormal basis of $\mathcal{H}_M$, a vector in $\mathcal{H}_M$ which can be represented by only $0 \leq k \leq M$ basis elements, is called *k-sparse*.

(2) Let $\{e_j\}_{j=1}^M$ be an orthonormal basis for $\mathcal{H}_M$. Then a frame $\{f_n\}_{n=1}^N$ for $\mathcal{H}_M$ is called *k-sparse* with respect to $\{e_j\}_{j=1}^M$, if for each $n \in \{1,\dots,N\}$ there exists $J_n \subseteq \{1,\dots,M\}$ such that $f_n \in \text{span}\{e_j : j \in J_n\}$ and $\sum_{n=1}^N |J_n| = k$.

In Example 2.12, column one of T^* is 1-sparse and the matrix T^* is 17-sparse.

The sparsity of T^* in Example 2.12 is not ad-hoc; it is shown in [10], that unit norm *tight* Spectral Tetris frames are optimally sparse in the sense that given $N \geq 2M$, the synthesis matrix of the N-element unit norm, tight Spectral Tetris frame (UNTSTF) for $\mathcal{H}_M$ is sparsest among all synthesis matrices of N-element unit norm, tight frames for $\mathcal{H}_M$.

Definition 2.14. Let F be a class of frames for $\mathcal{H}_M$, let $\{f_n\}_{n=1}^N \in F$, and let $\{e_j\}_{j=1}^M$ be an orthonormal basis for $\mathcal{H}_M$. Then $\{f_n\}_{n=1}^N$ is *optimally sparse* in F with respect to $\{e_j\}_{j=1}^M$ if $\{f_n\}_{n=1}^N$ is k_1-sparse with respect to $\{e_j\}_{j=1}^M$ and there does not exist a frame $\{\psi_n\}_{n=1}^N \in F$ which is k_2-sparse with respect to $\{e_j\}_{j=1}^M$ with $k_2 < k_1$.

To prove UNTSTFs are optimally sparse, the following definition and theorem are required.

Definition 2.15. A finite sequence of real values $\lambda_1,\dots,\lambda_M$ is *ordered blockwise*, if for any permutation π of $\{1,\dots,M\}$ the set of partial sums $\{\sum_{m=1}^s \lambda_m : s = 1,\dots M\}$ contains at least as many integers as the set $\{\sum_{m=1}^s \lambda_{\pi(m)} : s = 1,\dots,M\}$. The *maximal block number* of a finite sequence of real values $\lambda_1,\dots,\lambda_M$, denoted by $\mu(\lambda_1,\dots,\lambda_M)$, is the number of integers in $\{\sum_{m=1}^s \lambda_{\sigma(m)} : s = 1,\dots,M\}$, where σ is a permutation of $\{1,\dots,M\}$ such that $\lambda_{\sigma(1)},\dots,\lambda_{\sigma(M)}$ is ordered blockwise.

The following lemma provides a sparsity bound for any frame in $\mathcal{F}\left(N,\{\lambda_m\}_{m=1}^M\right)$ and is instrumental for proving that UNTSTFs are optimally sparse.

Lemma 2.16 ([10]). *Let $N \geq M > 0$ and let the real values $\lambda_1,\dots,\lambda_M \geq 2$ satisfy $\sum_{m=1}^M \lambda_m = N$. Then any frame in $\mathcal{F}\left(N,\{\lambda_m\}_{m=1}^M\right)$ has sparsity at least $N + 2(M - \mu(\lambda_1,\dots,\lambda_M))$ with respect to any orthonormal basis of $\mathcal{H}_M$.*

Theorem 2.17 ([10]). *Let $N \geq M > 0$, then the UNTSTF $\{f_n\}_{n=1}^N$ with real eigenvalues $\lambda_1,\dots,\lambda_M \geq 2$ ordered blockwise satisfying $\sum_{m=1}^M \lambda_m = N$ is optimally sparse in $\mathcal{F}\left(N,\{\lambda_m\}_{m=1}^M\right)$ with respect to the standard unit vector basis. That is, this frame is $N+2(M - \mu(\lambda_1,\dots,\lambda_M))$-sparse with respect to the standard unit vector basis.*

Proof ([10]). Let $\{f_n\}_{n=1}^N$ be a UNTSTF with eigenvalues $\lambda_1,\dots,\lambda_M \geq 2$. We will first show that its synthesis matrix has block decomposition of order $\mu := \mu(\lambda_1,\dots,\lambda_M)$. For this, let $k_0 = 0$, and let $k_1,\dots,k_\mu \in \mathbb{N}$ be chosen such that $m_i := \sum_{m=1}^{k_i} \lambda_m$ is an integer for every $i = 1,\dots,\mu$. Moreover, let $m_0 = 0$. Further, note that $k_\mu = M$ and $m_\mu = N$, since $\sum_{m=1}^M \lambda_m$ is an integer by hypothesis. The steps of Spectral Tetris for computing $\text{STF}(m_1;\lambda_1,\dots,\lambda_{k_1})$ and $\text{STF}(N;\lambda_1,\dots,\lambda_M)$ coincide until we reach the entry in the k_1^{th} row and m_1^{th} column when computing $\text{STF}(N;\lambda_1,\dots,\lambda_M)$. Therefore, the first k_1 entries of the first m_1 vectors of both constructions coincide. Continuing the computation

of STF$(N, \lambda_1, \ldots, \lambda_M)$ will set the remaining entries of the first m_1 vectors and also the first k_1 entries of the remaining vectors to zero. Thus, any of the first k_1 vectors have disjoint support from any of the vectors constructed later on. Repeating this argument for k_2 until k_μ, we obtain that the synthesis matrix has a block decomposition of order μ; the corresponding partition of the frame vectors being

$$\bigcup_{i=1}^{\mu} \{f_{m_{i-1}+1}, \ldots, f_{m_i}\}.$$

To compute the number of non-zero entries in the synthesis matrix generated by Spectral Tetris, we let $i \in \{1, \ldots, \mu\}$ be arbitrarily fixed and compute the number of non-zero entries of the vectors $f_{m_{i-1}+1}, \ldots, f_{m_i}$. Spectral Tetris ensures that each of the rows $k_{i-1} + 1$ up to $k_i - 1$ intersects the support of the subsequent row on a set of size 2, since in these rows Spectral tetris will always produce a 2×2 submatrix $A(x)$ for some $0 < x \leq 2$. Thus, there exist $2(k_i - k_{i-1} - 1)$ frame vectors with two non-zero entries. The remaining $(m_i - m_{i-1}) - 2(k_i - k_{i-1} - 1)$ frame vectors will have only one entry, yielding a total number of $(m_i - m_{i-1}) + 2(k_i - k_{i-1} - 1)$ non-zero entries in the vectors $f_{m_{i-1}+1}, \ldots, f_{m_i}$.

Summarizing, the total number of non-zero entries in the frame vectors of $\{f_n\}_{n=1}^N$ is

$$\sum_{i=1}^{\mu} (m_i - m_{i-1}) + 2(k_i - k_{i-1} - 1) =$$

$$\left(\sum_{i=1}^{\mu} (m_i - m_{i-1})\right) + 2\left(k_\mu - \left(\sum_{i=1}^{\mu} 1\right)\right) = N + 2(M - \mu),$$

which by Lemma 2.16 is the maximally achievable sparsity. □

Although Theorem 2.17 proves that unit norm tight Spectral Tetris frames are optimally sparse with respect to the standard unit vector basis, a slight modification to the Spectral Tetris algorithm will construct optimally sparse frames with respect to any desired orthogonal basis. This sparsity is, however, dependent on the ordering of the given sequence of eigenvalues for which the Spectral Tetris construction is performed, which is discussed in upcoming sections.

Remark 2.18. Optimally sparse UNTFs in $\mathcal{F}\left(N, \{\lambda_m\}_{m=1}^M\right)$ are not uniquely determined. The following is an example of two different UNTFs in $\mathcal{F}\left(9, \{\frac{9}{4}\}_{i=1}^9\right)$ which both achieve the optimal sparsity of 15:

$$\begin{bmatrix} 1 & 1 & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\frac{7}{8}} & -\sqrt{\frac{7}{8}} & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\frac{3}{4}} & -\sqrt{\frac{3}{4}} & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{5}{8}} & -\sqrt{\frac{5}{8}} & 1 \end{bmatrix}$$

and

$$\begin{bmatrix} 1 & \sqrt{\frac{5}{8}} & \sqrt{\frac{5}{8}} & 0 & 0 & 0 & 0 & 0\,0 \\ 0 & \sqrt{\frac{3}{8}} & -\sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & 0\,0 \\ 0 & 0 & 0 & \sqrt{\frac{5}{8}} & -\sqrt{\frac{5}{8}} & 0 & 0 & 1\,0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{5}{8}} & -\sqrt{\frac{5}{8}} & 0\,1 \end{bmatrix}.$$

Spectral Tetris not only provides optimally sparse UNTFs, it also yields orthogonality between numerous pairs of frame vectors due to their disjoint support. This can be seen in Example 2.12 where columns $t_{i,j}$ and $t_{i,j'}$ are orthogonal whenever $|j' - j| \geq 5$. More generally any UNTSTF, $\{f_n\}_{n=1}^N$, satisfies the orthogonality condition $\langle f_n, f_{n'} \rangle = 0$ whenever $|n' - n| \geq \lfloor \frac{N}{M} \rfloor + 3$. This is explicitly stated in the following theorem:

Theorem 2.19 ([9]). *For any $M, N \in \mathbb{N}$ such that $N \geq 2M$, there exists a UNTF, $\{f_n\}_{n=1}^N$, for $\mathcal{H}_M$ with the property that $\langle f_n, f_{n'} \rangle = 0$ whenever $|n' - n| \geq \lfloor \frac{N}{M} \rfloor + 3$.*

Sparse frames are instrumental in numerous applications as they reduce computational complexity and also ensure high compressibility of the synthesis matrix-which then is a sparse matrix. 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 utilizes this observation to show that such signals can be reconstructed from very few non-adaptive linear measurements by linear programming techniques. Finite frames thus play an essential role, both as sparsifying systems and in designing the measurement matrix. However, a drawback of Spectral Tetris is that it often generates multiple copies of the same frame vector. For practical purposes, this shall typically be avoided since the frame coefficients associated with a repeated frame vector do not provide any new information about the incoming signal.

6.2.5 *Spectral Tetris Constructions for Unit Norm Tight Frames with Redundancy Less than 2*

Spectral Tetris provides an easy to use construction method for UNTFs which are optimally sparse and possess orthogonality relations within the rows and columns. A potential drawback of Spectral Tetris is that in most cases it requires the frame to have at least twice as many vectors as the dimension. The following example illustrates the failure of Spectral Tetris when redundancy is less than 2.

Example 2.20. Use Spectral Tetris to construct a 5-element UNTF in $\mathcal{H}_3$. By Corollary 2.10 such a frame exists. The first step of Spectral Tetris forces column one to be e_1. Next, use the building block $A(x)$ for positions $t_{1,2}, t_{1,3}, t_{2,2}$ and $t_{2,3}$ to get:

$$T^* = \begin{bmatrix} 1 & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 & 0 \\ 0 & \sqrt{\frac{7}{8}} & -\sqrt{\frac{7}{8}} & 0 & 0 \\ 0 & 0 & 0 & \cdot & \cdot \\ 0 & 0 & 0 & \cdot & \cdot \end{bmatrix}.$$

Notice that row two square sums to $\frac{7}{4}$ exceeding the required eigenvalue of $\frac{5}{4}$. Thus Spectral Tetris cannot construct such a UNTF.

However in some scenarios Spectral Tetris can construct UNTFs with redundancy less than 2.

Theorem 2.21 ([14]). *For $M < N < 2M$ and $\lambda = \frac{N}{M}$ the following are equivalent:*

(1) *The Spectral Tetris construction will successfully produce a unit norm tight frame $\{f_n\}_{n=1}^N$ for $\mathcal{H}_M$.*
(2) *For all $1 \leq k \leq M-1$, if $k\lambda$ is not an integer, then we have $\lfloor k\lambda \rfloor \leq (k+1)\lambda - 2$, where $\lfloor x \rfloor$ is the greatest integer less than or equal to x.*

Theorem 2.21 completely characterizes when Spectral Tetris is able to construct UNTFs with redundancy less than 2 and because of this importance, the following example explicitly illustrates these conditions.

Example 2.22. In $\mathcal{H}_4$, construct a 6-element UNTF. The tight frame bound is $\lambda = \frac{6}{4} = \frac{3}{2} < 2$. Next, check if condition (2) holds: For all $1 \leq k \leq 3$,

- $1\left(\frac{3}{2}\right) = \frac{3}{2}$ is not at integer and $\left\lfloor \left(\frac{3}{2}\right) \right\rfloor = 1 \leq 1 = (1+1)\frac{3}{2} - 2$.
- $2\left(\frac{3}{2}\right) = 3$ is an integer.
- $3\left(\frac{3}{2}\right) = \frac{9}{2}$ is not at integer and $\left\lfloor \left(\frac{9}{2}\right) \right\rfloor = 4 \leq 4 = (3+1)\frac{3}{2} - 2$.

Thus condition (2) holds and therefore Spectral Tetris will construct this frame. Moreover, the frame constructed by Spectral Tetris is

$$\begin{bmatrix} 1 & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} & 0 & 0 & 0 \\ 0 & \sqrt{\frac{3}{4}} & -\sqrt{\frac{3}{4}} & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} \\ 0 & 0 & 0 & 0 & \sqrt{\frac{3}{4}} & -\sqrt{\frac{3}{4}} \end{bmatrix}.$$

The condition in Theorem 2.21 is completely determined by the value of the tight frame bound λ and as such an equivalent classification can be made.

Theorem 2.23 ([14]). *Spectral Tetris can be performed to generate a unit norm, tight frame with N vectors in $\mathcal{H}_M$ if and only if, when λ is in reduced form, one of the following occur:*

(1) $\lambda := \frac{N}{M} \geq 2$ *or*
(2) λ *is of the form* $\lambda = \frac{2L-1}{L}$ *for some positive integer* L.

Remark 2.24. The requirement that $\lambda = \frac{M}{N}$ is in reduced form is crucial to property (2) in Theorem 2.23. Also, if M and N are known to be relatively prime, then property (2) is equivalent to $M = 2N - 1$.

Example 2.25. Using Theorem 2.23, it is a straightforward check to see that Spectral Tetris can construct UNTFs with N-elements in $\mathcal{H}_4$ for all $N \geq 6$.

It is clear that there exist UNTFs with redundancy less than two for which the conditions of Theorem 2.23 are not satisfied. One method to construct such UNTFs is through the use of the Naimark Complement. First construct a UNTSTF satisfying $N \geq 2M$. It's Naimark Complement will be an N-element UNTF in $\mathcal{H}_{N-M}$ with redundancy less than 2. Hence Spectral Tetris ultimately constructs UNTFs for $N \geq M$.

Alternatively, a modified version of Spectral Tetris can be used directly to construct a UNTF with redundancy less than 2. In particular, a UNTF with redundancy greater than $\frac{j}{j-1}$ can be constructed using $J \times J$ discrete Fourier transform submatrices with scaled rows. However, through the use of these larger submatrices, we lose some sparsity within the frame, which inevitably reduces the orthogonality between the frame vectors.

Definition 2.26. Given $M \in \mathbb{N}$, let $\omega = exp\left(\frac{2\pi k}{M}\right)$ be a primitive M-th root of unity. The (non-normalized) discrete Fourier transform (DFT) matrix in $\mathcal{H}_{M\times M}$ is defined by $F_M = \left(\omega^{ij}\right)_{i,j=0}^{M-1}$.

Remark 2.27. DFT matrices possess the following properties:

(1) The rows are orthogonal.
(2) The columns are orthogonal.
(3) All entries have the same modulus.

Similar to the building block $A(x)$ used in Section 6.2.3, this adapted version of Spectral Tetris uses altered DFT submatrices for frame constructions where the rows of the DFT submatrices are multiplied by appropriate constants in order to get the correct row norm and unit norm columns. This scalar multiplication will not affect the pairwise orthogonality of the rows. It is important to note that the frames constructed using DFT submatrices will typically have complex entries.

Example 2.28. Construct a 5-element UNTF in $\mathcal{H}_4$. Recall, Example 2.20 showed that such a frame exists but the conventional Spectral Tetris method cannot construct this frame. Instead use altered DFT submatrices to construct such a UNTF. Define $\omega_M = exp\left(\frac{2\pi k}{M}\right)$.

Start by filling the desired 4×5 synthesis matrix with an altered 2×2 DFT matrix in the upper left corner. (Note a standard 2×2 matrix $A(x)$ could also be used here and in particular, when $w_2 = -1$ in the matrix below, this is exactly $A(\frac{5}{4})$.) To obtain the correct norms, multiply the entries of the first row by $\sqrt{\frac{5}{8}}$, thus making the first row have the desired norm $\sqrt{\frac{5}{4}}$.

In order to get unit norm columns, multiply the second row of the 2×2 DFT matrix by $\sqrt{\frac{3}{8}}$, yielding:

$$\begin{bmatrix} \sqrt{\frac{5}{8}} & \sqrt{\frac{5}{8}} & 0\,0\,0 \\ \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} \cdot \omega_2 & \cdots \\ 0 & 0 & \cdots \\ 0 & 0 & \cdots \end{bmatrix}.$$

The first two rows are orthogonal regardless of how the second row is completed. The second row, at this point, has norm $\sqrt{\frac{3}{4}}$, and thus it needs an additional weight of $\sqrt{\frac{2}{4}} = \sqrt{\frac{5}{4}} - \sqrt{\frac{3}{4}}$. We cannot insert a 1×1 block of $\sqrt{\frac{2}{4}}$ because the orthogonality of the rows would be lost when making this column unit norm. Also, if we attempt to insert an altered 2×2 DFT matrix in the same fashion we would have the following problem:

- To obtain the additional weight of $\sqrt{\frac{2}{4}}$ in row two, multiply the first row of a 2×2 DFT by the factor $\sqrt{\frac{2}{8}}$.
- Next, to obtain unit norm columns, multiply the second row of the DFT by the factor $\sqrt{\frac{6}{8}}$.
- Inserting this block into our synthesis matrix yields a norm of $\sqrt{\frac{12}{8}} = \sqrt{\frac{6}{4}} > \sqrt{\frac{5}{4}}$, the desired row norm.

To remedy this issue, we next attempt to utilize an altered 3×3 DFT. To obtain the correct altered 3×3 DFT proceed as follows:

- First to obtain the additional weight of $\sqrt{\frac{2}{4}}$ in row two, multiply the first row of a 3×3 DFT by $\sqrt{\frac{2}{12}} = \sqrt{\frac{1}{6}}$.
- Next, to obtain unit norm columns and row norms of $\sqrt{\frac{5}{4}}$ in the third and fourth row of the synthesis matrix, multiply the second and third row of the 3×3 DFT by $\sqrt{\frac{5}{12}}$.

This yields the desired 4×5 UNTF whose columns are normalized, rows are pairwise orthogonal and rows square sum to $\frac{5}{4}$.

$$\begin{bmatrix} \sqrt{\frac{5}{8}} & \sqrt{\frac{5}{8}} & 0 & 0 & 0 \\ \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} \cdot \omega_2 & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{6}} \\ 0 & 0 & \sqrt{\frac{5}{12}} & \sqrt{\frac{5}{12}} \cdot \omega_3 & \sqrt{\frac{5}{12}} \cdot \omega_3^2 \\ 0 & 0 & \sqrt{\frac{5}{12}} & \sqrt{\frac{5}{12}} \cdot \omega_3^2 & \sqrt{\frac{5}{12}} \cdot \omega_3^4 \end{bmatrix}.$$

Remark 2.29. As seen in Subsection 6.2.4, unit norm tight Spectral Tetris frames are optimally sparse when the original Spectral Tetris algorithm is implemented. However, when this altered version of Spectral Tetris with DFT submatrices is used to construct UNTFs with redundancy less than 2, then optimal sparsity is lost. The following example illustrates this.

Example 2.30. Constructing a 5-element UNTF in $\mathcal{H}_4$ using the DFT Spectral Tetris construction method yields one 2×2-block and one 3×3-block, resulting in 13 non-zero elements in the synthesis matrix (as seen in Example 2.28). However, the following matrix represents a sparser synthesis matrix for a 5-element UNTF in 4-dimensions:

$$\begin{bmatrix} \sqrt{\frac{5}{8}} & \sqrt{\frac{5}{8}} & 0 & 0 & 0 \\ \sqrt{\frac{3}{8}} & -\sqrt{\frac{3}{8}} & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{6}} & -\sqrt{\frac{1}{6}} \\ 0 & 0 & \sqrt{\frac{5}{8}} & -\sqrt{\frac{5}{8}} & 0 \\ 0 & 0 & \sqrt{\frac{5}{24}} & \sqrt{\frac{5}{24}} & 2\sqrt{\frac{5}{24}} \end{bmatrix}.$$

This matrix was constructed by starting with a 2×2 spectral tetris block and then adding the following 3×3 block for some $a, b, c \in \mathbb{C}$:

$$\begin{bmatrix} a & a & -a \\ b & -b & 0 \\ c & c & 2c \end{bmatrix}.$$

6.2.6 *Spectral Tetris for Non-Tight, Unit Norm Frames*

In general, sparse UNTFs represent a very small class of frames and hence a more general version of Spectral Tetris is necessary. In [1], the authors adapted Spectral Tetris to construct non-tight, unit norm frames with spectrum greater than or equal to two. The authors called this adaptation Sparse Unit Norm Frame Construction for Real Eigenvalues (SFR) and this was the first general construction method for non-tight frames. Note that the spectrum of a finite frame is necessarily positive and real. Also since the frames SFR will construct are not necessarily tight then the rows of the frame need not square sum to the same constant.

The SFR construction method also utilizes the 2×2 building block $A(x)$ and builds a unit norm frame with prescribed spectrum one or two vectors at a time. The sufficient conditions for when SFR can be implemented as well as the SFR algorithm follow.

Theorem 2.31 ([1]). *Suppose that real values* $\lambda_1 \geq \cdots \geq \lambda_M \geq 2$ *and* $N \in \mathbb{N}$ *satisfy:*

(1) $\sum_{j=1}^{M} \lambda_j = N$ *(i.e. unit norm frame vectors), and*
(2) *if* m_0 *is an integer in* $\{1, \ldots, M\}$*, for which* λ_{m_0} *is not an integer, then* $\lfloor \lambda_{m_0} \rfloor \leq N - 3$.

Then the eigenvalues of the frame operator of the frame $\{f_n\}_{n=1}^{N}$ *constructed by SFR are* $\{\lambda_m\}_{m=1}^{M}$ *and the frame vectors are at most* 2*-sparse.*

Remark 2.32. Note that the assumptions in Theorem 2.31 imply that only λ_1 could possibly be greater than $N - 3$ and therefore λ_{m_0} can be replaced by λ_1.

In [1], the authors provide an easily implementable algorithm, SFR, for constructing unit norm frames with prescribed spectrum.

SFR: Sparse Unit Norm Frame Construction for Real Eigenvalues

Parameters:

- Dimension $M \in \mathbb{N}$.
- Real eigenvalues $N \geq \lambda_1 \geq \cdots \geq \lambda_M \geq 2$, number of frame vectors N satisfying $\sum_{m=1}^{M} \lambda_j = N \in \mathbb{N}$.

Algorithm:

- Set $n = 1$
- For $m = 1, \ldots, M$ do

  (1) Repeat
    (a) If $\lambda_m < 1$ then
      (i) $f_n := \sqrt{\frac{\lambda_m}{2}} \cdot e_m + \sqrt{1 - \frac{\lambda_m}{2}} \cdot e_{m+1}$.
      (ii) $f_{n+1} := \sqrt{\frac{\lambda_m}{2}} \cdot e_m - \sqrt{1 - \frac{\lambda_m}{2}} \cdot e_{m+1}$.
      (iii) $n := n + 2$.
      (iv) $\lambda_{m+1} := \lambda_{m+1} - (2 - \lambda_m)$.
      (v) $\lambda_m := 0$.
    (b) else
      (i) $f_n := e_m$.
      (ii) $n := n + 1$.
      (iii) $\lambda_m := \lambda_m - 1$.
    (c) end
  (2) until $\lambda_m = 0$.

- end.

Output:

- Unit norm frame $\{f_n\}_{n=1}^{N}$ with eigenvalues $\{\lambda_m\}_{m=1}^{M}$

ALGORITHM 1: The SFR algorithm for constructing a unit norm frame with a desired spectrum.

Remark 2.33. For an explicit example of the SFR construction, see Remark 4.2, which is based on Example 4.1 in the Appendix of the present paper.

6.2.7 *Generalized Spectral Tetris Frame Constructions*

Spectral Tetris, in its original form, could only construct UNTFs and eventually it was modified to construct unit norm, non-tight frames with all eigenvalues greater than or equal to two. In [14], the authors adapted Spectral Tetris to construct highly sparse frames with specified eigenvalues and specified vector norms. In addition, the authors also proved necessary and sufficient conditions on the eigenvalues and vector norms of a frame for when this construction can be implemented and hence completely characterized the Spectral Tetris construction of a frame.

Similar to the Spectral Tetris construction method of Subsection 6.2.3, in this adaptation of Spectral Tetris called Prescribed Norms Spectral Tetris (PNSTC), a frame is built one or two vectors at a time and uses a 2×2 submatrix similar to $A(x)$. However, in order to allow for varied vector norms, modify property (1) of $A(x)$ so that the columns of $A(x)$ have varied norms; call these norms a_1 and a_2. Thus, the new 2×2 building block, denoted by $\hat{A}(x) := \hat{A}(x, a_1, a_2)$, is as follows:

$$\hat{A}(x) := \hat{A}(x, a_1, a_2) = \begin{bmatrix} \sqrt{\frac{x(a_1^2-y)}{x-y}} & \sqrt{\frac{x(x-a_1^2)}{x-y}} \\ \sqrt{\frac{y(x-a_1^2)}{x-y}} & -\sqrt{\frac{y(a_1^2-y)}{x-y}} \end{bmatrix},$$

where $y = a_1^2 + a_2^2 - x$.

Note, the existence of $\hat{A}(x)$ depends on x, a_1, and a_2.

Lemma 2.34 ([14]). *A real matrix $\hat{A}(x) := \hat{A}(x, a_1, a_2)$ satisfying*

$$\hat{A}(x)\hat{A}^*(x) = \begin{bmatrix} x & 0 \\ 0 & a_1^2 + a_2^2 - x \end{bmatrix}.$$

exists if and only if both of the following hold:

(1) $a_1^2 + a_2^2 \geq x > 0$, *and*
(2) *either* $a_1^2, a_2^2 \geq x$ *or* $a_1^2, a_2^2 \leq x$.

In order to satisfy the conditions in Lemma 2.34, a few restrictions on the eigenvalue sequence and the vector norm sequence are required. Note that the majorization condition of Theorem 2.9 is not sufficient in this scenario and hence there exist frames for which PNSTC cannot construct. In particular, a strengthening of majorization is required and is explicitly defined as follows:

Definition 2.35 ([14]). Two sequences $\{a_n\}_{n=1}^N$ and $\{\lambda_m\}_{m=1}^M$ are *Spectral Tetris ready* if $\sum_{n=1}^N a_n^2 = \sum_{m=1}^M \lambda_m$ and if there is a partition $0 \leq n_1 < \cdots < n_M = N$ of the set $\{0, 1, \ldots, N\}$ such that for all $k = 1, 2, \ldots, M-1$:

(1) $\sum_{n=1}^{n_k} a_n^2 \leq \sum_{m=1}^{k} \lambda_m < \sum_{n=1}^{n_k+1} a_n^2$ and
(2) if $\sum_{n=1}^{n_k} a_n^2 < \sum_{m=1}^{k} \lambda_m$, then $n_{k+1} - n_k \geq 2$ and $a_{n_k+2}^2 \geq \sum_{m=1}^{k} \lambda_m - \sum_{n=1}^{n_k} a_n^2$.

Since there is no assumption on the ordering of the sequence of eigenvalues nor the sequence of vector norms in Definition 2.35, it may be necessary to permute the sequences to make them Spectral Tetris ready. It is important to note that some permutations of the sequences may be Spectral Tetris ready while other permutations may not. This is illustrated in the following example:

Example 2.36. Given the eigenvalues $\{\lambda_m\}_{m=1}^3 = \{8, 6, 4\}$ and the vector norms $\{a_n\}_{n=1}^4 = \{3, 2, 2, 1\}$. Arranging the vectors norms as $\{a_n\}_{n=1}^4 = \{2, 1, 3, 2\}$ and taking the partition $n_1 = 1, n_2 = 3$ and $n_3 = 4$ yields Spectral Tetris ready sequences. However, arranging the vector norms as $\{a_n\}_{n=1}^4 = \{2, 2, 3, 1\}$ with eigenvalues $\{\lambda_m\}_{m=1}^3 = \{8, 6, 4\}$ yields no partition of the norms with Spectral Tetris ready sequences. Also, from this it is clear that the ordering of the sequences need not be monotone.

The properties given in Definition 2.35 completely characterize when PNSTC can be implemented.

Theorem 2.37 ([14]). *Given $\{a_n\}_{n=1}^N \subseteq (0, \infty)$ and $\{\lambda_m\}_{m=1}^M \subseteq (0, \infty)$, PNSTC can be used to construct a frame $\{f_n\}_{n=1}^N$ for $\mathcal{H}_M$ such that $\|f_n\| = a_n$ for $n = 1, \ldots, N$ and having eigenvalues $\{\lambda_m\}_{m=1}^M$ if and only if there exists a permutation which makes the sequences $\{a_n\}_{n=1}^N$ and $\{\lambda_m\}_{m=1}^M$ Spectral Tetris ready.*

Although PNSTC can construct a large class of sparse frames, there exist frames which fail to meet the conditions of Theorem 2.37.

Example 2.38. A 4-element tight frame in $\mathcal{H}_3$ with vector norms $\{a_n\}_{n=1}^4 = \{3, 3, 3, 1\}$ satisfies the majorization condition of Theorem 2.9 and hence such a frame exists; however, there is no arrangement of these eigenvalues and vector norms which is Spectral Tetris ready and thus PNSTC cannot construct such a frame.

The following algorithm, from [14], constructs a frame with prescribed norms and prescribed eigenvalues.

PNSTC: Prescribed Norms Spectral Tetris Construction

Parameters:

- Dimension $M \in \mathbb{N}$.
- Number of frame elements $N \in \mathbb{N}$.
- Eigenvalues $\{\lambda_m\}_{m=1}^M \subseteq (0,\infty)$ and norms of the frame vectors $\{a_n\}_{n=1}^N \subseteq (0,\infty)$ such that $\{\lambda_m\}_{m=1}^M$ and $\{a_n^2\}_{n=1}^N$ are Spectral Tetris ready.

Algorithm:

- Set $n = 1$
- For $m = 1, \ldots, M$ do

(1) Repeat
  (a) If $\lambda_m \geq a_n^2$ then
    (i) $f_n := a_n e_m$.
    (ii) $\lambda_m := \lambda_m - a_n^2$.
    (iii) $n := n + 1$.
  (b) else
    (i) If $2\lambda_m = a_n^2 + a_{n+1}^2$, then
      (A) $f_n := \sqrt{\frac{\lambda_m}{2}} \cdot (e_m + e_{m+1})$.
      (B) $f_{n+1} := \sqrt{\frac{\lambda_m}{2}} \cdot (e_m - e_{m+1})$.
    (ii) else
      (A) $y := a_n^2 + a_{n+1}^2 - \lambda_m$.
      (B) $f_n := \sqrt{\frac{\lambda_m\left(a_n^2 - y\right)}{\lambda_m - y}} \cdot e_m + \sqrt{\frac{y\left(\lambda_m - a_n^2\right)}{\lambda_m - y}} \cdot e_{m+1}$.
      (C) $f_{n+1} := \sqrt{\frac{\lambda_m\left(\lambda_m - a_n^2\right)}{\lambda_m - y}} \cdot e_m - \sqrt{\frac{y\left(a_n^2 - y\right)}{\lambda_m - y}} \cdot e_{m+1}$.
    (iii) end.
    (iv) $\lambda_{m+1} := \lambda_{m+1} - \left(a_n^2 + a_{n+1}^2 - \lambda_m\right)$.
    (v) $\lambda_m := 0$.
    (vi) $n := n + 2$.
  (c) end
(2) until $\lambda_m = 0$.

- end.

Output:

- Frame $\{f_n\}_{n=1}^N \subseteq \mathcal{H}_M$ with eigenvalues $\{\lambda_m\}_{m=1}^M$ and norms of the frame vectors $\{a_n\}_{n=1}^N$.

ALGORITHM 2: The PNSTC algorithm for constructing a frame with prescribed spectrum and prescribed vector norms.

Since PNSTC is the most general form of Spectral Tetris and hence the most useful, an example is included to illustrate the implementation of Algorithm 2. Within the following example, notice the requirement that the frame have redundancy at least 2 is no longer necessary.

Example 2.39. Construct an 8-element frame in $\mathcal{H}_5$ with:

$$\text{vector norms } \{a_n\}_{n=1}^8 = \{4, 1, 2, \sqrt{3}, 1, \sqrt{2}, 3, 2\}$$
$$\text{and eigenvalues } \{\lambda_m\}_{m=1}^5 = \{18, 6, 2, 10, 4\}.$$

First note, such a frame exists by Theorem 2.9. However, satisfying majorization does not guarantee that PNSTC can construct such a frame. Keeping the original arrangement of the sequences, $\{a_n\}_{n=1}^8 = \{4, 1, 2, \sqrt{3}, 1, \sqrt{2}, 3, 2\}$, $\{\lambda_m\}_{m=1}^5 = \{18, 6, 2, 10, 4\}$ and letting $n_1 = 2, n_2 = 4, n_3 = 5, n_4 = 7$, and $n_5 = 8$ yields Spectral Tetris ready sequences.

Next, implement PNSTC to construct the desired frame:

- Let $n = 1$.

(1) Let $m = 1$
  (a) Since $\lambda_1 = 18 \geq 16 = 4^2 = a_1^2$, then
    (i) $f_1 := a_1 e_1 = 4e_1$.
    (ii) $\lambda_1 := \lambda_1 - a_1^2 = 18 - 4^2 = 2$.
    (iii) $n := n + 1 = 1 + 1 = 2$.
    (iv) end.
  (b) Since $\lambda_1 \neq 0$, $(\lambda_1 = 2)$ then repeat with $\lambda_1 = 2$ and $n = 2$.
  (c) Since $\lambda_1 = 2 \geq 1^2 = a_2^2$, then
    (i) $f_2 := a_2 e_1 = 1e_1$.
    (ii) $\lambda_1 := \lambda_1 - a_1^2 = 2 - 1^2 = 1$.
    (iii) $n := n + 1 = 2 + 1 = 3$.
    (iv) end.
  (d) Since $\lambda_1 \neq 0$, $(\lambda_1 = 1)$ then repeat with $\lambda_1 = 1$ and $n = 3$.
  (e) Check: $\lambda_1 = 1 \not\geq 2^2 = a_3^2$
  (f) Check: $2\lambda_1 = 2 \neq 7 = 2^2 + \sqrt{3}^2 = a_3^2 + a_4^2$
  (g) Set $y := a_3^2 + a_4^2 - \lambda_1 = 2^2 + \sqrt{3}^2 - 1 = 6$, hence
    (i) $f_3 := \sqrt{\frac{1(2^2-6)}{1-6}} \cdot e_1 + \sqrt{\frac{6(1-2^2)}{1-6}} \cdot e_2 = \sqrt{\frac{2}{5}} \cdot e_1 + \sqrt{\frac{18}{5}} \cdot e_2$.
    (ii) $f_4 := \sqrt{\frac{1(1-2^2)}{1-6}} \cdot e_1 - \sqrt{\frac{6(2^2-6)}{1-6}} \cdot e_2 = \sqrt{\frac{3}{5}} \cdot e_1 - \sqrt{\frac{12}{5}} \cdot e_2$.
    (iii) end.
  (h) We also have the following:
    (i) $\lambda_2 := \lambda_2 - \left(a_3^2 + a_4^2 - \lambda_1\right) = 6 - (4 + 3 - 1) = 0$.
    (ii) $\lambda_1 := 0$.
    (iii) $n := n + 2 = 3 + 2 = 5$.
    (iv) end.
  (i) Now $\lambda_1 = 0$ and we end this loop.

- Next $m = 2$. (We still have $n = 5$).
- But $\lambda_2 = 0$ and we are done with this loop.
- Next $m = 3$. (We still have $n = 5$).

  (1) Since $\lambda_3 = 2 \geq 1 = 1^2 = a_5^2$, then
    (a) $f_5 := a_5 e_3 = 1e_3$.
    (b) $\lambda_3 := \lambda_3 - a_5^2 = 2 - 1^2 = 1$.
    (c) $n := n + 1 = 5 + 1 = 6$.
    (d) end.
  (2) Since $\lambda_3 \neq 0$, ($\lambda_3 = 1$) then repeat with $\lambda_3 = 1$ and $n = 6$.
  (3) Check: $\lambda_3 = 1 \not\geq 2 = a_6^2$
  (4) Check: $2\lambda_3 = 2 \neq 11 = \sqrt{2}^2 + 3^2 = a_6^2 + a_7^2$
  (5) Set $y := a_6^2 + a_7^2 - \lambda_3 = \sqrt{2}^2 + 3^2 - 1 = 10$, hence
    (a) $f_6 := \sqrt{\frac{1(2-10)}{1-10}} \cdot e_3 + \sqrt{\frac{10(1-2)}{1-10}} \cdot e_4 = \sqrt{\frac{8}{9}} \cdot e_3 + \sqrt{\frac{10}{9}} \cdot e_4$.
    (b) $f_7 := \sqrt{\frac{1(1-2)}{1-10}} \cdot e_3 - \sqrt{\frac{10(2-10)}{1-10}} \cdot e_4 = \sqrt{\frac{1}{9}} \cdot e_3 - \sqrt{\frac{80}{9}} \cdot e_4$.
    (c) end.
  (6) We also have the following:
    (a) $\lambda_4 := \lambda_4 - \left(a_6^2 + a_7^2 - \lambda_3\right) = 10 - (2 + 9 - 1) = 0$.
    (b) $\lambda_3 := 0$.
    (c) $n := n + 2 = 6 + 2 = 8$.
    (d) end.
  (7) Now $\lambda_3 = 0$ and we end this loop.

- Next $m = 4$. (We still have $n = 8$).
- But $\lambda_4 = 0$ and we are done with this loop.
- Next $m = 5$. (We still have $n = 8$).

  (1) Since $\lambda_5 = 4 \geq 4 = 2^2 = a_8^2$, then
    (a) $f_8 := a_8 e_5 = 2e_5$.
    (b) $\lambda_5 := \lambda_5 - a_8^2 = 4 - 4 = 0$.
    (c) $n := n + 1 = 8 + 1 = 9$.
    (d) end.
  (2) Now $\lambda_5 = 0$ and we end this loop.

- end.

Output: PNSTC created an 8-element frame $\{f_n\}_{n=1}^8$ in $\mathcal{H}_5$ with norms $\{a_n\}_{n=1}^8 = \{4, 1, 2, \sqrt{3}, 1, \sqrt{2}, 3, 2\}$ and eigenvalues $\{\lambda_m\}_{m=1}^5 = \{18, 6, 2, 10, 4\}$. This frame is represented in the following matrix:

$$\begin{bmatrix} 4 & 1 & \sqrt{\frac{2}{5}} & \sqrt{\frac{3}{5}} & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\frac{18}{5}} & -\sqrt{\frac{12}{5}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & \sqrt{\frac{8}{9}} & \sqrt{\frac{1}{9}} & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{10}{9}} & -\sqrt{\frac{80}{9}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 \end{bmatrix}.$$

Prior to implementing PNSTC, it is necessary to ensure that the sequences are Spectral Tetris ready; however, doing so can be a time-consuming and tedious task. To alleviate this task in specialized cases, the authors of [6] provide a simple systematic method for making sequences Spectral Tetris ready.

Proposition 2.40 ([6]). *Given a sequence of norms* $\{a_n\}_{n=1}^N$ *and a sequence of eigenvalues* $\{\lambda_m\}_{m=1}^M$ *where* $\sum_{n=1}^N a_n^2 = \sum_{m=1}^M \lambda_m$, *if*

$$\max_{i,j\in\{1,\dots,N\}} \left(a_i^2 + a_j^2\right) \leq \min_{m\in\{1,\dots,M\}} \lambda_m,$$

then the sequences can be made Spectral Tetris ready by systematically switching adjacent weights.

Proposition 2.40 allows PNSTC to construct a frame with sequences which are not initially Spectral Tetris ready and instead alters the sequences throughout the PNSTC process.

Example 2.41. Construct a Spectral Tetris frame on a sequence of vector norms and eigenvalues which are not Spectral Tetris ready. In $\mathcal{H}_2$, construct a 6-element frame with the sequence of norms $\{a_n\}_{n=1}^6 = \{\sqrt{3}, 2, \sqrt{3}, 1, 2, \sqrt{2}\}$ and eigenvalues $\{\lambda_m\}_{m=1}^2 = \{9, 8\}$.

First note that these sequences are not Spectral Tetris ready in the current order. Also, $\sum_{n=1}^6 a_n^2 = 17 = \sum_{m=1}^2 \lambda_m$ and

$$\max_{i,j\in\{1,\dots,6\}} \left(a_i^2 + a_j^2\right) = 8 \leq 8 \min_{m\in\{1,\dots,2\}} \lambda_m,$$

hence by Proposition 2.40 these sequences can be made Spectral Tetris ready by switching adjacent weights/norms. Therefore PNSTC can construct such a frame by possibly switching adjacent norms.

Starting the PNSTC construction of this frame yields $f_1 := \sqrt{3}\cdot e_1$ and $f_2 := 2\cdot e_1$.

Next, we need to add a weight of $9 - (\sqrt{3})^2 - (2)^2 = 2$ to row one. Since $\lambda = 2 \not\geq a_3^2 = 3$, in PNSTC we would typically add a 2×2 submatrix next. However, $a_4^2 = 1 < x = 2 < 3 = a_3^2$ and by Lemma 2.34 such a 2×2 submatrix does not exist. But, switching a_3 and a_4 yields the vector norm order $\{\sqrt{3}, 2, 1, \sqrt{3}, 2, \sqrt{2}\}$ and now $\lambda = 2 \geq 1 = a_3^2$. Hence, we assign $f_3 := 1 \cdot e_1$.

Now we need to add a weight of $\lambda = 1$ to row one. Since $\lambda = 1 \not\geq 3 = a_4^2$ then add a 2×2 submatrix to yield $f_4 := \sqrt{\frac{3}{5}}\cdot e_1 + \sqrt{\frac{12}{5}}\cdot e_2$ and $f_5 := \sqrt{\frac{2}{5}}\cdot e_1 - \sqrt{\frac{18}{5}}\cdot e_2$. Thus row one now has sufficient weight.

For row two, we need to add a weight of $8 - \left(\sqrt{\frac{12}{5}}\right)^2 - \left(\sqrt{\frac{18}{5}}\right)^2 = 2$ and hence let $f_6 := 2 \cdot e_2$. This yields the desired frame:

$$\begin{bmatrix} \sqrt{3} & 2 & 1 & \sqrt{\frac{3}{5}} & \sqrt{\frac{2}{5}} & 0 \\ 0 & 0 & 0 & \sqrt{\frac{12}{5}} & -\sqrt{\frac{18}{5}} & 2 \end{bmatrix}.$$

Notice this frame has orthogonal rows with norms $\{a_n\}_{n=1}^6 = \{\sqrt{3}, 2, 1, \sqrt{3}, 2, \sqrt{2}\}$ and eigenvalues $\{\lambda_m\}_{m=1}^2 = \{9, 8\}$.

As seen in Example 2.41, Proposition 2.40 is a modification of PNSTC which allows the algorithm to handle non-Spectral Tetris ready orderings. To use this re-ordering technique, simply insert the following Algorithm 3 between lines ((1)(b)(ii)) and ((1)(b)(ii)(A)) of the PNSTC algorithm, at Algorithm 2. This re-ordering procedure will be defined as Spectral Tetris Re-Ordering (STR).

STR: Spectral Tetris Re-Ordering Procedure

Parameters:

- Dimension $M \in \mathbb{N}$.
- Number of frame elements $N \in \mathbb{N}$.
- Eigenvalues $(\lambda_m)_{m=1}^M$ and vector norms $\{a_n\}_{n=1}^N$ such that $\sum_{n=1}^N a_n^2 = \sum_{m=1}^M \lambda_m$ and $\max_{i,j\in\{1,\dots,N\}} \left(a_i^2 + a_j^2\right) \leq \min_{m\in\{1,\dots,M\}} \lambda_m$.

Algorithm:

(1) If $\lambda_m > a_{n+1}^2$, then

  (a) temp $:= a_n$.
  (b) $a_{n+1} := a_n$.
  (c) a_{n+1} :=temp.
  (d) Go to PNSTC (1ai).

(2) end.

ALGORITHM 3: Procedure for running PNSTC on a non-spectral-tetris-ready ordering.

When the conditions of Proposition 2.40 are satisfied then STR always results in a Spectral Tetris ready ordering of the vector norms and eigenvalues. However, when these conditions are not satisfied, the authors of [14] provide alternative sufficient conditions on the prescribed norms and eigenvalues under which PNSTC can be implemented.

Theorem 2.42 ([14]). *Let* $\{a_n\}_{n=1}^N \subseteq (0,\infty)$ *and* $\{\lambda_m\}_{m=1}^M \subseteq (0,\infty)$ *be non-decreasing sequences such that* $\sum_{n=1}^N a_n^2 = \sum_{m=1}^M \lambda_m$ *and*

$$a_{N-2L}^2 + a_{N-2L-1}^2 \leq \lambda_{M-L}$$

for $L = 0, 1, \dots, M-1$. *Then* $\{a_n\}_{n=1}^N$ *and* $\{\lambda_m\}_{m=1}^M$ *are Spectral Tetris ready, hence by Theorem 2.37, PNSTC can construct a frame* $\{f_n\}_{n=1}^N$ *for* $\mathcal{H}_M$ *with* $\|f_n\| = a_n$ *for* $n = 1, \dots, N$ *and with eigenvalues* $\{\lambda_m\}_{m=1}^M$. *In particular, PNSTC can be performed if* $a_N^2 + a_{N-1}^2 \leq \lambda_1$.

Remark 2.43. In Theorem 2.42, the property $a_{N-2L}^2 + a_{N-2L-1}^2 \leq \lambda_{M-L}$ together with $\sum_{n=1}^N a_n^2 = \sum_{m=1}^M \lambda_m$ imply that $N \geq 2M$. Thus, this sufficient condition also requires redundancy of at least 2; whereas, the Spectral Tetris ready condition only required $N \geq M$.

The PNSTC algorithm can be specialized to construct tight and/or unit norm frames and in doing so the conditions in Theorem 2.42 become easier to check.

Theorem 2.44 ([14]). *Let $a_1 \geq a_2 \geq \cdots \geq a_N > 0$ and $\lambda = \frac{1}{M}\sum_{n=1}^{N} a_n^2$. If $a_1^2 + a_2^2 \leq \lambda$, then PNSTC constructs a λ-tight frame $\{f_n\}_{n=1}^{N}$ for $\mathcal{H}_M$ satisfying $\|f_n\| = a_n$ for all $n = 1, 2, \ldots, N$.*

Remark 2.45. The condition in Theorem 2.44 is an analog to the requirement that the frame have redundancy at least 2 in the original Spectral Tetris construction.

The condition in Theorem 2.44 is only a sufficient condition; hence, as the following example illustrates, there exist tight frames which fail this condition but satisfy the Spectral Tetris ready condition.

Example 2.46. Use PNSTC to construct a 6-element tight frame in $\mathcal{H}_3$ with vector norms $(\sqrt{6}, \sqrt{5}, \sqrt{5}, 1, 1, 1)$. The tight frame bound will be $\lambda = \frac{19}{3}$. Checking the conditions of Theorem 2.44 yields $a_1^2 + a_2^2 = 6 + 5 = 11 \not\leq \frac{19}{3}$ and hence Theorem 2.44 does not apply. However, arranging the norms $(\sqrt{6}, 1, \sqrt{5}, 1, 1, \sqrt{5})$ and taking the partition $n_1 = 1, n_2 = 3$ and $n_3 = 6$ yields sequences which are Spectral Tetris ready and hence PNSTC can construct such a frame.

A reformulation of Definition 2.35 and Theorem 2.37 to the case of tight frames with prescribed spectrum provides necessary and sufficient conditions for a sequence of norms to yield a tight frame via PNSTC.

Corollary 2.47 ([14]). *A tight frame for $\mathcal{H}_M$ with prescribed norms $\{a_n\}_{n=1}^{N}$ and eigenvalue $\lambda = \frac{1}{M}\sum_{n=1}^{N} a_n^2$ can be constructed via PNSTC if and only if there exists an ordering of $\{a_n^2\}_{n=1}^{N}$ for which there is a partition $0 \leq n_1 < \cdots < n_M = N$ of $\{0, 1, \ldots, N\}$ such that for all $k = 1, 2, \ldots, M-1$:*

(1) $\sum_{n=1}^{n_k} a_n^2 \leq k\lambda < \sum_{n=1}^{n_k+1} a_n^2$ for all $k = 1, \ldots, M-1$, and
(2) if $\sum_{n=1}^{n_k} a_n^2 < k\lambda$, then $n_{k+1} - n_k \geq$ and $a_{n_k+2}^2 \geq k\lambda - \sum_{n=1}^{n_k} a_n^2$.

Another specialized case of PNSTC is that of unit norm frames. Recall, Section 6.2.6 required the sufficient condition that the eigenvalues of a frame be greater than or equal to 2 for SFR to construct a unit norm frame; however, in [14], the authors found necessary and sufficient conditions for SFR to construct unit norm frames and relaxed this bound by reformulating Definition 2.35 and Theorem 2.37.

Corollary 2.48 ([14]). *Let $\sum_{m=1}^{M} \lambda_m = N$ where $N \in \mathbb{N}$ and $N \geq M$. Then SFR can be used to produce a unit norm frame for $\mathcal{H}_M$ with eigenvalues $\{\lambda_m\}_{m=1}^{M} \subseteq (0, \infty)$ if and only if there is some permutation of $\{\lambda_m\}_{m=1}^{M}$ and a partition $0 \leq n_1 < \cdots < n_M = N$ of $\{0, \ldots, N\}$ such that for each $k = 1, \ldots, M-1$,*

(1) $n_k \leq \sum_{m=1}^{k} \lambda_m < n_k + 1$ and
(2) if $n_k < \sum_{m=1}^{k} \lambda_m$, then $n_{k+1} - n_k \geq 2$.

The characterization in Corollary 2.48 provides a limitation on the location of the eigenvalues that can be strictly less than one, as the following Corollary proves.

Corollary 2.49 ([14]). *If SFR can be used to produce a unit norm frame for $\mathcal{H}_M$ with eigenvalues $(\lambda_m)_{m=1}^M$, then $\lambda_k < 1$ is only possible if $k = 1$ or if $n_{k-1} = \sum_{m=1}^{k-1} \lambda_m$.*

The PNSTC algorithm can also be applied to construct equal-norm frames.

Theorem 2.50 ([14]). *Let $\{\lambda_m\}_{m=1}^M \subseteq (0, \infty)$ be non-increasing. Then PNSTC can construct an equal-norm frame for $\mathcal{H}_M$ with eigenvalues $\{\lambda_m\}_{m=1}^M$.*

Through the development and adaptations of Spectral Tetris, we now have a complete characterization of the frames for which Spectral Tetris can construct. In continuing this study, we wish to further this construction technique to fusion frames. In the proceeding section, we will see that due to the sparsity of Spectral Tetris frames and the orthogonality of the frame vectors, we can generalize SFR and PNSTC to construct sparse fusion frames.

6.3 Spectral Tetris Fusion Frame Constructions

Now that Spectral Tetris frames have been completely characterized, the present section is dedicated to characterizing Spectral Tetris fusion frames. To do this, we first introduce and discuss applications of fusion frames. Next, we develop the original Spectral Tetris fusion frame construction technique as it is based on Spectral Tetris frame constructions and is restricted to unit-weighted equidimensional fusion frames. Making our way through the progression of Spectral Tetris and generalizing the algorithm within each subsection, we eventually completely characterize Spectral Tetris fusion frames in Section 6.3.7. We include algorithms, examples and proofs throughout to further illustrate the process.

6.3.1 Fusion Frames

Today, across numerous disciplines, scientists utilize vast amounts of data obtained from various networks which need to be analyzed at a central processor. However, due to low communication bandwidth and limited transit/computing power at each single node in the network, the data may not be able to be computed at one centralized processing system. Hence there has been a fundamental shift from centralized information processing to distributed processing, where network management is distributed and the reliability of individual links is less critical. Here the data processing is performed in two stages: (1) local processing at neighboring nodes, followed by (2) the integration of locally processed data streams at a central processor.

An example of distributed processing involves wireless sensor networks, which can provide cost-effective and reliable surveillance. Consider a large number of inexpensive, small sensors dispersed throughout an area in order to collect data or keep surveillance. Due to practical and economic factors such as the topography of

the land, limited signal processing power, low communication bandwidth, or short battery life, the sensors are not capable of transmitting their information to one central processor. Therefore, the sensors need to be deployed in smaller clusters, where in each cluster there is one higher powered sensor which collects all of the information from the signals in its cluster and then transmits this information to a central processor. In this two-stage model, information is first gathered locally in each cluster, then processed more globally at a central station. A similar two-stage (local-global) processing principle is also applicable in distributed sensing, parallel processing, packet encoding, optimal packings and in modeling the human visual cortex [3, 19].

Mathematically, this can be interpreted as follows: given data and a collection of subspaces, first project the data onto the subspaces then process the data within each subspace (this coincides with the local clusters of sensors gathering information locally). Next, combine or *fuse* all of the locally processed information (this coincides with the larger powered sensors in each cluster transmitting their information to a central processor).

This concept of a frame-like collection of subspaces is known as a *fusion frame* and provides a suitable mathematical framework to design and analyze two-stage processing (local-global). Fusion frames were first studied in [2] and further analyzed in [1, 3, 8].

Fusion frame theory is a generalization of frame theory. To illustrate this connection, for a given frame $\{f_n\}_{n=1}^N$ its frame operator can be viewed in the following manner:

$$Sx = T^*Tx = \sum_{n=1}^{N} \langle x, f_n \rangle f_n = \sum_{n=1}^{N} ||f_n||^2 \langle x, \frac{f_n}{||f_n||} \rangle \frac{f_n}{||f_n||}.$$

Notice that S is the sum of rank one projections each with weight given by the square norm of the respective frame vector. Generalizing this idea to consider weighted projections of arbitrary rank yields the definition of a fusion frame.

Definition 3.1. Let $\{W_i\}_{i=1}^D$ be a family of subspaces in $\mathcal{H}_M$, and let $\{w_i\}_{i=1}^D \subseteq \mathbb{R}^+$ be a family of weights. Then $\{(W_i, w_i)\}_{i=1}^D$ is a *fusion frame* for $\mathcal{H}_M$ if there exist constants $0 < A \leq B < \infty$ such that

$$A\|x\|_2^2 \leq \sum_{i=1}^{D} w_i^2 \|P_i(x)\|_2^2 \leq B\|x\|_2^2 \quad \text{for all } x \in \mathcal{H}_M,$$

where P_i denotes the orthogonal projection of $\mathcal{H}_M$ onto W_i for each $i \in \{1, \ldots, D\}$.

(1) In finite dimensions, a fusion frame is a spanning set of subspaces.
(2) The constants A and B are called the *lower fusion frame bound and upper fusion frame bound*, respectively.
(3) The largest lower fusion frame bound and the smallest upper fusion frame bound are called the *optimal lower fusion frame bound and optimal upper fusion frame bound*, respectively.

(4) If $A = B$ is possible then the family $\{(W_i, w_i)\}_{i=1}^D$ is called a *tight fusion frame*. Moreover, if $A = B = 1$ is possible then the family $\{(W_i, w_i)\}_{i=1}^D$ is called a *Parseval fusion frame*.
(5) If each subspace has unit weight, $w_i = 1$ for all $i = 1, \ldots, D$, then the family $\{(W_i, w_i)\}_{i=1}^D$ is denoted $(W_i)_{i=1}^D$ and is called a *unit weighted fusion frame*.
(6) The *fusion frame operator* $\tilde{S} : \mathcal{H}_M \to \mathcal{H}_M$ defined by $\tilde{S}x = \sum_{i=1}^D w_i^2 P_i(x)$ for all $x \in \mathcal{H}_M$ is a positive, self-adjoint, invertible operator, where P_i is the orthogonal projection of $\mathcal{H}_M$ onto W_i.
(7) The $\{W_i\}_{i=1}^D$ are called the *fusion frame subspaces*.

In two-stage processing, a signal can be reconstructed via a fusion frame; however, due to sensor failures, buffer over flows, added noise or subspace perturbations during processing, some information about the signal could be lost or corrupted. One might ask, how can a fusion frame reconstruct a signal when these problems are present? Clearly, redundancy between the subspaces helps to add resilience against erasures (or lost data); but what about other issues that could arise when a signal is being processed. Redundancy between these subspaces may not be sufficient to manage these issues and typically extra structure on the fusion frame is required, such as prescribing the subspace dimensions or prescribing the fusion frame operator. In particular, [17, 18] show that in order to minimize the mean-squared error in the linear minimum mean-squared error estimation of a random vector from its fusion frame measurements in white noise, the fusion frame needs to be Parseval or tight. Also to provide maximal robustness against erasures of one fusion frame subspace the fusion frame subspaces must also be equidimensional.

Within two stage processing, further issues could potentially arise due to economic factors which limit the available computing power and bandwidth for data processing. Because of this we need to be able to construct a fusion frame that enables signal decomposition with a minimal number of additions and multiplications and hence reduces computational costs. These numerous potential constraints on our data processing capabilities now motivates the need for fusion frames which not only have a desired fusion frame operator or subspace dimensions, but also possess some degree of *sparsity*. In particular, some of these issues could be alleviated if each subspace was spanned by a collection of sparse vectors with respect to a fixed orthonormal basis for $\mathcal{H}_M$.

Definition 3.2. Let $\{e_j\}_{j=1}^M$ be an orthonormal basis for $\mathcal{H}_M$. Then a fusion frame $\{(W_i, v_i)\}_{i=1}^D$ for $\mathcal{H}_M$ with $\dim W_i = d_i$ for all $i = 1, \ldots, D$ is called *k-sparse* with respect to $\{e_j\}_{j=1}^M$, if for each $i \in \{1, \ldots, N\}$ there exists an orthonormal basis $\{f_{i,l}\}_{l=1}^{d_i}$ for W_i and for each $l = 1, \ldots, d_i$ and $\{J_{i,l}\} \subset \{1, \ldots, M\}$ such that $f_{i,l} \in \text{span}\{e_j : j \in J_{i,l}\}$ and $\sum_{i=1}^M \sum_{l=1}^{d_i} |J_{i,l}| = k$. We refer to $\{f_{i,l}\}_{i=1,l=1}^{D,m_i}$ as an associated k-sparse frame.

Since fusion frames generalize the structure of a frame, it is natural to question if sparse representations in fusion frames possess similar properties as sparse representations in frames. In particular, do sparse fusion frames allow for precise signal reconstruction when using only an under determined set of equations? The

answer to this question is yes, which leads to a further question: how can such a sparse fusion frame be constructed? This motivates the need for the Spectral Tetris fusion frame construction.

To construct fusion frames, we will construct their associated frame matrix much like we did with conventional frames.

Theorem 3.3 ([2]). *The following are equivalent:*

(1) $\{(W_i, w_i)\}_{i=1}^{D}$ *is a fusion frame for* $\mathcal{H}_M$ *with lower and upper fusion frame bounds A and B, respectively.*
(2) *There exists an orthonormal basis* $\{e_{ij}\}_{j=1}^{d_i}$ *for* W_i*, for all* $i = 1, \ldots, D$*, so that the matrix C with column vectors* e_{ij} *for* $i \in \{1, \ldots, D\}$ *and* $j \in \{1, \ldots, d_i\}$ *satisfies:*

   (a) *The rows are orthogonal and*
   (b) *the square sums of the rows lie between A and B.*

Similar to our discussion of conventional frames, the square sum of the rows of the fusion frame matrix, as described in Theorem 3.3, yield the eigenvalues of the fusion frame operator where the smallest and largest eigenvalues of the fusion frame operator correspond to the optimal smallest and largest fusion frame bounds, A and B. Hence, if all of the rows of such a matrix square sum to the same value, then this is a tight fusion frame. In the present paper, when we discuss the eigenvalues of a fusion frame, we specifically mean the eigenvalues of its fusion frame operator.

6.3.2 Prior to the Spectral Tetris Fusion Frame Construction Method

Before Spectral Tetris was adapted to construct sparse fusion frames other construction methods existed; however these methods first required the knowledge of a given fusion frame. In particular, two general ways to construct a fusion frame from a given fusion frame are the *Spatial Complement Method* and the *Naimark Complement Method*.

Given a fusion frame, taking its spatial complement is a natural way of generating a new fusion frame. This requires the use of the *orthogonal fusion frame to a given fusion frame*.

Definition 3.4 ([1]). Let $\{(W_i, w_i)\}_{i=1}^{D}$ be a fusion frame for $\mathcal{H}_M$. If the family $\{(W_i^{\perp}, w_i)\}_{i=1}^{D}$, where $W_i^{\perp}$ is the orthogonal complement of W_i, is also a fusion frame, then we call $\{(W_i^{\perp}, w_i)\}_{i=1}^{D}$ the *orthogonal fusion frame* to $\{(W_i, w_i)\}_{i=1}^{D}$.

Theorem 3.5 (Spatial Complement Theorem [1]). *Let* $\{(W_i, w_i)\}_{i=1}^{D}$ *be a fusion frame for* $\mathcal{H}_M$ *with optimal fusion frame bounds* $0 < A \leq B < \infty$ *such that* $\sum_{i=1}^{D} w_i^2 < \infty$*. Then the following conditions are equivalent:*

(1) $\bigcap_{i=1}^{D} W_i = \{0\}$.
(2) $B < \sum_{i=1}^{D} w_i^2$.

(3) *The family* $\{(W_i^{\perp}, w_i)\}_{i=1}^{D}$ *is a fusion frame for* $\mathcal{H}_M$ *with optimal fusion frame bounds* $\sum_{i=1}^{D} w_i^2 - B$ *and* $\sum_{i=1}^{D} w_i^2 - A$.

Another fusion frame construction method, which also requires a given fusion frame, is the *Naimark Complement Method* and can be seen as an extension of Theorem 2.7. Consider the following relationship between frames and fusion frames. Let $\{(W_i, w_i)\}_{i=1}^{D}$ be a fusion frame for $\mathcal{H}_M$ with frame operator $\tilde{S}$. Let $\left(\psi_{i,j}\right)_{j=1}^{d_i}$ be an orthonormal basis for W_i for $i = 1, \ldots, D$ and let S be the frame operator for the frame $\{w_i\psi_{i,j}\}_{i=1,j=1}^{D,d_i}$. We have the following equivalence:

$$\tilde{S}x = \sum_{i=1}^{D} w_i^2 \left(P_i\left(x\right)\right) = \sum_{i=1}^{D}\sum_{j=1}^{d_i} w_i^2 \langle x, \psi_{i,j}\rangle \psi_{i,j} =$$

$$\sum_{i=1}^{D}\sum_{j=1}^{d_i} \langle x, w_i\psi_{i,j}\rangle w_i\psi_{i,j} = T^*Tx = Sx.$$

Thus every fusion frame arises from a conventional frame partitioned into equal-norm, orthogonal sets. This relationship validates defining the Naimark complement of a fusion frame via the Naimark complement of a conventional frame.

Definition 3.6. Let $\{(W_i, w_i)\}_{i=1}^{D}$ be a Parseval fusion frame for $\mathcal{H}_M$. Choose orthonormal bases $\left(\psi_{i,j}\right)_{j=1}^{d_i}$ for W_i, making $\{w_i\psi_{i,j}\}_{i=1,j=1}^{D,d_i}$ a Parseval frame for $\mathcal{H}_M$. By Theorem 2.7, $\{w_i\psi_{i,j}\}_{i=1,j=1}^{D,d_i}$ has a Naimark complement Parseval frame $\{\psi'_{i,j}\}_{i=1,j=1}^{D,d_i}$ for $\mathcal{H}_{D-M}$. The *Naimark Complement fusion frame* of $\{(W_i, w_i)\}_{i=1}^{D}$ is given by

$$\left\{\left(W_i', \sqrt{1 - w_i^2}\right)\right\}_{i=1}^{D},$$

which is a Parseval fusion frame for $\mathcal{H}_{\sum_{i=1}^{D} d_i - D}$, where $W_i' := \text{span}\left(\{\psi'_{i,j}\}_{j=1}^{d_i}\right)$.

Notice that the choice of the orthonormal bases for the subspaces W_i of a fusion frame will alter the corresponding Naimark complement fusion frame. However, it is shown in [13] that all choices yield unitarily equivalent Naimark complement fusion frames in the sense that there is a unitary operator mapping the corresponding fusion frame subspaces onto one another. The following theorem provides properties for when the Naimark Complement of a fusion frame exists.

Theorem 3.7 (Naimark Complement Method [1]). *Let* $\{(W_i, w_i)\}_{i=1}^{D}$ *be a Parseval fusion frame for* $\mathcal{H}_M$ *with* $0 < w_i < 1$, *for all* $i = 1, \ldots, D$. *Then there exists a Hilbert space* $\mathcal{H}_M \subseteq \mathcal{K}$ *and a Parseval fusion frame* $\{\left(W_i', \sqrt{1 - w_i^2}\right)\}_{i=1}^{D}$ *for* $\mathcal{K} \ominus \mathcal{H}_M$ *with* $\dim W_i' = \dim W_i$ *for all* $i = 1, \ldots, D$.

6.3.3 Spectral Tetris Fusion Frames Constructions

In general, the Spectral Tetris construction of a unit-weighted fusion frame first constructs a Spectral Tetris frame and then groups the vectors of this frame into orthonormal sets which span the subspaces of the desired fusion frame. This is explicitly seen in Algorithm 4 and Algorithm 6 in the following sections.

Definition 3.8. A frame constructed via the Spectral Tetris construction (SFR or PNSTC) is called a *Spectral Tetris frame*. A unit weighted fusion frame $(W_i)_{i=1}^D$ is called a *Spectral Tetris fusion frame* if there is a partition of a Spectral Tetris frame $\{f_{i,j}\}_{i=1,j=1}^{D,d_i}$ such that $\{f_{i,j}\}_{j=1}^{d_i}$ is an orthonormal basis for W_i for all $i = 1, \ldots, D$.

In Subsection 6.3.4 and Subsection 6.3.6 unit-weighted fusion frames are constructed and hence the restriction in Definition 3.8 to such fusion frames. However, in Subsection 6.3.7 non-unit weighted fusion frames are constructed and this requires a more general definition, which is developed in that section.

6.3.4 Spectral Tetris for Equidimensional, Unit-Weighted Fusion Frame Constructions

The authors of [1] were the first to adapt SFR to construct equidimensional, unit-weighted fusion frames for any given fusion frame operator with eigenvalues greater than or equal to two. Explicitly, in [1] they develop and analyze the following scenario:

Let $\lambda_1 \geq \cdots \geq \lambda_M \geq 2$ be real values and $M \in \mathbb{N}$ satisfy the factorization

$$\sum_{m=1}^{M} \lambda_m = kD \in \mathbb{N}.$$

The goal is to construct a sparse fusion frame $(W_i)_{i=1}^D$, $W_i \subseteq \mathcal{H}_M$, such that:

(G1) $\dim W_i = k$ for all $i = 1, \ldots, D$ and
(G2) the associated fusion frame operator has $\{\lambda_m\}_{m=1}^M$ as its eigenvalues.

To construct such a fusion frame, in [1] the authors generalize the SFR algorithm and develop a new algorithm called Sparse Fusion Frame Construction for Real Eigenvalues (SFFR). The SFFR algorithm follows the same construction formula as the SFR algorithm; however, in the output stage of SFFR, the vectors f_n are grouped in such a way so that the vectors assigned to each subspace form an orthonormal system.

Theorem 3.9 ([1]). *Suppose the real values $D \geq \lambda_1 \geq \cdots \geq \lambda_M \geq 2, D \in \mathbb{N}$, and $k \in \mathbb{N}$ satisfy:*

(1) $\sum_{m=1}^{M} \lambda_m = kD \in \mathbb{N}$,
(2) *If m_0 is the first integer in $\{1, \ldots, M\}$ for which λ_{m_0} is not an integer, then $\lfloor \lambda_{m_0} \rfloor \leq D - 3$.*

Then the fusion frame $\{W_i\}_{i=1}^D$ constructed by SFFR fulfills conditions (G1) and (G2) and the corresponding frame vectors are at most 2-sparse.

Remark 3.10. Note that the assumptions in Theorem 3.9 imply that only λ_1 could possibly be greater than $D-3$ and therefore λ_{m_0} can be replaced by λ_1.

The SFFR algorithm from [1] follows and constructs equidimensional, unit-weighted fusion frames.

SFFR: Sparse Fusion Frame Construction for Real Eigenvalues

Parameters:

- Dimension $M \in \mathbb{N}$.
- Real eigenvalues $D \geq \lambda_1 \geq \cdots \geq \lambda_M \geq 2$, number of subspaces D, and dimension of subspaces k satisfying $\sum_{m=1}^M \lambda_m = kD \in \mathbb{N}$.

Algorithm:

- Set $j := 1$
- For $m = 1, \ldots, M$ do

  (1) Repeat
  - (a) If $\lambda_m < 1$ then
    - (i) $f_j := \sqrt{\frac{\lambda_m}{2}} \cdot e_m + \sqrt{1 - \frac{\lambda_m}{2}} \cdot e_{m+1}$.
    - (ii) $f_{j+1} := \sqrt{\frac{\lambda_m}{2}} \cdot e_m - \sqrt{1 - \frac{\lambda_m}{2}} \cdot e_{m+1}$.
    - (iii) $j := j + 2$.
    - (iv) $\lambda_{m+1} := \lambda_{m+1} - (2 - \lambda_m)$.
    - (v) $\lambda_m := 0$.
  - (b) else
    - (i) $f_j := e_m$.
    - (ii) $j := j + 1$.
    - (iii) $\lambda_m := \lambda_m - 1$.
  - (c) end

  (2) until $\lambda_m = 0$.

- end.

Output:

- Equidimensional, unit weighted, fusion frame $\{W_i\}_{i=1}^D$ where $W_i := \text{span}\{f_{i+jD} : j = 0, \ldots, k-1\}$.

ALGORITHM 4: The SFFR algorithm for constructing an equidimensional, unit weighted, fusion frame with a desired frame operator.

Remark 3.11. For an explicit example of the SFFR construction method, see Example 4.1 in the Appendix of the present paper.

The SFFR algorithm constructs non-tight fusion frames; however due to the reconstruction properties of tight frames it is sometimes useful to extend such a fusion frame with additional subspaces until it becomes tight. The following theorem provides sufficient conditions for when and what types of subsets can be added to a fusion frame in order to obtain a tight fusion frame.

Theorem 3.12 ([1]). *Let $\{W_i\}_{i=1}^D$ be a fusion frame for $\mathcal{H}_M$ with $\dim W_i = k < M$ for all $i = 1, \dots, D$, and let $\tilde{S}$ be the associated fusion frame operator with eigenvalues $D \geq \lambda_1 \geq \cdots \geq \lambda_M \geq 2$ and eigenvectors $\{e_m\}_{m=1}^M$. Further, let A be the smallest positive integer, which satisfies the following conditions:*

(1) $\lambda_1 + 2 \leq A$.
(2) $AM = kN_0$ *for some* $N_0 \in \mathbb{N}$.
(3) $A \leq \lambda_M + N_0 - (D + 3)$.

Then there exists a fusion frame $\{V_i\}_{i=1}^{N_0-D}$ for $\mathcal{H}_M$ with $\dim V_i = k$ for all $i \in \{1, \dots, N_0 - D\}$ so that $\{W_i\}_{i=1}^D \cup \{V_i\}_{i=1}^{N_0-D}$ is an A-tight fusion frame.

The number of k-dimensional subspaces added in Theorem 3.12 to extend a fusion frame to a tight fusion frame is, in general, the smallest number that can be added.

6.3.5 Sparsity

The fusion frames constructed by SFFR are optimally sparse. We present analogous sparsity results to that of Subsection 6.2.4.

Definition 3.13. Let $M, D > 0$ and let the real values $\lambda_1, \dots, \lambda_M \geq 2$ satisfy $\sum_{j=1}^M \lambda_j = D$. Then the class of fusion frames $\{W_i\}_{i=1}^D$ in $\mathcal{H}_M$ with $\dim W_i = k$ for all $i = 1, \dots, D$ whose fusion frame operator has eigenvalues $\lambda_1, \dots, \lambda_M$ will be denoted by $FF(D, k, \{\lambda_i\}_{i=1}^M)$.

The notion of maximal block number as defined in Definition 2.15 can be extended to a decomposition property of the synthesis matrix of a fusion frame.

Definition 3.14. Let $M, D > 0$, and let $\{W_i\}_{i=1}^D$ be a fusion frame for $\mathcal{H}_M$ with associated frame $\{f_{i,l}\}_{l=1}^{d_i}$. Then we say that the synthesis matrix of $\{W_i\}_{i=1}^D$ associated with $\{f_{i,l}\}_{l=1}^{d_i}$ has *block decomposition of order M*, if there exists a partition $\{1, \dots, D\} = I_1 \cup \cdots \cup I_M$ such that, for any $k_1 \in I_{i_1}$ and $k_2 \in I_{i_2}$ with $i_1 \neq i_2$, we have $\operatorname{supp} \varphi_{k_1} \cap \operatorname{supp} \varphi_{k_2} = \emptyset$ and M is maximal.

The following result now connects the maximal block number of the sequence of eigenvalues of a fusion frame operator with the block decomposition order of an associated fusion frame.

Proposition 3.15 ([11]). *Let $M, k, D > 0$ and let the real values $\lambda_1, \ldots, \lambda_M \geq 2$ satisfy $\sum_{j=1}^{M} \lambda_j = kD$. Then the synthesis matrix of any fusion frame in the class $FF(D, k, \{\lambda_i\}_{i=1}^{M})$ with any associated frame has block decomposition of order at most $\mu(\lambda_1, \ldots, \lambda_M)$.*

Having introduced the required new notions, we are now in the position to state the exact value for the maximally achievable sparsity for a class $FF(D, k, \{\lambda_i\}_{i=1}^{M})$.

Theorem 3.16 ([11]). *Let $M, k, D > 0$, and let $\lambda_1, \ldots, \lambda_M \geq 2$ satisfy $\sum_{j=1}^{M} \lambda_j = kD$. Then any fusion frame in $FF(D, k, \{\lambda_i\}_{i=1}^{M})$ is at least $kD + 2(M - \mu(\lambda_1, \ldots, \lambda_M))$-sparse with respect to any orthonormal basis.*

Using the sparsity bound in Theorem 3.16, the following theorem proves that fusion frames constructed by SFFR do in fact achieve optimal sparsity.

Theorem 3.17 ([11]). *Let $M, k, D > 0$, and let $\lambda_1, \ldots, \lambda_M \geq 2$ be ordered blockwise and satisfy $\sum_{j=1}^{M} \lambda_j = kD$. Then the fusion frame $SFFR(D, k; \lambda_1, \ldots, \lambda_M)$ is optimally sparse in the class $FF(D, k, \{\lambda_i\}_{i=1}^{M})$ with respect to the standard unit vector basis.*

In particular, this fusion frame is $kD + 2(M - \mu(\lambda_1, \ldots, \lambda_M))$-sparse with respect to the standard unit vector basis, and the vectors generated by SFFR are an associated $kD + 2(M - \mu(\lambda_1, \ldots, \lambda_M))$-sparse frame.

Similar to conventional frames, if sparsity with respect to an orthonormal basis other that the standard unit basis is required, then the SFFR algorithm can easily be modified to accommodate this need by using this new basis instead.

6.3.6 Spectral Tetris for Unit Weighted Fusion Frame Constructions

In [1], the authors adapted SFR to construct optimally sparse, equidimensional, unit weighted fusion frames with all eigenvalues greater than or equal to two. However, it may be necessary to construct a fusion frame with fewer restrictions. In [12] the authors generalized Spectral Tetris through the use of PNSTC and a *Reference Fusion Frame* to construct unit-weighted fusion frames where the subspaces are not necessarily equidimensional and the eigenvalues need only to be positive. The authors also provide sufficient conditions for when this is possible, and provide necessary and sufficient conditions in the case of tight fusion frames with eigenvalues greater than or equal to two.

To implement this method and construct unit weighted fusion frames, first use PNSTC to construct a frame. Then use this Spectral Tetris frame to obtain a *reference fusion frame*. This reference fusion frame is not the desired fusion frame, it is however a major step in the construction of the fusion frame. Given a Spectral Tetris frame, $\{f_n\}_{n=1}^{N}$, the reference fusion frame is a first naive construction of a fusion frame from the Spectral Tetris frame. The procedure is to pick f_1 and then

choose the first vector after that which is orthogonal to f_1. Next, pick the first vector orthogonal to both of these vectors. Continue in this way until no more vectors can be chosen. This is the first subspace, V_1, of the reference fusion frame. For the second subspace, V_2, pick the first vector not in V_1 then repeat the above procedure and continue this process. The general procedure for constructing Spectral Tetris fusion frames is to then alter the dimensions of the subspaces of the reference fusion frame, one at a time, until the required dimensions, $\{d_i\}_{i=1}^D$, are met. To do this, for the first $i \in \{1, \ldots, D\}$ such that d_i does not equal the dimension of V_i, Algorithm 6 will give a procedure for increasing/decreasing the dimension of V_i until it equals d_i.

Definition 3.18. Let $N \geq M$ be positive integers, and let $\{\lambda_m\}_{m=1}^M \subseteq (0, \infty)$ have the property that $\sum_{m=1}^M \lambda_m = N$. The fusion frame constructed by RFF presented below in Algorithm 5 is called the *reference fusion frame* for the eigenvalues $(\lambda_m)_{m=1}^M$.

RFF: Reference Fusion Frame Spectral Tetris Construction

Parameters:

- Dimension $M \in \mathbb{N}$.
- Number of frame elements $N \in \mathbb{N}$.
- Eigenvalues $\{\lambda_m\}_{m=1}^M \subseteq (0, \infty)$ such that $\sum_{m=1}^M \lambda_m = N$ (unit norm).

Algorithm:

(1) Use PNSTC for $\{\lambda_m\}_{m=1}^M$ with unit norm vectors to get a Spectral Tetris frame $F = \{f_n\}_{n=1}^N$.
(2) $t :=$ maximal support size of the rows of F.
(3) $S_i := \emptyset$ for $i = 1, \ldots, t$.
(4) $k = 0$.
(5) Repeat.

(a) $k := k + 1$.
(b) $j :=$ $\min\{1 \leq r \leq t :$ $\text{supp} f_k \cap \text{supp} f_s = \emptyset$ for all $f_s \in S_r\}$.
(c) $S_j := S_j \cup \{f_k\}$.

(6) until $k = N$.

Output:

- Reference fusion frame $(V_i)_{i=1}^t$, where $V_i = \text{span}\,(S_i)$ for $i = 1, \ldots, t$.

ALGORITHM 5: The RFF algorithm for constructing the reference fusion frame.

The following example highlights a few important observations. First, in order to construct a reference fusion frame, the conventional frame needs to be unit norm but not necessarily tight. Secondly, different orderings of the eigenvalues of a frame will in general lead to different sequences of dimensions of the reference fusion frame and hence will alter the steps in the final fusion frame algorithm.

Example 3.19. Construct a 10-element unit norm frame in $\mathcal{H}_3$ with eigenvalues $\{\lambda_m\}_{m=1}^3 = \{\frac{13}{3}, \frac{10}{3}, \frac{7}{3}\}$ using PNSTC/SFR and then construct its reference fusion frame. It is a straightforward check of PNSTC or SFR to see that the corresponding 10-element frame is as follows:

$$[f_1 \cdots f_8] = \begin{bmatrix} 1 & 1 & 1 & 1 & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{6}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & \sqrt{\frac{5}{6}} & -\sqrt{\frac{5}{6}} & 1 & \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{2}{3}} & -\sqrt{\frac{2}{3}} & 1 \end{bmatrix}.$$

Thus the reference fusion frame constructed by RFF is as follows:

$$V_1 = \text{span}\{f_1, f_7, f_{10}\}, V_2 = \text{span}\{f_2, f_8\},$$
$$V_3 = \text{span}\{f_3, f_9\}, V_4 = \text{span}\{f_4\}, V_5 = \text{span}\{f_5\}, V_6 = \text{span}\{f_6\}.$$

However, if we reorder the same eigenvalues in the following way: $\{\lambda_m\}_{m=1}^3 = \{\frac{7}{3}, \frac{13}{3}, \frac{10}{3}\}$, then PNSTC yields the following frame:

$$[g_1 \cdots g_{10}] = \begin{bmatrix} 1 & 1 & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{6}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\frac{5}{6}} & -\sqrt{\frac{5}{6}} & 1 & 1 & \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{2}{3}} & -\sqrt{\frac{2}{3}} & 1 & 1 \end{bmatrix}.$$

Thus the reference fusion frame which RFF constructs for this frame is:

$$V_1 = \text{span}\{g_1, g_5, g_9\}, V_2 = \text{span}\{g_2, g_6, g_{10}\},$$
$$V_3 = \text{span}\{g_3\}, V_4 = \text{span}\{g_4\}, V_5 = \text{span}\{g_7\}, V_6 = \text{span}\{g_8\}.$$

The following Theorem 3.21 provides sufficient conditions for when a Spectral Tetris fusion frame can be constructed via a Reference fusion frame; but first the definition of a *chain* is necessary.

Definition 3.20. Let S be a set of vectors in $\mathcal{H}_M$, and $s \in S$. A subset $C \subseteq S$ is a *chain in S starting at s*, if $s \in S$ and the support of any element in S intersects the support of some other element of S. C is a *maximal chain in S starting at s* if C is not a proper subset of any other chain in S starting at s.

Theorem 3.21 ([12]). *Let $N \geq M$ be positive integers, $(\lambda_m)_{m=1}^M \subseteq (0, \infty)$ and let $(d_i)_{i=1}^D \subseteq \mathbb{N}$ be a non-increasing sequence of dimensions such that $\sum_{m=1}^M \lambda_m = \sum_{i=1}^D d_i = N$. Let $(V_i)_{i=1}^t$ be the reference fusion frame for $(\lambda_m)_{m=1}^M$. If we have the majorization $(\dim V_i)_{i=1}^t \succeq (d_i)_{i=1}^D$, then there exists a Spectral Tetris fusion frame $(W_i)_{i=1}^D$ for $\mathcal{H}_M$ with $\dim W_i = d_i$ for $i = 1, \ldots, D$ and eigenvalues $(\lambda_m)_{m=1}^M$.*

The proof of Theorem 3.21 from [12] is included as it is very constructive in nature and helps the reader to determine how the fusion frame subspaces are developed.

Proof. We show how to iteratively construct the desired fusion frame $(W_i)_{i=1}^D$. Let t and $V_1, \ldots, V_t$ be given by RFF for $(\lambda_m)_{m=1}^M$. Let $W_i^0 = V_i$ for $i = 1, \ldots, t$. We add empty sets if necessary to obtain a collection $\left(W_i^0\right)_{i=1}^D$ of D sets. If $\sum_{i=1}^D ||W_i^0| - d_i| = 0$ then $\left(W_i^0\right)_{i=1}^D$ is the desired fusion frame. Otherwise, starting from $\left(W_i^0\right)_{i=1}^D$, we will construct the spanning sets of the desired fusion frame.

Let

$$m = \max\{j \leq D : d_j \neq |W_j^0|\}.$$

Note that $\sum_{i=1}^m |W_i^0| = \sum_{i=1}^m d_i$ by the choice of m, and $\sum_{i=1}^{m-1} |W_i^0| > \sum_{i=1}^{m-1} d_i$ by the majorization assumption. Therefore, $d_m > |W_m^0|$ and there exists

$$k = \max\{j < m : |W_j^0| > d_j\}.$$

Notice that $|W_m^0| < d_m \leq d_k < |W_k^0|$ implies $|W_m^0| + 2 \leq |W_k^0|$.

We now have to consider two cases:

Case 1:

If there exists at least one element $w \in W_k^0$, which has disjoint support from every element in W_m^0, then pick one such $w \in W_k^0$ satisfying this property. Define $\left(W_i^1\right)_{i=1}^D$ by:

$$W_i^1 = \begin{cases} W_k^0 \setminus \{w\} & \text{if } i = k, \\ W_m^0 \cup \{w\} & \text{if } i = m, \\ W_i^0 & \text{else.} \end{cases}$$

Case 2: If there is *no* such element $w \in W_k^0$ which has disjoint support from every element in W_m^0, then partition $W_k^0 \cup W_m^0$ into maximal chains, say $C_1, \ldots, C_r$. Note that for each $i = 1, \ldots, r$, the cardinality of the sets $C_i \cap W_k^0$ and $C_i \cap W_m^0$ differ by at most one, since, given $v_k \in W_k^0$ and $v_m \in W_m^0$, we know that v_k and v_m either have disjoint support, or their support sets have intersection of size one. Since $|W_m^0| + 2 \leq |W_k^0|$ then there is a maximal chain C_j that contains one element more from W_k^0 than from W_m^0. Define $\left(W_i^1\right)_{i=1}^D$ by:

$$W_i^1 = \begin{cases} \left(W_k^0 \cup C_j\right) \setminus \left(W_k^0 \cap C_j\right) & \text{if } i = k, \\ \left(W_m^0 \cup C_j\right) \setminus \left(W_m^0 \cap C_j\right) & \text{if } i = m, \\ W_i^0 & \text{else.} \end{cases}$$

In both of the above cases, we have defined $\left(W_i^1\right)_{i=1}^D$ such that

$$\sum_{i=1}^{D} ||W_i^1| - d_i| < \sum_{i=1}^{D} ||W_i^0| - d_i|.$$

Note that $\left(W_i^1\right)_{i=1}^D$ satisfies the majorization condition $\left(|W_i^1|\right)_{i=1}^D \succeq (d_n)_{n=1}^N$. Thus if the sets of $\left(W_i^1\right)_{i=1}^D$ do not span the desired fusion frame, we can repeat the above procedure with $\left(W_i^1\right)_{i=1}^D$ instead of $\left(W_i^0\right)_{i=1}^D$ and get $\left(W_i^2\right)_{i=1}^D$ such that $\sum_{i=1}^D ||W_i^2| - d_i| < \sum_{i=1}^D ||W_i^1| - d_i|$. Continuing in this fashion we will, say after repeating the process l times, arrive at $\left(W_i^l\right)_{i=1}^D$ such that $\sum_{i=1}^D ||W_i^l| - d_i| = 0$, then the sets of $\left(W_i^l\right)_{i=1}^D$ span the desired fusion frame $(W_n)_{n=1}^D$. □

Combining the RFF algorithm for constructing a Reference fusion frame and the techniques in the proof of Theorem 3.21 yields a construction algorithm for unit weighted fusion frames. Explicitly, this algorithm is called the Unit-Weighted Fusion Frame Spectral Tetris Construction (UFF), as defined in [12], and follows.

Remark 3.22. For an explicit example of constructing a unit-weighted fusion frame via RFF and UFF see Example 4.3 in the Appendix of the present paper.

Although Theorem 3.21 only provides sufficient conditions for when UFF can construct a unit weighted fusion frame, the authors of [12] completely characterize the specialized case of unit weighted tight Spectral Tetris fusion frames. In particular, the majorization condition $(\dim V_i)_{i=1}^t \succeq (d_i)_{i=1}^D$ in Theorem 3.21 is also necessary for unit weighted tight fusion frames.

Theorem 3.23 ([12]). *Let $N \geq 2M$ be positive integers and $\{d_i\}_{i=1}^D \subseteq \mathbb{N}$ in non-increasing order such that $\sum_{i=1}^D d_i = N$. Let $(V_i)_{i=1}^t$ be the reference fusion frame for $\{\lambda_m\}_{m=1}^M = \{\frac{N}{M}, \cdots, \frac{N}{M}\}$. Then there exists a unit weighted tight Spectral Tetris fusion frame $(W_i)_{i=1}^D$ for $\mathcal{H}_M$ with $\dim W_i = d_i$ for $i = 1, \ldots, D$, if and only if $(\dim V_i)_{i=1}^t \succeq (d_i)_{i=1}^D$.*

Tight fusion frames are fitting in numerous applications of distributed processing because they are robust against additive noise and erasures. Also the fusion frame operator of a tight fusion frame is ideal for reconstruction purposes because it is a sequence of orthogonal projection operators which sum to a scalar multiple of the identity operator. Moreover, tight fusion frames are maximally robust against the loss of a single projection precisely when the tight fusion frame's projection operators are equidimensional, which is exactly the type of fusion frame Theorem 3.23 constructs. Hence, the complete characterization of unit weighted tight fusion frames in Theorem 3.23 is beneficial because this way researchers will know exactly when and how UFF can construct the tight fusion frames needed for their research.

UFF: Unit-Weighted Fusion Frame Spectral Tetris Construction

Parameters:

- Dimension $M \in \mathbb{N}$.
- Number of frame elements $N \in \mathbb{N}$.
- Eigenvalues $(\lambda_m)_{m=1}^M \subseteq (0, \infty)$ and dimensions $M > d_1 \geq d_2 \geq \cdots \geq d_D > 0$ such that $\sum_{m=1}^M \lambda_m = \sum_{i=1}^D d_i = N$.
- Reference fusion frame $(V_i)_{i=1}^t$ for $(\lambda_m)_{m=1}^M$ such that $(\dim V_i)_{i=1}^t \succeq (d_i)_{i=1}^D$.

Algorithm:

(1) Set $\ell := 0$
(2) Set $W_i^\ell := V_i$ for $0 < i \leq t$ and $W_i^\ell := \emptyset$ for $t < i \leq D$, do
(3) Repeat

(a) If $\sum_{i=1}^D ||W_i^\ell| - d_i| \neq 0$
(i) Set $m = \max\{j \leq D | d_j \neq |W_j^\ell|\}$
(ii) Set $k = \max\{j < m | |W_j^\ell| > d_j\}$
(iii) If $A = \{x \in W_k^\ell | \operatorname{supp}(x) \cap \operatorname{supp}(v) = \emptyset \text{ for all } v \in W_m^\ell\} \neq \emptyset$, then
(A) Pick one $\hat{x} \in A$
(B) $W_k^{\ell+1} := W_k^\ell \setminus \{\hat{x}\}$
(C) $W_m^{\ell+1} := W_m^\ell \cup \{\hat{x}\}$
(D) $W_i^{\ell+1} := W_i^\ell$ for all $i \neq k, m$
(iv) else
(A) Partition $W_k^\ell \cup W_m^\ell$ into maximal chains
(B) Pick one such maximal chain, C_j, which contains one more element from W_k^ℓ than from W_m^ℓ
(C) $W_k^{\ell+1} := \left(W_k^\ell \cup C_j\right) \setminus \left(W_k^\ell \cap C_j\right)$
(D) $W_m^{\ell+1} := \left(W_m^\ell \cup C_j\right) \setminus \left(W_m^\ell \cap C_j\right)$
(E) $W_i^{\ell+1} := W_i^\ell$ for all $i \neq k, m$
(v) Set $\ell := \ell + 1$
(b) end.

(4) Do until $\sum_{i=1}^D ||W_i^\ell| - d_i| = 0$
(5) end.

Output:

- The sets $\left(W_i^\ell\right)_{i=1}^D$ span the desired fusion frame $(W_i)_{i=1}^D$, where $W_i = \operatorname{span}\left(W_i^\ell\right)$ for all $i = 1, \ldots, D$.

ALGORITHM 6: The UFF algorithm for constructing a unit-weighted fusion frame.

6.3.7 Generalized Spectral Tetris Fusion Frame Constructions

Given a spectrum for a desired fusion frame operator and dimensions for the subspaces, UFF can be implemented to construct such a unit weighted fusion frame. However, since not all fusion frames are unit weighted, in [6] the authors developed the first construction method for fusion frames with prescribed weights through an adapted version of PNSTC/STR. The authors completely characterize Spectral Tetris fusion frames and in doing so provide the most general algorithm for Spectral Tetris fusion frames.

From previous discussions in Subsection 6.3.1, the fusion frame $\{(W_i, w_i)\}_{i=1}^D$, with frame operator $\tilde{S}$, arises from a conventional frame when we look at orthonormal bases $\left(\psi_{i,j}\right)_{j=1}^{d_i}$ of the fusion frame subspaces W_i. If it is further assumed that all subspaces have unit weight, i.e. $w_i = 1$ for all $i = \{1, \ldots, D\}$, then $\{\psi_{i,j}\}_{i=1,j=1}^{D,d_i}$ is a frame with unit-norm vectors and frame operator $\tilde{S} = S$. This relationship lead to the definition of a *Spectral Tetris fusion frame* as defined in Definition 3.8. However, to construct arbitrarily weighted Spectral Tetris fusion frames, this definition needs to be amended. In particular, to identify a non-unit weighted fusion frame with a conventional frame, use a tight frame within each subspace of the fusion frame instead of an orthonormal basis.

For a fusion frame $\{(W_i, w_i)\}_{i=1}^D$ in $\mathcal{H}_M$, recall our fusion frame operator $\tilde{S}x = \sum_{i=1}^D w_i^2 \left(P_i\left(x\right)\right)$ for any $x \in \mathcal{H}_M$. Let $\{f_{i,j}\}_{i=1}^{d_i}$ be a tight frame for W_i with frame operator S and let P_i be the orthogonal projection onto W_i. The fusion frame operator becomes:

$$\tilde{S}x = \sum_{i=1}^{D} w_i^2 \left(P_i\left(x\right)\right) = \sum_{i=1}^{D}\sum_{j=1}^{d_i} \langle P_i\left(x\right), f_{i,j}\rangle f_{i,j} = \sum_{i=1}^{D}\sum_{j=1}^{d_i} \langle x, f_{i,j}\rangle f_{i,j} = Sx.$$

Hence, a non-unit weighted fusion frame arises from a conventional frame by identifying a tight frame for each subspace of the fusion frame.

Theorem 3.24. *For $i \in \{1, \ldots, D\}$, let $w_i > 0$, W_i be a subspace of $\mathcal{H}_M$ and $\{f_{i,j}\}_{j=1}^{d_i}$ be a tight frame for W_i with tight frame bounds w_i^2. Then the following are equivalent.*

(1) *$\{(W_i, w_i)\}_{i=1}^D$ is a fusion frame whose fusion frame operator has spectrum $\{\lambda_m\}_{m=1}^M$.*
(2) *$\{f_{i,j}\}_{i=1,j=1}^{D,d_i}$ is a frame whose frame operator has spectrum $\{\lambda_m\}_{m=1}^M$.*

In light of this relationship, to construct arbitrarily weighted fusion frames via Spectral Tetris, first construct a Spectral Tetris frame and then partition this frame in such a way so that the corresponding partition is a tight frame for each subspace of the fusion frame.

Definition 3.25. Suppose $\{(W_i, w_i)\}_{i=1}^D$ is a fusion frame with fusion frame operator $\tilde{S}$. We say $\{(W_i, w_i)\}_{i=1}^D$ is a *Spectral Tetris fusion frame* if there exists

a Spectral Tetris frame $F = \{f_n\}_{n=1}^N$ with frame operator S and a partition $\{J_i\}_{i=1}^D$ of $\{1, \ldots, N\}$ such that $\{f_n\}_{n \in J_i}$ is a tight frame for W_i with tight frame bound w_i^2. Further, we say F and $\{J_i\}_{i=1}^D$ generate $\{(W_i, w_i)\}_{i=1}^D$.

We would like to construct a fusion frame which has a desired sequence of eigenvalues, subspace weights and dimensions. However, we make no mention of the norms of the vectors which span the subspaces of the fusion frame because, as we will see, different sequences of norms can produce the same fusion frame.

Example 3.26 ([6]). Construct a fusion frame in $\mathcal{H}_2$ with eigenvalues $\{2, 3\}$ and a sequence of weights $\{\sqrt{2}, 1\}$ with corresponding subspace dimensions $\{2, 1\}$. PNSTC can produce a variety of frames whose frame operator has this spectrum:

(1) The sequence of norms $\{\sqrt{2}, \sqrt{2}, 1\}$ produces the frame

$$\begin{bmatrix} f_1 \, f_2 \, f_3 \end{bmatrix} = \begin{bmatrix} \sqrt{2} & 0 & 0 \\ 0 & \sqrt{2} & 1 \end{bmatrix}.$$

(2) The sequence of norms $\left(1, 1, \sqrt{2}, 1\right)$ produces the frame

$$\begin{bmatrix} g_1 \; g_2 \; g_3 \; g_4 \end{bmatrix} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & \sqrt{2} & 1 \end{bmatrix}.$$

(3) The sequence of norms $\left(1, \sqrt{\frac{3}{2}}, \sqrt{\frac{3}{2}}, 1\right)$ produces the frame

$$\begin{bmatrix} h_1 \; h_2 \; h_3 \; h_4 \end{bmatrix} = \begin{bmatrix} 1 & \sqrt{\frac{1}{2}} & \sqrt{\frac{1}{2}} & 0 \\ 0 & 1 & -1 & 1 \end{bmatrix}.$$

A fusion frame $\{(W_i, w_i)\}_{i=1}^2$, with weights $w_1 = \sqrt{2}, w_2 = 1$, is then obtained via PNSTC by defining $W_1 = \text{span}\,(f_1, f_2)$, $W_2 = \text{span}\,(f_3)$, or $W_1 = \text{span}\,(g_1, g_2, g_3)$, $W_2 = \text{span}\,(g_4)$, or $W_1 = \text{span}\,(h_1, h_2, h_3)$, $W_1 = \text{span}\,(h_4)$. All three generate the same fusion frame.

The differences among the constructions in Example 3.26 are superficial; (2) simply splits a vector from (1) into two colinear vectors, and (3) takes two orthogonal vectors from (2) and combines them into a 2×2 block spanning the same 2-dimensional space. In fact, all Spectral Tetris frames which generate a given fusion frame are related in this manner. The following theorem explicitly states this and proves that every Spectral Tetris fusion frame can be generated from a Spectral Tetris frame, where each subspace of the fusion frame is spanned by equal norm, orthogonal frame vectors. Moreover, the weights of the subspaces of the Spectral Tetris fusion frame are the norms of the frame vectors from the Spectral Tetris frame.

Theorem 3.27 ([6]). *If $\{(W_i, w_i)\}_{i=1}^D$ is a spectral tetris fusion frame in $\mathcal{H}_M$, then there exists a spectral tetris frame $F = \{f_n\}_{n=1}^N$ and a partition $\{J_i\}_{i=1}^D$ of $\{1, \dots, N\}$ generating this fusion frame such that $\|f_n\| = w_i$ and $\langle f_n, f_{n'} \rangle = 0$, for $n, n' \in J_i$ for each $i \in \{1, \dots, D\}$.*

In order to apply the Spectral Tetris fusion frame construction, first a Spectral Tetris frame must be constructed and hence the sequence of weights/norms and eigenvalues must be Spectral Tetris ready. However, to construct the Spectral Tetris fusion frame, these sequences have further constraints as defined in the following theorem.

Theorem 3.28 ([6]). *Let $\{w_i\}_{i=1}^D$ be a sequence of weights, $\{\lambda_m\}_{m=1}^M$ a sequence of eigenvalues, and $\{d_i\}_{i=1}^D$ a sequence of dimensions. Let $N = \sum_{i=1}^D d_i$, and now consider each w_i repeated d_i times. We will use a double index to reference specific weights and a single index to emphasize the ordering:*

$$\{w_{i,j}\}_{i=1,j=1}^{D,d_i} = \{w_n\}_{n=1}^N.$$

Then Spectral Tetris can construct a fusion frame whose subspaces have the given weights and dimensions, and whose frame operator has the given spectrum if and only if there exists a Spectral Tetris ready (as in Definition 2.35) permutation of $\{w_n\}_{n=1}^N$ and $\{\lambda_m\}_{m=1}^M$, say $\{w_{\sigma n}\}_{n=1}^N$ and $\{\lambda_{\sigma' m}\}_{m=1}^M$ whose associated partition $1 \leq n_1 \leq \cdots \leq n_M = N$ satisfies:

(1) $\sum_{n=1}^{n_i} w_{\sigma n}^2 < \sum_{m=1}^{i} \lambda_{\sigma' m}$, *then*

- (a) *if* $\sum_{n=1}^{n_{i+1}} w_{\sigma n}^2 < \sum_{m=1}^{i+1} \lambda_{\sigma' m}$, *then for* $w_{u,v}, w_{p,q} \in \{w_{\sigma n}\}_{n=n_i}^{n_{i+1}+1}$, $v \neq q$
- (b) *if* $\sum_{n=1}^{n_{i+1}} w_{\sigma n}^2 = \sum_{m=1}^{i+1} \lambda_{\sigma' m}$, *then for* $w_{u,v}, w_{p,q} \in \{w_{\sigma n}\}_{n=n_i}^{n_{i+1}}$, $v \neq q$

(2) $\sum_{n=1}^{n_i} w_{\sigma n}^2 = \sum_{m=1}^{i} \lambda_{\sigma' m}$, *then*

- (a) *if* $\sum_{n=1}^{n_{i+1}} w_{\sigma n}^2 < \sum_{m=1}^{i+1} \lambda_{\sigma' m}$, *then for* $w_{u,v}, w_{p,q} \in \{w_{\sigma n}\}_{n=n_i+1}^{n_{i+1}+1}$, $v \neq q$
- (b) *if* $\sum_{n=1}^{n_{i+1}} w_{\sigma n}^2 = \sum_{m=1}^{i+1} \lambda_{\sigma' m}$, *then for* $w_{u,v}, w_{p,q} \in \{w_{\sigma n}\}_{n=n_i+1}^{n_{i+1}}$, $v \neq q$

for all $i = 1, \dots, M-1$.

Theorem 3.28 gives necessary and sufficient conditions for the construction of a Spectral Tetris fusion frame. Moreover, it is possible for a sequence of weights/norms and a sequence of eigenvalues to satisfy the Spectral Tetris ready condition and hence such a Spectral Tetris Frame exists but no partition of these sequences satisfies the orthogonality conditions (1)(a,b) and (2)(a,b) of Theorem 3.28. However this does not suggest that such a fusion frame cannot exist, it just implies that Spectral Tetris cannot construct such a fusion frame.

Example 3.29. Given the dimensions $\{d_i\}_{i=1}^5 = \{4, 2, 2, 2, 1\}$ and the eigenvalues $\{\lambda_m\}_{m=1}^6 = \{\frac{11}{6}, \frac{11}{6}, \frac{11}{6}, \frac{11}{6}, \frac{11}{6}, \frac{11}{6}\}$, PNSTC will construct the following unit norm frame:

$$\begin{bmatrix} 1 & \sqrt{\frac{5}{12}} & \sqrt{\frac{5}{12}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & \sqrt{\frac{7}{12}} & -\sqrt{\frac{7}{12}} & \sqrt{\frac{1}{3}} & \sqrt{\frac{1}{3}} & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & \sqrt{\frac{2}{3}} & -\sqrt{\frac{2}{3}} & \sqrt{\frac{1}{4}} & \sqrt{\frac{1}{4}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{3}{4}} & -\sqrt{\frac{3}{4}} & \sqrt{\frac{1}{6}} & \sqrt{\frac{1}{6}} & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{5}{6}} & -\sqrt{\frac{5}{6}} & \sqrt{\frac{1}{12}} & \sqrt{\frac{1}{12}} \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{11}{12}} & -\sqrt{\frac{11}{12}} \end{bmatrix}$$

A unit weighted fusion frame with these dimensions and spectrum is known to exist due to combinatorial arguments. However, the hypotheses of Theorem 3.28 cannot be satisfied in this case because no four columns can be chosen to be pairwise orthogonal. Hence there is no Spectral Tetris fusion frame with these properties.

Although the conditions in Theorem 3.28 completely characterize Spectral Tetris fusion frames, it can be a time consuming task to find a Spectral Tetris ready sequence which satisfies (1)(a,b) and (2)(a,b). The following theorem provides easier sufficient conditions for when a Spectral Tetris fusion frame can be constructed.

Theorem 3.30 ([6]). *Consider $\mathcal{H}_M$ and a sequence of weights $w_1 \leq w_2 \leq \cdots \leq w_D$ with corresponding subspace dimensions $\{d_i\}_{i=1}^D$, and a sequence of eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_M$. Let the doubly indexed sequence $\{w_{i,j}\}_{i=1,j=1}^{D,d_i}$ represent w_i each repeated d_i times. Now PNSTC/STR will build a weighted fusion frame $\{(W_i, w_i)\}_{i=1}^D$, with $\dim(W_i) = d_i$ and whose frame operator has the given spectrum if there exists an ordering $\{w_n\}_{n=1}^N$ of $\{w_{i,j}\}_{i=1,j=1}^{D,d_i}$ such that*

(1) $\sum_{n=1}^N w_n^2 = \sum_{m=1}^M \lambda_m$
(2) $w_{D-1,1}^2 + w_{D,1}^2 \leq \lambda_1$
(3) *If $w_l = w_{i,j}, w_{l'} = w_{i'j'}$ with $i = i'$ and $l < l'$, then $\sum_{n=l}^{l'-1} w_n^2 \geq 2\lambda_M$.*

The conditions in Theorem 3.30 are more relaxed than that of Theorem 3.28; however, finding an ordering of weights which achieves condition (3) is no small task. Intuitively, we would want to space like-weights as far apart as possible in our ordering in order to maximize $\sum_{n=l}^{l'-1} w_n^2$. When all of the subspaces have the same dimension then the ordering of the like-weights becomes obvious. We will start in this more obvious case and provide sufficient conditions for when PNSTC/STR can construct an equidimensional, tight fusion frame.

Corollary 3.31 ([6]). *Consider $\mathcal{H}_M$ and a sequence of weights $w_1 \leq w_2 \leq \cdots \leq w_D$. PNSTC/STR can construct a tight weighted fusion frame with the given weights, all subspaces of dimension k, (eigenvalue $\lambda = \frac{k}{M} \sum_{i=1}^D a_i^2$) provided both of the following hold:*

(1) $w_{D-1}^2 + w_D^2 \leq \lambda$
(2) $\frac{k}{M} \leq \frac{1}{2}$

Next, Theorem 3.30 is specialized to the case of equidimensional fusion frames and sufficient conditions are given for when PNSTC/STR can construct such fusion frames.

Corollary 3.32 ([6]). *Consider $\mathcal{H}_M$, a sequence of weights $w_1 \leq w_2 \leq \cdots \leq w_D$ and a sequence of eigenvalues $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_M$. PNSTC/STR can construct a weighted fusion frame $\{(W_i, w_i)\}_{i=1}^D$, all subspaces dimension k, and with the given spectrum provided all of the following hold:*

(1) $k\sum_{n=1}^D w_n^2 = \sum_{m=1}^M \lambda_m$
(2) $w_{D-1}^2 + w_D^2 \leq \lambda_1$
(3) $\sum_{n=1}^D w_n^2 \geq 2\lambda_M$

Remark 3.33. To construct the fusion frame in Corollary 3.31 and Corollary 3.32, write each weight w_i repeated k times and arrange these weights as follows: $\{a_1, \ldots, a_m, a_1, \ldots, a_m, \ldots\}$. Then proceed to use PNSTC/STR on this collection of norms. We provide this arrangement of the sequence of weights because it guarantees that such a fusion frame can be constructed so long as all of the conditions in the respective corollary are met. However, other arrangements are possible.

To help illustrate the generalized Spectral Tetris construction method of a fusion frame, an example is now included which constructs an equidimensional, weighted fusion frame, which utilizes Corollary 3.32.

Example 3.34. Construct a weighted fusion frame in $\mathcal{H}_5$ with 9 two-dimensional subspaces, weights $\{w_i\}_{i=1}^9 = \{1, 1, 1, 1, \sqrt{2}, \sqrt{2}, \sqrt{3}, \sqrt{3}, 2\}$ and spectrum $\{\lambda_m\}_{m=1}^5 = \{7, 7, 7, 7, 8\}$. Notice that conditions (1), (2), and (3) of Corollary 3.32 are met. Indeed:

(1) $k\sum_{i=1}^9 w_i^2 = 2\,(18) = 36 = \sum_{i=1}^5 \lambda_m$
(2) $w_{D-1}^2 + w_D^2 = 7 \leq 7 = \lambda_1$
(3) $\sum_{i=1}^9 w_i^2 = 18 \geq 16 = 2\lambda_5$

and hence such a construction is possible.

To construct such a fusion frame, first construct the corresponding Spectral Tetris frame via PNSTC. Write each norm k times and arrange these weights in the following order:

$$\{1, 1, 1, 1, \sqrt{2}, \sqrt{2}, \sqrt{3}, \sqrt{3}, 2, 1, 1, 1, 1, \sqrt{2}, \sqrt{2}, \sqrt{3}, \sqrt{3}, 2\}.$$

PNSTC constructs the following Spectral Tetris frame:

$$\begin{bmatrix} 1 & 1 & 1 & 1 & \sqrt{2} & \sqrt{\frac{2}{3}} & \sqrt{\frac{1}{3}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{4}{3}} & -\sqrt{\frac{8}{3}} & \sqrt{3} & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 2 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & \sqrt{2} & \sqrt{2} & 1 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{2} & -\sqrt{2} & 2 \end{bmatrix}.$$

Grouping the frame vectors in the following way will yield the desired fusion frame:

$$W_1 = \text{span}\{f_1, f_{10}\}; W_2 = \text{span}\{f_2, f_{11}\}; W_3 = \text{span}\{f_3, f_{12}\};$$
$$W_4 = \text{span}\{f_4, f_{13}\}; W_5 = \text{span}\{f_5, f_{14}\}; W_6 = \text{span}\{f_6, f_{15}\};$$
$$W_7 = \text{span}\{f_7, f_{16}\}; W_8 = \text{span}\{f_8, f_{17}\}; W_9 = \text{span}\{f_9, f_{18}\};$$

where each subspace is two-dimensional with respective desired weight and the spectrum of the fusion frame operator is $\{7, 7, 7, 7, 8\}$.

Remark 3.35 ([6]). In order for PNSTC/STR to build a desired fusion frame, a complex relationship among partial sums of weights, partial sums of eigenvalues, and dimensions of our subspaces must be satisfied according to Theorem 3.28. We simplified this relationship in Theorem 3.30 and its corollaries to achieve concrete constructions via PNSTC/STR. While these extra assumptions still allow a variety of fusion frames to be created, they are best suited for fusion frames with relatively flat spectrum. For example, (1) and (3) of Corollary 3.32 imply

$$\frac{\sum_{m=1}^{M} \lambda_m}{k} \geq 2\lambda_M,$$

and this can clearly be manipulated to

$$\frac{\text{Average } \left(\{\lambda_m\}_{m=1}^{M}\right)}{2\lambda_M} \geq \frac{k}{M}.$$

Hence if we desire PNSTC/STR to guarantee the construction of fusion frames with relatively large subspaces, our prescribed frame operator must have a relatively flat spectrum. However, the conditions used here are of the correct order for practical applications as we generally do not work with large subspaces or with eigenvalues for the frame operator which are very spread out.

6.4 Appendix

Additional examples of the SFR, SFFR, RFF and UFF construction methods are included here.

An explicit example of constructing an equidimensional unit-weighted fusion frame via SFFR is now included. The remark following this example explains how a frame can be constructed using SFR.

Example 4.1. Construct an equidimensional, unit-weighted fusion frame in $\mathcal{H}_3$ with 5 two-dimensional subspaces and spectrum $\{\lambda_m\}_{m=1}^3 = \{\frac{13}{3}, \frac{10}{3}, \frac{7}{3}\}$.

Note that $D = 5 \geq \frac{13}{3} \geq \frac{10}{3} \geq \frac{7}{3} \geq 2$ and $\sum_{m=1}^3 \lambda_m = 10 = 2\,(5) = kD \in \mathbb{N}$; hence the parameters of the algorithm are met.

- Set $j = 1$
- For $m = 1$ do

  (1) $\lambda_1 = \frac{13}{3} \geq 1$ then
  - (a) $f_1 := e_1$.
  - (b) $j := j + 1 = 1 + 1 = 2$.
  - (c) $\lambda_1 := \lambda_1 - 1 = \frac{13}{3} - 1 = \frac{10}{3}$.

  (2) $\lambda_1 = \frac{10}{3} \geq 1$ then
  - (a) $f_2 := e_1$.
  - (b) $j := 2 + 1 = 3$.
  - (c) $\lambda_1 := \frac{10}{3} - 1 = \frac{7}{3}$.

  (3) $\lambda_1 = \frac{7}{3} \geq 1$ then
  - (a) $f_3 := e_1$.
  - (b) $j := 3 + 1 = 4$.
  - (c) $\lambda_1 := \frac{7}{3} - 1 = \frac{4}{3}$.

  (4) $\lambda_1 = \frac{4}{3} \geq 1$ then
  - (a) $f_4 := e_1$.
  - (b) $j := 4 + 1 = 5$.
  - (c) $\lambda_1 := \frac{4}{3} - 1 = \frac{1}{3}$.

  (5) $\lambda_1 = \frac{1}{3} < 1$ then
  - (a) $f_5 := \sqrt{\frac{\frac{1}{3}}{2}} \cdot e_1 + \sqrt{1 - \frac{\frac{1}{3}}{2}} \cdot e_2 = \sqrt{\frac{1}{6}} \cdot e_1 + \sqrt{\frac{5}{6}} \cdot e_2$.
  - (b) $f_6 := \sqrt{\frac{\frac{1}{3}}{2}} \cdot e_1 - \sqrt{1 - \frac{\frac{1}{3}}{2}} \cdot e_2 = \sqrt{\frac{1}{6}} \cdot e_1 - \sqrt{\frac{5}{6}} \cdot e_2$.
  - (c) $j := 5 + 2 = 7$.
  - (d) $\lambda_{m+1} = \lambda_2 := \lambda_{m+1} - (2 - \lambda_m) = \frac{10}{3} - \left(2 - \frac{1}{3}\right) = \frac{5}{3}$.
  - (e) $\lambda_1 := 0$.

  (6) end.

- For $m = 2$ (we have $\lambda_m = \lambda_2 = \frac{5}{3}$ and $j = 7$) do

  (1) $\lambda_2 = \frac{5}{3} \geq 1$ then
  - (a) $f_7 := e_2$.
  - (b) $j := j + 1 = 8$.

(c) $\lambda_2 := \lambda_2 - 1 = \frac{2}{3}$.

(2) $\lambda_2 = \frac{2}{3} < 1$ then

(a) $f_8 := \sqrt{\frac{\frac{2}{3}}{2}} \cdot e_2 + \sqrt{1 - \frac{\frac{2}{3}}{2}} \cdot e_3 = \sqrt{\frac{1}{3}} \cdot e_2 + \sqrt{\frac{2}{3}} \cdot e_3$.

(b) $f_9 := \sqrt{\frac{\frac{2}{3}}{2}} \cdot e_2 - \sqrt{1 - \frac{\frac{2}{3}}{2}} \cdot e_3 = \sqrt{\frac{1}{3}} \cdot e_2 - \sqrt{\frac{2}{3}} \cdot e_3$.

(c) $j := j + 2 = 10$.

(d) $\lambda_{m+1} = \lambda_3 := \lambda_3 - (2 - \lambda_2) = \frac{7}{3} - \left(2 - \frac{2}{3}\right) = 1$.

(e) $\lambda_2 := 0$.

(3) end.

- For $m = 3$ (we have $\lambda_m = \lambda_3 = 1$ and $j = 10$) do

(1) $\lambda_3 = 1 \geq 1$ then

(a) $f_{10} := e_3$.

(b) $j := j + 1 = 11$.

(c) $\lambda_3 := \lambda_3 - 1 = 0$.

(2) end.

- end.

Output:

- Define the two-dimensional subspaces $\{W_i\}_{i=1}^5$ as the following

$$W_i := \text{span}\{f_{i+5j} : j = 0, 1\}.$$

Explicitly, this yields:

$$W_1 = \text{span}\{f_1, f_6\}, W_2 = \text{span}\{f_2, f_7\}, W_3 = \text{span}\{f_3, f_8\},$$
$$W_4 = \text{span}\{f_4, f_9\}, W_5 = \text{span}\{f_5, f_{10}\}.$$

It is straightforward to check that each of the subspaces $\{W_i\}_{i=1}^5$ are 2-dimensional and the spectrum of the fusion frame operator is $\{\frac{13}{3}, \frac{10}{3}, \frac{7}{3}\}$. Therefore $\{W_i\}_{i=1}^5$ is an equidimensional, unit weighted fusion frame with 5 two dimensional subspaces and spectrum $\{\frac{13}{3}, \frac{10}{3}, \frac{7}{3}\}$. Note that the corresponding frame vectors are at most 2-sparse.

Remark 4.2. Example 4.1 can be slightly simplified to also be an example of the SFR construction for a unit norm frame. Explicitly in Example 4.1, to adapt the SFFR algorithm construction to a construction for SFR our parameters and output would change to the following:

***New SFR* Parameters:**

- Dimension $3 \in \mathbb{N}$.
- Real eigenvalues $5 \geq \frac{13}{3} \geq \frac{10}{3} \geq \frac{7}{3} \geq 2$, number of frame vectors 10 satisfying $\sum_{m=1}^3 \lambda_m = 10 = D \in \mathbb{N}$.

Algorithm: The algorithm will be the exact same as in Example 4.1.
***New SFR* Output:**

- Unit norm frame $\{f_j\}_{j=1}^{10}$ with spectrum $\{\lambda_m\}_{m=1}^{3} = \{\frac{13}{3}, \frac{10}{3}, \frac{7}{3}\}$.

Next, an illustrative example of constructing a unit weighted fusion frame using UFF is included.

Example 4.3. Construct a unit-weighted fusion frame in $\mathcal{H}_4$ with 11 frame elements, eigenvalues $\left(\frac{11}{4}, \frac{11}{4}, \frac{11}{4}, \frac{11}{4}\right)$ and dimensions $3 \geq 3 \geq 2 \geq 1 \geq 1 \geq 1$.

Notice that $\sum_{m=1}^{4} \lambda_m = 11 = \sum_{i=1}^{6} d_i = N$.

Recall the unit norm tight frame with eigenvalue $\lambda = \frac{11}{4}$ constructed in Example 2.12,

$$T^* = [f_1 f_2 \cdots f_{11}] = \begin{bmatrix} 1 & 1 & \sqrt{\frac{3}{8}} & \sqrt{\frac{3}{8}} & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & \sqrt{\frac{5}{8}} & -\sqrt{\frac{5}{8}} & 1 & \sqrt{\frac{2}{8}} & \sqrt{\frac{2}{8}} & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{6}{8}} & -\sqrt{\frac{6}{8}} & 1 & \sqrt{\frac{1}{8}} & \sqrt{\frac{1}{8}} & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & \sqrt{\frac{7}{8}} & -\sqrt{\frac{7}{8}} & 1 \end{bmatrix}.$$

It is a straight forward check of the RFF algorithm to see that the reference fusion frame given for frame T^* is as follows:

$$V_1 = \text{span}\{f_1, f_5, f_8, f_{11}\}, V_2 = \text{span}\{f_2, f_6\},$$
$$V_3 = \text{span}\{f_3, f_9\}, V_4 = \text{span}\{f_4, f_{10}\}, V_5 = \text{span}\{f_7\}.$$

Note that the majorization condition, $(\dim V_i)_{i=1}^{5} \succeq (d_i)_{i=1}^{6}$, is also satisfied.

- $\ell := 0$
- $W_i^0 := \emptyset$ for $5 < i \leq 6$. Hence,

$$W_1^0 := \{f_1, f_5, f_8, f_{11}\}; W_2^0 := \{f_2, f_6\}; W_3^0 := \{f_3, f_9\};$$
$$W_4^0 := \{f_4, f_{10}\}; W_5^0 := \{f_7\}; W_6^0 := \emptyset.$$

- $\sum_{i=1}^{6} ||W_i^0| - d_i| = 4 \neq 0$
  (1) $m := \max\{j \leq 6 | d_j \neq |W_j^0|\} = 6$
  (2) $k := \max\{j < 6 || W_j^0| > d_j\} = 4$
  (3) Since $A = \{x \in W_4^0 | \text{ supp } (x) \cap \text{supp } (v) = \emptyset \text{ for all } v \in W_6^0\} \neq \emptyset$, then
    (a) Pick one $\hat{x} \in A$. We can pick f_{10}.
    (b) Yielding the new subspaces:

$$W_1^1 := W_1^0 = \{f_1, f_5, f_8, f_{11}\}; W_2^1 := W_2^0 = \{f_2, f_6\};$$
$$W_3^1 : W_3^0 = \{f_3, f_9\}; W_4^1 : W_4^0 \setminus \{f_{10}\} = \{f_4\};$$

$$W_5^1 := W_5^0 = \{f_7\}; W_6^1 := W_6^0 \cup \{f_{10}\} = \{f_{10}\}.$$

(4) $\ell := 0 + 1 = 1$.

- Repeat with $\ell = 1$.
- $\sum_{i=1}^{6} ||W_i^1| - d_i| = 2 \neq 0$

(1) $m := \max\{j \leq 6 | d_j \neq |W_j^1|\} = 2$.
(2) $k := \max\{j < 2 | |w_j^1| > d_j\} = 1$.
(3) $A = \{x \in W_1^1 | \operatorname{supp}(x) \cap \operatorname{supp}(v) = \emptyset \text{ for all } v \in W_2^1\} = \{f_{11}\} \neq \emptyset$.
(a) Pick one $\hat{x} \in A$. We can pick f_{11}.
(b) Yielding the new subspaces:

$$W_1^2 := W_1^1 \setminus \{f_{11}\} = \{f_1, f_5, f_8\};$$
$$W_2^2 := W_2^1 \cup \{f_{11}\} = \{f_2, f_6, f_{11}\};$$
$$W_3^2 : W_3^1 = \{f_3, f_9\}; W_4^2 : W_4^1 = \{f_4\};$$
$$W_5^2 := W_5^1 = \{f_7\}; W_6^2 := W_6^1 = \{f_{10}\}.$$

(4) $\ell := 1 + 1 = 2$

- Repeat with $\ell = 2$.
- $\sum_{i=1}^{6} ||W_i^2| - d_i| = 0$
- end.

Output:

- The sets $\left(W_i^2\right)_{i=1}^6$ span the desired fusion frame $(W_i)_{i=1}^D$, where $W_i = \operatorname{span}\left(W_i^2\right)$ for all $i = 1, \ldots, 6$.

Acknowledgements The authors are deeply indebted to the referees for their comprehensive reports which significantly shortened and improved the paper.

References

1. R. Calderbank, P.G. Casazza, A. Heinecke, G. Kutyniok, A. Pezeshki, Sparse fusion frames: existence and construction. Adv. Comput. Math. **35**(1), 1–31 (2011)
2. P.G. Casazza, G. Kutyniok, Frames of subspaces, in *Wavelets, Frames and Operator Theory* (College Park, MD, 2003), ed. by C. Heil, P.E.T. Jorgensen, D.R. Larson. Contemporary Mathematics, vol. 345 (American Mathematical Society, Providence, 2004), pp. 87–114
3. P.G. Casazza, G. Kutyniok (eds.), *Finite Frames: Theory and Applications*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2012)
4. P.G. Casazza, M. Leon, Existence and construction of finite tight frames. J. Concr. Appl. Math. **4**, 277–289 (2006)
5. P.G. Casazza, M. Leon, Existence and construction of finite frames with a given frame operator. Int. J. Pure Appl. Math **63**, 149–159 (2010)

6. P.G. Casazza, J. Peterson, Weighted fusion frame construction via spectral tetris. Adv. Comput. Math. **40**(2), 335–351 (2014)
7. P.G. Casazza, M. Fickus, J.J. Kovaĉević, M. Leon, J.C. Tremain, A physical interpretation of finite tight frames, in *Harmonic Analysis and Applications (in Honor of John Benedetto)*, ed. by C. Heil (Birkhaäuser, Boston, 2006), pp. 51–76
8. P.G. Casazza, G. Kutyniok, S. Li, Fusion frames and distributed processing. Appl. Comput. Harmon. Anal. **25**(1), 114–132 (2008)
9. P.G. Casazza, M. Fickus, D. Mixon, Y. Wang, Z. Zhou, Constructing tight fusion frames. Appl. Comput. Harmon. Anal. **30**(2), 175–187 (2011)
10. P.G. Casazza, A. Heinecke, F. Krahmer, G. Kutyniok, Optimally sparse frames. IEEE Trans. Inf. Theory **57**(11), 7279–7287 (2011)
11. P.G. Casazza, A. Heinecke, G. Kutyniok, Optimally sparse fusion frames: existence and construction, in *Proceedings of SampTA*, Singapore (2011)
12. P.G. Casazza, M. Fickus, A. Heinecke, Y. Wang, Z. Zhou, Spectral tetris fusion frame constructions. J. Fourier Anal. Appl. **18**(4), 828–851 (2012)
13. P.G. Casazza, M. Fickus, D. Mixon, J. Peterson, I. Smalyanau, Every Hilbert space frame has a Naimark complement. J. Math. Anal. Appl. **406**(1), 111–119 (2013)
14. P.G. Casazza, A. Heinecke, K. Kornelson, Y. Wang, Z. Zhou, Necessary and sufficient conditions to perform spectral tetris. Linear Algebra Appl. **438**(5), 2239–2255 (2013)
15. O. Christensen, *Frames and Riesz Bases*. Applied and Numerical Harmonic Analysis (Birkhäuser, Boston, 2002)
16. R. Duffin, A. Schaeffer, A class of nonharmonic Fourier series. Trans. Am. Math. Soc. **72**(2), 341–366 (1952)
17. G. Kutyniok, A. Pezeshki, R. Calderbank, Fusion frames and robust dimension reduction, in *Proceedings of 42nd Annual Conference on Information Sciences and Systems (CISS)* (Princeton University, Princton, 2008), pp. 264–268
18. G. Kutyniok, A. Pezeshki, R. Calderbank, T. Liu, Robust dimension reduction, fusion frames, and Grassmannian packings. Appl. Comput. Harmon. Anal. **26**(1), 64–76 (2009)
19. C.J. Rozell, D.H. Johnson, Analyzing the robustness of redundant population codes in sensory and feature extraction systems. Neurocomputing **69**(10), 1215–1218 (2006)

Part III
Bandlimitation and Generalizations

Chapter 7
System Approximations and Generalized Measurements in Modern Sampling Theory

Holger Boche and Volker Pohl

Abstract This chapter studies several aspects of signal reconstruction of sampled data in spaces of bandlimited functions. In the first part, signal spaces are characterized in which the classical sampling series uniformly converge, and we investigate whether adaptive recovery algorithms can yield uniform convergence in spaces where non-adaptive sampling series does not. In particular, it is shown that the investigation of adaptive signal recovery algorithms needs completely new analytic tools since the methods used for nonadaptive reconstruction procedures, which are based on the celebrated Banach–Steinhaus theorem, are not applicable in the adaptive case.

The second part analyzes the approximation of the output of stable linear time-invariant (LTI) systems based on samples of the input signal, and where the input is assumed to belong to the Paley–Wiener space of bandlimited functions with absolute integrable Fourier transform. If the samples are acquired by point evaluations of the input signal f, then there exist stable LTI systems H such that the approximation process does not converge to the desired output Hf even if the oversampling factor is arbitrarily large. If one allows generalized measurements of the input signal, then the output of every stable LTI system can be uniformly approximated in terms of generalized measurements of the input signal.

The last section studies the situation where only the amplitudes of the signal samples are known. It is shown that one can find specific measurement functionals such that signal recovery of bandlimited signals from amplitude measurement is possible, with an overall sampling rate of four times the Nyquist rate.

H. Boche (✉) • V. Pohl
Technische Universität München, Lehrstuhl für Theoretische Informationstechnik,
Arcisstraße 21, 80333 München, Germany
e-mail: boche@tum.de; volker.pohl@tum.de

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_7

7.1 Introduction

The great success of digital signal proceeding lies in the fact that analog signals observed in the physical world can equivalently be represented by a sequence of complex numbers. These digital signals can then be processed and filtered very quickly and efficiently on digital computers. Sampling theory is the theoretical foundation of the conversion from analog to digital signals and vice versa. Because of its fundamental importance for modern information theory, signal processing and communications, there exists a long and extensive list of impressive research results in this area starting with the seminal work of Shannon [60]. We refer to excellent survey articles and textbooks such as [38, 41, 59, 64, 73] and to [14, 15] for historical comments on the topic.

Sampling theory was originally formulated for bandlimited signals with finite energy. Later these results were extended to non-bandlimited signals [16, 25, 26, 29] and to broader classes of bandlimited functions, in particular to functions which do not necessarily have finite energy [19, 52, 71]. But these results often took into consideration only the pointwise convergence of the reconstruction series. From a practical point of view, however, it is often necessary to control the peak value of the reconstructed signal because electronic circuits and devices (like amplifiers, antennas, etc.) have only limited dynamic ranges. Moreover, the energy efficiency of these devices usually and largely depends on this dynamic range [68]. To control the peak value of the signals, one has to investigate the uniform convergence of the sampling series. This will be done in some detail in the first part of this chapter (Sec. 7.4). The starting point will be a classical result [13] which shows that the uniform Shannon sampling series is locally uniformly convergent for all bandlimited signals with an absolute integrable Fourier transform. Then we present several extensions of this result to larger signal spaces and we investigate whether it is possible to have global uniform convergence on the entire real axis. In particular, we discuss the influence of the sampling points and we investigate whether it is possible to apply adaptive reconstruction algorithms to obtain signal recovery methods which are uniformly convergent. Classical sampling series, like the one of Shannon, are fixed for the whole signal space under consideration. These series may or may not converge for all signals in this signal space. But even if the sampling series does not converge for all functions in the space, it might be possible to adapt the reconstruction series to the actual signal to obtain a signal approximation which converges uniformly to the desired signal. However, it will be shown that for the common signal spaces of bandlimited signals, such an adaption of the recovery series essentially gives no improvement of the global uniform convergence behavior. These investigations in Sec. 7.4 are strongly related to one of the cornerstones of functional analysis, namely to the theorem of Banach–Steinhaus. This important theorem is a very powerful and elegant tool to investigate nonadaptive algorithms, and in particular to prove the divergence of sampling series considered in this chapter. For adaptive algorithms, however, the Banach–Steinhaus theorem cannot

be applied. Therefore completely new tools are necessary for the investigation of such recovery algorithms, which are related to some early works of Paul Erdős.

The second part of this chapter (Sec. 7.5) investigates another aspect of sampling-based signal processing. Whereas the sampling theorem deals with the reconstruction of a signal f from its samples $\{f(\lambda_n)\}$, the main task of digital signal processing is often not to reconstruct f, but to process the sampled data $\{f(\lambda_n)\}$ such that an approximation of the processed signal $g = \mathrm{H}f$ is obtained, where H is a certain linear transformation. In other words, one wants a digital implementation of the analog system H. It seems to be widely accepted that such a digital implementation is always possible, at least for stable linear systems. This is certainly true for bandlimited signals with finite energy. However, for more general signal spaces, a digital implementation of stable LTI systems may fail, even on such spaces where the sampling series uniformly converges. This remarkable observation is presented and discussed in Sec. 7.5. Moreover, we also discuss the influence of data acquisition for possible signal-based signal processing. Usually, signals are sampled by point evaluations $f \mapsto \{f(\lambda_n)\}_{n\in\mathbb{Z}}$. However, more general measurement functionals are possible to digitize an analog signal. It is shown that generalized measurement functionals will enable the digital implementation of analog signal processing schemes on spaces where a digital implementation based on point evaluations fail.

The last part (Sec. 7.6) considers another application where generalized measurement functionals are necessary to guarantee signal recovery from signal samples. Here we consider the situation where only the amplitude of the signal samples is available, but not the phase. This problem, known as *phase retrieval*, plays an important role in many different applications. Even though there is a long history of research [31] on phase retrieval, it is still not clear whether signal reconstruction from the amplitudes of point evaluations is in general, possible. Here we will show, however, that specifically designed measurement functionals will allow signal recovery of bandlimited functions from the knowledge of its amplitudes only. The recovery algorithm will be partially based on the sampling series investigated in Sec. 7.4.

7.2 Preliminaries and Signal Models

7.2.1 General Notations

We use standard notations. In particular, the set of all continuous functions on the real axis $\mathbb{R}$ is denoted by $\mathscr{C}(\mathbb{R})$, and $\mathscr{C}_0(\mathbb{R})$ stands for all $f \in \mathscr{C}(\mathbb{R})$ which vanish at infinity. Both spaces are equipped with the supremum norm $\|f\|_\infty = \sup_{t\in\mathbb{R}} |f(t)|$. If $1 \le p < \infty$ or $p = \infty$, then

$$\|f\|_p = \left(\int_{-\infty}^{\infty} |f(t)|^p \,\mathrm{d}t\right)^{1/p} \qquad \text{and} \qquad \|f\|_\infty = \operatorname*{ess\,sup}_{t\in\mathbb{R}} |f(t)|$$

is the L^p norm of f, and $L^p(\mathbb{R})$ stands for the Banach space of all measurable functions on $\mathbb{R}$ with finite L^p norm. Similarly, if $\mathbb{S} \subset \mathbb{R}$ is a finite interval of length $|\mathbb{S}|$ on $\mathbb{R}$, then $L^p(\mathbb{S})$ with $1 \leq p \leq \infty$ stands for the set of measurable functions on $\mathbb{S}$ with finite norm, defined by

$$\|f\|_p = \left(\frac{1}{|\mathbb{S}|} \int_{\mathbb{S}} |f(t)|^p \, \mathrm{d}t \right)^{1/p} \qquad \text{and} \qquad \|f\|_\infty = \frac{1}{|\mathbb{S}|} \operatorname*{ess\,sup}_{t \in \mathbb{S}} |f(t)| \, .$$

In particular, $L^2(\mathbb{R})$ and $L^2(\mathbb{S})$ are Hilbert spaces with the inner products

$$\langle f, g \rangle = \int_{\mathbb{R}} f(t) \, \overline{g(t)} \, \mathrm{d}t \qquad \text{and} \qquad \langle f, g \rangle = \frac{1}{|\mathbb{S}|} \int_{\mathbb{S}} f(t) \, \overline{g(t)} \, \mathrm{d}t \, , \tag{7.1}$$

respectively. For any $f \in L^1(\mathbb{R})$, its *Fourier transform* is defined by

$$\hat{f}(\omega) = (\mathscr{F} f)(\omega) = \int_{-\infty}^{\infty} f(t) \, \mathrm{e}^{-\mathrm{i}t\omega} \, \mathrm{d}t \, , \qquad \omega \in \mathbb{R} \, .$$

Because $L^1(\mathbb{R}) \cap L^2(\mathbb{R})$ is a dense subset of $L^2(\mathbb{R})$, *Plancherel's theorem* extends $\mathscr{F}$ to a unitary operator on $L^2(\mathbb{R})$. There, $\mathscr{F}$ satisfies *Parseval's formula* $\langle \hat{f}, \hat{g} \rangle = 2\pi \langle f, g \rangle$ for all $f, g \in L^2(\mathbb{R})$. By Riesz–Thorin interpolation, $\mathscr{F}$ can be extended to any $L^p(\mathbb{R})$ with $1 < p < 2$, and for $p > 2$ it can be defined in the distributional sense (see, e.g., [39, 58]).

7.2.2 Spaces of Bandlimited Functions

In many applications, and especially in communications, bandlimited signals are the prevailing signal model. In order to set our discussion in the context of known results, we consider two families of bandlimited functions, namely Paley–Wiener and Bernstein spaces.

Paley–Wiener Spaces For $\sigma > 0$ and $1 \leq p \leq \infty$, the *Paley–Wiener space* $\mathscr{PW}_\sigma^p$ is the set of all functions f that can be represented as

$$f(z) = \frac{1}{2\pi} \int_{-\sigma}^{\sigma} \hat{f}(\omega) \, \mathrm{e}^{\mathrm{i}\omega z} \, \mathrm{d}\omega \, , \qquad z \in \mathbb{C} \tag{7.2}$$

for some $\hat{f} \in L^p([-\sigma, \sigma])$. Thus f can be represented as the inverse Fourier transform of a function $\hat{f}$ in $L^p([-\sigma, \sigma])$, and we say that f has *bandwidth* σ. The norm in $\mathscr{PW}_\sigma^p$ is defined by $\|f\|_{\mathscr{PW}_\sigma^p} = \|\hat{f}\|_{L^p([-\sigma,\sigma])}$.

The Paley–Wiener spaces are nested. Indeed, Hölder's inequality implies that $\|f\|_{\mathscr{PW}_\sigma^q} \leq \|f\|_{\mathscr{PW}_\sigma^p}$ for all $f \in \mathscr{PW}_\sigma^p$ and all $1 \leq q \leq p$. This yields the inclusions $\mathscr{PW}_\sigma^p \subset \mathscr{PW}_\sigma^q$ for all $1 \leq q \leq p < \infty$. In particular, $\mathscr{PW}_\sigma^1$ is the largest space

in the family of Paley–Wiener spaces. By the properties of the Fourier transform, it follows easily from (7.2) that any Paley–Wiener function f is continuous on $\mathbb{R}$ with $\|f\|_\infty \leq \frac{1}{2\pi}\|f\|_{\mathcal{PW}^1_\sigma}$ and the *Riemann–Lebesgue lemma* shows that any Paley–Wiener function vanishes at infinity such that $\mathcal{PW}^p_\sigma \subset \mathscr{C}_0(\mathbb{R})$ for every $1 \leq p \leq \infty$.

Similarly, as for the $L^p(\mathbb{R})$ spaces, the Paley–Wiener space $\mathcal{PW}^2_\sigma$ plays a particular role since it is a Hilbert space with the L^2 inner product given on the left-hand side of (7.1). Moreover, it is even a reproducing kernel Hilbert space. This means that for every $\lambda \in \mathbb{R}$, and for all $f \in \mathcal{PW}^2_\sigma$, the point evaluation $f(\lambda)$ can be written as an inner product

$$f(\lambda) = \langle f, r_\lambda \rangle \qquad \text{with} \qquad r_\lambda(t) = \frac{\sin(\sigma[t-\lambda])}{\pi[t-\lambda]} ,$$

and with the *reproducing kernel* $r_\lambda \in \mathcal{PW}^2_\sigma$ with $\|r_\lambda\|_{\mathcal{PW}^2_\pi} = \sqrt{\sigma/\pi}$.

Bernstein Spaces For any $\sigma > 0$, the set $\mathscr{B}_\sigma$ contains all entire functions of exponential type $\leq \sigma$, i.e., to every $f \in \mathscr{B}_\sigma$ and every $\epsilon > 0$ there is a constant $C = C(f, \epsilon)$ such that

$$|f(z)| \leq C\, \mathrm{e}^{(\sigma+\epsilon)|z|} \quad \text{for all} \quad z \in \mathbb{C} .$$

Then for $1 \leq p \leq \infty$, the *Bernstein space* $\mathscr{B}^p_\sigma$ is the set of all $f \in \mathscr{B}_\sigma$ whose restriction to the real axis belongs to $L^p(\mathbb{R})$. The norm in $\mathscr{B}^p_\sigma$ is defined as the $L^p(\mathbb{R})$ norm of f on $\mathbb{R}$.

Functions in the Bernstein spaces $\mathscr{B}^p_\sigma$ are also bandlimited in the sense that they have a Fourier transform with finite support. This follows from the *Paley–Wiener theorem* [39, 58]. It states that f is an entire function of exponential type $\leq \sigma$, if and only if it is the Fourier transform of a distribution with compact support which is contained in the interval $[-\sigma, \sigma]$. Finally, we remark that the *theorem of Plancherel–Pólya* implies that there exists a constant $C = C(p, \sigma)$, such that for all $f \in \mathscr{B}^p_\sigma$

$$|f(t + \mathrm{i}\tau)| \leq C\, \|f\|_p\, \mathrm{e}^{\sigma|\tau|} \quad \text{for all } t, \tau \in \mathbb{R} . \tag{7.3}$$

Thus every $f \in \mathscr{B}^p_\sigma$ is uniformly bounded on every line parallel to the real axis. It follows that $\mathscr{B}^p_\sigma \subset \mathscr{B}^q_\sigma \subset \mathscr{B}^\infty_\sigma$ for all $1 \leq p \leq q \leq \infty$. In particular, $\mathscr{B}^\infty_\sigma$ is the largest space in the family of Bernstein spaces. The space of all functions $f \in \mathscr{B}^\infty_\sigma$ for which $f(t) \to 0$ as $t \to \pm\infty$ will be denoted by $\mathscr{B}^\infty_{\sigma,0}$, and we have the relations $\mathscr{B}^\infty_{\sigma,0} \subset \mathscr{C}_0(\mathbb{R})$ and $\mathscr{B}^p_\sigma \subset \mathscr{C}(\mathbb{R})$ for every $1 \leq p \leq \infty$. Note also that Plancherel's theorem shows that $\mathscr{B}^2_\sigma = \mathcal{PW}^2_\sigma$. So overall we have the relation

$$\mathscr{B}^2_\sigma = \mathcal{PW}^2_\sigma \subset \mathcal{PW}^1_\sigma \subset \mathscr{B}^\infty_\sigma .$$

Without any loss of generality, we normalize the bandwidth σ of our signals to $\sigma = \pi$ throughout this chapter.

In applications, it is often important to control the peak value of the signals because of limited dynamic ranges of power amplifiers and other hardware components [68]. Moreover, the energy efficiency is an increasingly important aspect for mobile communication networks, and since the efficiency of high power amplifiers is directly related to the peak-to-average power ratio of the signals, it is necessary to control the peak values of the signals to design energy efficient systems. For such applications, $\mathscr{B}_{\pi}^{\infty}$ is the appropriated signal space. Moreover, in sampling and reconstruction of stochastic processes, the space $\mathscr{PW}_{\pi}^{1}$ plays a fundamental role [8, 17] because the spectral densities of such processes are L^1 functions, in general. Consequently, one has to investigate the behavior of the reconstruction series for functions in $\mathscr{PW}^{1}$. For these reasons, the spaces $\mathscr{B}_{\pi}^{\infty}$ and $\mathscr{PW}_{\pi}^{1}$ are the primary signal spaces considered in this chapter.

7.3 Classical Sampling Theory: A Short Introduction

Sampling theory deals with the reconstruction of functions f in terms of their values (*samples*) $f(\lambda_n)$ on an appropriated set $\{\lambda_n\}$ of sampling points. The particular choice of the set $\{\lambda_n\}$ and in particular the density of the points λ_n determine whether it is possible to reconstruct f from its samples [59]. This theory is the foundation of all modern digital signal processing [60]. Moreover, in [30], *Feynman* discusses sampling theory in a much wider context (*"physics of computations"*) and its importance for theoretical physics. This section reviews some of the most important results in sampling theory, as far as they will be needed in the subsequent discussions.

7.3.1 Uniform Sampling

We start our discussion with the situation where the sampling points $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ are distributed uniformly on the real axis, i.e., where $\lambda_n = n$ for all $n \in \mathbb{Z}$. Then the fundamental initial result in sampling theory is the so-called *cardinal series* [37] which is also known by the name *Whittaker–Shannon–Kotelnikov sampling theorem*. Let $f \in \mathscr{PW}_{\pi}^{p}$ be a bandlimited function and consider the *Shannon series* of degree N

$$(\mathrm{S}_N f)(t) = \sum_{n=-N}^{N} f(n) r_n(t) = \sum_{n=-N}^{N} f(n) \frac{\sin(\pi[t-n])}{\pi[t-n]} \tag{7.4}$$

with the reproducing kernels r_n of $\mathscr{PW}_{\pi}^{2}$. This series is intended to approximate the given function f from its samples $\{f(n)\}_{n=-N}^{N}$. The question is whether, and in which sense, $\mathrm{S}_N f$ converges to the given function f as $N \to \infty$.

Original research was mainly focused on functions in the Hilbert space $\mathscr{PW}_\pi^2$, and it is well known that

$$\lim_{N\to\infty} \|f - \mathrm{S}_N f\|_{\mathscr{PW}_\pi^2} = 0 \quad \text{for all} \quad f \in \mathscr{PW}_\pi^2 \,. \tag{7.5}$$

This result easily follows by observing that the reproducing kernels $\{r_n\}_{n\in\mathbb{Z}}$ form an orthonormal basis for $\mathscr{PW}_\pi^2$ [36]. Moreover, since $\mathscr{PW}_\pi^2$ is a reproducing kernel Hilbert space, Cauchy–Schwarz inequality immediately gives

$$|f(t) - (\mathrm{S}_N f)(t)| = |\langle f - \mathrm{S}_N f, r_t\rangle| \le \|r_t\|_{\mathscr{PW}_\pi^2} \|f - \mathrm{S}_N f\|_{\mathscr{PW}_\pi^2} \,.$$

Since $\|r_t\|_{\mathscr{PW}_\pi^2} = 1$ for all $t \in \mathbb{R}$, (7.5) also implies

$$\lim_{N\to\infty} \max_{t\in\mathbb{R}} |f(t) - (\mathrm{S}_N f)(t)| = 0 \quad \text{for all} \quad f \in \mathscr{PW}_\pi^2 \,.$$

In other words, $\mathrm{S}_N f$ converges uniformly to f for all $f \in \mathscr{PW}_\pi^2$, and one can even show that $S_N f$ converges absolutely. Moreover, it is fairly easy to extend the above results for $\mathscr{PW}_\pi^2$ to all Paley–Wiener space $\mathscr{PW}_\pi^p$ with $1 < p \le \infty$. More precisely, one has the following statement [38]:

Theorem 1. *For each $1 < p \le \infty$ and for all $f \in \mathscr{PW}_\pi^p$, we have*

$$f(t) = \lim_{N\to\infty} (\mathrm{S}_N f)(t) = \sum_{n=-\infty}^{\infty} f(n) \frac{\sin(\pi[t-n])}{\pi[t-n]}$$

where the sum converges absolutely and uniformly on the whole real axis $\mathbb{R}$, and also in the norm of $\mathscr{PW}_\pi^p$.

Theorem 1 gives a simple and convenient reconstruction formula for all functions in the Paley–Wiener spaces $\mathscr{PW}_\pi^p$ with $p > 1$. However, we want to stress that Theorem 1 does not hold for the largest Paley–Wiener space $\mathscr{PW}_\pi^1$. The convergence of the sampling series in this space will be considered in more detail in Section 7.4. In particular, Corollary 1 below will show that with oversampling, S_N converges uniformly on $\mathscr{PW}_\pi^1$.

7.3.2 Nonuniform Sampling Series

The Shannon sampling series (7.4) is based on uniform signal samples taken at integer values $\lambda_n = n$, $n \in \mathbb{Z}$. To gain more flexibility, one may consider series which reconstruct a function f from samples $\{f(\lambda_n)\}_{n\in\mathbb{Z}}$ taken at a set $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ of nonuniform sampling points. To choose an appropriate sampling set Λ, we start again with the Hilbert space $\mathscr{B}_\pi^2 = \mathscr{PW}_\pi^2$. For this signal space, the so-called complete interpolating sequences are suitable sampling sets.

Definition 1 (Interpolating and Sampling Sequence). Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be a sequence in $\mathbb{C}$ and let $\tilde{\mathrm{S}} : \mathscr{B}^2_\pi \to \ell^2$ be the associated sampling operator defined by $\tilde{\mathrm{S}} : f \mapsto \{f(\lambda_n)\}_{n\in\mathbb{Z}}$. We call Λ

- a *sampling sequence* for $\mathscr{B}^2_\pi$ if $\tilde{\mathrm{S}}$ is injective.
- an *interpolation sequence* for $\mathscr{B}^2_\pi$ if $\tilde{\mathrm{S}}$ is surjective.
- *complete interpolating* for $\mathscr{B}^2_\pi$ if $\tilde{\mathrm{S}}$ is bijective.

In the following, we always use as sampling sets complete interpolating sequences for $\mathscr{B}^2_\pi$. Such sequences were completely characterized by *Minkin* [47] after [49, 51] already gave characterizations under mild constrains on Λ.

Over- and Undersampling Assuming that Λ is complete interpolating for $\mathscr{B}^2_\pi$, then $\tilde{\mathrm{S}}$ is an isomorphism between the signal space $\mathscr{B}^2_\pi$ and the sequence space ℓ^2 such that the interpolation problem $f(\lambda_n) = c_n$, $n \in \mathbb{Z}$ has a unique solution $f \in \mathscr{B}^2_\pi$ for every sequence $\{c_n\}_{n\in\mathbb{Z}} \in \ell^2$. Now, let $\beta \in \mathbb{R}$ and consider the signal space $\mathscr{B}^2_{\beta\pi}$. If $\beta > 1$, then $\mathscr{B}^2_\pi$ is a proper subset of $\mathscr{B}^2_{\beta\pi}$ and $\tilde{\mathrm{S}}$ will no longer be injective viewed as an operator on $\mathscr{B}^2_{\beta\pi} \to \ell^2$. Therefore it will not be possible to reconstruct every $f \in \mathscr{B}^2_{\beta\pi}$ uniquely from its samples $\tilde{\mathrm{S}}f$. In this case, we say that $\mathscr{B}^2_{\beta\pi}$ is *undersampled* by Λ. Conversely, if $\beta < 1$, then $\mathscr{B}^2_{\beta\pi}$ is a proper subset of $\mathscr{B}^2_\pi$ and the sampling operator $\tilde{\mathrm{S}} : \mathscr{B}^2_{\beta\pi} \to \ell^2$ is injective but not surjective. In this case, every function $f \in \mathscr{B}^2_{\beta\pi}$ can uniquely be reconstructed from its samples $\tilde{\mathrm{S}}$ but there exist sequences $\{c_n\}_{n\in\mathbb{Z}} \in \ell^2$ such that the interpolation problem $f(\lambda_n) = c_n$, $n \in \mathbb{Z}$ has no solution in $\mathscr{B}^2_{\beta\pi}$. In this case, we say that $\mathscr{B}^2_{\beta\pi}$ is *oversampled* by Λ and $1/\beta$ is the *oversampling factor*.

Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be a complete interpolating sequence for $\mathscr{B}^2_\pi$ and define the function

$$\varphi(z) = z^{\delta_\Lambda} \lim_{R\to\infty} \prod_{\substack{|\lambda_n|<R \\ \lambda_n\neq 0}} \left(1 - \frac{z}{\lambda_n}\right) \qquad \text{with} \qquad \delta_\Lambda = \begin{cases} 1 & \text{if } 0 \in \Lambda \\ 0 & \text{otherwise} \end{cases} . \tag{7.6}$$

One can show [43, 70] that the product in (7.6) converges uniformly on every compact subset of $\mathbb{C}$ and that φ is an entire function of exponential type π. It follows from (7.6) that $\varphi(\lambda_n) = 0$, i.e., Λ is the zero set of the function φ which is often called the *generating function* of Λ. Based on the function (7.6), one defines for every $n \in \mathbb{Z}$

$$\varphi_n(z) := \frac{\varphi(z)}{\varphi'(\lambda_n)(z - \lambda_n)} , \qquad z \in \mathbb{C} . \tag{7.7}$$

Again, these are entire functions of exponential type π which solve the interpolation problem $\varphi_n(\lambda_n) = 1$ and $\varphi_n(\lambda_k) = 0$ if $k \neq n$. Following the ideas of classical Lagrange interpolation, one considers for any $N \in \mathbb{N}$, the approximation operator

$$(\mathrm{A}_N f)(z) = \sum_{n=-N}^{N} f(\lambda_n)\, \varphi_n(z)\ , \tag{7.8}$$

which only involves $2N+1$ sampling values. The aim is to approximate any function $f \in \mathscr{B}_\pi^2$ by $\mathrm{A}_N f$ in such a way that the approximation error $\|f - \mathrm{A}_N f\|_{\mathscr{B}_\pi^2}$ becomes less than any arbitrary given bound ϵ as soon as $N \in \mathbb{N}$ is sufficiently large.

By the definition of the interpolation kernels φ_n, it is clear that $\mathrm{A}_N f$ satisfies the interpolation condition $(\mathrm{A}_N f)(\lambda_n) = f(\lambda_n)$ for all $n = 0, \pm 1, \pm 2, \ldots, \pm N$, and one can show that for every $f \in \mathscr{B}_\pi^2$, the sequence $\{\mathrm{A}_N f\}_{N\in\mathbb{N}}$ converges in $\mathscr{B}_\pi^2$ as $N \to \infty$. Since Λ is completely interpolating, we therefore have $\mathrm{A}_N f$ converging to f for every $f \in \mathscr{B}_\pi^2$ [70].

Theorem 2. *Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be a complete interpolating sequence for $\mathscr{B}_\pi^2$ and let $\{\varphi_n\}_{n\in\mathbb{Z}}$ be the functions defined by (7.7) based on the generating function (7.6). Then*

$$f(t) = \lim_{N\to\infty} (\mathrm{A}_N f)(t) = \sum_{n=-\infty}^{\infty} f(\lambda_n)\, \varphi_n(t)$$

for all $f \in \mathscr{B}_\pi^2$ where the sum converges in the norm of $\mathscr{B}_\pi^2$, and uniformly on $\mathbb{R}$.

In general, the characterization of complete interpolating sequences Λ is fairly complicated and the calculation of the corresponding generating function φ via (7.6) can be computationally difficult. Fortunately, an important subset of complete interpolating sequences is known which simplifies the entire procedure considerably. These are the zero sets of the so-called sine-type functions.

Definition 2 (Sine-type function). An entire function φ of exponential type π is said to be a *sine-type function* if it has simple and separated zeros and if there exist positive constants A, B, H such that

$$A\, \mathrm{e}^{\sigma|\eta|} \le |\varphi(\xi + \mathrm{i}\eta)| \le B\, \mathrm{e}^{\sigma|\eta|} \quad \text{for all} \quad \xi \in \mathbb{R} \text{ and } |\eta| \ge H\ .$$

Any sine-type function can be determined from its zero set Λ by (7.6).

Example 1. The uniform sampling considered above is a special case of nonuniform sampling. The sampling set Λ is obtained as the zero set of the sine-type function $\varphi(z) = \sin(\pi z)$, which is equal to $\Lambda = \{\lambda_n = n\}_{n\in\mathbb{Z}}$. The corresponding interpolation kernels (7.7) become $\varphi_n(z) = \sin(\pi[z-n])/(\pi[z-n])$ and (7.8) becomes equal to (7.4).

If the sampling set Λ is chosen as the zero set of a sine-type function, then Theorem 2 can be extended to all Bernstein spaces $\mathscr{B}_\pi^p$ with $1 < p < \infty$. Thus, the nonuniform sampling series (7.8) reconstructs every function in $\mathscr{B}_\pi^p$. More precisely, one has the following statement [43, Lect. 22].

Theorem 3. *Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be the zero set of a sine-type function φ and let $\{\varphi_n\}_{n\in\mathbb{Z}}$ and A_N be defined as in (7.7) and (7.8), respectively. Then for each $1 \le p < \infty$*

$$f(t) = \lim_{N\to\infty} (\mathrm{A}_N f)(t) = \sum_{n=-\infty}^{\infty} f(\lambda_n)\,\varphi_n(t)\,, \qquad \text{for all } f \in \mathscr{B}_\pi^p$$

where the sum converges uniformly on $\mathbb{R}$ and for $1 < p < \infty$, it also converges in the norm of $\mathscr{B}_\pi^p$.

Also here we would like to stress that Theorem 3 does not hold for the largest space $\mathscr{B}_\pi^\infty$ in the family of Bernstein spaces, and we will discuss the $\mathscr{B}_\pi^\infty$-case later in Sec. 7.4 (cf. Conjecture 2 and Theorem 13 below).

7.4 On the Global Uniform Convergence of Sampling Series

Theorems 1 and 3 establish the uniform convergence of the sampling series on $\mathscr{PW}_\pi^p$ for $1 < p \le \infty$ and on $\mathscr{B}_\pi^p$ for $1 \le p < \infty$, respectively. However, both results cannot easily be extended to the largest spaces $\mathscr{PW}_\pi^1$ and $\mathscr{B}_\pi^\infty$. This section reviews and discusses some recent results which investigate on which signal spaces and under which conditions the sampling series converges uniformly on $\mathbb{R}$. So we are going to investigate the behavior of the quantity

$$\max_{t\in\mathbb{R}} |f(t) - (\mathrm{S}_N f)(t)| \qquad \text{or} \qquad \max_{t\in\mathbb{R}} |f(t) - (\mathrm{A}_N f)(t)|$$

as N tends to infinity. This quantity is an important measure for the stability of the reconstruction process since it allows us to control the peak value of the error between the approximation $\mathrm{A}_N f$ and the function f itself. The question is whether the maximum error can be made arbitrarily small for a sufficiently large approximation degree N.

For the signal space $\mathscr{PW}_\pi^1$, a classical theorem due to *Brown* states that the Shannon sampling series S_N converges uniformly on compact sets of $\mathbb{R}$.

Theorem 4 (Brown [13]). *For all $f \in \mathscr{PW}_\pi^1$ and for all $T > 0$, we have*

$$\lim_{N\to\infty} \left(\max_{t\in[-T,T]} |f(t) - (\mathrm{S}_N f)(t)| \right) = 0\,.$$

Based on this result we consider now the following two questions.

1. Is it possible to have even uniform convergence on the entire real axis, i.e., is it possible to replace the interval $[-T, T]$ by $\mathbb{R}$ in Theorem 4?
2. Is it possible to extend Theorem 4 to the larger space $\mathscr{B}_\pi^\infty$ of bounded bandlimited functions?

Since Theorem 4 is based on uniform sampling without oversampling, we may hope to achieve these extensions by replacing the uniform sampling series S_N with a non-uniform series A_N and by using oversampling.

7.4.1 Weak Divergence of the Shannon Sampling Series

We begin by asking whether the Shannon sampling series (7.4) converges uniformly on the whole real axis $\mathbb{R}$ for every function $f \in \mathscr{PW}_\pi^1$. The negative answer is given by the following theorem [7].

Theorem 5. *There exists a signal $f_0 \in \mathscr{PW}_\pi^1$ such that*

$$\limsup_{N\to\infty} \|\mathrm{S}_N f_0\|_\infty = \infty \,. \tag{7.9}$$

Remark 1. Since $\mathscr{PW}_\pi^1 \subset \mathscr{C}_0(\mathbb{R})$, all functions in $\mathscr{PW}_\pi^1$ are bounded on $\mathbb{R}$. Therefore Theorem 5 implies in particular that there exists an $f_0 \in \mathscr{PW}_\pi^1$ such that $\limsup_{N\to\infty} \|f_0 - \mathrm{S}_N f_0\|_\infty = \infty$.

Remark 2. In fact, the divergence behavior described by Theorem 5 is not a particular property of the Shannon sampling series but holds for a large class of approximation processes which rely on uniform sampling. More precisely, [7] proved Theorem 5 not only for the sampling series (7.4) but also for all sampling series with the general form

$$(\mathrm{R}_N f)(t) = (\mathrm{T} f)(t) + \sum_{n=-N}^{N} f(n)\, \phi_n(t) \,, \tag{7.10}$$

where $\mathrm{T} : \mathscr{PW}_\pi^1 \to \mathscr{B}_\pi^\infty$ is linear and bounded, and $\phi_n \in \mathscr{B}_\pi^\infty$ are certain interpolation kernels. If the series R_N satisfies[1] the following three properties:

- The kernels ϕ_n are uniformly bounded, i.e., $\|\phi_n\|_\infty \le C_\phi < \infty$ for all $n \in \mathbb{Z}$,
- $(\mathrm{R}_N f)(t)$ converges pointwise to $f(t)$ for all $f \in \mathscr{PW}_\pi^2$,
- The operator T is such that there exist two constants $C, D > 0$ such that for all $f \in \mathscr{PW}_\pi^1$ always $\sup_{t\in\mathbb{R}} |(\mathrm{T} f)(t)| \le C \max_{|z|\le D} |f(z)|$,

then it shows the same divergence behavior as in Theorem 5. Moreover, the particular function $f_0 \in \mathscr{PW}_\pi^1$, for which $\|\mathrm{S}_N f_0\|_\infty$ diverges, is universal in the sense that all interpolation series R_N with the above properties diverge for f_0.

[1] Interpolation series R_N which satisfy these conditions include the so-called *Valiron series* [5, 37, 66] or *Tschakaloff series* [37, 63].

The proof of Theorem 5 relies on an explicit construction of $f_0 \in \mathscr{PW}_\pi^1$ and a corresponding subsequence $\{N_k\}_{k=1}^\infty$ such that

$$(\mathrm{S}_{N_k} f_0)(N_k + 1/2) \geq C_1 \sqrt{k^3} + C_2 \to \infty \quad \text{as } k \to \infty \,.$$

Because of this construction, one has the lim sup-divergence in (7.9). In a sense, this is a weak notion of divergence, because one designs a very specific function f_0 and a corresponding subsequence of approximations $\{\mathrm{S}_{N_k} f_0\}_{k\in\mathbb{N}}$ such that divergence emerges. This notion of (weak) divergence is sufficient to show that the approximation procedure is not always convergent. However, it does not show that there exists no recovery procedure at all. In particular, the divergence result of Theorem 5 does not allow to answer the following two questions:

Q-1 Let $f \in \mathscr{PW}_\pi^1$ be arbitrary. Does there exist a specific sequence $\mathscr{N}(f) = \{N_k = N_k(f)\}_{k\in\mathbb{N}}$, depending on f, such that

$$\sup_{k\in\mathbb{N}} \|f - \mathrm{S}_{N_k} f\|_\infty < \infty \,. \tag{7.11}$$

Q-2 Does there exist a universal approximation sequence $\mathscr{N} = \{N_k\}_{k\in\mathbb{N}}$ such that (7.11) holds for all $f \in \mathscr{PW}_\pi^1$?

Remark 3. Note that a negative answer to Q-1 implies a negative answer to Q-2. Conversely, a positive answer to Q-2 implies a positive answer to Q-1.

With that said, Theorem 5 gives only a weak statement about the global divergence behavior of the Shannon series on $\mathscr{PW}_\pi^1$. Because if Q-2 were to have a positive answer, then one would have a globally convergent method to reconstruct every $f \in \mathscr{PW}_\pi^1$ from its sampled values $\{f(n)\}_{n\in\mathbb{Z}}$. But even if only question Q-1 were to have a positive answer, signal recovery would still be possible, but with an adaptive approximation process which depends on the actual function.

Divergence results like those in Theorem 5 are usually proved using the *uniform boundedness principle*. This principle is one of the cornerstones of functional analysis and it may be formulated as follows (see, e.g., [57]):

Theorem 6 (Banach–Steinhaus, [3]). *Let $\{\mathrm{T}_n\}_{n\in\mathbb{N}}$ be a sequence of linear operators $\mathrm{T}_n : \mathscr{X} \to \mathscr{Y}$ mapping a Banach space $\mathscr{X}$ into a normed space $\mathscr{Y}$ with the operator norms*

$$\|\mathrm{T}_n\| = \sup_{f\in\mathscr{X}} \frac{\|\mathrm{T}_n f\|_{\mathscr{Y}}}{\|f\|_{\mathscr{X}}} \,.$$

If $\sup_{n\in\mathbb{N}} \|\mathrm{T}_n\| = \infty$ then there exists an $x_0 \in \mathscr{X}$ such that

$$\sup_{n\in\mathbb{N}} \|\mathrm{T}_n x_0\|_{\mathscr{Y}} = \infty \,. \tag{7.12}$$

In fact, the set $\mathscr{D}$ of all $x_0 \in \mathscr{X}$ which satisfy (7.12) is a residual set in $\mathscr{X}$.

Remark 4. In a Banach space, a *residual set* is the complement of a set of first category (a meager set) and therefore it is a set of second category (i.e., a nonmeager set). In the following we use that the countable intersection of residual sets is again a residual set. In particular, the countable intersection of open dense subsets is a residual set [42].

Here, we shortly discuss how the theorem of Banach–Steinhaus can be used to investigate the two questions Q-1 and Q-2. In particular, we want to show the limitations of the Banach–Steinhaus theorem for answering question Q-1.

To prove Theorem 5, based on the uniform boundedness principle, it is sufficient to shown that the norms

$$\|\mathrm{S}_N\| = \sup\left\{\|\mathrm{S}_N f\|_\infty \ : \ f \in \mathscr{PW}^1_\pi \ , \ \|f\|_{\mathscr{PW}^1_\pi} \leq 1\right\}$$

of the operators $\mathrm{S}_N : \mathscr{PW}^1_\pi \to \mathscr{B}^\infty_\pi$, defined in (7.4), are not uniformly bounded. This was done in [7], where it was shown that there exists a constant C_S such that

$$\|\mathrm{S}_N\| \geq C_S \, \log N \qquad \text{for all } N \in \mathbb{N} \, . \tag{7.13}$$

Then Theorem 6 implies immediately that there exits a residual set $\mathscr{D} \subset \mathscr{PW}^1_\pi$ such that

$$\limsup_{N\to\infty} \|\mathrm{S}_N f\|_\infty = +\infty \quad \text{for all} \quad f \in \mathscr{D} \, .$$

Next we use Theorem 6 to investigate question Q-2. Since (7.13) holds for all $N \in \mathbb{N}$, the same reasoning can be applied to any subsequence $\mathscr{N} = \{N_k\}_{k\in\mathbb{N}}$ of $\mathbb{N}$. Then the Banach–Steinhaus theorem states that there exists a residual set $\mathscr{D}(\mathscr{N}) \subset \mathscr{PW}^1_\pi$ such that

$$\limsup_{k\to\infty} \|\mathrm{S}_{N_k} f\|_\infty = +\infty \quad \text{for all} \quad f \in \mathscr{D}(\mathscr{N}) \, .$$

This shows that the answer to question Q-2 is actually negative, i.e., there exists no universal subsequence $\mathscr{N} = \{N_k\}_{k\in\mathbb{N}}$ such that $\mathrm{S}_{N_k} f$ converges uniformly for every $f \in \mathscr{PW}^1_\pi$. One can even say more about the size of the divergence set. Let $\{\mathscr{N}_v\}_{v\in\mathbb{N}}$ be a countable collection of subsequences of $\mathbb{N}$. Then to every $\mathscr{N}_v = \{N_{v,k}\}_{k\in\mathbb{N}}$ there exists a subset $\mathscr{D}(\mathscr{N}_v) \subset \mathscr{PW}^1_\pi$ such that

$$\limsup_{k\to\infty} \|\mathrm{S}_{N_{v,k}} f\|_\infty = +\infty \quad \text{for all} \quad f \in \mathscr{D}(\mathscr{N}_v) \, .$$

Since each $\mathscr{D}(\mathscr{N}_v)$ is a residual sets in $\mathscr{PW}^1_\pi$ and since we have only countably many sets, *Baire's category theorem* (see, e.g., [57]) implies that the intersection of these sets

$$\bigcap_{v} \mathscr{D}(\mathscr{N}_v) \neq \varnothing \tag{7.14}$$

is again a (nonempty) residual set in $\mathscr{PW}_{\pi}^{1}$. So given a countable collection of subsets $\{\mathscr{N}_v\}_{v\in\mathbb{N}}$, the set of functions $f \in \mathscr{PW}_{\pi}^{1}$ for which

$$\limsup_{k\to\infty} \|\mathrm{S}_{N_{v,k}} f\|_\infty = +\infty \quad \text{for all} \quad \mathscr{N}_v = \{N_{v,k}\}_{k\in\mathbb{N}} \in \{\mathscr{N}_v\}_{v\in\mathbb{N}}$$

is nonmeager (of second category) in $\mathscr{PW}_{\pi}^{1}$.

However, the above reasoning cannot be extended to give a definite answer to question Q-1. Because for a negative answer to Q-1, we need to show that

$$\bigcap_{\substack{\mathscr{N}_v=\{N_{v,k}\}_{k\in\mathbb{Z}} \\ \mathscr{N}_v \text{ is a subsequence of } \mathbb{N}}} \mathscr{D}(\mathscr{N}_v) \neq \varnothing \,.$$

In other words, we have to show that there exists a function $f_* \in \mathscr{PW}_{\pi}^{1}$ such that $\lim_{k\to\infty} \|\mathrm{S}_{N_{v,k}} f_*\|_\infty = +\infty$ for *every* subsequence $\mathscr{N}_v = \{N_{v,k}\}_{k\in\mathbb{N}}$ of $\mathbb{N}$. However, in contrast to (7.14), the set of all subsequences of $\mathbb{N}$ contains uncountably many elements and the uncountable intersection of residual sets may not be of second category. It even may be empty. Therefore, using the above technique, it is not possible to decide whether this intersection is empty or not. This way we are not able to answer question Q-1.

In the next subsection we are going to investigate question Q-1 for the Shannon sampling series in more detail, using completely new techniques. Before that, we give an example of an operator sequence which is (weakly) divergent, but for which question Q-2 can be answered positively.

Example 2 (Approximation by Walsh functions). Consider the usual Lebesgue space $L^2([0,1])$ of square integrable functions on the interval $[0,1]$ and let $\{\psi_n\}_{n=0}^{\infty}$ be the orthonormal system of Walsh functions [67] in $L^2([0,1])$, where the functions are indexed as in [33]. Let $\mathrm{P}_N : L^2([0,1]) \to \overline{\mathrm{span}}\{\psi_n : n = 0,1,\ldots,N\}$ be the orthogonal projection onto the first $N+1$ Walsh functions. Now we view P_N as a mapping $L^\infty([0,1]) \to L^\infty([0,1])$ with the corresponding norm

$$\|\mathrm{P}_N\| = \sup\{\|\mathrm{P}_N f\|_\infty \,:\, f \in L^\infty([0,1]) \,,\ \|f\|_\infty \le 1\} \,.$$

Then one can show [33] that

$$\limsup_{N\to\infty} \|\mathrm{P}_N\| = +\infty \qquad \text{but} \qquad \|\mathrm{P}_{2^k}\| = 1 \text{ for all } k \in \mathbb{N} \,.$$

So the sequence $\{\mathrm{P}_N\}_{N\in\mathbb{N}}$ of linear operators is not uniformly bounded. Therefore the uniform boundedness principle yields a divergence result similar to Theorem 5.

However, since there exists a uniformly bounded subsequence $\{P_{2^k}\}_{k\in\mathbb{N}}$, the question Q-2 has a positive answer for this operator sequence $\{P_N\}_{N\in\mathbb{N}}$.

7.4.2 *Strong Divergence of the Shannon Sampling Series*

The difficulties in answering Q-1, based on the Banach–Steinhaus theorem, may also be viewed as follows. Under Q-1, the approximation series $\{N_k(f)\}_{k\in\mathbb{N}}$ can be chosen subject to the actual function f, i.e., one is allowed to adapt the reconstruction method to the actual function. Therefore, the overall approximation procedure $\{S_{N_k(f)}f\}_{k\in\mathbb{Z}}$ depends nonlinearly on the function f. Hence, one essential requirement for applying the Banach–Steinhaus theorem (the linearity of the operators) is no longer satisfied. So even though the theorem of Banach–Steinhaus is a very powerful tool for proving (weak) divergence results as in Theorem 5, it cannot be used to answer question Q-1. Thus, for the investigation of adaptive recovery algorithms completely different techniques are needed.

We will show below, that for the Shannon sampling series S_N, question Q-1 has a negative answer. To this end, it is necessary and sufficient to show that the sequence $\{S_N\}_{N\in\mathbb{N}}$ diverges *strongly*.

Definition 3 (Strong divergence). Let $\mathscr{X}$ and $\mathscr{Y}$ be Banach spaces, and let $\{T_N\}_{N\in\mathbb{N}}$ be a sequence of bounded operators $T_N : \mathscr{X} \to \mathscr{Y}$. We say that T_N *diverges strongly* if

$$\lim_{N\to\infty} \|T_N f_1\|_{\mathscr{Y}} = \infty \quad \text{for some} \quad f_1 \in \mathscr{X} .$$

So the strong divergence is in contrast to the weaker statement of the lim sup divergence used in Theorem 5. As explained above, it is not possible to show the strong divergence of S_N using the Banach–Steinhaus theorem, and even though several extensions [23, 24, 61] of the Banach–Steinhaus theorem were developed in the past, there currently exists no systematic way to show strong divergence. For the Shannon sampling series (7.4) on $\mathscr{PW}_\pi^1$, its strong divergence is established by the following theorem.

Theorem 7. *The Shannon sampling series* $S_N : \mathscr{PW}_\pi^1 \to \mathscr{B}_\pi^\infty$ *given in* (7.4) *diverges strongly, i.e., there exists a function* $f_1 \in \mathscr{PW}_\pi^1$ *for which*

$$\lim_{N\to\infty} \left(\max_{t\in\mathbb{R}} |(S_N f_1)(t)| \right) = \lim_{N\to\infty} \|S_N f_1\|_\infty = \infty . \tag{7.15}$$

Clearly, this is a much stronger statement than Theorem 5, with important practical implications for adaptive signal processing algorithms. It rules out the possibility that the divergence in (7.9) occurs only because of a divergent subsequence. In particular, it implies a negative answer to question Q-1. Consequently, there exists

no (adaptive) signal recovery procedure which converges uniformly on the entire real axis $\mathbb{R}$.

The structure of the divergence sets. For nonadaptive linear methods, the Banach–Steinhaus theorem is a very powerful and well established tool in functional analysis to investigate nonadaptive approximation methods. In particular, if one can show that there exists one function $f \in \mathscr{X}$ such that the sequence $\mathrm{T}_N f$ diverges (weakly) in $\mathscr{Y}$, then the Banach–Steinhaus theorem immediately implies that there exists a whole set $\mathscr{D}_{\mathrm{weak}}$ of second category for which $\mathrm{T}_N f$ diverges weakly for every $f \in \mathscr{D}_{\mathrm{weak}}$.

We established in Theorem 7 that there exists one function f_1 such that the Shannon sampling series diverges strongly. The question is now whether it is possible to say something about the size or structure of the set of all functions for which S_N diverges strongly. Since the Banach–Steinhaus theorem cannot be applied in the case of strong divergence, other techniques have to be developed. Because of the close relation between strong divergence and adaptive signal processing methods, we believe that it is an important research topic to develop general tools for the investigation of strong divergence, similar to the Banach–Steinhaus technique for weak divergence, i.e., for nonadaptive systems.

Here we start with such an investigation and study the structure of the weak and strong divergence sets of approximation series. To this end, we consider linear approximation operators $\mathrm{T}_N : \mathscr{X} \to \mathscr{Y}$ mapping a Banach space $\mathscr{X}$ into a Banach space $\mathscr{Y}$. Since $\mathrm{T}_N f$ should be a good approximation of $f \in \mathscr{X}$, measured in the topology of $\mathscr{Y}$, it is natural to assume that $\mathscr{X} \subset \mathscr{Y}$. Additionally, we assume that there exists a dense subset $\mathscr{X}_0 \subset \mathscr{X}$ such that

$$\lim_{N\to\infty} \|\mathrm{T}_N f_0 - f_0\|_{\mathscr{Y}} = 0 \quad \text{for all } f_0 \in \mathscr{X}_0 \,, \tag{7.16}$$

i.e., such that $\mathrm{T}_N f_0$ converges to f_0 in the norm of $\mathscr{Y}$. For such linear operators, the next theorem studies the structure of the divergence sets

$$\begin{aligned} \mathscr{D}_{\mathrm{weak}} &= \left\{ f \in \mathscr{X} \;:\; \limsup_{N\to\infty} \|\mathrm{T}_N f\|_{\mathscr{Y}} = \infty \right\} \quad \text{and} \\ \mathscr{D}_{\mathrm{strong}} &= \left\{ f \in \mathscr{X} \;:\; \lim_{N\to\infty} \|\mathrm{T}_N f\|_{\mathscr{Y}} = \infty \right\} \end{aligned} \tag{7.17}$$

of functions for which weak and strong divergence emerges, respectively.

Theorem 8. *Let $\mathscr{X}$ and $\mathscr{Y}$ be two Banach spaces such that $\mathscr{X}$ is continuously embedded in $\mathscr{Y}$, and let $\mathrm{T}_N : \mathscr{X} \to \mathscr{Y}$ be a sequence of bounded linear operators such that (7.16) holds for a dense subset $\mathscr{X}_0 \subset \mathscr{X}$. For any $M, N \in \mathbb{N}$ define the set*

$$D(M,N) := \left\{ f \in \mathscr{X} \;:\; \|\mathrm{T}_N f\|_{\mathscr{Y}} > M \right\} .$$

1. *If $\mathscr{D}_{\mathrm{weak}}$ is nonempty, then for all $M, N_0 \in \mathbb{N}$ the set*

$$\bigcup_{N=N_0}^{\infty} D(M,N) \tag{7.18}$$

is open and dense in $\mathscr{X}$.

2. *For the divergence sets defined in* (7.17), *hold*

$$\mathscr{D}_{\text{weak}} = \bigcap_{M=1}^{\infty} \limsup_{N\to\infty} D(M,N) = \bigcap_{M=1}^{\infty} \bigcap_{N_0=1}^{\infty} \bigcup_{N=N_0}^{\infty} D(M,N) \tag{7.19}$$

$$\mathscr{D}_{\text{strong}} = \bigcap_{M=1}^{\infty} \liminf_{N\to\infty} D(M,N) = \bigcap_{M=1}^{\infty} \bigcup_{N_0=1}^{\infty} \bigcap_{N=N_0}^{\infty} D(M,N)\,. \tag{7.20}$$

Remark 5. For completeness and for later reference, the straightforward proofs of these statements are provided in the Appendix.

Remark 6. It is easy to see that the operators $\mathrm{S}_N : \mathscr{PW}_{\pi}^{1} \to \mathscr{B}_{\pi}^{\infty}$, associated with the Shannon sampling series and defined in (7.4), satisfy the requirements of Theorem 8. Indeed, Theorem 1 shows that $\mathscr{PW}_{\pi}^{2}$ is a dense subset of $\mathscr{PW}_{\pi}^{1}$ such that $\lim_{N\to\infty} \|\mathrm{S}_N f_0 - f_0\|_{\infty} = 0$ for all $f_0 \in \mathscr{PW}_{\pi}^{2}$, and Theorem 5 implies $\mathscr{D}_{\text{weak}} \neq \varnothing$.

At a first sight, the structure of both divergence sets seems to be fairly similar. The only difference is that the inner intersection and union in (7.19) and (7.20) are interchanged. However, the different order of these two operations has a distinct consequence. The first statement of Theorem 8 shows that sets (7.18) are open and dense subsets of $\mathscr{X}$, provided that $\mathscr{D}_{\text{weak}}$ is nonempty. Then Theorem 8 states that $\mathscr{D}_{\text{weak}}$ is a countable intersection of these sets, and Baire's category theorem implies that a countable intersection of open and dense subsets of a Banach space $\mathscr{X}$ is a set of second category, i.e., a nonmeager set. Consequently, if one is able to show that there exists one function in $\mathscr{D}_{\text{weak}}$, representation (7.19) together with the category theorem of Baire implies immediately that $\mathscr{D}_{\text{weak}}$ is nonmeager.

For $\mathscr{D}_{\text{strong}}$ the situation is completely different. There we have on the right-hand side the countable intersection of the open sets $D(M,N)$. However, the intersection of open sets is generally no longer open, and it may even be empty. So representation (7.20) gives only little information about the size of $\mathscr{D}_{\text{strong}}$. If there exists one function $f \in \mathscr{D}_{\text{strong}}$, then it is easy to show (using the same ideas as in part one of the proof of Theorem 8) that $\mathscr{D}_{\text{strong}}$ is dense in $\mathscr{X}$. Nevertheless, even if it is dense, it might be a set of first category, i.e., a meager set.

Finally, we shortly discuss the oscillatory behavior of the Shannon series [6]. This will give further insight into its divergence behavior.

Theorem 9. *Let* $f \in \mathscr{PW}_{\pi}^{1}$ *be a function for which the Shannon sampling series* (7.4) *diverges strongly. Then*

$$\lim_{N\to\infty} \left(\max_{t\in\mathbb{R}} (\mathrm{S}_N f)(t) \right) = +\infty \tag{7.21}$$

and

$$\lim_{N\to\infty}\left(\min_{t\in\mathbb{R}}(\mathrm{S}_N f)(t)\right) = -\infty\ .$$

This result not only implies the statement of Theorem 7 but it additionally shows the oscillatory behavior of the Shannon series and its unlimited growth as N tends to infinity. To the best of our knowledge, Theorem 9 is the only example which proves the strong oscillatory behavior of a practically relevant reconstruction method, and we will also see in Sec. 7.4.3 that nonuniform sampling series diverge strongly (cf. Theorem 11 below and the corresponding discussion).

We close this section with two examples which illustrate that there are many more problems where the question of strong divergence is of importance.

Example 3 (Lagrange interpolation on Chebyshev nodes). In 1941, *Paul Erdős* tried to show a behavior like (7.21) for the Lagrange interpolation on Chebyshev nodes. In [27], he claimed that a statement like (7.21) holds for the Lagrange interpolation of continuous functions. However, in [28] he observed that his proof was erroneous, and he was not able to present a correct proof. He presented a result equivalent to Theorem 7, and it seems that the original problem is still open.

Example 4 (Calculation of the Hilbert transform). For any function $f \in L^1([-\pi,\pi])$ the *Hilbert transform* H is defined by

$$(\mathrm{H}f)(t) = \lim_{\epsilon\to 0}\frac{1}{2\pi}\int_{\epsilon\le|\tau|\le\pi}\frac{f(t+\tau)}{\tan(\tau/2)}\,\mathrm{d}\tau\ . \tag{7.22}$$

This operation plays a very important role in different areas of communications, control theory, physics, and signal processing [35, 54, 65]. In practical applications, Hf has to be determined based on discrete samples $\{f(t_n)\}_{n=-N}^{N}$ of f. To this end, let $\{\mathrm{T}_N\}_{N\in\mathbb{N}}$, any sequence of linear operators which determines an approximation of Hf from the samples of $\{f(t_n)\}_{n=-N}^{N}$ of f. It was shown in [11] that for any such operator sequence, there exists a function $f_0 \in \mathscr{B} := \{f \in \mathscr{C}([-\pi,\pi]) : \mathrm{H}f \in \mathscr{C}([-\pi,\pi])\}$ such that

$$\limsup_{N\to\infty}\|\mathrm{T}_N f_0\|_\infty = \infty\ .$$

In other words, any (nonadaptive) linear method which determines the Hilbert transform from a discrete set of sampled values diverges weakly. Here, it would also be interesting to investigate whether the question Q-1 has positive answers or not, i.e., whether there exist adaptive methods to approximate the Hilbert transform from interpolated data.

7.4.3 Convergence for Oversampling

So far we saw that the Shannon sampling series diverges strongly on $\mathscr{PW}_\pi^1$. However, applying nonuniform sampling patterns and increasing the sampling rate induces additional degrees of freedom, which might give a better convergence behavior. The question is whether nonuniform sampling resolves the divergence problems observed for uniform sampling.

We start our investigations by showing that the result of Brown (Theorem 4) on the local uniform approximation behavior of the Shannon sampling series can be extended to nonuniform sampling series, provided that the sampling pattern is equal to the zero set of a sine-type function [9].

Theorem 10. *Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be the zero set of a sine-type function φ, let $\{\varphi_n\}_{n\in\mathbb{Z}}$ be the corresponding interpolation kernels, defined in (7.7), and let $\mathrm{A}_N : \mathscr{PW}_\pi^1 \to \mathscr{B}_\pi^\infty$ be defined as in (7.8). Then for every $T > 0$, one has*

$$\lim_{N\to\infty} \left(\max_{t\in[-T,T]} |f(t) - (\mathrm{A}_N f)(t)| \right) = 0 \qquad \text{for all } f \in \mathscr{PW}_\pi^1 \ .$$

So if sampling patterns derived from sine-type functions are used, then we do not need oversampling to obtain local convergence for all $f \in \mathscr{PW}_\pi^1$. It seems natural to ask whether we can apply other sampling patterns to achieve local convergence of $\mathrm{A}_N f$. However, we believe that Theorem 10 is sharp with respect to the chosen sampling pattern. If more general sampling patterns are used, the sampling series $\mathrm{A}_N f$ may no longer converge to f. More precisely, we believe that the following statement is true.

Conjecture 1. There exist a complete interpolating sequence $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ with generating function φ, corresponding interpolation kernels $\{\varphi_n\}_{n\in\mathbb{Z}}$, a point $t_* \in \mathbb{R}$, and a function $f_* \in \mathscr{PW}_\pi^1$ such that

$$\limsup_{N\to\infty} \left| (\mathrm{A}_N f_*)(t_*) \right| = \limsup_{N\to\infty} \left| \sum_{n=-N}^{N} f_*(\lambda_n)\, \varphi_n(t_*) \right| = +\infty \ .$$

Next, we ask whether the nonuniform sampling series $\mathrm{A}_N f$ even converges globally uniformly on $\mathbb{R}$. It turns out that the answer is negative. More precisely, one can even show that the nonuniform sampling series $\mathrm{A}_N : \mathscr{PW}_\pi^1 \to \mathscr{B}_\pi^\infty$, given in (7.8), *diverges strongly*. To prove this statement, [6] used nonuniform sampling patterns derived from a special type of sine-type function: For any $g \in \mathscr{PW}_\pi^1$, we define the function

$$\varphi(z) = A\, \sin(\pi z) - g(z) \ , \qquad z \in \mathbb{C} \tag{7.23}$$

with a constant $A > \|g\|_{\mathscr{PW}_\pi^1} \geq \|g\|_\infty$. This is a sine-type function and we say that φ is determined by the *sine wave crossings* of g. Such functions are used in

sampling theory and communications [4, 53] by methods which try to reconstruct the signal from its sine wave crossings.

Theorem 11. *Let φ be a sine-type function of the form (7.23) with zero set $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ and let $\{\varphi_n\}_{n\in\mathbb{Z}}$ be the corresponding interpolation kernels (7.7). Then the nonuniform sampling series (7.8) diverges strongly, i.e., there exists a function $f \in \mathscr{PW}^1_\pi$ such that*

$$\lim_{N\to\infty} \|\mathrm{A}_N f\|_\infty = \lim_{N\to\infty} \left(\max_{t\in\mathbb{R}} \left| \sum_{n=-N}^{N} f(\lambda_n)\, \varphi_n(t) \right| \right) = \infty \,.$$

So a nonuniform sampling series alone gives no improvement with respect to the global uniform convergence as compared to the uniform sampling considered in Theorem 7. Both the uniform and the nonuniform sampling series diverge strongly on $\mathscr{PW}^1_\pi$. Note that the previous theorem was formulated only for sampling patterns arising from the zero set of a sine-type function of the form (7.23). This is only a small subset of all complete interpolating sequences. Nevertheless, there is strong evidence that Theorem 11 also holds for arbitrary complete interpolating sequences. This gives the following conjecture [6].

Conjecture 2. Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be an arbitrary complete interpolating sequence with generator φ and corresponding interpolation kernels (7.7). Then there exists an $f \in \mathscr{PW}^1_\pi$ such that

$$\lim_{N\to\infty} \|\mathrm{A}_N f\|_\infty = \lim_{N\to\infty} \left(\max_{t\in\mathbb{R}} \left| \sum_{n=-N}^{N} f(\lambda_n)\, \varphi_n(t) \right| \right) = \infty \,.$$

Nonuniform sampling pattern alone does not resolve the divergence problem of the sampling series on $\mathscr{PW}^1_\pi$. Since $\mathscr{PW}^1_\pi \subset \mathscr{B}^\infty_\pi$, the same statement holds for $\mathscr{B}^\infty_\pi$. Next we want to investigate whether oversampling improves the global convergence behavior of the nonuniform sampling series (7.8). This is done for the largest signal spaces $\mathscr{B}^\infty_\pi$.

Our first theorem in this direction, taken from [48], shows that if oversampling is applied, then the result of Brown (Theorem 4) on the local uniform convergence on $\mathscr{PW}^1_\pi$ can be extended to the larger space $\mathscr{B}^\infty_\pi$ and to nonuniform sampling series of the form (7.8):

Theorem 12. *Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be the zero set of a sine-type function φ, and let $\{\varphi_n\}_{n\in\mathbb{Z}}$ be defined as in (7.7). Then for every $T > 0$ and any $0 < \beta < 1$, we have*

$$\lim_{N\to\infty} \max_{t\in[-T,T]} |f(t) - (\mathrm{A}_N f)(t)| = 0 \quad \textit{for all } f \in \mathscr{B}^\infty_{\beta\pi}$$

where A_N is defined in (7.8).

Remark 7. Note that this theorem allows sampling patterns from arbitrary sine-type functions. The oversampling is expressed by the fact that the above result holds only for functions in $\mathscr{B}^{\infty}_{\beta\pi}$ with $\beta < 1$, i.e., for functions with bandwidth $\beta\pi < \pi$.

Remark 8. If no oversampling is applied, i.e., if $\beta = 1$, then the above result is not true, in general. Then one can only show [48] that the approximation error remains locally bounded, i.e., $\sup_{N\in\mathbb{N}} \max_{t\in[-T,T]} |f(t)-(\mathrm{A}_N f)(t)| \le C\,\|f\|_\infty$ for all $f \in \mathscr{B}^{\infty}_{\pi}$.

The question is whether we can also have uniform convergence on the entire real axis. So what happens if we let T go to infinity in Theorem 12? The answer is given by the next theorem [48]. It shows that we only have uniform convergence on $\mathbb{R}$ for the subset $\mathscr{B}^{\infty}_{\beta\pi,0}$ of all $f \in \mathscr{B}^{\infty}_{\beta\pi}$ which vanish at infinity. However, in general, the approximation error remains uniformly bounded for every $f \in \mathscr{B}^{\infty}_{\beta\pi}$.

Theorem 13. *Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be the zero set of a sine-type function φ, and let A_N be defined as in (7.8) with interpolation kernels $\{\varphi_n\}_{n\in\mathbb{Z}}$ given in (7.7). Then for any $0 < \beta < 1$, we have*

$$\lim_{N\to\infty} \max_{t\in\mathbb{R}} |f(t) - (\mathrm{A}_N f)(t)| = 0 \qquad \text{for all } f \in \mathscr{B}^{\infty}_{\beta\pi,0}\,,$$

and there exists a constant $C > 0$ such that

$$\lim_{N\to\infty} \max_{t\in\mathbb{R}} |f(t) - (\mathrm{A}_N f)(t)| \le C\,\|f\|_\infty \qquad \text{for all } f \in \mathscr{B}^{\infty}_{\beta\pi}\,.$$

Since $\mathscr{PW}^1_\sigma \subset \mathscr{B}^{\infty}_{\sigma,0}$, Theorem 13 includes in particular the following corollary [9] on the global uniform convergence of A_N on $\mathscr{PW}^1_{\beta\pi}$.

Corollary 1. *Let φ be a sine-type function with zero set $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ and let $\{\varphi_n\}_{n\in\mathbb{Z}}$ be the corresponding interpolation kernels (7.7). Then for every $f \in \mathscr{PW}^1_{\beta\pi}$ with $0 < \beta < 1$ holds*

$$\lim_{N\to\infty} \|f - \mathrm{A}_N f\|_\infty = \lim_{N\to\infty} \left(\max_{t\in\mathbb{R}} \left| f(t) - \sum_{n=-N}^{N} f(\lambda_n)\,\varphi_n(t) \right| \right) = 0\,.$$

So with oversampling, i.e., for functions in $\mathscr{PW}^1_{\beta\pi}$ with $\beta < 1$, the sampling series (7.8) converges uniformly on the entire real axis. This result cannot be extended to the larger signal space $\mathscr{B}^{\infty}_{\beta\pi}$. For these functions, the approximation error is only bounded in general, but does not go to zero as N tends to infinity. This even holds for any arbitrary large oversampling factor $1/\beta$.

7.5 Sampling-Based Signal Processing

Up to now, our discussion was based on the goal to reconstruct a certain function, say $f \in \mathcal{PW}_\pi^1$, from its values $\{f(\lambda_n)\}_{n\in\mathbb{Z}}$ at the sampling points $\{\lambda_n\}_{n\in\mathbb{Z}}$. However, in applications one is often not interested in f itself, but in some processed version of f, i.e., one wants to determine $g = \mathrm{H}f$ where $\mathrm{H} : \mathcal{PW}_\pi^1 \to \mathcal{PW}_\pi^1$ is a certain linear system, for example the Hilbert transform (7.22) or the derivation operator $f(t) \mapsto \mathrm{d}f/\mathrm{d}t$ [18]. Since signals in the physical world are usually analog, the system H is often described in the analog domain. Then for a given function f, it would be easy determining $g = \mathrm{H}f$. However, if only samples $\{f(\lambda_n)\}_{n\in\mathbb{Z}}$ of f are available, then it seems to be more desirable to implement the system H directly in the digital domain, based on the signal samples $\{f(\lambda_n)\}_{n\in\mathbb{Z}}$. Thus, we look for a mapping $\mathrm{H_D} : \{f(\lambda_n)\} \mapsto \mathrm{H}f$ which determines $g = \mathrm{H}f$ directly from the available samples $\{f(\lambda_n)\}$. We call $\mathrm{H_D}$ the *digital implementation* of the analog system H.

Example 5 (Sensor Networks). In a sensor network, many sensors which are distributed nonuniformly in space (and time) measure (i.e., sample) a certain physical quantity (e.g., temperature, pressure, the electric or magnetic field strength, and velocity). For concreteness, assume that we measure temperature. Then the aim is not necessarily to reconstruct the entire temperature distribution in the observed area, but only to determine, say, the maximum temperature or the maximum temperature difference.

The interesting question now is whether it is possible to find for a given analog system H, a digital implementation $\mathrm{H_D}$. The answer depends strongly on the system H under consideration. Here we only investigate this problem for a fairly simple but very important class of mappings H, namely for stable linear, time-invariant systems $\mathrm{H} : \mathcal{PW}_\pi^1 \to \mathcal{PW}_\pi^1$.

7.5.1 Linear Time-Invariant Systems

In our context, a *linear system* is always a linear operator $\mathrm{H} : \mathcal{PW}_\pi^1 \to \mathcal{PW}_\pi^1$. Such a system is called *stable*, if H is bounded, i.e., if

$$\|\mathrm{H}\| = \sup\left\{\|\mathrm{H}f\|_{\mathcal{PW}_\pi^1} : f \in \mathcal{PW}_\pi^1,\ \|f\|_{\mathcal{PW}_\pi^1} \le 1\right\} < \infty ,$$

and H is said to be *time invariant* if it commutes with the *translation operator* $\mathrm{T}_a : f(t) \mapsto f(t-a)$, i.e., if $\mathrm{HT}_a f = \mathrm{T}_a \mathrm{H}f$ for all $a \in \mathbb{R}$ and for every $f \in \mathcal{PW}_\pi^1$.

It is known that for every stable, linear, time-invariant (LTI) system $\mathrm{H} : \mathcal{PW}_\pi^1 \to \mathcal{PW}_\pi^1$ there exists a unique function $\hat{h} \in L^\infty([-\pi,\pi])$ such that for all $f \in \mathcal{PW}_\pi^1$

$$(\mathrm{H}f)(t) = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{f}(\omega)\,\hat{h}(\omega)\,\mathrm{e}^{\mathrm{i}\omega t}\,\mathrm{d}\omega , \quad t \in \mathbb{R} \tag{7.24}$$

and with $\|H\| = \|\hat{h}\|_\infty$. Conversely, every function $\hat{h} \in L^\infty([-\pi,\pi])$, defines via (7.24) a stable LTI system H. In engineering, the function $\hat{h}$ is often called the *transfer function* of the LTI system H, whereas its inverse Fourier transform $h = \mathscr{F}^{-1}\hat{h}$ is said to be the *impulse response* of H. Since $L^\infty([-\pi,\pi]) \subset L^2([-\pi,\pi])$, we have that $h \in \mathscr{PW}^2_\pi$.

Digital implementation for $\mathscr{PW}^2_\pi$. Now we want to find a digital implementation $\mathrm{H_D}$ for a stable LTI system H. To this end we first consider the situation on the Hilbert space $\mathscr{PW}^2_\pi$. There the obvious way to define $\mathrm{H_D}$ is by applying H to $\mathrm{A}_N f$. This yields

$$(\mathrm{HA}_N f)(t) = \sum_{n=-N}^{N} f(\lambda_n)\,(\mathrm{H}\varphi_n)(t) = \sum_{n=-N}^{N} f(\lambda_n)\,\psi_n(t) =: (\mathrm{H}_N f)(t) \tag{7.25}$$

with kernels $\psi_n := \mathrm{H}\varphi_n \in \mathscr{PW}^2_\pi$, for all $n \in \mathbb{Z}$, and where the sampling set Λ is chosen to be a complete interpolating for $\mathscr{PW}^2_\pi$. If H is a stable system $\mathscr{PW}^2_\pi \to \mathscr{PW}^2_\pi$, then it follows from Theorem 2 that

$$\|\mathrm{H}f - \mathrm{H}_N f\|_{\mathscr{PW}^2_\pi} = \|\mathrm{H}f - \mathrm{HA}_N f\|_{\mathscr{PW}^2_\pi} \le \|\mathrm{H}\|\,\|f - \mathrm{A}_N f\|_{\mathscr{PW}^2_\pi} \to 0$$

as $N \to \infty$ for every $f \in \mathscr{PW}^2_\pi$. Since $\mathscr{PW}^2_\pi$ is a reproducing kernel Hilbert space, the norm convergence again implies the uniform convergence on $\mathbb{R}$.

Digital implementation in $\mathscr{PW}^1_\pi$. Since $\mathscr{PW}^2_\pi$ is a dense subset of $\mathscr{PW}^1_\pi$, we may hope that the implementation for $\mathscr{PW}^2_\pi$ extends in some sense to $\mathscr{PW}^1_\pi$.

Let us first consider a very special stable LTI system, namely the identity operator $\mathrm{H} = \mathrm{I}_{\mathscr{PW}^1_\pi}$ on $\mathscr{PW}^1_\pi$. For this particular system, its digital implementation is easily derived. Following the above ideas for $\mathscr{PW}^2_\pi$, its digital implementation is simply $\mathrm{H}_N = \mathrm{HA}_N = \mathrm{A}_N$, and Corollary 1 implies

$$\lim_{N\to\infty} \|\mathrm{H}f - \mathrm{H}_N f\|_\infty = \lim_{N\to\infty} \|\mathrm{H}f - \mathrm{A}_N f\|_\infty = \lim_{N\to\infty} \max_{t\in\mathbb{R}} \left| \mathrm{H}f - \sum_{n=-N}^{N} f(\lambda_n)\,\varphi_n(t) \right| = 0$$

for all $f \in \mathscr{PW}^1_{\beta\pi}$ with $\beta < 1$, and provided that the sampling set $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ was chosen to be the zero set of a sine-type function. So for the identity operator, we found a digital implementation. Since Corollary 1 was used in the above arguments, it is clear that this digital implementation is based on the oversampled input signal f. Theorem 11 shows then that if Λ is the zero set of sine-type functions of the form (7.23), oversampling is indeed necessary even for the global approximation of the simple LTI system $\mathrm{H} = \mathrm{I}_{\mathscr{PW}^1_\pi}$. Moreover, if Conjecture 2 turns out to be true, then it would imply that oversampling is necessary for all complete interpolating sequences Λ.

Our next question is whether the digital implementation (7.25) converges for any arbitrary stable LTI system H on $\mathscr{PW}^1_{\beta\pi}$, i.e., whether $\mathrm{H}_N f \to \mathrm{H}f$ as $N \to \infty$ for all

$f \in \mathcal{PW}^1_{\beta\pi}$. In particular, we investigate whether $\mathrm{H}_N f$ converges locally uniformly to $\mathrm{H}f$ or even globally uniformly as the identity operator.

7.5.2 Sampling via Point Evaluations

So the digital implementation H_N of any stable LTI system $\mathrm{H} : \mathcal{PW}^1_\pi \to \mathcal{PW}^1_\pi$ is defined as in (7.25), based on a complete interpolating sequence $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ for $\mathcal{PW}^2_\pi$ and based on the interpolation kernels $\{\varphi_n\}_{n\in\mathbb{Z}}$ given in (7.7) with φ as in (7.6).

The next theorem taken from [10] shows that there exist stable LTI systems for which such a digital implementation is not possible, even if we allow arbitrarily large oversampling. More precisely, it shows that there exist stable LTI systems such that the approximation of its digital implementation $\mathrm{H}_N f$ diverges even pointwise for some $f \in \mathcal{PW}^1_\pi$.

Theorem 14. *Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be a complete interpolating sequence for $\mathcal{PW}^2_\pi$ and let $\{\varphi_n\}_{n\in\mathbb{Z}}$ be the interpolation kernels defined in (7.7). Let $t \in \mathbb{R}$ be arbitrary, then there exists a stable LTI system $\mathrm{H} : \mathcal{PW}^1_\pi \to \mathcal{PW}^1_\pi$ such that for every $0 < \beta < 1$ there exists a signal $f \in \mathcal{PW}^1_{\beta\pi}$ such that*

$$\limsup_{N\to\infty} |(\mathrm{H}_N f)(t)| = \limsup_{N\to\infty} \left| \sum_{n=-n}^{N} f(\lambda_n)\, (\mathrm{H}\varphi_n)(t) \right| = \infty \,.$$

So for any fixed $t \in \mathbb{R}$ there are stable LTI systems $\mathrm{H} : \mathcal{PW}^1_\pi \to \mathcal{PW}^1_\pi$ such that for every $\beta \in (0, 1]$ the corresponding digital approximation H_N diverges at t for some signals $f \in \mathcal{PW}^1_{\beta\pi}$. But on the other side, since H is stable, we have for any $t \in \mathbb{R}$ and any $f \in \mathcal{PW}^1_\pi$

$$|(\mathrm{H}f)(t)| \le \|\mathrm{H}f\|_\infty \le \|\mathrm{H}f\|_{\mathcal{PW}^1_\pi} \le \|\mathrm{H}\|\, \|f\|_{\mathcal{PW}^1_\pi} < \infty \,.$$

So the divergence observed in Theorem 14 is indeed a property of the digital implementation of H based on (time domain) samples of f and not a property of the system H itself. Moreover, if it were possible to sample f in the frequency domain, then we could approximate the integral in its analog implementation (7.24) by its Riemann sum. This sum would converge to $(\mathrm{H}f)(t)$ for every $t \in \mathbb{R}$.

Overall, we see that there exists no general answer to the question whether every LTI system $\mathrm{H} : \mathcal{PW}^1_\pi \to \mathcal{PW}^1_\pi$ can be implemented digitally. Of course there are stable LTI systems which allow such digital implementation. The identity operator discussed above is one example of such a system. However, Theorem 14 shows that there exist stable systems for which such a digital implementation is not possible.

We come back to the discussion at the end of Section 7.4.1 and ask whether question Q-1 or Q-2 may have a positive answer for the approximation operators H_N, i.e., whether there exist subsequences $\{N_k\}_{k\in\mathbb{N}}$ (dependent on the function f, or

not) such that $\{\mathrm{H}_{N_k} f\}_{k\in\mathbb{N}}$ converges to $\mathrm{H}f$. To this end, let H be a stable LTI system and let $t \in \mathbb{R}$ be a fixed point. Then $(\mathrm{H}_N f)(t)$ defines a sequence of linear functionals on $\mathscr{PW}^1_{\beta\pi}$, for every $\beta \in (0, 1]$, with the norm

$$\|\mathrm{H}_N\|_{t,\beta} = \sup\left\{|(\mathrm{H}_N f)(t)| : f \in \mathscr{PW}^1_{\beta\pi}\ ,\ \|f\|_{\mathscr{PW}^1_{\beta\pi}} \le 1\right\}\ .$$

Then Theorem 14 implies that for every $t \in \mathbb{R}$ there exists a stable LTI system H such that for all $\beta \in (0, 1]$

$$\limsup_{N\to\infty} \|\mathrm{H}_N\|_{t,\beta} = +\infty\ .$$

However, since we have no statement for $\liminf_{N\to\infty} \|\mathrm{H}_N\|_{t,\beta}$, we do not know at the moment whether $\|\mathrm{H}_N\|_{t,\beta}$ satisfies an inequality similar to (7.13) for some $t \in \mathbb{R}$. If a lower bound like (7.13) were to exist, then question Q-2 would have a negative answer, i.e., no subsequence $\{N_k\}_{k\in\mathbb{N}}$ would exist such that $H_{N_k} f$ converges globally uniformly to $\mathrm{H}f$ for all $f \in \mathscr{PW}^1_{\pi}$. Indeed, we believe that the following statement is true, which would imply a negative answer to Q-2 (see discussion in Section 7.4.1).

Conjecture 3. Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be a complete interpolating sequence for $\mathscr{PW}^2_{\pi}$, and let $t \in \mathbb{R}$ be arbitrary. Then there exists a stable LTI system $\mathrm{H} : \mathscr{PW}^1_{\pi} \to \mathscr{PW}^1_{\pi}$ such that for every $\beta \in (0, 1]$

$$\lim_{N\to\infty} \|\mathrm{H}_N\|_{t,\beta} = +\infty\ .$$

It would also be interesting to investigate question Q-1, i.e., to ask whether the sequence $\{\mathrm{H}_N : \mathscr{PW}^1_{\beta\pi} \to \mathscr{B}^\infty_{\pi}\}$ diverges strongly. We believe that this is indeed the case, i.e., we think that the following conjecture is true.

Conjecture 4. Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be a complete interpolating sequence for $\mathscr{PW}^2_{\pi}$. There exists a stable LTI system $\mathrm{H} : \mathscr{PW}^1_{\pi} \to \mathscr{PW}^1_{\pi}$ such that for every $\beta \in (0, 1]$ there exists an $f_\beta \in \mathscr{PW}^1_{\beta\pi}$ for which

$$\lim_{N\to\infty} \left\|\mathrm{H}_N f_\beta\right\|_\infty = \lim_{N\to\infty} \left(\max_{t\in\mathbb{R}} \left|(\mathrm{H}_N f_\beta)(t)\right|\right) = +\infty\ .$$

Remark 9. Note that the LTI system $\mathrm{H} : \mathscr{PW}^1_{\pi} \to \mathscr{PW}^1_{\pi}$ for which the approximation H_N diverges is universal with respect to β. In other words, we believe that it is not possible to find a digital implementation of H, regardless of the amount of oversampling.

If this conjecture is true, it would exclude the existence of an adaptive algorithm which chooses the approximation sequence $\{N_k(f)\}_{k\in\mathbb{N}}$ subject to the actual function f to approximate the output $\mathrm{H}f$ of the system H from the signal samples $\{f(\lambda_n)\}_{n\in\mathbb{Z}}$.

7.5.3 Sampling by Generalized Measurement

The fundamental concept of digital signal processing is to represent analog (i.e., continuous) signals as a sequence of numbers. In the previous discussions it was always assumed that the conversion from the analog to the digital domain is based on point evaluations of the analog signal. Thus the measurement functionals were assumed to be of the form

$$\gamma_n : f \mapsto f(\lambda_n) \,, \qquad n \in \mathbb{Z} \tag{7.26}$$

with a certain sequence $\{\lambda_n\}_{n\in\mathbb{Z}}$ of sampling points. However, more general measurement methods are possible, which we want to investigate next.

Although we depart from the point evaluations (7.26), we still require that our measurements are based on bounded linear functionals on the specific function space. Again, we first consider the situation on the Hilbert space $\mathscr{PW}_\pi^2$. By the Riesz representation theorem, we know that any bounded linear functional $\gamma_n : \mathscr{PW}_\pi^2 \to \mathbb{C}$ can be written as an inner product with a certain function $s_n \in \mathscr{PW}_\pi^2$, i.e.,

$$\gamma_n(f) = \langle f, s_n \rangle_{\mathscr{PW}_\pi^2} = \int_{-\infty}^{\infty} f(t)\,\overline{s_n(t)}\,\mathrm{d}t = \frac{1}{2\pi}\int_{-\pi}^{\pi} \hat{f}(\omega)\,\overline{\hat{s}_n(\omega)}\,\mathrm{d}\omega \,, \tag{7.27}$$

where the last equation follows from Parseval's formula, and Cauchy–Schwarz inequality gives immediately $\|\gamma_n\| = \|s_n\|_{\mathscr{PW}_\pi^2}$. In this respect, any generalized sampling process on $\mathscr{PW}_\pi^2$ is based on a sequence $\{s_n\}_{n\in\mathbb{N}}$ of sampling functions in $\mathscr{PW}_\pi^2$, which defines via (7.27) a sequence $\{\gamma_n\}_{n\in\mathbb{N}}$ of measurement functionals. A stable reconstruction of any $f \in \mathscr{PW}_\pi^2$ from the samples $\{\gamma_n(f)\}_{n\in\mathbb{N}}$ is possible if $\{s_n\}_{n\in\mathbb{N}}$ is at least a *frame* [22, 70] for $\mathscr{PW}_\pi^2$. Let $\{\sigma_n\}_{n\in\mathbb{N}}$ be the dual frame of $\{s_n\}_{n\in\mathbb{N}}$, then f can be reconstructed from its samples $\{\gamma_n(f)\}_{n\in\mathbb{N}}$ by

$$f(t) = \lim_{N\to\infty} (\mathrm{A}_N f)(t) \qquad \text{where} \qquad (\mathrm{A}_N f)(t) = \sum_{n=1}^{N} \gamma_n(f)\,\sigma_n(t)$$

and where the sum converges in the norm of $\mathscr{PW}_\pi^2$ and uniformly on $\mathbb{R}$. If $\{s_n\}_{n\in\mathbb{Z}}$ is even an orthonormal basis for $\mathscr{PW}_\pi^2$, then we simply have $\sigma_n = s_n$ for all $n \in \mathbb{Z}$.

Example 6. The point evaluations (7.26) can be written as in (7.27) by choosing s_n to be equal to the reproducing kernels r_{λ_n} of $\mathscr{PW}_\pi^2$. Moreover, it is known [70] that $\{s_n\}_{n\in\mathbb{Z}}$ is a Riesz basis for $\mathscr{PW}_\pi^2$ if and only if $\{\lambda_n\}_{n\in\mathbb{Z}}$ is complete interpolating for $\mathscr{PW}_\pi^2$. Note that the measurement functionals associated with the point evaluations are uniformly bounded, because $\|\gamma_n\| = \|r_{\lambda_n}\|_{\mathscr{PW}_\pi^2} = 1$ for all $n \in \mathbb{Z}$.

Now we apply again a stable LTI system H to the approximation operator A_N. This gives an approximation of the digital implementation $\mathrm{H_D}$ of H

$$(\mathrm{H}_N f)(t) := (\mathrm{H}\mathrm{A}_N f)(t) = \sum_{n=1}^{N} \gamma_n(f)\,(\mathrm{H}\sigma_n)(t)\ . \tag{7.28}$$

If H is a stable LTI system $\mathscr{PW}^2_\pi \to \mathscr{PW}^2_\pi$, then it is again easy to see that $\mathrm{H}_N f \to \mathrm{H}f$ as $N \to \infty$ in the norm of $\mathscr{PW}^2_\pi$ and uniformly on $\mathbb{R}$ for every $f \in \mathscr{PW}^2_\pi$.

Now we consider the approximation operator (7.28) on $\mathscr{PW}^1_\pi$. To this end, $\{\gamma_n\}_{n\in\mathbb{N}}$ has to be a sequence of bounded linear functionals on $\mathscr{PW}^1_\pi$. It is known that every bounded linear functional $\gamma_n : \mathscr{PW}^1_\pi \to \mathbb{C}$ has the form (7.27) but with a function $\hat{s}_n \in L^\infty([-\pi,\pi])$ and such that $\|\gamma_n\| = \|\hat{s}_n\|_\infty$. As in the case of point evaluations on $\mathscr{PW}^2_\pi$, we require that all measurement functionals are uniformly bounded, i.e., we require that there exists a positive constant C_γ such that

$$\|\gamma_n\| = \|\hat{s}_n\|_\infty \le C_\gamma \quad \textit{for all } n \in \mathbb{N}\ . \tag{7.29}$$

The question is whether we can find a frame $\{s_n\}_{n\in\mathbb{N}}$ for $\mathscr{PW}^2_\pi$ such that the series (7.28) converges to $\mathrm{H}f$ for any stable LTI system $\mathrm{H} : \mathscr{PW}^1_\pi \to \mathscr{PW}^1_\pi$ and for every $f \in \mathscr{PW}^1_\pi$. The answer is affirmative, provided oversampling is applied. Moreover, appropriate measurement functionals $\{\gamma_n\}_{n\in\mathbb{N}}$ are generated by an orthonormal sequence $\{s_n\}_{n\in\mathbb{N}}$ in $\mathscr{PW}^2_\pi$. More precisely, the following statement can be proved [10].

Theorem 15. *Let $0 < \beta < 1$ be arbitrary. There exists an orthonormal basis $\{s_n\}_{n\in\mathbb{N}}$ for $\mathscr{PW}^2_\pi$ with the associated measurement functionals (7.27) which satisfy (7.29) such that for all stable LTI systems $\mathrm{H} : \mathscr{PW}^1_\pi \to \mathscr{PW}^1_\pi$ and for all $f \in \mathscr{PW}^1_{\beta\pi}$*

$$\lim_{N\to\infty} \|\mathrm{H}f - \mathrm{H}_N f\|_\infty = \lim_{N\to\infty} \left(\sup_{t\in\mathbb{R}} \left| (\mathrm{H}f)(t) - \sum_{n=1}^{N} \gamma_n(f)(\mathrm{H}s_n)(t) \right| \right) = 0\ . \tag{7.30}$$

Moreover, there exists a constant C_s such that

$$\|\mathrm{H}_N f\|_\infty = \sup_{t\in\mathbb{R}} \left| \sum_{n=1}^{N} \gamma_n(f)\,(\mathrm{H}s_n)(t) \right| \le C_s\,\|\mathrm{H}\|\,\|f\|_{\mathscr{PW}^1_\pi} \quad \textit{for all } f \in \mathscr{PW}^1_{\beta\pi}\ .$$

This theorem shows that there exists a set $\{\gamma_n\}_{n\in\mathbb{Z}}$ of generalized measurement functionals such that, in connection with oversampling, every stable LTI system on $\mathscr{PW}^1_\pi$ possesses a digital implementation. The measurement functionals γ_n are defined via (7.27) by a specific orthonormal basis $\{s_n\}_{n\in\mathbb{Z}}$ for $\mathscr{PW}^2_\pi$. It should be noted that this orthonormal basis depends on β, i.e., on the amount of oversampling. Theorem 14 shows that $\{s_n\}$ cannot be a sequence of reproducing kernels because this would yield point evaluations as measurement functionals. The proof of Theorem 15 in [10] is constructive in the sense that it provides an explicit construction of an orthonormal basis $\{s_n\}_{n\in\mathbb{N}}$ such that (7.30) holds. This construction is based on the *Olevskii system* [50] which is an orthonormal basis for $\mathscr{C}([0,1])$.

Thus for the signal space $\mathscr{PW}_{\pi}^{1}$, a digital implementation of a stable LTI system can only be guaranteed if generalized measurement functionals with oversampling are used. If the data acquisition is based on simple point evaluations, then there are stable LTI systems which possess no digital implementation. This demonstrates in particular the limitation of digital implementations based on measurements from sensor networks, because the sampling process in such a network is basically a point evaluation at the particular sensor position.

7.6 Signal Recovery from Amplitude Samples

Sampling theory as discussed in the previous sections is based on signal samples taken by a sequence of linear functionals $\{\gamma_n\}_{n\in\mathbb{Z}}$. Then the reconstruction method is linear, namely a simple interpolation series of the form (7.8) or (7.10). If the measurement functionals are nonlinear then signal recovery will become more involved and in particular nonlinear, in general.

This section discusses a particular case of nonlinear measurements which is of considerable interest in many applications. Assume again that $\{\gamma_n\}$ is a set of linear functionals on our signal space and $\{\gamma_n(f)\}_{n\in\mathbb{Z}}$ is the sequence of complex-valued samples of a signal f. In many different applications it is not possible to measure the magnitude and the phase of $\gamma_n(f)$, but only the squared modulus $|\gamma_n(f)|^2$. In this case, the sampling operator $f \mapsto \{|\gamma_n(f)|^2\}_{n\in\mathbb{Z}}$ is nonlinear. However, the nonlinearity arises only due to the intensity measurement $|\cdot|^2$ but it is often possible to design the linear functionals $\{\gamma_n\}_{n\in\mathbb{Z}}$ by an appropriate measurement setup. The interesting question is now whether f can be reconstructed from the intensity samples $\{|\gamma_n(f)|^2\}_{n\in\mathbb{Z}}$ and how we have to choose the functionals $\{\gamma_n\}_{n\in\mathbb{Z}}$ such that signal recovery becomes possible.

The described problem, also known as *phase retrieval*, arises especially in optics, where, because of the short wavelength, only the intensity of the electromagnetic wave can be measured, but not its actual phase. Applications where such problems appear range from X-ray crystallography [45, 46], astronomical imaging [32], radar, [40], speech processing [1] to quantum tomography [34], to mention only some.

Phase retrieval for signals from finite-dimensional spaces $\mathbb{C}^N$ were considered extensively in the last years. Now there exists necessary and sufficient conditions on the number of samples as well as different recovery algorithms ranging from algebraic methods to algorithms based on convex optimization [1, 2, 12, 20, 21, 55]. For infinite-dimensional signal spaces, only a few results exist up to now. Nevertheless, it seems natural to ask whether it is possible to obtain results for bandlimited signals which are similar to the sampling series considered in the previous section, but which are based on the sampled amplitude only.

Since only the amplitudes of the signal samples are available, some oversampling has to be used to compensate this information loss. However, several questions arise: How much oversampling is necessary and what are sufficient conditions on

the measurement functionals $\{\gamma_n\}$ such that signal recovery can be guaranteed? In particular, can we recover every signal from the amplitudes of point evaluations $\gamma_n(f) = f(\lambda_n)$ or do we need generalized measurement functionals (as discussed in Sec. 7.5.3)?

Before we start our discussion, we want to mention that in the considered situation, signal recovery will only be possible up to an unknown global phase factor. To see this, let $\{\gamma_n\}_{n\in\mathbb{Z}}$ be the set of linear measurement functionals. Then it is not possible to recover f perfectly from the intensity measurements $\{|\gamma_n(f)|^2\}$, but only up to a unitary constant. Because if $\tilde{f}(z) = cf(z)$, where c is a unitary constant, then both functions f and $\tilde{f}$ will give the same measurements, i.e., $|\gamma_n(\tilde{f})|^2 = |\gamma_n(f)|^2$ for all n. Consequently, we consider here only signal recovery up to a global unitary constant, which is sufficient in most applications.

Real-valued bandlimited functions For real-valued bandlimited signals there exists a remarkable result [62] in the spirit of classical Shannon sampling theory. It shows that any signal in the Bernstein spaces $\mathscr{B}_\pi^p$ with $0 < p \leq \infty$ can be reconstructed from amplitude samples taken at an average rate of at least twice the Nyquist rate. The result in [62] implies in particular the following statement.

Theorem 16. *Let $\Lambda = \{\lambda_n\}_{n\in\mathbb{Z}}$ be the zero set of a sine-type function φ, and for $1 \leq p \leq \infty$ let $f \in \mathscr{B}_{\pi/2}^p$ be real valued on $\mathbb{R}$. Then f can uniquely be determined from the samples $c_n = |\gamma_n(f)|^2 = |f(\lambda_n)|^2$, $n \in \mathbb{Z}$, up to a sign factor.*

The proof of Theorem 16 in [62] also provides a reconstruction algorithm. It is noteworthy that in the case of real-valued functions, point evaluations are sufficient as measurement functionals γ_n. Unfortunately, the technique used to prove Theorem 16 in [62] cannot be easily extended to the complex-valued functions.

Complex-valued bandlimited functions In the complex case, simple point evaluation does not seem to be sufficient, but rather very specific measurement functionals have to be chosen. In [69], measurement functionals $\{\gamma_n\}$ for phase retrieval in $\mathscr{B}_\pi^2$ were proposed, which consist of linear combinations of point evaluations at specific sampling points. This approach was later extended to all Bernstein spaces $\mathscr{B}_\pi^p$ with $1 < p < \infty$ in [56]. More precisely, let $f \in \mathscr{B}_\pi^p$ and let $K \geq 2$ be an arbitrary integer, then the measurement functionals proposed in [69] are given by

$$\gamma_{n,m}(f) = \sum_{k=1}^{K} \overline{\alpha_{k,m}} f(n\beta + \lambda_k) \,, \qquad n \in \mathbb{Z} \,,\ m = 1, 2, \ldots, K^2 \tag{7.31}$$

where the constant $\beta > 0$, the complex numbers $\{\lambda_k\}_{k=1}^K$, and the complex coefficients $\{\alpha_{k,m}\}$ are chosen in a very specific way. To formulate sufficient conditions on the functionals (7.31) such that signal recovery is possible, we define the $\mathbb{C}^K$ vectors

$$a_m = (\alpha_{1,m}, \dots, \alpha_{K,M})^{\mathrm{T}} , \qquad m = 1, \dots, K^2 .$$

Therewith the requirements on the functionals (7.31) can be formulated as follows:

Definition 4 (Recovery condition). We say that the measurement functionals (7.31) satisfy the *recovery condition*, if

1) $\lambda_K = \lambda_1 + \beta$
2) $\Lambda = \{n\beta + \lambda_k : n \in \mathbb{Z}\, , k = 1, \dots, K-1\}$ is the zero set of a sine-type function
3) $\{a_m\}_{m=1}^{K^2}$ forms a 2-uniform K^2/K tight frame for $\mathbb{C}^K$.

Remark 10. The particular form of the measurement functionals arises from a concrete measurement setup for a particular phase retrieval problem. Therefore, there exists a fairly simple practical implementation of these functionals. We also remark that the conditions on the functional can be slightly weakened, c.f. [56, 69].

Remark 11. It is fairly easy to find coefficients for the measurement functionals (7.31) such that the recovery condition is satisfied. In particular, constructions for 2-uniform tight frames can be found in [72]. To get appropriate λ_k one can choose $\Lambda = \{\lambda_n : n \in \mathbb{Z}\}$ as the zero set of the sine-type function $\varphi(z) = sin(\pi z)$. Then $\lambda_n = n$, $n = 1, \dots, K$ and $\beta = K - 1$.

Theorem 17. *For any $K \geq 2$ let $\{\gamma_{n,m}\}$ be the measurement functionals given in (7.31) such that they satisfy the recovery condition. Let $1 < p < \infty$ and set*

$$\mathscr{R}_\pi^p := \{f \in \mathscr{B}_\pi^p : f(n\beta + \lambda_1) \neq 0 \, for \;\; all \, n \in \mathbb{Z}\} .$$

Then every $f \in \mathscr{R}_\pi^p$ can be recovered from the amplitude measurements

$$c_{n,m} = |\gamma_{n,m}(f)|^2 , \qquad n \in \mathbb{Z}\, , \; m = 1, \dots, K^2$$

up to a global unitary factor.

The recovery procedure which belongs to Theorem 17 consists basically of a two-step procedure.

1. In the first step, one determines all values $f(n\beta + \lambda_k)$ from the amplitude measurements $|\gamma_{n,m}(f)|^2$ using ideas and algorithms from finite-dimensional phase retrieval.
2. Since $\Lambda = \{n\beta + \lambda_k : n \in \mathbb{Z}\, , k = 1, \dots, K-1\}$ is the zero set of a sine-type function, we can use the sampling series discussed in Section 7.4 to recover f from its samples at the sampling set Λ.

The first step in this recovery procedure relies only on an appropriate choice of the coefficients $\{\alpha_{k,m}\}$, whereas the second step only relies on an appropriate choice of the sampling set Λ. For this reason, it is also easy to extend Theorem 17 to other signal spaces. For example, applying Corollary 1 in the second step of the recovery procedure, we immediately obtain a phase retrieval result for functions in $\mathscr{PW}_\pi^1$.

Corollary 2. *For any $K \geq 2$ let $\{\gamma_{n,m}\}$ be the measurement functionals given in (7.31) such that they satisfy the recovery condition. Let $0 < \beta < 1$ and set*

$$\mathscr{P}_\pi^1 := \{f \in \mathscr{PW}_\pi^1 : f(n\beta + \lambda_1) \neq 0 \text{ for all } n \in \mathbb{Z}\} .$$

Then every $f \in \mathscr{P}_\pi^1$ can be recovered from the amplitude measurements $c_{n,m} = |\gamma_{n,m}(f)|^2$, $n \in \mathbb{Z}$, $m = 1, \ldots, K^2$ up to a global unitary factor.

The overall sampling rate in Theorem 17 is mainly determined by the constant K, which can be an arbitrary natural number $K \geq 2$. We see from (7.31) that we apply K^2 measurement functional $\gamma_{n,m}$ in every interval of length β, i.e. the overall sampling rate becomes $R = K^2/\beta$, and β has to be chosen such that the sequence $\{n\beta + \lambda_k : n \in \mathbb{Z}, k = 1, \ldots, K-1\}$ is the zero set of a sine-type function. This implies (see also Remark 11) that $\beta \leq K - 1$. Therefore, the overall sampling rate has to be at least $R \geq K^2/(K-1) \geq 4$, where $R = 4$ is achieved for $K = 2$. So we have found a sufficient condition on the sampling rate.

In Theorem 17, functions $f \in \mathscr{B}_\pi^p$ which have a zero in the set $\{n\beta + \lambda_1 : n \in \mathbb{Z}\}$ cannot be recovered. However, on the one hand, it is not hard to see that the set of these functions is fairly small, namely it is a set of first category. On the other hand, it was also shown in [56] that this restriction on the recoverable functions can be avoided if the desired signal f is preprocessed in a specific way, namely by adding a known sine-type function u prior to the amplitude measurements.

Theorem 18. *Let $A_{\max} > 0$ be arbitrary, and let $0 < \beta < \beta_1 < 1$. For any $1 \leq p \leq \infty$ set*

$$\mathscr{S}^p := \{f \in \mathscr{B}_{\beta\pi}^p : \|f\|_{\mathscr{B}_{\beta\pi}^p} \leq A_{\max}\} .$$

Then there exists a sine-type function $u \in \mathscr{B}_{\beta_1\pi}^\infty$ and a sequence of measurement functionals of the form (7.31) which satisfy the recovery condition such that every $f \in \mathscr{S}^p$ can be recovered from the amplitude measurements

$$c_{n,m} = |\gamma_{n,m}(f+u)|^2 , \qquad n \in \mathbb{Z} , \; m = 1, \ldots, K^2$$

up to a global unitary factor.

Remark 12. The restriction on the norm in the definition of the signal space $\mathscr{S}^p$ requires only that we need to know an upper bound on the signal norm. By the Theorem of Plancharel-Pólya (7.3), this is equivalent to a restriction on the peak value of the signal. This knowledge is necessary to choose an appropriated function u.

The proof of Theorem 18 is based on the following two facts:

1. Let $\Lambda = \{\lambda_n = \xi_n + i\eta_n\}_{n\in\mathbb{Z}}$ be the zero set of an arbitrary sine-type function. If one changes the imaginary parts of every λ_n, the resulting sequence is again the zero set of a sine-type function [44].

2. Fix $1 \leq p \leq \infty$ and $\beta < \pi$. Then to every $H_u > 0$ there exists a sine-type function u such that for every $f \in \mathscr{B}^p_{\beta\pi}$ with $\|f\|_{\mathscr{B}^p_{\beta\pi}} \leq A_{\max}$, the function $v = f + u$ satisfies $|v(\xi + \mathrm{i}\eta)| > 0$ for all $\xi \in \mathbb{R}$ and all $|\eta| > H_u$ [56]. So all zeros of $v = f + u \in \mathscr{B}^\infty_\pi$ are concentrated in a strip parallel to the real axis.

So based on these two observations, we can choose $\Lambda = \{\lambda_n = \xi_n + \mathrm{i}\eta_n\}$ such that the measurement functionals satisfy the recovery condition. Then we choose an arbitrary $H_u > 0$ and increase (if necessary) all η_n such that $|\eta_n| > H_u$ for all $n \in \mathbb{Z}$. The resulting sequence will still satisfy the recovery condition. Then we choose u such that the zeros of all function $v = f + u$ with $f \in \mathscr{B}^p_{\beta\pi}$ lie close to the real axis. In this way, we can achieve that the functions $v = f + u$ will definitely have no zero on the set Λ and we can recover v from the measurements $|\gamma_n(v)|^2$. Since u is known, we can finally determine f.

The second step of the recovery algorithm consist in the interpolation of v from the samples $\{v(\lambda_n\}_n$. Since $v \in \mathscr{B}^\infty_\pi$, we necessarily need to apply the results for the sampling series on $\mathscr{B}^\infty_\pi$ as discussed in Section 7.4.3. In particular, it follows from Theorem 13 that we necessarily need oversampling, i.e., Theorem 18 does not hold for functions in $\mathscr{B}^p_\pi$.

Acknowledgements We thank Joachim Hagenauer and Sergio Verdú for drawing our attention to [15] and for related discussions and Ullrich Mönich for carefully reading the manuscript and for helpful comments. The first author thanks the referees of the German Research Foundation (DFG) grant BO 1734/13-2 for highlighting the importance of understanding the strong divergence behavior addressed in Section 7.4.2 of this chapter. He also likes to thank Rudolf Mathar for his insistence in several conversations on the significance of these questions.

The authors gratefully acknowledge support by the DFG through grants BO 1734/22-1 and PO 1347/2-1.

Appendix

This appendix provides a short proof of Theorem 8 in Section 7.4.2

Proof (Theorem 8).

1. First, we prove the statement for the sets (7.18). To this end, let $g \in \mathscr{X}$ and $\epsilon > 0$ be arbitrary. For all $M, N_0 \in \mathbb{N}$, we have to show that there exists a functions f_* in the set (7.18) such that $\|g - f_*\|_{\mathscr{X}} < \epsilon$. Since $\mathscr{X}_0$ is a dense subset of $\mathscr{X}$, there exists a $q \in \mathscr{X}_0$ such that $\|g - q\|_{\mathscr{X}} < \epsilon/2$. Therewith, we define $f_* := q + \frac{\epsilon}{2} f_0$ with $f_0 \in \mathscr{D}_{weak}$ and with $\|f_0\|_{\mathscr{X}} = 1$. Then we get

$$\|g - f_*\|_{\mathscr{X}} \leq \|g - q\|_{\mathscr{X}} + \tfrac{\epsilon}{2}\|f_0\|_{\mathscr{X}} < \tfrac{\epsilon}{2} + \tfrac{\epsilon}{2} = \epsilon\,.$$

Let $M, N_0 \in \mathbb{N}$ be arbitrary. We still have to show that f_* is contained in set (7.18). To this end, we observe that for every $N \in \mathbb{N}$

$$\|\mathrm{T}_N f_*\|_{\mathscr{Y}} = \|\mathrm{T}_N q + \tfrac{\epsilon}{2}\mathrm{T}_N f_0\|_{\mathscr{Y}} \geq \tfrac{\epsilon}{2}\|\mathrm{T}_N f_0\|_{\mathscr{Y}} - \|\mathrm{T}_N q\|_{\mathscr{Y}} \ .$$

Since $q \in \mathscr{X}_0$, (7.16) implies that there is an $N_1 \geq N_0$ such that $1 \geq \|\mathrm{T}_N q - q\|_{\mathscr{Y}} \geq \|\mathrm{T}_N q\|_{\mathscr{Y}} - \|q\|_{\mathscr{Y}}$ for all $N \geq N_1$. Consequently

$$\|\mathrm{T}_N q\|_{\mathscr{Y}} \leq 1 + \|q\|_{\mathscr{Y}} \leq 1 + C_0\|q\|_{\mathscr{X}} \quad \textit{for all } N \geq N_1 \ ,$$

using for the last inequality that $\mathscr{X}$ is continuously embedded in $\mathscr{Y}$ with a certain constant $C_0 < \infty$. Combining the last two inequalities, we get $\|\mathrm{T}_N f_*\|_{\mathscr{Y}} \geq \frac{\epsilon}{2}\|\mathrm{T}_N f_0\|_{\mathscr{Y}} - 1 - C_0\|q\|_{\mathscr{X}}$ for all $N \geq N_1$. Since $f_0 \in \mathscr{D}_{\text{weak}}$ there exits an $N_2 \geq N_1$ such that

$$\|\mathrm{T}_{N_2} f_*\|_{\mathscr{Y}} \geq \tfrac{\epsilon}{2}\|\mathrm{T}_{N_2} f_0\| - 1 - C_0\|q\|_{\mathscr{X}} > M$$

which shows that $f_* \in D(M, N_2) \subset \bigcup_{N \geq N_0} D(M, N)$. Thus the sets (7.18) are dense in $\mathscr{X}$ and it remains to show that these sets are open. To this end, let $M, N \in \mathbb{N}$ and $f_* \in D(M, N)$ be arbitrary, i.e., $\|\mathrm{T}_N f_*\|_{\mathscr{Y}} > M$. Since T_N is a continuous linear operator $\mathscr{X} \to \mathscr{Y}$, there exists a $\delta > 0$ and a neighborhood

$$U_\delta(f_*) = \{f \in \mathscr{X} : \|f - f_*\|_{\mathscr{X}} < \delta\}$$

of f_* such that $\|\mathrm{T}_N f\|_{\mathscr{Y}} > M$ for all $f \in U_\delta$. Thus $D(M, N)$ is open for all $M, N \in \mathbb{N}$ and since the union of (countable many) open sets is again open, the sets (7.18) are also open.

2. We prove (7.19). By the definition of the lim sup operation, the set $\mathscr{D}_{\text{weak}}$ can be written as

$$\mathscr{D}_{\text{weak}} = \left\{ f \in \mathscr{X} \ : \ \lim_{N_0 \to \infty} \sup_{N \geq N_0} \|\mathrm{T}_N f\|_{\mathscr{Y}} = \infty \right\}$$

and we note that for every fixed $f \in \mathscr{X}$ the sequence $\{\sup_{N \geq N_0} \|\mathrm{T}_N f\|_{\mathscr{Y}}\}_{N_0=1}^{\infty}$ is monotone decreasing. Assume that $f \in \mathscr{D}_{\text{weak}}$ and choose $M \in \mathbb{N}$ arbitrary. Then, by the above definition of $\mathscr{D}_{\text{weak}}$, it follows that for arbitrary N_0 there exists an $N \geq N_0$ such that $\|\mathrm{T}_N f\|_{\mathscr{Y}} > M$, i.e., $f \in \bigcup_{N \geq N_0} D(M, N)$, and since this holds for all $M, N \in \mathbb{N}$ we have

$$f \in \bigcap_{M=1}^{\infty} \bigcap_{N_0=1}^{\infty} \bigcup_{N=N_0}^{\infty} D(M, N)$$

which shows that $\mathscr{D}_{\text{weak}} \subset \bigcap_{M=1}^{\infty} \bigcap_{N_0=1}^{\infty} \bigcup_{N=N_0}^{\infty} D(M, N)$. Conversely, assume that $f \in \bigcap_{M=1}^{\infty} \bigcap_{N_0=1}^{\infty} \bigcup_{N=N_0}^{\infty} D(M, N)$. Then to every arbitrary $M \in \mathbb{N}$ and

$N_0 \in \mathbb{N}$ there exists an $N > N_0$ such that $f \in D(M,N)$, i.e., that $\|\mathrm{T}_N f\|_{\mathscr{Y}} > M$. Thus $f \in \mathscr{D}_{\mathrm{weak}}$.

Finally, we prove (7.20). Assume first that $f \in \mathscr{D}_{\mathrm{strong}}$. Then to every $M \in \mathbb{N}$ there exists an $N_0 = N_0(M)$ such that $\|\mathrm{T}_N f\|_{\mathscr{Y}} > M$ for all $N \geq N_0$. In other words

$$f \in \bigcap_{N=N_0(M)}^{\infty} D(M,N) \subset \bigcup_{N_0=1}^{\infty} \bigcap_{N=N_0}^{\infty} D(M,N) \qquad \text{for every } M \in \mathbb{N}\,,$$

which shows that $f \in \bigcap_{M=1}^{\infty} \bigcup_{N_0=1}^{\infty} \bigcap_{N=N_0}^{\infty} D(M,N)$. Conversely, assume that $f \in \bigcap_{M=1}^{\infty} \bigcup_{N_0=1}^{\infty} \bigcap_{N=N_0}^{\infty} D(M,N)$. This means that for any arbitrary $M \in \mathbb{N}$ the function f belongs to $\bigcup_{N_0=1}^{\infty} \bigcap_{N=N_0}^{\infty} D(M,N)$, i.e., there exists an N_0 such that

$$f \in \bigcap_{N=N_0}^{\infty} D(M,N), \qquad \text{i.e.,} \qquad \|\mathrm{T}_N f\|_{\mathscr{Y}} > M \quad \text{for all } N \geq N_0\,.$$

Thus $\lim_{N\to\infty} \|\mathrm{T}_N f\|_{\mathscr{Y}} = \infty$, i.e., $f \in \mathscr{D}_{\mathrm{strong}}$. □

References

1. R. Balan, P.G. Casazza, D. Edidin, On signal reconstruction without phase. Appl. Comput. Harmon. Anal. **20**(3), 345–356 (2006)
2. R. Balan, B.G. Bodmann, P.G. Casazza, D. Edidin, Painless reconstruction from magnitudes of frame coefficients. J. Fourier Anal. Appl. **15**(4), 488–501 (2009)
3. S. Banach, H. Steinhaus, Sur le principle de la condensation de singularités. Fund. Math. **9**, 50–61 (1927)
4. I. Bar-David, An implicit sampling theorem for bounded bandlimited functions. Inf. Control **24**(1), 36–44 (1974)
5. R.P. Boas, *Entire Functions* (Academic, New York, 1954)
6. H. Boche, B. Farrell, Strong divergence of reconstruction procedures for the Paley-Wiener space $\mathscr{PW}_{\pi}^{1}$ and the Hardy space $\mathscr{H}^{1}$. J. Approx. Theory **183**, 98–117 (2014)
7. H. Boche, U.J. Mönich, There exists no globally uniformly convergent reconstruction for the Paley-Wiener space PW_{π}^{1} of bandlimited functions sampled at Nyquist rate. IEEE Trans. Signal Process. **56**(7), 3170–3179 (2008)
8. H. Boche, U.J. Mönich, Approximation of wide-sense stationary stochastic processes by Shannon sampling series. IEEE Trans. Inf. Theory **56**(12), 6459–6469 (2010)
9. H. Boche, U.J. Mönich, Convergence behavior of non-equidistant sampling series. Signal Process. **90**(1), 145–156 (2010)
10. H. Boche, U.J. Mönich, Signal and system approximation from general measurements, in *New Perspectives on Approximation and Sampling Theory: Festschrift in honor of Paul Butzer's 85th birthday* ed. by A.I. Zayed, G. Schmeisser (Applied and Numerical Harmonic Analysis) (Birkhäuser, Basel, 2014)
11. H. Boche, V. Pohl, On the calculation of the Hilbert transform from interpolated data. IEEE Trans. Inf. Theory **54**(5), 2358–2366 (2008)

12. B.G. Bodmann, N. Hammen, Stable phase retrieval with low-redundancy frames. Adv. Compt. Math. **41**(2), 317–331 (2015)
13. J. Brown, On the error in reconstruction a non-bandlimited function by means of the bandpass sampling theorem. J. Math. Anal. Appl. **18**, 75–84 (1967)
14. P.L. Butzer, P.J.S.G. Ferreira, J.R. Higgins, S. Saitoh, G. Schmeisser, R. L. Stens, Interpolation and sampling: E.T. Whittaker, K. Ogura and their Followers. J. Fourier Anal. Appl. **17**(2), 320–354 (2011)
15. P.L. Butzer, M.M. Dodson, P.J.S.G. Ferreira, J.R. Higgins, O. Lange, P. Seidler, R.L. Stens, Multiplex signal transmission and the development of sampling techniques: the work of Herbert Raabe in contrast to that of Claude Shannon. Appl. Anal. **90**(3–4), 643–688 (2011)
16. P.L. Butzer, R.L. Stens, Sampling theory for not necessarily band-limited functions: a historical overview. SIAM Rev. **34**(1), 40–53 (1992)
17. P.L. Butzer, W. Splettstsser, R.L. Stens, The sampling theorem and linear prediction in signal analysis. Jahresber. Deutsch. Math.-Verein. **90**(1), 1–70 (1988)
18. P.L. Butzer, G. Schmeisser, R.L. Stens, Shannon's sampling theorem for bandlimited signals and their Hilbert transform, boas-type formulae for higher order derivatives - the aliasing error involved by their extensions from bandlimited to non-bandlimited signals. Entropy **14**(11), 2192–2226 (2012)
19. L.L. Campbell, Sampling theorem for the Fourier transform of a distribution with bounded support. SIAM J. Appl. Math. **16**(3), 626–636 (1968)
20. E.J. Candès, Y.C. Eldar, T. Strohmer, V. Voroninski, Phase retrieval via matrix completion. SIAM J. Imaging Sci. **6**(1), 199–225 (2013)
21. E.J. Candès, T. Strohmer, V. Voroninski, Phase lift: exact and stable signal recovery from magnitude measurements via convex programming. Commun. Pure Appl. Math. **66**(8), 1241–1274 (2013)
22. O. Christensen, *An Introduction to Frames and Riesz Bases* (Birkhäuser, Bosten, 2003)
23. W. Dickmeis, R.J. Nessel, A quantitative condensation of singularities on arbitrary sets. J. Approx. Theory **43**(4), 383–393 (1985)
24. W. Dickmeis, R.J. Nessel, E. van Wickeren, A quantitative condensation of singularities on arbitrary sets. Manuscripta Math. **52**, 1–20 (1985)
25. Y.C. Eldar, T. Michaeli, Beyond bandlimited sampling: nonlinearities, smoothness and sparsity. IEEE Signal Process. Mag. **26**(3), 48–68 (2009)
26. Y.C. Eldar, V. Pohl, Recovering signals from lowpass data. IEEE Trans. Signal Process. **58**(5), 2636–2646 (2010)
27. P. Erdős, On divergence properties of the Lagrange interpolation parabolas. Ann. Math. **42**(1), 309–315 (1941)
28. P. Erdős, Corrections to two of my papers. Ann. Math. **44**(4), 647–651 (1943)
29. P.J.S.G Ferreira, Nonuniform sampling of nonbandlimited signals. IEEE Signal Process. Lett. **2**(5), 89–91 (1995)
30. R.P. Feynman, *Feynman Lectures on Computation* (Addison-Wesley, Reading, 1996)
31. J.R. Fienup, Phase retrieval algorithms: a comparison. Appl. Opt. **21**(15), 2758–2769 (1982)
32. J.R. Fienup, J.C. Marron, T.J. Schulz, J.H. Seldin, Hubble space telescope characterized by using phase-retrieval algorithms. Appl. Opt. **32**(10), 1747–1767 (1993)
33. N.J. Fine, On the walsh functions. Trans. Am. Math. Soc. **65**(3), 372–414 (1949)
34. J. Finkelstein, Pure-state informationally complete and "really" complete measurements. Phys. Rev. A **70**, 052107 (2004)
35. D. Gabor, Theory of communication. J. IEE **93**(26), 429–441 (1946)
36. G.H. Hardy, Notes on special systems of orthogonal functions (IV): the orthogonal functions of Whittaker's cardinal series. Math. Proc. Cambridge Philos. Soc. **37**(4), 331–348 (1941)
37. J.R. Higgins, Five short stories about the cardinal series. Bull. Am. Math. Soc. **12**(1), 45–89 (1985)
38. J.R. Higgins, *Sampling Theory in Fourier and Signal Analysis – Foundations* (Clarendon Press, Oxford, 1996)
39. L. Hörmander, *Linear Partial Differential Operators* (Springer, Berlin, 1976)

40. P. Jaming, Phase retrieval techniques for radar ambiguity problems. J. Fourier Anal. Appl. **5**(4), 309–329 (1999)
41. A.J. Jerri, The Shannon sampling theorem–its various extensions and applications: a tutorial review. Proc. IEEE **65**(11), 1565–1596 (1977)
42. L.V. Kantorovich, G.P. Akilov, *Functional Analysis in Normed Spaces* (Pergamon Press, New York, 1964)
43. B.Y. Levin, *Lectures on Entire Functions* (American Mathematical Society, Providence, 1997)
44. B.Y. Levin, I.V. Ostrovskii, Small perturbations of the set of roots of sine-type functions. Izv. Akad. Nauk SSSR Ser. Mat **43**(1), 87–110 (1979)
45. J. Miao, T. Ishikawa, Q. Shen, T. Earnest, Extending X-ray crystallography to allow the imaging of noncrystalline materials, cells, and single protein complexes. Annu. Rev. Phys. Chem. **59**, 387–410 (2008)
46. R.P. Millane, Phase retrieval in crystallography and optics. J. Opt. Soc. Am. A **7**(3), 394–411 (1990)
47. A.M. Minkin, The reflection of indices and unconditional bases of exponentials. St. Petersburg Math. J. **3**(5), 1043–1064 (1992)
48. U.J. Mönich, H. Boche, Non-equidistant sampling for bounded bandlimited signals. Signal Process. **90**(7), 2212–2218 (2010)
49. N.K. Nikol'skii, Bases of exponentials and the values of reproducing kernels. Dokl. Akad. Nauk SSSR **252**, 1316–1320 (1980) [English translation, Sov. Math. Dokl. **21**, 937–941 (1980)]
50. A.M. Olevskii, On an orthonormal system and its applications. Mat. Sb. (N.S.) **71(113)**(3), 297–336 (1966)
51. B.S. Pavlov, Basicity of an exponential system and Muckenhoupt's condition. Dokl. Akad. Nauk SSSR **247**, 37–40 (1979) [English translation, Sov. Math. Dokl. **20** 655–659 (1979)]
52. E. Pfaffelhuber, Sampling series for band-limited generalized functions. IEEE Trans. Inf. Theory **17**(6), 650–654 (1971)
53. K. Piwnicki, Modulation methods related to sine-wave crossings. IEEE Trans. Commun. **31**(4), 503–508 (1983)
54. V. Pohl, H. Boche, Advanced topics in system and signal theory: a mathematical approach, in *Foundations in Signal Processing*. Communications and Networking, vol. 4 (Springer, Berlin, 2009)
55. V. Pohl, F. Yang, H. Boche, Phase retrieval from low-rate samples. Sampling Theory Signal Image Process. **14**(1), 71–99 (2015)
56. V. Pohl, F. Yang, H. Boche, Phaseless signal recovery in infinite dimensional spaces using structured modulations. J. Fourier Anal. Appl. **20**(6), 1212–1233 (2014)
57. W. Rudin, *Real and Complex Analysis*, 3rd edn. (McGraw-Hill, Boston, 1987)
58. W. Rudin, *Functional Analysis*, 2nd edn. (McGraw-Hill, Boston, 1991)
59. K. Seip, *Interpolation and Sampling in Spaces of Analytic Functions* (American Mathematical Society, Providence, 2004)
60. C.E. Shannon, Communication in the presence of noise. Proc. IRE **37**(1), 10–21 (1949)
61. E.M. Stein, On limits of sequences of operators. Ann. Math. (2) **74**(1), 140–170 (1961)
62. G. Thakur, Reconstruction of bandlimited functions from unsigned samples. J. Fourier Anal. Appl. **17**(4), 720–732 (2011)
63. L. Tschakaloff, Zweite Lösung der Aufgabe 105. Jahresber. Deutsch. Math.-Verein. **43**, 11–13 (1934)
64. M. Unser, Sampling–50 Years after Shannon. Proc. IEEE **88**(4), 569–587 (2000)
65. D.Ye. Vakman, On the definition of concepts of amplitude, phase and instantaneous frequency of a signal. Radio Eng. Electron. Phys. **17**(5), 754–759 (1972)
66. G. Valiron, Sur la formule d'interpolation de Lagrange. Bull. Sci. Math. **49**(2), 181–192 (1925)
67. J.L. Walsh, A closed set of normal orthogonal functions. Am. J. Math. **45**(1), 5–24 (1923)
68. G. Wunder, R.F.H. Fischer, H. Boche, S. Litsyn, J.-S. No, The PAPR problem in OFDM transmission. IEEE Signal Process. Mag. **30**(6), 130–144 (2013)

69. F. Yang, V. Pohl, H. Boche, Phase retrieval via structured modulations in Paley-Wiener spaces, in *Proceedings of 10th International Conference on Sampling Theory and Applications (SampTA)*, July 2013
70. R.M. Young, *An Introduction to Nonharmonic Fourier Series* (Academic, New York, 1980)
71. M. Zakai, Band-limited functions and the sampling theorem. Inf. Control **8**(2), 143–158 (1965)
72. G. Zauner, Quantum designs: foundations of a noncommutative design theory. Int. J. Quantum Inf. **9**(1), 445–507 (2011)
73. A.I. Zayed, *Advances in Shannon's Sampling Theory* (CRC Press, Boca Raton, 1993)

Chapter 8
Entire Functions in Generalized Bernstein Spaces and Their Growth Behavior

Brigitte Forster and Gunter Semmler

Abstract For an $L^2(\mathbb{R})$ function, the famous theorem by Paley and Wiener gives a beautiful relation between extensibility to an entire function of exponential type and the line support of its Fourier transform. However, there is a huge class of entire functions of exponential type which are not square integrable on an axis, but do have integrability properties on certain half lines. In this chapter we investigate such functions, their growth behavior, and their integrability properties in L^p-norms. We show generalizations of a theorem of J. Korevaar and the Paley-Wiener theorem.

8.1 Entire functions of exponential type, the Paley-Wiener theorem and the theorem of Korevaar — the classical case

The famous Paley-Wiener theorem gives a relation between $L^2(\mathbb{R})$ functions with compact support in Fourier domain and the growth behavior of their extensions as an entire functions in the complex plane. In fact, it states:

Theorem 1 (Paley–Wiener). *[24, p. 101] Let f be an entire function of exponential type $A > 0$. In addition, let f be square-integrable on the real line. Then there exists a function $F \in L^2([-A, A])$ such that*

$$f(z) = \int_{-A}^{A} F(t) e^{izt}\, dt. \tag{8.1}$$

Conversely, if $F \in L^2([-A, A])$ then the function f defined by (8.1) is an entire function of exponential type.

B. Forster (✉)
Fakultät für Informatik und Mathematik, Universität Passau, Passau, Germany
e-mail: brigitte.forster@uni-passau.de

G. Semmler
Fakultät für Mathematik und Informatik, Technische Universität Bergakademie Freiberg, Freiberg, Germany
e-mail: semmler@math.tu-freiberg.de

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_8

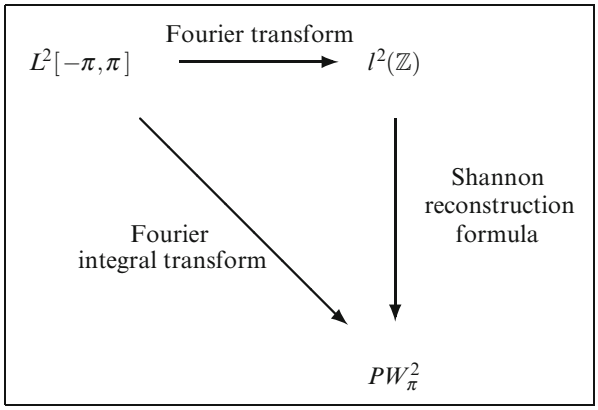


Fig. 8.1 The isometries in the Fourier transform, in the Theorem of Paley–Wiener, i.e., in the Fourier integral transform, and in the sampling theorem via the Shannon reconstruction formula generate a commutative diagram.

The theorem is closely related to the sampling theorem of Shannon, Whittaker, and Kotel'nikov [13] on the reconstruction of band-limited L^2 functions. In fact, according to the Paley-Wiener theorem and the sampling theorem we have the commutative diagram in Fig. 8.1, see [8, p. 30]. The Nyquist rate, giving the maximal recoverable frequency, is identical to the constant A in the growth condition. The space of all band-limited functions is called a Paley-Wiener space and is denoted by PW_A^2, where A is the constant from Theorem 1.

Obviously, changes in one of the transforms in the diagram effect at least one of the other two. For example, changing the interpolation function $\frac{\sin(\pi x)}{\pi x}$ in the Sampling theorem to another appropriate function of sine type serving as Lagrange interpolator has the effect that the Fourier transform must be changed to the so-called non-harmonic Fourier series $\sum_{n\in\mathbb{Z}} \hat{f}_{\lambda_n} e^{\lambda_n x}$. Here, $\{\lambda_n\}_{n\in\mathbb{Z}}$ is the sequence of not necessarily real-valued zeros of the Lagrange interpolator. Considering such generalized Fourier and sampling series has led to a wide class of research articles giving better understanding, what sampling really is, which sets of sampling and interpolation points are appropriate, and how stable they are with respect to small errors, e.g., jitter errors. Good starting points in this research area are, e.g., [1, 11, 15, 17, 24] and the references given therein.

In this chapter, we are interested in the effects which a change of the Paley-Wiener mapping of this commutative diagram evokes. J. Korevaar proved inequalities for the growth of entire functions of exponential type. His main result states:

Theorem 2 (Theorem of Korevaar). *([12], see also [6, Theorem 6.7.17]) If $f(z)$ is an entire function of exponential type τ, $1 \le p < \infty$, and*

$$\int_{-\infty}^{\infty} |f(x)|^p \, dx = M^p < \infty,$$

then

$$|f(x+iy)|^p \le A_p M^p y^{-1} \sinh p\tau y, \tag{8.2}$$

with

$$A_1 = \frac{1}{\pi},\ A_p = \frac{2^k}{p\pi} < \frac{1}{\pi} \quad \text{for} \quad 2^k < p \le 2^{k+1},\ k = 0, 1, 2, \dots \tag{8.3}$$

For $p = 2$ the constant $A_2 = \frac{1}{2\pi}$ in (8.3) is best possible. For $p \neq 2$ the best possible value B_p of the constant A_p in (8.2) satisfies

$$\frac{1}{2p\pi} \le B_p \le \frac{1}{p\pi} \quad (1 \le p < 2), \quad \frac{1}{p\pi} < B_p < \frac{1}{\pi} \quad (p > 2). \tag{8.4}$$

There is a large class of entire functions of exponential type τ, which are not elements of $L^p(\mathbb{R})$ for any $1 \le p \le \infty$. Simple examples are entire functions with triangular indicator diagram. The aim of this chapter is to establish analog results for entire functions with polygonal indicator diagram.

The organization of the chapter is as follows: First, we consider the general setting of entire functions with indicator diagram contained in some polygon in the complex plane and define corresponding generalized Bernstein spaces. Then, in Section 8.3, we give an extension of the Paley-Wiener theorem to this setting for the L^p-norm, $1 \le p \le 2$, as well as its converse. These results are essential for Section 8.4, where we prove the generalization of Korevaar's theorem. The last section is devoted to relations to the Phragmén-Lindelöf theorems.

Parts of this work have been mentioned in the 4-page article [9], which has been presented at the SampTA conference 2011. The proofs of the results are given in this chapter, here.

8.2 Entire functions in generalized Bernstein spaces

Whereas the L^2-theory of the commutative diagram of sampling theorem, Fourier transform and Paley-Wiener theorem, is well understood, the complete characterization in the L^p- theory, $1 \le p \le \infty$, is still an open question. However, a good part of the developments in sampling theory is driven by results from function theory. For example, the standard proof of the Paley-Wiener Theorem 1 is based on the one hand on the Theorem of Morera and on the other hand on a limiting process of curves encycling the straight-line interval $[-A, A]$ in the complex plane. In this chapter, we consider a special class of these curves, i.e., polygonal curves.

8.2.1 Growth and indicator function

Let f be an entire function of exponential type $\tau > 0$, i.e., τ is the smallest number such that for all $\varepsilon > 0$ there exist constants $A(\varepsilon)$ with

$$|f(z)| < A(\varepsilon)e^{(\tau+\varepsilon)|z|}.$$

The indicator function of f is defined as

$$h_f(\theta) := \limsup_{r\to\infty} \frac{1}{r} \ln |f(re^{i\theta})|. \tag{8.5}$$

It represents the growth exponent in direction θ and is bounded by the type of f, i.e., $h(\theta) \leq \tau$.

The Borel transform of the entire function $f(z) = \sum_{n=0}^{\infty} a_n z^n$ of exponential type τ is defined as

$$Bf(w) = \sum_{n=0}^{\infty} n! \frac{a_n}{w^{n+1}} \quad \text{for } |w| > \tau. \tag{8.6}$$

Let E^* denote the smallest closed convex set outside which $Bf(w)$ is regular. Then E^* is called conjugate indicator diagram of f and the Pólya representation

$$f(z) = \frac{1}{2\pi i} \int_C Bf(w)e^{zw}\, dw \tag{8.7}$$

holds for all $z \in \mathbb{C}$, where C is a contour containing E^* in its interior. Conversely, in the half plane $\{z\colon \operatorname{Re}(ze^{-i\theta}) > h_f(-\theta)\}$, the Borel transform Bf can be computed from f using the Laplace integral

$$Bf(z) = \int_0^{\infty} e^{-zw} f(w) dw, \tag{8.8}$$

where the integration is along the ray $\{w = te^{-i\theta} : t \geq 0\}$. The supporting function

$$k_K(\theta) := \max_{z\in K} \operatorname{Re}(ze^{-i\theta})$$

of a compact set $K \subset \mathbb{C}$ designates the (signed) maximum distance of its projection onto the ray $\{z\colon \arg z = \theta\}$ from the origin. If K is the conjugate indicator diagram of an entire function f of exponential type τ then $h_f(-\theta) = k_{E^*}(\theta)$ is true for all $\theta \in \mathbb{R}$. Equivalently, $h_f(\theta) = k_E(\theta)$ holds for the indicator diagram $E := \{z : z^* \in E^*\}$ where z^* denotes the complex conjugate of a number $z \in \mathbb{C}$. This gives rise to the idea to consider all entire functions with growth behavior related to convex sets. In the following, we restrict ourselves to convex polygons D.

A good overview on entire functions of exponential type and the properties of their indicator diagrams is given in the monograph of R. P. Boas, Jr. [6, Ch. 5].

8.2.2 *Generalized Bernstein spaces*

Let D be a closed convex polygon with $N \geq 2$ vertices and ∂D the boundary of D. For $N = 2$, the polygon D is degenerated to a straight line segment. We assume that the origin lies in D. Let θ_j denote the angles between the normals N_j to the sides l_j of D ($j = 1, \ldots, N$) and the positive real axis, such that

$$0 \leq \theta_1 < \theta_2 < \ldots < \theta_N < 2\pi.$$

The vertices $a_j, j = 1, \ldots, N$, are numbered in the mathematical positive sense, such that $l_j = [\, a_j, a_{j+1}\,]$ for all $j = 1, \ldots, N$, where $a_{N+1} := a_1$ (see Figure 8.2). By L_j we denote for $j = 1, \ldots, N$ the straight line containing the segment l_j.

Let $k_D(\theta)$ denote the supporting function of D. The normals N_j to the sides of D split the complex plane into the angles $\Gamma_j := \{z : \arg(z) \in]\, \theta_{j-1}, \theta_j\, [\,\}$ where $\theta_0 := \theta_N$. In each of these angles we have

$$k_D(\theta) = \mathrm{Re}\,(a_j e^{-i\theta}) \quad \text{for all } e^{i\theta} \in \overline{\Gamma}_j.$$

Consider for $1 \leq p < \infty$ the set $B_p(D)$ of all entire functions f of exponential type for which

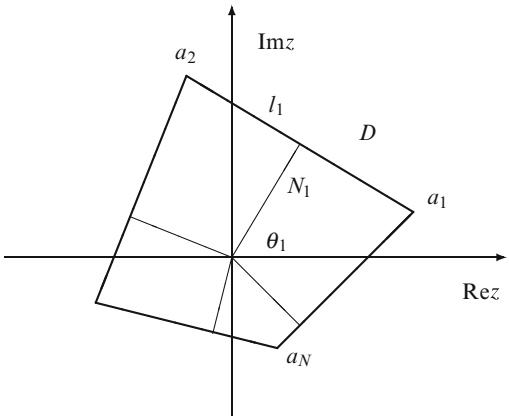


Fig. 8.2 Convex polygon D in the complex plane with vertices a_k, sides l_k, and normals N_k, $k = 1, \ldots, N$.

$$\|f\|_{p,D} := \sup_{\theta\in[0,2\pi[} \left\{ \int_0^\infty |f(re^{i\theta})|^p e^{-prk_D(\theta)}\, dr \right\}^{1/p} \tag{8.9}$$

is finite. The set $B_p(D)$ constitutes a Banach space with norm $\|\cdot\|_{B_p(D)} = \|\cdot\|_{p,D}$ [15]. In fact, an entire function f of exponential type belongs to $B_p(D)$ if the functions

$$r \mapsto f(re^{i\theta_j})e^{-rk_D(\theta_j)}, \quad j = 1,\dots,N, \tag{8.10}$$

belong to $L^p([0,\infty[)$, in which case we have

$$\|f\|_{p,D} = \max_{j=1,\dots,N} \left\{ \int_0^\infty |f(re^{i\theta_j})|^p e^{-prk_D(\theta_j)}\, dr \right\}^{1/p}, \tag{8.11}$$

see [15, p. 632], or as an equivalent norm

$$|||f|||_{p,D} := |||f|||_{B_p(D)} := \sum_{j=1}^N \left\{ \int_0^\infty |f(re^{i\theta_j})|^p e^{-prk_D(\theta_j)}\, dr \right\}^{1/p}.$$

In fact, $\|f\|_{p,D} \le |||f|||_{p,D} \le N\|f\|_{p,D}$. For $p = \infty$, the entire functions with norm

$$\|f\|_{\infty,D} = \sup_{z\in\mathbb{C}} \left\{ |f(z)|e^{-|z|k_D(\arg z)} \right\} < \infty$$

constitute a Banach space $B_\infty(D)$ [15].

If D is degenerated to the interval $[-i\tau, i\tau]$ then $B_p(D)$ for $1 \le p < \infty$ is identical to the classical Bernstein space L^p_τ consisting of those entire functions f of exponential type at most τ whose restrictions to the real line belong to $L^p(\mathbb{R})$. Moreover, the norms $\|\cdot\|_{p,D}$ and $\|\cdot\|_{L^p(\mathbb{R})}$ turn out to be equivalent [15, p. 626].

Similarly, $B_\infty(D)$ coincides in that case with the Bernstein space B_τ of entire functions of exponential type at most τ that are bounded on $\mathbb{R}$ [15, p.634].

Lemma 1. *Let $f \in B_p(D)$ and $1 \le p < \infty$. Then*

$$h_f(\theta) \le k_D(\theta) \qquad \forall\theta \in \mathbb{R}.$$

Thus, the indicator diagram of f is contained in D.

Proof. Define entire functions $g_j, j = 1,\dots,N$, of exponential type by $g_j(z) := f(z)e^{-a_j^* z}$. Then $r \mapsto g(re^{i\theta})$ belongs to $L^p([0,\infty[)$ for $\theta \in [\theta_{j-1}, \theta_j]$, since

$$|g_j(re^{i\theta})| = |f(re^{i\theta})|e^{-\mathrm{Re}\,(a_j^* e^{i\theta})} = |f(re^{i\theta})|e^{-rk_D(\theta)}.$$

From [6, Theorem 6.7.8] (cf. [5] for a proof), we infer that

$$\lim_{r\to\infty} g_j(re^{i\theta}) = 0 \qquad \forall \theta \in [\theta_{j-1}, \theta_j]. \tag{8.12}$$

If an entire function of exponential type is bounded on the sides of an angle that is smaller than a half plane, then the Phrágmen-Lindelöf theorem (cf. [6, Theorem 1.4.2.]) implies that this function is bounded in the entire angle by the same constant. Another similar result (cf. [6, Theorem 1.4.4.]) even implies that relation (8.12) holds uniformly with respect to θ. Consequently, there is a constant $C > 0$ with

$$|f(z)| \le C|e^{a_j^* z}| = Ce^{|z|k_D(\arg z)}, \qquad z \in \overline{\Gamma}_j, \quad j = 1, \dots, N.$$

Hence we obtain

$$h_f(\theta) = \limsup_{r\to\infty} \frac{1}{r} \ln |f(re^{i\theta})| \le k_D(\theta).$$

If E denotes the indicator diagram, we have thus proved $k_E(\theta) \le k_D(\theta)$, which implies $E \subseteq D$.

There is a partial converse of the preceding lemma.

Lemma 2. *Let D be a convex polygon with $N > 2$ vertices. Let D contain the origin in its interior. If f is an entire function of exponential type so that its indicator diagram E is properly included in D (i.e., $E \subset D, E \cap \partial D = \emptyset$), then $f \in B_p(D)$ for all $p \in [1, \infty[$.*

Proof. Since E is compact there is $\varepsilon > 0$ such that $k_D(\theta) - k_E(\theta) > \varepsilon$ for all $\theta \in \mathbb{R}$. Using again $h_f(\theta) = k_E(\theta)$ we infer from definition (8.5) of the indicator function that

$$|f(re^{i\theta_j})| \le e^{r(h_f(\theta_j)+\varepsilon/2)} < e^{r(k_D(\theta_j)-\varepsilon/2)}, \qquad j = 1, \dots, N,$$

for all sufficiently large r, say $r \ge r_0$. Thus

$$\begin{aligned}
&\int_0^\infty |f(re^{i\theta_j})|^p e^{-prk_D(\theta_j)}\, dr \\
&\le \int_0^{r_0} |f(re^{i\theta})|^p e^{-prk_D(\theta_j)}\, dr + \int_{r_0}^\infty e^{pr(k_D(\theta_j)-\varepsilon/2)} e^{-prk_D(\theta_j)}\, dr \\
&= C(r_0) + \int_{r_0}^\infty e^{-pr\varepsilon/2}\, dr \\
&< \infty.
\end{aligned}$$

Hence the functions (8.10) belong to $L^p([0, \infty[)$, and $f \in B_p(D)$ according to what was said above.

This lemma cannot be improved in the sense that the inclusion $E \subseteq D$ of the indicator diagram E in some polygon D would imply that $f \in B_p(D)$ for any p. As a counterexample, consider the entire function $f(z) := \cos(z) + \cos(iz)$ of exponential type. An elementary consideration yields $h_f(\theta) = \max(|\cos(\theta)|, |\sin(\theta)|)$ and hence the indicator diagram is the square $E := \{z : |\mathrm{Re}\,(z)| + |\mathrm{Im}\,(z)| \leq 1\}$. But the divergence of the integral $\int_0^\infty |\cos(x) + \cos(ix)|^p e^{-px}\,dx$ shows that f does not belong to $B_p(E)$ for any $p \in [1, \infty[$.

8.3 Extensions of the Paley-Wiener theorem

Let $G \subset \mathbb{C}$ be a region bounded by a closed rectifiable Jordan curve. We denote by $E^p(G)$, $p > 0$, the class of all functions $f(z)$ analytic in G for which there is a sequence of closed rectifiable Jordan contours $\Gamma_n \subset G$ converging to the boundary ∂G of G such that

$$\sup_n \int_{\Gamma_n} |f(z)|^p |dz| < \infty. \tag{8.13}$$

Convergence to the boundary means here that each compact set $K \subset G$ is surrounded by almost all contours Γ_n.

In the same way, we denote by $E^p(\mathrm{ext}\,G)$, $p > 0$, the class of all functions $f(z)$ analytic in $\mathrm{ext}\,G := \mathbb{C} \setminus \overline{G}$ with $f(z) \to 0$ as $|z| \to \infty$ such that (8.13) is true for a sequence of closed rectifiable Jordan contours $\Gamma_n \subset \mathrm{ext}\,G$ converging to ∂G from outside.

For each function $f \in E^p(G)$ and each function $g \in E^p(\mathrm{ext}\,G)$ the angular boundary values exist almost everywhere on ∂G, and the functions defined on the boundary by these values are elements of $L^p(\partial D)$. Moreover, for $p \geq 1$, both classes constitute Banach spaces with norm

$$\|f\|_{E^p(G)} = \|f\|_{L^p(\partial G)} \quad \text{resp.} \quad \|g\|_{E^p(\mathrm{ext}\,G)} = \|g\|_{L^p(\partial G)}.$$

More properties of those spaces can be found in [7] and [20]. Sedletskii showed in [22] that for $1 < p < \infty$ and $1/p + 1/q = 1$

$$(E^p(G))' \cong E^q(\mathrm{ext}\,G) \quad \text{resp.} \quad (E^p(\mathrm{ext}\,G))' \cong E^q(G), \tag{8.14}$$

where the prime denotes the dual space.

The following theorem is an extension of the Paley–Wiener theorem. Levin [14, p. 392] has given a generalization of this theorem in the L^2 setting. It gives a relation between entire functions of exponential type with indicator diagram of polygonal form and their Borel transform. We check that these ideas carry over to the L^p-setting if $1 < p \leq 2$. Denote by D^* the complex conjugate of the polygon D, i.e., $D^* := \{z^* : z \in D\}$.

Theorem 3. *Let D be a closed convex polygon with $N \geq 2$ vertices and ∂D the boundary of D. We assume that the origin lies in D. D^* denotes the complex conjugate polygon. Let $f \in B_p(D)$, $1 < p \leq 2$, $1/p + 1/q = 1$. Then the Borel transform $\psi := Bf$ belongs to $E^q(\operatorname{ext} D^*)$ and f has the representation*

$$f(z) = \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi(\lambda)\, d\lambda. \tag{8.15}$$

Moreover, we have the estimate

$$\|\psi\|_{E^q(\operatorname{ext} D^*)} \leq M_p \sqrt[q]{2\pi}\, |||f|||_{B_p(D)}, \tag{8.16}$$

where $M_p = (p^{1/p}/q^{1/q})^{1/2}$ is the Babenko-Beckner constant.

From Lemma 1 we know that under the assumptions of the theorem the indicator diagram E of f is contained in D, so that the Borel transform is defined outside the polygon D^*. As soon as $\psi \in E^q(\operatorname{ext} D^*)$ is established we know that the Borel transform has an extension to ∂D^* and is an L^q function there, so that it makes sense to consider the integral occurring on the right-hand side of (8.15). In contrast to Levin [14], who requires the indicator diagram to be polygonal, we make no assumptions on the form of E.

It is well known (cf. [6, Section 5.3]) that f possesses the Pólya representation (8.7) where Γ is some Jordan curve inclosing the conjugate indicator diagram of f. So the theorem claims that this formula remains true if Γ is shrunken to the boundary of D.

In the special case $D = [-i\tau, i\tau]$, Theorem 3 says that an entire function f of degree not larger than τ with $f|\mathbb{R} \in L^p(\mathbb{R})$ $(1 < p \leq 2)$ has a representation

$$f(z) = \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi(\lambda)\, d\lambda = \frac{1}{2\pi i} \int_{-i\tau}^{i\tau} e^{\lambda z} \psi_+(\lambda)\, d\lambda + \frac{1}{2\pi i} \int_{i\tau}^{-i\tau} e^{\lambda z} \psi_-(\lambda)\, d\lambda$$

where

$$\psi_+(it) := \lim_{\varepsilon \to 0+0} \psi(it + \varepsilon), \qquad \psi_-(it) := \lim_{\varepsilon \to 0+0} \psi(it - \varepsilon), \qquad t \in [-\tau, \tau] \quad \text{a.e.}$$

Setting $\hat{f}(t) := \psi_+(it) - \psi_-(it) \in L^q([-\tau, \tau])$ we arrive at the representation

$$f(z) = \frac{1}{2\pi} \int_{-\tau}^{\tau} e^{itz} \hat{f}(t)\, dt, \tag{8.17}$$

i.e., f can be represented as the inverse Fourier transform of an L^q function with bounded support. The original Paley–Wiener theorem [21] treats thus in our terminology $B_2(D)$ functions with polygons degenerated to a line segment. Analog theorems for $B_p(D)$ functions with line-like polygons D and $p \neq 2$ have been proved

in [18, 19], and [4], see also Zygmund [25]. In the following, we show the result for $1 < p \le 2$ and an arbitrary convex polygon containing the origin.

Proof (Theorem 3). We follow the proof in [14]. Let us consider the Borel transform of f on the line $\{(\varepsilon + it + k_D(\theta_j))e^{-i\theta_j} : t \in \mathbb{R}\}$ which is parallel at a distance $\varepsilon > 0$ to the side $l_j^* = [a_j^*, a_{j+1}^*]$ of D^*. Setting $w = re^{i\theta_j}$, $dw = e^{i\theta_j}\,dr$ in (8.8), the Borel transform ψ on this line can be computed by

$$\begin{aligned}\psi_j^\varepsilon(t) &:= \psi((\varepsilon + it + k_D(\theta_j))e^{-i\theta_j}) \\ &= e^{i\theta_j}\int_0^\infty \exp(-(\varepsilon + it + k_D(\theta_j))e^{-i\theta_j}\cdot re^{i\theta_j})f(re^{i\theta_j})\,dr \\ &= e^{i\theta_j}\int_0^\infty e^{-(\varepsilon+it)r} f_j(r)\,dr\end{aligned}$$

with the abbreviation

$$f_j(r) := f(re^{i\theta_j})e^{-k_D(\theta_j)r}, \qquad r \ge 0.$$

The convergence condition

$$\operatorname{Re}\,((\varepsilon + it + k_D(\theta_j))e^{-i\theta_j}\cdot e^{i\theta_j}) > h_f(\theta_j)$$

for this integral is fulfilled for all $\varepsilon > 0$ in view of Lemma 1. We have $f_j \in L^p([0,\infty[)$ because $f \in B_p(D)$. Moreover, $\psi_j^\varepsilon \in L^q(\mathbb{R})$. This can be seen from the Hausdorff-Young inequality with sharp Babenko-Beckner constant $M_p = (p^{1/p}/q^{1/q})^{1/2}$ [2, 3]:

$$\begin{aligned}\|\psi_j^\varepsilon\|_{L_q(\mathbb{R})}^q = \int_{\mathbb{R}}\left|\int_0^\infty e^{-\varepsilon r} f_j(r)e^{-irt}\,dr\right|^q dt \quad &\le \quad 2\pi M_p^q\left(\int_0^\infty \left|e^{-\varepsilon r}f_j(r)\right|^p dr\right)^{q/p} \\ \le 2\pi M_p^q\left(\int_0^\infty |f_j(r)|^p\,dr\right)^{q/p} \quad &= \quad 2\pi M_p^q\|f_j\|_{L^p([0,\infty[)}^q.\end{aligned}$$

By the same means we see that

$$\|\psi_j^{\varepsilon_1} - \psi_j^{\varepsilon_2}\|_{L^q(\mathbb{R})}^q \le 2\pi M_p^q\left(\int_0^\infty |e^{-\varepsilon_1 r} - e^{-\varepsilon_2 r}|^p\,|f_j(r)|^p\,dr\right)^{q/p}$$

so that by the dominated convergence theorem the sequence ψ_j^ε converges in L^q-norm to some $\psi_j \in L^q(\mathbb{R})$ for $\varepsilon \to 0+$. Moreover,

$$\|\psi_j\|_{L^q(\mathbb{R})}^q \le 2\pi M_p^q\|f_j\|_{L^p([0,\infty[)}^q. \tag{8.18}$$

Putting

$$\psi(z) := \psi_j(-i(ze^{i\theta_j} - k_D(\theta_j)), \qquad z \in l_j^*,$$

we define an extension of ψ onto ∂D^* such that $\psi \in L^q(\partial D^*)$. From (8.6) we know that $\lim_{|z|\to\infty} \psi(z) = 0$ holds for every Borel transform of an entire function of exponential type. Hence $\psi \in E^q(\operatorname{ext} D^*)$ and (8.18) yields

$$\|\psi\|_{E^q(\operatorname{ext} D^*)} = \|\psi\|_{L^q(\partial D^*)} = \sum_{j=1}^{N} \|\psi\|_{L^q([a_j^*, a_{j+1}^*])} \le \sum_{j=1}^{N} \|\psi_j\|_{L^q(\mathbb{R})}$$

$$\le M_p \sqrt[q]{2\pi} \sum_{j=1}^{N} \|f_j\|_{L^p([0,\infty[)} = M_p \sqrt[q]{2\pi} \|f\|_{B_p(D)},$$

and thus (8.16). From the relation

$$f(z) = \lim_{\varepsilon\to 0+0} \frac{1}{2\pi i} \int_{(1+\varepsilon)\partial D^*} e^{\lambda z} \psi(\lambda)\, d\lambda = \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi(\lambda)\, d\lambda$$

(cf. [14, Anhang 1, § 3] for further details), we finally deduce (8.15).

The preceding theorem has a counterpart.

Theorem 4. *Let $1 \le q \le 2$, and let D be a closed convex polygon with $N \ge 2$ vertices. Let D contain the origin in its interior. Let $\psi \in E^q(\operatorname{ext} D^*)$, and set*

$$f(z) := \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi(\lambda)\, d\lambda.$$

Then $f \in B_p(D)$ for $1/p + 1/q = 1$, and there is some constant $C > 0$ independent of ψ such that

$$\|f\|_{B_p(D)} \le C \|\psi\|_{E^q(\operatorname{ext} D^*)}. \tag{8.19}$$

Proof. The estimate

$$|f(z)| \le \frac{1}{2\pi} \max_{\lambda\in\partial D^*} |e^{\lambda z}| \int_{\partial D^*} |\psi(\lambda)|\, |d\lambda| = \frac{1}{2\pi} e^{|z| k_D(\arg z)} \int_{\partial D^*} |\psi(\lambda)|\, |d\lambda|$$

shows that f is an entire function of exponential type. In order to verify $f \in B_p(D)$ it remains to prove that each of the functions

$$f_j(r) := f(re^{i\theta_j}) e^{-rk_D(\theta_j)}, \qquad j = 1, \dots, N,$$

belongs to $L^p([0, \infty[)$.

First, we consider the case where $p = q = 2$. Let D_1 be a disk with center 0 so small that it is contained in the interior of D as well as in the interiors of

the reflections of D in the normals N_j. Define S to be the space of all functions ψ holomorphic in $\operatorname{ext} D_1^*$, continuous in $\overline{\operatorname{ext} D_1^*}$, and such that $\psi(z) \to 0$ for $|z| \to \infty$. Clearly, $S \subset E^2(\operatorname{ext} D^*)$, and S is dense in $E^2(\operatorname{ext} D^*)$. This follows from the fact that for $\psi \in E^2(\operatorname{ext} D^*)$ the function $g(z) := \psi(1/z)$ belongs to $E^2(G)$ where G is the domain

$$G := \{1/z \colon z \in \operatorname{ext} D^*\} \cup \{0\}. \tag{8.20}$$

Since $g(0) = 0$ also $g(z)/z \in E^2(G)$. Moreover, since G is a Smirnov domain, there is a sequence $p_n(z)$ of polynomials with

$$\|p_n(z) - g(z)/z\|_{E^q(G)} \to 0 \tag{8.21}$$

for $n \to \infty$, see [7, chapter 10.3]. Then $\psi_n(z) := p_n(1/z)/z$ is a sequence in S tending to $\psi \in E^2(\operatorname{ext} D^*)$, which shows the density.

We are now going to prove the existence of a constant $C > 0$ with (8.19) for all $\psi \in S$ and $p = q = 2$. By the Cauchy integral theorem we know that for $\psi \in S$

$$f(z) = \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi(\lambda)\, d\lambda = \frac{1}{2\pi i} \int_{\partial D_1^*} e^{\lambda z} \psi(\lambda)\, d\lambda.$$

Hence we obtain

$$\begin{aligned}
\int_0^\infty |f_j(r)|^2\, dr &= \int_0^\infty f(re^{i\theta_j}) f(re^{i\theta_j})^* e^{-2rk_D(\theta_j)}\, dr \\
&= \int_0^\infty \left(\frac{1}{2\pi i} \int_{\partial D_1^*} e^{re^{i\theta_j} z} \psi(z)\, dz \right) \left(\frac{1}{2\pi i} \int_{\partial D^*} e^{re^{i\theta_j} w} \psi(w)\, dw \right)^* e^{-2rk_D(\theta_j)}\, dr \\
&= \frac{1}{4\pi^2} \int_{\partial D^*} \int_{\partial D_1^*} \int_0^\infty e^{re^{i\theta_j} z + re^{-i\theta_j} w^* - 2rk_D(\theta_j)} \psi(z) \psi(w)^*\, dr\, dz\, dw^* \\
&= \frac{1}{4\pi^2} \int_{\partial D^*} \int_{\partial D_1^*} \frac{1}{e^{i\theta_j} z + e^{-i\theta_j} w^* - 2k_D(\theta_j)} \psi(z) \psi(w)^*\, dz\, dw^*.
\end{aligned}$$

The integrand is holomorphic with respect to z in $\operatorname{ext} D_1^* \cup \{\infty\}$ with exception of $z_j := 2e^{-i\theta_j} k_D(\theta_j) - w^* e^{-2i\theta_j}$, where it has a simple pole. Hence the residue theorem yields

$$\begin{aligned}
\int_0^\infty |f_j(r)|^2\, dr &= -\frac{1}{2\pi i} \int_{\partial D^*} e^{i\theta_j} \psi(2e^{-i\theta_j} k_D(\theta_j) - w^* e^{-2i\theta_j}) \psi(w)^* dw^* \\
&= \left(\frac{1}{2\pi i} \int_{\partial D^*} e^{-i\theta_j} \psi(2e^{-i\theta_j} k_D(\theta_j) - w^* e^{-2i\theta_j})^* \psi(w)\, dw \right)^*.
\end{aligned}$$

The point z_j is obtained from w by means of a reflection in the straight line L_j^*, i.e., the line defined by the segment l_j^*. Since the integrand of the last integral is

holomorphic, we can replace the integration along ∂D^* by the integration along any closed curve Γ that has winding number one about the points of D_1^* and winding number 0 about the image of D_1^* under the two reflections just described. In particular, we can choose Γ to consist of a sufficiently large section of L_j^* that is closed up by a semicircle. Since

$$\psi(2e^{-i\theta_j}k_D(\theta_j) - w^* e^{-2i\theta_j})^* \psi(w) = O(1/|w|^2),$$

the integral along this semicircle tends to 0 if its radius tends to infinity. Consequently, we can just integrate over L_j^*. But on this line, the reflection mentioned above is the identity map. Thus, choosing the parameterization $w = e^{-i\theta_j}(k_D(\theta_j) + ix)$, we obtain

$$\int_0^\infty |f_j(r)|^2 \, dr = \frac{1}{2\pi} \int_{-\infty}^\infty |\psi(e^{-i\theta_j}(k_D(\theta_j) + ix))|^2 \, dx = \frac{1}{2\pi} \|\psi\|^2_{L_2(L_j^*)}.$$

Postponing the existence of a positive constant C_1 independent of ψ with

$$\|\psi\|_{L_2(L_j^*)} \le C_1 \|\psi\|_{E^2(\operatorname{ext} D^*)}$$

to the subsequent Lemma 3, and taking (8.11) into account, we conclude that (8.19) holds for $\psi \in S$ and $p = q = 2$. The linear and bounded operator $\psi \mapsto f$ from S to $B_2(D)$ can therefore be extended to an operator from $E^2(\operatorname{ext} D^*)$ to $B_2(D)$ with the same norm in the usual manner: To $\psi \in E^2(\operatorname{ext} D^*)$ we find a sequence $\psi_n \in S$ with $\psi_n \to \psi$ in $E^2(\operatorname{ext} D^*)$. The associated functions

$$f_n(z) := \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi_n(\lambda) \, d\lambda$$

are by (8.19) a Cauchy sequence in $B_2(D)$ and hence convergent to some $F \in B_2(D)$ satisfying $\|F\|_{B_2(D)} \le C \|\psi\|_{E^2(\operatorname{ext} D^*)}$. Moreover, the sequence $f_n(z)$ converges to $f(z)$ uniformly on compact subsets of $\mathbb{C}$ since

$$\begin{aligned} |f_n(z) - f(z)| &= \left| \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi_n(\lambda) \, d\lambda - \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi(\lambda) \, d\lambda \right| \\ &\le \frac{1}{2\pi} \|\psi_n - \psi\|_{E^2(\operatorname{ext} D^*)} \left(\int_{\partial D^*} |e^{2\lambda z}| \, |d\lambda| \right)^{1/2}. \end{aligned}$$

We investigate the consequences of the convergence of f_n to F in $B_2(D)$. The sequences

$$f_j^{(n)}(r) := f_n(re^{i\theta_j}) e^{-rk_D(\theta_j)}, \qquad j = 1, \dots, N,$$

converge in $L^2([0, \infty[)$ to

$$F_j(r) := F(re^{i\theta_j})e^{-rk_D(\theta_j)}, \qquad j = 1, \dots, N,$$

for $n \to \infty$. Hence a subsequence $f_j^{(n_k)}$ converges to F_j almost everywhere for $k \to \infty$. But if F coincides with f almost everywhere on the rays $\{z\colon \arg z = \theta_j\}$ we infer from the identity theorem $f = F \in B_2(D)$ and thus also (8.19) for arbitrary $\psi \in E^2(\operatorname{ext} D^*)$.

Now we show the theorem for $q = 1, p = \infty$. Indeed,

$$\begin{aligned}
\|f\|_{\infty,D} &= \sup_{z\in\mathbb{C}} \left\{ |f(z)| e^{-|z|k_D(\arg z)} \right\} \\
&= \sup_{z\in\mathbb{C}} \left\{ \left| \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} \psi(\lambda)\, d\lambda \right| e^{-|z|k_D(\arg z)} \right\} \\
&\le \frac{1}{2\pi} \sup_{z\in\mathbb{C}} \left\{ \int_{\partial D^*} |\psi(\lambda)|\, |d\lambda| \max_{\lambda\in\partial D} \left| e^{\lambda^* z} \right| e^{-|z|k_D(\arg z)} \right\} \\
&= \frac{1}{2\pi} \sup_{z\in\mathbb{C}} \left\{ \int_{\partial D^*} |\psi(\lambda)|\, |d\lambda| \right\} \\
&= \frac{1}{2\pi} \|\psi\|_{E^1(\operatorname{ext} D^*)}.
\end{aligned}$$

The validity of our assertions for $1 < q < 2$ is now a consequence of the Riesz-Thorin Interpolation Theorem.

It cannot be expected that Theorem 4 also holds for $q > 2$, since for a function $\hat{f} \in L^q([-\tau, \tau])$, $q > 2$ formula (8.17) does not always define a function $f \in L^p(\mathbb{R})$, see [23, p. 111]. A counterexample is constructed in the recent article [16]. There, also an explicit converse of Theorem 3 is given for $f \in B_p([-\sigma, \sigma])$, $\sigma > 0$, $1 < p < 2$, with additional summability assumptions on the Fourier coefficients of $\hat{f}$ or the corresponding sampling sequence of f.

If $\mathbb{D} := \{z\colon |z| < 1\}$ is the unit disk, it is common to write $H^p := E^p(\mathbb{D})$ for the Smirnov space, and H^p is then called Hardy space. Equipped with these notations we are ready to give a proof for the lemma that closes the gap in the derivation of the preceding theorem.

Lemma 3. *Let the origin be contained in the interior of the convex polygon D. Let D have $N \ge 2$ vertices. Then there is a constant $C > 0$ depending only on D such that for every function $\psi \in E^2(\operatorname{ext} D^*)$ we have*

$$\|\psi\|_{L_2(L_j^*)} \le C \|\psi\|_{E^2(\operatorname{ext} D^*)}, \qquad j = 1, \dots, N.$$

Proof. For $\varepsilon > 0$ we subdivide L_j^* into

$$M_j := \{z \in L_j^* : \min_{w\in l_j^*} |z - w| \le \varepsilon\} \qquad \text{and} \qquad N_j := \{z \in L_j^* : \min_{w\in l_j^*} |z - w| > \varepsilon\},$$

and estimate the L^2-norm of ψ on M_j and N_j separately. The Cauchy integral formula

$$\psi(z) = \frac{1}{2\pi i}\int_{\partial D^*} \frac{\psi(w)}{w-z}dw, \qquad z \in \operatorname{ext} D^*,$$

and the Cauchy-Schwarz inequality yield the estimate

$$|\psi(z)|^2 \le \frac{1}{2\pi}\int_{\partial D^*} \frac{|dw|}{|w-z|^2}\int_{\partial D^*} |\psi(w)|^2\,|dw| \le \frac{|\partial D|}{2\pi}\max_{w\in\partial D^*}\frac{1}{|w-z|^2}\,\|\psi\|^2_{L^2(\partial D^*)},$$

where also $z \in \operatorname{ext} D^*$. Here, we have denoted the length of the boundary curve of D by $|\partial D|$. Now we find

$$\|\psi\|^2_{L^2(N_j)} \le C_1\|\psi\|^2_{E^2(\operatorname{ext}\partial D^*)}, \tag{8.22}$$

with the constant

$$C_1 := \frac{|\partial D|}{2\pi}\int_{N_j}\max_{w\in\partial D^*}\frac{1}{|w-z|^2}\,|dz| < \infty.$$

The function $g(z) := \psi(1/z)$ belongs to $E^2(G)$ where G is the domain (8.20). If we had proved that for $\tilde{M}_j := \{1/z : z \in M_j\}$ holds

$$\|g\|_{L^2(\tilde{M}_j)} \le C_2\|g\|_{E^2(G)}, \tag{8.23}$$

we could conclude that also

$$\begin{aligned}\|\psi\|^2_{L^2(M_j)} &= \int_{M_j}|\psi(w)|^2\,|dw| = \int_{\tilde{M}_j}|g(z)|^2\left|\frac{dz}{z^2}\right| \le \max_{z\in\tilde{M}_j}\left|\frac{1}{z^2}\right|\|g\|^2_{L^2(\tilde{M}_j)}\\ &\le C_3C_2\|g\|_{E^2(G)} = C_2C_3\int_{\partial G}|g(z)|^2\,|dz| = C_2C_3\int_{\partial D^*}|\psi(w)|^2\left|\frac{dw}{w^2}\right|\\ &\le C_2C_3\max_{w\in\partial D^*}\left|\frac{1}{w^2}\right|\int_{\partial D^*}|\psi(w)|^2\,|dw| = C_2C_3C_4\|\psi\|_{L^2(\partial D^*)},\end{aligned}$$

and thus

$$\|\psi\|^2_{L^2(M_j)} \le C_2C_3C_4\|\psi\|^2_{E^2(\operatorname{ext}\partial D^*)} \tag{8.24}$$

would be true. In order to prove (8.23) we transform the problem once more by means of a conformal mapping $\varphi : \mathbb{D} \to G$. The function

$$f(z) := g(\varphi(z))[\varphi'(z)]^{1/2}, \qquad z \in \mathbb{D},$$

belongs to H^2 and

$$\|f\|_{H^2} = \|g\|_{E^2(G)}, \qquad \|f\|_{L^2(\varphi^{-1}(\tilde{M}_j))} = \|g\|_{L^2(\tilde{M}_j)}. \tag{8.25}$$

If ε is sufficiently small, the preimage $\varphi^{-1}(\tilde{M}_j) \subset \overline{\mathbb{D}}$ of $\tilde{M}_j$ under the continuous extension of the conformal mapping φ onto $\overline{\mathbb{D}}$ can be projected onto $\partial\mathbb{D}$ from the origin, i.e., it can be parameterized by a function $\eta(e^{i\theta}) := \varrho(\theta)e^{i\theta}$, $\alpha \le \theta \le \beta$, where $\varrho : [\alpha, \beta] \to]0, 1]$ is piecewise C^1. Thus the derivative η' is bounded by some constant $C_5 > 0$. Introducing the maximal function $F : [0, 2\pi] \to \mathbb{R}^+$ by

$$F(\theta) := \sup_{0 \le r < 1} |f(re^{i\theta})|,$$

the Hardy-Littlewood Maximal Theorem [7, Theorem 1.9] asserts the existence of a positive constant C_6 with

$$\|F\|_{L^2([0,2\pi])} \le C_6 \|f\|_{H^2}. \tag{8.26}$$

Putting these facts together we obtain

$$\begin{aligned} \|f\|^2_{L^2(\varphi^{-1}(\tilde{M}_j))} &= \int_{\varphi^{-1}(\tilde{M}_j)} |f(z)|^2 \, |dz| = \int_\alpha^\beta |f(\eta(e^{i\theta})|^2 |\eta'(e^{i\theta})| \, d\theta \\ &\le C_5 \int_0^{2\pi} F(\theta)^2 \, d\theta \le C_5 C_6^2 \|f\|^2_{H^2}. \end{aligned} \tag{8.27}$$

Recalling (8.25) and putting $C_2 := C_6 C_5^{1/2}$ the inequality (8.23) is now established. From (8.22) and (8.24) follows easily the lemma with $C := C_1^{1/2} + (C_2 C_3 C_4)^{1/2}$.

We note that careful study of the proofs in [7, Thm. 1.8, 1.9 and Appendix B] reveals that the estimate (8.26) holds with $C_6 = 8$. However, the estimate

$$\|f\|_{L^2(\varphi^{-1}(\tilde{M}_j))} \le C \|f\|_{H^2}$$

from the last proof (cf. (8.27)) has been shown with $C = \sqrt{2}$ as best possible constant by different means also by Gabriel [10].

As an immediate consequence of the two preceding theorems, we obtain the following

Corollary 1. *Let the origin be contained in the interior of the polygon D. Then the linear operator associating to each function $f \in B_2(D)$ its Borel transform $\psi \in E^2(\operatorname{ext} D^*)$ defines an isomorphism between Banach spaces.*

This fact is also contained in [15, Lemma 3.1], which refers to [14, Anhang 1, § 3] for a proof.

8.4 Growth estimate of Korevaar type

The following generalization of Korevaar's theorem is our main result.

Theorem 5 (Generalization of Korevaar's Theorem). *Let D be a closed convex polygon with $N \geq 2$ vertices and let the origin be a point of D. Let $f \in B_p(D)$, $1 \leq p < \infty$. Then*

$$|f(z)|^p \leq K_p \|f\|^p_{B_p(D)} \cdot \sum_{j=1}^{N} \frac{\exp(p\operatorname{Re} za_{j+1}) - \exp(p\operatorname{Re} za_j)}{p\operatorname{Re} za_{j+1} - p\operatorname{Re} za_j} |a_{j+1} - a_j|.$$

The constant K_p depends on p only: For $1 \leq p \leq 2$, $K_p = \frac{M_p^p}{2\pi}$; for $2 \leq p < \infty$, $K_p = 2^k \frac{M_s^s}{2\pi}$, where $s = \frac{p}{2^k}$ and $2^k < p \leq 2^{k+1}$ for some $k \in \mathbb{N}$, and M_p stands again for the Babenko-Beckner constant.

For the case of the 2-gone, i.e., the interval $D = [-ia, ib]$, $a, b > 0$, our result is a little more general than Korevaar's, since we take into account the different growth behavior in the half plane $\operatorname{Re} z > 0$ and $\operatorname{Re} z < 0$. In fact, let f be an entire function in $B_p([-ia, ib])$. This implies in particular that $f|\mathbb{R} \in L^p(\mathbb{R})$. Then Theorem 5 yields

$$\begin{aligned} |f(z)|^p = |f(x+iy)|^p &\leq K_p \|f\|^p_{B_p(D)} 2 \frac{\exp(p\operatorname{Re}(-zia)) - \exp(p\operatorname{Re}(zib))}{p\operatorname{Re}(-zia) - p\operatorname{Re}(zib)} |-ia - ib| \\ &= K_p \|f\|^p_{B_p(D)} 2 \frac{\exp(pay) - \exp(-pby)}{pay + pby} (a+b) \\ &= K_p \|f\|^p_{B_p(D)} \frac{2}{p} \frac{\exp(pay) - \exp(-pby)}{y}. \end{aligned}$$

This last estimate is sharper than Korevaar's, since

$$\exp(pay) - \exp(-pby) \leq 2 \sinh(p \max\{a, b\} y).$$

Hence, for $p = 2$, with this estimate and the constant K_p given in Theorem 5 we get

$$\begin{aligned} |f(x+iy)|^2 &\leq K_2 \|f\|^2_{B_2(D)} 2 \frac{\sinh(2 \max\{a, b\} y)}{y} \\ &= \frac{1}{2\pi} \|f\|^2_{B_2(D)} \frac{\sinh(2 \max\{a, b\} y)}{y}. \end{aligned} \tag{8.28}$$

This is the result of Korevaar with respect to the $L^2(\mathbb{R})$-norm, since for the special case of a 2-gone $D = [-ia, ib]$ we have $\|f\|_{L^2(\mathbb{R})} = \|f\|_{B_2(D)}$.

Theorem 5 shows that functions bounded in the norm of a Bernstein space allow for pointwise estimates in the whole plane. The geometry of the indicator diagram

determines the size of these estimates. Therefore, the theorem yields a specific growth estimate for every direction.

Proof (Theorem 5). We split the proof in several cases with respect to p.

First we consider the special case $p = 2$. Let $f \in B_2(D)$. According to Theorem 3, f has the representation

$$f(z) = \frac{1}{2\pi i} \int_{\partial D^*} \psi(\lambda) e^{z\lambda}\, d\lambda,$$

where $\psi \in E^2(\operatorname{ext} D^*)$ denotes the Borel transform of f. We can estimate because of the duality (8.14) of inner and outer Smirnov spaces

$$|f(z)| \le \frac{1}{2\pi} \|\psi\|_{E^2(\operatorname{ext} D^*)} \|e^{z\bullet}\|_{E^2(D^*)}. \tag{8.29}$$

We compute the norm

$$\begin{aligned} \|e^{z\bullet}\|^2_{E^2(D^*)} &= \int_{\partial D^*} |e^{z\omega}|^2\, |d\omega| = \sum_{k=1}^{N} \int_{a_k^*}^{a_{k+1}^*} |e^{z\omega}|^2\, |d\omega| \\ &= \sum_{k=1}^{N} \int_0^1 |e^{z(a_k^* + t(a_{k+1}^* - a_k^*))}|^2\, dt \cdot |a_{k+1} - a_k| \\ &= \sum_{k=1}^{N} \frac{e^{2\operatorname{Re} z a_{k+1}^*} - e^{2\operatorname{Re} z a_k^*}}{2\operatorname{Re} z a_{k+1}^* - 2\operatorname{Re} z a_k^*} \cdot |a_{k+1} - a_k|, \end{aligned}$$

where $|d\omega|$ is an element of arc length. Hence we find, using (8.16), i.e., $\|\psi\|_{E^2(D^*)} \le M_2 \sqrt{2\pi}\, |||f|||_{B_p(D)} = \sqrt{2\pi}\, |||f|||_{B_p(D)}$,

$$\begin{aligned} |f(z)| &\le \frac{1}{2\pi} \|\psi\|_{E^2(\operatorname{ext} D^*)} \left(\sum_{k=1}^{N} \frac{e^{2\operatorname{Re} z a_{k+1}^*} - e^{2\operatorname{Re} z a_k^*}}{2\operatorname{Re} z a_{k+1}^* - 2\operatorname{Re} z a_k^*} \cdot |a_{k+1} - a_k| \right)^{\frac{1}{2}} \\ &\le \frac{1}{\sqrt{2\pi}} |||f|||_{B_2(D)} \left(\sum_{k=1}^{N} \frac{e^{2\operatorname{Re} z a_{k+1}^*} - e^{2\operatorname{Re} z a_k^*}}{2\operatorname{Re} z a_{k+1}^* - 2\operatorname{Re} z a_k^*} \cdot |a_{k+1} - a_k| \right)^{\frac{1}{2}}. \end{aligned} \tag{8.30}$$

Now, let $1 < p \le 2$ and let $f \in B_p(D)$. Then by Theorem 3 there exists some $\psi \in E^p(\operatorname{ext} D^*)$ with

$$f(z) = \frac{1}{2\pi i} \int_{\partial D^*} \psi(\lambda) e^{\lambda z}\, d\lambda.$$

By the Minkowski inequality and (8.16),

$$|f(z)| \le \frac{1}{2\pi}\|\psi\|_{L_q(\partial D^*)}\|e^{z^\bullet}\|_{L_p(\partial D^*)} \le \frac{M_p}{2\pi}\sqrt[q]{2\pi}\,|||f|||_{B_p(D)}\|e^{\bullet z}\|_{E^p(\operatorname{ext} D^*)}.$$

Since

$$\|e^{\bullet z}\|^p_{E^p(\operatorname{ext} D^*)} = \sum_{k=1}^{N} \frac{\exp(p\operatorname{Re} za^*_{k+1}) - \exp(p\operatorname{Re} za^*_k)}{p\operatorname{Re} za^*_{k+1} - p\operatorname{Re} za^*_k} \cdot |a_{k+1} - a_k|,$$

it is

$$|f(z)|^p \le \frac{M_p^p}{2\pi}|||f|||^p_{B_p(D)} \sum_{k=1}^{N} \frac{\exp(p\operatorname{Re} za^*_{k+1}) - \exp(p\operatorname{Re} za^*_k)}{p\operatorname{Re} za^*_{k+1} - p\operatorname{Re} za^*_k} \cdot |a_{k+1} - a_k|. \quad (8.31)$$

For $2 \le p < \infty$, let $2^k < p \le 2^{k+1}$, $k \in \mathbb{N}$. If $f \in B_p(D)$ and $s = p/2^k$, then $1 < s \le 2$ and

$$g = (f)^{2^k} \in B_{\frac{p}{2^k}}(2^k D) = B_s(2^k D).$$

From the previous cases we see

$$|g(z)|^s \le \frac{M_s^s}{2\pi}|||g|||^s_{B_s(2^k D)} \sum_{k=1}^{N} \frac{\exp(s\operatorname{Re} z2^k a^*_{k+1}) - \exp(r\operatorname{Re} z2^k a^*_k)}{s\operatorname{Re} z2^k a^*_{k+1} - s\operatorname{Re} z2^k a^*_k} \cdot |2^k a_{k+1} - 2^k a_k|.$$

Thus

$$|f(z)|^p \le \frac{M_s^s}{2\pi}2^k|||f|||^p_{B_p(D)} \sum_{k=1}^{N} \frac{\exp(p\operatorname{Re} za^*_{k+1}) - \exp(p\operatorname{Re} za^*_k)}{p\operatorname{Re} za^*_{k+1} - p\operatorname{Re} za^*_k} \cdot |a_{k+1} - a_k|.$$

Special case $p = 1$: Let $f \in B_1(D)$. Then by [15, Lemma 2.8], we have

$$\lim_{|z|\to\infty} |f(z)| \exp(-|z|k_D(-\arg z)) = 0.$$

For $|z|$ large enough and all $p \ge 1$ we deduce

$$|f(z)|^p \exp(-p|z|k_D(-\arg z)) \le |f(z)| \exp(-|z|k_D(-\arg z)).$$

Hence, from equation (8.31) and the limit $p \to 1$,

$$|f(z)| \le \frac{M_1}{2\pi}|||f|||_{B_1(D)} \sum_{k=1}^{N} \frac{\exp(\operatorname{Re} za^*_{k+1}) - \exp(\operatorname{Re} za^*_k)}{\operatorname{Re} za^*_{k+1} - \operatorname{Re} za^*_k} |a_{k+1} - a_k|.$$

Note that $M_1 = 1$. This concludes the proof.

Note 1. The upper bound in Theorem 5 is never attained. Indeed, let $p = 2$ and assume

$$|f(z_0)| = \frac{1}{\sqrt{2\pi}} \|f\|_{B_2(D)} \left(\sum_{k=1}^{N} \frac{e^{2\operatorname{Re} z_0 a_{k+1}^*} - e^{2\operatorname{Re} z_0 a_k^*}}{2\operatorname{Re} z_0 a_{k+1}^* - 2\operatorname{Re} z_0 a_k^*} \cdot |a_{k+1} - a_k| \right)^{1/2}$$

for some function $f \in B_2(D)$ and some $z_0 \in \mathbb{D}$. Since $E^2(\operatorname{ext} D^*)$ and $E^2(D^*)$ are both subspaces of $L^2(\partial D^*)$ with norms induced by $L^2(\partial D^*)$, we can read (8.29) as Cauchy-Schwarz inequality. There, equality holds if and only if

$$Bf(\lambda) = \psi(\lambda) = c(e^{z_0 \lambda})^*$$

for some constant c. Hence $f(z) = c \cdot \frac{1}{2\pi i} \int_{\partial D^*} e^{\lambda z} e^{(\lambda z_0)^*} \, d\lambda$ and we find

$$\begin{aligned}
f(\xi) &= \frac{c}{2\pi i} \int_{\partial D^*} \exp((z_0 \lambda)^*) \exp(\xi \lambda) \, d\lambda \\
&= \frac{c}{2\pi i} \sum_{j=1}^{N} \int_{a_j^*}^{a_{j+1}^*} \exp(z_0^* \lambda^* + \xi \lambda) \, d\lambda \\
&= \frac{c}{2\pi i} \sum_{j=1}^{N} (a_{j+1}^* - a_j^*) \int_0^1 \exp(z_0^* (a_j + t(a_{j+1} - a_j)) + \xi (a_j^* + t(a_{j+1}^* - a_j^*))) \, dt \\
&= \frac{c}{2\pi i} \sum_{j=1}^{N} (a_{j+1}^* - a_j^*) \frac{\exp(z_0^* a_{j+1} + \xi a_{j+1}^*) - \exp(z_0^* a_j + \xi a_j^*)}{z_0^* (a_{j+1} - a_j) + \xi (a_{j+1}^* - a_j^*)}.
\end{aligned}$$

But the function $\psi(\lambda) = c(e^{z_0} \lambda)^*$ is not holomorphic off D^* and does not vanish at infinity. Therefore, it is not the Borel transform of some function in $B_2(D)$. In particular, $f \notin B_2(D)$. Equality is not attained in $B_2(D)$.

Therefore, the question for best constants in Theorem 5 is an open problem.

8.5 Special case $p = \infty$

For the 2-gone, i.e., an interval $D = [-ia, ib]$, $a, b > 0$, the limit case $p = \infty$ of Korevaar's Theorem 2 is closely related to a result by Phragmén and Lindelöf:

Theorem 6 ([24, p. 82]). *Let f be an entire function of exponential type $B > 0$. If, in addition, f is bounded on the real axis by $|f(x)| \le M$ for all $x \in \mathbb{R}$, then*

$$|f(x + iy)| \le M e^{B|y|}.$$

If, in our case, we assume that g is an entire function of exponential type bounded on the rays $re^{i\theta_j}$, $r > 0, j = 1, \ldots, N$, then g is also bounded in the sectors between the rays. This is due to the more general Phragmén–Lindelöf theorem:

Theorem 7 (Phragmén–Lindelöf). *([24, p. 80]) Let f be continuous on a closed sector of opening π/α and analytic in the open sector. Suppose that on the bounding rays of the sector, $|f(z)| \leq M$.*

If for some $\beta < \alpha$, $|f(z)| \leq \exp(r^\beta)$ whenever z lies in the sector and $|z|$ large enough, then $|f(z)| \leq M$ throughout the sector.

Therefore, the above considered entire function g of exponential type is bounded everywhere and by Liouville's theorem a constant.

If we suppose the weaker conditions that f has a certain exponential growth on the rays,

$$|f(re^{i\theta_j})| \leq M \exp(rh(\theta_j)), \quad r > 0,$$

for constants $h(\theta_j)$ satisfying $k_D(-\theta_j) - h(\theta_j) \geq 0$ for all $j = 1, \ldots, N$, we cannot deduce $f \in B_\infty(D)$. A counterexample is the function

$$f(z) = \sum_{k=1}^{N} e^{a_k z} + e^{(a_1+a_2)z}.$$

Obviously, f satisfies the growth condition. But the conjugate indicator diagram E of f is not contained in D and thus there exists $\theta \in [0, 2\pi]$ with $k_D(-\theta) - h(\theta) < 0$. In this sense, Theorem 7 is best possible.

8.6 Conclusion

In this chapter, we considered variations on the commutative diagram (Fig. 8.1) consisting of the Fourier transform, the Sampling Theorem, and the Paley-Wiener Theorem. We started from a generalization of the Paley-Wiener theorem and considered entire functions with specific growth properties along half lines. Our main result showed that the growth exponents are directly related to the shape of the corresponding indicator diagram, e.g., its side lengths. Since many results from sampling theory are derived with the help from a more general function theoretic point of view (the most prominent example for this is the Paley-Wiener Theorem itself), we believe that a closer examination and understanding of the Bernstein spaces and the corresponding commutative diagrams can — via a a limiting process to the straight-line interval $[-A, A]$ — yield new insights into the $L^p(\mathbb{R})$-sampling theory.

Acknowledgements This work was partially supported by the grant MEXT-CT-2004-013477, Acronym MAMEBIA, of the European Commission and by the DFG grant FO 792/2-1 awarded to Brigitte Forster.

References

1. S.A. Avdonin, S.A. Ivanov, *Families of Exponentials* (Cambridge University Press, Cambridge, 1995)
2. K.I. Babenko, An inequality in the theory of Fourier integrals. Am. Math. Soc. Transl. II. Ser. **44**, 115–128 (1965)
3. W. Beckner, Inequalities in Fourier analysis. Ann. Math. **102**(1), 159–182 (1975)
4. R.P. Boas Jr., Representation for entire functions of exponential type. Ann. Math. 2nd Ser. **39**(2), 269–286 (1938). A correction, Ann. Math. 2nd Ser. **2**(40), 948 (1939)
5. R.P. Boas Jr., Inequalities between series and integrals involving entire functions. J. Indian Math. Soc. (N.S.) **16**, 127–135 (1952)
6. R.P. Boas Jr., *Entire Functions* (Academic, New York, 1954)
7. P.L. Duren, Theory of H^p spaces, in *Pure and Applied Mathematics*, ed. by P.A. Smith, S. Eilenberg. Monographs and Textbooks, vol. 38 (Academic, New York/London, 1970)
8. B. Forster, P. Massopust (eds.), *Four Short Courses on Harmonic Analysis* (Birkhäuser, Basel, 2010)
9. B. Forster, G. Semmler, Growth estimates of Korevaar type for entire functions in generalized Bernstein spaces, in *Proceedings of the SampTA Conference on Sampling Theory and Applications, Singapore (Online Resource)*, 2011
10. R.M. Gabriel, Some results concerning the integrals of moduli of regular functions along certain curves. J. Lond. Math. Soc. **2**, 112–117 (1927)
11. J.R. Higgins, *Sampling Theory in Fourier and Signal Analysis: Foundations* (Oxford University Press, Oxford, 1996)
12. J. Korevaar, An inequality for entire functions of exponential type. Nieuw Arch. Wiskd. **23**, 55–62 (1949)
13. R. Lasser, *Introduction to Fourier Series* (Marcel Dekker, New York, 1996)
14. B.J. Lewin, *Nullstellenverteilung ganzer Funktionen* (Akademie-Verlag, Berlin, 1962) (German)
15. B.J. Lewin, J.I. Ljubarskiĭ, Interpolation by means of special classes of entire functions and related expansions in series of exponentials. Math. USSR Izvestija **9**, 621–662 (1975)
16. L.S. Maergoiz, An Analog of the Paley–Wiener theorem for entire functions of the space W_σ^p, $1 < p < 2$, and some applications. Comput. Methods Funct. Theory **6**(2), 459–469 (2006)
17. F.A. Marvasti, (ed.), *Nonuniform Sampling: Theory and Practice* (Kluwer Academic/Plenum Publishers, New York, 2001)
18. M. Plancherel, G. Pólya, Fonctions entières et intégrales de Fourier multiples. Comment. Math. Helv. **9**, 224–248 (1937) (French)
19. M. Plancherel, G. Pólya, Fonctions entières et intégrales de Fourier multiples (seconde partie). Comment. Math. Helv. **10**, 110–163 (1938) (French)
20. I.I. Privalov, *Randeigenschaften analytischer Funktionen* (Duetschen Verlag der Wissenschaften, Berlin, 1956)
21. E. Raymond, A.C. Paley, N. Wiener, *Fourier Transforms in the Complex Domain*. American Mathematical Society Colloquium Publications, vol. 19 (American Mathematical Society, New York, 1934)
22. A.M. Sedletskii, Bases of exponential functions in the space E^p on convex polygons. Math. USSR Izvestija **13**(2), 387–404 (1979)
23. E.C. Titchmarsh, *Introduction to the Theory of Fourier Integrals* (Oxford University Press, Oxford, 1937)

24. R.M. Young, *An Introduction to Nonharmonic Fourier Series* (Academic, New York, 1980)
25. A. Zygmund, *Trigonometric Series*, vols. I & II, 2nd edn. (Cambridge University Press, Cambridge, 1988) First edition Warschau 1935

Chapter 9
Sampling in Euclidean and Non-Euclidean Domains: A Unified Approach

Stephen D. Casey and Jens Gerlach Christensen

Abstract Sampling theory is a fundamental area of study in harmonic analysis and signal and image processing. The purpose of this paper is to connect sampling theory with the geometry of the signal and its domain. It is relatively easy to demonstrate this connection in Euclidean spaces, but one quickly gets into open problems when the underlying space is not Euclidean. We focus primarily on Euclidean and hyperbolic geometries.

There are numerous motivations for extending sampling to non-Euclidean geometries. Applications of sampling in non-Euclidean geometries are showing up areas from EIT to cosmology. Irregular sampling of bandlimited functions by iteration in hyperbolic space is possible, as shown by Feichtinger and Pesenson. Sampling in spherical geometry has been analyzed by many authors, e.g., Driscoll, Healy, Keiner, Kunis, McEwen, Potts, and Wiaux, and brings up questions about tiling the sphere. In Euclidean space, the minimal sampling rate for Paley-Wiener functions on $\mathbb{R}^d$, the Nyquist rate, is a function of the bandwidth. No such rate has yet been determined for hyperbolic or spherical spaces. We look to develop a structure for the tiling of frequency spaces in both Euclidean and non-Euclidean domains. In particular, we establish *Nyquist tiles* and *sampling groups* in Euclidean geometry, and discuss the extension of these concepts to hyperbolic and spherical geometry and general orientable surfaces.

S.D. Casey (✉)
Department of Mathematics and Statistics, American University, Washington, DC 20016-8050, USA
e-mail: scasey@american.edu

J.G. Christensen
Department of Mathematics, Colgate University, Hamilton, NY 13346, USA
e-mail: jchristensen@colgate.edu

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_9

9.1 Introduction: Nyquist Tiles and Sampling Groups

Sampling Theory is the distinctive branch of mathematics which sets up and solves the interpolation problem of a function with bounded growth from known sampled values. The theory is fundamental in the field of information theory, particularly in telecommunications, signal processing and image processing. Sampling is the process of converting a signal (for example, a function continuous in time or space) into the sample values (a numeric sequence, which is a function of discrete time or space), storing and/or transmitting these values, and then reconstructing the original function when this is required. The theory is a subset of the general theory of interpolation.

The purpose of this paper is to connect sampling theory with the geometry of the signal and its domain. It is relatively easy to demonstrate this connection in Euclidean spaces, but one quickly gets into open problems when the underlying space is not Euclidean. We discuss the extension to hyperbolic and spherical geometry and general orientable surfaces. The establishment of the exact Nyquist rate in non-Euclidean spaces is an open problem. We use two tools to work on the problem – the Beurling-Landau density and Voronoi cells. Using these tools, we establish a relation in Euclidean domains, connecting Beurling-Landau density to sampling lattices and hence dual lattice groups, and then use these dual lattices to define Voronoi cells, which become our tiles in frequency. We then discuss how to extend this to hyperbolic geometry.

There are numerous motivations for extending sampling to non-Euclidean geometries, and in particular, hyperbolic and spherical and geometries. Hyperbolic space and its importance in Electrical Impedance Tomography (EIT) [4, 5] and Network Tomography [6] has been mentioned in several papers of Berenstein et. al. and some methods developed in papers of Kuchment, e.g., [25]. Irregular sampling of bandlimited functions by iteration in hyperbolic space is possible, as shown by Feichtinger and Pesenson [12, 13] and Christensen and Ólafsson [7]. Applications where data are defined inherently on the sphere are found in computer graphics, planetary science, geophysics, quantum chemistry, and astrophysics [9, 28]. In many of these applications, a harmonic analysis of the data is insightful. For example, spherical harmonic analysis has been remarkably successful in cosmology, leading to the emergence of a standard cosmological model [8, 28].

The sphere is compact, and its study requires different tools. Fourier analysis on $\mathbb{S}^2$ amounts to the decomposition of $L^2(\mathbb{S}^2)$ into minimal subspaces invariant under all rotations in $SO(3)$. Bandlimited functions on the sphere are spherical polynomials. Sampling on the sphere is how to sample a band-limited function, an Nth degree spherical polynomial, at a finite number of locations, such that all of the information content of the continuous function is captured. Since the frequency domain of a function on the sphere is discrete, the spherical harmonic coefficients describe the continuous function exactly. A sampling theorem thus describes how to exactly recover the spherical harmonic coefficients of the continuous function from its samples. Developing sampling lattices leads to questions on how to efficiently

tile the sphere, a subject in its own right. We refer to the work of Driscoll and Healy [9], Keiner, Kunis, and Potts [24], and McEwen and Wiaux [28] for results on the sphere.

For Paley-Wiener functions on Euclidean spaces, the minimal sampling rate, the *Nyquist rate*, is a function of the bandwidth. No such rate has yet been determined for hyperbolic or spherical spaces. The establishment of the Nyquist rate in non-Euclidean spaces is an important open question.

The Nyquist rate allows us to develop an efficient tiling of frequency space. A *tiling* or a tessellation of a flat surface is the covering of the plane or region in the plane using one or more geometric shapes, called tiles, with no overlaps and no gaps. This generalizes to higher dimensions. We look to develop *Nyquist tiles* and *sampling groups* for Euclidean, hyperbolic, and spherical spaces. We assume throughout the paper that all signals are single band and symmetric in frequency, i.e., that the transform of the signal can be contained in a simply connected region centered at the origin. Symmetry can be achieved by shifting, and multiband signals can be addressed by the techniques in this paper, but there are techniques to more cleverly deal with multiband signals, e.g., see [22].

The paper is structured as follows. We establish the concepts of Nyquist tiles and sampling groups for Euclidean spaces in this section, and demonstrate their intrinsic relation to sampling. Section 9.2 gives an discussion of the geometry of orientable surfaces, concluding with a discussion of the *Uniformization Theorem*. The Uniformization Theorem gives that all orientable surfaces inherit their intrinsic geometry from their *universal covers*. There are only three of these covers – the plane $\mathbb{C}$ (Euclidean geometry), the Riemann sphere $\tilde{\mathbb{C}}$ (spherical geometry), and the hyperbolic disk $\mathbb{D}$ (hyperbolic geometry). The third section is on Fourier analysis and sampling in hyperbolic space. We develop our analysis in terms of the Fourier-Helgason transform, and discuss results on the Beurling-Landau densities of sampling lattices. We describe two approaches to sampling in hyperbolic space, the first using operator theory, the second Beurling-Landau densities. We then include a discussion of the sphere, and close with a discussion about sampling on general orientable surfaces.

9.1.1 Nyquist Tiling in $\mathbb{R}$

We work with square integrable functions on the real line ($f \in L^2(\mathbb{R})$). References for the material on harmonic analysis and sampling in $\mathbb{R}$ include Benedetto [3], Dym and McKean [10], Grafakos [15], Gröchenig [18], Higgins [22], Hörmander [23], Levin [27], and Young [37].

Fourier series and Fourier transforms are defined as follows (see [3, 10], and [18]). Let f be a periodic, integrable function on $\mathbb{R}$, with period 2Φ, i.e., $f \in L^1(\mathbb{T}_{2\Phi})$. The *Fourier coefficients* of f, $\hat{f}[n]$, are defined by

$$\hat{f}[n] = \frac{1}{2\Phi} \int_{-\Phi}^{\Phi} f(t) \exp(-i\pi nt/\Phi)\, dt\,.$$

If $\{\hat{f}[n]\}$ is absolutely summable ($\{\hat{f}[n]\} \in l^1$), then the *Fourier series* of f is

$$f(t) = \sum_{n\in\mathbb{Z}} \hat{f}[n] \exp(i\pi nt/\Phi)\,.$$

Let f be a function in L^1. The *Fourier transform* of f is defined as

$$\hat{f}(\omega) = \int_{\mathbb{R}} f(t)e^{-2\pi it\omega}dt$$

for $t \in \mathbb{R}$ (time), $\omega \in \hat{\mathbb{R}}$ (frequency). The *Fourier inversion formula*, for $\hat{f} \in L^1(\hat{\mathbb{R}})$, is

$$f(t) = (\hat{f})^{\vee}(t) = \int_{\hat{\mathbb{R}}} \hat{f}(\omega)e^{2\pi i\omega t}d\omega.$$

Formally, we can think of the transform and the coefficient integral as *analysis*, and the inverse transform and series as *synthesis*. The choice to have 2π in the exponent simplifies certain expressions, e.g., for $f, g \in L^1 \cap L^2(\mathbb{R})$, $\hat{f}, \hat{g} \in L^1 \cap L^2(\hat{\mathbb{R}})$, we have the *Parseval-Plancherel* equations – $\|f\|_{L^2(\mathbb{R})} = \|\hat{f}\|_{L^2(\hat{\mathbb{R}})}$ and $\langle f, g\rangle = \langle \hat{f}, \hat{g}\rangle$. Extend the transform from $L^1 \cap L^2$ to L^2 via a density argument. We also define the *periodization* of a function of finite support. Let $T > 0$ and let $f(t)$ be a function such that $\text{supp} f \subseteq [0, T]$. The *T-periodization* of f is $[f]^\circ(t) = \sum_{n=-\infty}^{\infty} f(t - nT)$.

Classical sampling theory applies to functions that are square integrable and band-limited. A function in $L^2(\mathbb{R})$ whose Fourier transform is compactly supported has several smoothness and growth properties given in the Paley-Wiener Theorem (see, e.g., [10, 22, 29, 34, 35]).

Definition 1 (Paley-Wiener Space $\mathbb{PW}_\Omega$).

$$\mathbb{PW}_\Omega = \{f \text{continuous} : f, \hat{f} \in L^2, \text{supp}(\hat{f}) \subset [-\Omega, \Omega]\}$$

Theorem 1 (W-K-S Sampling Theorem). *Let* $f \in \mathbb{PW}_\Omega$, $\underset{T}{\text{sinc}}(t) = \dfrac{\sin(\frac{\pi}{T}t)}{\pi t}$, *and* $\delta_{nT}(t) = \delta(t - nT)$.

1.) *If* $T \le 1/2\Omega$, *then for all* $t \in \mathbb{R}$,

$$f(t) = T\sum_{n\in\mathbb{Z}} f(nT)\frac{\sin(\frac{\pi}{T}(t - nT))}{\pi(t - nT)} = T\left(\left[\sum_{n\in\mathbb{Z}} \delta_{nT}\right] f\right) * \underset{T}{\text{sinc}}(t)\,.$$

2.) *If* $T \le 1/2\Omega$ *and* $f(nT) = 0$ *for all* $n \in \mathbb{Z}$, *then* $f \equiv 0$.

A beautiful way to prove the W-K-S Sampling Theorem is to use the Poisson Summation Formula. Let $T > 0$ and for $f \in L^1([0, T))$, let $[f]^\circ(t) = \sum_{n\in\mathbb{Z}} f(t - nT)$

be the T-periodization of f. We can then expand $[f]^\circ(t)$ in a Fourier series. The sequence of Fourier coefficients of this T-periodic function are given by $\widehat{[f]^\circ}[n] = \frac{1}{T}\hat{f}\left(-\frac{n}{T}\right)$. We have

$$\sum_{n\in\mathbb{Z}} f(t+nT) = \frac{1}{T}\sum_{n\in\mathbb{Z}} \hat{f}(n/T)e^{2\pi int/T}. \tag{PSF1}$$

Therefore

$$\sum_{n\in\mathbb{Z}} f(nT) = \frac{1}{T}\sum_{n\in\mathbb{Z}} \hat{f}(n/T).$$

Thus, the Poisson Summation Formula allows us to compute the Fourier series of $[f]^\circ$ in terms of the Fourier transform of f at equally spaced points. This extends to the Schwartz class of distributions as

$$\widehat{\sum_{n\in\mathbb{Z}} \delta_{nT}} = \frac{1}{T}\sum_{n\in\mathbb{Z}} \delta_{n/T}. \tag{PSF2}$$

If $f \in \mathbb{PW}_\Omega, \hat{f}$ is compactly supported, and we can periodically extend the function. If $T \le 1/2\Omega$,

$$\hat{f}(\omega) = \left(\sum_{n\in\mathbb{Z}} \hat{f}(\omega - \frac{n}{T})\right)\cdot \chi_{[-1/2T,1/2T)}(\omega).$$

But, by computing inverse transforms and applying (PSF2),

$$\hat{f}(\omega) = \left(\sum_{n\in\mathbb{Z}} \hat{f}(\omega - \frac{n}{T})\right)\cdot \chi_{[-1/2T,1/2T)}(\omega) = \left(\sum_{n\in\mathbb{Z}} \left[\delta_{n/T}\right]\hat{f}\right)\cdot \chi_{[-1/2T,1/2T)}(\omega)$$

if and only if

$$f(t) = T\left(\left[\sum_{n\in\mathbb{Z}} \delta_{nT}\right] f\right) * \operatorname*{sinc}_T(t).$$

An additional bonus to this derivation is that it gives a direct method for analyzing reconstruction errors.

There are several errors associated with the sampling reconstruction. Complete reconstruction requires samples over all time. If only a finite number of the samples are used, we get *truncation error*. If sample values are not measured at intended points, we can get *jitter error*. If we have a uniformly spaced sampling set, the main error is *aliasing*. The sampling rate $1/2\Omega$ is called the *Nyquist rate*. Sampling sub-Nyquist results in *aliasing error* $\mathcal{E}_A$, described in the following. If f has bandlimit Ω,

and we sample at rate $T > 1/2\Omega$, high frequencies of one block of $e^{2\pi nt/T} f(t)$ intersect with low frequencies of the next block $e^{2\pi(n+1)t/T} f(t)$. Aliasing results in a stroboscopic effect [3], an effect which is visualized as jumps in the output signal. The high and low frequencies of adjacent blocks alias each other. To analyze aliasing error, we compute the pointwise estimate. For simplicity, assume $f \in \mathbb{PW}_1$. If $T = \frac{1}{2}$, applying (PSF1) and integrating gives us that $f(t)$ equals

$$\int_{-1/2}^{1/2} [\hat{f}]^\circ(\omega) e^{2\pi i t\omega}\, d\omega = \sum_{n\in\mathbb{Z}} f(n) \int_{-1/2}^{1/2} e^{2\pi i(t-n)\omega}\, d\omega = \sum_{n\in\mathbb{Z}} f(n) \frac{\sin(\pi(t-n))}{\pi(t-n)}\,.$$

If $T > \frac{1}{2}$, then

$$\int_{-1/2}^{1/2} [\hat{f}]^\circ(\omega) e^{2\pi i t\omega}\, d\omega = \sum_{n\in\mathbb{Z}} \int_{-1/2}^{1/2} \hat{f}(\omega+n) e^{2\pi i t\omega}\, d\omega$$

$$= \sum_{n\in\mathbb{Z}} \int_{n-1/2}^{n+1/2} \hat{f}(u) e^{2\pi i t(u-n)}\, du = \sum_{n\in\mathbb{Z}} e^{2\pi i t(-n)} \int_{n-1/2}^{n+1/2} \hat{f}(u) e^{2\pi i t u}\, du\,.$$

Now,

$$f(t) = \sum_{n\in\mathbb{Z}} \int_{n-1/2}^{n+1/2} \hat{f}(u) e^{2\pi i t u}\, du\,.$$

Thus,

$$\begin{aligned}
\mathcal{E}_A &= \sup \left| f(t) - \int_{-1/2}^{1/2} [\hat{f}]^\circ(\omega) e^{2\pi i t\omega}\, d\omega \right| \\
&= \sup \left| \sum_{n\neq 0} \left(1 - e^{2\pi i t(-n)}\right) \int_{n-1/2}^{n+1/2} \hat{f}(u) e^{2\pi i t(u)}\, du \right| \\
&\leq 2 \sup \left[\sum_{n\neq 0} \int_{n-1/2}^{n+1/2} |\hat{f}(u)|\, du \right] = 2 \int_{|u|\geq 1/2} |\hat{f}(u)| du\,.
\end{aligned}$$

The constant 2 is sharp. An analysis of this error bound in terms of operators can be found in Chapter 11 of [22].

So, for $f \in \mathbb{PW}_\Omega$, if we sample at exactly Nyquist

$$f(t) = \frac{1}{2\Omega} \left(\left[\sum_{n\in\mathbb{Z}} \delta_{(\frac{1}{2\Omega})} \right] f \right) * \underset{(\frac{1}{2\Omega})}{\operatorname{sinc}}(t)$$

if and only if

$$\hat{f}(\omega) = \left(\sum_{n \in \mathbb{Z}} \hat{f}(\omega - 2n\Omega) \right) \cdot \chi_{[-\Omega,\Omega)}(\omega) = \left(\sum_{n \in \mathbb{Z}} \left[\delta_{2n\Omega} \right] \hat{f} \right) \cdot \chi_{[-\Omega,\Omega)}(\omega) \,.$$

The interval $[-\Omega, \Omega)$ is simply connected and symmetric to the origin. It is spread by the group of translations to form a tiling of frequency space – $\{[(k-1)\Omega, (k+1)\Omega)\}$. We refer to $[-\Omega, \Omega)$ as a *sampling interval*. Note, sampling intervals are "half open, half closed," with length determined by the Nyquist rate. The inverse transform of the characteristic functions of the tiles are sinc functions, which form an orthonormal (o.n.) basis for $\mathbb{PW}_\Omega$. Sampling is expressed in terms of this basis. We can now define the following.

Definition 2 (Nyquist Tiles for $f \in \mathbb{PW}_\Omega$). Let f be a nontrivial function in $\mathbb{PW}_\Omega$. The *Nyquist Tile* $\mathbb{NT}(f)$ for f is the sampling interval of minimal length in $\hat{\mathbb{R}}$ such that $\text{supp}(\hat{f}) \subseteq \mathbb{NT}(f)$. A *Nyquist Tiling* for f is the set of translates $\{\mathbb{NT}(f)_k\}_{k \in \mathbb{Z}}$ of Nyquist tiles which tile $\hat{\mathbb{R}}$.

We are assuming throughout the paper that all signals are single band and symmetric in frequency, i.e., that the transform of the signal can be contained in a simply connected region centered at the origin. Symmetry can be achieved by shifting. For example, consider the function $g(t) = e^{i\pi t} \frac{\sin(\pi t)}{\pi t}$. The Fourier transform is $\hat{g}(\omega) = \chi_{[0,1)}(\omega)$. By modulating the original function g by $e^{-i\pi t}$, we get $f(t) = \frac{\sin(\pi t)}{\pi t}$, whose transform is $\hat{f}(\omega) = \chi_{[-1/2,1/2)}(\omega)$. The Nyquist tile for both g and f is $[-1/2, 1/2)$.

The Nyquist tile is transported by a group of motions to cover the transform domain.

Definition 3 (Sampling Group for $f \in \mathbb{PW}_\Omega$). Let $f \in \mathbb{PW}_\Omega$ with Nyquist Tile $\mathbb{NT}(f)$. The *Sampling Group* $\mathbb{G}(f)$ is a group of translations such that $\mathbb{NT}(f)$ tiles $\hat{\mathbb{R}}$.

The group $\mathbb{G}$ is clearly isomorphic to $\mathbb{Z}$.

9.1.2 Nyquist Tiles and Sampling Groups in $\mathbb{R}^d$

Let $f \in L^1(\mathbb{R}^d)$. We define the *Fourier transform* as

$$\hat{f}(\omega) = \int_{\mathbb{R}^d} f(t) e^{-2\pi i t \cdot \omega} dt$$

for $t \in \mathbb{R}^d$ (time), $\omega \in \hat{\mathbb{R}}^d$ (frequency). The *Fourier inversion formula*, for $\hat{f} \in L^1(\hat{\mathbb{R}}^d)$, is

$$f(t) = (\hat{f})^\vee(t) = \int_{\hat{\mathbb{R}}^d} \hat{f}(\omega) e^{2\pi i \omega \cdot t} d\omega \,.$$

Again, the choice to have 2π in the exponent simplifies certain expressions, e.g., for $f, g \in L^1 \cap L^2(\mathbb{R}^d), \hat{f}, \hat{g} \in L^1 \cap L^2(\widehat{\mathbb{R}^d})$, we have the *Parseval-Plancherel* equations $-\|f\|_{L^2(\mathbb{R}^d)} = \|\hat{f}\|_{L^2(\hat{\mathbb{R}}^d)}$ and $\langle f, g\rangle = \langle \hat{f}, \hat{g}\rangle$. Extend the transform from $L^1 \cap L^2$ to L^2 via a density argument.

We again define the *periodization* of a function of finite support. Let $T > 0$ and let $f(t)$ be a function such that $\text{supp} f \subseteq [0, T]^k$. The *T-periodization* of f is $[f]^\circ(t) = \sum_{n\in\mathbb{Z}^d} f(t - nT)$. We can expand a T-periodic function $[f]^\circ(t)$ in a Fourier series. Denote the lattice $\Lambda = \mathbf{T}\mathbb{Z}^d$, where $\mathbf{T}$ is the $n \times n$ matrix with T on the main diagonal and zeroes elsewhere. The sequence of Fourier coefficients of this periodic function on the lattice $\Lambda = \mathbf{T}\mathbb{Z}^d$ are given by

$$\widehat{[f]^\circ}[n] = \frac{1}{T^d}\hat{f}\left(-\frac{n}{T}\right).$$

We have

$$\sum_{n\in\mathbb{Z}^d} f(t + nT) = \frac{1}{T^d}\sum_{n\in\mathbb{Z}^d} \hat{f}(n/T)e^{2\pi i n\cdot t/T}. \tag{PSF1}$$

Therefore,

$$\sum_{n\in\mathbb{Z}^d} f(nT) = \frac{1}{T^d}\sum_{n\in\mathbb{Z}^d} \hat{f}(n/T).$$

We can write the Poisson summation formula for an arbitrary lattice by a change of coordinates. Let $\mathbf{A}$ be an invertible $d \times d$ matrix, $\Lambda = \mathbf{A}\,\mathbb{Z}^d$, and $\Lambda^\perp = (\mathbf{A}^T)^{-1}\mathbb{Z}^d$ be the dual lattice. Then

$$\begin{aligned}\sum_{\lambda\in\Lambda} f(t + \lambda) &= \sum_{n\in\mathbb{Z}^d} (f \circ \mathbf{A})(\mathbf{A}^{-1}t + n) = \sum_{n\in\mathbb{Z}^d} \widehat{(f \circ \mathbf{A})}(n)e^{2\pi i n\cdot\mathbf{A}^{-1}(t)} \\ &= \frac{1}{|\det\mathbf{A}|}\sum_{n\in\mathbb{Z}^d} \hat{f}((\mathbf{A}^T)^{-1}(n))e^{2\pi i(\mathbf{A}^T)^{-1}(n)\cdot t}.\end{aligned}$$

Note, $|\det\mathbf{A}| = \text{vol}(\Lambda)$. This last expression can be expressed more directly as

$$\sum_{\lambda\in\Lambda} f(t + \lambda) = \frac{1}{\text{vol}(\Lambda)}\sum_{\beta\in\Lambda^\perp} \hat{f}(\beta)e^{2\pi i\beta\cdot t}.$$

This extends again to the Schwartz class of distributions as

$$\widehat{\sum_{\lambda\in\Lambda} \delta_\lambda} = \frac{1}{\text{vol}(\Lambda)}\sum_{\beta\in\Lambda^\perp} \delta_\beta. \tag{PSF2}$$

The sampling formula again follows from computations and an application of (PSF2). We assume a single band signal. Let Λ be a regular sampling lattice in $\mathbb{R}^d$,and let $\Lambda^\perp$ be the dual lattice in $\widehat{\mathbb{R}^d}$. Then Λ has generating vectors $\{\boldsymbol{\tau_1}, \boldsymbol{\tau_2}, \ldots, \boldsymbol{\tau_d}\}$, and the sampling lattice can be written as $\Lambda = \{\lambda : \lambda = z_1\boldsymbol{\tau_1} + z_2\boldsymbol{\tau_2} + \ldots + z_d\boldsymbol{\tau_d}\}$ for $(z_1, z_2, \ldots, z_d) \in \mathbb{Z}^d$. Let $\{\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots, \boldsymbol{\Omega}_d\}$ be the generating vectors for the dual lattice $\Lambda^\perp$. The dual sampling lattice can be written as $\Lambda^\perp = \{\lambda^\perp : \lambda^\perp = z_1\boldsymbol{\Omega}_1 + z_2\boldsymbol{\Omega}_2 + \ldots + z_d\boldsymbol{\Omega}_d\}$ for $(z_1, z_2, \ldots, z_d) \in \mathbb{Z}^d$. The vectors $\{\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots, \boldsymbol{\Omega}_d\}$ generate a parallelepiped. We want to use this parallelepiped to create a tiling, and therefore we make the parallelepiped "half open, half closed" as follows. If we shift the parallelepiped so that one vertex is at the origin, we include all of the boundaries that contain the origin, and exclude the other boundaries. We denote this region as a *sampling parallelepiped* $\boldsymbol{\Omega}_{\mathcal{P}}$.

If the region $\boldsymbol{\Omega}_{\mathcal{P}}$ is a hyper-rectangle, we get the familiar sampling formula

$$f(t) = \frac{1}{\mathrm{vol}(\Lambda)} \sum_{n\in\mathbb{Z}^d} f(\frac{n_1}{\omega_1}, \ldots, \frac{n_d}{\omega_d}) \frac{\sin(\frac{\pi}{\omega_1}(t - n_1\omega_1))}{\pi(t - n_1\omega_1)} \cdot \ldots \cdot \frac{\sin(\frac{\pi}{\omega_d}(t - n_d\omega_d))}{\pi(t - n_d\omega_d)} .$$

If, however, the sampling parallelepiped $\boldsymbol{\Omega}_{\mathcal{P}}$ a general parallelepiped, we first have to compute the inverse Fourier transform of $\chi_{\boldsymbol{\Omega}_{\mathcal{P}}}$. Let $\mathcal{S}$ be the generalized sinc function

$$\mathcal{S} = \frac{1}{\mathrm{vol}(\Lambda)} (\chi_{\boldsymbol{\Omega}_{\mathcal{P}}})^\vee .$$

Then, the sampling formula (see [22]) becomes

$$f(t) = \sum_{\lambda\in\Lambda} f(\lambda)\mathcal{S}(t - \lambda) .$$

Definition 4 (Nyquist Tiles for $f \in \mathbb{PW}_{\boldsymbol{\Omega}_{\mathcal{P}}}$). Let

$$\mathbb{PW}_{\boldsymbol{\Omega}_{\mathcal{P}}} = \{f \text{continuous} : f \in L^2(\mathbb{R}^d), \hat{f} \in L^2(\widehat{\mathbb{R}^d}), \mathrm{supp}(\hat{f}) \subset \boldsymbol{\Omega}_{\mathcal{P}}\} ,$$

where $\{\boldsymbol{\Omega}_1, \boldsymbol{\Omega}_2, \ldots, \boldsymbol{\Omega}_d\}$ be the generating vectors for the dual lattice $\Lambda^\perp$. Let f be a nontrivial function in $\mathbb{PW}_{\boldsymbol{\Omega}_{\mathcal{P}}}$. The *Nyquist Tile* $\mathbb{NT}(f)$ for f is the sampling parallelepiped of minimal area in $\widehat{\mathbb{R}^d}$ centered at the origin such that $\mathrm{supp}(\hat{f}) \subseteq \mathbb{NT}(f)$. A *Nyquist Tiling* is the set of translates $\{\mathbb{NT}(f)_k\}_{k\in\mathbb{Z}^d}$ of Nyquist tiles which tile $\widehat{\mathbb{R}^d}$.

Definition 5 (Sampling Group for $f \in \mathbb{PW}_{\boldsymbol{\Omega}_{\mathcal{P}}}$). Let $f \in \mathbb{PW}_{\boldsymbol{\Omega}_{\mathcal{P}}}$ with *Nyquist Tile* $\mathbb{NT}(f)$. The *Sampling Group* $\mathbb{G}$ is a symmetry group of translations such that $\mathbb{NT}(f)$ tiles $\widehat{\mathbb{R}^d}$.

Remark. Note that the sampling group $\mathbb{G}$ of $f \in \mathbb{PW}_{\boldsymbol{\Omega}_{\mathcal{P}}}$ will be isomorphic to $\mathbb{Z} \oplus \mathbb{Z} \oplus \ldots \oplus \mathbb{Z}$, d-times.

9.1.3 Beurling-Landau Density for Euclidean Space

If sample values are not measured at intended points, we can get jitter error. Let $\{\epsilon_n\}$ denote the error in the nth sample point. First we note that if $f \in \mathbb{PW}_1$, then, by *Kadec's 1/4 Theorem*, the set $\{n \pm \epsilon_n\}_{n\in\mathbb{Z}}$ is a stable sampling set if $|\epsilon_n| < 1/4$. Moreover, this bound is sharp. The sampling set $\Lambda = \{\lambda_k \in \mathbb{R} : |\lambda_k - k| < 1/4\}_{k\in\mathbb{Z}}$ in Kadec's theorem is just a perturbation of $\mathbb{Z}$. For more general sampling sets, the work of Beurling and Landau provide a deep understanding of the one-dimensional theory of nonuniform sampling of band-limited functions.

A sequence Λ is *separated* or *uniformly discrete* if $q = \inf_k(\lambda_{k+l} - \lambda_k) > 0$. The value q is referred to as the *separation constant* of Λ. With a separated sequence Λ we associate a distribution function $n_\Lambda(t)$ defined such that for $a < b$,

$$n_\Lambda(b) - n_\Lambda(a) = \text{card}(\Lambda \cap (a, b]) \,,$$

and normalized such that $n_\Lambda(0) = 0$. There is clearly a one-to-one correspondence between Λ and n_Λ. A discrete set Λ is a *set of sampling* for $\mathbb{PW}_\Omega$ if there exists a constant C such that $\|f\|_2^2 \le C \sum_{\lambda_k\in\Lambda} |f(\lambda_k)|^2$ for every $f \in \mathbb{PW}_\Omega$. The set Λ is called a *set of interpolation* for $\mathbb{PW}_\Omega$ if for every square summable sequence $\{a_\lambda\}_{\lambda\in\Lambda}$, there is a solution $f \in \mathbb{PW}_\Omega$ to $f(\lambda) = a_\lambda$, $\lambda \in \Lambda$. Clearly, all complete interpolating sequences are separated. Landau showed that if Λ is a sampling sequence for $\mathbb{PW}_\Omega$, then there exists constants A and B, independent of a, b such that

$$n_\Lambda(b) - n_\Lambda(a) \ge (b - a) - A \log^+(b - a) - B \,.$$

Definition 6 (Beurling-Landau Densities).

1.) The *Beurling-Landau lower density*

$$D^-(\Lambda) = \liminf_{r\to\infty} \inf_{t\in\mathbb{R}} \frac{(n_\Lambda(t + r)) - n_\Lambda(t)}{r}$$

2.) The *Beurling-Landau upper density*

$$D^+(\Lambda) = \limsup_{r\to\infty} \sup_{t\in\mathbb{R}} \frac{(n_\Lambda(t + r)) - n_\Lambda(t)}{r}$$

The densities are defined similarly in higher dimensions. Specifically, for the exact and stable reconstruction of a bandlimited function f from its samples $\{f(\lambda_k) : \lambda_k \in \Lambda\}$, it is sufficient that the Beurling-Landau lower density satisfies $D^-(\Lambda) > 1$. A set fails to be a sampling set if $D^-(\Lambda) < 1$. Conversely, if f is uniquely and stably determined by its samples on Λ, then $D^-(\Lambda) \ge 1$. Note, a sampling set for which the reconstruction is stable in this sense is called a (stable) set of sampling. This terminology is used to contrast a set of sampling with the weaker notion of a set of uniqueness. Λ is a set of uniqueness for $\mathbb{PW}_\Omega$ if $f|_\Lambda = 0$ implies that $f = 0$.

Whereas a set of sampling for $\mathbb{PW}_{\Omega}$ has a density $D^{-}(\Lambda) \geq 1$, there are sets of uniqueness with arbitrarily small density. We also have that if the Beurling-Landau upper density satisfies $D^{+}(\Lambda) \leq 1$, then Λ is a set of interpolation.

The canonical case is when $\Omega = 2\pi$ and $\Lambda = \mathbb{Z}$. Since $\{e^{int}\}$ in an o.n. basis for $L^2[-\pi, \pi]$, it follows from Parseval that Λ is both a set of sampling and a set of interpolation. This scales by a change of variable, and so $\Lambda = \frac{1}{\Omega}\mathbb{Z}$ is both a set of sampling and a set of interpolation for $\mathbb{PW}_{2\pi\Omega}$. Moreover, general lattices can be compared to the canonical results as follows. If Λ is a set of sampling for $\mathbb{PW}_{2\pi\Omega}$, then Λ is everywhere at least as dense as the lattice $\frac{1}{\Omega}\mathbb{Z}$. If Λ is a set of interpolation for $\mathbb{PW}_{2\pi\Omega}$, then Λ is everywhere at least as sparse as the lattice $\frac{1}{\Omega}\mathbb{Z}$.

This generalizes to $\mathbb{R}^d$. Let $\boldsymbol{\Omega}_{\mathcal{P}}$ be a hyper-rectangle with side lengths Ω. If we normalize the density of $\mathbb{Z}^d$ to be one, then the density of the canonical lattice for $\mathbb{PW}_{2\pi\boldsymbol{\Omega}_{\mathcal{P}}}$ is $1/(2\pi)^d$ times the volume of the spectrum $\boldsymbol{\Omega}_{\mathcal{P}}$. Then, if Λ is a set of sampling for $\mathbb{PW}_{2\pi\boldsymbol{\Omega}_{\mathcal{P}}}$, then Λ is everywhere at least as dense as the lattice $\frac{1}{\Omega^d}\mathbb{Z}^d$. If Λ is a set of interpolation for $\mathbb{PW}_{2\pi\boldsymbol{\Omega}_{\mathcal{P}}}$, then Λ is everywhere at least as sparse as the lattice $\frac{1}{\Omega^d}\mathbb{Z}^d$.

9.1.4 Voronoi Cells for Euclidean Space

We use our sampling lattices to develop *Voronoi cells* corresponding to the sampling lattice. These cells will be, in the Euclidean case, our Nyquist tiles.

Definition 7 (Voronoi Cells in $\widehat{\mathbb{R}^d}$). Let $\hat{\Lambda} = \{\hat{\lambda}_k \in \widehat{\mathbb{R}^d} : k \in \mathbb{N}\}$ be a discrete set in $\widehat{\mathbb{R}^d}$. Then, the Voronoi cells $\{\Phi_k\}$, the Voronoi partition $\mathcal{VP}(\hat{\Lambda})$, and partition norm $\|\mathcal{VP}(\hat{\Lambda})\|$ corresponding to this set are defined as follows. Here, dist is the Euclidean distance.

1.) The *Voronoi cells* $\Phi_k = \{\omega \in \widehat{\mathbb{R}^d} : \text{dist}(\omega, \hat{\lambda}_k) \leq \inf_{j \neq k} \text{dist}(\omega, \hat{\lambda}_j)\}$,
2.) The *Voronoi partition* $\mathcal{VP}(\hat{\Lambda}) = \{\Phi_k \in \widehat{\mathbb{R}^d}\}_{k \in \mathbb{Z}^d}$,
3.) The *partition norm* $\|\mathcal{VP}(\hat{\Lambda})\| = \sup_{k \in \mathbb{Z}^d} \sup_{\omega, \nu \in \Phi_k} \text{dist}(\omega, \nu)$.

Given $f, \hat{f} \in L^2(\mathbb{R}^d)$ such that $f \in \mathbb{PW}_{\boldsymbol{\Omega}_{\mathcal{P}}}$, if the signal is sampled on a lattice exactly at Nyquist, we get a sampling grid $\Lambda = \{\lambda_k \in \mathbb{R}^d\}_{k \in \mathbb{Z}^d}$ that is both a sampling set and a set of interpolation. The Beurling-Landau lower density and the Beurling-Landau upper density are equal for Λ. The dual lattice $\Lambda^{\perp}$ in frequency space can be used to create Voronoi cells $\{\Phi_k\}$, a Voronoi partition $\mathcal{VP}(\Lambda^{\perp})$, and partition norm $\|\mathcal{VP}(\Lambda^{\perp})\|$. If we sample on a lattice exactly at Nyquist, each sample point will correspond to an element in the dual lattice which is at the center of a Nyquist tile $\mathbb{NT}(f)$ for f. The set of Nyquist tiles will cover $\widehat{\mathbb{R}^d}$. If, however, we develop the Voronoi cells $\{\Phi_k\}$ for $\Lambda^{\perp}$, we get $\mathcal{VP}(\Lambda^{\perp}) = \{\Phi_k \in \widehat{\mathbb{R}^d}\}_{k \in \mathbb{Z}^d}$ such that for all k, $\Phi_k = \{\omega \in \widehat{\mathbb{R}^d} : \text{dist}(\omega, \lambda_k^{\perp}) \leq \inf_{j \neq k} \text{dist}(\omega, \lambda_j^{\perp})\}$. But this puts $\lambda_k^{\perp}$ in the center of the cell. Then, if we construct the Voronoi cell containing this point, we will get, up to the boundary, the exact Nyquist tile corresponding to this

point. Nyquist tiles are "half open, half closed." If we shift a Nyquist tile so that one vertex is at the origin, we include all of the boundaries that contain the origin and exclude the other boundaries. To get the exact correspondence between $\mathbb{NT}(f)_k$ and Φ_k, we make Φ_k "half open, half closed" and denote it as $\widetilde{\Phi_k}$. We denote the adjusted Voronoi partition as $\widetilde{\mathcal{VP}}$.

Theorem 2 (Nyquist Tiling for Euclidean Space). *Let f be a nontrivial function in $\mathbb{PW}_{\boldsymbol{\Omega}_{\mathcal{P}}}$, and let $\Lambda = \{\lambda_k \in \mathbb{R}^d\}_{k\in\mathbb{Z}^d}$ be the sampling grid which samples f exactly at Nyquist. Let $\Lambda^\perp$ be the dual lattice in frequency space. Then the adjusted Voronoi partition $\widetilde{\mathcal{VP}}(\Lambda^\perp) = \{\widetilde{\Phi_k} \in \widehat{\mathbb{R}^d}\}_{k\in\mathbb{Z}^d}$ equals the Nyquist Tiling, i.e.,*

$$\{\widetilde{\Phi_k} \in \widehat{\mathbb{R}^d}\}_{k\in\mathbb{Z}^d} = \{\mathbb{NT}(f)_k\}_{k\in\mathbb{Z}^d}\,.$$

Moreover, the partition norm equals the volume of $\Lambda^\perp$, i.e.,

$$\|\widetilde{\mathcal{VP}}(\Lambda^\perp)\| = \sup_{k\in\mathbb{Z}^d}\sup_{\omega,\nu\in\widetilde{\Phi_k}} \operatorname{dist}(\omega,\nu) = \operatorname{vol}(\Lambda^\perp)\,,$$

and the sampling group $\mathbb{G}$ is exactly the group of motions that preserve $\Lambda^\perp$.

This connects, in the Euclidean case, sampling theory with the geometry of the signal and its domain. Given a function $f \in \mathbb{PW}_\Omega$, sampling of such a function is the process of tiling the frequency domain by translated identical copies of the parallelepiped of minimal area, the *Nyquist Tile*, which contains the frequency support of $\hat{f}$. The relation between the geometry and sampling problem in the Euclidean case is as follows: the set of the corresponding translations – the *Sampling Group* – forms a symmetry group. The corresponding sampling set, which is simply the annihilator of the sampling group, is also a symmetry group of translations on $\mathbb{R}^d$. The set of copies of the Nyquist tile, obtained by applying the sampling group, is the *Nyquist Tiling*.

The situation is considerably different when the underlying space is not Euclidean. We quickly get into open problems. Theorem 2 gives an approach for solving the problem in non-Euclidean spaces. We suggest using the two tools we just established – the Beurling-Landau density and Voronoi cells. Our next section discusses the geometry of orientable surfaces. In particular, it provides insight into why a focus on Euclidean, spherical, and, especially, hyperbolic geometries is important.

9.2 Geometry of Surfaces

A surface is a generalization of Euclidean space. From the viewpoint of harmonic analysis, there is a natural interest in both the theory and applications of the study of integrable and square integrable functions on surfaces. This section discusses the geometry of surfaces. Background material for this section can be found in Ahlfors [1, 2], Farkas and Kra [11], Forster [14], Lee [26], and Singer and Thorpe [36].

We assume our surfaces are connected and orientable. Therefore, we can choose a coordinate system so that differential forms are positive [36]. We consider *Riemann surfaces*, but our discussion carries through to connected and orientable Riemannian manifolds of dimension two [26]. Riemann surfaces allow us to discuss the *Uniformization Theorem*, which gives that all orientable surfaces inherit their intrinsic geometry from their *universal coverings*. There are only three universal covers – the plane $\mathbb{C}$ (Euclidean geometry), the Riemann sphere $\tilde{\mathbb{C}}$ (spherical geometry), and the hyperbolic disk $\mathbb{D}$ (hyperbolic geometry).

Recall that a *Jordan curve* Γ is a simple closed continuous path. The *interior* of Γ is the union of all open sets contained inside of Γ. We say that an open set U is *simply connected* if its boundary ∂U is a Jordan curve whose interior contains only points in U.

Klein's Erlangen program sought to characterize and classify the different geometries on the basis of projective geometry and group theory. Since there is a lot of freedom in projective geometry, due to the fact that its properties do not depend on a metric, projective geometry became the unifying frame of all other geometries. Also, group theory provided a useful way to organize and abstract the ideas of symmetry for each geometry. The different geometries need their own appropriate languages for their underlying concepts, since objects like circles and angles were not preserved under projective transformations. Instead, one could talk about the subgroups and normal subgroups created by the different concepts of each geometry and use this to create relations between other geometries. The underlying group structure is the group of isometries under which the geometry is invariant. Isometries are functions that preserve distances and angles of all points in the set. A property of surfaces in $\mathbb{R}^3$ is said to be *intrinsic* if it is preserved by isometry, i.e., if it can be determined from any point on the surface. Isometries can be modeled as the groups of symmetries of the geometry. Thus, the hierarchies of the symmetry groups give a way for us to define the hierarchies of the geometries. We explore the groups of isometries for three geometries – Euclidean, spherical, and hyperbolic. In the next subsection, we present the Uniformization Theorem, which shows that for connected and orientable surfaces, these are the only intrinsic geometries.

The motions that preserve lengths in Euclidean geometry are rotations and translations. Shortest paths, or geodesics, are line segments. Let Γ be a path in $\mathbb{C}$. The *Euclidean length* of Γ is $\mathcal{L}_E(\Gamma) = \int_\Gamma |dz|$. Let $\alpha \in \mathbb{C}$, and let $\varphi_{\theta,\alpha} = e^{i\theta}z + \alpha$. Then $\varphi_{\theta,\alpha}$ preserves the Euclidean length, i.e.,

$$\mathcal{L}_E(\varphi_{\theta,\alpha}(\Gamma)) = \mathcal{L}_E(\Gamma) \ .$$

The motions that preserve lengths in spherical geometry are normalized Möbius maps. Shortest paths, or geodesics, are subarcs of great circles, which are images of the equator of $\tilde{\mathbb{C}}$ under isometries. The metric is weighted by $\lambda(z) = 2/(1 + |z|^2)$. Let Γ be a path on the Riemann sphere $\tilde{\mathbb{C}}$. The *spherical length* of Γ is

$$\mathcal{L}_S(\Gamma) = \int_\Gamma \frac{2\,|dz|}{1+|z|^2}\,.$$

Let $\alpha, \beta \in \mathbb{C}$ and let

$$\varphi_{\alpha,\beta} = \frac{\alpha z + \beta}{-\overline{\beta} z + \overline{\alpha}}\,,$$

where $|\alpha|^2 + |\beta|^2 = 1$. Then $\varphi_{\alpha,\beta}$ preserves the spherical length, i.e.,

$$\mathcal{L}_S(\varphi_{\alpha,\beta}(\Gamma)) = \mathcal{L}_S(\Gamma)\,.$$

The spherical distance δ between two points z_1, z_2 in $\tilde{\mathbb{C}}$ is

$$\delta(z_1, z_2) = \frac{2\,|z_1 - z_2|}{[(1+|z_1|^2)(1+|z_2|^2)]^{1/2}}\,.$$

The motions that preserve lengths in hyperbolic geometry are Möbius-Blaschke maps. They preserve the unit circle $\partial\mathbb{D}$. Shortest paths, or geodesics, are subarcs paths that intersect $\partial\mathbb{D}$ at right angles, which are images of $\mathbb{R} \cap \mathbb{D}$ under isometries. The metric is weighted by $\lambda(z) = 2/(1-|z|^2)$. Let Γ be a smooth path in the unit disk $\mathbb{D}$. The *hyperbolic length* of Γ is

$$\mathcal{L}_H(\Gamma) = \int_\Gamma \frac{2\,|dz|}{1-|z|^2}\,.$$

Let $\alpha \in \mathbb{D}$, and let

$$\varphi_{\theta,\alpha} = e^{i\theta}\,\frac{z-\alpha}{1-\overline{\alpha} z}$$

(a Möbius-Blaschke transformation of $\mathbb{D}$ onto $\mathbb{D}$). Then $\varphi_{\theta,\alpha}$ preserves the hyperbolic length, i.e.,

$$\mathcal{L}_H(\varphi_{\theta,\alpha}(\Gamma)) = \mathcal{L}_H(\Gamma)\,.$$

Let r be a real number, $0 < r < 1$. Assuming that the geodesic with respect to the hyperbolic metric joining 0 to r is the line segment $[-1, 1]$, then the hyperbolic distance ρ between two points z_1, z_2 in $\mathbb{D}$ is

$$\rho(z_1, z_2) = 2\,\operatorname{arctanh}\left(\frac{|z_1 - z_2|}{|1-\overline{z}_2 z_1|}\right) = \log\left(\frac{1+\frac{|z_1-z_2|}{|1-\overline{z}_2 z_1|}}{1-\frac{|z_1-z_2|}{|1-\overline{z}_2 z_1|}}\right)\,.$$

To see these formulae in the hyperbolic case, let Γ be some smooth curve in $\mathbb{D}$, and let $\varphi_{\theta,\alpha}$ be a hyperbolic isometry. Then

$$\mathcal{L}_H(\varphi_{\theta,\alpha}(\Gamma)) = \int_{\varphi_{\theta,\alpha}(\Gamma)} \frac{2\,|dz|}{1-|z|^2} = \int_\Gamma \frac{2\,|\varphi'_{\theta,\alpha}|\,|dz|}{1-|\varphi_{\theta,\alpha}(z)|^2}$$

$$= \int_\Gamma \frac{2\,|e^{i\theta}|\frac{|1-\overline{\alpha}z+\overline{\alpha}(z-\alpha)|}{|1-\overline{\alpha}z|^2}\,|dz|}{1-|e^{i\theta}|^2\frac{|z-\alpha|^2}{|1-\overline{\alpha}z|^2}} . = \int_\Gamma \frac{2\,\frac{|1-|\alpha|^2|}{|1-\overline{\alpha}z|^2}\,|dz|}{\frac{|1-\overline{\alpha}z|^2-|z-\alpha|^2}{|1-\overline{\alpha}z|^2}} .$$

Thus,

$$\mathcal{L}_H(\varphi_{\theta,\alpha}(\Gamma)) = \int_\Gamma \frac{2\,|1-|\alpha|^2|\,|dz|}{|1-\overline{\alpha}z|^2-|z-\alpha|^2}$$

$$= \int_\Gamma \frac{2\,|1-|\alpha|^2|\,|dz|}{(1-\overline{\alpha}z)(1-\alpha\overline{z})-(z-\alpha)(\overline{z}-\overline{\alpha})}$$

$$= \int_\Gamma \frac{2\,(1-|\alpha|^2)\,|dz|}{(1-|\alpha|^2)(1-|z|^2)} = \int_\Gamma \frac{2\,|dz|}{1-|z|^2} = \mathcal{L}_H(\Gamma) .$$

We have

$$\mathcal{L}_H(\Gamma) = \int_\Gamma \frac{2\,|dz|}{1-|z|^2} = \int_\Gamma \frac{|dz|}{1-|z|} + \int_\Gamma \frac{|dz|}{1+|z|} = \log\left(\frac{1+|z|}{1-|z|}\right).$$

To compute hyperbolic distance, let $z_1, z_2 \in \mathbb{D}$, let Γ be the geodesic between z_1 and z_2, and let $\varphi_{\theta,z_1}(\Gamma)$, where θ is chosen so that we rotate z_2 onto the value r on the positive real axis.

$$\mathcal{L}_H(\Gamma) = \mathcal{L}_H(\varphi_{\theta,z_1}(\Gamma)) = \int_{\varphi_{\theta,z_1}(\Gamma)} \frac{2\,|dz|}{1-|z|^2} .$$

Since $\varphi_{\theta,z_1}(\Gamma)$ goes from 0 to r, we have

$$\int_0^r \frac{2\,|dz|}{1-|z|^2} = \log\left(\frac{1+|z|}{1-|z|}\right)\Big|_0^r = \log\left(\frac{1+|r|}{1-|r|}\right).$$

Since $r = e^{i\theta}\frac{z_2-z_1}{1-\overline{z_1}z_2}$,

$$\rho(z_1,z_2) = 2\,\operatorname{arctanh}\left(\frac{|z_1-z_2|}{|1-\overline{z_2}z_1|}\right) = \log\left(\frac{1+\frac{|z_1-z_2|}{|1-\overline{z_2}z_1|}}{1-\frac{|z_1-z_2|}{|1-\overline{z_2}z_1|}}\right) .$$

The metric in $\mathbb{D}$ is derived from the differential $ds_{\mathbb{D}} = \frac{2\,|dz|}{1-|z|^2}$. A parallel development for hyperbolic geometry is in the upper half plane $\mathbb{H} = \{z = x + iy : \mathrm{Im}(z) = y > 0\}$. The corresponding differential in this metric is $ds_{\mathbb{H}} = \frac{|dz|}{\mathrm{Im}(z)}$. The mapping $T(z) = \frac{z-i}{z+i}$ conformally maps $\mathbb{H}$ to $\mathbb{D}$, with

$$\frac{2\,|T'(z)|}{1-|T(z)|^2} = \frac{|dz|}{\mathrm{Im}(z)}$$

for all $z \in \mathbb{H}$, i.e., T is an isometry from $(\mathbb{H}, ds_{\mathbb{H}})$ to $(\mathbb{D}, ds_{\mathbb{D}})$. Some authors use the model $\mathbb{H}$, e.g., [13], while others use $\mathbb{D}$, e.g., [21].

9.2.1 The Uniformization Theorem

The Uniformization Theorem is one of the most important theorems in both the geometry of surfaces and the theory of functions of one complex variable. It plays the same role for Riemann surfaces that the Riemann Mapping Theorem plays for regions in the complex plane $\mathbb{C}$.

We say that two simply connected domains Ω and Δ in $\mathbb{C}$ are *analytically equivalent* if there exists a bijective analytic mapping $\varphi : \Omega \longrightarrow \Delta$. The *Riemann Mapping Theorem* gives the result that if Ω is a simply connected proper subset of $\mathbb{C}$, then Ω is analytically equivalent to the unit disk $\mathbb{D}$. *Riemann surfaces* are generalizations of the complex domain $\mathbb{C}$. The term is used with two different but related meanings. Riemann introduces the concept in his thesis to explain multi-valued analytic functions by letting their domains be multiple copies of the complex plane $\mathbb{C}$. The axiomatic formalization of these leads to *covering spaces*.

Definition 8. Let $\mathcal{S}$ be a connected orientable one-dimensional complex surface. An **atlas** of $\mathcal{S}$ is a collection $\{(U_\alpha, \varphi_\alpha)\}$ on $\mathcal{S}$ such that each U_α is an open set, every $s \in \mathcal{S}$ is contained in some U_α ($\{U_\alpha\}$ forms an **open cover** of $\mathcal{S}$), and

$$\varphi_\alpha : U_\alpha \longrightarrow \mathbb{C}$$

is a one-to-one, onto continuous mapping with a continuous inverse (a **homeomorphism**), mapping U_α onto some open subset of $\mathbb{C}$ such that the **transition functions**

$$f_{\alpha\beta} = \varphi_\alpha \circ \varphi_\beta^{-1} : \varphi_\beta(U_\alpha \cap U_\beta) \longrightarrow \varphi_\alpha(U_\alpha \cap U_\beta)$$

are analytic whenever $U_\alpha \cap U_\beta \neq \emptyset$. Each $(U_\alpha, \varphi_\alpha)$ is referred to a **chart**.

Definition 9. Given a surface $\mathcal{S}$, two atlases are **compatible** if the transition functions between their elements are analytic. We can create a partial ordering of compatible atlases by set containment. By Zorn's Lemma, this collection of partially ordered sets has a maximal element. This *maximal set of charts* of $\mathcal{S}$ will be referred

to as the **maximal atlas** of $\mathcal{S}$ and will be denoted as $\{(U_\alpha^*, \varphi_\alpha)\}$. Then, for this maximal atlas, $\left\langle \mathcal{S}\,,\ \{(U_\alpha^*, \varphi_\alpha)\}_\alpha \right\rangle$ is a **Riemann surface**.

We could also define a Riemann surface without using the maximal atlas. Because we want to discuss uniformization, we will assume, for a given surface, that the atlas is maximal, and we will denote charts without the $*$. Note that the charts are a key component of the surface. For a given α, the pair $(U_\alpha, \varphi_\alpha)$ is also called a *local coordinate*.

Definition 10. Let $\mathcal{S}, \mathcal{T}$ be two Riemann surfaces. A continuous mapping

$$f : \mathcal{S} \longrightarrow \mathcal{T}$$

is called **analytic** if for every local coordinate (U, φ) on $\mathcal{S}$ and every local coordinate (V, ψ) on $\mathcal{T}$ with $U \cap f^{-1}(V) \neq \emptyset$, the mapping

$$\psi \circ f \circ \varphi^{-1} : \varphi(U \cap f^{-1}(V)) \longrightarrow \psi(V)$$

is analytic as a mapping $\mathbb{C} \to \mathbb{C}$. The map is called **conformal** if it is also one-to-one and onto. Two conformally equivalent Riemann surfaces are regarded as equivalent.

Ahlfors [1] efficiently develops the theory of Riemann surfaces using coverings. This idea goes back to Riemann's original idea of a surface, that is, as a way to explain multi-valued analytic functions by letting their domains be multiple copies of the complex plane $\mathbb{C}$. We first define a general covering.

Definition 11. Let X, Y be Hausdorff topological spaces. A **covering** is a continuous, surjective mapping f between X and Y. A covering $f : X \longrightarrow Y$ is said to be **smooth** or **unramified** if f is a local homeomorphism. A covering $f : X \longrightarrow Y$ is said to be **unlimited** if every point of Y possesses a neighborhood U such that the preimage of U under f is a disjoint union of open subsets of X.

Thus, for an unlimited, unramified covering $f : X \longrightarrow Y$, every point in Y is contained in an admissible open neighborhood.

Definition 12. Let $\tilde{\mathcal{S}}, \mathcal{S}$ be two Riemann surfaces, and let $f : \tilde{\mathcal{S}} \longrightarrow \mathcal{S}$ be a covering. Let $s = f(\tilde{s})$. Then, given a local coordinate (U, φ) for $\tilde{s} \in \tilde{\mathcal{S}}$, there exists a local coordinate (V, ψ) for $s \in \mathcal{S}$ such that $\varphi(\tilde{s}) = \psi(s) = 0, f(U) \subset V$, and there exists a natural number n such that f is given locally by the nth power of the complex variable z, i.e.,

$$\psi \circ f \circ \varphi^{-1}(z) = z^n\,,\ z \in \varphi(U)\,.$$

The integer n depends only on the point $\tilde{s}$. If $n > 1$, $\tilde{s}$ is called a **branch point of order** $n - 1$ or a **ramification point of order** n. If $n = 1$ for all $\tilde{s} \in \tilde{\mathcal{S}}$, the cover unramified.

We say that $\tilde{\mathcal{S}}$ is an *unlimited* covering of $\mathcal{S}$ provided that for every curve γ on $\mathcal{S}$ and every $\tilde{\zeta} \in \tilde{\mathcal{S}}$ with $f(\tilde{\zeta}) = \gamma(0)$, there exists a curve $\tilde{\gamma}$ on $\tilde{\mathcal{S}}$ with initial point $\tilde{\zeta}$ and $f(\tilde{\gamma}) = \gamma$. The curve $\tilde{\gamma}$ is called a *lift* of γ. This is generally referred to as the *curve lifting property*, and it follows directly from the unlimited, unramified covering.

Given a point z_0 on a Riemann surface $\mathcal{S}$, we consider all closed curves on $\mathcal{S}$ passing through z_0. We say that any two of these paths are equivalent whenever they are homotopic. The set of these equivalence classes forms a group with the operation of multiplication of equivalence classes of paths. This group is called the *fundamental group of* $\mathcal{S}$ *based at* z_0 and denoted as $\pi_1(\mathcal{S}, z_0)$. Since all Riemann surfaces are connected, given any two points z_0, z_1 on $\mathcal{S}$, the groups $\pi_1(\mathcal{S}, z_0)$ and $\pi_1(\mathcal{S}, z_1)$ are isomorphic. This allows us to refer to the *fundamental group of* $\mathcal{S}$ ($\pi_1(\mathcal{S})$) by picking any base point on $\mathcal{S}$. Note, if $\mathcal{S}$ is simply connected, $\pi_1(\mathcal{S})$ is trivial.

There is an important connection between $\pi_1(\mathcal{S})$ and the smooth unlimited covering spaces $\tilde{\mathcal{S}}$ of $\mathcal{S}$. If $\tilde{\mathcal{S}}$ is a smooth unlimited covering space of $\mathcal{S}$, then $\pi_1(\tilde{\mathcal{S}})$ is isomorphic to a subgroup of $\pi_1(\mathcal{S})$. Conversely, every subgroup of $\pi_1(\mathcal{S})$ determines a smooth unlimited covering corresponding to the space $\tilde{\mathcal{S}}$. Given that the trivial group is a subgroup of every group, the group of $\pi_1(\mathcal{S})$ determines a simply connected smooth unlimited covering space $\tilde{\mathcal{S}}$, which is called the *universal cover*, i.e., the universal covering space is the covering space corresponding to the trivial subgroup of $\pi_1(\mathcal{S})$.

Given connected Riemann surface $\mathcal{S}$ and its universal covering space $\tilde{\mathcal{S}}$, $\mathcal{S}$ is isomorphic to $\tilde{\mathcal{S}}/G$, where the group G is isomorphic to the fundamental group of $\mathcal{S}$, $\pi_1(\mathcal{S})$ (see [14], Section 5). The corresponding universal covering is simply the quotient map which sends every point of $\tilde{\mathcal{S}}$ to its orbit under G. Thus, the fundamental group of $\mathcal{S}$ determines its universal cover. Moreover, the universal covering is indeed the "biggest" smooth unlimited covering of a connected Riemann surface, in the sense that all other unramified unlimited covering space of a Riemann surface can be covered unlimitedly and without ramification by the universal covering of this surface.

The Uniformization Theorem allows us to classify all universal covers of all Riemann surfaces. This in turn allows us to understand the geometry of every Riemann surface. An open Riemann surface is called *hyperbolic* if the maximum principle is not valid. This is equivalent to the existence of a Green's function and a harmonic measure. An open Riemann surface is called *parabolic* if it does not have these properties. Closed Riemann surfaces are *elliptic*.

Theorem 3 (The Uniformization Theorem). *Let $\mathcal{S}$ be a Riemann surface.*

1.) *Every surface admits a Riemannian metric of constant Gaussian curvature κ.*
2.) *Every simply connected Riemann surface is conformally equivalent to one of the following:*

a.) $\mathbb{C}$ *with Euclidean Geometry (parabolic)* – $\kappa = 0$ – *with isometries*

$$\left\langle \left\{ e^{i\theta} z + \alpha \right\} , \circ \right\rangle , \text{ where } \theta \in [0, 2\pi) ,$$

b.) $\tilde{\mathbb{C}}$ *with Spherical Geometry (elliptic)* – $\kappa = 1$ – *with isometries*

$$\left\langle \left\{ \frac{\alpha z + \beta}{-\overline{\beta} z + \overline{\alpha}} \right\} , \circ \right\rangle , \text{ where } |\alpha|^2 + |\beta|^2 = 1 ,$$

c.) $\mathbb{D}$ *with Hyperbolic Geometry (hyperbolic)* – $\kappa = -1$ – *with isometries*

$$\left\langle \left\{ e^{i\theta} \frac{z - \alpha}{1 - \overline{\alpha} z} \right\} , \circ \right\rangle , \text{ where } |\alpha| < 1 \text{ and } \theta \in [0, 2\pi) .$$

Proofs are given in Ahlfors [1], Chapter 10, Forster [14], Section 27, and Farkas and Kra [11], Section IV.6. Ahlfors [1] states the theorem by saying that every simply connected Riemann surface is conformally equivalent to $\mathbb{D}, \mathbb{C}$ or $\hat{\mathbb{C}}$. Also see Table 7.1 on page 214 of Singer and Thorpe's *Lecture Notes on Elementary Topology and Geometry* [36]. Chapter 7 of [36] is on the intrinsic Riemannian geometry of surfaces. They also feature Table 7.1 on the front cover of the book. The discussion in [26, 36] allows us to extend Uniformization to orientable Riemannian manifolds of dimension two.

We finish this section by computing the Gaussian curvature κ of $\tilde{\mathbb{C}}$, $\mathbb{C}$, and $\mathbb{D}$. Gauss' *Theorema Egregium* gives the deep result that κ is intrinsic to every Riemann surface [26]. Moreover, a surface inherits its geometry from the geometry of its universal cover. Given that $\mathcal{S}$ is isomorphic to $\tilde{\mathcal{S}}/G$, where the group G is isomorphic to the fundamental group of $\mathcal{S}$, $\pi_1(\mathcal{S})$, the metric is preserved (see [11], section IV.9). The *Riemannian metrics* for $\tilde{\mathbb{C}}$, $\mathbb{C}$, and $\mathbb{D}$ are $\lambda(z)|dz|$, where $\lambda(z)$ equals

$$\frac{2}{1 + |z|^2} \text{ for } \tilde{\mathbb{C}} , \ 1 \text{ for } \mathbb{C} , \ \frac{2}{1 - |z|^2} \text{ for } \mathbb{D} .$$

The *Gaussian curvature* κ of a surface $\mathcal{S}$ measures the amount of rotation obtained in parallel transporting vectors around small Jordan curves on $\mathcal{S}$. Given the Riemannian metrics for $\mathbb{C}$, $\tilde{\mathbb{C}}$ and $\mathbb{D}$, the curvature is given by

$$\kappa(\lambda) = -\frac{\Delta \log(\lambda)}{\lambda^2} ,$$

where Δ is the Laplacian. The curvatures κ for $\tilde{\mathbb{C}}$, $\mathbb{C}$, and $\mathbb{D}$ are $1, 0, -1$, respectively. To see this, first note that since $\log(1) = 0$, $\kappa = 0$ for $\mathbb{C}$. For $\tilde{\mathbb{C}}$, $\lambda(x, y) = 2/(1 + x^2 + y^2)$. Computing,

$$\frac{\partial^2}{\partial x^2} \lambda(x, y) = \frac{-2 + 2x^2 - 2y^2}{(1 - x^2 - y^2)^2} , \ \frac{\partial^2}{\partial y^2} \lambda(x, y) = \frac{-2 - 2x^2 + 2y^2}{(1 - x^2 - y^2)^2} .$$

Adding gives

$$\Delta \log(\lambda) = \frac{-4}{(1-x^2-y^2)^2}\,.$$

Thus

$$\kappa(\lambda) = -\frac{\Delta \log(\lambda)}{\lambda^2} = 1\,.$$

For $\mathbb{D}$, $\lambda(x,y) = 2/(1-x^2-y^2)$. Computing,

$$\frac{\partial^2}{\partial x^2}\lambda(x,y) = \frac{2+2x^2-2y^2}{(1-x^2-y^2)^2}\,,\quad \frac{\partial^2}{\partial y^2}\lambda(x,y) = \frac{2-2x^2+2y^2}{(1-x^2-y^2)^2}\,.$$

Adding gives

$$\Delta \log(\lambda) = \frac{4}{(1-x^2-y^2)^2}\,.$$

Thus

$$\kappa(\lambda) = -\frac{\Delta \log(\lambda)}{\lambda^2} = -1\,.$$

9.3 Sampling in Hyperbolic Space

We begin by stating the Fourier transform, its inversion, and the Plancherel formula for hyperbolic space [20, 21].

Let dz denote the area measure on the unit disc $\mathbb{D} = \{z \mid |z| < 1$, and let the measure dv be given by then the SU(1, 1)-invariant measure on $\mathbb{D}$ is given by $dv(z) = dz/(1-|z|^2)^2$. For functions $f \in L^1(\mathbb{D}, dv)$ the *Helgason-Fourier transform* is defined as

$$\hat{f}(\lambda, b) = \int_{\mathbb{D}} f(z) e^{(-i\lambda+1)\langle z,b\rangle}\, dv(z)$$

for $\lambda > 0$ and $b \in \mathbb{T}$. Here $\langle z, b\rangle$ denotes the minimal hyperbolic distance from the origin to the horocycle through z and a point $b \in \partial\mathbb{D}$. The mapping $f \mapsto \hat{f}$ extends to an isometry $L^2(\mathbb{D}, dv) \to L^2(\mathbb{R}^+ \times \mathbb{T}, (2\pi)^{-1}\lambda \tanh(\lambda\pi/2) d\lambda\, db)$, i.e., the Plancherel formula becomes

$$\int_{\mathbb{D}} |f(z)|^2 \frac{dz}{(1-|z|^2)^2} = \frac{1}{2\pi}\int_{\mathbb{R}^+\times\mathbb{T}} |\hat{f}(\lambda, b)|^2 \lambda \tanh(\lambda\pi/2) d\lambda\, db.$$

Here db denotes the normalized measure on the circle $\mathbb{T}$, such that $\int_{\mathbb{T}} db = 1$, and $d\lambda$ is Lebesgue measure on $\mathbb{R}$. The *Helgason-Fourier inversion formula* is

$$f(z) = \frac{1}{2\pi} \int_{\mathbb{R}+} \int_{\mathbb{T}} \hat{f}(\lambda, b) e^{(i\lambda+1)\langle z,b\rangle} \lambda \tanh(\lambda\pi/2)\, d\lambda\, db\,.$$

A function $f \in L^2(\mathbb{D}, dv)$ is called *bandlimited* if its Helgason-Fourier transform $\hat{f}$ is supported inside a bounded subset $[0, \Omega]$ of $\mathbb{R}^+$. The collection of bandlimited functions with bandlimit inside a set $[0, \Omega]$ will be denoted $\mathbb{PW}_\Omega = \mathbb{PW}_\Omega(\mathbb{D})$. This definition of bandlimit coincides with the definitions given in [12] and [7] which both show that sampling is possible for bandlimited functions. The Laplacian on $\mathbb{D}$ is symmetric and given by

$$\Delta = (1 - x^2 - y^2)^{-2} \left(\frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2} \right),$$

and we note that

$$\widehat{\Delta f}(\lambda, b) = -(\lambda^2 + 1)\hat{f}(\lambda, b).$$

Therefore, if $f \in \mathbb{PW}_\Omega(\mathbb{D})$, we see that the following Bernstein inequality is satisfied

$$\|\Delta^n f\| \leq (1 + |\Omega|^2)^{n/2} \|f\|.$$

In the following section we will describe sampling results for band-limited functions on hyperbolic space, which, it must be stressed, do not deal with optimal densities.

9.3.1 Sampling via Operator Theory in $\mathbb{D}$

The work in [12] defines bandlimits using the spectrum of the Laplacian on a manifold, while [7] builds on representation theory which for the case at hand gives the explicit form of the Fourier transform on $\mathbb{D}$ as defined above. We also refer to the paper [13] which provides the same results in the setting of the upper half plane (which is bi-holomorphically equivalent to $\mathbb{D}$). These papers build on Neumann series for an operator based on sampling as well as the Bernstein inequality. The sampling operators have previously been explored in [16, 17].

According to Pesenson [30] there is a natural number N such that for any sufficiently small r there are points $x_j \in \mathbb{D}$ for which $B(x_j, r/4)$ are disjoint, $B(x_j, r/2)$ cover $\mathbb{D}$, and $1 \leq \sum_j \chi_{B(x_j,r)} \leq N$. Such a collection of $\{x_j\}$ will be called an (r, N)-lattice.

Let ϕ_j be smooth nonnegative functions which are supported in $B(x_j, r/2)$ and satisfy that $\sum_j \phi_j = 1_{\mathbb{D}}$ and define the operator

$$Tf(x) = P_\Omega \left(\sum_j f(x_j)\phi_j(x) \right),$$

where P_Ω is the orthogonal projection from $L^2(\mathbb{D}, dv)$ onto $\mathbb{PW}_\Omega(\mathbb{D})$. By decreasing r (and thus choosing x_j closer) one can obtain the inequality $\|I - T\| < 1$, in which case T can be inverted by

$$T^{-1}f = \sum_{k=0}^{\infty} (I - T)^k f.$$

For given samples we can calculate Tf, and the Neumann series provides the recursion formula

$$f_{n+1} = f_n + Tf - Tf_n$$

and then $\lim_{n\to\infty} f_n = f$ with norm convergence. The rate of convergence is determined by the estimate $\|f_n - f\| \leq \|I - T\|^{n+1}\|f\|$.

The paper [13] further provides a necessary condition for the set $\{x_i\}$ to be a sampling set. They find that there is a constant C which is determined by the geometry of $\mathbb{D}$, such that if $r < C^{-1}(1 + |\Omega|^2)^{k/2})^{-1}$ for any $k > 1$, then any (N, r)-lattice $\{x_i\}$ is a sampling set. The paper [7] obtains similar results, but removes some restrictions on the functions ϕ_j. In particular the partitions of unity do not need to be smooth and can actually be chosen as characteristic functions $\phi_j = \chi_{U_j}$ for a cover of disjoint sets U_j contained in the balls $B(x_j, r/2)$. This is done by lifting the functions to the group of isometries (which in this case is SU(1, 1)) and by estimating local oscillations using Sobolev norms for left-invariant vector fields on this group.

9.3.2 Beurling density for Bergman spaces

In this section we describe a collection of celebrated sampling theorems for Bergman spaces on the unit disc by Schuster and Seip [31–33]. Let $\mathcal{H}(\mathbb{D})$ be the space of holomorphic functions on $\mathbb{D}$. Let $1 \leq p < \infty$ be given and equip the unit disc $\mathbb{D}$ with normalized area measure $d\sigma(z)$. We define the Bergman space $A^p(\mathbb{D}) = L^p(\mathbb{D}, d\sigma) \cap \mathcal{H}(\mathbb{D})$. This is a reproducing kernel Banach space with reproducing kernel

$$K(z, w) = \frac{1}{(1 - \overline{w}z)^2}.$$

By [32] and [31] sampling and interpolation sets for $A^p(\mathbb{D})$ are characterized by the upper and lower Beurling densities

$$D^+(Z) = \limsup_{r\to 1} \sup_{w\in\mathbb{D}} D(\phi_w(Z), r),$$
$$D^-(Z) = \liminf_{r\to 1} \inf_{w\in\mathbb{D}} D(\phi_w(Z), r).$$

Here $\phi_w(z) = \frac{w-z}{1-\overline{w}z}$ and $D(Z,r) = (\sum_{|z_k|<r} \log|z_k|)/(\log(1-r))$. Let $\rho(z,w) = |\phi_w(z)|$ be the pseudo-hyperbolic distance from z to w, then a sequence $Z = \{z_i\}$ is called uniformly discrete if there is a $\delta > 0$ such that $\rho(z_i, z_j) > \delta$ for $i \neq j$.

Theorem 4. *Let Λ be a set of distinct points in $\mathbb{D}$.*

1.) *A sequence Λ is a set of sampling for A^p if and only if it is a finite union of uniformly discrete sets and it contains a uniformly discrete subsequence Λ' for which $D^-(\Lambda') > 1/p$.*

2.) *A sequence Λ is a set of interpolation for A^p if and only if it is uniformly discrete and $D^+(\Lambda) < 1/p$.*

These results show there can be no Nyquist density for the Bergman spaces, since the sampling sets are always sharply separated from the interpolating sets. We note that the results of Seip and Schuster are for a particular class of holomorphic functions, to which the bandlimited functions $\mathbb{PW}_{\Omega(\mathbb{D})}$ do not belong. It is an open question whether it is possible to establish a Nyquist density for bandlimited functions on $\mathbb{D}$ and to use this information to create regular lattices and dual lattices determined by the size of the bandlimit Ω.

9.3.3 Voronoi Cells and Beurling-Landau Density for $\hat{\mathbb{D}}$

We develop our model for hyperbolic space on the Poincaré disk $\mathbb{D}$. The motions that preserve lengths in hyperbolic geometry are Möbius-Blaschke maps. Geodesics are subarcs of paths that intersect $\partial\mathbb{D}$ at right angles. Let Γ be a smooth path in the unit disk $\mathbb{D}$. The hyperbolic length of Γ is $\mathcal{L}_H(\Gamma) = \int_\Gamma \frac{2\,|dz|}{1-|z|^2}$. Let $\alpha \in \mathbb{D}$, and let $\varphi_{\theta,\alpha} = e^{i\theta}\frac{z-\alpha}{1-\overline{\alpha}z}$ (a Möbius-Blaschke transformation of $\mathbb{D}$ onto $\mathbb{D}$). Then $\varphi_{\theta,\alpha}$ preserves the hyperbolic length, i.e., $\mathcal{L}_H(\varphi_{\theta,\alpha}(\Gamma)) = \mathcal{L}_H(\Gamma)$. The hyperbolic distance ρ between two points z_1, z_2 in $\mathbb{D}$ is

$$\rho(z_1,z_2) = 2\,\operatorname{arctanh}\left(\frac{|z_1-z_2|}{|1-\overline{z_2}z_1|}\right) = \log\left(\frac{1+\frac{|z_1-z_2|}{|1-\overline{z_2}z_1|}}{1-\frac{|z_1-z_2|}{|1-\overline{z_2}z_1|}}\right).$$

The distance ρ will be used to determine distance for the sampling lattice Λ. Note that, because we cannot establish the Beurling-Landau densities, we cannot create regular lattices and dual lattices.

The Helgason-Fourier transform maps $L^2(\mathbb{D})$ to

$$L^2(\mathbb{R}^+ \times \mathbb{T}\,,\ \frac{1}{2\pi}\lambda \tanh(\lambda\pi/2)d\lambda\, db)\,,$$

which is isomorphic to the space of $L^2(\mathbb{T})$-vector-valued square integrable functions with measure $\lambda \tanh(\lambda\pi/2)d\lambda$, in short denoted by

$$L^2(\mathbb{R}^+; L^2(\mathbb{T}), \lambda \tanh(\lambda\pi/2)d\lambda)\,.$$

The negative Laplacian $-\Delta$ is positive with spectrum $\mathbb{R}^+$, and therefore we define Voronoi cells based on a distance on $\mathbb{R}^+$. This distance is denoted dist, and it is an open question in which manner it is related to the measure $\lambda \tanh(\lambda\pi/2)\, d\lambda$. With an appropriate distance function dist, we can define the following.

Definition 13 (Voronoi Cells in $\hat{\mathbb{D}}$). Let $\hat{\Lambda} = \{\hat{\lambda}_k \in \hat{\mathbb{D}} = \mathbb{R}^+ \times \mathbb{T} : k \in \mathbb{N}\}$ be a discrete set in frequency space. Then, the Voronoi cells $\{\Phi_k\}$, the Voronoi partition $\mathcal{VP}(\hat{\Lambda})$, and partition norm $\|\mathcal{VP}(\hat{\Lambda})\|$ corresponding to this set are defined as follows:

1.) The *Voronoi cells* $\Phi_k = \{\omega \in \hat{\mathbb{D}} : \text{dist}(\omega, \hat{\lambda}_k) \leq \inf_{j\neq k} \text{dist}(\omega, \hat{\lambda}_j)\}$,
2.) The *Voronoi partition* $\mathcal{VP}(\hat{\Lambda}) = \{\Phi_k \subseteq \hat{\mathbb{D}}\}_{k\in\mathbb{Z}^d}$,
3.) The *partition norm* $\|\mathcal{VP}(\hat{\Lambda})\| = \sup_{k\in\mathbb{Z}^d} \sup_{\omega,\nu\in\Phi_k} \text{dist}(\omega, \nu)$.

A crucial step in answering the question of Nyquist density using Voronoi cells is to determine an appropriate candidate for the distance on $\hat{\mathbb{D}}$.

9.4 Sampling on the Sphere

One perspective of Fourier analysis is to think of it as a systematic use of symmetry to simplify and understand linear operators. The unit sphere $\mathbb{S}^2$ admits the special orthogonal group of three variables, $SO(3)$ – proper rotations of $\mathbb{R}^3$ about the origin – as a transitive group of symmetries. Fourier analysis on $\mathbb{S}^2$ amounts to the decomposition of $L^2(\mathbb{S}^2)$ into minimal subspaces invariant under all rotations in $SO(3)$. The rotations of the sphere induce operators on functions by rotating the graphs over $\mathbb{S}^2$. The Hilbert space $L^2(\mathbb{S}^2)$ is defined with the usual inner product, using the rotation-invariant area element μ. Background for this section can be found in Driscoll and Healy [9], Keiner, Kunis, and Potts [24], and McEwen and Wiaux [28].

Bandlimited functions on the sphere are spherical polynomials. The corresponding sampling problem is the computation of Fourier coefficients of a function from sampled values at scattered nodes. If we consider the problem of reconstructing a spherical polynomial of degree $N \in \mathbb{N}$ from sample values, one might set up a linear system of equations with $M = (N+1)^2$ interpolation constraints which has to be

solved for the unknown vector of Fourier coefficients $\hat{\mathbf{f}} \in \mathbb{C}^{(N+1)^2}$. If the nodes for interpolation are chosen such that the interpolation problem always has a unique solution, the sampling set is called a fundamental system.

Let $\mathbb{S}^2 = \{x \in \mathbb{R}^3 : \|x\|_2 = 1\}$ be the two-dimensional unit sphere embedded in $\mathbb{R}^3$. A point $\rho \in \mathbb{S}^2$ is identified in spherical coordinates by $\eta = (\sin(\theta)\cos(\phi), \sin(\theta)\sin(\phi), \cos(\theta))^T$, where the angles (θ, ϕ) are the co-latitude and longitude of η. Topologically, $\mathbb{S}^2 = \tilde{\mathbb{C}}$. Geodesics are great circles, and the geodesic distance can be most directly written as

$$\operatorname{dist}(\eta, \xi) = \arccos(\eta \cdot \xi)\,.$$

For $\eta, \xi \in \mathbb{S}^2$, $\|\eta - \xi\|_2^2 = 2 - 2(\eta \cdot \xi)$. The distance to the "north pole" $n = (0, 0, 1)^T$ of $\mathbb{S}^2$ is $\arccos(\eta \cdot n) = \theta$.

The *spherical harmonics* Y_k^n form an o.n. basis for $L^2(\mathbb{S}^2)$. We can define them as follows. The *Legendre polynomials* $P_k : [-1, 1] \longrightarrow \mathbb{R}$ are generated by applying the Gram-Schmidt method to $\{x^k\}_{k=0}^{\infty}$. They are given by the Rodrigues formula $P_k(t) = 1/(2^k k!) d^k/dt^k (t^2 - 1)^k$. The *associated Legendre functions* are defined by

$$P_k^n(t) = \sqrt{\frac{(k-n)!}{(k+n)!}} (t^2 - 1)^{\frac{n}{2}} \frac{d^n}{dt^n} P_k(t)\,.$$

The *spherical harmonics* $Y_k^n : \mathbb{S}^2 \longrightarrow \mathbb{C}$ of degree $k \in \mathbb{N} \cup \{0\}$ and order $n \in \mathbb{Z}$, $|n| \leq k$, are the functions

$$Y_k^n(\eta) = Y_k^n(\theta, \phi) = \sqrt{\frac{2k+1}{4\pi}} P_k^{|n|}(\cos(\theta)) e^{in\phi}\,.$$

We have that

$$\int_0^{2\pi} \int_0^{\pi} Y_k^n(\theta, \phi) Y_l^m(\theta, \phi) \sin(\theta)\, d\theta\, d\phi = \delta_{k,l} \cdot \delta_{m,n}\,,$$

i.e., Y_k^n form an o.n. basis for $L^2(\mathbb{S}^2)$. We say that f is a *spherical polynomial of degree* N if $f(\theta, \phi) = \sum_{k=0}^{N} \sum_{n=-k}^{k} \hat{f}_k^n Y_k^n$. The space of spherical polynomials of degree at most N has dimension $(N+1)^2$.

The Fourier transform is the spherical Fourier matrix

$$f(\theta, \phi) = \sum_{k=0}^{\infty} \sum_{n=-k}^{k} \hat{f}_k^n Y_k^n\,,$$

with coefficients given by

$$\hat{f}_k^n = \int_{\mathbb{S}^2} f \overline{Y_k^n}\, d\mu\,.$$

The dual space of $L^2(\mathbb{S}^2)$ is discrete. The inverse Fourier transform is the construction of a spherical polynomial from the coefficients. The function f is N bandlimited ($N \in \mathbb{N}$) if $\hat{f}_k^n = 0$ for $k > N$. Thus, $f(\theta, \phi) = \sum_{k=0}^{N} \sum_{n=-k}^{k} \hat{f}_k^n Y_k^n$. For the problem of solving for a spherical polynomial f of degree N from sample values, we are looking to solve for the unknown Fourier coefficients $\{\hat{f}_k^n\} = \hat{\mathbf{f}} \in \mathbb{C}^{(N+1)^2}$.

Let $\Lambda = \{\lambda_k\}_{k=1}^{M}$ be a sampling set on $\mathbb{S}^2$. The *mesh norm* δ_Λ and the *separation distance* q_Λ are defined by

$$\delta_\Lambda = 2 \max_{\eta \in \mathbb{S}^2} \min_{k=1,\dots,M} \operatorname{dist}(\eta, \lambda_k) \ , \ q_\Lambda = \min_{j \neq k} \operatorname{dist}(\lambda_j, \lambda_k) \, .$$

A sampling set Λ is called δ *dense* if for some $0 < \delta \leq 2\pi$, $\delta_\Lambda \leq \delta$, and called q *separated* if there exists $0 < q \leq 2\pi$ such that $q_\Lambda \geq q$. We assume that our sampling set is separated. Finally, a sampling set is called quasi-uniform if there exists a constant C independent of the number on sample points M such that $\delta_\Lambda \leq C\, q_\Lambda$.

Sampling on the sphere is how to sample a bandlimited function, an Nth degree spherical polynomial, at a finite number of locations, such that all of the information content of the continuous function is captured. Since the frequency domain of a function on the sphere is discrete, the spherical harmonic coefficients describe the continuous function exactly. A sampling theorem thus describes how to exactly recover the spherical harmonic coefficients of the continuous function from its samples. Given Λ, the *spherical Fourier transform matrix* is

$$\mathbf{Y} = (Y_k^n(\lambda_j))_{j=1,\dots,M;k=0,\dots,N;|n| \leq k} \, .$$

Let $\mathbf{Y^H}$ denote its complex conjugate transpose. The inverse Fourier transform matrix is the construction of a spherical polynomial of degree N from given data points $(\lambda_j, y_j) \in \mathbb{S}^2 \times \mathbb{C}$ such that the identity $f(\lambda_j) = y_j$ is solved. This is solving the linear system $\mathbf{Y}\hat{\mathbf{f}} = \mathbf{y}$, $\mathbf{y} = (y_1, y_2, \dots, y_M)$ for the vector of Fourier coefficients $\hat{\mathbf{f}} = \{\hat{f}_k^n\}$ of the spherical polynomial. Essentially, it is the inverse problem to $f = \mathbf{Y}\hat{\mathbf{f}}$, which corresponds to evaluating a spherical polynomial on Λ.

The open question again is the establishment of the optimal Beurling-Landau densities. This leads to questions about sphere tiling. The papers [24] and [28] address the problem of finding optimal sampling lattices.

9.5 Conclusion

The purpose of this chapter is to connect sampling theory with the geometry of the signal and its domain. We have demonstrated this connection in Euclidean spaces. This chapter gives an outline for an approach to carrying this over to non-Euclidean spaces. We look to get the exact Nyquist rates in both hyperbolic and spherical geometries. We have used two tools to work on the problem – the Beurling-Landau density and Voronoi cells. Given a sampling lattice Λ in either a Euclidean or non-

Euclidean geometry, we can define Voronoi cells using the dual lattice $\Lambda^{\perp}$. These cells then become our tiles in frequency. Working in Euclidean domains, we can connect Beurling-Landau density to sampling lattices and hence the lattice groups, and then using the dual lattices to define Voronoi cells, which become our tiles in frequency. The open questions boil down to the establishment of exact the Beurling-Landau densities for functions in Paley-Wiener spaces in spherical and hyperbolic geometries. These densities are the key to extending sampling to more general settings (see, e.g., [19]). This program can extend to general Riemann surfaces.

9.5.1 Surface Redux

Given connected Riemann surface $\mathcal{S}$ and its universal covering space $\tilde{\mathcal{S}}$, $\mathcal{S}$ is isomorphic to $\tilde{\mathcal{S}}/G$, where the group G is isomorphic to the fundamental group of $\mathcal{S}$, $\pi_1(\mathcal{S})$ (see [14], Section 5). The corresponding universal covering is simply the quotient map which sends every point of $\tilde{\mathcal{S}}$ to its orbit under G. Forster [14] (Section 27) gives the consequences of the Uniformization Theorem very succinctly. The only covering surface of Riemann sphere $\tilde{\mathbb{C}}$ is itself, with the covering map being the identity. The plane $\mathbb{C}$ is the universal covering space of itself, the once punctured plane $\mathbb{C} \setminus \{z_0\}$ (with covering map $\exp(z - z_0)$), and all tori $\mathbb{C}/\Gamma$, where Γ is a parallelogram generated by $z \longmapsto z + n\gamma_1 + m\gamma_2$, $n, m \in \mathbb{Z}$ and γ_1, γ_2 are two fixed complex numbers linearly independent over $\mathbb{R}$. *The universal covering space of every other Riemann surface is the hyperbolic disk* $\mathbb{D}$. Therefore, the establishment of exact the Beurling-Landau densities for functions in Paley-Wiener spaces in spherical and especially hyperbolic geometries will allow the development of sampling schemes on arbitrary Riemann surfaces.

Acknowledgements The authors would like to thank the referees for their valuable input. First author's research was partially supported by US Army Research Office Scientific Services program, administered by Battelle (TCN 06150, Contract DAAD19-02-D-0001) and US Air Force Office of Scientific Research Grant Number FA9550-12-1-0430.

References

1. L.V. Ahlfors, *Conformal Invariants* (McGraw-Hill, New York, 1973)
2. L.V. Ahlfors, *Complex Analysis*, 3rd edn. (McGraw-Hill, New York, 1979)
3. J.J. Benedetto, *Harmonic Analysis and Applications* (CRC Press, Boca Raton, FL, 1997)
4. C.A. Berenstein, Local tomography and related problems, in *Radon Transforms and Tomography*, ed. by E.T. Quinto, L. Ehrenpreis, A. Faridani, F. Gonzalez, E. Grinberg. Contemporary Mathematics, vol. 278 (American Mathematical Society, Providence, RI, 2001), pp. 3–14
5. C.A. Berenstein, E.C. Tarabusi, Integral geometry in hyperbolic spaces and electrical impedance tomography. SIAM J. Appl. Math. **56**(3), 755–764 (1996)
6. C.A. Berenstein, F. Gavilánez, J. Baras, *Network Tomography*. Contemporary Mathematics, vol. 405 (American Mathematical Society, Providence, RI, 2006), pp. 11–17

7. J.G. Christensen, G. Ólafsson, Sampling in spaces of bandlimited functions on commutative spaces, in *Excursions in Harmonic Analysis*, ed. by T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, K.A. Okoudjou. Applied and Numerical Harmonic Analysis, vol. 1 (Birkhäuser/Springer, New York, 2013), pp. 35–69
8. N.J. Cornish, J.R. Weeks, Measuring the shape of the universe. Not. Am. Math. Soc. **45**(11), 1463–1471 (1998)
9. J.R. Driscoll, D.M. Healy, Computing Fourier transforms and convolutions on the 2-sphere. Adv. Appl. Math. **15**(2), 202–250 (1994)
10. H. Dym, H.P. McKean, *Fourier Series and Integrals* (Academic, Orlando, FL, 1972)
11. H.M. Farkas, I. Kra, *Riemann Surfaces* (Springer, New York, 1980)
12. H. Feichtinger, I. Pesenson, Recovery of band-limited functions on manifolds by an iterative algorithm, in *Wavelets, Frames and Operator Theory*, ed. by C. Heil, P.E.T. Jorgensen, D.R. Larson. Contemporary Mathematics, vol. 345 (American Mathematical Society, Providence, RI, 2004), pp. 137–152
13. H. Feichtinger, I. Pesenson, A reconstruction method for band-limited signals in the hyperbolic plane. Sampling Theory Signal Image Process. **4**(3), 107–119 (2005)
14. O. Forster, *Lectures on Riemann Surfaces* (Springer, New York, 1981)
15. L. Grafakos, *Classical and Modern Fourier Analysis* (Pearson Education, Upper Saddle River, NJ, 2004)
16. K. Gröchenig, Describing functions: atomic decompositions versus frames. Monatsh. Math. **112**(1), 1–42 (1991)
17. K. Gröchenig, Reconstruction algorithms in irregular sampling. Math. Comput. **59**(199), 181–194 (1992)
18. K. Gröchenig, *Foundations of Time-Frequency Analysis* (Birkhäuser, Boston, 2000)
19. K. Gröchenig, G. Kutyniok, K. Seip, Landau's necessary density conditions for LCA groups. J. Funct. Anal. **255**, 1831–1850 (2008)
20. S. Helgason, *Geometric Analysis on Symmetric Spaces* (American Mathematical Society, Providence, RI, 1994)
21. S. Helgason, *Groups and Geometric Analysis* (American Mathematical Society, Providence, RI, 2000)
22. J.R. Higgins, *Sampling Theory in Fourier and Signal Analysis: Foundations* (Clarendon Press, Oxford, 1996)
23. L. Hörmander, *The Analysis of Linear Partial Differential Operators I (Distribution Theory and Fourier Analysis)*, 2nd edn. (Springer, New York, 1990)
24. J. Keiner, S. Kunis, D. Potts, Efficient reconstruction of functions on the sphere from scattered data. J. Fourier Anal. App. **13**(4), 435–458 (2007)
25. P. Kuchment, Generalized transforms of Radon type and their applications, in *Proceedings of Symposia in Applied Mathematics*, vol. 63 (American Mathematical Society, Providence, RI, 2006), pp. 67–98
26. J.M. Lee, *Riemannian Manifolds: An Introduction to Curvature* (Springer, New York, 1997)
27. B.Y. Levin, *Lectures on Entire Functions* (American Mathematical Society, Providence, RI, 1996)
28. J.D. McEwen, E. Wiaux, A novel sampling theorem on the sphere. IEEE Trans. Signal Process. **59**(12), 617–644 (2011)
29. H. Nyquist, Certain topics in telegraph transmission theory. AIEE Trans. **47**, 617–644 (1928)
30. I. Pesenson, A sampling theorem of homogeneous manifolds. Trans. Am. Math. Soc. **352**(9), 4257–4269 (2000)
31. A.P. Schuster, Sets of sampling and interpolation in Bergman spaces. Proc. Am. Math. Soc. **125**(6), 1717–1725 (1997)

32. K. Seip, Beurling type density theorems in the unit disk. Invent. Math. **113**, 21–39 (1993)
33. K. Seip, Regular sets of sampling and interpolation for weighted Bergman spaces. Proc. Am. Math. Soc. **117**(1), 213–220 (1993)
34. C.E. Shannon, A mathematical theory of communication. Bell Syst. Tech. J. **27**, 379–423 (1948)
35. C.E. Shannon, Communications in the presence of noise. Proc. IRE. **37**, 10–21 (1949)
36. I.M. Singer, J.A. Thorpe, *Lecture Notes on Elementary Topology and Geometry* (Springer, New York, 1967)
37. R. Young, *An Introduction to Nonharmonic Fourier Series* (Academic, New York, 1980)

Chapter 10
A Sheaf-Theoretic Perspective on Sampling

Michael Robinson

Abstract Sampling theory has traditionally drawn tools from functional and complex analysis. Past successes, such as the Shannon-Nyquist theorem and recent advances in frame theory, have relied heavily on the application of geometry and analysis. The reliance on geometry and analysis means that these results are dependent on the symmetries of the space of samples. There is a subtle interplay between the topology of the domain of the functions being sampled, and the class of functions themselves. Bandlimited functions are somewhat limiting; often one wishes to sample from other classes of functions. The correct topological tool for modeling all of these situations is the *sheaf*; a tool which allows local structure and consistency to derive global inferences. This chapter develops a general sampling theory for sheaves using the language of exact sequences, recovering the Shannon-Nyquist theorem as a special case. It presents sheaf-theoretic approach by solving several different sampling problems involving non-bandlimited functions. The solution to these problems show that the topology of the domain has a varying level of importance depending on the class of functions and the specific sampling question being studied.

10.1 Introduction

Sampling theory has traditionally drawn tools from functional and complex analysis. Past successes, such as the Shannon-Nyquist theorem and recent advances in frame theory, have relied heavily on the application of geometry and analysis. The traditional perspective is that reconstruction of functions from samples relies on an appropriate notion of bandlimitedness. This chapter advances a complementary perspective that the topology of the underlying space can have a strong impact as well.

The reliance on geometry and analysis means that the results are usually dependent on symmetries of the space of samples. For instance, the space of samples

M. Robinson (✉)
American University, 4400 Massachusetts Ave NW, Washington, DC 20016, USA
e-mail: michaelr@american.edu

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_10

in uniform sampling has a particular translation invariance. Nonuniform sampling breaks this translation invariance, thereby permitting it to exceed the performance of uniform sampling. This fuller expressive power and generality comes at a cost, in that the theory required for nonuniform sampling is much more intricate. Topological methods offer a different perspective by reducing the dependence on symmetries and by providing a more flexible framework in which to pose sampling and reconstruction questions.

There is evidence that topology has an important – and largely unexplored – impact on sampling problems. For instance, the space of bandlimited functions associated to the Laplace-Beltrami operator on the real line with the usual metric is infinite-dimensional, while the space of bandlimited functions over a compact subset of the real line is finite dimensional. There is a subtle interplay between the topology of the domain of the functions being sampled and the class of functions themselves. The correct algebraic tool for modeling all of these situations is the *sheaf*; a tool which is sensitive to the topology of the domain and allows local structure to derive global inferences.

Sheaves permit greater generality in specifying sampling procedures and provide more general conditions for reconstruction than otherwise possible. They highlight both the importance of local control in reconstruction as well as the importance of topology in sampling. Most sampling problems that have been studied in the literature assume that samples are scalar valued and are collected in a geometrically aware fashion. Sheaves formalize and generalize the sampling process, allowing each sample to be vector valued, of different dimensions, and located arbitrarily. In this way, sheaves gracefully permit the study of functions over general topological spaces using samples of varying types, richness, and rates. This generality suggests that sheaves are the appropriate tool for unifying traditional notions of bandlimitedness with topological aspects of sampling.

In order to build a sampling theory using sheaves, it is necessary to specify the *base space* that serves as the domain of functions to be sampled. Both the topology and the combinatorial structure of the base space are important for specifying practical examples of sampling. Sheaves which specify the function space and the sampling procedure are written over the space. Once specified, these sheaves will be analyzed using exact sequences for cohomology, which lead to the most general conditions for reconstruction.

This chapter makes several contributions to sampling theory. It discusses the use of sheaves in sampling and reconstruction through a three-part procedure:

1. Represent the appropriate function spaces as sheaves,
2. Construct a sampling morphism, and
3. Compute the cohomology of the ambiguity sheaf.

This procedure leads to a general sampling theorem for sheaves using the language of exact sequences. The Shannon-Nyquist theorem is a special case of this more general sampling theorem. This chapter shows how a sheaf-theoretic approach emphasizes the impact of the topology of the domain by solving three different sampling problems:

1. *Bandlimited functions on the real line*, in which reconstruction is global. Topology strongly impacts the number of samples required: if we instead consider bandlimited functions on a compact space, we obtain finite Fourier series. (The sampling *rate* is unchanged, however.)
2. *Quantum graphs*, in which reconstruction is somewhat local. Sometimes non-trivial topology in the domain is detected, sometimes not.
3. *Splines* written over a coarse topological space describing a fixed knot sequence, in which there remain only local constraints on the functions. Topology plays almost no further role in the reconstruction of splines from their samples.

10.1.1 Historical context

Sampling theory has a long and storied history, about which a number of recent survey articles [3, 13, 38, 40] have been written. Since sampling plays an important role in applications, substantial effort has been expended on practical algorithms. Our approach is topologically motivated, which is similar to other topological approaches to sampling (for instance [6, 23]) in that it is not constrained by specific timing constraints. Relaxed timing constraints are an important feature of bandpass [42] and multirate [41] algorithms.

Some signals have local or partially local control, of which splines [39] are an excellent example. There are several subtly different perspectives on splines, which can be characterized by their underlying knot sequences. If the knot sequences are fixed – as is common in the computer graphics literature [12] – the underlying topology can be quite coarse. The resulting splines exhibit strictly local behavior. If the knot sequence is allowed to vary over a Riemannian manifold or a stratified space, then splines can reflect both the global and local topology of the space [25, 27, 30]. In order to emphasize the impact of the topology of the underlying space, this chapter will discuss splines with a fixed knot sequence.

Sheaf theory has not been used in applications until fairly recently. The catalyst for new applications was the technical tool of *cellular sheaves*, developed in [37]. Since that time, an applied sheaf theory literature has emerged, for instance [8, 15, 22, 32, 33].

Our sheaf-theoretic approach has sufficient generality to treat sampling on non-Euclidean spaces. Others have also studied sampling on non-Euclidean spaces, for instance general Hilbert spaces [24], Riemann surfaces [36], symmetric spaces [10], the hyperbolic plane [14], combinatorial graphs [31], and quantum graphs [26, 28]. Each of these methods are specific to a particular kind of space; sheaves provide unified sufficiency conditions for perfect reconstruction on abstract simplicial complexes, which encompass all of the above cases.

A large class of local signals are those with *finite rate of innovation* [17, 43]. Our ambiguity sheaf is a generalization of the Strang-Fix conditions as identified in [9]. With our approach, one can additionally consider reconstruction using richer samples.

10.2 A unifying example

Vector spaces of functions such as $C^k(U, \mathbb{C})$ are rather global in nature – an element of such a space *is* a function. In contrast, evaluating a function at a particular point x corresponds to a linear transformation that is only sensitive to a function's value at or near x. Function evaluation is a local process, so without further knowledge of the type of function being sampled, a single sample is a weak constraint. Even reconstructing a function from a discrete collection of samples therefore appears counterintuitive.

The local sampling versus global reconstruction paradox is resolved because reconstruction theorems only exist for certain suitably constrained vector spaces. An extreme example of reconstruction is that of the Taylor series of a holomorphic function. Evaluating such a function and all its derivatives at a point determines its value anywhere in a connected component of its domain. Analytic continuation is therefore a very strong kind of reconstruction from a single sample. Analytic continuation relies both on (1) a restricted space of functions (merely smooth functions do not suffice) and (2) a rather large amount of information at the sample point (not just the value of the function but also all of its derivatives). These two constraints are essential to understand the nature of reconstruction from samples, so the admittedly special case of analytic continuation is informative.

Consider the space of holomorphic functions $C^\omega(U, \mathbb{C})$ on a connected open set $U \subseteq \mathbb{C}$. Without loss of generality, suppose that U contains the origin. Then the function $a : C^\omega(U, \mathbb{C}) \to l^1$ given by

$$a(f) = \left(f(0), f'(0), \ldots, \frac{f^{(n)}(0)}{n!}, \ldots \right)$$

for $f \in C^\omega(U, \mathbb{C})$ is a linear transformation. Because a computes the Taylor series of f, whenever $a(f) = a(g)$ it must follow that $f = g$ on U. This means that as a linear transformation, the sampling function a has a trivial kernel.

Conversely, the trivial kernel of a witnesses the fact that the original $f \in C^\omega(U, \mathbb{C})$ can be recovered from the sampled value $a(f)$. This is by no means necessarily true for all sampling functions. For instance, the sampling function $b : C^\omega(U, \mathbb{C}) \to l^1$ given by

$$b(f) = \left(f'(0), \ldots, \frac{f^{(n)}(0)}{n!}, \ldots \right)$$

has a one-dimensional kernel. This means that reconstruction of an analytic function from its image through b is ambiguous – it is known only up to the addition of a constant. Since the kernel of b is a subspace of $C^\omega(U, \mathbb{C})$, there is more information available than merely its dimension. If we restrict the domain of b to be the subspace $Z \subset C^\omega(U, \mathbb{C})$ of analytic functions whose value at the origin is zero, then the

intersection $Z \cap \ker b$ is trivial. Reconstruction succeeds on Z using b for sampling even though using b on its whole domain is ambiguous.

Sampling a function in $C^\omega(U, \mathbb{C})$ can be represented generally as a function $s : C^\omega(U, \mathbb{C}) \to l^1$. Observe that s could take the form of the functions a or b above, in which a function is evaluated in the immediate vicinity of a single point. However, s could also be given by

$$s(f) = (\ldots, f(-1), f(0), f(1), \ldots)$$

or by many other possibilities. In this general setting, the simplest way to determine whether reconstruction is ambiguous is to examine $\ker s$. Recognizing that we may wish to restrict the class of functions under discussion, it is useful to understand how the subspace $\ker s$ is included within $C^\omega(U, \mathbb{C})$. This situation can also be described as the following *exact sequence* of linear functions

$$0 \to A \xrightarrow{\ i\ } C^\omega(U, \mathbb{C}) \xrightarrow{\ s\ } l^1 \to 0,$$

which means that $\ker s = \text{image}\, i$. Observe that the zero at the beginning of the sequence indicates that the map i is injective so that $A = \ker s$. Likewise, the zero at the end of the sequence indicates that s is surjective.

If $f \in C^\omega(U, \mathbb{C})$ is fixed (but unknown) and $s(f)$ is known, then clearly f can only be known to be one of the preimages $i^{-1}(f) \subseteq A$. If there is only one preimage (as in the case of a above), then reconstruction is said to be *unambiguous*.

This example contrasts sharply with the situation of sampling data from the space of *all* smooth functions. The analytic functions are a subset $C^\omega(U, \mathbb{C}) \subset C^\infty(U, \mathbb{C})$, so if s' represents s extended to the space of all continuous functions, one has a diagram like

$$0 \to A' \xrightarrow{\ i\ } C^\infty(U, \mathbb{C}) \xrightarrow{\ s'\ } l^1 \to 0.$$

A' is quite large when compared to the case of sampling analytic functions using the function a above. This can be expressed diagrammatically as

$$\begin{array}{ccccccccc}
 & & & & 0 & & 0 & & \\
 & & & & \downarrow & & \downarrow & & \\
 & & 0 & \xrightarrow{\ i\ } & C^\omega(U, \mathbb{C}) & \xrightarrow{\ a\ } & l^1 & \longrightarrow & 0 \\
 & & \downarrow & & \downarrow & & \downarrow & & \\
0 & \longrightarrow & A' & \xrightarrow{\ i\ } & C^\infty(U, \mathbb{C}) & \xrightarrow{\ s'\ } & l^1 & \longrightarrow & 0
\end{array}$$

in which all possible compositions of linear maps with the same domain and codomain are equal. This shows how the two classes of functions and their samples are related, and the technique will be used in later sections as a kind of algebraic bound.

Although the example of sampling analytic functions will be generalized considerably throughout the rest of the chapter, the overall structure of constructing an exact sequence will remain. We begin by replacing the vector spaces $C^\omega(U, \mathbb{C})$ and l^1 with sheaves and replacing the linear sampling map a with a sheaf morphism. This allows us to constrain the domain of influence of an individual sample. This refined control is not usually visible in a function space, since the topology of the domain of the function to be sampled is hidden. Because the local degrees of freedom of the function and the samples can vary over the base space (heterogeneous sampling) it is useful to place maximum and minimum bounds on the amount of information available. This generalizes the diagrammatic construction of sampling for $C^\omega(U, \mathbb{C}) \subset C^\infty(U, \mathbb{C})$ that was given above.

The effect of the domain's topology is easily identified using sheaves and is visible through the size of the ambiguity space A. Although it leads to weaker invariants, it is usually easier to compute the size of the ambiguity space A instead of the kernel of the sampling map a. The most precise sampling conditions come from constraints on the cohomology of A. This leads to something unanticipated by the example in this section – the cohomology of A separates the influence of ambiguity from the influence of redundancy on reconstruction.

10.3 Local data

This section formalizes the intuition in the previous section, by showing how sheaves are the correct mathematical formalism for discussing local information. From this formal structure, general sampling reconstruction theorems can be proven which place bounds on necessary and sufficient sampling rates.

Section 10.3.1 distills an axiomatic framework that precisely characterizes what "local" means. Section 10.3.2 defines the cohomology functor for sheaves, which assembles this local information into global information. With the definition of a *sheaf morphism* in Section 10.3.3, these tools allow the statement of general conditions under which a sampling suffices to reconstruct a function in a particular space in Section 10.4.

10.3.1 Sheaves represent local data

A local model of data should be flexible enough to capture both analytic and non-analytic functions. Because portions of the data in one region will not necessarily be related to those farther away, the model should allow us to infer global effects only when they are appropriate to the kind of function under study.

Spaces of continuous functions exhibit several properties related to locality. As a concrete example, consider the following properties of $C^k(U, \mathbb{C})$ when $k \geq 0$:

1. *Restriction*: Whenever $V \subseteq U$ are open sets, there is a linear map $C^k(U, \mathbb{C}) \to C^k(V, \mathbb{C})$ that is given by restricting the domain of a function defined on U to one defined on V.
2. *Uniqueness*: Whenever a function is the zero function on some open set, then all of its restrictions are zero functions also. The converse is true also: suppose $f \in C^k(V, \mathbb{C})$ and that $\{U_1, \dots\}$ is an open cover[1] of V. If the restriction of f to each U_k is the zero function on U_k, then f has to be the zero function on V.
3. *Gluing*: If U and V are open sets and $f \in C^k(U, \mathbb{C})$, $g \in C^k(V, \mathbb{C})$ then whenever $f(x) = g(x)$ for all $x \in U \cap V$ there is a function $h \in C^k(U \cup V, \mathbb{C})$ that restricts to f and g.

The gluing property provides a condition by which local information (the elements $f \in C^k(U, \mathbb{C})$, $g \in C^k(V, \mathbb{C})$) can be assembled into global information in $C^k(U \cup V, \mathbb{C})$, provided a consistency condition is met. We will call this specification of f and g a *section* when they restrict to the same element in $C^k(U \cap V, \mathbb{C})$. This can also be illustrated diagrammatically as shown in Figure 10.1, where the arrows represent the restrictions of functions from one domain to the next. When two functions on the middle level are mapped to the same function on the bottom level, they are both images of a function on the top level.

As is clear from the above construction, open sets – the topology of the domain – play a central role in the description of continuous functions. Change the topology and the space of continuous functions changes. Therefore, sampling and reconstruction problems will reflect topological properties. Analytical methods usually make topological properties of the domain implicit and its geometric properties explicit – sheaf theory reverses this: the topology of the domain is explicit, while its geometry is implicit.

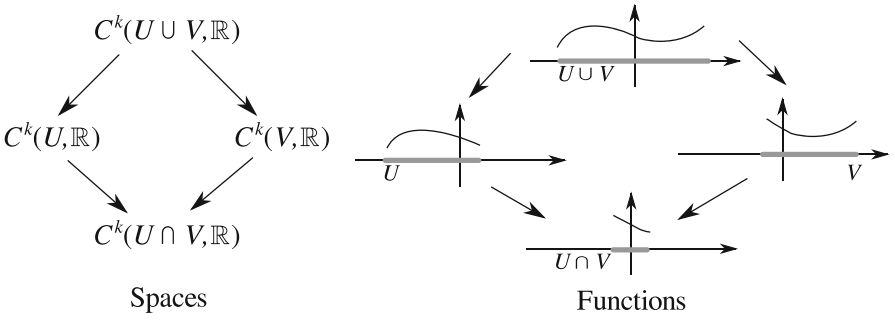


Fig. 10.1 A diagram of spaces of functions over two intersecting sets (left) and a particular function within those spaces (right)

[1] An *open cover* of a topological space X is a collection of open sets whose union is X.

It is usually unnecessary to consider *all* open sets; what is really relevant is the intersection lattice. In this chapter, we need a concept of space that is convenient for computations. The most efficient such definition is that of a simplicial complex, which specifies a decomposition of a space into simplices. The combinatorial structure of a simplicial complex makes most constructions easier, though it can be limiting if not chosen carefully. Multiple constructions are often possible, which usually leads to isomorphic descriptions of sampling problems. As a rule of thumb, our constructions will place vertices everywhere there are samples.

Let us formalize these properties to obtain a more general construction, in which the data are not necessarily encoded as continuous functions. We will encode the data locally, by assigning a vector space to each face. When the gluing rule above indicates that these data are consistent across the entire space, we obtain the analog of a function whose domain is the whole space. That is, functions are gluings of information specified locally and consistently at all parts of the space. Distances, area, and volume are not explicitly included in the construction of a simplicial complex. These important properties are defined implicitly within the definition of the sheaf of functions. Instead, the sheaf-based perspective showcases the importance of topology on sampling.

Definition 1. An *abstract simplicial complex* X on a set A is a collection of ordered subsets of A that is closed under the operation of taking subsets. We call each element of X a *face*. A face with $k+1$ elements is called a k-dimensional face (or a k-face), though we usually call a 0-face a *vertex* and a 1-face an *edge*. If all of the faces of an abstract simplicial complex X are of dimension n or less, we say that X is an n-dimensional simplicial complex. If X is a 1-dimensional simplicial complex, we usually call X a *graph*.

If a and b are two faces in an abstract simplicial complex X with a a proper subset of b, we will write $a \rightsquigarrow b$ and say that a is *attached* to b. Finally, a collection Y of faces of X is called a *closed subcomplex* if whenever $b \in Y$ and $a \rightsquigarrow b$, then $a \in Y$ also.

The ordering of the vertices within a face is called its *orientation*, which generalizes the notion of direction on a graph. For this chapter, an *orientation index* plays an important algebraic role. Suppose that a is a k-face and b is a $k+1$-face of an abstract simplicial complex X. If a is a face of b, suppose that $a = (v_0, \cdots, v_k)$ and $b = (v_{\sigma(0)}, \cdots, \perp, \cdots, v_{\sigma(k)})$, where $\perp$ represents a vertex not appearing in a and σ is a permutation on $k+1$ elements. The *orientation index* is a number given by

$$[b:a] = \begin{cases} (-1)^m \text{sign}(\sigma) & \text{if } \perp \text{ appears in slot } m \text{ (starting with 0) of } b, \text{ or} \\ 0 & \text{if } a \text{ is not a face of } b. \end{cases}$$

Example 1. Consider the graph shown in Figure 10.2. This is a visual representation of the simplicial complex

$$\{\{A,B\},\{C,B\},\{C,D\},\{C,E\},\{A\},\{B\},\{C\},\{D\},\{E\},\emptyset\}.$$

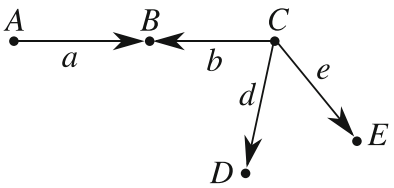


Fig. 10.2 A small directed graph for Example 1

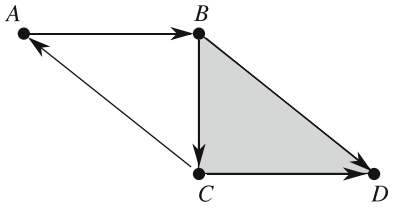


Fig. 10.3 A small simplicial complex for Example 2

Observe that because B appears second in $\{A, B\}$, then $[\{A, B\}, \{B\}] = (-1)^0 \times 1 = +1$ and $[\{A, B\}, \{A\}] = (-1)^1 \times 1 = -1$.

Example 2. Consider the abstract simplicial complex given by

$$\{\{A\}, \{B\}, \{C\}, \{D\}, \{A, B\}, \{B, C\}, \{C, A\}, \{C, D\}, \{B, D\}, \{B, C, D\}, \emptyset\},$$

which is shown in Figure 10.3. Observe that this simplicial complex contains a 2-dimensional face $\{B, C, D\}$. The orientation index relating the 2-dimensional face's edges is given by

$$\begin{aligned}
[\{B, C, D\}, \{B, C\}] &= (-1)^2 \times 1 = +1, \\
[\{B, C, D\}, \{B, D\}] &= (-1)^1 \times 1 = -1, \\
[\{B, C, D\}, \{C, D\}] &= (-1)^0 \times 1 = +1.
\end{aligned}$$

Sometimes simplicial complexes arise naturally from the problem, for instance the connection graph for a network, but it is helpful to have a procedure to obtain a simplicial complex from a topological space. Suppose that X is a topological space and that $\mathscr{U} = \{U_1, \dots\}$ is an open cover of X.

Definition 2. The *nerve* $N(\mathscr{U})$ is the abstract simplicial complex whose vertices are given by the elements of $\mathscr{U}$ and whose k-faces $\{U_{i_0}, \dots, U_{i_k}\}$ are given by the nonempty intersections $U_{i_0} \cap \cdots \cap U_{i_k}$.

Example 3. Figure 10.4 shows two covers and their associated nerves. In the left diagram, the sets A, B, and C have nonempty pairwise intersections and a nonempty

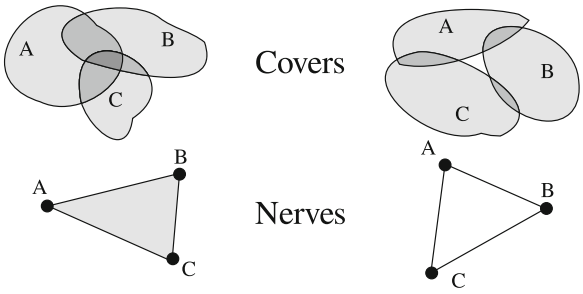


Fig. 10.4 The nerve of two covers: (left) with a nonempty triple intersection (right) without a triple intersection

triple intersection $A \cap B \cap C$, so the nerve is a 2-dimensional abstract simplicial complex. In the right diagram, $A \cap B \cap C$ is empty, so the nerve is only 1 dimensional.

The concept of local information over a simplicial complex is a straightforward generalization of the three properties (restriction, uniqueness, and gluing) for continuous functions. The resulting mathematical object is called a *sheaf*.

Definition 3. A *sheaf* $\mathscr{F}$ on an abstract simplicial complex X is an assignment of the following collection of data to the faces and attachments of X:

- for each element a of X, $\mathscr{F}(a)$ is a vector space, called the *stalk at a*,
- for each attachment of two faces $a \rightsquigarrow b$ of X, $\mathscr{F}(a \rightsquigarrow b)$ is a linear function from $\mathscr{F}(a) \to \mathscr{F}(b)$ called a *restriction map* (or *restriction*), and
- for every composition of attachments $a \rightsquigarrow b \rightsquigarrow c$, the restrictions satisfy $\mathscr{F}(b \rightsquigarrow c) \circ \mathscr{F}(a \rightsquigarrow b) = \mathscr{F}(a \rightsquigarrow b \rightsquigarrow c)$.

We will usually refer to X as the *base space* for $\mathscr{F}$.

Remark 1. Although sheaves have been extensively studied over topological spaces (see [4] or the appendix of [19] for a modern, standard treatment), the resulting definition is ill suited for application to sampling. Instead, we follow a substantially more combinatorial approach introduced in the 1980 thesis of Shepard [37].

Example 4. The space of continuous functions over a topological space can be represented as a sheaf. For instance, Figure 10.5 shows one way to organize the space of continuous functions over the interval $(-2, 2)$ in terms of spaces of continuous functions over smaller intervals. (See Example 8 for another encoding of continuous functions as a sheaf.) In this particular sheaf model, the base space is given by an abstract simplicial complex X over three abstract vertices,

$$\{\{-2\}, \{0\}, \{2\}, \{-2, 0\}, \{0, 2\}, \emptyset\}.$$

We define the sheaf $\mathscr{C}$ over X by assigning spaces of continuous functions to each face. Over vertices, we assign the stalks

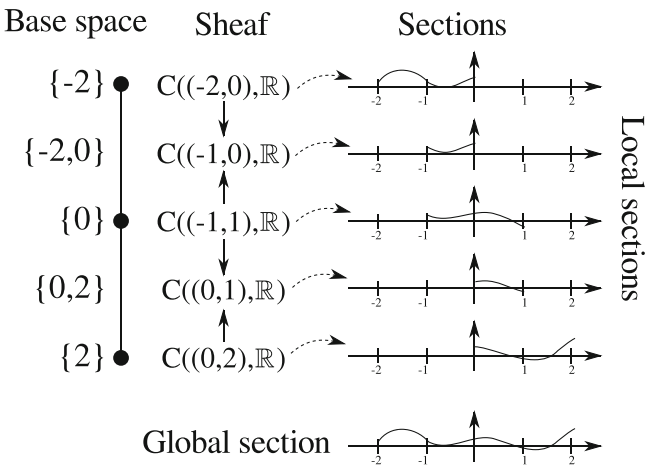


Fig. 10.5 A sheaf of continuous functions over an interval (compare with Figure 10.6)

$$\begin{aligned}\mathscr{C}(\{-2\}) &= C((-2,0),\mathbb{R}),\\ \mathscr{C}(\{0\}) &= C((-1,1),\mathbb{R}), \text{ and}\\ \mathscr{C}(\{2\}) &= C((0,2),\mathbb{R}).\end{aligned}$$

Notice that each of these are spaces of continuous functions over intervals of length 2 and that they overlap. The stalks over the edges specify these overlapping regions, so they are spaces of continuous functions over intervals of length 1 as follows:

$$\begin{aligned}\mathscr{C}(\{-2,0\}) &= C((-1,0),\mathbb{R}) \text{and}\\ \mathscr{C}(\{0,2\}) &= C((0,1),\mathbb{R}).\end{aligned}$$

The restriction maps between stalks are given by the process of "actually" restricting the domains of the functions.

Example 5. Coming back to Section 10.2, a sheaf of analytic functions can be constructed as a subsheaf of the previous example by merely replacing the stalks with spaces of analytic functions defined over the appropriate intervals.

Notice that the definition of a sheaf captures the *restriction* locality property, but does not formalize the *uniqueness* or *gluing* properties. Some authors [4, 7, 16, 20] explicitly require these properties from the outset, calling the object defined in Definition 3 a *presheaf*, regarding it as incomplete. Although the difference between sheaves and presheaves is useful in navigating certain technical arguments, every presheaf has a unique *sheafification*. Because of this, our strategy follows the somewhat more economical treatment set forth in [35, 37], which removes this

distinction. As a consequence of this choice, we explicitly define collections of *sections*, which effectively implement the uniqueness and gluing properties.

Definition 4. Suppose $\mathscr{F}$ is a sheaf on an abstract simplicial complex X and that $\mathscr{U}$ is a collection of faces of X. An assignment s which assigns an element of $\mathscr{F}(a)$ to each face $a \in \mathscr{U}$ is called a *section* supported on $\mathscr{U}$ when for each attachment $a \rightsquigarrow b$ of faces in $\mathscr{U}$, $\mathscr{F}(a \rightsquigarrow b)s(a) = s(b)$. We will denote the space of sections of $\mathscr{F}$ over $\mathscr{U}$ by $\mathscr{F}(\mathscr{U})$, which is easily checked to be a vector space. A *global section* is a section supported on X. If r and s are sections supported on $\mathscr{U} \subset \mathscr{V}$, respectively, in which $r(a) = s(a)$ for each $a \in \mathscr{U}$ we say that s *extends* r.

Example 6. The space of global sections in the sheaf $\mathscr{C}$ given in Example 4 is $C((-2,2),\mathbb{R})$. A global section of $\mathscr{C}$ consists of five continuous functions on the intervals $(-2,0), (-1,0), (-1,1), (0,1), (0,2)$ that all restrict to the same two functions on $(-1,0)$ and $(0,1)$. Since the union of these intervals is $(-2,2)$ and is connected, these five continuous functions must be restrictions of a single continuous function over $(-2,2)$.

The following sheaf plays a central role in this chapter, as it is used to represent discrete time series.

Example 7. Consider $Y \subseteq X$ a subset of the vertices of an abstract simplicial complex. A sheaf $\mathscr{S}$ which assigns a vector space V to vertices in Y and the trivial vector space to every other face is called a *V-sampling sheaf supported on Y*. To every attachment of faces of different dimension, $\mathscr{S}$ will assign the zero function. For a finite abstract simplicial complex X, the space of global sections of a V-sampling sheaf supported on Y is isomorphic to $\bigoplus_{y \in Y} V$.

Example 8. Figure 10.6 shows a sheaf that is essentially dual to the one in Example 4 and whose global sections are continuous functions. (Although it is straightforward to generalize the construction to cell complexes of arbitrary dimension, we will work over an interval to keep the exposition simple.) Consider a simplicial complex with two vertices v_1 and v_2 and one edge e between them. The stalk over each vertex is a space of continuous functions as in Example 4, though we require the functions to be continuous over a *closed* interval. However, the stalk

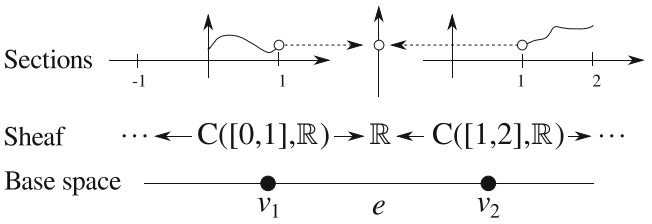


Fig. 10.6 Another sheaf of continuous functions over an interval (compare with Figure 10.5)

over the edge is merely $\mathbb{R}$. The restriction in this case evaluates functions at an appropriate endpoint. If we name the sheaf $\mathscr{C}$, then for $f \in \mathscr{C}(v_1) = C([0,1],\mathbb{R})$,

$$(\mathscr{C}(v_1 \nrightarrow e))(f) = f(1),$$

and

$$(\mathscr{C}(v_2 \nrightarrow e))(g) = g(1),$$

for $g \in \mathscr{C}(v_2) = C([1,2],\mathbb{R})$. Observe that the global sections of this sheaf are precisely functions that are continuous on $[0,2]$.

Sheaves can also describe spaces of piecewise continuous functions, as the next example shows.

Example 9. Suppose G is a graph in which each vertex has finite degree. Let $\mathscr{PL}$ be the sheaf constructed on G that assigns $\mathscr{PL}(v) = \mathbb{R}^{1+\deg v}$ to each edge v of degree $\deg v$ and $\mathscr{PL}(e) = \mathbb{R}^2$ to each edge e. The stalks of $\mathscr{PL}$ specify the value of the function (denoted y below) at each face and the slopes of the function on the edges (denoted $m_1, \ldots, m_k$ below).

To each attachment of a degree k vertex v into an edge e, let $\mathscr{PL}$ assign the linear function

$$(\mathscr{PL}(v \nrightarrow e))(y, m_1, \cdots, m_e, \cdots, m_k) = \begin{pmatrix} y + ([e:v]-1)\frac{1}{2}m_e \\ m_e \end{pmatrix}.$$

The global sections of this sheaf are *piecewise linear functions* on G, which is discussed extensively in Section 10.5.3. Figure 10.7 shows an example of this sheaf on a graph model of $\mathbb{R}$. Consider the stalk at the vertex v_0 at the origin. Since this vertex has degree 2, the stalk is given by the space of ordered triples (y, m_1, m_2). The y specifies the value of a section at that vertex, while m_1 specifies the slope to the left, while m_2 specifies the slope to the right. The left restriction is

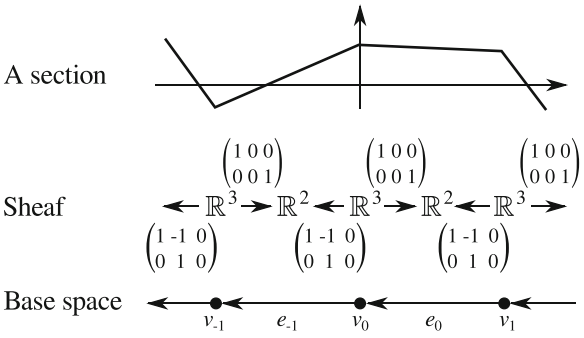


Fig. 10.7 An example of a sheaf $\mathscr{PL}$ over a graph

$$(\mathscr{P}\mathscr{L}(v_0 \rightsquigarrow e_{-1}))\,(y, m_1, m_2) = \begin{pmatrix} y - m_1 \\ m_1 \end{pmatrix}$$

because $[e_{-1} : v_0] = [\{v_0, v_{-1}\} : v_0] = -1$. The right restriction is

$$(\mathscr{P}\mathscr{L}(v_0 \rightsquigarrow e_0))\,(y, m_1, m_2) = \begin{pmatrix} y \\ m_2 \end{pmatrix}$$

because $[e_0 : v_0] = [\{v_1, v_0\} : v_0] = +1$. Notice that in both cases, the value of a section over an edge $\{v_1, v_2\}$ specifies the value of the linear function at v_2 and the slope of the line along the edge.

10.3.2 Sheaf cohomology

The space of global sections of a sheaf is important in applications. Although Definition 4 is not constructive, one can compute this space algorithmically. Consider the abstract simplicial complex X shown in Figure 10.8, which consists of an edge e between two vertices v_1 and v_2. Suppose s is a global section of a sheaf $\mathscr{S}$ on X. This means that

$$\mathscr{S}(v_1 \rightsquigarrow e)s(v_1) = s(e) = \mathscr{S}(v_2 \rightsquigarrow e)s(v_2).$$

Since the above equation is written in a vector space, we can rearrange it to obtain the equivalent specification

$$\mathscr{S}(v_1 \rightsquigarrow e)s(v_1) - \mathscr{S}(v_2 \rightsquigarrow e)s(v_2) = 0,$$

which could be written in matrix form as

$$\begin{pmatrix} \mathscr{S}(v_1 \rightsquigarrow e) & -\mathscr{S}(v_2 \rightsquigarrow e) \end{pmatrix} \begin{pmatrix} s(v_1) \\ s(v_2) \end{pmatrix} = 0.$$

This purely algebraic manipulation shows that computing the space of global sections of a sheaf is equivalent to computing the kernel of a particular matrix as in Section 10.2. Clearly this procedure ought to work for arbitrary sheaves over arbitrary abstract simplicial complexes, though it could get quite complicated.

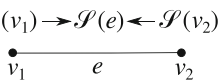


Fig. 10.8 A sheaf over a small abstract simplicial complex

Cohomology is a systematic way to perform this computation, and it results in additional information as we will see in later sections.

The vector $\begin{pmatrix} s(v_1) \\ s(v_2) \end{pmatrix}$ above suggests that we should define the following formal *cochain* vector spaces $C^k(X;\mathscr{F}) = \bigoplus_{a \text{ a } k-\text{face of } X} \mathscr{F}(a)$ to represent the possible assignments of data to the k-faces. In the same way, the matrix

$$\big(\mathscr{S}(v_1 \rightsquigarrow e) \; -\mathscr{S}(v_2 \rightsquigarrow e)\big)$$

generalizes into the *coboundary map* $d^k : C^k(X;\mathscr{F}) \to C^{k+1}(X;\mathscr{F})$, which we now define. The coboundary map d^k takes an assignment s on the k-faces to a different assignment $d^k s$ whose value at a $(k+1)$-face b is

$$(d^k s)(b) = \sum_{a \text{ a } k-\text{face of } X} [b : a]\, \mathscr{F}(a \rightsquigarrow b)\, s(a). \tag{10.1}$$

(Notice that the orientation index $[b : a]$ supplies the minus sign in the matrix equations above.) Together, we have a sequence of linear maps

$$0 \to C^0(X;\mathscr{F}) \xrightarrow{d^0} C^1(X;\mathscr{F}) \xrightarrow{d^1} C^2(X;\mathscr{F}) \xrightarrow{d^2} \cdots$$

called the *cochain complex*.

As in the simple example described above, the kernel of d^k consists of data specified on k-faces that are consistent when tested on the $(k+1)$-faces. However, because of the orientation index in the coboundary map, it can be shown that $d^{k+1} \circ d^k = 0$, so that the image of d^k is a subspace of the kernel of d^{k+1}. This means that the image of d^k is essentially redundant information, since it is already known to be consistent when tested on the $(k+2)$-faces. Because of this fact, only those elements of the kernel of d^k that are *not* already known to be consistent are really worth mentioning. This leads to the definition of sheaf cohomology:

Definition 5. The kth *sheaf cohomology* of $\mathscr{F}$ on an abstract simplicial complex X is

$$H^k(X;\mathscr{F}) = \ker d^k / \operatorname{image} d^{k-1}.$$

As an immediate consequence of this construction, we have the following useful statement.

Proposition 1. $H^0(X;\mathscr{F}) = \ker d^0$ *consists precisely of those assignments s which are global sections, so a global section is determined entirely by its values on the vertices of X.*

10.3.3 Transformations of local data

Sheaves can be used to represent local data and cohomology can be used to infer the resulting globally consistent data. We now connect this theory to the process of sampling. As envisioned in Section 10.2, sampling is a *transformation* between two spaces of functions – from functions with a continuous domain to functions with a discrete domain. Such a transformation arising from sampling respects the local structure of the function spaces. This kind of transformation is called a *sheaf morphism*. There are two aspects to a sheaf morphism: (1) its effect on the base space and (2) its effect on stalks. The effect on the base space should be to respect local neighborhoods, which means that a sheaf morphism must at least specify a continuous map. Since we have restricted our attention to abstract simplicial complexes rather than general topological spaces, the analog of a continuous map is a simplicial map.

Definition 6. A *simplicial map* from one abstract simplicial complex X to another Y is a function f from the set of faces of X to the faces of Y that additionally satisfies two properties:

1. If $a \rightsquigarrow b$ is an attachment of two faces in X, then $f(a) \rightsquigarrow f(b)$ is an attachment of faces in Y and
2. The dimension of $f(a)$ is no more than the dimension of a, a face in X.

The last condition means a simplicial map takes vertices to vertices, edges either to edges or vertices, and so on.

Example 10. Consider the simplicial complexes X and Y shown in Figure 10.9. The function $F : X \to Y$ given by

$$F(v_1) = w_1,\ F(v_2) = w_2,\ F(v_3) = w_2$$

determines a simplicial map, in which $F(e_1) = w_2, F(e_2) = F(e_3) = F(f) = g_1$.

In contrast, any function that takes v_1 to w_1, v_2 to w_2, and v_3 to w_3 cannot be a simplicial map because the image of e_2 should be an edge from w_1 to w_3, but no such edge exists.

Definition 7. Suppose that $f : X \to Y$ is a simplicial map, that $\mathscr{F}$ is a sheaf on Y, and that $\mathscr{G}$ is a sheaf on X. A *sheaf morphism* (or simply a *morphism*) $m : \mathscr{F} \to \mathscr{G}$

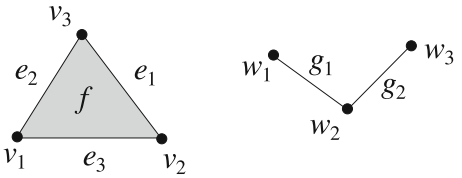


Fig. 10.9 The simplicial complexes X (left) and Y (right) for Example 10

along f assigns a linear map $m_a : \mathscr{F}(f(a)) \to \mathscr{G}(a)$ to each face $a \in X$ so that for every attachment $a \rightsquigarrow b$ of X, $m_b \circ \mathscr{F}(f(a) \rightsquigarrow f(b)) = \mathscr{G}(a \rightsquigarrow b) \circ m_a$.

Usually, we describe a morphism by way of a commutative diagram like the one below.

$$\begin{array}{ccc} \mathscr{F}(f(a)) & \xrightarrow{m_a} & \mathscr{G}(a) \\ {\scriptstyle \mathscr{F}(f(a) \rightsquigarrow f(b))} \downarrow & & \downarrow {\scriptstyle \mathscr{G}(a \rightsquigarrow b)} \\ \mathscr{F}(f(b)) & \xrightarrow{m_b} & \mathscr{G}(b) \end{array}$$

Remark 2. The reader is cautioned that a sheaf morphism and its underlying simplicial map "go opposite ways."

$$\text{Sheaf morphism :} \qquad \mathscr{G} \xleftarrow{m} \mathscr{F}$$

$$\text{Simplicial map :} \qquad X \xrightarrow{f} Y$$

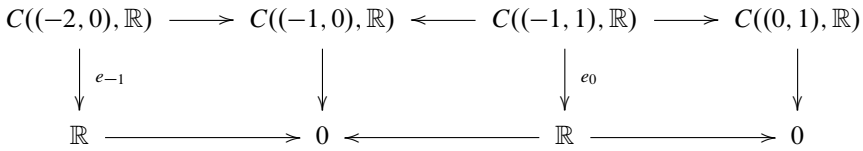
Cohomology is a functor from the category of sheaves and sheaf morphisms to the category of vector spaces. This indicates that cohomology preserves and reflects the underlying relationships between data stored in sheaves.

Proposition 2. *Suppose that $\mathscr{R}$ is a sheaf on X and that $\mathscr{S}$ is a sheaf on Y. If $m : \mathscr{R} \to \mathscr{S}$ is a morphism of these sheaves, then m induces linear maps $m^k : H^k(X, \mathscr{R}) \to H^k(Y, \mathscr{S})$ for each k. (Note that the simplicial map associated to m is a function $Y \to X$.)*

As a consequence, m^0 is a linear map from the space of global sections of $\mathscr{R}$ to the space of global sections of $\mathscr{S}$. Because of this, it is possible to describe the process of sampling using a sheaf morphism.

Definition 8. Suppose that $\mathscr{F}$ is a sheaf on an abstract simplicial complex X, and that $\mathscr{S}$ is a V-sampling sheaf on X supported on a closed subcomplex Y. A *sampling morphism* (or *sampling*) of $\mathscr{F}$ is a morphism $s : \mathscr{F} \to \mathscr{S}$ that is surjective on every stalk.

Example 11. The diagram below shows a morphism (vertical arrows) between two sheaves, namely the sheaf of continuous functions defined in Example 4 (top row) and the sampling sheaf defined in Example 7 (bottom row).



In the diagram, e_x represents the operation of evaluating a continuous function at x. As in Section 10.2, this sampling morphism takes a continuous function $f \in C((-2,1),\mathbb{R})$ to a vector $(f(-1), f(0))$.

In algebraic topology, special emphasis is placed on *sequences* of maps of the form

$$\cdots \longrightarrow A_1 \xrightarrow{m_1} A_2 \xrightarrow{m_2} A_3 \xrightarrow{m_3} A_4 \xrightarrow{m_4} \cdots,$$

where the A_k are vector spaces and the m_k are linear maps. We will denote this sequence by $(A_\bullet, m_\bullet)$. For instance, the cochain complex described in the previous section is a sequence of vector spaces. A linear map satisfies the dimension theorem, which relates the size of its kernel, cokernel, and image. In some sequences, the dimension theorem is extremely useful – these are the exact sequences.

Definition 9. A sequence $(A_\bullet, m_\bullet)$ of vector spaces is called *exact* if $\ker m_k = \operatorname{image} m_{k-1}$ for all k.

Via the dimension theorem, exact sequences can encode information about linear maps, namely

1. $0 \to A \xrightarrow{m} B$ is exact if and only if m is injective,
2. $A \xrightarrow{m} B \to 0$ is exact if and only if m is surjective, and
3. $0 \to A \xrightarrow{m} B \to 0$ is exact if and only if m is an isomorphism.

Observe that the cochain complex $(C^\bullet(X;\mathscr{S}), d^\bullet)$ is exact if and only if $H^k(X;\mathscr{S}) = 0$ for all k.

Remark 3. Sequences of sheaf morphisms (instead of just vector spaces) are surprisingly powerful and play an important role in the general theory of sheaves. If the direction of the morphisms is allowed to change across the sequence, like

$$\mathscr{A} \longleftarrow \mathscr{B} \longrightarrow \mathscr{C},$$

the resulting construction can represent all linear, shift-invariant filters [34, 35].

10.4 The general sampling theorems

Given a sampling morphism $\mathscr{F} \to \mathscr{S}$, we can construct its *ambiguity sheaf* $\mathscr{A}$, which characterizes lost information. The ambiguity sheaf is constructed over the same space X as $\mathscr{F}$. Each stalk of the ambiguity sheaf $\mathscr{A}(a)$ at a face $a \in X$ is given by the kernel of the map $\mathscr{F}(a) \to \mathscr{S}(a)$. Each restriction map of the ambiguity sheaf $\mathscr{A}(a \rightsquigarrow b)$ is given by restricting the domain of $\mathscr{F}(a \rightsquigarrow b)$ to $\mathscr{A}(a)$ whenever $a \rightsquigarrow b$ is an attachment of faces in X. This implies that the exact sequence of sheaves

$$0 \to \mathscr{A} \longrightarrow \mathscr{F} \xrightarrow{s} \mathscr{S} \to 0$$

induces short exact sequences of cochain spaces

$$0 \to C^k(X; \mathscr{A}) \to C^k(X; \mathscr{F}) \to C^k(X; \mathscr{S}) \to 0,$$

one for each k. Together, these sequences of cochain spaces induce a long exact sequence (via the well-known Snake lemma; see [18] for instance)

$$0 \to H^0(X; \mathscr{A}) \to H^0(X; \mathscr{F}) \to H^0(X; \mathscr{S}) \to H^1(X; \mathscr{A}) \to \cdots$$

An immediate consequence is therefore

Corollary 1 (Sheaf-theoretic Nyquist theorem). *The global sections of $\mathscr{F}$ are identical with the global sections of $\mathscr{S}$ if and only if $H^k(X; \mathscr{A}) = 0$ for $k = 0$ and* 1.

The cohomology space $H^0(X; \mathscr{A})$ characterizes the *ambiguity* in the sampling. When $H^0(X; \mathscr{A})$ is nontrivial, there are multiple global sections of $\mathscr{F}$ that result in the same set of samples. In contrast, $H^1(X; \mathscr{A})$ characterizes the *redundancy* of the sampling. When $H^1(X; \mathscr{A})$ is nontrivial, then there are sets of samples that correspond to *no* global section of $\mathscr{F}$. Optimal sampling therefore consists of identifying minimal closed subcomplexes Y so the resulting ambiguity sheaf $\mathscr{A}$ has $H^0(X; \mathscr{A}) = H^1(X; \mathscr{A}) = 0$.

Remark 4. Corollary 1 is also useful for describing boundary value problems for differential equations, as we will see in Section 10.5.2. The sheaf $\mathscr{F}$ can be taken to be a sheaf of solutions to a differential equation [11]. The sheaf $\mathscr{S}$ can be taken to have support only at the boundary of the region of interest, and therefore specifies the possible boundary conditions. In this case, the space of global sections of the ambiguity sheaf $\mathscr{A}$ consists of all solutions to the differential equation that *also* satisfy the boundary conditions.

Let us place bounds on the cohomologies of the ambiguity sheaf. To do so, we construct two new sheaves associated to a given sheaf $\mathscr{F}$ and a closed subcomplex $Y \subseteq X$. These new sheaves allow us to study reconstruction from a collection of rich samples.

Definition 10. For a closed subcomplex Y of X, let $\mathscr{F}^Y$ be the sheaf whose stalks are the stalks of $\mathscr{F}$ on Y and zero elsewhere, and whose restrictions are either those of $\mathscr{F}$ on Y or zero as appropriate. There is a surjective sheaf morphism $\mathscr{F} \to \mathscr{F}^Y$ and an induced ambiguity sheaf $\mathscr{F}_Y$ which can be constructed in exactly the same way as $\mathscr{A}$ before.

Thus the dimension of each stalk of $\mathscr{F}^Y$ is larger than that of any sampling sheaf supported on Y, and the dimension of stalks of $\mathscr{F}_Y$ are therefore as small as or smaller than that of any ambiguity sheaf. Because global sections are determined by their values at the vertices (Proposition 1), obtaining rich samples from $\mathscr{F}^Y$ at all vertices evidently allows reconstruction. This idea works for all degrees of cohomology, which generalizes the notion of oversampling.

Proposition 3 (Oversampling theorem). *If X^k is the closed subcomplex generated by the k-faces of X, then $H^k(X^{k+1}; \mathscr{F}_{X^k}) = 0$.*

Proof. By direct computation, the k-cochains of $\mathscr{F}_{X^k}$ are

$$\begin{aligned} C^k(X^{k+1}; \mathscr{F}_{X^k}) &= C^k(X^{k+1}; \mathscr{F}) / C^k(X^k; \mathscr{F}) \\ &= \bigoplus_{a \text{ a } k\text{-face of } X} \mathscr{F}(a) \Big/ \bigoplus_{a \text{ a } k\text{-face of } X} \mathscr{F}(a) \\ &= 0. \end{aligned}$$

□

As an immediate consequence, $H^0(X; \mathscr{F}_Y) = 0$ when Y is the set of vertices of X. On the other hand, not taking enough samples leads to an ambiguous reconstruction problem. This can be detected by the presence of nontrivial global sections of the ambiguity sheaf.

Theorem 1 (Sampling obstruction theorem). *Suppose that Y is a closed subcomplex of X and $s : \mathscr{F} \to \mathscr{S}$ is a sampling of sheaves on X supported on Y. If $H^0(X, \mathscr{F}_Y) \neq 0$, then the induced map $H^0(X; \mathscr{F}) \to H^0(X; \mathscr{S})$ is not injective.*

Succinctly, $H^0(X, \mathscr{F}_Y)$ is an obstruction to the recovery of global sections of $\mathscr{F}$ from its samples.

Proof. We begin by constructing the ambiguity sheaf $\mathscr{A}$ as before so that

$$0 \to \mathscr{A} \to \mathscr{F} \xrightarrow{s} \mathscr{S} \to 0$$

is a short exact sequence of sheaves. Observe that $\mathscr{S} \to \mathscr{F}^Y$ can be chosen to be injective, because the stalks of $\mathscr{S}$ have dimension not more than the dimension of $\mathscr{F}$ (and hence $\mathscr{F}^Y$ also). Thus the induced map $H^0(X;\mathscr{S}) \to H^0(X;\mathscr{F}^Y)$ is also injective. Therefore, by a diagram chase on

$$\begin{array}{ccccc} 0 \to H^0(X;\mathscr{A}) & \longrightarrow & H^0(X;\mathscr{F}) & \xrightarrow{s} & H^0(X;\mathscr{S}) \\ & & \downarrow \cong & & \downarrow \\ 0 \to H^0(X;\mathscr{F}_Y) & \longrightarrow & H^0(X;\mathscr{F}) & \longrightarrow & H^0(X;\mathscr{F}^Y) \end{array}$$

we infer that there is a surjection $H^0(X;\mathscr{A}) \to H^0(X;\mathscr{F}_Y)$. By hypothesis, this means that $H^0(X;\mathscr{A}) \neq 0$, so in particular $H^0(X;\mathscr{F}) \to H^0(X;\mathscr{S})$ cannot be injective. □

10.5 Examples

This section shows the unifying power of a sheaf-theoretic approach to sampling, by focusing on three rather different examples. The examples differ in terms of how "local" the reconstruction is; those that are less local show a greater impact of the topology of the base space on reconstruction. We examine

1. *The Paley-Wiener space PW_B on the real line*, which leads to a sheaf-theoretic reinterpretation of the Shannon-Nyquist sampling theorem. Because of the intimate connection between the usual Laplace-Beltrami operator and the topology of the base space, global topology strongly impacts the number of samples required.
2. *Quantum graphs*, which reflect an intermediate case in which nontrivial topology in the domain is detected, sometimes not.
3. *Splines* with a fixed knot sequence, which exhibit a substantially coarsened base space topology. The resulting functions are determined locally and do not reflect much of the global topology of the base space.

The case of the Paley-Wiener space PW_B is rather well known – but we show that it has a sheaf-theoretic interpretation. In rather stark contrast to the case of PW_B is the vector space consisting of the B-splines associated to a fixed knot sequence. The functions in this space are determined via a locally finite, piecewise polynomial partition of unity. Since the resulting B-splines are determined locally with respect to a much coarser topology than the usual one, it makes sense that reconstructing them from local samples is possible. Importantly, sampling theorems obtained for spaces of B-spline are less sensitive to global topological properties.

Spaces of solutions to linear differential equations have intermediate sampling behavior between PW_B and the space of B-splines. While a degree k differential

equation defines its solution locally, there are k linearly independent such solutions. Additionally, the topology of the underlying space on which the differential equation is written impacts the process of reconstruction from samples [26, 28, 32].

The unifying power of sheaf theory means that all of the examples in this section can be treated in the same way, according to the following procedure:

1. Representing the base space and the functions to be sampled in a sheaf,
2. Constructing a sampling morphism between sheaves, and
3. Analyzing the cohomology of the resulting ambiguity sheaf.

10.5.1 Bandlimited functions

In this section, we prove the traditional form of the Nyquist theorem by showing that an appropriate bandlimit is a sufficient condition for $H^0(X;\mathscr{A}) = 0$, where $\mathscr{A}$ is an ambiguity sheaf and X is an abstract simplicial complex for the real line $\mathbb{R}$.

Recall that the Paley-Wiener space PW_B consists of functions f whose Fourier transform

$$\hat{f}(\omega) = \int_{-\infty}^{\infty} f(x)e^{-2\pi i\omega x}dx$$

is supported on $[-B, B]$. We say that each $f \in PW_B$ has *bandwidth* B. The Shannon-Nyquist theorem asserts that functions in $PW_{1/2}$ are uniquely determined by their values on the integers, which is best explained by the fact that every $f \in PW_{1/2}$ has a cardinal series decomposition

$$f(x) = \sum_{n=-\infty}^{\infty} f(n)\frac{\sin \pi(x-n)}{\pi(x-n)}.$$

Moreover, the set of sinc functions is orthonormal over the usual inner product in $PW_{1/2}$, so we have that

$$f(n) = \int_{-\infty}^{\infty} f(x)\frac{\sin \pi(x-n)}{\pi(x-n)}\,dx. \tag{10.2}$$

Even though the support of $\sin \pi(x-n)/(\pi(x-n))$ is $\mathbb{R}$, it decays away from n. This means that in (10.2), the effect of values of f far away from n will have little effect on $f(n)$. So in the case of PW_B, sampling via (10.2) is only *approximately* local. Because of this, global constraints – such as those arising from compactness – on the function space play an important role in sampling theorems.

We begin by specifying the following 1-dimensional simplicial complex X. This simplicial complex should be a model for the real line – the domain of the functions we will be sampling. In order to facilitate the construction of a sampling morphism

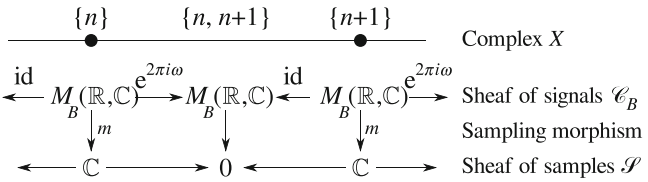


Fig. 10.10 The sheaves used in proving the traditional Nyquist theorem

that will sample functions at the integers, let the set of vertices be given by $X^0 = \mathbb{Z}$. The edges connect each consecutive pair of vertices so that the set of edges is given by $X^1 = \{\{n, n+1\} : n \in \mathbb{Z}\}$, which yields the same base space as in Example 4.

The main property of bandlimited functions is that their Fourier transform has bounded support. Although the sampling to be performed is local, bandlimitedness is not a local property. The sheaf of bandlimited signals will be somewhat trivial – all of the stalks will be the same. Therefore, we construct the sheaf $\mathscr{C}_B$ of signals according to their Fourier transforms (see Figure 10.10) so that for every face, the stalk of $\mathscr{C}_B$ is the vector space $M_B(\mathbb{R}, \mathbb{C})$ of complex-valued measures on $\mathbb{R}$ whose support is contained in $[-B, B]$.

Intuitively, the stalk over a vertex n represents functions that are localized to a vicinity of that vertex. Following the inspiration of Example 4, moving from one vertex to the next amounts to translating the function. Therefore, moving from one vertex in $\mathscr{C}_B$ should apply a time translation of one sample. Since the stalks represent the function's frequency domain, time translation applies a multiplication by a unit complex number. Without loss of generality, each restriction to the left is chosen to be the identity, and each restriction to the right is chosen to be multiplication by $e^{2\pi i\omega}$. In essence, $\mathscr{C}_B$ is the sheaf of *local Fourier transforms* of functions on $\mathbb{R}$. Observe that the space of global sections of $\mathscr{C}_B$ is therefore just $M_B(\mathbb{R}, \mathbb{C})$, because the restrictions do not change the bandwidth of the function being represented.

Example 12. A simple example of a bandlimited function is $f(x) = \sin(2\pi x)$, which is represented as a global section F of $\mathscr{C}_1$. The value of F over a vertex n is given by the measure

$$F(\{n\}) = e^{2\pi i n\omega}\left(\frac{1}{2i}\delta(\omega+1) - \frac{1}{2i}\delta(\omega-1)\right).$$

By Proposition 1, the values of F at the edges of X are merely the same values – since the restrictions are given by multiplication by $e^{2\pi\omega}$.

Another familiar example of a bandlimited function is the sinc function $g(x) = \sin(\pi x)/(\pi x)$. This is represented as a global section G of $\mathscr{C}_{1/2}$ in which the value over each vertex n is given by the measure

$$G(\{n\}) = e^{2\pi i n\omega}\mathrm{rect}(\omega),$$

where rect(ω) is the measure that takes the value 1 for $-1/2 \le \omega \le 1/2$ and zero elsewhere.

Construct the sampling sheaf $\mathscr{S}$ whose stalk on each vertex is $\mathbb{C}$ and each edge stalk is zero. We construct a sampling morphism $m : \mathscr{C}_B \to \mathscr{S}$ by the zero map on each edge, and by the integral

$$m(f) = \int_{-\infty}^{\infty} f(\omega) d\omega = f((-\infty, \infty))$$

on each vertex.

Then the ambiguity sheaf $\mathscr{A}_B$ has stalks $M_B(\mathbb{R}, \mathbb{C})$ on each edge, and $\{f \in M_B(\mathbb{R}, \mathbb{C}) : m(f) = 0\}$ on each vertex $\{n\}$.

Example 13. Continuing Example 12, consider the function $f(x) = \sin(2\pi x)$ sampled at the integers. The sampling morphism m takes the global section F to a global section mF of the sampling sheaf $\mathscr{S}$. Then at a vertex n,

$$\begin{aligned} mF(n) &= e^{2\pi i n\omega} \left(\frac{1}{2i}\delta(\omega + 1) - \frac{1}{2i}\delta(\omega - 1) \right) ((-\infty, \infty)) \\ &= e^{2\pi i n\omega} \frac{1}{2i}\delta(\omega + 1)(-\infty, \infty) - e^{2\pi i n\omega} \frac{1}{2i}\delta(\omega - 1)(-\infty, \infty) \\ &= e^{-2\pi i n} \frac{1}{2i} - e^{2\pi i n} \frac{1}{2i} \\ &= \sin(-2\pi n) = 0, \end{aligned}$$

which means that mF is the zero section of $\mathscr{S}$. Thus F lies in the kernel of m, and therefore pulls back to a nontrivial section of the ambiguity sheaf $\mathscr{A}_1$. The correct interpretation is that the function f cannot be sampled unambiguously.

Theorem 2 (Traditional Nyquist theorem). *If $B \le 1/2$, the ambiguity sheaf $\mathscr{A}_B$ has $H^0(X; \mathscr{A}_B) = 0$. Therefore, each such function can be recovered uniquely from its samples on $\mathbb{Z}$.*

Proof. The elements of $H^0(X; \mathscr{A}_B)$ are given by the measures f supported on $[-B, B]$ for which

$$\int_{-B}^{B} f(\omega) e^{2\pi i n\omega} d\omega = (e^{2\pi i n\omega} f)([-B, B]) = 0$$

for all n. Observe that if $B \le 1/2$, this is precisely the statement that the Fourier series coefficients of f all vanish; hence f must vanish. This means that the only global section of $\mathscr{A}_B$ is the zero function. (Ambiguities can arise if $B > 1/2$, because the set of functions $\{e^{-2\pi i n\omega}\}_{n \in \mathbb{Z}}$ is then *not* complete.) □

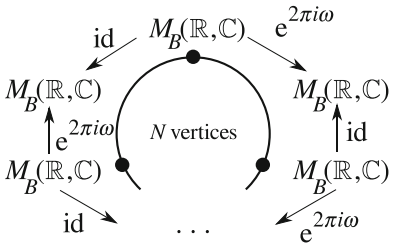


Fig. 10.11 The sheaf $\mathscr{C}_B$ of local Fourier transforms of functions on a circle with N vertices

Sampling on the circle can be addressed by a related construction of a sheaf $\mathscr{C}_B$. As indicated in Figure 10.11, the stalk over each edge and vertex is still $M_B(\mathbb{R},\mathbb{C})$. Again, the restrictions are chosen so that left-going restrictions are identities and the right-going restrictions consist of multiplying measures by $e^{2\pi i\omega}$. As in the case of functions on a line, this restriction map explains the effect of translation on the Fourier transform of a function. This also means that functions that are local to an edge or a vertex do not reflect any nontrivial topology since the restriction maps are identical to what they were in the case of the line.

Since the topology is no longer that of a line, there are some important consequences. The space of global sections of $\mathscr{C}_B$ on the circle is *not* $M_B(\mathbb{R},\mathbb{C})$, and now depends on the number N of vertices on the circle.[2] One may conclude from a direct computation that the value of any global section at a vertex must be a measure f satisfying

$$e^{2\pi i\omega N} f(\omega) = f(\omega).$$

This means that the support of f must be no larger than the set of fractions $\frac{1}{N}\mathbb{Z}$ because at each ω either $f(\omega)$ must vanish or $(e^{2\pi i\omega N} - 1)$ must vanish. Hence a global section describes a function whose Fourier transform is discrete. Thus resulting space of global sections of $\mathscr{C}_B$ is finite dimensional. Perhaps surprisingly, this does not impact the required sampling rate.

Corollary 2. *If $\mathscr{C}_B$ is sampled at each vertex, then a sampling morphism will fail to be injective on global sections if $B > 1/2$.*

(If $B < 1/2$, some sampling morphisms – such as the zero morphism – are still not injective.)

Proof. Merely observe that for a sampling sheaf $\mathscr{S}$ a necessary condition for injectivity is that

[2] N must be at least 3 to use an abstract simplicial complex model of the circle. If N is 1 or 2, one must instead use a CW complex. This does not change the analysis presented here.

$$\begin{aligned} \dim H^0(\mathscr{C}_B) &\leq \dim H^0(\mathscr{S}) \\ (2N+1)B &\leq N \\ B &\leq N/(2N+1) < 1/2. \end{aligned}$$ □

10.5.2 Wave propagation (quantum graphs)

A rich source of interesting sheaves arise in the context of differential equations [11]. Sampling problems are interesting in spaces of solutions to differential equations, because they are restricted enough to have relatively relaxed sampling rates. Although a differential equation describes a function locally, continuity and boundary conditions allow topology to influence which of these locally defined functions can be extended globally.

Consider the differential equation

$$\frac{\partial^2 u}{\partial x^2} + k^2 u = 0 \tag{10.3}$$

on the real line, in which k is a complex scalar parameter called the *wavenumber*. The general solution to this differential equation is the linear combination of two traveling waves, namely

$$u(x) = c_1 e^{ikx} + c_2 e^{-ikx},$$

a right-going and a left-going wave. This means that locally and globally, a given solution is described by an element of $\mathbb{C}^2$.

Although the definition of a sheaf in this chapter is combinatorial, it can be an accurate model of the space of solutions of a differential equation. Under an appropriate definition of the Laplacian operator on the geometric realization of X (such as is given in [1, 2, 32, 44]), we will construct the sheaf of solutions to (10.3). There are sensible definitions for bandlimitedness in this geometric realization, which give rise to Shannon-Nyquist theorems [26, 28]. However, our focus will remain combinatorial and topological. We note that others [29, 31] have obtained results in general combinatorial settings, though we will focus on the impact of the topology of X on sampling requirements.

Let us now generalize to the case of solutions to (10.3) over a graph X, written as a 1-dimensional abstract simplicial complex. In order for the space of solutions to be well defined, it is necessary to assign a *length* to each edge. We shall write $L(e)$ for the length of edge e, which is a positive real number. We will make use of the orientation of the edges of X to help keep track of the direction waves are moving along them. Since the differential equation is insensitive to orientation, this is merely a bookkeeping tool – the orientations of the edges are arbitrary.

The differential equation (10.3) as specified along the edges of X requires boundary conditions at each vertex in order to have unique solutions. There are many possible conditions that can be placed at each vertex; for concreteness, we will consider Kirchhoff conditions [21]:

1. The solution is continuous and
2. The sum of the derivatives at a vertex (facing outward) is zero.

These two conditions limit the number of complex degrees of freedom in specifying a solution at a vertex to be the degree of the vertex, as follows. Suppose that a vertex v of X has degree n. Without loss of generality, suppose that the edges are given by $\{v_1, v\}, \cdots, \{v_n, v\}$ and that the solutions to (10.3) on these edges are given by

$$\begin{aligned} u_1 &= c_{1,1}e^{ikx} + c_{1,2}e^{-ikx}, \\ &\vdots \\ u_n &= c_{n,1}e^{ikx} + c_{n,2}e^{-ikx}, \end{aligned}$$

where x is the distance from vertex v. There are therefore $2n$ complex degrees of freedom in the above equations. The continuity condition (1) above means that

$$\begin{aligned} u_1(0) = \cdots = u_n(0) \text{ so that} \\ c_{1,1} + c_{1,2} = \cdots = c_{n,1} + c_{n,2}, \end{aligned}$$

which reduces the number of degrees of freedom to $n + 1$. The derivative condition (2) above is given by

$$0 = \sum_{m=1}^{n} u'_m(0) = \sum_{m=1}^{n} ik\left(c_{m,1} - c_{m,2}\right),$$

which further reduces the degrees of freedom to n. Without loss of generality, this means that a solution to (10.3) at v is determined by the incoming wave amplitudes $c_{1,1}, \cdots, c_{n,1}$. Given these amplitudes, the others can be computed. (For a derivation of the appropriate formulas, see for instance [32].)

The number of degrees of freedom at an edge or vertex determines the dimension of the stalks of the sheaf. The restriction maps are given by the formulas for computing the outgoing wave amplitudes $c_{1,2}, \cdots, c_{n,2}$ from the incoming wave amplitudes, and they additionally account for the propagation of the phase of the waves along each edge.

This sheaf is called a transmission line sheaf and is defined as follows.

Definition 11. The *transmission line sheaf* $\mathscr{T}$ on X has stalks given by

1. $\mathscr{T}(e) = \mathbb{C}^2$ over each edge e and
2. $\mathscr{T}(v) = \mathbb{C}^{\deg v}$ over each vertex v, whose degree is $\deg v$.

If e_m is the mth edge attached to vertex v, the restriction $\mathscr{T}(v \nrightarrow e_m)$ is given by

$$\mathscr{T}(v \nrightarrow e_m)(u_1, \ldots, u_{\deg v}) = \begin{cases} \left(u_m, e^{-ikL(e_m)}\left(\frac{2}{\deg v}\sum_{j=1}^{\deg v} u_j - u_m\right)\right) & \text{if}[e_m : v] = 1; \\ \left(e^{ikL(e_m)}\left(\frac{2}{\deg v}\sum_{j=1}^{\deg v} u_j - u_m\right), u_m\right) & \text{if}[e_m : v] = -1. \end{cases}$$

The transmission line sheaf encodes the space of solutions to (10.3) in that every global section is a solution and *vice versa*. Reconstruction from samples of a transmission line sheaf therefore corresponds to solving a specific boundary value problem.

The easiest sampling result for transmission line sheaves follows immediately from Proposition 1, namely that a global section of a transmission line sheaf $\mathscr{T}$ on X is completely specified by its values on the vertices of X. This result is clearly inefficient; merely consider the simplicial complex X for the real line with vertices $X^0 = \mathbb{Z}$ and edges $X^1 = \{n, n+1\}$. We have already seen that the space of sections of a transmission line sheaf on X is merely $\mathbb{C}^2$, yet Proposition 1 would have us collect samples at infinitely many vertices. The missing insight is that the topology of X impacts the global sections of a transmission line sheaf. Changing the edge length in the simplicial complex model for $\mathbb{R}$ does not change the space of global sections of a transmission line sheaf. Another situation in which edge length does not matter is shown in the next example.

Example 14. Consider the space Y shown at right in Figure 10.12. The coboundary map d^0 for a general sheaf is given by (10.1). It is a block matrix, in which each block is a restriction map between attached faces, and whose sign comes from the orientation index. The rows correspond to the stalks over each edge, and the columns correspond to the stalks over each vertex. In the case of a transmission line sheaf over Y, this matrix has the form

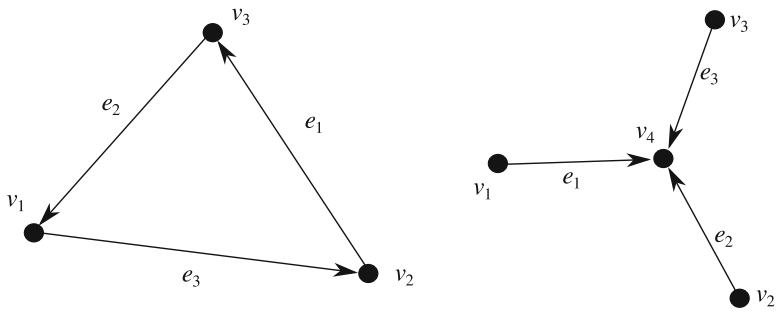


Fig. 10.12 The graph X for Example 15 (left) and Y for Example 14 (right)

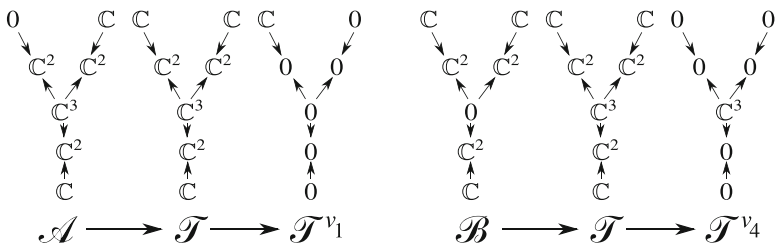


Fig. 10.13 Sampling a star at a leaf (left) and at its center (right)

$$
\begin{array}{c|c|c|c|cc|cc|}
 & \mathscr{T}(v_1) & \mathscr{T}(v_2) & \mathscr{T}(v_3) & \multicolumn{5}{c|}{\mathscr{T}(v_4)} \\
\mathscr{T}(e_1) & -e^{ikL(e_1)} & 0 & 0 & \multicolumn{2}{c|}{1} & \multicolumn{2}{c|}{0} \\
 & -1 & 0 & 0 & \multicolumn{2}{c|}{-\tfrac{1}{3}e^{-ikL(e_1)}} & \multicolumn{2}{c|}{\tfrac{2}{3}e^{-ikL(e_1)}} \\
\mathscr{T}(e_2) & 0 & -e^{ikL(e_2)} & 0 & \multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{1} \\
 & 0 & -1 & 0 & \multicolumn{2}{c|}{\tfrac{2}{3}e^{-ikL(e_2)}} & \multicolumn{2}{c|}{-\tfrac{1}{3}e^{-ikL(e_2)}} \\
\mathscr{T}(e_3) & 0 & 0 & -e^{ikL(e_3)} & \multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{0} \\
 & 0 & 0 & -1 & \multicolumn{2}{c|}{\tfrac{2}{3}e^{-ikL(e_3)}} & \multicolumn{2}{c|}{\tfrac{2}{3}e^{-ikL(e_3)}}
\end{array}
\quad
\begin{array}{|cc|}
\multicolumn{2}{|c|}{\mathscr{T}(v_4)} \\
0 & 0 \\
\tfrac{2}{3}e^{-ikL(e_1)} & \tfrac{2}{3}e^{-ikL(e_1)} \\
0 & 0 \\
\tfrac{2}{3}e^{-ikL(e_2)} & \tfrac{2}{3}e^{-ikL(e_2)} \\
1 & 0 \\
-\tfrac{1}{3}e^{-ikL(e_3)} & \tfrac{2}{3}e^{-ikL(e_3)}
\end{array}
$$

(Notice the presence of the minus signs in the blocks corresponding to the attachments $v_1 \rightsquigarrow e_1$, $v_2 \rightsquigarrow e_2$, and $v_3 \rightsquigarrow e_3$ due to the orientation indices $[e_1, v_1] = -1$, $[e_2, v_2] = -1$, and $[e_3, v_3] = -1$.)

Because d^0 has rank 5, $\dim H^0(Y; \mathscr{T}) = 1$ for any transmission sheaf $\mathscr{T}$ regardless of edge lengths so that the space of solutions to (10.3) with Kirchhoff conditions is 1 dimensional.

This means that reconstruction of sections requires at least one dimension of measurements, a lower bound. If we consider a sampling morphism $\mathscr{T} \to \mathscr{T}^Z$ for some set of vertices Z, this induces an injective map on global sections. To see this, consider sampling at any one of the leaf nodes or at the center.

If we sample at a leaf node only, as shown in Figure 10.13 at left, the ambiguity sheaf $\mathscr{A}$ has a coboundary matrix given by

$$
\begin{array}{c|c|c|cc|cc|}
 & \mathscr{A}(v_2) & \mathscr{A}(v_3) & \multicolumn{5}{c|}{\mathscr{A}(v_4)} \\
\mathscr{A}(e_1) & 0 & 0 & \multicolumn{2}{c|}{1} & \multicolumn{2}{c|}{0} \\
 & 0 & 0 & \multicolumn{2}{c|}{-\tfrac{1}{3}e^{-ikL(e_1)}} & \multicolumn{2}{c|}{\tfrac{2}{3}e^{-ikL(e_1)}} \\
\mathscr{A}(e_2) & -e^{ikL(e_2)} & 0 & \multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{1} \\
 & -1 & 0 & \multicolumn{2}{c|}{\tfrac{2}{3}e^{-ikL(e_2)}} & \multicolumn{2}{c|}{-\tfrac{1}{3}e^{-ikL(e_2)}} \\
\mathscr{A}(e_3) & 0 & -e^{ikL(e_3)} & \multicolumn{2}{c|}{0} & \multicolumn{2}{c|}{0} \\
 & 0 & -1 & \multicolumn{2}{c|}{\tfrac{2}{3}e^{-ikL(e_3)}} & \multicolumn{2}{c|}{\tfrac{2}{3}e^{-ikL(e_3)}}
\end{array}
\quad
\begin{array}{|cc|}
\multicolumn{2}{|c|}{\mathscr{A}(v_4)} \\
0 & 0 \\
\tfrac{2}{3}e^{-ikL(e_1)} & \tfrac{2}{3}e^{-ikL(e_1)} \\
0 & 0 \\
\tfrac{2}{3}e^{-ikL(e_2)} & \tfrac{2}{3}e^{-ikL(e_2)} \\
1 & 0 \\
-\tfrac{1}{3}e^{-ikL(e_3)} & \tfrac{2}{3}e^{-ikL(e_3)}
\end{array}
$$

which has rank 5, implying that the ambiguity sheaf has only trivial global sections. Again, this means that there are only trivial solutions to (10.3) with Kirchhoff conditions.

On the other hand, sampling at the center yields a different ambiguity sheaf $\mathscr{B}$, whose coboundary matrix is

$$\begin{array}{c} \\ \mathscr{B}(e_1) \\ \\ \mathscr{B}(e_2) \\ \\ \mathscr{B}(e_3) \\ \end{array} \begin{array}{c} \mathscr{B}(v_4) \\ \left(\begin{array}{ccc} -e^{ikL(e_1)} & 0 & 0 \\ -1 & 0 & 0 \\ \hline 0 & -e^{ikL(e_2)} & 0 \\ 0 & -1 & 0 \\ \hline 0 & 0 & -e^{ikL(e_3)} \\ 0 & 0 & -1 \end{array}\right) \end{array},$$

which also has trivial kernel.

If we instead consider a different topology, for instance a circle, then edge lengths do have an impact on the global sections of the resulting transmission line sheaf as the following example shows. When wave solutions to (10.3) traverse a loop they must start and end the loop with the same phase – if not, they will violate the Kirchhoff continuity condition.

Example 15. Consider the simplicial complex X shown at left in Figure 10.12, in which the edges are oriented as marked. Because X has a nontrivial loop, the lengths of the edges impact the space of global sections of a transmission line sheaf over X. If $\mathscr{T}$ is a transmission line sheaf, its coboundary $d^0 : C^0(X;\mathscr{T}) \to C^1(X;\mathscr{T})$ is given by

$$\begin{array}{c} \\ \mathscr{T}(e_1) \\ \\ \mathscr{T}(e_2) \\ \\ \mathscr{T}(e_3) \\ \end{array} \begin{array}{c} \mathscr{T}(v_1) \qquad\qquad \mathscr{T}(v_2) \qquad\qquad \mathscr{T}(v_3) \\ \left(\begin{array}{cc|cc|cc} 0 & 0 & -e^{ikL(e_1)} & 0 & 1 & 0 \\ 0 & 0 & 0 & -1 & 0 & e^{-ikL(e_1)} \\ \hline 1 & 0 & 0 & 0 & -e^{ikL(e_2)} & 0 \\ 0 & e^{-ikL(e_2)} & 0 & 0 & 0 & -1 \\ \hline -e^{ikL(e_3)} & 0 & 1 & 0 & 0 & 0 \\ 0 & -1 & 0 & e^{-ikL(e_3)} & 0 & 0 \end{array}\right) \end{array}.$$

This matrix has full rank unless $e^{-ik(L(e_1)+L(e_2)+L(e_3))} = 1$, a condition called *resonance*. Therefore, the space of global sections of $\mathscr{T}$ has dimension

$$\dim H^0(X;\mathscr{T}) = \begin{cases} 2 & \text{if } k\ (L(e_1)+L(e_2)+L(e_3)) \in 2\pi\mathbb{Z}; \\ 0 & \text{otherwise,} \end{cases}$$

and an easy calculation shows that sampling at any one of the vertices results in an injective map on global sections. This means that the space of solutions to (10.3) with Kirchhoff conditions is sensitive to edge lengths when there is a loop – if the phase of a wave does not match its value after propagating around a loop, it is not a solution.

Based on the previous examples, a sound procedure is to consider the dimension of the space of global sections of $\mathscr{T}$ to be a lower bound on how much information is to be obtained through sampling. (Clearly, this may not be enough in some

situations, especially if the sampling morphisms are not injective on stalks.) As described in [32], a general lower bound on the dimension of $H^0(\mathscr{T})$ is $n + 1$, where n is the number of resonant loops. (A tighter lower bound exists, but its expression is complicated by the presence of degree 1 vertices.) Therefore, topology plays an important role in acquiring enough information to recover global sections of $\mathscr{T}$ from samples.

10.5.3 Polynomial splines

Section 10.5.1 showed how limiting the support of the Fourier transform of a function permitted it to be reconstructed by its values at a discrete subset. Because of the Paley-Wiener theorem, the smoothness of a function is reflected in the decay of its Fourier transform. On the other hand, Section 10.5.2 showed that applying smoothness constraints directly to the function also enables perfect recovery. This suggests that as we consider smoother functions, we can reconstruct them from more widely spaced samples.

In this section, we consider sampling from polynomial splines, which are functions whose smoothness is explicitly controlled. A C^k degree n polynomial spline has k continuous derivatives and is constructed piecewise from degree n polynomial segments (see Figure 10.14). Because of this, a polynomial spline is infinitely differentiable on all of its domain except at a discrete set of *knot points*, where it has k continuous derivatives.

Unlike variational splines traditionally used for approximation, the splines in this section employ a fixed set of knot points. The fixed choice of knot points means that in effect we are explicitly choosing a coarse topology that differs from the usual topology. The polynomial splines that are constructed in this section are local with respect to this coarse topology and not with respect to the usual topology. The reader is cautioned that if the knot points are allowed to vary, then the resulting space of splines is sensitive to *both* local and global topological features, since they

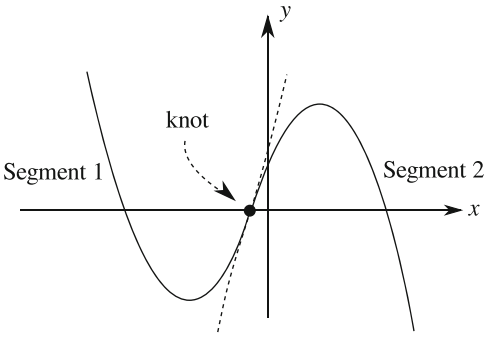


Fig. 10.14 A polynomial spline with two quadratic segments, joined at a knot with continuous first derivatives

can approximate solutions to differential equations. For instance, the discussion on quantum graphs given above can be lifted to the context of approximating splines, as is discussed in [25, 27, 30].

Consider a degree n polynomial spline that has two knots: one at 0 and one at L. Require it to have $(n-1)$ continuous derivatives across its three segments: $(-\infty, 0)$, $(0, L)$, and (L, ∞). To obtain $n-1$ continuous derivatives at $x = 0$, such a spline should have the form

$$f(x) = \begin{cases} a_n^- x^n + \sum_{k=0}^{n-1} a_k x^k & \text{for } x \le 0; \\ a_n^+ x^n + \sum_{k=0}^{n-1} a_k x^k & \text{for } 0 \le x \le L. \end{cases}$$

In a similar way, to obtain $n-1$ continuous derivatives at $x = L$, the spline should be of the form

$$f(x) = \begin{cases} b_n^-(x-L)^n + \sum_{k=0}^{n-1} b_k (x-L)^k & \text{for } 0 \le x \le L; \\ b_n^+(x-L)^n + \sum_{k=0}^{n-1} b_k (x-L)^k & \text{for } x \ge L. \end{cases}$$

But clearly on $0 \le x \le L$, these two definitions should agree so that f is a well-defined function. This means that for all x,

$$\begin{aligned} a_n^+ x^n + \sum_{k=0}^{n-1} a_k x^k &= b_n^-(x-L)^n + \sum_{k=0}^{n-1} b_k (x-L)^k \\ &= \sum_{i=0}^{n-1} b_i \sum_{k=0}^{i} \binom{i}{k} x^k (-L)^{i-k} + b_n^- \sum_{k=0}^{n} \binom{n}{k} x^k (-L)^{n-k} \\ &= \sum_{k=0}^{n-1} x^k \sum_{i=k}^{n-1} \binom{i}{k} (-L)^{i-k} b_i + \sum_{k=0}^{n} x^k \binom{n}{k} (-L)^{n-k} b_n^- \\ &= \sum_{k=0}^{n-1} x^k \left(\sum_{i=k}^{n-1} \binom{i}{k} (-L)^{i-k} b_i + \binom{n}{k} (-L)^{n-k} b_n^- \right) + b_n^- x^n. \end{aligned}$$

By linear independence, this means that

$$\begin{aligned} a_n^+ &= b_n^- \\ a_k &= \sum_{i=k}^{n-1} \binom{i}{k} (-L)^{i-k} b_i + \binom{n}{k} (-L)^{n-k} b_n^-. \end{aligned}$$

Notice that if the a variables are given, then this is a triangular system for the b variables.

Using this computation, we can define $\mathscr{PS}^n$, the C^{n-1} sheaf of n-degree polynomial splines on $\mathbb{R}$ with knots at each of the integers. This sheaf is built on

the simplicial complex X for $\mathbb{R}$, whose vertices are $X^0 = \mathbb{Z}$ and whose edges are $X^1 = \{m, m+1\}$. Because a degree n spline at any given knot can be defined by $n+2$ real values on the two segments adjacent to that knot, we assign

$$\mathscr{PS}^n(\{m\}) = \mathbb{R}^{n+2}$$

for each $m \in \mathbb{Z}$. We will think of these as defining $(a_0, \ldots, a_{n-1}, a_n^-, a_n^+)$ in our calculation above. The spline on each segment is merely a degree n polynomial so that

$$\mathscr{PS}^n(\{m, m+1\}) = \mathbb{R}^{n+1}.$$

For each knot, there are two restriction maps: one to the left and one to the right. They are given by

$$\mathscr{PS}^n(\{m\} \rightsquigarrow \{m, m+1\}) = \begin{pmatrix} 1 \cdots 0 & 0 & 0 \\ \vdots \quad\quad & \vdots & \vdots \\ 0 \cdots 1 & 0 & 0 \\ 0 \cdots 0 & 0 & 1 \end{pmatrix}$$

and

$$\mathscr{PS}^n(\{m+1\} \rightsquigarrow \{m, m+1\}) = \begin{pmatrix} 1 & -L & L^2 & -L^3 & \cdots & (-L)^n & 0 \\ & 1 & -2L & 3L^2 & \cdots & \binom{n}{1}(-L)^{n-1} & 0 \\ & & 1 & -3L & \cdots & \binom{n}{2}(-L)^{n-2} & 0 \\ & & & 1 & \cdots & \binom{n}{3}(-L)^{n-3} & 0 \\ & & & & & \vdots & \vdots \\ & & & & & \binom{n}{k}(-L)^{n-k} & 0 \\ & & & & & \vdots & \vdots \\ & & & & & 1 & 0 \end{pmatrix}.$$

Remark 5. Observe that $L = 1$, $\mathscr{PS}^1$ reduces to the sheaf $\mathscr{PL}$ given in Example 9 for the special case of the graph being a line (so all vertices have degree 2).

Lemma 1. *Consider $\mathscr{PS}^n$ on the simplicial complex shown in Figure 10.15, which has $k+2$ vertices and $k+1$ edges. The sheaf has nontrivial global sections*

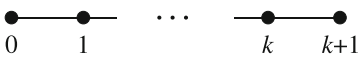


Fig. 10.15 The simplicial complex used in Lemma 1

that vanish at the endpoints if and only if $k > n+1$. If $k \leq n+1$, then the only global section which vanishes at both endpoints is the zero section.

Proof. Consider the ambiguity sheaf $\mathscr{A}$ associated to the sampling morphism $\mathscr{PS}^n \to (\mathscr{PS}^n)^Y$ where Y consists of the two endpoints. Observe that the global sections of $\mathscr{A}$ correspond to global sections of $\mathscr{PS}^n$ that vanish at the endpoints, so the lemma follows by reasoning about $H^0(\mathscr{A})$. The matrix for the coboundary map $d^0 : C^0(\mathscr{A}) \to C^1(\mathscr{A})$ has a block structure

$$\begin{pmatrix} A & 0 & & 0 \\ B & A & \cdots & 0 \\ 0 & B & & 0 \\ & & & \vdots \\ & & & A \\ & & & B \end{pmatrix},$$

where the $(n+1) \times (n+2)$ blocks are

$$A = \begin{pmatrix} 1 & \cdots & 0 & 0 & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & 1 & 0 & 0 \\ 0 & \cdots & 0 & 0 & 1 \end{pmatrix}, \quad B = \begin{pmatrix} -1 & \cdots & * & * & 0 \\ \vdots & & \vdots & \vdots & \vdots \\ 0 & \cdots & -1 & * & 0 \\ 0 & \cdots & 0 & -1 & 0 \end{pmatrix}.$$

Clearly both such blocks are of full rank. Thus the coboundary matrix has a nontrivial kernel whenever it has more rows than columns. Namely, whenever the following is satisfied:

$$\begin{aligned} (n+2)k &> (n+1)(k+1) \\ nk + 1k &> nk + n + k + 1 \\ k &> n+1 \end{aligned}$$

as desired. □

This lemma implies that unambiguous reconstruction from samples is possible provided the gaps between samples are small enough. Increased smoothness allows the gaps to be larger without inhibiting reconstruction. Because of this, it is convenient to define distances between vertices.

Definition 12. On a graph G, define the *edge distance* between two vertices v, w to be

$$\mathrm{ed}(v,w) = \begin{cases} \min_p\{\#\text{ edges in } p \text{ such that } p \text{ is a} \\ \mathrm{PL}-\text{continuous path from } v \to w\} \text{ or} \\ \infty \text{if no such path exists.} \end{cases}$$

From this, the maximal distance to a vertex set Y is

$$\mathrm{med}(Y) = \max_{x\in X^0}\{\min_{y\in Y} \mathrm{ed}(x,y)\}.$$

Corollary 3. *Suppose that $Y \subseteq \mathbb{Z}$, which we take to be a subset of vertices of X. If $\mathrm{med}(Y) \leq n+2$, then the sampling morphism $\mathscr{PS}^n \to (\mathscr{PS}^n)^Y$ induces an injective linear map on global sections.*

If we instead consider sampling of polynomial splines on the circle, very little changes. The proof of Lemma 1 does not change at all. However, it is tedious to ensure the continuity of many derivatives on general abstract simplicial complexes, and there are many inequivalent sets of continuity conditions[3] that make sense.

Because of the limited impact of the topology on splines, the rest of this section will focus on the sheaf of piecewise linear functions $\mathscr{PL}$. This allows us to examine splines over spaces with nontrivial topology in the base space, since the results are fundamentally similar for most generalizations of $\mathscr{PS}^n$. We will focus on the special case of a sampling morphism $s : \mathscr{PL} \to \mathscr{PL}^Y$ where Y is a subset of the vertices of X. Excluding one or two vertices from Y does not prevent reconstruction in this case, because the samples include information about slopes along adjacent edges.

Lemma 2. *Consider $\mathscr{PL}_Y$, the subsheaf of $\mathscr{PL}$ whose sections vanish on a vertex set Y, and the graphs G_1, G_2, and G_3 as shown in Figure 10.16. There are no nontrivial sections of $\mathscr{PL}_Y$ on G_1 and G_2, but there are nontrivial sections of $\mathscr{PL}_Y$ on G_3.*

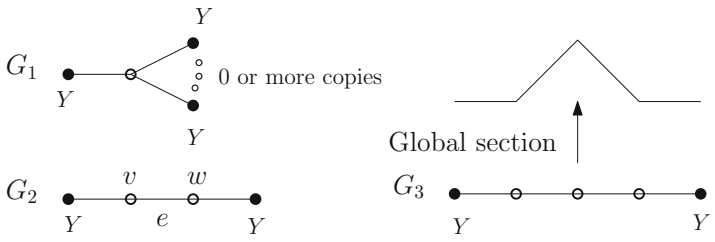


Fig. 10.16 Graphs G_1 and G_2 (left) and G_3 (right) for Lemma 2. Filled vertices represent elements of Y, empty ones are in the complement of Y.

[3]The interested reader should consider [12] for a practical discussion of several of these conditions in dimensions 1 and 2.

Proof. If a section of $\mathscr{PL}$ vanishes at a vertex x with degree n, this means that the value of the section there is an $(n+1)$-dimensional zero vector. The value of the section on every edge adjacent to x is then the 2-dimensional zero vector. Since the dimensions in each stalk of $\mathscr{PL}$ represent the value of the piecewise linear function and its slopes, linear extrapolation to the center vertex in G_1 implies that its value is zero too.

Lemma 1 shows that G_2 has no nontrivial sections. It also shows that G_3 has nontrivial sections, which are spanned by the one shown in Figure 10.16. □

Proposition 4 (Unambiguous sampling). *Consider the sheaf $\mathscr{PL}$ on a graph X and $Y \subseteq X^0$. Then $H^0(X; \mathscr{F}_Y) = 0$ if and only if* $\text{med}(Y) \leq 1$.

Proof. ($\Leftarrow$) Suppose that $x \in X^0 \backslash Y$ is a vertex not in Y. Then there exists a path with one edge connecting it to Y. Whence we are in the case of G_1 of Lemma 2, so any section at x must vanish.

($\Rightarrow$) By contradiction. Assume $\text{med}(Y) > 1$. Without loss of generality, consider $x \in X^0 \backslash Y$, whose distance to Y is exactly 2. Then one of the subgraphs shown in Figure 10.17 must be present in X. The case of G_3 in Lemma 2 makes it each of these has nontrivial sections at x, merely looking at sections over the subgraph. □

Proposition 5 (Nonredundant sampling). *Consider the case of $s : \mathscr{PL} \to \mathscr{PL}^Y$. If $Y = X^0$, then $H^1(X; \mathscr{A}) \neq 0$. If Y is such that* $\text{med}(Y) \leq 1$ *and the number of edges*[4] $|X^1|$ *satisfies* $|X^0 \backslash Y| + \sum_{y \notin Y} \deg y = 2|X^1|$, *then* $H^1(X; \mathscr{A}) = 0$.

Proof. The stalk of $\mathscr{A}$ over each edge is $\mathbb{R}^2$, and the stalk over a vertex in Y is trivial. However, the stalk over a vertex of degree n not in Y is $\mathbb{R}^{n+1}$. Observe that if $H^0(X; \mathscr{A}) = 0$, then $H^1(X; \mathscr{A}) = C^1(X; \mathscr{A})/C^0(X; \mathscr{A})$. Using the degree sum formula in graph theory, we compute that $H^1(X; \mathscr{A})$ has dimension $2|X^1| - \sum_{y \notin Y}(\deg y + 1)$. □

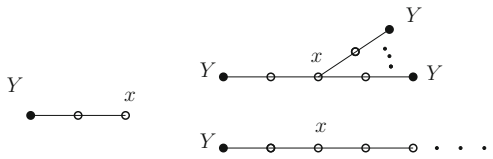


Fig. 10.17 The three families of subgraphs that arise when $\text{med}(Y) > 1$. Filled vertices represent elements of Y, empty ones are in the complement of Y.

[4] $|A|$ represents the cardinality of a set A.

10.6 Conclusions

This chapter has shown that exact sequences of sheaves are a unifying principle for sampling theory. These tools reveal general, precise conditions under which reconstruction succeeds and are not limited to uniform sampling or other symmetries. Several sampling problems for bandlimited and non-bandlimited functions were discussed. The use of sheaves in sampling is essentially unexplored otherwise, and there remain many open questions. We discuss two such questions here with a relationship to bandlimited functions:

1. "Is the $B < 1/2$ in the Nyquist theorem invariant with respect to changes in topology and geometry?"
2. "Is there a general notion of bandwidth for sheaves of Hilbert spaces?"

The first question is connected to the existing literature on sampling. We showed that although the number of samples required for the reconstruction of a bandlimited function may vary, the sampling *rate* necessary appears to be the same. We found that the bandwidth B required for reconstructing functions on the real line and the circle was constrained to be less than $1/2$. Others (for instance [26, 31]) have found that this remains the case for other domains as well.

The second question is a bit more subtle. If a function is bandlimited, this means that its Fourier transform has bounded support. The Fourier transform is intimately related to the spectrum of the Laplacian operator. On the other hand, because cohomological obstructions play a role in sampling, we had to consider the cochain complex for sheaves of functions. These two threads of study are in fact closely related through Hodge theory via the study of Hilbert complexes [5]. Indeed, if $\mathscr{F}$ and $\mathscr{S}$ are sheaves of Hilbert spaces then their cochain complexes are Hilbert complexes, and they have a Hodge decomposition, along with an associated Laplacian operator.

Acknowledgements This work was partly supported under Federal Contract No. FA9550-09-1-0643. The author also wishes to thank the editor for the invitation to write this chapter. Portions of this chapter appeared in the proceedings of SampTA 2013, published by EURASIP. Finally, the author wishes to thank Isaac Pesenson for insightful comments that greatly improved the readability of this chapter.

References

1. M. Baker, X. Faber, Metrized graphs, Laplacian operators, and electrical networks, in *Quantum Graphs and Their Applications* (American Mathematical Society, Providence, 2006), pp. 15–34
2. M. Baker, R. Rumely, Harmonic analysis on metrized graphs. Can. J. Math. **59**(2), 225–275 (2007)
3. J. Benedetto, W. Heller, Irregular sampling and the theory of frames: I. Note Math. **10**(1), 103–125 (1990)
4. G. Bredon, *Sheaf Theory* (Springer, New York, 1997)

5. J. Brüning, M. Lesch, Hilbert complexes. J. Funct. Anal. **108**, 88–132 (1992)
6. F. Chazal, D. Cohen-Steiner, A. Lieutier, A sampling theory for compact sets in euclidean space. Discret. Comput. Geom. **41**, 461–479 (2009)
7. J. Curry, Sheaves, cosheaves and applications (2013) [arxiv:1303.3255]
8. J. Curry, R. Ghrist, M. Robinson, Euler calculus and its applications to signals and sensing, in *Proceedings of Symposia in Applied Mathematics: Advances in Applied and Computational Topology*, ed. by A. Zomorodian (American Mathematical Society, Providence, 2012)
9. P. Dragotti, M. Vetterli, T. Blue, Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang–Fix. IEEE Trans. Signal Process **55**(5), (2007)
10. M. Ebata, M. Eguchi, S. Koizumi, K. Kumahara, Analogues of sampling theorems for some homogeneous spaces. Hiroshima Math. J. **36**, 125–140 (2006)
11. L. Ehrenpreis, Sheaves and differential equations. Proc. Am. Math. Soc. **7**(6), 131–1138 (1956)
12. G. Farin, *Curves and Surfaces for CAGD* (Elsevier, Amsterdam, 1985)
13. H.G. Feichtinger, K. Gröchenig, Theory and practice of irregular sampling. *Wavelets: Mathematics and Applications* (CRC Press, Boca Raton, 1994), pp. 305–363
14. H. Feichtinger, I. Pesenson, A reconstruction method for band-limited signals on the hyperbolic plane. Sampling Theory Signal Image Process. **4**(2), 107–119 (2005)
15. R. Ghrist, Y. Hiraoka, Applications of sheaf cohomology and exact sequences to network coding, in *Proc. NOLTA* (2011)
16. R. Godement, *Topologie algebrique et théorie des faisceaux* (Herman, Paris, 1958)
17. K. Gröchening, Reconstruction algorithms in irregular sampling. Math. Comput. **59**(199), 181–194 (1992)
18. A. Hatcher, *Algebraic Topology* (Cambridge University Press, Cambridge, 2002)
19. J.H. Hubbard, *Teichmüller Theory*, vol. 1 (Matrix Editions, Ithaca, 2006)
20. B. Iverson, *Cohomology of Sheaves*. Aarhus universitet, Matematisk institut (1984)
21. P. Kuchment, Quantum graphs: an introduction and a brief survey, in *Analysis on Graphs and Its Applications*. (Isaac Newton Institute for Mathematical Sciences, Cambridge, 2007), pp. 291–312
22. J. Lilius, Sheaf semantics for Petri nets. Technical report, Helsinki University of Technology, Digital Systems Laboratory (1993)
23. P. Niyogi, S. Smale, S. Weinberger, Finding the homology of submanifolds with high confidence from random samples, in *Twentieth Anniversary Volume*, ed. by R. Pollack, J. Pach, J.E. Goodman (Springer, New York, 2009), pp. 1–23
24. I. Pesenson, Sampling of band-limited vectors. J. Fourier Anal. Appl. **7**(1), 93–100 (2001)
25. I. Pesenson, An approach to spectral problems on riemannian manifolds. Pac. J. Math. **215**(1), 183–199 (2004)
26. I. Pesenson, Band limited functions on quantum graphs. Proc. Am. Math. Soc. **133**(12), 3647–3656 (2005)
27. I. Pesenson, Polynomial splines and eigenvalue approximations on quantum graphs. J. Approx. Theory **135**(2), 203–220 (2005)
28. I. Pesenson, Analysis of band-limited functions on quantum graphs. Appl. Comput. Harmon. Anal. **21**(2), 230–244 (2006)
29. I. Pesenson, Sampling in Paley-Wiener spaces on combinatorial graphs. Trans. Am. Math. Soc. **360**(10), 5603 (2008)
30. I.Z. Pesenson, Removable sets and approximation of eigenvalues and eigenfunctions on combinatorial graphs. Appl. Comput. Harmon. Anal. **29**, 123–133 (2010)
31. I.Z. Pesenson, M.Z. Pesenson, Sampling, filtering and sparse approximations on combinatorial graphs. J. Fourier Anal. Appl. **16**(6), 921–942 (2010)
32. M. Robinson, Inverse problems in geometric graphs using internal measurements (2010) [arxiv:1008.2933]
33. M. Robinson, Asynchronous logic circuits and sheaf obstructions. Electron. Notes Theor. Comput. Sci. **283**, 159–177 (2012)
34. M. Robinson, Understanding networks and their behaviors using sheaf theory, in *GlobalSIP* (2013)

35. M. Robinson, *Topological Signal Processing* (Springer, Heidelberg, 2014)
36. A. Schuster, D. Varolin, Interpolation and sampling for generalized Bergman spaces on finite Riemann surfaces. Rev. Mat. Iberoam.. **24**(2), 499–530 (2008)
37. A. Shepard, A cellular description of the derived category of a stratified space. Ph.D. thesis, Brown University, 1980
38. S. Smale, D.X. Zhou, Shannon sampling and function reconstruction from point values. Bull. Am. Math. Soc. **41**(3), 279–306 (2004)
39. M. Unser, Splines: a perfect fit for signal and image processing. IEEE Signal Process. Mag. **16**(6), 22–38 (1999)
40. M. Unser, Sampling–50 years after Shannon. Proc. IEEE **88**(4), 569–587 (2000)
41. M. Unser, J. Zerubia, A generalized sampling theory without band-limiting constraints. IEEE Trans. Circuits Syst. II Analog Digit. Signal Process. **45**(8), 959–969 (1998)
42. R.G. Vaughan, N.L. Scott, D.R. White, The theory of bandpass sampling. IEEE Trans. Signal Process. **39**(9), 1973–1984 (1991)
43. M. Vetterli, P. Marziliano, T. Blu, Sampling signals with finite rate of innovation. IEEE Trans. Signal Process. **50**(6), 1417–1428 (2002)
44. S. Zhang, Admissible pairing on a curve. Invent. Math. **112**(1), 171–193 (1993)

Part IV
Sampling and Parametric Partial Differential Equations

Chapter 11
How To Best Sample a Solution Manifold?

Wolfgang Dahmen

Abstract Model reduction attempts to guarantee a desired "model quality," e.g. given in terms of accuracy requirements, with as small a model size as possible. This chapter highlights some recent developments concerning this issue for the so-called Reduced Basis Method (RBM) for models based on parameter-dependent families of PDEs. In this context the key task is to sample the *solution manifold* at judiciously chosen parameter values usually determined in a *greedy fashion*. The corresponding *space growth* concepts are closely related to the so-called *weak greedy* algorithms in Hilbert and Banach spaces which can be shown to give rise to convergence rates comparable to the best possible rates, namely the *Kolmogorov n-width* rates. Such algorithms can be interpreted as *adaptive sampling* strategies for approximating compact sets in Hilbert spaces. We briefly discuss the results most relevant for the present RBM context. The applicability of the results for weak greedy algorithms has however been confined so far essentially to well-conditioned coercive problems. A critical issue is therefore an extension of these concepts to a wider range of problem classes for which the conventional methods do not work well. A second main topic of this chapter is therefore to outline recent developments of RBMs that do realize n-width rates for a much wider class of variational problems covering indefinite or singularly perturbed unsymmetric problems. A key element in this context is the design of *well-conditioned variational formulations* and their numerical treatment via saddle point formulations. We conclude with some remarks concerning the relevance of uniformly approximating the whole solution manifold also when the *quantity of interest* is only the value of a *functional* of the parameter-dependent solutions.

1991 *Mathematics Subject Classification.* 65J10, 65N12, 65N15, 35B30

This work has been supported in part by the DFG SFB-Transregio 40, and by the DFG Research Group 1779, the Excellence Initiative of the German Federal and State Governments, and NSF grant DMS 1222390.

W. Dahmen (✉)
Institut für Geometrie und Praktische Mathematik, RWTH Aachen, Aachen, Germany
e-mail: dahmen@igpm.rwth-aachen.de

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_11

11.1 Introduction

Many engineering applications revolve around the task of identifying a configuration that in some sense best fits certain objective criteria under certain constraints. Such design or optimization problems typically involve (sometimes many) *parameters* that need to be chosen so as to satisfy given optimality criteria. An optimization over such a parameter domain usually requires a frequent evaluation of the states under consideration which typically means to frequently solve a *parameter-dependent* family of operator equations

$$B_y u = f, \quad y \in \mathcal{Y}. \tag{11.1}$$

In what follows the parameter set $\mathcal{Y}$ is always assumed to be a compact subset of $\mathbb{R}^p$ for some fixed $p \in \mathbb{N}$ and B_y should be thought of as a (linear) partial differential operator whose coefficients depend on the parameters $y \in \mathcal{Y}$. Moreover, B_y is viewed as an operator taking some Hilbert space U one-to-one and onto the *normed dual* V' of some (appropriate) Hilbert space V where U and V are identified through a variational formulation of (11.1) as detailed later, see for instance (11.30). Recall also that the normed dual V' is endowed with the norm

$$\|w\|_{V'} := \sup_{v \in V, v \neq 0} \frac{\langle w, v \rangle}{\|v\|_V}, \tag{11.2}$$

where $\langle \cdot, \cdot \rangle$ denotes the dual pairing between V and V'.

Given a parametric model (11.1) the above mentioned design or optimization problems concern now the *states* $u(y) \in U$ which, as a function of the parameters $y \in \mathcal{Y}$, form what we refer to as the *solution manifold*

$$\mathcal{M} := \{u(y) := B_y^{-1} f : y \in \mathcal{Y}\}. \tag{11.3}$$

Examples of (11.1) arise, for instance, in geometry optimization when a transformation of a variable finitely parametrized domain to a reference domain introduces parameter-dependent coefficients of the underlying partial differential equation (PDE) over such domains, see, e.g., [14]. Parameters could describe conductivity, viscosity, or convection directions, see, e.g., [10, 23, 25]. As an extreme case, parametrizing the random diffusion coefficients in a stochastic PDE, e.g., by Karhunen-Loew or polynomial chaos expansions, leads to a deterministic parametric PDE involving, in principle, even *infinitely* many parameters, $p = \infty$, see, e.g., [7] and the literature cited there. We will, however, not treat this particular problem class here any further since, as will be explained later, it poses different conceptual obstructions than those in the focus of this chapter, namely the absence of ellipticity

which makes conventional strategies fail. In particular, we shall explain why for other relevant problem classes, e.g., those dominated by transport processes, $\mathcal{M}$ is not "as visible" as for elliptic problems and how to restore "full visibility."

11.1.1 General Context - Reduced Basis Method

A conventional way of searching for a specific state in $\mathcal{M}$ or optimize over $\mathcal{M}$ is to compute approximate solutions of (11.1) possibly for a large number of parameters y. Such approximations would then reside in a sufficiently large trial space $U_{\mathcal{N}} \subset U$ of dimension $\mathcal{N}$, typically a finite element space. Ideally one would try to assure that $U_{\mathcal{N}}$ is large enough to warrant sufficient accuracy of whatever conclusions are to be drawn from such a discretization. A common terminology in reduced order modeling refers to $U_{\mathcal{N}}$ as the *truth space* providing accurate computable information. Of course, each such parameter query in $U_{\mathcal{N}}$ is a computationally expensive task so that many such queries, especially in an online context, would be practically infeasible. On the other hand, solving for each $y \in \mathcal{Y}$ a problem in $U_{\mathcal{N}}$ would just treat each solution $u(y)$ as some "point" in the infinite-dimensional space U, viz. in the very large finite-dimensional space $U_{\mathcal{N}}$. This disregards the fact that all these points actually belong to a possibly much thinner and more coherent set, namely the low-dimensional manifold $\mathcal{M}$ which, for compact $\mathcal{Y}$ and well-posed problems (11.1), is compact. Moreover, if the solutions $u(y)$, as functions of $y \in \mathcal{Y}$, depend smoothly on y there is hope that one can approximate all elements of $\mathcal{M}$ uniformly over $\mathcal{Y}$ with respect to the Hilbert space norm $\|\cdot\|_U$ by a relatively small but judiciously chosen linear space U_n. Here "small" means that $n = \dim U_n$ is significantly smaller than $\mathcal{N} = \dim U_{\mathcal{N}}$, often by orders of magnitude. As detailed later the classical notion of *Kolmogorov n-widths* quantifies how well a compact set in a Banach space can be approximated in the corresponding Banach norm by a linear space and therefore can be used as a *benchmark* for the effectiveness of a model reduction strategy.

Specifically, the core objective of the *Reduced Basis Method* (RBM) is to find for a given *target accuracy* ε a possibly small number $n = n(\varepsilon)$ of basis functions $\phi_j, j = 0, \ldots, n$, whose linear combinations approximate each $u \in \mathcal{M}$ within accuracy at least ε. This means that ideally for each $y \in \mathcal{Y}$ one can find coefficients $c_j(y)$ such that the expansion

$$u_n(x, y) := \sum_{j=0}^{n(\varepsilon)} c_j(y)\phi_j(x) \tag{11.4}$$

satisfies

$$\|u(y) - u_n(y)\|_U \leq \varepsilon, \quad y \in \mathcal{Y}. \tag{11.5}$$

Thus projecting (11.1) into the small space $U_n := \text{span}\{\phi_0, \ldots, \phi_n\}$ reduces each parameter query to solving a small $n \times n$ system of equations where typically $n \ll \mathcal{N}$.

11.1.2 Goal Orientation

Recall that the actual goal of reduced modeling is often not to recover the full fields $u(y) \in \mathcal{M}$ but only some *quantity of interest* $I(y)$ typically given as a *functional* $I(y) := \ell(u(y))$ of $u(y)$ where $\ell \in U'$. Asking just the value of such a functional is possibly a weaker request than approximating all of $u(y)$ in the norm $\|\cdot\|_U$. In other words, one may have $|\ell(u(y)) - \ell(u_n(y))| \leq \varepsilon$ without insisting on the validity of (11.5) for a tolerance of roughly the same size. Of course, one would like to exploit this in favor of online efficiency. Duality methods as used in the context of *goal-oriented* finite element methods [3] are indeed known to offer ways of economizing the approximate evaluation of functionals. Such concepts apply in the RBM context as well, see, e.g., [16, 21]. However, as we shall point out later, guaranteeing that $|\ell(u(y)) - \ell(u_n(y))| \leq \varepsilon$ holds for $y \in \mathcal{Y}$ ultimately reduces to tasks of the type (11.5) as well. So, in summary, understanding how to ensure (11.5) for possibly small $n(\varepsilon)$ remains the core issue and therefore guides the subsequent discussions.

Postponing for a moment the issue of how to actually compute the ϕ_j, it is clear that they should intrinsically depend on $\mathcal{M}$ rendering the whole process highly nonlinear. To put the above approach first into perspective, viewing $u(x, y)$ as a function of the spatial variables x and of the parameters y, (11.4) is just *separation of variables* where the factors $c_j(y)$, $\phi_j(x)$ are a priori unknown. It is perhaps worth stressing though that, in contrast to other attempts to find good *tensor approximations*, in the RBM context explicit representations are only computed for the spatial factors ϕ_j while for each y the weight $c_j(y)$ has to be *computed* by solving a small system in the reduced space U_n. Thus the computation of $\{\phi_0, \ldots, \phi_{n(\varepsilon)}\}$ could be interpreted as *dictionary learning* and, loosely speaking, $n = n(\varepsilon)$ being relatively small for a given target accuracy, means that all elements in $\mathcal{M}$ are *approximately sparse* with respect to the dictionary $\{\phi_0, \ldots, \phi_n, \ldots\}$.

The methodology just outlined has been pioneered by Y. Maday, T.A. Patera, and collaborators, see, e.g., [6, 21, 23, 25]. As indicated before, RBM is one variant of a *model order reduction* paradigm that is specially tailored to parameter dependent problems. Among its distinguishing constituents one can name the following. There is usually a careful division of the overall computational work into an *offline phase*, which could be computationally intense but should remain manageable, and an *online phase*, which should be executable with highest efficiency taking advantage of a precomputed basis and matrix assemblations during the offline phase. It is important to note that while the offline phase is accepted to be computationally expensive it should remain *offline feasible* in the sense that a possibly extensive search over the parameter domain $\mathcal{Y}$ in the offline phase requires for each query solving only problems in the small reduced space. Under which circumstances this is possible and how to realize such division concepts has been worked out in the literature, see, e.g., [23, 25]. Here we are content with stressing that an important role is played by the way how the operator B_y depends on the parameter y, namely in an *affine* way as stated in (11.18) later below. Second, and this is perhaps the

most distinguishing constituent, along with each solution in the reduced model one strives to provide a *certificate* of accuracy, i.e., computed bounds for incurred error tolerances [23, 25].

11.1.3 Central Objectives

When trying to quantify the performance of such methods aside from the above-mentioned structural and data organization aspects, among others, the following questions come to mind:

(i) for which type of problems do such methods work very well in the sense that $n(\varepsilon)$ in (11.5) grows only slowly when ε decreases? This concerns quantifying the sparsity of solutions.
(ii) How can one compute reduced bases $\{\phi_0, \ldots, \phi_{n(\varepsilon)}\}$ for which $n(\varepsilon)$ is *nearly minimal* in a sense to be made precise below?

Of course, the better the sparsity quantified by (i) the better could be the payoff of an RBM. However, as one may expect, an answer to (i) depends strongly on the problem under consideration. This is illustrated also by the example presented in §11.5.4. Question (ii), instead, can be addressed independently of (i) in the sense that, no matter how many basis functions have to be computed in order to meet a given target accuracy, can one come up with methods that guarantee generating a *nearly minimal number* of such basis functions? This has to do with *how to sample* the solution manifold and is the central theme in this chapter.

The most prominent way of generating the reduced bases is a certain *greedy sampling* of the manifold $\mathcal{M}$. Contriving *greedy sampling strategies* that give rise to reduced bases of nearly minimal length, in a sense to be made precise below, also for *noncoercive or unsymmetric singularly perturbed problems* is the central objective in this chapter. We remark though that a greedy parameter search in its standard form is perhaps not suitable for very high-dimensional parameter spaces without taking additional structural features of the problem into account. The subsequent discussions do therefore not target specifically the large amount of recent work on stochastic elliptic PDEs, since while greedy concepts are in principle well understood for elliptic problems they are per se not necessarily adequate for infinitely many parameters without exploiting specific problem-dependent structural information.

First, we recall in §11.2 a *greedy space growth* paradigm commonly used in all established RBMs. To measure its performance in the sense of (ii) we follow [6] and compare the corresponding distances $\mathrm{dist}_U(\mathcal{M}, U_n)$ to the smallest possible distances achievable by linear spaces of dimension n, called *Kolmogorov n-widths*. The fact that for *elliptic problems* the convergence rates for the greedy errors are essentially those of the n-widths, and hence *rate-optimal*, is shown in §11.3 to be ultimately reduced to analyzing the so-called *weak greedy algorithms* in Hilbert spaces, see also [4, 13]. However, for indefinite or strongly unsymmetric

and singularly perturbed problems this method usually operates far from optimality. We explain why this is the case and describe in §11.4 a remedy proposed in [10]. A pivotal role is played by certain *well-conditioned variational formulations* of (11.1) which are then shown to lead again to an optimal *outer greedy* sampling strategy also for non-elliptic problems. An essential additional ingredient consists of certain stabilizing *inner greedy loops*, see §11.5. The obtained rate-optimal scheme is illustrated by a numerical example addressing convection-dominated convection-diffusion problems in §11.5.4. We conclude in §11.6 with applying these concepts to the efficient evaluation of quantities of interest.

11.2 The Greedy Paradigm

The by far most prominent strategy for constructing reduced bases for a given parameter-dependent problem (11.1) is the following greedy procedure, see, e.g., [23]. The basic idea is that, having already constructed a reduced space U_n of dimension n, find an element $u_{n+1} = u(y_{n+1})$ in $\mathcal{M}$ that is farthest away from the current space U_n, i.e., that maximizes the best approximation error from U_n and then grow U_n by setting $U_{n+1} := U_n + \mathrm{span}\{u_{n+1}\}$. Hence, denoting by P_{U,U_n} the U-orthogonal projection onto U_n,

$$y_{n+1} := \mathrm{argmax}_{y\in\mathcal{Y}} \|u(y) - P_{U,U_n}u(y)\|_U, \quad u_{n+1} := u(y_{n+1}). \tag{11.6}$$

Unfortunately, determining such an exact maximizer is computationally way too expensive even in an offline phase because one would have to compute for a sufficiently dense sampling of $\mathcal{Y}$ the exact solution $u(y)$ of (11.1) in U (in practice in $U_\mathcal{N}$). Instead one tries to construct more efficiently computable *surrogates* $R(y, U_n)$ satisfying

$$\|u(y) - P_{U,U_n}u(y)\|_U \le R(y, U_n), \quad y \in \mathcal{Y}. \tag{11.7}$$

Recall that "efficiently computable" in the sense of offline feasibility means that for each $y \in \mathcal{Y}$, the surrogate $R(y, U_n)$ can be evaluated by solving only a problem of size n in the reduced space U_n. Deferring an explanation of the nature of such surrogates, Algorithm 1 described below is a typical offline feasible *surrogate-based greedy algorithm* (SGA). Clearly, the maximizer in (11.8) below is not necessarily unique. In case several maximizers exist it does not matter which one is selected.

Strictly speaking, the scheme SGA is still idealized since:

(a) computations cannot be carried out in U;
(b) one cannot parse through all of a continuum $\mathcal{Y}$ to maximize $R(y, U_n)$.

Concerning (a), as mentioned earlier computations in U are to be understood as synonymous to computations in a sufficiently large truth space $U_\mathcal{N}$ satisfying all targeted accuracy tolerances for the underlying application. Solving problems in

Algorithm 1 Surrogate-based greedy algorithm

function SGA
  Set $U_0 := \{0\}$, $n = 0$,
  while $\operatorname{argmax}_{y\in\mathcal{Y}} R(y, U_n) \geq tol$ **do**

$$\begin{aligned} y_{n+1} &:= \operatorname{argmax}_{y\in\mathcal{Y}} R(y, U_n), \\ u_{n+1} &:= u(y_{n+1}), \\ U_{n+1} &:= \operatorname{span}\left\{U_n, \{u(y_{n+1})\}\right\} = \operatorname{span}\{u_1, \ldots, u_{n+1}\} \end{aligned} \tag{11.8}$$

  end while
end function

$U_{\mathcal{N}}$ is strictly confined to the offline phase and the number of such solves should remain of the order of $n = \dim U_n$. We will not distinguish in what follows between U and $U_{\mathcal{N}}$ unless such a distinction matters.

As for (b), the maximization is usually performed with the aid of a exhaustive search over a *discrete* subset of $\mathcal{Y}$. Again, we will not distinguish between a possibly continuous parameter set and a suitable training subset. In fact, continuous optimization methods that would avoid a complete search have so far not proven to work well since each greedy step increases the number of local maxima of the objective functional. Now, how fine such a discretization for a exhaustive search should be, depends on how smoothly the $u(y)$ depend on y. But even when such a dependence is very smooth a coarse discretization of a *high-dimensional* parameter set $\mathcal{Y}$ would render a exhaustive search infeasible so that, depending on the problem at hand, one has to resort to alternate strategies such as, for instance, random sampling. However, since it seems that (b) can only be answered for a specific problem class we will not address this issue in this chapter any further.

Instead, we focus on general principles which guarantee the following. Loosely speaking the reduced spaces based on sampling $\mathcal{M}$ should perform *optimally* in the sense that the resulting spaces U_n have the (near) "smallest dimension" needed to satisfy a given target tolerance while the involved offline and online cost remains feasible in the sense indicated above. To explain first what is meant by "optimal" let us denote the *greedy error* produced by SGA as

$$\sigma_n(\mathcal{M})_U := \max_{v\in\mathcal{M}} \inf_{\bar{u}\in U_n} \|v - \bar{u}\|_U = \max_{y\in\mathcal{Y}} \|u(y) - P_{U,U_n} u(y)\|_U. \tag{11.9}$$

Note that if we replace in (11.9) the space U_n by *any* linear subspace $W_n \subset U$ and infimize the resulting distortion over *all* subspaces of U of dimension at most n, we obtain the classical *Kolmogorov n-widths* $d_n(\mathcal{M})_U$ quantifying the "thickness" of a compact set, see (11.21). One trivially has

$$d_n(\mathcal{M})_U \leq \sigma_n(\mathcal{M})_U, \quad n \in \mathbb{N}. \tag{11.10}$$

Of course, it would be best if one could reverse the above inequality. We will discuss in the next section to what extent this is possible.

To prepare for such a discussion we need more information about how the surrogate $R(y, U_n)$ relates to the actual error $\|u(y) - P_{U,U_n}u(y)\|_U$ because the surrogate drives the greedy search and one expects that the quality of the snapshots found in SGA depends on how "tight" the upper bound in (11.7) is.

To identify next the essential conditions on a "good" surrogate it is instructive to consider the case of *elliptic* problems. To this end, suppose that

$$\langle B_y u, v\rangle = b_y(u, v) = \langle f, v\rangle, \quad u, v \in U,$$

is a uniformly U-coercive bounded bilinear form and $f \in U'$, i.e., there exist constants $0 < c_1 \le C_1 < \infty$ such that

$$c_1\|v\|_U^2 \le b_y(v, v), \quad |b_y(u, v)| \le C_1\|u\|_U\|v\|_U, \ u, v \in U, \ y \in \mathcal{Y}, \tag{11.11}$$

holds uniformly in $y \in \mathcal{Y}$. The operator equation (11.1) is then equivalent to: given $f \in U'$ and a $y \in \mathcal{Y}$, find $u(y) \in U$ such that

$$b_y(u(y), v) = \langle f, v\rangle, \quad v \in U. \tag{11.12}$$

Ellipticity has two important well-known consequences. First, since (11.11) implies $\|B_y\|_{U\to U'} \le C_1$, $\|B_y^{-1}\|_{U'\to U} \le c_1^{-1}$ the operator $B_y : U \to U'$ has a finite condition number

$$\kappa_{U,U'}(B_y) := \|B_y\|_{U\to U'}\|B_y^{-1}\|_{U'\to U} \le C_1/c_1 \tag{11.13}$$

which, in particular, means that residuals in U' are uniformly comparable to errors in U

$$C_1^{-1}\|f - B_y\bar{u}\|_{U'} \le \|u(y) - \bar{u}\|_U \le c_1^{-1}\|f - B_y\bar{u}\|_{U'}, \quad \bar{u} \in U, \ y \in \mathcal{Y}. \tag{11.14}$$

Second, by Céa's Lemma, the Galerkin projection Π_{y,U_n} onto U_n is up to a constant as good as the *best approximation*, i.e., under assumption (11.11)

$$\|u(y) - \Pi_{y,U_n}u(y)\|_U \le \frac{C_1}{c_1}\inf_{v\in U_n}\|u(y) - v\|_U. \tag{11.15}$$

(When $b_y(\cdot,\cdot)$ is in addition symmetric C_1/c_1 can be replaced by $(C_1/c_1)^{1/2}$.) Hence, by (11.14) and (11.15),

$$R(y, U_n) := c_1^{-1}\sup_{v\in U}\frac{\langle f, v\rangle - b_y(\Pi_{y,U_n}u(y), v)}{\|v\|_U} \tag{11.16}$$

satisfies more than just (11.7), namely it provides also a uniform *lower bound*

$$\frac{c_1}{C_1}R(y, U_n) \le \|u(y) - P_{U,U_n}u(y)\|_U, \quad y \in \mathcal{Y}. \tag{11.17}$$

Finally, suppose that $b_y(\cdot,\cdot)$ depends *affinely* on the parameters in the sense that

$$b_y(u,v) = \sum_{k=1}^{M} \theta_k(y) b_k(u,v), \tag{11.18}$$

where the θ_k are smooth functions of $y \in \mathcal{Y}$ and the bilinear forms $b_k(\cdot,\cdot)$ are independent of y. Then, based on suitable precomputations (in $U_{\mathcal{N}}$) in the offline phase, the computation of $\Pi_{y,U_n} u(y)$ reduces for each $y \in \mathcal{Y}$ to the solution of a rapidly assembled $(n \times n)$ system, and $R(y, U_n)$ can indeed be computed very efficiently, see [16, 23, 25].

An essential consequence of (11.7) and (11.17) can be formulated as follows.

Proposition 2.1. *Given $U_n \subset U$, the function u_{n+1} generated by (11.8) for $R(y, U_n)$ defined by (11.16) has the property that*

$$\|u_{n+1} - P_{U,U_n} u_{n+1}\|_U \geq \frac{c_1}{C_1} \max_{v \in \mathcal{M}} \min_{\bar{u} \in U_n} \|v - \bar{u}\|_U. \tag{11.19}$$

Hence, maximizing the residual based surrogate $R(y, U_n)$ (over a suitable discretization of $\mathcal{Y}$) is a computationally feasible way of determining, up to a fixed factor $\gamma := c_1/C_1 \leq 1$, the maximal distance between $\mathcal{M}$ and U_n and performs in this sense almost as well as the "ideal" but computationally infeasible surrogate $R^*(\mu, U_n) := \|u(y) - P_{U,U_n} u(y)\|_U$.

Proof of Proposition 2.1. Suppose that $\bar{y} = \operatorname{argmax}_{y \in \mathcal{Y}} R(y, U_n)$, $y^* := \operatorname{argmax}_{y \in \mathcal{Y}} \|u(y) - P_{U,U_n} u(y)\|_U$ so that $u_{n+1} = u(\bar{y})$. Then, keeping (11.17) and (11.15) in mind, we have

$$\begin{aligned} \|u_{n+1} - P_{U,U_n} u_{n+1}\|_U &= \|u(\bar{y}) - P_{U,U_n} u(\bar{y})\|_U \geq \frac{c_1}{C_1} R(\bar{y}, U_n) \geq \frac{c_1}{C_1} R(y^*, U_n) \\ &\geq \frac{c_1}{C_1} \|u(y^*) - P_{U,U_n} u(y^*)\|_U = \frac{c_1}{C_1} \max_{y \in \mathcal{Y}} \|u(y) - P_{U,U_n} u(y)\|_U, \end{aligned}$$

where we have used (11.7) in the second but last step. This confirms the claim. □

Property (11.19) turns out to play a key role in the analysis of the performance of the scheme SGA.

11.3 Greedy Space Growth

Proposition 2.1 allows us to view the algorithm SGA as a special instance of the following scenario. Given a compact subset $\mathcal{K}$ of a Hilbert space H with inner product $(\cdot,\cdot)$ inducing the norm $\|\cdot\|^2 = (\cdot,\cdot)$, consider the *weak greedy* Algorithm 2 (WGA) below.

Algorithm 2 Weak greedy algorithm

1: **function** WGA
2: Set $H_0 := \{0\}$, $n = 0$, $u_0 := 0$, fix any $0 < \gamma \leq 1$,
3: given H_n, choose some $u_{n+1} \in \mathcal{K}$ for which

$$\min_{v_n \in H_n} \|v_n - u_{n+1}\| \geq \gamma \max_{v \in \mathcal{K}} \min_{v_n \in U_n} \|v - v_n\| =: \gamma \sigma_n(\mathcal{K})_H, \tag{11.20}$$

and set $H_{n+1} := H_n + \text{span}\,\{u_{n+1}\}$.
4: **end function**

Note that again the choice of u_{n+1} is not necessarily unique and what follows holds for *any* choice satisfying (11.20).

Greedy strategies have been used in numerous contexts and variants. The current version is not to be confused though with the *weak orthogonal greedy algorithm* introduced in [26] for *approximating* a *function* by a linear combination of n terms from a *given* dictionary. In contrast, the scheme WGA described in Algorithm 2 aims at *constructing* a (problem dependent) dictionary with the aid of a PDE model. While greedy function approximation is naturally compared with the *best n-term approximation* from the underlying dictionary (see [2, 26] for related results), a natural question here is to compare the corresponding greedy errors

$$\sigma_n(\mathcal{K})_H := \max_{v \in \mathcal{K}} \min_{v_n \in U_n} \|v - v_n\| =: \max \text{dist}_H(\mathcal{K}, U_n)$$

incurred when approximating a compact set $\mathcal{K}$ with the smallest possible deviation of $\mathcal{K}$ from any n-dimensional linear space, given by the Kolmogorov n-widths

$$d_n(\mathcal{K})_H := \inf_{\dim V = n} \sup_{v \in \mathcal{K}} \inf_{v_n \in V} \|v - v_n\| = \inf_{\dim V = n} \max \text{dist}_H(\mathcal{K}, V), \tag{11.21}$$

mentioned earlier in the preceding section. One trivially has $d_n(\mathcal{K})_H \leq \sigma_n(\mathcal{K})_H$ for all $n \in \mathbb{N}$ and the question arises whether there actually exists a constant C such that

$$\sigma_n(\mathcal{K})_H \leq C d_n(\mathcal{K})_H, \quad n \in \mathbb{N}. \tag{11.22}$$

One may doubt such a relation to hold for several reasons. First, orthogonal greedy *function approximation* performs in a way comparable to best n-term approximation only under rather strong assumptions on the underlying given dictionary. Intuitively, one expects that errors made early on in the iteration are generally hard to correct later although this intuition turns out to be misleading in the case of the present *set approximation*. Second, the spaces U_n generated by the greedy growth are restricted by being generated only from snapshots in $\mathcal{K}$ while the best spaces can be chosen freely, see the related discussion in [4].

The comparison (11.22) was addressed first in [6] for the ideal case $\gamma = 1$. In this case a bound of the form $\sigma_n(\mathcal{K})_H \leq Cn2^n d_n(\mathcal{K})_H$ could be established for some

absolute constant C. This is useful only for cases where the n-widths decay faster than $n^{-1}2^{-n}$ which indeed turns out to be possible for elliptic problems (11.12) with a sufficiently smooth affine parameter dependence (11.18). In fact, in such a case the $u(y)$ can be even shown to be *analytic* as a function of y, see [7] and the literature cited there. It was then shown in [4] that the slightly better bound

$$\sigma_n(\mathcal{K})_H \leq \frac{2^{n+1}}{\sqrt{3}} d_n(\mathcal{K})_H, \quad n \in \mathbb{N}, \tag{11.23}$$

holds. More importantly, these bounds cannot be improved in general. Moreover, the possible exponential loss in accuracy is not due to the fact the greedy spaces are generated by snapshots from $\mathcal{K}$. In fact, denoting by $\bar{d}_n(\mathcal{K})_H$ the restricted "inner" widths, obtained by allowing only subspaces spanned by snapshots of $\mathcal{K}$ in the competition, one can prove that $\bar{d}_n(\mathcal{K})_H \leq n d_n(\mathcal{K})_H$, $n \in \mathbb{N}$, which is also sharp in general [4].

While these findings may be interpreted as limiting the use of reduced bases generated in a greedy fashion to problems where the n-widths decay exponentially fast the situation turns out to be far less dim if one does *not* insist on a *direct comparison* of the type (11.22) with n being *the same* on both sides of the inequality. In [4, 13] the question is addressed whether a certain *convergence rate* of the n-widths $d_n(\mathcal{K})_H$ implies some convergence rate of the greedy errors $\sigma_n(\mathcal{K})_H$. The following result from [4] gave a first affirmative answer.

Theorem 3.1. *Let $0 < \gamma \leq 1$ be the parameter in (11.20) and assume that $d_0(\mathcal{K})_H \leq M$ for some $M > 0$. Then*

$$d_n(\mathcal{K})_H \leq M n^{-\alpha}, \quad n \in \mathbb{N},$$

for some $\alpha > 0$, implies

$$\sigma_n(\mathcal{K})_H \leq C M n^{-\alpha}, \quad n > 0, \tag{11.24}$$

where $C := q^{\frac{1}{2}}(4q)^{\alpha}$ and $q := \lceil 2^{\alpha+1}\gamma^{-1}\rceil^2$.

This means that the weak greedy scheme may still be highly profitable even when the n-widths do not decay exponentially. Moreover, as expected, the closer the weakness parameter γ is to one, the better, which will later guide the sampling strategies for constructing reduced bases.

Results of the above type are not confined to polynomial rates. A sub-exponential decay of the $d_n(\mathcal{K})_H$ with a rate $e^{-cn^{\alpha}}$, $\alpha \leq 1$ is shown in [4] to imply a rate

$$\sigma_n(\mathcal{K})_H \leq C(\alpha, \gamma) e^{-\tilde{c}n^{\tilde{\alpha}}}, \quad \tilde{\alpha} = \alpha/(1+\alpha), \quad n \in \mathbb{N}. \tag{11.25}$$

The principle behind the estimates (11.24), (11.25) is to exploit a "flatness" effect or what one may call "conditional delayed comparison." More precisely, given any $\theta \in (0, 1)$ and defining $q := \lceil 2(\gamma\theta)^{-1}\rceil^2$, one can show that ([4, Lemma 2.2])

$$\sigma_{n+qm}(\mathcal{K})_H \geq \theta \sigma_n(\mathcal{K})_H \quad \Rightarrow \quad \sigma_n(\mathcal{K})_H \leq q^{1/2} d_m(\mathcal{K})_H, \quad n \in \mathbb{N}.$$

Thus a comparison between greedy errors and n-widths is possible when the greedy errors do not decay too quickly. This is behind the diminished exponent $\tilde{\alpha}$ in (11.25).

These results have been improved upon in [13] in several ways employing different techniques yielding improved comparisons. Abbreviating $\sigma_n := \sigma_n(\mathcal{K})_H, d_n := d_n(\mathcal{K})_H$, a central result in the present general Hilbert space context states that for any $N \geq 0, K \geq 1, 1 \leq m < K$ one has

$$\prod_{i=1}^{K} \sigma_{N+i}^2 \leq \gamma^{-2K} \Big(\frac{K}{M}\Big)^m \Big(\frac{K}{K-m}\Big)^{K-m} \sigma_{N+1}^{2m} d_m^{2(K-m)}. \tag{11.26}$$

As a first important consequence, one derives from these inequalities a nearly direct comparison between σ_n and d_n without any constraint on the decay of σ_n or d_n. In fact, taking $N = 0, K = n$, and any $1 \leq m < n$ in (11.26), using the monotonicity of the σ_n, one shows that $\sigma_n^{2n} \leq \gamma^{-2n} \Big(\frac{n}{m}\Big)^m \Big(\frac{n}{n-m}\Big)^{n-m} d_m^{2(n-m)}$ from which one deduces

$$\sigma_n \leq \sqrt{2}\gamma^{-1} \min_{1 \leq m < n} d_m^{\frac{n-m}{n}}, \quad n \in \mathbb{N}. \tag{11.27}$$

This, in particular, gives the direct unconditional comparison

$$\sigma_{2n}(\mathcal{K})_H \leq \gamma^{-1} \sqrt{2 d_n(\mathcal{K})_H}, \quad n \in \mathbb{N}.$$

The estimate (11.27) is then used in [13] to improve on (11.25) establishing the bounds

$$d_n(\mathcal{K})_H \leq C_0 e^{-c_0 n^\alpha} \quad \Rightarrow \quad \sigma_n(\mathcal{K})_H \leq \sqrt{2C_0}\gamma^{-1} e^{-c_1 n^\alpha}, \quad n \in \mathbb{N}, \tag{11.28}$$

i.e., the exponent α is preserved by the rate for the greedy errors. Moreover, one can recover (11.24) from (11.26) (with different constants).

Although not needed in the present context the second group of results in [13] should be mentioned that concerns the extension of the weak greedy algorithm WGA to *Banach* spaces X in place of the Hilbert space H. Remarkably, a direct comparison between $\sigma_n(\mathcal{K})_X$ and $d_n(\mathcal{K})_X$ similar to (11.26) is also established in [13]. The counterpart to (11.27) reads $\sigma_{2n} \leq 2\gamma^{-1}\sqrt{n d_n}$, i.e., one loses a factor $\sqrt{n}$ which is shown, however, to be necessary in general.

All the above results show that the smaller the weakness parameter γ the stronger the derogation of the rate of the greedy errors in comparison with the n-widths.

11.4 What are the Right Projections?

As shown by (11.24) and (11.28), the weak greedy algorithm WGA realizes optimal rates for essentially all ranges of interest. A natural question is under which circumstances a surrogate-based greedy algorithm SGA is in this sense also

rate-optimal, namely ensures the validity of (11.24) and (11.28). Obviously, this is precisely the case when new snapshots generated through maximizing the surrogate have the *weak greedy property* (11.20). Note that Proposition 2.1 says that the *residual-based surrogate* (11.16) in the case of *coercive problems* does ensure the weak-greedy property so that SGA is indeed rate-optimal for coercive problems. Note also that the weakness parameter $\gamma = c_1/C_1$ is in this case the larger the smaller the condition number of the operator B_y is, see (11.13). Obviously, the key is that the surrogate not only yields an upper bound for the best approximation error but also, up to a constant, a lower bound (11.17), and the more tightly the best approximation error is sandwiched by the surrogate the better the performance of SGA. Therefore, even if the problem is coercive for a very small $\gamma = c_1/C_1$, as is the case for convection-dominated *convection-diffusion problems*, in view of the dependence of the bounds in (11.24) and (11.28) on γ^{-1}, one expects that the performance of a greedy search based on (11.16) degrades significantly.

In summary, as long as algorithm SGA employs a *tight surrogate* in the sense that

$$c_S R(y, U_n) \le \inf_{v \in U_n} \|u(y) - v\|_U \le R(y, U_n), \quad y \in \mathcal{Y}, \tag{11.29}$$

holds for some constant $c_S > 0$, independent of $y \in \mathcal{Y}$, algorithm SGA is *rate-optimal* in the sense of (11.24), (11.28), i.e., it essentially realizes the n-width rates over all ranges of interest, see [10]. We refer to $c_S^{-1} := \kappa_n(R)$ as the *condition* of the surrogate $R(\cdot, U_n)$. In the RBM community the constant c_S^{-1} is essentially the *stability factor* which is usually computed along with an approximate reduced solution. Clearly, the bounds in §11.3 also show that the quantitative performance of SGA is expected to be the better the smaller the condition of the surrogate, i.e., the larger c_S.

As shown so far, coercive problems with a small condition number $\kappa_{U,U'}(B_y)$ represent an ideal setting for RBM and standard Galerkin projection combined with the *symmetric surrogate* (11.16), based on measuring the residual in the dual norm $\|\cdot\|_{U'}$ of the "error norm" $\|\cdot\|_U$, identifies rate-optimal snapshots for a greedy space growth. Of course, this marks a small segment of relevant problems. Formally, one can still apply these projections and surrogates for any variational problem (11.12) for which a residual can be computed. However, in general, for indefinite or unsymmetric singularly perturbed problems, the tightness relation (11.29) may no longer hold for surrogates of the form (11.16) or, if it holds the condition $\kappa_n(R)$ becomes prohibitively large. In the latter case, the upper bound of the best approximation error is too loose to direct the search for proper snapshots. A simple example is the *convection-diffusion* problem: for $f \in (H_0^1(\Omega))'$ find $u \in H_0^1(\Omega)$, $\Omega \subset \mathbb{R}^d$, such that

$$\varepsilon(\nabla u, \nabla v) + (\vec{b} \cdot \nabla u, v) + (cu, v) =: b_y(u, v) = \langle f, v \rangle, \quad v \in H_0^1(\Omega), \tag{11.30}$$

where, for instance, $y = (\varepsilon, \vec{b}) \in \mathcal{Y} := [\varepsilon_0, 1] \times S^{d-1}$, S^{d-1} the $(d-1)$-sphere.

Remark 4.1. It is well known that when $c - \frac{1}{2}\mathrm{div}\vec{b} \geq 0$ problem (11.30) has for any $f \in H^{-1}(\Omega) := (H_0^1(\Omega))'$ a unique solution. Thus for $U := H_0^1(\Omega)$ (11.11) is still valid but with $\kappa_{U,U'}(B_y) \sim \varepsilon^{-1}$ which becomes arbitrarily large for a correspondingly small diffusion lower bound ε_0.

The standard scheme SGA indeed no longer performs nearly as well as in the well-conditioned case. The situation is even less clear when $\varepsilon = 0$ (with modified boundary conditions) where no "natural" variational formulation suggests itself (we refer to [10] for a detailed discussion of these examples). Moreover, for *indefinite problems* the Galerkin projection does generally perform like the best approximation which also adversely affects tightness of the standard symmetric residual based surrogate (11.16).

Hence, to retain rate-optimality of SGA also for the above-mentioned extended range of problems one has to find a better surrogate than the one based on the symmetric residual bound in (11.16). We indicate in the next section that such better surrogates can indeed be obtained at affordable computational cost for a wide range of problems through combining *Petrov-Galerkin projections* with appropriate *unsymmetric* residual bounds. The approach can be viewed as *preconditioning* the continuous problem already on the infinite-dimensional level.

11.4.1 Modifying the Variational Formulation

We consider now a wider class of (not necessarily coercive) variational problems

$$b(u, v) = \langle f, v\rangle, \quad v \in V, \tag{11.31}$$

where we assume at this point only for each $f \in V'$ the existence of a unique solution $u \in U$, i.e., the operator $B : U \to V'$, induced by $b(\cdot,\cdot)$, is bijective. This is well known to be equivalent to the validity of

$$\begin{cases} \inf\limits_{w\in W}\sup\limits_{v\in V}\dfrac{b(w,v)}{\|w\|_U\|v\|_V} \geq \beta, \quad \sup\limits_{v\in V}\sup\limits_{w\in U}\dfrac{b(w,v)}{\|w\|_U\|v\|_V} \leq C_b, \\ \text{for } v \in V\, \exists\, w \in W, \ \text{ such that } b(w,v) \neq 0, \end{cases} \tag{11.32}$$

for some constants β, C_b. However, one then faces two principal obstructions regarding an RBM based on the scheme SGA:

(a) first, as in the case of (11.30) for small diffusion, $\kappa_{U,V'}(B) \leq C_b/\beta$ could be very large so that the corresponding error-residual relation

$$\|u - v\|_U \leq \beta^{-1}\|f - Bv\|_{V'}, \quad v \in U, \tag{11.33}$$

renders a corresponding residual-based surrogate ill conditioned.

(b) When $b(\cdot,\cdot)$ is not coercive, the Galerkin projection does, in general, not perform as well as the best approximation.

The following approach has been used in [10] to address both (a) and (b). The underlying basic principle is not new, see [1], and variants of it have been used for different purposes in different contexts such as least squares finite element methods [18] and, more recently, in connection with *discontinuous Petrov Galerkin methods* [9, 11, 12]. In the context of RBMs the concept of *natural norms* goes sort of half way by sticking in the end to Galerkin projections [25]. This marks an essential distinction from the approach in [10] discussed later below.

The idea is to change the topology of one of the spaces so as to (ideally) make the corresponding induced operator an *isometry*, see also [9]. Following [10], fixing for instance, $\|\cdot\|_V$, one can define

$$\|w\|_{\hat{U}} := \sup_{v\in V} \frac{b(w,v)}{\|v\|_V} = \|Bw\|_{V'}, \quad w \in U, \tag{11.34}$$

which means that one has for $Bu = f$

$$\|u - w\|_{\hat{U}} = \|f - Bw\|_{V'}, \quad w \in U, \tag{11.35}$$

a perfect error-residual relation. It also means that replacing $\|\cdot\|_U$ in (11.32) by $\|\cdot\|_{\hat{U}}$ yields the inf-sup constant $\hat{\beta} = 1$. Alternatively, fixing $\|\cdot\|_U$, one may set

$$\|v\|_{\hat{V}} := \sup_{w\in U} \frac{b(w,v)}{\|w\|_U} = \|B^* v\|_{U'}, \quad v \in V, \tag{11.36}$$

to again arrive at an isometry $B : U \to \hat{V}'$, meaning

$$\|u - w\|_U = \|f - Bw\|_{\hat{V}'}, \quad w \in U. \tag{11.37}$$

Whether the norm for U or for V is prescribed depends on the problem at hand and we refer to [8–10] for examples of both types.

Next note that for any subspace $W \subset U$ one has

$$u_W = \operatorname*{argmin}_{w\in W} \|u - w\|_{\hat{U}} = \operatorname*{argmin}_{w\in W} \|f - Bw\|_{V'}, \tag{11.38}$$

and analogously for the pair $(U, \hat{V})$, i.e., the best approximation in the $\hat{U}$ norm is a *minimum residual solution* in the V' norm.

To use residuals in V' as surrogates requires fixing a suitable discrete projection for a given trial space. In general, in particular when $V \neq U$, the Galerkin projection is no longer appropriate since inf-sup stability of the *infinite-dimensional* problem is no longer inherited by an arbitrary pair of *finite-dimensional* trial and test spaces. To see which type of projection would be ideal, denote by $R_U : U' \to U$ the Riesz map defined for any linear functional $\ell \in U'$ by

$$\langle \ell, w\rangle = (R_U \ell, w)_U, \quad w \in U.$$

Then, by (11.34), for any $w \in W \subset U$, taking $v := R_V Bw \in V$ one has

$$b(w, v) = \langle Bw, v\rangle = \langle Bw, R_V Bw\rangle = (Bw, Bw)_{V'} = (w, w)_{\hat{U}}.$$

Thus, in particular,

$$b(u - u_h, R_V Bw) = (u - u_h, w)_{\hat{U}},$$

i.e., given $W \subset U$, using $V_W := R_V B(W)$ as a test space in the *Petrov-Galerkin* scheme

$$b(u_h, v) = \langle f, v\rangle, \quad v \in V_W := R_V B(W), \tag{11.39}$$

is equivalent to computing the $\hat{U}$-orthogonal projection of the exact solution u of (11.31) and hence the best $\hat{U}$ approximation to u. One readily sees that this also means

$$\inf_{w\in W} \sup_{v\in V(W)} \frac{b(w, v)}{\|w\|_{\hat{U}}\|v\|_V} = 1, \tag{11.40}$$

i.e., we have a Petrov-Galerkin scheme for the pair of spaces W, V_W with perfect stability and the Petrov-Galerkin projection is the best $\hat{U}$-projection. Unfortunately, this is not of much help yet, because computing the *ideal test space* $V_W = R_V B(W) = B^{-*} R_{\hat{U}}^{-1}(W)$ is not numerically feasible. Nevertheless, it provides a useful orientation for finding good and practically realizable pairs of trial and test spaces, as explained next.

11.4.2 A Saddle Point Formulation

We briefly recall now from [9, 10] an approach to deriving from the preceding observations a practically feasible numerical scheme which, in particular, fits into the context of RBMs. Taking (11.38) as point of departure we notice that the minimization of $\|f - Bw\|_{V'}$ over W is a least squares problem whose normal equations read: find $u_W \in W$ such that (with $R_{V'} = R_V^{-1}$)

$$0 = (f - Bu_W, Bw)_{V'} = \langle R_V(f - Bu_W), Bw\rangle, \quad w \in W. \tag{11.41}$$

Introducing the auxiliary variable $r := R_V(f - Bu_W)$ which is equivalent to

$$\langle R_{V'} r, v\rangle = (r, v)_V = \langle f - Bu_w, v\rangle, \quad v \in V_W = R_V B(W), \tag{11.42}$$

the two relations (11.41) and (11.42) can be rewritten in form of the *saddle point problem*

$$\begin{aligned} (r, v)_V + b(u_W, v) &= \langle f, v\rangle, \ v \in V_W. \\ b(w, r) &= 0, \qquad w \in W. \end{aligned} \tag{11.43}$$

The corresponding inf-sup constant is still one (since the supremum of $b(w, v)$ over V_W equals for each $w \in W$ the supremum over all of V) and $(\cdot, \cdot)_V$ is a scalar product so that (11.43) has a unique solution u_W, see e.g. [5]. Taking for any $w \in W$ the test function $v = R_V Bw \in V_W$ in the first line of (11.43), one obtains

$$(r, v)_V = (r, R_V Bw)_V = \langle r, Bw\rangle = b(w, r) = 0,$$

by the second line in (11.43) so we see that $\langle f, R_V Bw\rangle = b(u_W, R_V Bw)$ holds for all $w \in W$ which means that u_W solves the ideal Petrov-Galerkin problem (11.39). Thus (11.43) is equivalent to the ideal Petrov Galerkin scheme (11.39).

Of course, (11.43) is still not realizable since the space V_W is still not computable at affordable cost. One more step to arrive at a realizable scheme is based on the following: given the finite-dimensional space W, replacing V_W in (11.43) by some (accessible) space $Z \subset V$, amounts to a Petrov-Galerkin formulation with test space $P_{V,Z}V_W$, where again $P_{V,Z}$ denotes the V-orthogonal projection to Z. Thus, when Z is large enough the (computable) projection $P_{V,Z}V_W$ is close enough to V_W so that one obtains a *stable* finite-dimensional saddle point problem which is the same as saying that its inf-sup constant is safely bounded away from zero. Since $Z = V$ would yield perfect stability the choice of $Z \subset V$ can be viewed as a *stabilization*. To quantify this we follow [10] and say that for some $\delta \in (0, 1)$, $Z \subset V$ is *δ-proximal* for $W \subset U$ if Z is sufficiently close to the ideal test space $V_W = R_V B(W)$ in the sense that

$$\|(I - P_{V,Z})R_V Bw\|_V \le \delta \|R_V Bw\|_V, \quad w \in W. \tag{11.44}$$

The related main findings from [10] can be summarized as follows.

Theorem 4.2. *(i) The pair $(u_{W,Z}, r_{W,Z}) \in W \times Z \subset U \times V$ solves the saddle point problem*

$$\begin{aligned} (r_{W,Z}, v)_V + b(u_{W,Z}, v) &= \langle f, v\rangle, \ v \in Z, \\ b(w, u_{W,Z}) &= 0, \qquad w \in W, \end{aligned} \tag{11.45}$$

if and only if $u_{W,Z}$ solves the Petrov-Galerkin problem

$$b(u_{W,Z}, v) = \langle f, v\rangle, \quad v \in P_{V,Z}(R_V B(W)). \tag{11.46}$$

(ii) If Z is δ-proximal for W, (11.46) is solvable and one has

$$\begin{aligned} \|u - u_{W,Z}\|_{\hat{U}} &\le \tfrac{1}{1-\delta} \inf_{w\in W} \|u_{W,Z} - w\|_{\hat{U}}, \\ \|u - u_{W,Z}\|_{\hat{U}} + \|r_{W,Z}\|_V &\le \tfrac{2}{1-\delta} \inf_{w\in W} \|u_{W,Z} - w\|_{\hat{U}}. \end{aligned} \tag{11.47}$$

(iii) Z is δ-proximal for W if and only if

$$\inf_{w\in W} \sup_{v\in Z} \frac{b(w, v)}{\|w\|_{\hat{U}} \|v\|_V} \ge \sqrt{1 - \delta^2}. \tag{11.48}$$

Note that (11.45) involves ordinary bilinear forms and finite-dimensional spaces W, Z and (iii) says that the V-projection of the ideal test space $R_V B(W)$ onto Z is a good test space if and only if Z is δ-proximal for W. Loosely speaking, Z is large enough to "see" a substantial part of the ideal test space $R_V B(W)$ under projection. The perhaps most important messages to be taken home regarding the RBM context read as follows.

Remark 4.3. *(i)* The Petrov-Galerkin scheme (11.46) is realized through the saddlepoint problem (11.45) *without* explicitly computing the test space $P_{V,Z}(R_V B(W))$.

(ii) Moreover, given W, by compactness and (11.44), one can in principle enlarge Z so as to make δ as small as possible, a fact that will be exploited later.

(iii) The solution component $u_{W,Z}$ is a near best approximation to the exact solution u in the $\hat{U}$ norm.

(iv) $r_{W,Z}$ can be viewed as a *lifted residual* which tends to zero in V when W grows and can be used for a posteriori error estimation, see [9]. In the Reduced Basis context this can be exploited for certifying the accuracy of the truth solutions and for constructing computationally feasible surrogates for the construction of the reduced bases.

11.5 The Reduced Basis Construction

We point out next how to use the preceding results for sampling the solution manifold $\mathcal{M}$ of a given *parametric family of variational problems*: given $y \in \mathcal{Y}$, $f \in V_y'$, find $u(y) \in U_y$ such that

$$b_y(u(y), v) = \langle f, v\rangle, \quad v \in V_y, \tag{11.49}$$

in a way that the corresponding subspaces are rate-optimal. We will always assume that the dependence of the bilinear form $b_y(\cdot,\cdot)$ on $y \in \mathcal{Y}$ is affine in the sense of (11.18).

As indicated by the notation the spaces U_y, V_y for which the variational problems are well posed in the sense that the induced operator $B_y : U_y \to V_y'$ is bijective, could depend on y through y-dependent norms. However, to be able to speak of a "solution manifold" $\mathcal{M}$ as a compact subset of some "reference Hilbert space," the norms $\|\cdot\|_{U_y}$ should be *uniformly* equivalent to some *reference* norm $\|\cdot\|_U$ which has to be taken into account when formulating (11.49). In fact, under this condition, as shown in [10], for well-posed variational formulations of pure transport problems the dependence of the test spaces V_y on $y \in \mathcal{Y}$ is essential, in that

$$V := \bigcap_{y\in\mathcal{Y}} V_y \tag{11.50}$$

is a strict subset of each individual V_y. This complicates the construction of a tight surrogate. We refer to [10] for ways of dealing with this obstruction and confine the subsequent discussion for simplicity to cases where the test norms $\|\cdot\|_{V_y}$ are also uniformly equivalent to a single reference norm $\|\cdot\|_V$, see the example later below.

Under the above assumptions, the findings of the preceding section will be used next to contrive a well-conditioned tight surrogate even for non-coercive or severely ill-conditioned variational problems which is then in general unsymmetric, i.e., $V_y \neq U_y$. These surrogates will then be used in SGA. To obtain such a residual-based well-conditioned surrogate in the sense of (11.29), we first *renorm* the pairs of spaces U_y or V_y according to (11.34) or (11.36). In anticipation of the example below, for definiteness we concentrate on (11.34) and refer to [10] for a discussion of (11.36). As indicated above, we assume further that the norms $\|\cdot\|_{\hat{U}_y}, \|\cdot\|_{V_y}$ are equivalent to reference norms $\|\cdot\|_{\hat{U}}, \|\cdot\|_V$, respectively.

11.5.1 The Strategy

Suppose that we have already constructed a pair of spaces $U_n \subset U_y, V_n \subset V_y, y \in \mathcal{Y}$, such that for a given $\delta < 1$

$$\inf_{w\in U_n} \sup_{v\in V_n} \frac{b_y(w,v)}{\|w\|_{\hat{U}_y}\|v\|_{V_y}} \geq \sqrt{1-\delta^2}, \quad y \in \mathcal{Y}, \tag{11.51}$$

i.e., $V_n \subset V$ is δ-proximal for $U_n \subset U$. Thus, by Theorem 4.2, the parametric saddle point problem

$$\begin{aligned} (r_n(y), v)_{V_y} + b_y(u_n(y), v) &= \langle f, v\rangle, \ v \in V_n, \\ b(w, r_n(y)) &= 0, \quad w \in U_n, \end{aligned} \tag{11.52}$$

has for each $y \in \mathcal{Y}$ a unique solution $(u_n(y), r_n(y)) \in U_n \times V_n$. By the choice of norms we know that

$$\|u(y) - u_n(y)\|_{\hat{U}_y} = \|f - B_\mu u_n(y)\|_{V_y'}, \quad y \in \mathcal{Y}, \tag{11.53}$$

i.e.,

$$R(y, U_n \times V_n) := \|f - B_\mu u_n(y)\|_{V_y'}, \quad y \in \mathcal{Y}, \tag{11.54}$$

suggests itself as a surrogate. There are some subtle issues about how to evaluate $R(y, U_n \times V_n)$ in the dual $V_{\mathcal{N}}'$ of a sufficiently large truth space $V_{\mathcal{N}} \subset V_y, y \in \mathcal{Y}$, so as to faithfully reflect errors in $\hat{U}_\mu$, not only in the truth space $U_{\mathcal{N}} \subset U_y$ but also in $\hat{U}$, and how these quantities are actually related to the auxiliary variable $\|r_n(y)\|_{V_y}$ which is computed anyway. As indicated before, these issues are aggravated when

the norms $\|\cdot\|_{V_y}$ are *not* all equivalent to a single reference norm. We refer to a corresponding detailed discussion in [10, §5.1] and continue working here for simplicity with the idealized version (11.54) and assume its offline feasibility.

Thus we can evaluate the errors $\|u(y)-u_n(y)\|_{\hat{U}_y}$ and can determine a maximizing parameter y_{n+1} for which

$$\|u(y_{n+1}) - u_n(y_{n+1})\|_{\hat{U}_y} = \max_{y\in\mathcal{Y}} \|f - B_\mu u_n(y)\|_{V_y'}. \tag{11.55}$$

Now relation (11.47) in Theorem 4.2 tells us that for each $y \in \mathcal{Y}$

$$\|u(y) - u_n(y)\|_{\hat{U}_y} \le (1-\delta)^{-1} \inf_{w\in U_n} \|u(y) - w\|_{\hat{U}_y}, \tag{11.56}$$

i.e., $u_n(y)$ is a near best approximation to $u(y)$ from U_n which is, in fact, the closer to the best approximation the smaller δ. By (11.53) and (11.56), the surrogate (11.54) is indeed well conditioned with condition number close to one for small δ.

A natural strategy is now to enlarge U_n to $U_{n+1} := U_n + \operatorname{span}\{u(y_{n+1})\}$. In fact, this complies with the *weak greedy* step (11.20) in §11.3 with weakness parameter $\gamma = (1-\delta)$ as close to one as one wishes, when δ is chosen accordingly small, provided that the pair of spaces U_n, V_n satisfies (11.51). A repetition would therefore, in principle, be a realization of Algorithm 1, SGA, establishing rate-optimality of this RBM. Obviously, the critical condition for such a procedure to work is to ensure at each stage the validity of the weak greedy condition (11.20) which in the present situation means that the companion space V_n is at each stage δ-proximal for U_n. So far we have not explained yet how to grow V_n along with U_n so as to ensure δ-proximality. This is explained in the subsequent section.

Remark 5.1. One should note that, due to the possible parameter dependence of the norms $\|\cdot\|_{\hat{U}_y}, \|\cdot\|_{V_y}$ on y, obtaining tight surrogates with the aid of an explicit Petrov-Galerkin formulation would be infeasible in an RBM context because one would have to recompute the corresponding (parameter dependent) test basis for each parameter query which is not online feasible. It is therefore actually crucial to employ the saddle point formulation in the context of RBMs since this allows us to determine a space V_n of somewhat larger dimension than U_n which stabilizes the saddle point problem *for all y simultaneously*.

11.5.2 A Greedy Stabilization

A natural option is to enlarge V_n by the second component $r_n(y_{n+1})$ of (11.52). Note though that the lifted residuals r_n tend to zero as $n \to \infty$. Hence, the solution manifold of the (y-dependent version of the) saddle point formulation (11.43) has the form

$$\mathcal{M} \times \{0\},$$

where $\mathcal{M}$ is the solution manifold of (11.49) (since $r(y) = 0$ for $y \in \mathcal{Y}$). Thus the spaces V_n are *not* needed to approximate the solution manifold. Instead the sole purpose of the space V_n is to guarantee stability. At any rate, the grown pair $U_{n+1}, V_n + \operatorname{span}\{r_n(y_{n+1})\} =: V_{n+1}^0$ may fail to satisfy now (11.51).

Therefore, in general one has to further enrich V_{n+1}^0 by additional *stabilizing* elements again in a greedy fashion until (11.51) holds for the desired δ. For problems that initially arise as natural saddle point problems such as the Stokes system, enrichments by the so-called *supremizers* (to be defined in a moment) have been proposed already in [14, 15, 22]. In these cases it is possible to enrich V_{n+1}^0 by a *fixed* a priori known number of such supremizers to guarantee inf-sup stability. As shown in [10], this is generally possible when using fixed (parameter independent) reference norms $\|\cdot\|_{\hat{U}}$, $\|\cdot\|_V$ for U and V. For the above more general scope of problems a greedy strategy was proposed and analyzed in [10], a special case of which is also considered in [15] without analysis. The strategy in [10] adds only as many stabilizing elements as are actually needed to ensure stability and works for a much wider range of problems including singularly perturbed ones. In cases where not all parameter-dependent norms $\|\cdot\|_{V_y}$ are equivalent such a strategy is actually necessary and its convergence analysis is then more involved, see [10].

To explain the procedure, suppose that after growing U_n to U_{n+1} we have already generated an enrichment V_{n+1}^k of V_{n+1}^0 (which could be, for instance, either $V_{n+1}^0 := V_n + \operatorname{span}\{r_n(y_{n+1})\}$ or $V_{n+1}^0 := V_n$) but the pair U_{n+1}, V_{n+1}^k still fails to satisfy (11.51) for the given $\delta < 1$. To describe the next enrichment from V_{n+1}^k to V_{n+1}^{k+1} we first search for a parameter $\bar{y} \in \mathcal{Y}$ and a function $\bar{w} \in U_{n+1}$ for which the inf-sup condition (11.51) is worst, i.e.,

$$\sup_{v \in V_{n+1}^k} \frac{b_{\bar{y}}(\bar{w}, v)}{\|v\|_{V_{\bar{y}}} \|\bar{w}\|_{\hat{U}_{\bar{y}}}} = \inf_{y \in \mathcal{Y}} \left(\inf_{w \in U_{n+1}} \sup_{v \in V_{n+1}^k} \frac{b_y(w, v)}{\|v\|_{V_y} \|w\|_{\hat{U}_y}} \right). \tag{11.57}$$

If this worst case inf-sup constant does not exceed yet $\sqrt{1-\delta^2}$, the current space V_{n+1}^k does not contain an effective supremizer for $\bar{y}, \bar{w}$, yet. However, since the truth space satisfies the uniform inf-sup condition (11.51) there *must exist* a good supremizer in the truth space which can be seen to be given by

$$\bar{v} = \operatorname{argmax}_{v \in V_{\bar{y}}} \frac{b_{\bar{y}}(\bar{w}, v)}{\|v\|_{V_{\bar{y}}} \|\bar{w}\|_{\hat{U}_{\bar{y}}}}, \tag{11.58}$$

providing the next enrichment

$$V_{n+1}^{k+1} := \operatorname{span}\{V_{n+1}^k, \bar{v}\}. \tag{11.59}$$

We defer some comments on the numerical realization of finding $\bar{y}, \bar{v}$ in (11.57), (11.58) to the next section.

This strategy can now be applied recursively until one reaches a satisfactory uniform inf-sup condition for the reduced spaces. Again, the termination of this stabilization loop is easily ensured when (11.18) holds and the norms $\|\cdot\|_{\hat{U}_y}$, $\|\cdot\|_{V_y}$ are uniformly equivalent to reference norms $\|\cdot\|_{\hat{U}}$, $\|\cdot\|_V$, respectively, but is more involved in the general case [10].

11.5.3 The Double Greedy Scheme and Main Result

Thus, in summary, to ensure that the greedy scheme SGA with the particular surrogate (11.54), based on the corresponding *outer greedy* step for extending U_n to U_{n+1}, has the *weak greedy property* (11.20), one can employ an *inner stabilizing greedy* loop producing a space $V_{n+1} = V_{n+1}^{k^*}$ which is δ-proximal for U_{n+1}. Here $k^* = k^*(\delta)$ is the number of enrichment steps needed to guarantee the validity of (11.51) for the given δ. A sketchy version of the corresponding "enriched" SGA, developed in [10], looks is given below in Algorithm 3.

As indicated above, both Algorithm 1, SGA, and Algorithm 3, SGA-DOU, are surrogate-based greedy algorithms. The essential difference is that for non-coercive problems or problems with an originally large variational condition number in SGA-DOU an additional interior greedy loop provides a tight well-conditioned (unsymmetric) surrogate which guarantees the desired weak greedy property (with weakness constant γ as close to one as one wishes) needed for rate-optimality.

Of course, the viability of Algorithm SGA-DOU hinges mainly on two questions:

(a) how to find the worst inf-sup constant in (11.57) and how to compute the supremizer in (11.58)?
(b) does the inner greedy loop terminate (early enough)?

As for (a), it is well known that, fixing bases for U_n, V_n^k, finding the worst inf-sup constant amounts to determine for $y \in \mathcal{Y}$ the cross-Gramian with respect to $b_y(\cdot,\cdot)$ and compute its smallest singular value. Since these matrices are of size $n \times (n+k)$ and hence (presumably) of "small" size, a search over $\mathcal{Y}$ requires solving only problems in the reduced spaces and are under assumption (11.18) therefore

Algorithm 3 Double greedy algorithm

```
1: function SGA-DOU
2:     Initialize U_1, V_1^0, δ ∈ (0,1), target accuracy tol, n ← 1,
3:     while σ_n(M) > tol do
4:         while U_n, V_n^0 fail to satisfy (11.51) do
5:             compute V_n with the aid of the inner stabilizing greedy loop,
6:         end while
7:         given U_n, V_n, satisfying (11.51), compute U_{n+1}, V_{n+1}^0 with the aid of the outer greedy
             step 4, (11.8) in algorithm SGA for the surrogate (11.54),
8:     end while
9: end function
```

offline feasible. The determination of the corresponding supremizer $\bar{v}$ in (11.58), in turn, is based on the well-known observation that

$$\operatorname{argmax}_{v\in V_{\bar{y}}} \frac{b_{\bar{y}}(\bar{w},v)}{\|v\|_{V_{\bar{y}}}} = R_{V_{\bar{y}}} B_{\bar{y}} \bar{w},$$

which is equivalent to solving the Galerkin problem

$$(\bar{v},z)_{V_{\bar{y}}} = b_{\bar{y}}(\bar{w},z), \quad z \in V_{\bar{y}}.$$

Thus each enrichment step requires one offline Galerkin solve in the truth space.

A quantitative answer to question (b) is more involved. We are content here with a few related remarks and we refer to a detailed discussion of this issue in [10]. As mentioned before, when all the norms $\|\cdot\|_{\hat{U}_y}, \|\cdot\|_{V_y}$, $y \in \mathcal{Y}$, are equivalent to reference norms $\|\cdot\|_{\hat{U}}, \|\cdot\|_V$, respectively, the inner loop terminates after at most the number of terms in (11.18). When the norms $\|\cdot\|_{V_y}$ are no longer uniformly equivalent to a single reference norm termination is less clear. Of course, since all computations are done in a truth space which is finite dimensional, compactness guarantees termination after finitely many steps. However, the issue is that the number of steps should not depend on the truth space dimension. The reasoning used in [10] to show that (under mild assumptions) termination happens after a finite number of steps, *independent* of the truth space dimension, is based on the following fact. Defining $U_n^1(y) := \{w \in U_n : \|w\|_{\hat{U}_y} = 1\}$, solving the problem

$$(\bar{y},\bar{w}) := \operatorname*{argmax}_{y\in\mathcal{Y}; w\in U_n^1(y)} \inf_{\phi\in V_n^k} \|R_{V_y} B_y w - \phi\|_{V_y}, \tag{11.60}$$

when all the $\|\cdot\|_{\hat{U}_y}$ norms are equivalent to a single reference norm, can be shown to be equivalent to a greedy step of the type (11.57) and can hence again be reduced to similar small eigenvalue problems in the reduced space. Note, however, that (11.60) is similar to a greedy space growth used in the outer greedy loop and for which some understanding of convergence is available. Therefore, successive enrichments based on (11.60) are studied in [10] regarding their convergence. The connection with the inner stabilizing loop based on (11.57) is that

$$\operatorname*{argmax}_{y\in\mathcal{Y}; w\in U_n^1(y)} \inf_{\phi\in V_n^k} \|R_{V_{\bar{y}}} B_{\bar{y}} w - \phi\|_{V_{\bar{y}}} \le \delta$$

just means

$$\inf_{\phi\in V_n^k} \|R_{V_y} B_y w - \phi\|_{V_y} \le \delta \|R_{V_y} B_y\|_{V_y} = \delta \|w\|_{\hat{U}_y}, \quad w \in U_n,\ y \in \mathcal{Y},$$

which is a statement on δ-proximality known to be equivalent to inf-sup stability, see Theorem 4.2, and (11.44).

A central result from [10] can be formulated as follows, see [10, Theorem 5.5].

Theorem 5.2. *If (11.18) holds and the norms* $\|\cdot\|_{\hat{U}_y}, \|\cdot\|_{V_y}$ *are all equivalent to a single reference norm* $\|\cdot\|_{\hat{U}}, \|\cdot\|_V$*, respectively, and the surrogates (11.54) are used, then the scheme* SGA-DOU *is rate-optimal, i.e., the greedy errors* $\sigma_n(\mathcal{M})_{\hat{U}}$ *decay at the same rate as the n-widths* $d_n(\mathcal{M})_{\hat{U}}$*,* $n \to \infty$*.*

Recall that the quantitative behavior of the greedy error rates are directly related to those of the n-widths by $\gamma = c_S$, see Theorem 3.1. This suggests that a fast decay of $d_n(\mathcal{M})_{\hat{U}}$ is reflected by the corresponding greedy errors already for moderate values of n which is in the very interest of reduced order modeling. This will be confirmed by the examples below. In this context an important feature of SGA-DOU is that through the choice of the δ-proximality parameter the weakness parameter γ can be driven toward one, of course, at the expense of somewhat larger spaces V_{n+1}. Hence, stability constants close to one are built into the method. This is to be contrasted by the conventional use of SGA based on surrogates that are not ensured to be well conditioned and for which the computation of the certifying stability constants tends to be computationally expensive, see e.g. [21].

11.5.4 A Numerical Example

The preceding theoretical results are illustrated next by a numerical example that brings out some of the main features of the scheme. While the double greedy scheme applies to noncoercive or indefinite problems (e.g., see [10] for pure transport) we focus here on a classical *singularly perturbed* problem because it addresses also some principal issues for RBMs regarding problems with *small scales*. Specifically, we consider the *convection-diffusion* problem (11.30) on $\Omega = (0,1)^2$ for a simple *parameter-dependent convection* field

$$\vec{b}(y) := \begin{pmatrix} \cos y \\ \sin y \end{pmatrix}, \quad y \in [0, 2\pi), \quad c = 1,$$

keeping for simplicity the diffusion level ε fixed but allowing it to be arbitrarily small. All considerations apply as well to variable and parameter-dependent diffusion with any arbitrarily small but strictly positive lower bound. The "transition" to a pure transport problem is discussed in detail in [10, 28]. Parameter-dependent convection directions mark actually the more difficult case and are, for instance, of interest with regard to kinetic models.

Let us first briefly recall the main challenges posed by (11.30) for very small diffusion ε. The problem becomes obviously dominantly unsymmetric and singularly perturbed. Recall that for each positive ε the problem possesses for each $y \in \mathcal{Y}$ a unique solution $u(y)$ in $U = H_0^1(\Omega)$ that has a zero trace on the boundary $\partial\Omega$. However, as indicated earlier, the condition number $\kappa_{U,U'}(B_y)$ of the underlying convection-diffusion operator B_y, viewed as an operator from $U = H_0^1(\Omega)$ onto $U' = H^{-1}(\Omega)$, behaves like ε^{-1}, that is, it becomes increasingly *ill conditioned*.

This has well known consequences for the performance of numerical solvers but above all for the stability of corresponding discretizations.

We emphasize that the conventional mesh-dependent stabilizations like SUPG (cf. [17]) do *not* offer a definitive remedy because the corresponding condition, although improved, remains very large for very small ε. In [19] SUPG stabilization for the offline truth calculations as well as for the low-dimensional online Galerkin projections are discussed for moderate Peclét numbers of the order of up to 10^3. In particular, comparisons are presented when only the offline phase uses stabilization while the un-stabilized bilinear form is used in the online phase, see also the references in [19] for further related work.

As indicated earlier, we also remark in passing that the singularly perturbed nature of the problem poses an additional difficulty concerning the choice of the truth space $U_{\mathcal{N}}$. In fact, when ε becomes very small one may not be able to afford resolving correspondingly thin layers in the truth space which increases the difficulty of capturing essential features of the solution by the reduced model.

This problem is addressed in [10] by resorting to a weak formulation that does not use $H_0^1(\Omega)$ (or a renormed version of it) as a trial space but builds on the results from [8]. A central idea is to enforce the boundary conditions on the outflow boundary $\Gamma_+(y)$ only weakly. Here $\Gamma_+(y)$ is that portion of $\partial\Omega$ for which the inner product of the outward normal and the convection direction is positive. Thus solutions are initially sought in the larger space $H^1_{0,\Gamma_-(y)}(\Omega) =: U_-(y)$ enforcing homogeneous boundary conditions only on the *inflow* boundary $\Gamma_-(y)$. Since the outflow boundary and hence also the inflow boundary depend on the parameter y, this requires subdividing the parameter set into smaller sectors, here four, for which the outflow boundary $\Gamma_+ = \Gamma_+(y)$ remains unchanged. We refer in what follows for simplicity to one such sector denoted again by $\mathcal{Y}$.

The following prescription of the *test space* falls into the category (11.34) where the norm for U is adapted. Specifically, choosing

$$
\begin{aligned}
s_y(u,v) &:= \frac{1}{2}\big(\langle B_y u, v\rangle + \langle B_y v, u\rangle\big),\\
\|v\|^2_{V_y} &:= s_y(v,v) = \epsilon|v|^2_{H^1(\Omega)} + \Big\|\Big(c - \frac{1}{2}\mathrm{div}\,\vec{b}(y)\Big)^{1/2} v\Big\|^2_{L_2(\Omega)},
\end{aligned}
$$

in combination with a boundary penalization on Γ_+, we follow [8, 28] and define

$$
\|u\|^2_{\bar{U}_y} := \|\bar{B}_y u\|^2_{\bar{V}_y'} = \|\bar{B}_y u\|^2_{V_y'} + \lambda\|u\|^2_{H_b(\mu)},
$$

where $H_b(y) = H^{1/2}_{00}(\Gamma_+(y))$, $\bar{V}_y := V_y \times H_b(y)'$ and $\bar{B}_y$ denotes the operator induced by this weak formulation over $\bar{U}_y := H^1_{0,\Gamma_-(y)}(\Omega) \times H_b(y)$. The corresponding variational formulation is of minimum residual type (cf. (11.38)) and reads

$$
u(y) = \underset{w\in U_-(y)}{\mathrm{argmin}} \big\{\|\bar{B}_y w - f\|^2_{V_y'} + \lambda\|w\|^2_{H_b(y)}\big\}. \tag{11.61}
$$

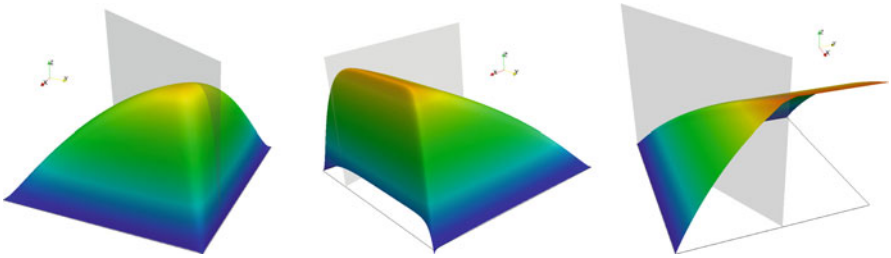


Fig. 11.1 Left: $\varepsilon = 2^{-5}, n = 6, n_V = 13$; middle: $\varepsilon = 2^{-7}, n = 7, n_V = 20$; right: $\varepsilon = 2^{-26}, n = 20, n_V = 57$.

One can show that its (infinite-dimensional) solution, whenever being sufficiently regular, solves also the strong form of the convection diffusion problem (11.30). Figure 11.1 illustrates the effect of this formulation where we set $n = \dim U_n, n_V := \dim V_n$.

The shaded planes shown in Figure 11.1 indicate the convection direction for which the snapshot is taken. For moderately large diffusion the boundary layer at Γ_+ is resolved by the truth space discretization and the boundary conditions at the outflow boundary are satisfied exactly. For smaller diffusion in the middle example the truth space discretization can no longer resolve the boundary layer and for very small diffusion (right) the solution is close to the one for pure transport. The rationale of (11.61) is that all norms commonly used for convection-diffusion equations resemble the one chosen here, for instance in the form of a mesh-dependent "broken norm," which means that most part of the incurred error of an approximation is concentrated in the layer region, see, e.g., [24, 27]. Hence, when the layers are not resolved by the discretization, enforcing the boundary conditions does not improve accuracy and, on the contrary, may degrade accuracy away from the layer by causing oscillations. The present formulation instead avoids any nonphysical oscillations and enhances accuracy in those parts of the domain where this is possible for the afforded discretization, see [8, 10, 28] for a detailed discussion. The following table quantifies the results for the case of small diffusion $\varepsilon = 2^{-26}$ and a truth discretization whose a posteriori error bound is 0.002.

Columns 3 and 8 show the δ governing the condition of the saddle point problems (and hence of the corresponding Petrov-Galerkin problems), see (11.51), the greedy space growth is based upon (Table 11.1). Hence, the surrogates are very tight giving rise to weakness parameters very close to one. As indicated in Remark 4.3 one can use also an a posteriori bound for the truth solution based on the corresponding lifted residual. Columns 5 and 10 show therefore the relative accuracy of the current reduced model and the truth model. This corresponds to the stability constants computed by conventional RBMs. Even for elliptic problems these latter ones are significantly larger than the ones for the present singularly perturbed problem which are guaranteed to be close to one by the method itself. Based on the a posteriori bounds for the truth solution (which are also obtained with the aid of tailored

Table 11.1 Convection-diffusion equation, $\varepsilon = 2^{-26}$, maximal a posteriori error 0.00208994

n	n_V	δ	Surrogate	Surr/a-post	n	n_V	δ	Surrogate	Surr/a-post
2	5	1.36e-03	2.10e-01	1.01e+02	14	39	1.17e-04	8.15e-03	3.90e+00
4	9	1.10e-02	7.51e-02	3.59e+01	16	45	9.79e-05	7.56e-03	3.62e+00
6	15	1.75e-03	4.95e-02	2.37e+01	18	51	6.32e-05	7.40e-03	3.54e+00
8	21	9.16e-04	2.34e-02	1.12e+01	20	57	4.74e-05	6.09e-03	2.92e+00
10	27	3.65e-04	2.05e-02	9.82e+00	22	63	2.36e-05	5.43e-03	2.60e+00
12	33	3.34e-04	1.56e-02	7.45e+00	24	65	2.36e-05	4.73e-03	2.27e+00

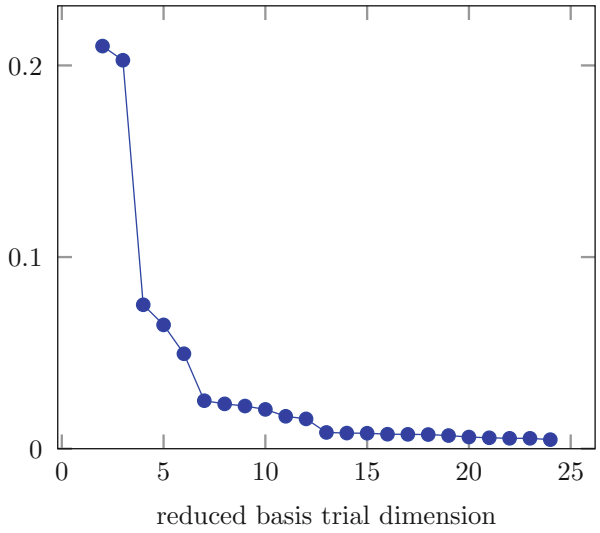


Fig. 11.2 Convection-diffusion equation, $\varepsilon = 2^{-26}$, maximal a posteriori error 0.00208994

well-conditioned variational formulations, see [8]), the greedy space growth is stopped when the surrogates reach the order of the truth accuracy. As illustrated in Figure 11.2, in the present example this is essentially already the case for ≤ 20 trial reduced basis functions and almost three times as many test functions. To show this "saturation effect" we have continued the space growth formally up to $n = 24$ showing no further significant improvement which is in agreement with the resolution provided by the truth space. These relations agree with the theoretical predictions in [10]. Figure 11.2 illustrates also the rapid gain of accuracy by the first few reduced basis functions which supports the fact that the solution manifold is "well seen" by the Petrov-Galerkin surrogates. More extensive numerical tests shown in [10] show that the achieved stability is independent of the diffusion but the larger the diffusion the smaller become the dimensions $n = \dim U_n, n_V = \dim V_n$ for the reduced spaces. This indicates the expected fact that the larger the diffusion the smoother is the dependence of $u(y)$ on the parameter y. In fact, when $\varepsilon \to 0$ one

approaches the regime of pure transport where the smoothness of the parameter dependence is merely Hölder continuity requiring for a given target accuracy a larger number of reduced basis functions, see [10].

11.6 Is it Necessary to Resolve All of $\mathcal{M}$?

The central focus of the preceding discussion has been to control the maximal deviation

$$\sigma_n(\mathcal{M})_U = \max_{y\in\mathcal{Y}} \|u(y) - P_{U,U_n}u(y)\|_U \tag{11.62}$$

and to push this deviation below a given tolerance for n as small as possible. However, in many applications one is not interested in the whole solution field but only in a *quantity of interest* $I(y)$, typically of the form $I(y) = \ell(u(y))$ where $\ell \in U'$ is a bounded linear functional. Looking then for some desired optimal state $I^* = \ell(u(y^*))$ one is interested in a guarantee of the form

$$|\ell(u_n(y)) - \ell(u(y))| \le \text{tol}, \quad y \in \mathcal{Y}, \tag{11.63}$$

where the states $u_n(y)$ belong to a possibly small reduced space U_n in order to be then able to carry out the optimization over $y \in \mathcal{Y}$ in the small space $U_n \subset U$. Asking only for the values of just a *linear functional* of the solution seems to be much less demanding than asking for the whole solution and one wonders whether this can be exploited in favor of even better online efficiency.

Trying to reduce computational complexity by exploiting the fact that retrieving only a linear functional of an unknown state - a scalar quantity - may require less information than recovering the whole state is the central theme of *goal-oriented* adaptation in finite element methods, see [3]. Often the desired accuracy is indeed observed to be reached by significantly coarser discretizations than needed to approximate the whole solution within a corresponding accuracy. The underlying effect, sometimes referred to as "squared accuracy" is well understood and exploited in the RBM context as well, see [16, 21]. We briefly sketch the main ideas for the current larger scope of problems and point out that, nevertheless, a guarantee of the form (11.63) ultimately requires controlling the maximal deviation of a reduced space in the sense of (11.62). Hence, an optimal sampling of a solution manifold remains crucial.

First, a trivial estimate gives for $\ell \in U'$

$$|\ell(u_n(y)) - \ell(u(y))| \le \|\ell\|_{U'}\|u_n(y) - u(y)\|_U \tag{11.64}$$

so that a control of $\sigma_n(\mathcal{M})_U$ would indeed yield a guarantee. However, the n needed to drive $\|\ell\|_{U'}\sigma_n(\mathcal{M})_U$ below tol is usually larger than necessary.

To explain the principle of improving on (11.64) we consider again a variational problem of the form (11.31) (suppressing any parameter dependence for a moment) for a pair of spaces U, V where we assume now that $\kappa_{U,V'}(B) \leq C_b/c_b$ is already small, possibly after renorming an initial less favorable formulation through (11.34) or (11.36). Let $u \in U$ again denote the exact solution of (11.31). Given a $\ell \in U'$ we wish to approximate $\ell(u)$, using an *approximate solution* $\bar{u} \in W \subset U$ defined by

$$b(\bar{u}, v) = \langle f, v\rangle, \quad v \in \tilde{V}_W \subset V, \tag{11.65}$$

where $\tilde{V}_W$ is a suitable test space generated by the methods discussed in §11.4.1. In addition we will use the solution $z \in V$ of the *dual problem*:

$$b(w, z) = -\ell(w), \quad w \in U, \tag{11.66}$$

together with an approximation $\bar{z} \in Z \subset V$ defined by

$$b(w, \bar{z}) = -\ell(w), \quad w \in \tilde{W}_Z \subset U, \tag{11.67}$$

again with a suitable test space $\tilde{W}_Z$. Recall that we need not determine the test spaces $\tilde{V}_W, \tilde{W}_Z$ explicitly but rather realize the corresponding Petrov-Galerkin projections through the equivalent saddle-point formulations with suitable δ-proximal auxiliary spaces generated by a greedy stabilization.

Then, defining the primal residual functional

$$r_{\bar{u}}(v) := r(\bar{u}, v) := b(u - \bar{u}, v) = \langle f, v\rangle - b(\bar{u}, v) \tag{11.68}$$

and adapting the ideas in [16, 21] for the symmetric case $V = U$ to the present slightly more general setting, we claim that

$$\hat{\ell}(\bar{u}) := \ell(\bar{u}) - r(\bar{u}, \bar{z}) \tag{11.69}$$

is an approximation to the true value $\ell(u)$ satisfying

$$|\hat{\ell}(\bar{u}) - \ell(u)| \leq C \inf_{w\in W} \|u - w\|_U \inf_{v\in Z} \|z - v\|_V, \tag{11.70}$$

where C depends only on the inf-sup constant of the finite-dimensional problems. In fact, since by (11.66),

$$\ell(u) - \ell(\bar{u}) = b(\bar{u} - u, z) = -r(\bar{u}, z),$$

one has $\ell(u) = \ell(\bar{u}) - r(\bar{u}, z)$ and hence

$$\begin{aligned} |\hat{\ell}(\bar{u}) - \ell(u)| &= |\ell(\bar{u}) - r(\bar{u}, \bar{z}) - \ell(\bar{u}) + r(\bar{u}, z)| = |r(\bar{u}, z - \bar{z})| = |b(u - \bar{u}, z - \bar{z})| \\ &\leq C_b \|u - \bar{u}\|_U \|z - \bar{z}\|_V, \end{aligned}$$

which confirms the claim since $\bar{u}, \bar{z}$ are near best approximations due to the asserted inf-sup stability of the finite-dimensional problems.

Clearly, (11.70) says that in order to approximate $\ell(u)$ the primal approximation in U need not resolve u at all as long as the dual solution z is approximated well enough. Moreover, when ℓ is a local functional, e.g., a local average approximating a point evaluation, z is close to the corresponding Green's function with (near) singularity in the support of ℓ. In the elliptic case z would be very smooth away from the support of ℓ and hence well approximable by a relatively small number of degrees of freedom concentrated around the support of ℓ. Thus it may very well be more profitable to spend less effort on approximating u than on approximating z.

Returning to parameter-dependent problems (11.49), the methods in §11.5 can now be used as follows to construct possibly small reduced spaces for a frequent online evaluation of the quantities $I(y) = \ell(u(y))$. We assume that we already have properly renormed families of norms $\|\cdot\|_{U_y}, \|\cdot\|_{V_y}$, $y \in \mathcal{Y}$, with uniform inf-sup constants close to one. We also assume now that both families of norms are equivalent (by compactness of $\mathcal{Y}$ uniformly equivalent) to reference norms $\|\cdot\|_U, \|\cdot\|_V$, respectively. Hence, we can consider two solution manifolds

$$\mathcal{M}_{\rm pr} := \{u(y) = B_y^{-1} f,\ y \in \mathcal{Y}\} \subset U, \quad \mathcal{M}_{\rm dual} := \{z(y) := B_y^{-*}\ell,\ y \in \mathcal{Y}\} \subset V,$$

and use Algorithm 3, SGA-DOU, to generate (essentially in parallel) two sequences of pairs of reduced spaces

$$(U_n, V_n),\ (Z_n, W_n), \quad n \in \mathbb{N}.$$

Here $V_n \subset V, W_n \subset U$ are suitable stabilizing spaces such that for $m < n$ and for the corresponding reduced solutions $u_m(y) \in U_m, z_{n-m}(y) \in Z_{n-m}$ the quantity

$$I_{n,m}(y) := \ell(u_m(y)) - r(u_m(y), z_{n-m}(y)) \tag{11.71}$$

satisfies

$$|I(y) - I_{n,m}(y)| \le C\sigma_m(\mathcal{M}_{\rm pr})_U \sigma_{n-m}(\mathcal{M}_{\rm dual})_V, \tag{11.72}$$

with a constant C independent of n, m. The choice of $m < n$ determines how to distribute the computational effort for computing the two sequences of reduced bases and their stabilizing companion spaces. By Theorem 5.2, one can see that whichever n-width rate $d_n(\mathcal{M}_{\rm pr})_U$ or $d_n(\mathcal{M}_{\rm dual})_V$ decays faster one can choose $m < n$ to achieve for a total of $\dim U_m + \dim Z_{n-m} = n$ the smallest error bound. Of course, the rates are not known and one can use the tight surrogates to bound and estimate the respective errors very accurately. For instance, when $d_n(\mathcal{M}_{\rm pr})_U \le Cn^{-\alpha}$, $d_n(\mathcal{M}_{\rm dual})_V \le Cn^{-\beta}$, $m = \left\lfloor \left(\frac{\alpha}{\alpha+\beta}\right) n \right\rfloor$ yields an optimal distribution with a bound

$$|I(y) - I_{n,m}(y)| \le C\Big(\frac{\alpha+\beta}{\beta}\Big)^{\beta}\Big(\frac{\alpha+\beta}{\alpha}\Big)^{\alpha} n^{-(\alpha+\beta)}. \tag{11.73}$$

In particular, when $\beta > \alpha$ the dimensions on the reduced bases for the dual problem should be somewhat larger but essentially using the same dimensions for the primal and dual reduced spaces yields the rate $n^{-(\alpha+\beta)}$ confirming the "squaring" when $\alpha = \beta$. In contrast, as soon as either one of the n-width rates decays exponentially it is best to grow only the reduced spaces for the faster decay while keeping a fixed space for the other side.

11.7 Summary

We have reviewed recent developments concerning reduced basis methods with the following main focus. Using Kolmogorov n-width as a benchmark for the performance of reduced basis methods in terms of minimizing the dimensions of the reduced models for a given target accuracy, we have shown that this requires essentially to construct tight well-conditioned surrogates for the underlying variational problem. We have explained how *renormation* in combination with *inner stabilization loops* can be used to derive such residual-based surrogates even for problem classes not covered by conventional schemes. This includes in a fully robust way indefinite as well as ill-conditioned (singularly perturbed) coercive problems. Greedy strategies based on such surrogates are then shown to constitute an optimal sampling strategy, i.e., the resulting snapshots span reduced spaces whose distances from the solution manifold decay essentially at the same rate as the Kolmogorov n-widths. This means, in particular, that stability constants need not be determined by additional typically expensive computations but can be pushed by the stabilizing inner greedy loop as close to one as one wishes. Finally, we have explained why the focus on uniform approximation of the entire solution manifold is equally relevant for applications where only functionals of the parameter-dependent solutions have to be approximated.

References

1. J.W. Barrett, K.W. Morton, Approximate symmetrization and Petrov-Galerkin methods for diffusion-convection problems. Comput. Method. Appl. Mech. **45**, 97–12 (1984)
2. A. Barron, A. Cohen, W. Dahmen, R. DeVore, Approximation and learning by greedy algorithms. Ann. Stat. **3**(1), 64–94 (2008)
3. R. Becker, R. Rannacher, An optimal error control approach to a–posteriori error estimation. Acta Numer. **10**, 1–102 (2001)
4. P. Binev, A. Cohen, W. Dahmen, R. DeVore, G. Petrova, P. Wojtaszczyk, Convergence rates for greedy algorithms in reduced basis methods. SIAM J. Math. Anal. **43**, 1457–1472 (2011)

5. F. Brezzi, M. Fortin, *Mixed and Hybrid Finite Element Methods*. Springer Series in Computational Mathematics, vol. 15 (Springer, New York, 1991) [ISBN: 978-1-4612-7824-5 (Print) 978-1-4612-3172-1 (Online)]
6. A. Buffa, Y. Maday, A.T. Patera, C. Prud'homme, G. Turinici, A Priori convergence of the greedy algorithm for the parameterized reduced basis. ESAIM Math. Model. Numer. Anal. **46**(03), 595–603 (2012)
7. A. Cohen, R. DeVore, C. Schwab, Convergence rates of best N-term Galerkin approximations for a class of elliptic sPDEs. Found. Comput. Math. **10**, 615–646 (2010)
8. A. Cohen, W. Dahmen, G. Welper, Adaptivity and variational stabilization for convection-diffusion equations. ESAIM Math. Model. Numer. Anal. **46**(5), 1247–1273 (2012)
9. W. Dahmen, C. Huang, C. Schwab, G. Welper, Adaptive Petrov-Galerkin methods for first order transport equations. SIAM J. Numer. Anal. **50**(5), 2420–2445 (2012)
10. W. Dahmen, C. Plesken, G. Welper, Double greedy algorithms: reduced basis methods for transport dominated problems. ESAIM Math. Model. Numer. Anal. **48**(3), 623–663 (2014). doi:10.1051/m2an/2013103
11. L.F. Demkowicz, J. Gopalakrishnan, A class of discontinuous Petrov-Galerkin methods I: the transport equation. Comput. Methods Appl. Mech. Eng. **199**(23–24), 1558–1572 (2010)
12. L. Demkowicz, J. Gopalakrishnan, A class of discontinuous Petrov-Galerkin methods. Part II: optimal test functions. Numer. Methods Partial Differ. Equ. **27**(1), 70–105 (2011)
13. R. DeVore, G. Petrova, P. Wojtaszczyk, Greedy algorithms for reduced bases in Banach spaces. Constr. Approx. **37**, 455–466 (2013)
14. A.-L. Gerner, K. Veroy, Reduced basis a posteriori error bounds for the Stokes equations in parameterized domains: a penalty approach. Math. Models Methods Appl. Sci. **21**(10), 2103–2134 (2011)
15. A. Gerner, K. Veroy-Grepl, Certified reduced basis methods for parametrized saddle point problems. Preprint, SIAM J. Sci. Comput. **34**(5), A2812–A2836 (2012)
16. M.A. Grepl, A.T. Patera, A posteriori error bounds for reduced-basis approximations of parametrized parabolic partial differential equations. Math. Model. Numer. Anal. (M2AN) **39**(1), 157–181 (2005)
17. T. Hughes, G. Sangalli, Variational multiscale analysis: the fine-scale Green's function, projection, optimization, localization, and stabilized methods. SIAM J. Numer. Anal. **45**(2), 539–557 (2007)
18. T. Manteuffel, S. McCormick, J. Ruge, J.G. Schmidt, First-order system $\mathcal{LL}^*$ $(FOSLL)^*$ for general scalar elliptic problems in the plane. SIAM J. Numer. Anal. **43**, 2098–2120 (2005)
19. P. Pacciarini, G. Rozza, Stabilized reduced basis method for parametrized advection–diffusion PDEs. Comput. Methods Appl. Mech. Eng. **274**, 1–18 (2014)
20. A.T. Patera, G. Rozza, Reduced Basis Approximation and A Posteriori Error Estimation for Parametrized Partial Differential Equations, Version 1.0, Copyright MIT 2006–2007 (tentative rubric). MIT Pappalardo Graduate Monographs in Mechanical Engineering. Available at http://augustine.mit.edu
21. C. Prud'homme, D.V. Rovas, K. Veroy, L. Machies, Y. Maday, A.T. Patera, G. Turinici, Reliable real-time solution of parametrized partial differential equations: reduced basis output-bound methods. Trans. ASME **124**, 70–80 (2002)
22. G. Rozza, K. Veroy, On the stability of reduced basis techniques for Stokes equations in parametrized domains. Comput. Methods Appl. Mech. Eng. **196**(7), 1244–1260 (2007)
23. G. Rozza, D.B.P. Huynh, A.T. Patera, Reduced basis approximation and a posteriori error estimation for affinely parametrized elliptic coercive partial differential equations. Arch. Comput. Methods Eng. **15**, 229–275 (2008)
24. G. Sangalli, A uniform analysis of non-symmetric and coercive linear operators. SIAM J. Math. Anal. **36**(6), 2033–2048 (2005)

25. S. Sen, K. Veroy, D.B.P. Huynh, S. Deparis, N.C. Nguyn, A.T. Patera, "Natural norm" a-posteriori error estimators for reduced basis approximations. J. Comput. Phys. **217**, 37–62 (2006)
26. V. Temlyakov, Weak greedy algorithms. Adv. Comput. Math. **12**, 213–227 (2000)
27. R. Verfürth, Robust a posteriori error estimates for stationary convection-diffusion equations. SIAM J. Numer. Anal. **43**(4), 1766–1782 (2005)
28. G. Welper, Infinite dimensional stabilization of convection-dominated problems. Ph.D. Thesis, RWTH Aachen, 2012

Chapter 12
On the Stability of Polynomial Interpolation Using Hierarchical Sampling

Albert Cohen and Abdellah Chkifa

Abstract Motivated by the development of nonintrusive methods for high-dimensional parametric PDEs, we study the stability of a sparse high-dimensional polynomial interpolation procedure introduced in Chkifa et al. (Found. Comput. Math. 1–33, 2013). A key aspect of this procedure is its hierarchical structure: the sampling set is progressively enriched together with the polynomial space. The evaluation points are selected from a grid obtained by tensorization of a univariate sequence. The Lebesgue constant that quantifies the stability of the resulting interpolation operator depends on the choice of this sequence. Here we study $\Re$-Leja sequences, obtained by the projection of Leja sequences on the complex unit disk, with initial value 1, onto $[-1, 1]$. For this sequence, we prove cubic growth in the number of points for the Lebesgue constant of the multivariate interpolation operator, independently of the number of variable and of the shape of the polynomial space.

12.1 Introduction

This chapter deals with a high-dimensional interpolation process, for which the sampling set is hierarchically enriched, in parallel with the polynomial space. Our main motivation for considering this process is the development of non-intrusive methods for high-dimensional parametric PDE.

Parametric PDEs are equations with the general form

$$\mathscr{D}(u, y) = 0, \tag{12.1}$$

where $\mathscr{D}$ is a differential operator and $y := (y_1, \ldots, y_d)$ is a parameter vector in a tensor product domain X^d. Up to a change of variable, typical choices for X are the real interval $[-1, 1]$ or the complex unit disk $\{|z| \leq 1\}$. The solution u to such PDEs

A. Cohen (✉) • A. Chkifa
Laboratoire Jacques-Louis Lions, UPMC Univ Paris 06, UMR 7598, 75005 Paris, France

Laboratoire Jacques-Louis Lions, CNRS, UMR 7598, 75005 Paris, France
e-mail: cohen@ann.jussieu.fr; chkifa@ann.jussieu.fr

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_12

is therefore a function of y, which may be deterministic or stochastic depending on the context of application, in addition to the usual space and time variable. Assuming well posedness of the problem for all $y \in X^d$ in some Banach space V, we may define the solution map

$$y \mapsto u(y), \tag{12.2}$$

acting from X^d to V. For certain relevant parametric PDEs, this map is uniformly bounded, and therefore belongs to $L^\infty(X^d, V)$. This is the case for instance for the linear diffusion equation

$$-\mathrm{div}(a(y)u) = f, \tag{12.3}$$

set on a bounded Lipschitz domain $D \subset R^m$, with $f \in L^2(D)$ and boundary conditions $u = 0$ on ∂D, provided that for some fixed $r > 0$ the diffusion coefficient $a(y) \in L^\infty(D)$ satisfies the ellipticity condition

$$r \leq a(y), \tag{12.4}$$

for all $y \in X^d$.

Parametric PDEs raise significant computational challenges in the high dimensional context, that is when $d >> 1$ or $d = +\infty$. Recent results such as in [8–10] have shown the effectiveness of approximating the map $y \mapsto u(y)$ to certain such PDEs by multivariate polynomials in the parametric variables $(y_1, \ldots, y_d)$. Here, the multivariate polynomial spaces are of the general form

$$\mathbb{P}_\Lambda := \mathrm{Span}\{y^\nu = y_1^{\nu_1} \ldots y_d^{\nu_d} \ : \ \nu = (\nu_1, \ldots, \nu_d) \in \Lambda\}, \tag{12.5}$$

where $\Lambda \in \mathbb{N}^d$ is an index set that is assumed to be downward closed (also called lower set), in the sense that for $\nu := (\nu_1, \ldots, \nu_d), \mu := (\mu_1, \ldots, \mu_d) \in \mathbb{N}^d$, we have

$$\nu \in \Lambda \ \text{ and } \ \mu_i \leq \nu_i, \ \ i = 1, \ldots, d \quad \Rightarrow \quad \mu \in \Lambda. \tag{12.6}$$

It was shown in [7, 8] that for relevant classes of parametric PDEs, certain sequences of downward closed index sets

$$\Lambda_1 \subset \Lambda_2 \subset \cdots \subset \mathbb{N}^d \tag{12.7}$$

with $\#(\Lambda_k) = k$ break the curse of dimensionality in the sense that the polynomial approximation error decays with k at a rate k^{-s} that does not deteriorates as d gets large, in the sense that it remains valid even when $d = \infty$.

One practical way to construct such polynomial approximations is by interpolation, based on the evaluation of u at certain points $y^i \in X^d$. One attractive feature of such an approach is that it is nonintrusive and therefore can benefit from existing numerical codes for evaluating $y \mapsto u(y)$ pointwise. An important issue

for computational simplicity and economy is that the sampling and interpolation procedure should be hierarchical: the solution u is evaluated at only one new point in X^d when Λ_k is updated to Λ_{k+1}.

Such a procedure was recently proposed and analyzed in [6]. It is based on the data of a sequence $Z := (z_i)_{i\geq 0}$ of pairwise distinct points in X and the univariate interpolation operators I_k onto $\mathbb{P}_k$ associated with the sections $\{z_0, \cdots, z_k\}$. The corresponding multivariate interpolation operator I_Λ onto $\mathbb{P}_\Lambda$ is constructed by the Smolyack process of tensorization and sparsification based on the difference operators $\Delta_k := I_k - I_{k-1}$, which is described in §12.2 of this chapter. We also show that there is a simple relation between the algebraic growth of the Lebesgue constant $\mathbb{L}_\Lambda := \|I_\Lambda\|_{L^\infty \to L^\infty}$ in terms of $\#(\Lambda)$ and that of its univariate counterpart $\mathbb{L}_k := \|I_k\|_{L^\infty \to L^\infty}$ or of $\mathbb{D}_k := \|\Delta_k\|_{L^\infty \to L^\infty}$ in terms of $(k+1)$.

This motivates the search for "good" univariate sequences Z of points on $[-1, 1]$ such that the Lebesgue constant $\mathbb{L}_k$ or the norm of the difference operator $\mathbb{D}_k$, has moderate algebraic growth, controlled by $(1+k)^\theta$ for a small θ. Note that it is well known that the Lebesgue constant grows logarithmically with k for certain choices of non-nested sets of points on $[-1, 1]$, such as Chebychev and Gauss-Lobatto points; however, it is not clear that such a very slow growth is possible for nested sets corresponding to the sections of a sequence Z.

In this chapter, we consider the so-called $\Re$-Leja sequences, obtained by the projection of Leja sequences on the complex unit disk, with initial value 1, onto $[-1, 1]$, and studied in [3, 4]. We recall in §12.3 some main properties of these sequences. We then obtain in §12.4 the bound $\mathbb{L}_k \leq 8\sqrt{2}(1+k)^2$, which improves on the $\mathcal{O}((k+1)^3 \log(k+1))$ bound given in [3] and on the $\mathcal{O}((k+1)^2 \log(k+1))$ bound given in [4]. Then in §12.5, we establish the improved bound $\mathbb{D}_k \leq (1+k)^2$ for the difference operator, which could not be obtained directly from $\mathbb{D}_k \leq \mathbb{L}_k + \mathbb{L}_{k-1}$. A consequence of this last result is that using the $\Re$-Leja sequence, the resulting multivariate interpolation operator has Lebesgue constant with bound

$$\mathbb{L}_\Lambda \leq (\#\Lambda)^3, \tag{12.8}$$

whatever the dimension d and the shape of the finite downward closed set Λ.

12.2 Sparse polynomial interpolation

In this section, we recall the construction of the multivariate interpolation operator proposed in [6]. Given an infinite sequence $Z := (z_i)_{i\geq 0}$ of pairwise distinct points in X, we define I_k the univariate interpolation operator onto $\mathbb{P}_k$ associated with the section $\{z_0, \cdots, z_k\}$. We may express I_k as the telescoping sum

$$I_k = \sum_{l=0}^{k} \Delta_l, \qquad \Delta_l := I_l - I_{l-1}, \tag{12.9}$$

with the convention that $I_{-1} = 0$, which corresponds to the Newton form with

$$\Delta_k f = \Big(f(z_k) - I_{k-1}f(z_k)\Big)h_k, \quad h_0(z) = 1, \;\; h_k(z) = \prod_{j=0}^{k-1} \frac{z - z_j}{z_k - z_j}. \tag{12.10}$$

Now, for an arbitrary downward closed set $\Lambda \subset \mathbb{N}^d$, we introduce the grid of points

$$\Gamma_\Lambda := \Big\{z_\nu : \nu \in \Lambda\Big\} \quad \text{where} \quad z_\nu := (z_{\nu_j})_{j=1,\dots,d} \in X^d. \tag{12.11}$$

We also introduce the operator

$$I_\Lambda := \sum_{\nu\in\Lambda} \Delta_\nu, \quad \text{where} \quad \Delta_\nu := \otimes_{j=1,\dots,d}\Delta_{\nu_j}. \tag{12.12}$$

We observe that this coincides with (12.9) for the univariate case $d = 1$ when $\Lambda = \{0, 1, \dots, k\}$. We also observe that when Λ is a rectangular block, that is

$$\Lambda = \mathscr{B}_\mu := \{\nu \;:\; \nu \le \mu\}, \tag{12.13}$$

for some μ, then

$$I_\Lambda = \sum_{\nu_1=1}^{\mu_1} \cdots \sum_{\nu_d=1}^{\mu_d} \otimes_{j=1}^d \Delta_{\nu_j} = \otimes_{j=1,\dots,d}\Big(\sum_{\nu_j=1}^{\mu_j} \Delta_{\nu_j}\Big) = \otimes_{j=1,\dots,d} I_{\mu_j} \tag{12.14}$$

is the interpolation operator for the tensor product polynomial space $\mathbb{P}_\Lambda := \otimes_{j=1,\dots,d}\mathbb{P}_{\mu_j}$ and the tensor product grid $\Gamma_\Lambda = \otimes_{j=1,\dots,d}\{z_0, \dots, z_{\mu_j}\}$.

The following result is given in [6] but its first appearance dates back from [13] in the bi-dimensional case. It shows that the previous observation generalizes to any downward closed set.

Theorem 1. *The grid Γ_Λ is unisolvent for the polynomial space $\mathbb{P}_\Lambda$ and the interpolation operator is given by I_Λ.*

Proof. Since $\#(\Lambda) = \dim(\mathbb{P}_\Lambda)$ and the image of I_Λ is obviously contained in $\mathbb{P}_\Lambda$, it suffices to show that I_Λ is the interpolation operator, that is, $I_\Lambda f(z_\mu) = f(z_\mu)$ for all $\mu \in \Lambda$. This is shown by splitting $I_\Lambda f$ into

$$I_\Lambda f = I_{\mathscr{B}_\mu} f + (I_\Lambda - I_{\mathscr{B}_\mu})f, \tag{12.15}$$

where $\mathscr{B}_\mu$ is the rectangular block in (12.13). For the first, we have already observed that $I_{\mathscr{B}_\mu} f(z_\mu) = f(z_\mu)$. The second part in the above splitting is a sum of terms $\Delta_\nu f$ where ν is such that $\nu_j > \mu_j$ for at least one value of j. For this value we have $\Delta_{\nu_j} f(z_{\mu_j}) = I_{\nu_j} f(z_{\mu_j}) - I_{\nu_j - 1} f(z_{\mu_j}) = f(z_{\mu_j}) - f(z_{\mu_j}) = 0$, which implies that $\Delta_\nu f(z_\mu) = 0$. Therefore $(I_\Lambda - I_{\mathscr{B}_\mu})f(z_\mu) = 0$ which concludes the proof. □

One main interest of the above construction is that it is hierarchical in the sense that the enrichment of Λ by a new index μ corresponds to adding one sampling point z_μ to the grid Γ_Λ. In a similar way to the univariate case, the hierarchical computation of the interpolant is possible, based on the formula

$$\Delta_\nu f = \Big(f(z_\nu) - I_\Lambda f(z_\nu)\Big) H_\nu, \quad H_\nu(z) = \prod_{j=1}^{d} h_{\nu_j}(z_j), \tag{12.16}$$

which holds whenever Λ is any downward closed set such that $\nu \notin \Lambda$ and $\Lambda \cup \{\nu\}$ is also a downward closed set. This hierarchical form allows us to develop adaptive interpolation algorithms: given a certain set Λ_n of cardinality n, one picks a new index ν^{n+1} which maximizes the contribution $\Delta_\nu f$ in some norm of interest (typically L^p for $p = 1, 2$ or ∞) among those $\nu \notin \Lambda_n$ such that $\Lambda_n \cup \{\nu\}$ is a downward closed set. The numerical behavior of such adaptive algorithms is studied in [6].

The stability of the operators I_Λ is critical for numerical applications such as the nonintrusive treatment of parametric PDEs. It is measured by the Lebesgue constant

$$\mathbb{L}_\Lambda := \max_{f \in C(X^d) - \{0\}} \frac{\|I_\Lambda f\|_{L^\infty(X^d)}}{\|f\|_{L^\infty(X^d)}}. \tag{12.17}$$

In particular, we have the classical estimate

$$\|f - I_\Lambda f\|_{L^\infty(X^d)} \leq (1 + \mathbb{L}_\Lambda) \inf_{g \in \mathbb{P}_\Lambda} \|f - g\|_{L^\infty(X^d)}. \tag{12.18}$$

This constant depends on the sequence Z, in particular through the Lebesgue constant of the univariate interpolation operators

$$\mathbb{L}_k := \max_{f \in C(X) - \{0\}} \frac{\|I_k f\|_{L^\infty(X)}}{\|f\|_{L^\infty(X)}} = \max_{z \in X} \lambda_k(z), \tag{12.19}$$

where λ_k is the Lagrange function for the section $\{z_0, \ldots, z_k\}$ defined by

$$\lambda_k(z) := \sum_{i=0}^{k} |l_{i,k}(z)|, \quad z \in X, \tag{12.20}$$

where

$$l_{i,k}(z) := \prod_{\substack{j=0,\ldots,k,\\ j \neq i}} \frac{z - z_j}{z_i - z_j}, \quad z \in X, \tag{12.21}$$

for $j = 0, \ldots, k$ are the Lagrange polynomials associated with $\{z_0, \ldots, z_k\}$.

It is shown in [6] that algebraic growth of $\mathbb{L}_k$ yields algebraic growth of the Lebesgue constant $\mathbb{L}_\Lambda$. More precisely, given any $\theta \geq 1$

$$\mathbb{L}_k \leq (1+k)^\theta, \quad \text{for any} \quad k \geq 1 \quad \Longrightarrow \quad \mathbb{L}_\Lambda \leq (\#\Lambda)^{\theta+1}. \tag{12.22}$$

Surprisingly, the previous implication is valid whatever the dimension d and the shape of the finite downward closed set Λ.

A more straightforward computation shows that we also have

$$\mathbb{D}_k \leq (1+k)^\theta, \quad \text{for any} \quad k \geq 1 \quad \Longrightarrow \quad \mathbb{L}_\Lambda \leq (\#\Lambda)^{\theta+1}, \tag{12.23}$$

where

$$\mathbb{D}_k := \max_{f \in C(X)-\{0\}} \frac{\|\Delta_k f\|_{L^\infty(X)}}{\|f\|_{L^\infty(X)}}. \tag{12.24}$$

Indeed, by triangle inequality, we find that

$$\mathbb{L}_\Lambda \leq \sum_{\nu \in \Lambda} \prod_{j=1}^{d} \mathbb{D}_{\nu_j} \leq \sum_{\nu \in \Lambda} \prod_{j=1}^{d} (1+\nu_j)^\theta = \sum_{\nu \in \Lambda} (\#(\mathscr{B}_\nu))^\theta \leq \sum_{\nu \in \Lambda} (\#\Lambda)^\theta = (\#\Lambda)^{\theta+1}, \tag{12.25}$$

where in the fourth inequality, we have used the fact that $\mathscr{B}_\nu \subset \Lambda$ for any $\nu \in \Lambda$ because Λ is downward closed.

The construction of sequences with algebraic growth of the Lebesgue constant is then essential. In all the following, without loss of generality, we consider the interval $X = [-1,1]$, for which the classical choices of Chebyshev and Gauss-Lobatto points give univariate Lebesgue constants that grow logarithmically, hence polynomially, with k. However, these choices are of no use for our purposes since they do not correspond to the sections of a single sequence Z.

A possible alternative is provided by the so-called Leja sequences $A := (a_j)_{j \geq 0}$ constructed according to $a_0 \in [-1,1]$ arbitrary and a_k satisfying

$$|a_k - a_0| \ldots |a_k - a_{k-1}| = \max_{t \in [-1,1]} |t - a_0| \ldots |t - a_{k-1}|. \tag{12.26}$$

Numerical evidence shows that such sequences have moderate growth of the Lebesgue constant, the bound $\mathbb{L}_k \leq (k+1)$ seems valid, see [4]. However, no rigorous proof supports this evidence. It is only known that the growth of the Lebesgue constants is sub-exponential, i.e., $(\mathbb{L}_k)^{\frac{1}{k}} \to_{k \to \infty} 0$, see [14]. In the rest of this chapter, we provide estimates on the growth of Lebesgue constants for slightly different sequences, namely Leja points for the complex unit disk and their projections on the interval $[-1,1]$.

Remark 1. In the remainder of the chapter, we work with sections of length k, more precisely given a sequence $Z = (z_j)_{j \geq 0}$ of pairwise distinct points in X, we study the

growth of the Lebesgue constant of the k-section $Z_k := (z_0, \ldots, z_{k-1})$. In order to avoid confusion with the previous notations which deal rather with $\{z_0, \ldots, z_k\}$, we denote, when needed, by I_{Z_k} the interpolation operator associated with Z_k, by λ_{Z_k} the Lebesgue function associated with Z_k and by $\mathbb{L}_{Z_k}$ the Lebesgue constant associated with Z_k.

12.3 Leja sequences and their projections

12.3.1 Leja sequence on the unit disk

Recently, Calvi and Phung [2, 3] have shown that the Lebesgue constants of Leja sequences on the unit disk and their real projections on $[-1, 1]$, the so-called $\Re$-Leja sequences, are moderate and have growths that are bounded asymptotically in $\mathscr{O}(k \log k)$ and $\mathscr{O}(k^3 \log k)$, respectively. In addition, unlike Leja sequences on $[-1, 1]$, theses sequences are easy to construct and have explicit formulas. In [4], their bounds were improved to $2k$ and $5k^2 \log k$, respectively. In this chapter, we improve further these bounds and give direct bounds for the norms $\mathbb{D}_k$ of the difference operators, which are useful in view of the discussion in the previous section. Our techniques of proof share several common points with those developed in [2–4], yet they are shorter and exploit to a considerable extent the properties of Leja sequences on the unit disk.

We introduce the notations $\mathscr{U}$ and $\partial\mathscr{U}$ for the closed complex unit disk and the complex unit circle, respectively, and the notation $\mathscr{U}_N$ for the set of N-root of unity. Given an infinite sequence $Z := (z_j)_{j\geq 0}$, we introduce the notation

$$Z_k := (z_0, \cdots, z_{k-1}) \quad \text{and} \quad Z_{l,m} := (z_l, \cdots, z_{m-1}), \quad l \leq m-1. \tag{12.27}$$

Given two finite sequence S_1 and S_2, we denote by $S_1 \wedge S_2$ the concatenation of S_1 and S_2. For any finite set $S = (s_0, \cdots, s_l)$ of complex numbers, we introduce the notation

$$\rho S := (\rho s_0, \cdots, \rho s_l), \ \rho \in \mathbb{C}, \quad \Re(S) := (\Re(s_0), \cdots, \Re(s_l)), \quad \overline{S} := (\overline{s_0}, \cdots, \overline{s_l}). \tag{12.28}$$

Throughout this chapter, to any finite set S of numbers, we associate the polynomial

$$w_S(x) := \prod_{s\in S}(x - s). \tag{12.29}$$

Any integer $k \geq 1$ can be uniquely expanded according to

$$k = \sum_{j=0}^{n} a_j 2^j, \quad a_j \in \{0, 1\} \tag{12.30}$$

We denote, respectively, by $\sigma_1(k)$, $\sigma_0(k)$, and $p_0(k)$ the number of ones in the binary expansion of k, the number of zeros in the binary expansion of k, and the largest integer p such that 2^p divide k. For $k = 2^n, \dots, 2^{n+1} - 1$ with binary expansion as above, one has

$$\sigma_1(k) = \sum_{j=0}^{n} a_j, \quad \sigma_0(k) = \sum_{j=0}^{n} (1 - a_j) = n + 1 - \sigma_1(k), \quad p_0(k) = \inf\{j = 0, \dots, n : a_j \neq 0\}. \tag{12.31}$$

We recall also that for any $n \geq 1$ and any $0 < l < 2^n$, one has

$$\sigma_1(l) + \sigma_1(2^n - l) = n + 1 - p_0(l). \tag{12.32}$$

The proof is simple and can be found in [4].

Leja sequences $E = (e_j)_{j \geq 0}$ on $\mathscr{U}$ considered in [2, 4] have all their initial value $e_0 \in \partial\mathscr{U}$ the unit circle. In view of definition (12.26), the maximum principle implies $e_j \in \partial\mathscr{U}$ for any $j \geq 1$. The sequence considered in [2] are actually Leja sequences on the unit circle.

A Leja sequence on the unit circle $E = (e_j)_{j \geq 0}$ is defined inductively by pick $e_0 \in \partial\mathscr{U}$ arbitrary and for $k \geq 1$

$$e_k = \operatorname{argmax}_{z \in \partial\mathscr{U}} |z - e_{k-1}| \dots |z - e_0|. \tag{12.33}$$

The previous argmax problem might admit many solutions and e_k is one of them. We call a k-Leja section every finite sequence $(a_0, \dots, a_{k-1})$ obtained by the same recursive procedure. In particular, when $E := (e_j)_{j \geq 1}$ is a Leja sequence then the section $E_k = (e_0, \dots, e_{k-1})$ is k-Leja section.

In contrast to the interval $[-1, 1]$ where even the first points of a Leja sequence cannot be computed explicitly, Leja sequences on $\partial\mathscr{U}$ are much easier to compute. For instance, suppose that $e_0 = 1$, then we can immediately check that $e_1 = -1$ and $e_2 = \pm i$. Assume that $e_2 = i$ then e_3 maximizes $|z^2 - 1||z - i|$, so that $e_3 = -i$ because $-i$ maximizes jointly $|z^2 - 1|$ and $|z - i|$. Then e_4 must maximize $|z^4 - 1|$, etc. We observe that a "binary" pattern on the distribution of E begins to appear.

Since the elements of $\partial\mathscr{U}$ have all the same modulus 1, an arbitrary Leja sequence $E = (e_0, e_1, \dots)$ on $\partial\mathscr{U}$ is merely the product, i.e. geometric rotation, by e_0 of a Leja sequence with initial value 1. The latter are completely determined according to the following theorem, see [1, 2, 4].

Theorem 2. *Let $n \geq 0$, $2^n < k \leq 2^{n+1}$, $l = k - 2^n$, and $e_0 = 1$. The finite sequence $E_k = (e_0, \dots, e_{k-1})$ is a k-Leja section if and only if $E_{2^n} = (e_0, \dots, e_{2^n-1})$ and $U_l = (e_{2^n}, \dots, e_{k-1})$ are, respectively, 2^n-Leja and l-Leja sections and e_{2^n} is any 2^n-root of -1.*

The most natural construction of a Leja sequence in ∂U consists then in defining $E := (e_j)_{j \geq 0}$ inductively by

$$E_1 := (e_0 = 1) \quad \text{and} \quad E_{2^{n+1}} := E_{2^n} \wedge e^{\frac{i\pi}{2^n}} E_{2^n}, \quad n \geq 0. \tag{12.34}$$

This "uniform" construction of the sequence E yields an interesting distribution of its elements. Indeed, by an immediate induction, see [1], it can be shown that the elements e_k are given by

$$e_k = \exp\left(i\pi \sum_{l=0}^{n} a_j 2^{-j}\right) \quad \text{for} \quad k = \sum_{j=0}^{n} a_j 2^j, \quad a_j \in \{0, 1\}. \tag{12.35}$$

The construction yields then a low-discrepancy sequence on $\partial\mathscr{U}$ based on the bit-reversal Van der Corput enumeration. This sequence was known to be a Leja sequence over $\partial\mathscr{U}$ in many earlier works.

As stated above, Theorem 2 characterizes completely Leja sequences on the unit circle. It has many implications that turn out to be very useful in the analysis of the growth of Lebesgue constants. We have

Theorem 3. *Let $E := (e_j)_{j\geq 0}$ be a Leja sequence on $\mathscr{U}$ starting at $e_0 = 1$. We have:*

- *For any $n \geq 0$, $E_{2^n} = \mathscr{U}_{2^n}$ in the set sense.*
- *For any $k \geq 1$, $|w_{E_k}(e_k)| = \sup_{z\in\partial\mathscr{U}} |w_{E_k}(z)| = 2^{\sigma_1(k)}$.*
- *For any $n \geq 0$, $E_{2^n,2^{n+1}} := (e_{2^n}, \cdots, e_{2^{n+1}-1})$ is a 2^n-Leja section.*
- *For any $n \geq 0$, $\mathscr{B}(E_{2^n}) := (e_{2^n-1}, \cdots, e_1, e_0)$ is a 2^n-Leja section.*
- *The sequence $E^2 := (e_{2j}^2)_{j\geq 0}$ is a Leja sequence on $\partial\mathscr{U}$ starting at 1.*

The proof of these properties can be found in [2, 4, 5].

Using the implications of the Leja definition (12.33) on the growth of the Lebesgue constants $\mathbb{L}_{E_k}$ and the previous structural properties of Leja sequences on the unit disk, it was proved in [4] that for any Leja sequence E on $\partial\mathscr{U}$, we have

$$\lambda_{E_k}(e_k) \leq k \quad \text{and} \quad \mathbb{L}_{E_k} \leq 2k, \quad k \geq 1. \tag{12.36}$$

For further use, let us note that given E a Leja section starting at $\rho \in \partial U$, $n \geq 1$, and k such that $1 \leq k \leq 2^n$, one has for any $z, \xi \in \partial\mathscr{U}$ with $\xi \notin E_k$

$$\frac{|w_{E_k}(z)|}{|w_{E_k}(\xi)|} = \frac{|w_{E_k}(z)||w_{\mathscr{B}(E_{k,2^n})}(\xi)|}{|w_{E_{2^n}}(\xi)|} \leq \frac{2^{\sigma_1(k)}2^{\sigma_1(2^n-k)}}{|\xi^{2^n} - \rho^{2^n}|} = \frac{2^{n+1-p_0(k)}}{|\xi^{2^n} - \rho^{2^n}|}. \tag{12.37}$$

We have used that $E_k \cup \mathscr{B}(E_{k,2^n}) = E_k \cup E_{k,2^n} = E_{2^n} = \rho\mathscr{U}_{2^n}$ in the set sense, that $\mathscr{B}(E_{k,2^n})$ is a $\{2^n - k\}$-Leja section according to the forth property above, and the identity (12.32).

12.3.2 $\mathfrak{R}$-Leja sequences on $[-1, 1]$

We consider a Leja sequence $E = (e_j)_{j\geq 0}$ on the unit circle with $e_0 = 1$ and project it onto the real interval $[-1, 1]$ and denote by $R = (r_j)_{j\geq 0}$ the sequence obtained. Since $E = (1, -1, \alpha, -\alpha, \cdots)$ with $\alpha = \pm i$, one should make sure that no point is repeated

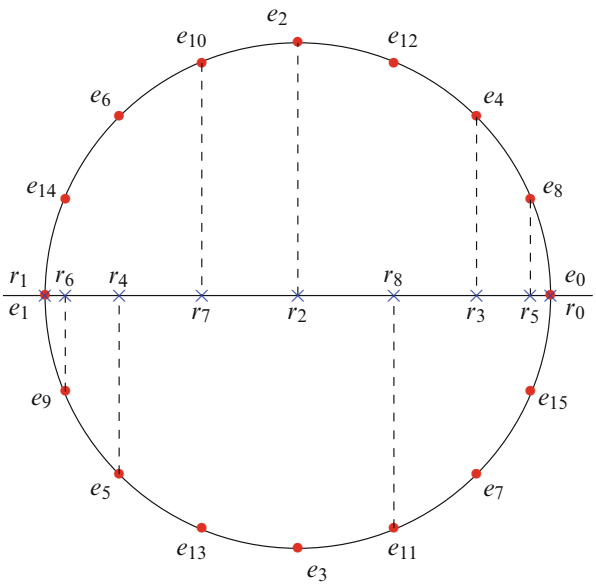


Fig. 12.1 Distribution of the first elements of the Leja sequence E defined in (12.35) and R the associated $\Re$-Leja sequence.

on R simply by not projecting a point e_j such that $e_j = \overline{e}_i$ for some $i < j$. Such sequences R were named $\Re$-Leja sequences in [3]. The projection rule that prevents the repetition is well understood, it was already explained in [3, Theorem 2.4] and we also provide below in Theorem 4. We explicit in Fig. 12.1 this rule for the first elements of the Leja sequence E defined in (12.35) and the associated $\Re$-Leja sequence. First let us observe that given E a Leja sequence starting at $e_0 = 1$, we have $e_1 = 1$, $e_2 = -e_3 = \pm i$, and $e_{2j} = -e_{2j+1}$ for any $j \geq 2$, therefore the associated $\Re$-Leja sequence R satisfies

$$r_{2j-1} = -r_{2j}, \qquad j \geq 2 \tag{12.38}$$

Theorem 4. *Let E be a Leja sequence on $\partial\mathscr{U}$ with $e_0 = 1$ and R the associated $\Re$-Leja sequence. Then*

$$R = \Re(\Xi), \quad \text{with} \quad \Xi := (\xi_j)_{j\geq 0} = (1,-1) \wedge \bigwedge_{j=1}^{\infty} E_{2^j,2^j+2^{j-1}}. \tag{12.39}$$

The previous theorem says essentially that the section $E_{2^n,2^{n+1}}$, considered as a set, is the union of the first half $E_{2^n,2^n+2^{n-1}}$ and of its element-wise conjugate $\overline{E_{2^n,2^n+2^{n-1}}}$ defined as in (12.28). We have $r_0 = 1$, $r_1 = -1$. In addition, for $n \geq 0$ and k such that $2^n \leq k-1 < 2^{n+1}$, using the simple identity $k = 2 + \sum_{j=1}^{n} 2^{j-1} + (k-1-2^n)$, we deduce that $R_k = \Re(\Xi_k)$ and the following element $r_k = \Re(\xi_k)$ are obtained from

$$\Xi_k = (1,-1) \wedge \bigwedge_{j=1}^{n} E_{2^j,2^j+2^{j-1}} \wedge E_{2^{n+1},2^n+k-1} \quad \text{and} \quad \xi_k = e_{2^n+k-1}. \tag{12.40}$$

The particular structure of the Leja sequences E yields useful properties for $\Re$-Leja sequences. First, in view of the first property in Theorem 3, since $E_{2^{n+1}} = \mathscr{U}_{2^{n+1}}$ in the set sense, the projection on $[-1,1]$ gives

$$R_{2^n+1} = \left\{ \cos\left(\frac{j\pi}{2^n}\right) : j = 0, \ldots, 2^n \right\}, \qquad n \geq 0 \tag{12.41}$$

in the set sense. Therefore R_{2^n+1} coincides as a set with the Gauss-Lobatto abscissas. We have also the following result.

Lemma 1. *Let $R := (r_j)_{j\geq 0}$ be an $\Re$-Leja sequence. The sequence*

$$R^2 := (2r_{2j}^2 - 1)_{j\geq 0} \tag{12.42}$$

is also an $\Re$-Leja sequence.

Proof. We consider $E = (e_j)_{j\geq 0}$ to be a Leja sequence associated with R and recall that by Theorem 3, the sequence $E^2 = (e_{2j}^2)_{j\geq 0}$ is also Leja sequence starting at 1 since $e_0 = 1$. The sequence R^2 can be obtained by projection of E^2 onto $[-1,1]$. Indeed, the first two elements of R^2 are 1 and -1 because $r_0 = 1$, $r_2 = 0$, so that we only need to show that (12.40) holds with R^2 and E^2. For $n \geq 0$ and $2^n \leq k-1 < 2^{n+1}$, one has $2^{n+1} \leq (2k-1)-1 < 2^{n+2}$ so that by the second equality in (12.40),

$$r_{2k-1} = \Re(e_{2^{n+1}+2k-1-1}) = \Re(e_{2(2^n+k-1)}).$$

Since $2k \geq 4$ then according to (12.38), we have $r_{2k} = -r_{2k-1}$, hence

$$2r_{2k}^2 - 1 = 2r_{2k-1}^2 - 1 = \Re(e_{2(2^n+k-1)}^2),$$

where we have used $\Re(z^2) = 2\Re(z)^2 - 1$ for $z \in \partial\mathscr{U}$. The proof is then complete. □

We should note that the notation R^2 in the previous lemma does not match the notation E^2 given in Theorem 3 for Leja sequences. It is however natural and convenient in the sense that if R is an $\Re$-Leja sequence associated with E then R^2 is also an $\Re$-Leja sequence, yet associated with E^2.

The previous lemma has certain implications on the polynomials w_{R_k} associated with the sections R_k which are essential to the study of the norms of the difference operators discussed in section §12.5. In order to clarify our notation, we find it convenient to work with normalized versions of these polynomials w_{R_k} that we define by

$$W_{R_k}(x) := 2^k w_{R_k}(x), \quad x \in [-1,1]. \tag{12.43}$$

We are interested in the relation between these polynomials for sections of the sequences R and R^2. First, since all $\Re$-Leja sequences have initial elements 1 and -1, it is immediate that

$$W_{R_1^2}(2x^2-1) = W_{R_2}(x) \quad x \in [-1,1]. \tag{12.44}$$

For higher value of k, we have the following

Lemma 2. *Let R be an $\Re$-Leja sequence and denote $S := R^2$. For any $k \geq 2$*

$$W_{S_k}(2x^2-1) = 2x\, W_{R_{2k-1}}(x), \quad x \in [-1,1]. \tag{12.45}$$

Consequently $W'_{S_k}(-1) = W'_{R_{2k-1}}(0)$, $W'_{S_k}(1) = \frac{1}{2}W'_{R_{2k-1}}(1) = \frac{1}{2}W'_{R_{2k-1}}(-1)$, and

$$W'_{S_k}(s_j) = \frac{1}{2}W'_{R_{2k-1}}(r_{2j}) = \frac{1}{2}W'_{R_{2k-1}}(r_{2j-1}), \;\; j = 2,\dots,k-1 \tag{12.46}$$

Proof. The verification of (12.45) for $k = 2$ is immediate. Now, from the definition of R^2, we have for $k \geq 3$,

$$w_{S_k}(2x^2-1) = \prod_{j=0}^{k-1}\Big(2x^2-1-(2r_{2j}^2-1)\Big) = 2^k\prod_{j=0}^{k-1}(x+r_{2j})(x-r_{2j}).$$

Since $r_0 = 1,\ r_1 = -1,\ r_2 = 0$, and $r_{2j} = -r_{2j-1}$ for any $j \geq 2$,

$$w_{S_k}(2x^2-1) = 2^k(x+1)(x-1)x^2\prod_{j=2}^{k-1}(x-r_{2j-1})(x-r_{2j}) = 2^k x\, w_{R_{2k-1}}(x),$$

which implies (12.45) after multiplication by 2^k. The derivation with respect to x gives

$$4x\, W'_{S_k}(2x^2-1) = 2\Big(x\, W'_{R_{2k-1}}(x) + W_{R_{2k-1}}(x)\Big).$$

Since $W_{R_{2k-1}}(0) = 0$, the first result on derivatives is obtained when dividing by x and letting $x \to 0$. The second result is obtained by the substitution of x by 1 or -1. In order to obtain (12.46), we substitute x by r_{2j} and $r_{2j-1} = -r_{2j}$ for $j = 2\dots,k-1$. □

The previous Lemma has also implications on the growth of $W_{R_k}(r_k)$, which we will use in §12.4.

Lemma 3. *Let R be an $\Re$-Leja sequence and denote $S := R^2$. For any $N \geq 1$, we have $W_{R_2}(r_2) = W_{R_1}(r_1) = 4$ and*

$$2r_k\, W_{R_k}(r_k) = W_{S_{N+1}}(s_{N+1}), \quad k = 2N+1, \;\; N \geq 1, \tag{12.47}$$

and

$$W_{R_k}(r_k) = 2W_{S_N}(s_N), \quad k = 2N, \ \ N \geq 2. \tag{12.48}$$

Proof. The first equality follows from (12.45) applied with $x = r_k$ since $k = 2(N+1) - 1$ and $2r_k^2 - 1 = 2r_{2(N+1)}^2 - 1 = s_{N+1}$. The second equality can be checked easily for $N = 1$. For $N \geq 2$, using the fact $r_k = -r_{2N-1}$ and $s_N = 2r_k^2 - 1$, formula (12.45) implies

$$W_{R_k}(r_k) = 2(r_k - r_{2N-1})W_{R_{2N-1}}(r_k) = 4r_k W_{R_{2N-1}}(r_k) = 2W_{S_N}(s_N).$$

□

12.4 Growth of Lebesgue constants for $\Re$-Leja sections

As stated above in (12.41), for any $\Re$-Leja sequence R, the sections R_{2^n+1} coincide in the set sense with the Gauss-Lobatto abscissas. This type of abscissas is known to have Lebesgue constant with logarithmic growth $\mathbb{L}_{R_{2^n+1}} \sim \frac{2}{\pi}\log(2^n+1)$. More precisely, we have the bound

$$\mathbb{L}_{R_{2^n+1}} \leq 1 + \frac{2}{\pi}\log(2^n). \tag{12.49}$$

See [12, Formulas 5 and 13]. In [4], using the previous bound and classical trigonometric arguments as the one used in the bounding of Lebesgue constant of Tchybeshev abscissas, e.g., [11], it is established that for any $n \geq 0$ and any k such that $2^n + 1 \leq k < 2^{n+1} + 1$

$$\mathbb{L}_{R_k} \leq 4^{n-p_0(k')}(5 + \frac{8}{\pi}\log 2^n) \quad \text{where} \quad k' := k - (2^n+1). \tag{12.50}$$

Although the effect of the binary pattern on the distribution of the Leja sequence E on $\partial\mathscr{U}$ is somehow reflected by the term $2^{n-p_0(k')}$, we observe that if k is an even number, we only have the bound $\mathbb{L}_{R_k} \leq \frac{8}{\pi}k^2\log k$.

Through a novel analysis, we propose to relate the analysis of the Lebesgue constants $\mathbb{L}_{R_k}$ to the analysis of the Lebesgue constants $\mathbb{L}_{E_k}$ where E is any Leja sequence associated with R, which allows us to improve the bound on $\mathbb{L}_{R_k}$.

The sections R_k of length $k = 2^n + 1$ having been already treated, see (12.49), we only discuss the case of k such that $2^n + 1 < k < 2^{n+1} + 1$. For such values we have $R_k = \Re(\Xi_k)$ where Ξ_k is the section obtained from E_{2^n+k-1} by the elimination procedure described in (12.40). Observe that E_{2^n+k-1} is the shortest section of E that yields R_k when projected onto $[-1, 1]$. We have the following result

Theorem 5. *Let $n \geq 0$ and $k \geq 3$ such that $2^n + 1 < k < 2^{n+1} + 1$. One has*

$$\mathbb{L}_{R_k} \leq 2^{n+\frac{3}{2}-p_0(k')}\mathbb{L}_{E_{2^n+k-1}} \quad \text{where} \quad k' := k - (2^n + 1). \tag{12.51}$$

In view of (12.36), the previous theorem implies in particular

$$\mathbb{L}_{R_k} \leq 2^{n+\frac{3}{2}} \times 2(2^n + k - 1) \leq 8\sqrt{2}k^2. \tag{12.52}$$

In order to prove the theorem, we must bound the Lebesgue function associated with the real section R_k using the Lebesgue function or constant associated with the complex section E_{k+2^n-1}. To this end, we propose to bound the Lagrange polynomials associated with R_k using those associated with E_{k+2^n-1}. For notational simplicity, we introduce

$$G_k = E_{2^n+k-1}, \qquad 2^n + 1 < k < 2^{n+1} + 1, \tag{12.53}$$

where G_k is considered as a set. The section Ξ_k is obtained from E_{2^n+k-1} by the elimination procedure described in (12.40). The following lemma describes how G_k can be obtained from Ξ_k.

Lemma 4. *Let E be a Leja sequence with $e_0 = 1$ and $\Xi = (\xi_j)_{j\geq 0}$ the sequence defined in Theorem 4. For any $n \geq 0$ and any k with $2^n + 1 < k < 2^{n+1} + 1$, we have*

$$G_k = \{\xi_0, \xi_1\} \cup \{\xi_2, \overline{\xi_2}, \cdots, \xi_{2^n}, \overline{\xi_{2^n}}\} \cup F_k \qquad F_k := \Xi_{2^n+1,k} = \{\xi_{2^n+1}, \cdots, \xi_{k-1}\}. \tag{12.54}$$

Proof. We have that

$$G_k = E_{2^n+k-1} = E_{2^{n+1}} \wedge E_{2^{n+1},2^n+k-1} = E_{2^{n+1}} \wedge \Xi_{2^n+1,k}.$$

Therefore, we only need to show that $E_{2^{n+1}} = \{\xi_0, \xi_1, \xi_2, \overline{\xi_2}, \cdots, \xi_{2^n}, \overline{\xi_{2^n}}\}$ in the set sense. Since $E_{2^{n+1}}$ coincides with the set of 2^{n+1}-root of unity, E_{2^n+1} is the union of $\{1, -1\}$ and $\{\xi_2, \ldots, \xi_{2^n}\}$ and their conjugates, which finishes the proof. □

The previous lemma allows us to relate the polynomials W_{R_k} defined in (12.43) and the polynomials w_{G_k} and their derivatives.

Lemma 5. *Let E be a Leja sequence on $\mathscr{U}$ with $e_0 = 1$ and Ξ and R the associated sequence as in Theorem 4. Let n, k, F_k, and G_k as in the previous lemma. For any $z \in \partial\mathscr{U}$ and $x = \Re(z)$*

$$|W_{R_k}(x)| = |z^2 - 1||w_{G_k}(z)||w_{\overline{F_k}}(z)| = |\overline{z}^2 - 1||w_{G_k}(\overline{z})||w_{\overline{F_k}}(\overline{z})|. \tag{12.55}$$

Consequently, for any $j = 0, \cdots, k-1$

$$|W'_{R_k}(r_j)| = 2\alpha_j|w'_{G_k}(\xi_j)||w_{\overline{F_k}}(\xi_j)|, \quad S \tag{12.56}$$

where $\alpha_j = 1$ for every j except for $j = 0$ and $j = 1$ for which it is equal to 2.

Proof. Given $z, z' \in \partial\mathscr{U}$, $x = \frac{1}{2}(z+\overline{z})$, and $x' = \frac{1}{2}(z'+\overline{z}')$, one has

$$2|x-x'| = 2\left|\frac{z+z^{-1}}{2} - \frac{z'+z'^{-1}}{2}\right| = \left|z-z' + \frac{1}{z} - \frac{1}{z'}\right| = |z-z'|\left|1-\frac{1}{zz'}\right| = |z-z'||z-\overline{z'}|. \tag{12.57}$$

Since $r_j = \Re(\xi_j)$ and $\xi_j \in \partial\mathscr{U}$ for any $j \geq 0$,

$$|W_{R_k}(x)| = \prod_{j=0}^{k-1} 2|x-r_j| = \prod_{j=0}^{k-1} |z-\xi_j| \prod_{j=0}^{k-1} |z-\overline{\xi_j}|.$$

In view of (12.54), taking into account that $\xi_0 = 1$ and $\xi_1 = -1$ are repeated twice in the previous product, the first part in (12.55) follows. The second part is immediate since z and $\overline{z}$ play symmetric roles. This result combined with identity (12.57) shows that for every $j = 1, \cdots, k-1$

$$|W'_{R_k}(r_j)| = \lim_{x\to r_j} \frac{|W_{R_k}(x)|}{|x-r_j|} = \lim_{z\to\xi_j} \frac{|z^2-1||w_{G_k}(z)||w_{\overline{F_k}}(z)|}{\frac{1}{2}|z-\xi_j||z-\overline{\xi_j}|},$$

where the limit $\lim_{z\to\xi_j}$ is meant in the circle $\partial\mathscr{U}$. The second result follows then from the fact that $\lim_{z\to\xi} |z^2-1|/|z-\overline{\xi}|$ is equal to 1 for every $\xi \in \partial\mathscr{U}$, except for $\xi = 1$ and $\xi = -1$ for which it is equal to 2. □

In view of the above, we are now able to relate the Lagrange polynomials associated with the sections R_k and those associated with the set G_k, hence the Lebesgue functions associated with R_k and G_k. First, we introduce the quotient notation

$$q_k(z,\xi) := \frac{|w_{\overline{F_k}}(z)|}{|w_{\overline{F_k}}(\xi)|}, \quad z \in \partial\mathscr{U}, \quad \xi \in \partial\mathscr{U} \setminus F_k. \tag{12.58}$$

Lemma 6. *We have*

$$\mathbb{L}_{R_k} \leq 2\mathbb{L}_{G_k} \sup_{\substack{z\in\partial\mathscr{U} \\ \xi\in G_k}} q_k(z,\xi). \tag{12.59}$$

Proof. We denote by $l_0, \ldots, l_{k-1}$ the Lagrange polynomials associated with the section R_k and by $L_0, L_1, L_{(2,1)}, L_{(2,2)}, \cdots, L_{(2^n,1)}, L_{(2^n,2)}, L_{2^n+1}, \cdots, L_{k-1}$, the Lagrange polynomials associated with the set G_k following the order given in (12.54). For convenience, we write the first polynomials as

$$l_j(x) := \frac{W_{R_k}(x)}{W'_{R_k}(r_j)(x-r_j)}, \qquad x \in [-1,1].$$

In view of Lemma 5 and identity (12.57), we have for $j = 0, \ldots, k-1$, $z \in \partial\mathscr{U}$ and $x = \Re(z)$

$$|l_j(x)| = \frac{1}{\alpha_j}\Big|\frac{z^2-1}{(z-\xi_j)(z-\overline{\xi_j})}\Big|\frac{|w_{G_k}(z)|}{|w'_{G_k}(\xi_j)|}\frac{|w_{\overline{F_k}}(z)|}{|w_{\overline{F_k}}(\xi_j)|} \tag{12.60}$$

where α_j are defined as in Lemma 5. We observe that

$$\Big|\frac{z^2-1}{(z-\xi)(z-\overline{\xi})}\Big| = \Big|\frac{z-\overline{z}}{(z-\xi)(\overline{z}-\xi)}\Big| = \Big|\frac{z-\xi+\xi-\overline{z}}{(z-\xi)(\overline{z}-\xi)}\Big| \le \frac{1}{|z-\xi|}+\frac{1}{|\overline{z}-\xi|}. \tag{12.61}$$

The last inequality applied with the real values $\xi = \xi_0 = 1$ and $\xi = \xi_1 = -1$ and injected in (12.60), with $\alpha_0 = \alpha_1$=2, yields

$$|l_0(x)| \le q_k(z,\xi_0)|L_0(z)| \quad \text{and} \quad |l_1(x)| \le q_k(z,\xi_1)|L_1(z)|. \tag{12.62}$$

Now for the indices $j = 2, \ldots, 2^n$, since ξ_j and $\overline{\xi_j}$ play symmetric roles in the sense $\Re(\xi_j) = \Re(\overline{\xi_j}) = r_j$ and $\xi_j, \overline{\xi_j} \in G_k$, one observes that (12.56) yields

$$|w'_{G_k}(\xi_j)||w_{\overline{F_k}}(\xi_j)| = \frac{1}{2}|W'_{R_k}(r_j)| = |w'_{G_k}(\overline{\xi_j})||w_{\overline{F_k}}(\overline{\xi_j})|.$$

Taking this equality into account when injecting (12.61) into (12.60) and the fact that $\alpha_j = 1$, we deduce

$$|l_j(x)| \le q_k(z,\xi_j)L_{(j,1)}(z) + q_k(z,\overline{\xi_j})L_{(j,2)}(z). \tag{12.63}$$

Finally for the indices $j = 2^n+1, \ldots, k-1$, taking account of $|z-\overline{\xi}| = |\overline{z}-\xi|$ and the identity (12.55), which shows that $|w_{G_k}(z)||w_{\overline{F_k}}(z)| = |w_{G_k}(\overline{z})||w_{\overline{F_k}}(\overline{z})|$, when injecting (12.61) into (12.60), we obtain

$$|l_j(x)| \le q_k(z,\xi_j)L_j(z) + q_k(\overline{z},\xi_j)L_j(\overline{z}). \tag{12.64}$$

Summing the inequalities (12.62), (12.63), and (12.64), we conclude the proof. □

In view of the previous lemma, we can derive Theorem 5 through a study of the growth of the quotient function q_k. By the main structure of Leja sequences on $\mathscr{U}$ described by Theorem 2, we have that $F_k = E_{2^{n+1},2^n+k-1}$ is a k'-Leja section with $k' = k-(2^n+1)$ and $0 < k' < 2^n$, therefore by (12.37), we derive

$$q_k(z,\xi) = \frac{|w_{F_k}(\overline{z})|}{|w_{F_k}(\overline{\xi})|} \le \frac{2^{n+1-p_0(k')}}{|\overline{\xi}^{2n} - e^{2^n}_{2^{n+1}}|}.$$

Since $e_{2^{n+1}}$ is a 2^{n+1}-root of -1, $(e_{2^{n+1}})^{2^n} = \pm i$. As for $\xi \in G_k$, since $G_k \subset E_{2^{n+2}} = \mathscr{U}_{2^{n+2}}$, $\xi^{2^n} \in \{1, -1, i, -i\}$. This shows that necessarily $|\overline{\xi}^{2^n} - e_{2^{n+1}}^{2^n}| \geq \sqrt{2}$, so that

$$\sup_{\substack{z \in \partial \mathscr{U} \\ \xi \in G_k}} q_k(z, \xi) \leq 2^{n+\frac{1}{2}-p_0(k')}. \tag{12.65}$$

This bound injected in (12.59) completes the proof of Theorem 5.

12.5 Growth of the norms of the difference operators

In this section, we focus our attention on the difference operators

$$\Delta_0 = I_0, \quad \text{and} \quad \Delta_k = I_k - I_{k-1}, \quad k \geq 1. \tag{12.66}$$

associated with interpolation on Leja sequences on $\partial \mathscr{U}$ and $\Re$-Leja sequences on $[-1, 1]$. We are interested in estimating their norm

$$\mathbb{D}_k := \sup_{f \in C(X) - \{0\}} \frac{\|\Delta_k f\|_{L^\infty(X)}}{\|f\|_{L^\infty(X)}}. \tag{12.67}$$

We write $\mathbb{D}_k(Z)$ when needed to emphasize the dependence on the sequence Z. It is immediate that $\mathbb{D}_0 = \mathbb{L}_0 = 1$ and $\mathbb{D}_k \leq \mathbb{L}_k + \mathbb{L}_{k-1}$ any for $k \geq 1$. We shall sharpen the previous bound when Z has a particular structure, for instance, if Z is a Leja or an $\Re$-Leja sequence.

We recall that $I_k = I_{Z_{k+1}}$. Similar to the expression of Lebesgue constant in (12.19), we can express $\mathbb{D}_k$ using Lagrange polynomials. Indeed, using Lagrange interpolation formula in $\{z_0, \ldots, z_k\}$, it can be easily checked that for any $k \geq 1$

$$\Delta_k f(z) = \Big(f(z_k) - I_{Z_k} f(z_k)\Big) \frac{w_{Z_k}(z)}{w_{Z_k}(z_k)}, \quad z \in X. \tag{12.68}$$

This implies that

$$\mathbb{D}_k(Z) = \sup_{z \in X} \frac{|w_{Z_k}(z)|}{|w_{Z_k}(z_k)|} \sup_{f \in C(X) - \{0\}} \frac{|f(z_k) - I_{Z_k} f(z_k)|}{\|f\|_{L^\infty(X)}}. \tag{12.69}$$

The second supremum in the previous equality is obviously bounded by $1 + \lambda_{Z_k}(z_k)$, where λ_{Z_k} is the Lebesgue function as defined in §12.2, see Remark 1. This bound is actually attained: to see this, take f a function in $C(X)$ having a maximum value equal to 1 and satisfying $f(z_k) = -1$ and $f(z_j) = \frac{|l_j(z_k)|}{l_j(z_k)}$ for every $j = 0, \ldots, k-1$ where $l_0, \ldots, l_{k-1}$ are the Lagrange polynomials associated with Z_k. Therefore

$$\mathbb{D}_k(Z) = \Big(1 + \lambda_{Z_k}(z_k)\Big) \sup_{z \in X} \frac{|w_{Z_k}(z)|}{|w_{Z_k}(z_k)|}. \tag{12.70}$$

The previous formula shows in particular that if Z is a Leja sequence on X, then

$$\mathbb{D}_k(Z) = 1 + \lambda_{Z_k}(z_k). \tag{12.71}$$

In particular, in view of the results on Leja sequences on the unit disk, more precisely (12.36), we have

Theorem 6. *Let E be a Leja sequence in $\mathscr{U}$ with initial value $e_0 \in \partial\mathscr{U}$. The norms of the difference operators associated with E satisfy $\mathbb{D}_0 = 1$ and for $k \geq 1$*

$$\mathbb{D}_k \leq 1 + k \tag{12.72}$$

Combining this result with (12.23), we obtain the following stability estimate for the multivariate interpolation operator.

Corollary 1. *With $X = \mathscr{U}$ and E any Leja sequence with initial value $e_0 \in \partial\mathscr{U}$, one has*

$$\mathbb{L}_\Lambda \leq (\#(\Lambda))^2, \tag{12.73}$$

for any downward closed set Λ.

Formula (12.70) is convenient in the case of Leja sequences since it yields exact values of the quantities $\mathbb{D}_k$. In the case of $\Re$-Leja sequences, we opt for a different expression of (12.70). From the formulas of Lagrange polynomials associated with Z_k, we may write (12.70) as

$$\mathbb{D}_k = \Big(\frac{1}{|w_{Z_k}(z_k)|} + \sum_{j=0}^{k-1} \frac{1}{|w'_{Z_k}(z_j)||z_k - z_j|}\Big) \sup_{z\in X} |w_{Z_k}(z)|. \tag{12.74}$$

We remark that $|w_{Z_k}(z_k)| = |w'_{Z_{k+1}}(z_k)|$ and $|w'_{Z_k}(z_j)||z_k - z_j| = |w'_{Z_{k+1}}(z_j)|$ for any $j = 0, \ldots, k-1$, we may then rewrite (12.70) in a more compact form

$$\mathbb{D}_k = \Big(\sum_{j=0}^{k} \frac{1}{|w'_{Z_{k+1}}(z_j)|}\Big) \sup_{z\in X} |w_{Z_k}(z)| \tag{12.75}$$

Now, we let $R = (r_j)_{j\geq 0}$ be an $\Re$-Leja sequence. Using for this sequence the polynomials W_{R_k} defined in (12.43) instead of w_{R_k}, we may rewrite (12.75) for R as

$$\mathbb{D}_k(R) = 2\beta_k(R) \sup_{x\in[-1,1]} |W_{R_k}(x)| \qquad \text{where} \qquad \beta_k(R) := \sum_{j=0}^{k} \frac{1}{|W'_{R_{k+1}}(r_j)|}. \tag{12.76}$$

We propose to bound separately the quantities $\beta_k(R)$ and $\sup_{x\in[-1,1]} |W_{R_k}(x)|$ in this order.

Lemma 7. *Let R be a $\Re$-Leja sequence. We have $\beta_{2^n}(R) = \frac{1}{4}$ for any $n \geq 0$. More generally, for $k \neq 2^n \geq 1$,*

$$\beta_k(R) \leq \frac{2^{\sigma_0(k)-p_0(k)}}{2}, \tag{12.77}$$

where $\sigma_0(k)$ and $p_0(k)$ are defined in (12.31).

Proof. We first assume that $k = 2N \geq 4$ is an even integer. We have

$$\beta_k(R) = \frac{1}{|W'_{R_{2N+1}}(1)|} + \frac{1}{|W'_{R_{2N+1}}(-1)|} + \frac{1}{|W'_{R_{2N+1}}(0)|} + \sum_{j=2}^{N} \Big(\frac{1}{|W'_{R_{2N+1}}(r_{2j-1})|} + \frac{1}{|W'_{R_{2N+1}}(r_{2j})|} \Big). \tag{12.78}$$

We introduce the shorthand $S = R^2$ where R^2 is the sequence depending on R according to (12.42). Using Lemma 2, we deduce that

$$\beta_k(R) = \frac{1}{|W'_{S_{N+1}}(1)|} + \frac{1}{|W'_{S_{N+1}}(-1)|} + \sum_{j=2}^{N} \frac{1}{|W'_{S_{N+1}}(s_j)|} = \beta_N(S).$$

The same argument implies that $\beta_2(R) = \beta_1(S)$, so that $\beta_{2N}(R) = \beta_N(S)$ is valid for any $N \geq 1$. Since S is also an $\Re$-Leja sequence, see Lemma 1, the verification $\beta_2(S) = \beta_1(S) = \frac{1}{4}$ for any $\Re$-Leja sequence S implies the first result in the lemma $\beta_{2^n}(R) = \frac{1}{4}$ for any $n \geq 0$.

We now assume that $k = 2N + 1 \geq 3$ is an odd integer. First, we isolate the last quotient in the sum giving $\beta_k(R)$ and multiply the other quotients by $\frac{2|r_j - r_{k+1}|}{2|r_j - r_{k+1}|}$ yielding

$$\beta_k(R) = \frac{1}{W_{R_k}(r_k)} + \sum_{j=0}^{k-1} \frac{2|r_j - r_{k+1}|}{|W'_{R_{k+2}}(r_j)|}.$$

Since $k = 2(N+1) - 1$ and $k + 2 = 2(N+2) - 1$, regrouping the sum as in (12.78) and using Lemmas 2 and 3, we deduce

$$\beta_k(R) = \frac{2|r_k|}{|W_{S_{N+1}}(s_{N+1})|} + 2\frac{|1 - r_{2N+2}| + |-1 - r_{2N+2}|}{2|W'_{S_{N+2}}(1)|} + 2\frac{|r_{2N+2}|}{|W'_{S_{N+2}}(-1)|} + 2\Big(\sum_{j=2}^{N} \frac{|r_{2j-1} - r_{2N+2}| + |r_{2j} - r_{2N+2}|}{2|W'_{S_{N+2}}(s_j)|} \Big)$$

Since $|x-r|+|x+r| \leq 2$ for any $x, r \in [-1,1]$ and $r_{2j-1} = -r_{2j}$ for every $j \geq 2$, we deduce that

$$\beta_k(R) \leq \frac{2}{|W_{S_{N+1}}(s_{N+1})|} + \frac{2}{|W'_{S_{N+2}}(1)|} + \frac{2}{|W'_{S_{N+2}}(-1)|} + \sum_{j=2}^{N} \frac{2}{|W'_{S_{N+2}}(s_j)|} \leq 2\beta_{N+1}(S).$$

We introduce the sequence $(\beta_k)_{k\geq 1}$ defined by

$$\beta_k := \sup\left\{\beta_k(R) : R \ \text{ isan } \ \Re\text{−Lejasequence}\right\}, \qquad k \geq 1.$$

Since $S = R^2$ is also an $\Re$-Leja sequence, in view of the previous discussion, we have $\beta_1 = 1/4$ and

$$\beta_{2N} = \beta_N, \qquad \beta_{2N+1} \leq 2\beta_{N+1}, \qquad N \geq 1.$$

The sequence $(\beta_k)_{k\geq 1}$ is positive therefore it is bounded by the sequence $\frac{1}{4}(u_k)_{k\geq 1}$ where $(u_k)_{k\geq 1}$ is defined inductively by $u_1 = 1$ and

$$u_{2N} = u_N, \quad u_{2N+1} = 2u_{N+1}, \quad N \geq 1.$$

The sequence $(u_k)_{k\geq 1}$ is given by

$$u_{2^n} = 1, \ \ n \geq 0, \quad u_k = 2^{\sigma_0(k)-p_0(k)+1}, \ \ k \neq 2^n \geq 3.$$

Indeed, we check easily that $u_1 = 1$ and $u_{2N} = u_N$. Let now $N \geq 1$. If $N+1 = 2^n$ for some n, then $2N+1 = 2^{n+1}-1$, so that $u_{2N+1} = 2 = 2u_{N+1}$. Else if $N+1 \neq 2^n$, then by "binary" subtraction we have

$$\sigma_0(2N+1) = \sigma_0(N) = \sigma_0((N+1)-1) = \sigma_0(N+1) - p_0(N+1) + 1,$$

so that $u_{2N+1} = 2u_{N+1}$. As a consequence, the sequence $(\beta_k)_{k\geq 1}$ satisfies

$$\beta_k \leq \frac{u_k}{4} = \frac{2^{\sigma_0(k)-p_0(k)}}{2},$$

for any $k \neq 2^n$, which finishes the proof. □

Lemma 8. *Let R be an $\Re$-Leja sequence. For any $k \geq 2$,*

$$\sup_{x\in[-1,1]} |W_{R_k}(x)| \leq 4^{\sigma_1(k)+p_0(k)-1}. \tag{12.79}$$

Proof. Let $n \geq 0$, $2^n + 1 < k < 2^{n+1} + 1$, and $k' = k - (2^n + 1)$. By Lemma 5, for $z \in \partial\mathscr{U}$ and $x \in [-1,1]$, we have

$$|W_{R_k}(x)| = |z^2-1||w_{G_k}(z)||w_{\overline{F_k}}(z)| \leq 2 \times 2^{\sigma_1(2^{n+1}+k')} 2^{\sigma_1(k')} = 4^{\sigma_1(k')+1},$$

where we have used that G_k and F_k are, respectively, $\{2^{n+1}+k'\}$-Leja and k'-Leja section of the unit disk and the second point in Theorem 3. When $k = 2^n + 1$, the section R_k corresponds to the Gauss-Lobatto points as in (12.41) and the set G_k to the 2^{n+1}-roots of 1, which shows that the above estimate remains valid since $k' = 0$ and

$$|W_{R_k}(x)| = |z^2 - 1||w_{G_k}(z)| = |(z^2 - 1)(z^{2^{n+1}} - 1)| \leq 4.$$

In both cases, since $0 \leq k' < 2^n$ and $k = k' + 2^n + 1$, the number of ones in the binary expansion of k' satisfies $\sigma_1(k') + 1 = \sigma_1(k' + 2^n) = \sigma_1(k-1)$. Observe that $\sigma_1(k-1) = \sigma_1(k) - 1$ if k is odd and $\sigma_1(k-1) = p_0(k) + \sigma_1(k) - 1$ for k even, therefore $\sigma_1(k') + 1 = \sigma_1(k) + p_0(k) - 1$. We deduce that for any $k \geq 2$, we have the bound

$$|W_{R_k}(x)| \leq 4^{\sigma_1(k)+p_0(k)-1},$$

which completes the proof. □

In view of the two previous lemmas, we are now able to provide a bound on the growth of the norms of the difference operators for $\Re$-Leja sequences.

Theorem 7. *Let R be an $\Re$-Leja sequence in $[-1,1]$. The norms of the difference operators associated with R satisfy $\mathbb{D}_0 = 1$ and for $k \geq 1$*

$$\mathbb{D}_k \leq (1+k)^2 \tag{12.80}$$

Proof. For $k = 1$, we have $\beta_1(R) = 1/4$ and $W_{R_1}(x) = 2(x-1)$, therefore in view of (12.76), we get $\mathbb{D}_1(R) \leq 2$. For the values $2^n < k < 2^{n+1}$, combining formula (12.76) and the bounds (12.77) and (12.79) obtained in the two previous lemmas, we deduce

$$\mathbb{D}_k(R) \leq 2 \times \frac{2^{\sigma_0(k)-p_0(k)}}{2} 4^{\sigma_1(k)+p_0(k)-1} = 2^{2\sigma_1(k)+p_0(k)+\sigma_0(k)-2} \leq 2^{2\sigma_1(k)+2\sigma_0(k)-2}.$$

Since $\sigma_0(k) + \sigma_1(k) = n+1$ for the values $2^n < k < 2^{n+1}$, for such values $\mathbb{D}_k(R) \leq 2^{2n} \leq k^2$. The previous bound can be checked for $k = 2^n$ since $\beta_k = \frac{1}{4}$. We observe then that the bound $(k+1)^2$ is valid for any $k \geq 1$. □

Combining this result with (12.23), we obtain the following stability estimate for the multivariate interpolation operator.

Corollary 2. *With $X = [-1,1]$ and Z an $\Re$-Leja sequence on $[-1,1]$, one has*

$$\mathbb{L}_\Lambda \leq (\#(\Lambda))^3, \tag{12.81}$$

for any downward closed set Λ.

References

1. L. Bialas-Ciez, J.P. Calvi, Pseudo Leja sequences. Ann. Mat. Pura Appl. **191**, 53–75 (2012)
2. J.P. Calvi, V.M. Phung, On the Lebesgue constant of Leja sequences for the unit disk and its applications to multivariate interpolation. J. Approx. Theory **163–5**, 608–622 (2011)
3. J.P. Calvi, V.M. Phung, Lagrange interpolation at real projections of Leja sequences for the unit disk. Proc. Am. Math. Soc. **140**(12), 4271–4284 (2012)
4. A. Chkifa, On the Lebesgue constant of Leja sequences for the complex unit disk and of their real projection. J. Approx. Theory **166**, 176–200 (2013)
5. A. Chkifa, Méthodes polynomiales parcimonieuses en grande dimension. Application aux EDP Paramétriques. Ph.D. thesis, Laboratoire Jacques Louis Lions, 2014
6. A. Chkifa, A. Cohen, C. Schwab, High-dimensional adaptive sparse polynomial interpolation and applications to parametric PDEs. Found. Comput. Math. **14**(4), 601–633 (2013)
7. A. Chkifa, A. Cohen, R. DeVore, C. Schwab, Sparse adaptive Taylor approximation algorithms for parametric and stochastic elliptic PDEs. Math. Model. Numer. Anal. **47**, 253–280 (2013)
8. A. Chkifa, A. Cohen, C. Schwab, Breaking the curse of dimensionality in parametric PDEs. Math. Pures Appl. **103**(2), 400–428 (2015)
9. A. Cohen, R. DeVore, C. Schwab, Convergence rates of best *N*-term Galerkin approximations for a class of elliptic PDEs. Found. Comput. Math. **10**, 615–646 (2010)
10. A. Cohen, R. DeVore, C. Schwab, Analytic regularity and polynomial approximation of parametric and stochastic PDE's. Anal. Appl. **9**, 11–47 (2011)
11. R.A. Devore, G.G. Lorentz, *Constructive Approximation* (Springer, Berlin, 1993)
12. V.K. Dzjadyk, V.V. Ivanov, On asymptotics and estimates for the uniform norms of the Lagrange interpolation polynomials corresponding to the Chebyshev nodal points. Anal. Math. **9–11**, 85–97 (1983)
13. J. Kuntzman, *Méthodes Numériques - Interpolation, Dérivées* (Dunod, Paris, 1959)
14. R. Taylor, Lagrange interpolation on Leja points. Ph.D. thesis, University of South Florida, 2008

Part V
Data Acquisition

Chapter 13
OperA: Operator-Based Annihilation for Finite-Rate-of-Innovation Signal Sampling

Chandra Sekhar Seelamantula

Abstract We consider the problem of finite-rate-of-innovation (FRI) signal sampling, which received a lot of attention from the sampling community in the past decade. Specifically, we consider the mechanism of reconstruction based on the notion of annihilation and show that one can design annihilators based on linear differential operators and translation operators. By working in the continuous domain, we show that annihilation can be achieved on nonuniform grids using derivative-type sampling approaches and on interleaved sampling grids using translation-operator-based annihilators. The standard annihilation procedure operating in the discrete domain becomes a special case of this approach. We show perfect reconstruction results with the sampling approaches considered and present simulation results to support the theoretical calculations. We also establish a link between annihilation and exponential-spline construction. Monte Carlo performance analysis in the presence of noise shows that annihilation on interleaved sampling grids leads to more noise-robust estimates than annihilation on uniform sampling grids.

13.1 Introduction

Shannon's sampling theorem [1] for bandlimited signals set the stage for analog-to-digital conversion and revolutionized the way electrical engineers addressed the problem of information transmission using digital means. Shannon's sampling theorem guarantees that bandlimited signals can be reconstructed exactly from samples taken at integer multiples of the sampling period provided that the sampling period is chosen to satisfy the Nyquist criterion. The notion of perfect bandlimitedness is not practical, as most real-world signals are not bandlimited. However, essential bandlimitedness can be ensured by using an analog lowpass prefilter to suppress frequencies beyond a chosen frequency. In the decades that followed Shannon's

C.S. Seelamantula (✉)
Department of Electrical Engineering, Indian Institute of Science, Bangalore 560 012, Karnataka, India
e-mail: chandra.sekhar@ieee.org

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_13

celebrated paper, many attempts have been made to generalize Shannon's theory [2, 3]. One such generalization was provided by Papoulis (*Papoulis' Generalized Sampling Theorem* [4]), who considered the problem of reconstructing from the samples of filtered versions of a bandlimited signal. Each channel operates at the corresponding Nyquist rate and the overall sampling scheme operates at the Nyquist rate corresponding to the full-band, bandlimited signal.

Another type of generalization came in the context of sampling and reconstructing signals that are not bandlimited, but lie in a shift-invariant subspace spanned by an appropriately chosen generator kernel [3, 5]. The signal subspace structure in this case is similar to that of Shannon's where the generator kernel is a sinc function. Some examples of generator kernels are basis splines (B-splines) [6–8], exponential splines (E-splines) [9, 10], Gaussian functions, etc.

In the past one decade, extensive research has gone into the sampling and reconstruction of signals that are not bandlimited, but possess a certain structure. The structure could be in the form of sparsity in time or spatial domains or parsimony of representation in a suitably chosen basis. In this direction, various techniques and algorithms have been developed within the framework of *Compressed Sensing* (CS) [11–15]. This is different from the classical sampling framework where one addresses the question of analog-to-discrete-signal conversion in the process of sampling, and vice versa, during reconstruction. The CS literature largely focuses on finite-dimensional measurements and sequence recovery from projections, which may also be random. Another direction of research within the framework of sparsity goes by the name of *finite-rate-of-innovation (FRI) signal sampling*. The notion of FRI was introduced by Vetterli et al. [16] to quantify the degrees of freedom possessed by certain classes of parametric, not necessarily bandlimited, signals. Specifically, a signal is said to have a finite rate of innovation if it has a fixed number of degrees of freedom per unit time interval or spatial extent. Typical examples of FRI signals are a stream of Dirac impulses, stream of differentiated Dirac impulses, piecewise-constant signals [16], piecewise-polynomial or trigonometric signals [17], nonuniform splines, superposition of amplitude-scaled and shifted pulses, etc. The sampling mechanism for such signals comprises computing inner-products (equivalently, filtering) with a suitable sampling kernel, resulting in a sequence of measurements. Often, the number of such measurements required for capturing the degrees of freedom is finite. The sampling kernel is carefully designed such that the measurements or their linear combinations can be expressed in the form of a *power-sum* sequence [18]. Some important examples of such kernels are Gaussians, sinc functions [16], sum-of-sincs function [19], polynomial-reproducing kernels (such as B-splines), exponential-reproducing kernels (such as E-splines), and kernels associated with rational transfer functions [20], etc. In particular, practically realizable kernels such as causal exponentials have given rise to stable reconstruction algorithms in the multichannel setting [21, 22]. In many formulations, the problem of FRI signal reconstruction is eventually reduced to one of solving for the parameters of a sequence of the form $f(n) = \sum_{\ell=1}^{L} a_\ell \, u_\ell^n$,

where the parameters a_ℓs and u_ℓs are unknown and have to be solved for. Such problems are encountered in various disciplines of engineering and science. In signal processing, this problem has been extensively studied within the framework of high-resolution spectral estimation. One of the early methods to solve the problem in the context of harmonic retrieval dates back to the eighteenth century, when de Prony [23] introduced the so-called *annihilating filter* methodology to solve for the parameters. Subsequent developments in high-resolution spectral estimation led to many important contributions such as the multiple signal classification (MUSIC) algorithm [24], estimation of signal parameters via rotational invariance technique (ESPRIT) [25], minimum-norm method [26], and their numerous variants [27, 28].

While there are strong similarities between the FRI methodology and high-resolution spectral estimation, there are also differences and unique challenges within the FRI context. For example, the kernel design is unique to FRI sampling methods and is intimately tied to the FRI signal and the reconstruction methodology adopted. In most cases, if one would like to deploy annihilation-based methods for reconstruction, it would be convenient to have a kernel that is capable of reproducing exponentials or polynomials (*Strang-Fix condition* [20], or its generalized version [29]) so that the measurements can be expressed using linear operators as a power-sum. There is a duality between harmonic retrieval and the reconstruction of a periodic stream of Dirac impulses, but the link becomes weak when one is interested in sampling more sophisticated and practically relevant FRI signals such as piecewise-constant signals, piecewise-polynomial/sinusoidal signals, etc. In other words, generic FRI signals do not have counterparts in the spectral estimation paradigm. FRI methods have proved to be useful in improving the image reconstruction quality in imaging modalities such as ultrasound [19, 30] and optical-coherence tomography [31, 32].

13.1.1 This chapter

In this chapter, we address the core reconstruction methodology based on the concept of annihilation. In most FRI methods, the reconstruction problem is reduced to solving for the parameters of a linear combination of exponential functions. Therefore, we focus on the sum-of-exponential signals and develop operator-based annihilation (codenamed *OperA*) strategies. We work largely in the continuous domain and address the reconstruction issue based on discrete measurements. We take an *ab initio* approach and show that annihilation can be directly achieved using linear differential operators (Section 13.2), which gives rise to a new type of *Derivative Sampling*, in which one measures the function (which is a sum of exponentials or linearly modulated exponentials) and a minimal number of derivatives (determined by the order of the model) at synchronized time instants, and then computes the parameters of the function within the annihilation framework. The interesting aspect is that exact reconstruction is guaranteed in theory and the sampling grid is not constrained to be uniform. We next demonstrate annihilation using translation

operators (Section 13.3), which may be viewed as the finite-difference counterpart of the differential operators. This approach has two advantages: (i) the annihilation equation manifests in continuous time, which makes it possible to deviate from the uniform sampling pattern to a more generic *interleaved sampling* pattern; and (ii) it establishes a direct link between annihilation operators and exponential splines, which play an important role in wavelet theory [33] and kernel design for FRI problems. More precisely, we show that exponential annihilation and exponential-spline localization filter design problems are equivalent. Monte Carlo performance analysis in the presence of noise shows that annihilation on an interleaved sampling grid leads to a significant improvement in estimation performance over annihilation carried out on a uniform sampling grid.

13.1.2 Notations

The first-order derivative operator is denoted as $\mathrm{D} = \dfrac{\mathrm{d}}{\mathrm{d}t}$, while the higher-order ones are denoted as $\mathrm{D}^k = \dfrac{\mathrm{d}^\mathrm{k}}{\mathrm{d}t^k}, k = 1, 2, 3, \cdots$. The identity operator is denoted by I. Since D^0 does not involve computing any derivative, naturally $\mathrm{D}^0 = I$. The shift/translation operator is denoted by S_τ, where the subscript denotes the quantum of shift; for example, $\mathrm{S}_\tau\{f(t)\} = f(t-\tau)$.

13.2 Annihilation based on differential operators

13.2.1 Sum of exponentials

Consider the differential equation $\mathrm{D}f = \alpha f$, where $\alpha \in \mathbb{C}$. The nontrivial solution of this differential equation is $f(t) = \exp(\alpha t)$. In fact, f is also an eigenfunction of the operator D. This fact can also be expressed equivalently as follows:

$$(\mathrm{D} - \alpha I)f = 0. \tag{13.1}$$

Thus, the operator $L \stackrel{\Delta}{=} (\mathrm{D} - \alpha I)$ maps f to 0. In other words, $(\mathrm{D} - \alpha I)$ *annihilates* f. If we turn the problem the other way round and ask for what values of β will $(\mathrm{D} - \beta I)$ annihilate f?, the answer would be $\beta = \alpha$. Therefore, $(\mathrm{D} - \beta I)f(t) = 0$, $\forall t \in \mathbb{R}$ if and only if $\beta = \alpha$. Thus, in principle, given access to f and its derivative, the parameter α can be computed directly by asking for what value of β, $(\mathrm{D} - \beta I)f(t)$ vanishes, and that value would be α (up to phase-wrapping ambiguity). This property carries over to the multi-exponential generalization that we shall soon attempt. Since the annihilation happens for all values of t, the function

and its derivative are required to be measured only over a short interval or even at a single point. Thus, given f and its derivative at a point, and the knowledge that f is an exponential helps us determine the function for all values of t.

The function $f(t)$ is, in general, an unbounded function. For $\mathrm{Re}\{\alpha\} < 0$, where Re denotes the real part, the function blows up for $t < 0$ and vice versa for $\mathrm{Re}\{\alpha\} > 0$. For purely imaginary α, the function is oscillatory. In all these cases, the function is neither integrable nor square integrable, a sought-after property in most signal processing paradigms, which does not seem to be required here. Thus, *annihilation* has little to do with function stability or boundedness. In practice, however, one works with functions over a finite duration interval.

By taking into account linearity of the differential operator, we make a generalization. Consider the function

$$f(t) = \sum_{i=1}^{p} a_i \exp(\alpha_i t). \tag{13.2}$$

Each of the constituents $a_i \exp(\alpha_i t)$ is annihilated by the operator $L_i \stackrel{\Delta}{=} (\mathrm{D} - \alpha_i I)$. Hence, $f(t)$ is annihilated by the composite operator $L = \prod_{i=1}^{p} L_i$, which can be expressed as $\sum_{i=0}^{p} \gamma_i \mathrm{D}^i$ with $\mathrm{D}^0 = I$ and $\gamma_p = 1$. Consider the question of determining the parameters of $f(t)$ given that it is of the form shown in (13.2). Constructing $L = \sum_{i=0}^{p} \gamma_i \mathrm{D}^i$, and enforcing $Lf(t) = 0$, γ_i can be determined. Since the annihilation holds for all values of t, and the number of unknowns in $\boldsymbol{\gamma}$ is p, we need $Lf(t_i) = 0, i = 1, 2, \cdots, p$, where t_i are the observation/sampling instants, which are not required to be uniformly spaced. In principle, any set of p points would suffice for estimating the γ_i s. The system of equations that one has to solve is the following:

$$\underbrace{\begin{pmatrix} f(t_1) & f^{(1)}(t_1) & f^{(2)}(t_1) & \cdots & f^{(p)}(t_1) \\ f(t_2) & f^{(1)}(t_2) & f^{(2)}(t_2) & \cdots & f^{(p)}(t_2) \\ f(t_3) & f^{(1)}(t_3) & f^{(2)}(t_3) & \cdots & f^{(p)}(t_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(t_p) & f^{(1)}(t_p) & f^{(2)}(t_p) & \cdots & f^{(p)}(t_p) \end{pmatrix}}_{\mathbb{F}} \underbrace{\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{pmatrix}}_{\boldsymbol{\gamma}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{0}}.$$

Thus, $\boldsymbol{\gamma}$ is a vector that lies in the null-space of the matrix $\mathbb{F}$. Incidentally, the matrix $\mathbb{F}$ has an *alternant structure*. This is a linear system of equations involving exponentials and one is interested in a non-trivial solution. In most cases, with distinct exponentials, a unique non-trivial solution may exist, but generic guarantees

on uniqueness are not yet available. Hence, given a function and its p derivatives at p points, it is possible to estimate the coefficient vector, from which one can construct the annihilator polynomial $\sum_{i=0}^{p} \gamma_i \, \mathrm{D}^i$. By expressing the polynomial in terms of its factors $\sum_{i=0}^{p} \gamma_i \, \mathrm{D}^i = \prod_{i=1}^{p} (\mathrm{D} - \alpha_i \, I)$, the parameters α_i may be computed by following a *root-finding procedure*. Once the αs are obtained, one can obtain the a_is in (13.2) by solving the following linear system of equations:

$$\underbrace{\begin{pmatrix} \exp(\alpha_1 t_1) & \exp(\alpha_2 t_1) & \exp(\alpha_3 t_1) & \cdots & \exp(\alpha_p t_1) \\ \exp(\alpha_1 t_2) & \exp(\alpha_2 t_2) & \exp(\alpha_3 t_2) & \cdots & \exp(\alpha_p t_2) \\ \exp(\alpha_1 t_3) & \exp(\alpha_2 t_3) & \exp(\alpha_3 t_3) & \cdots & \exp(\alpha_p t_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \exp(\alpha_1 t_p) & \exp(\alpha_2 t_p) & \exp(\alpha_3 t_p) & \cdots & \exp(\alpha_p t_p) \end{pmatrix}}_{\mathbb{E}} \underbrace{\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix}}_{\mathbf{a}} = \underbrace{\begin{pmatrix} f(t_1) \\ f(t_2) \\ f(t_3) \\ \vdots \\ f(t_p) \end{pmatrix}}_{\mathbf{f}},$$

where $\mathbf{f}$ denotes the vector of measurements. Instead of $\mathbf{f}$, one could also employ the vector of derivatives or the higher-order derivatives $\mathbf{f}^{(\mathrm{k})}$, with matrix $\mathbb{E}$ constructed accordingly. In any case, we have perfect reconstruction of the sum of exponentials. A schematic of the differential-operator-based sampling circuit is shown in Figure 13.1.

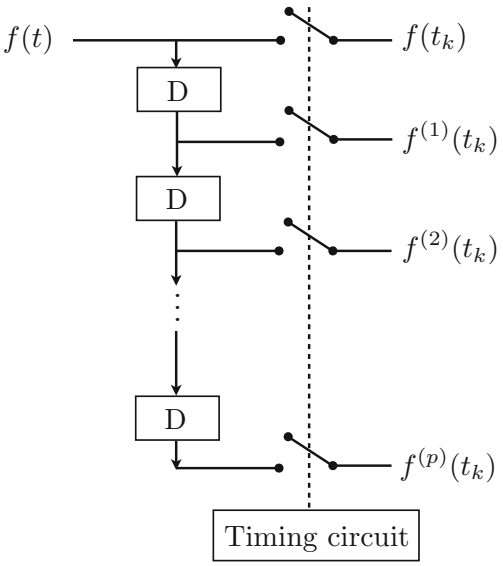


Fig. 13.1 A schematic of the differential-operator-based sampling circuit. The set of measurements have to be in time-synchrony for every t_k, which is ensured by the timing circuit.

We summarize the preceding analysis in the form of the following proposition:

Proposition 1. *Given a function $f(t)$ to be of the form of a linear combination of distinct exponentials: $f(t) = \sum_{i=1}^{p} a_i \exp(\alpha_i t)$, p samples of the function and its p derivatives, at p distinct known instants, are sufficient to fully characterize the function f for all values of t.*

The above proposition is reminiscent of the *derivative sampling* approach, which, in turn is a special case of Papoulis' generalized sampling theorem. These approaches consider the reconstruction of a bandlimited function from the knowledge of the function and its derivatives [34]. The difference between the two approaches is that we are dealing with nonbandlimited functions in a finite dimensional setting with no specific requirements of stability.

The key to reconstruction is annihilation using the differential operator; however, annihilation is not the goal in itself, which can possibly be obtained in many different ways. The goal is to come up with a signal-dependent annihilator in a fashion that enables reliable estimation of the signal parameters.

13.2.2 Simulation results

To illustrate the accurate reconstruction capability of the operator-based sampling approach, we consider a sum of five exponentials, with random parameters. In theory, the exponentials are not required to be bounded, but keeping numerical precision issues in mind, we have considered exponential parameters with negative real parts. The poles are complex-valued in general and their locations corresponding to one such instantiation of the exponentials are shown in Figure 13.2.

The samples of the function and the derivatives were measured at fifty random locations, generated by perturbing a uniform sampling grid at the integers 1 to 50. The perturbation has a uniform distribution over $[0, 0.5]$. The resulting nonuniform locations are known to the reconstruction algorithm. The amplitudes of the exponentials were chosen from a standard normal distribution (zero mean, unity variance). The number of exponentials were assumed to be known in the reconstruction process. The estimated poles coincided quite accurately with the ground truth as shown in Figure 13.2. The complex-valued ground truth signal $f(t)$ and the reconstructed counterpart $\hat{f}(t)$ is shown in Figure 13.3. The reconstruction error is quite low, of the order of 10^{-12}.

13.2.3 Sum of linearly modulated exponentials

Consider the function

$$f(t) = t \exp(\alpha t), \tag{13.3}$$

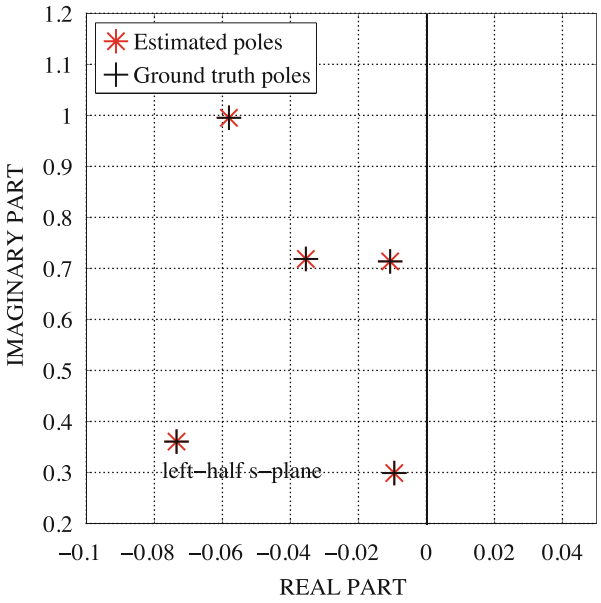


Fig. 13.2 Pole configurations corresponding to the synthesized signal vis-a-vis the poles estimated by the differential-operator-based-annihilation technique.

which is a ramp multiplied by an exponential. The function f satisfies the differential equation

$$(\mathrm{D} - \alpha\, I)^2 f = 0, \tag{13.4}$$

and the corresponding annihilator polynomial has repeated roots at α. Extending to the multicomponent case, we have

$$f(t) = \sum_{i=1}^{p} a_i\, t\, \exp(\alpha_i\, t), \tag{13.5}$$

which is annihilated by the composite operator $L = \prod_{i=1}^{p} L_i^2 = \sum_{i=0}^{2p} \gamma_i\, \mathrm{D}^i$, where $\gamma_{2p} = 1$. To determine the parameters of $f(t)$ given that it is of the form given in (13.5), we enforce $Lf(t) = 0, \forall t \in \mathbb{R}$. Since the annihilation holds for all values of t, and the number of unknowns in the vector γ is $2p$, we need $Lf(t_i) = 0, i = 1, 2, \cdots, 2p$, where t_is are the sampling instants. Observe that t_i are not required to be uniformly spaced. In principle, any set of $2p$ points would suffice for estimating the γs. The system of equations that one has to solve is the following:

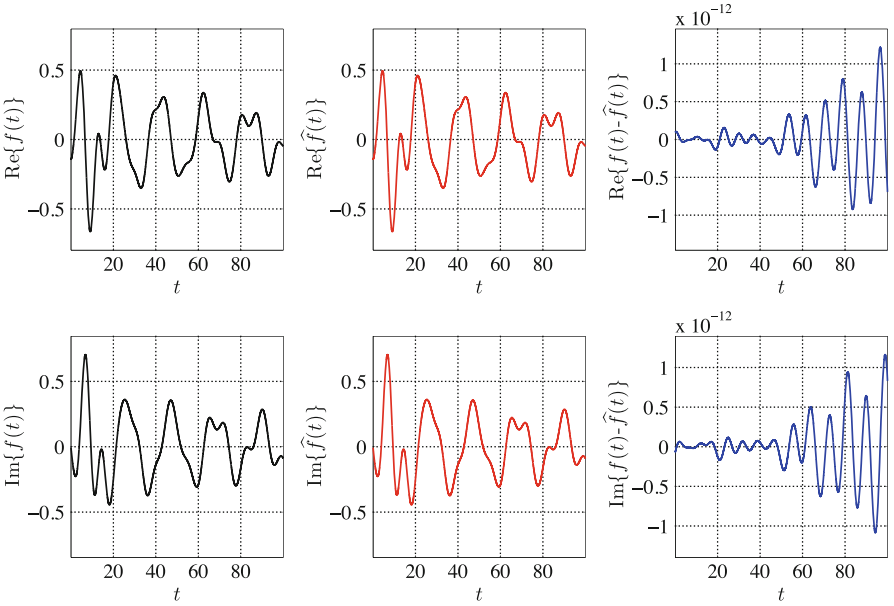


Fig. 13.3 A comparison of the reconstructed signal (sum of exponentials) with the ground truth. The first column shows the real and imaginary parts of the ground truth signal, the second column that of the reconstruction, and the third column corresponds to the error signal in the real and imaginary parts. Although theoretically, exact reconstruction is guaranteed, in practice, numerical precision constraints imposed by the root-finding technique limit the achievable accuracy, and hence the error signal shown in the third column is sufficiently small, but not vanishing.

$$\underbrace{\begin{pmatrix} f(t_1) & f^{(1)}(t_1) & f^{(2)}(t_1) & \cdots & f^{(2p)}(t_1) \\ f(t_2) & f^{(1)}(t_2) & f^{(2)}(t_2) & \cdots & f^{(2p)}(t_2) \\ f(t_3) & f^{(1)}(t_3) & f^{(2)}(t_3) & \cdots & f^{(2p)}(t_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(t_{2p}) & f^{(1)}(t_{2p}) & f^{(2)}(t_{2p}) & \cdots & f^{(2p)}(t_{2p}) \end{pmatrix}}_{\mathbb{F}} \underbrace{\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_{2p} \end{pmatrix}}_{\boldsymbol{\gamma}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{0}}.$$

Thus, $\boldsymbol{\gamma}$ lies in the null-space of the matrix $\mathbb{F}$. This is a linear system of equations with a non-trivial solution in general. Thus, given a function and its $2p$ derivatives at $2p$ distinct points, it is possible to estimate the coefficient vector, from which one can construct the annihilator polynomial $\sum_{i=0}^{2p} \gamma_i \, \mathrm{D}^i$. By expressing the polynomial in terms of its factors $\sum_{i=0}^{2p} \gamma_i \, \mathrm{D}^i = \prod_{i=1}^{p} (\mathrm{D} - \alpha_i \, I)^2$, we compute the parameters α_i (which

have multiplicity 2) essentially by following a *root-finding procedure*. Once the α_i s are obtained, one can obtain the a_is in (13.2) by solving the following linear system of equations:

$$\underbrace{\begin{pmatrix} t_1\exp(\alpha_1 t_1) & t_1\exp(\alpha_2 t_1) & t_1\exp(\alpha_3 t_1) & \cdots & t_1\exp(\alpha_p t_1) \\ t_2\exp(\alpha_1 t_2) & t_2\exp(\alpha_2 t_2) & t_2\exp(\alpha_3 t_2) & \cdots & t_2\exp(\alpha_p t_2) \\ t_3\exp(\alpha_1 t_3) & t_3\exp(\alpha_2 t_3) & t_3\exp(\alpha_3 t_3) & \cdots & t_3\exp(\alpha_p t_3) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ t_p\exp(\alpha_1 t_p) & t_p\exp(\alpha_2 t_p) & t_p\exp(\alpha_3 t_p) & \cdots & t_p\exp(\alpha_p t_p) \end{pmatrix}}_{\mathbb{E}} \underbrace{\begin{pmatrix} a_1 \\ a_2 \\ a_3 \\ \vdots \\ a_p \end{pmatrix}}_{\mathbf{a}} = \underbrace{\begin{pmatrix} f(t_1) \\ f(t_2) \\ f(t_3) \\ \vdots \\ f(t_p) \end{pmatrix}}_{\mathbf{f}},$$

where $\mathbf{f}$ denotes the vector function measurements. Instead of $\mathbf{f}$, we could also employ the vector of derivatives or the higher-order derivatives $\mathbf{f}^{(k)}$, with corresponding matrix $\mathbb{E}$ constructed accordingly. The preceding analysis leads to the following proposition.

Proposition 2. *Given a function $f(t)$ to be of the form of a linear combination of p distinct exponentials: $f(t) = \sum_{i=1}^{p} a_i\, t\, \exp(\alpha_i t)$, $2p$ samples of the function and its derivatives at $2p$ distinct known instants are sufficient to fully characterize the function f for all values of t.*

The above proposition may also be viewed as a finite-dimensional counterpart of Papoulis' generalized sampling theorem for nonbandlimited signals. By considering higher-order roots of the corresponding annihilator polynomial, we can develop similar results, which can all be summarized in the following proposition.

Proposition 3. *Given a function $f(t)$ to be of the form of a linear combination of distinct polynomial modulated exponentials: $f(t) = \sum_{i=1}^{p} a_i\, t^n\, \exp(\alpha_i t)$, where n is a natural number, np samples of the function and its derivatives at np distinct known instants are sufficient to fully characterize the function f for all values of t.*

13.2.4 Simulation results

We consider four randomly placed double (repeated) poles on the left-half s-plane shown in Figure 13.4. The sampling instants were first chosen at the integers 1 to 50, followed by a uniformly distributed random perturbation over $[0, 0.5]$, resulting overall in a nonuniform sampling grid. The corresponding signal $f(t)$ (which is a sum of linearly modulated exponentials), the reconstruction $\hat{f}(t)$, and the reconstruction error in both real and imaginary parts are shown in Figure 13.5. The error is of the order of 10^{-6} in comparison with the signal amplitude and

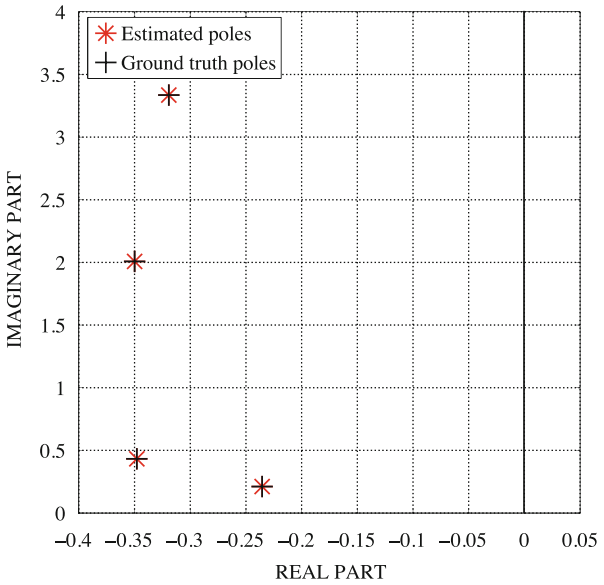


Fig. 13.4 Pole configuration (repeated poles) for a sum of linearly modulated exponentials. The estimated poles overlap accurately with the ground truth.

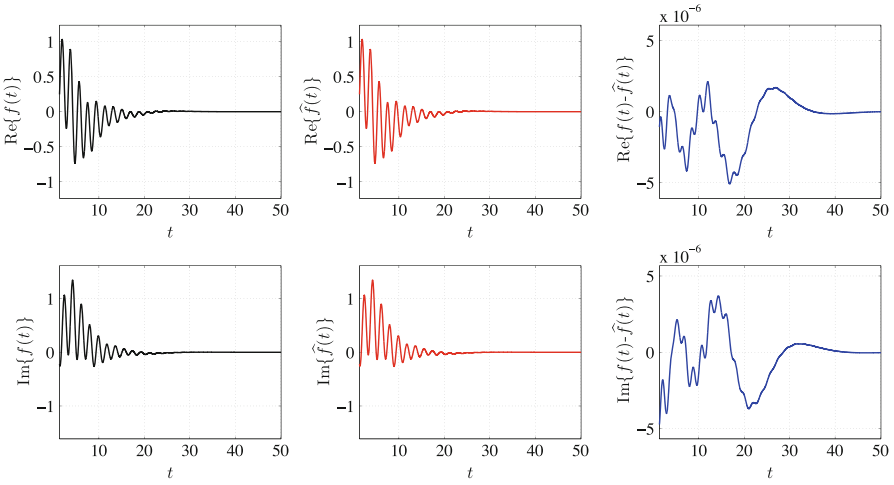


Fig. 13.5 A comparison of the reconstructed signal (sum of linearly modulated exponentials) with the ground truth. The first column shows the real and imaginary parts of the ground truth signal, the second column that of the reconstruction, and the third column corresponds to the error signal in the real and imaginary parts.

acceptable for practical applications. However, in comparison with the error for the sum-of-exponentials signal, the error is higher in this case. We believe this is due to numerical precision limitations in root-finding procedures.

13.2.5 Causal exponentials

The function types considered in the preceding analysis are not stable since they contain exponentials that extend for all values of time. In practice, we have to deal with their causal counterparts (which may not be stable), which can be obtained by multiplying with the Heaviside function. This causes a small hurdle in applying the results directly. The operator $(\mathrm{D} - \alpha\, I)$ annihilates $f(t) = \exp(\alpha\, t),\ \forall t \in \mathbb{R}$, but not $\exp(\alpha\, t)u(t)$, where $u(t)$ denotes the Heaviside function, because

$$(\mathrm{D} - \alpha\, I)f(t) = \delta(t), \tag{13.6}$$

where δ denotes the Dirac delta, which is caused by the derivative operator acting on the Heaviside step discontinuity at $t = 0$. Since the discrepancy between the two-sided exponentials and the causal ones is localized around the origin, the problem can be overcome by taking measurements at instants away from the origin.

Practical linear circuits are governed by linear differential equations and their responses are causal. The response can be sampled and the system behavior can be analyzed by using the measurements. The exponential parameters are directly related to the degrees of damping or oscillation in the system and hence can be computed directly using the proposed approach.

13.3 Annihilation based on translation operators

Consider $f(t) = \exp(\alpha\, t)$, and the translated version $S_T\{f\}(t) \stackrel{\Delta}{=} f(t - T) = \exp(\alpha\,(t - T)) = \exp(-\alpha\, T)f(t)$. Hence, $(I - \exp(\alpha\, T)S_T)\{f\}(t) = 0, \forall t$. Thus, the operator $M \stackrel{\Delta}{=} (I - \exp(\alpha\, T)S_T)$ is also an annihilator of the exponential. Since it is also linear and shift-invariant, the annihilation property holds for a linear combination of exponentials. Thus, there is more than one way to annihilate exponentials. Given a function $f(t)$ with an unknown parameter α, with the goal of estimating α, we could construct an operator $(I - \exp(\beta\, T)S_T)$ and determine for the value of β for which $(I - \exp(\beta\, T)S_T)\{f\}(t) = 0$. Comparing with the annihilation based on the differential operator, we can view the operator $(I - \exp(\alpha\, T)S_T)$ as a weighted finite-difference operator, and as an approximation to the continuous-domain derivative operator.

Similar to the single exponent case, it is straightforward to verify that the multi-exponent function $f(t) = \sum_{i=1}^{p} a_i \exp(\alpha_i t)$ is annihilated for all time by the operator $M \stackrel{\Delta}{=} \prod_{i=1}^{p} \underbrace{(I - \exp(\alpha_i T) S_T)}_{M_i}$, which can be expressed in additive form as $M = \sum_{i=0}^{p} \gamma_i S_{iT}$, where $\gamma_0 = 1$, and S_{iT} is the i-times dilated version of the translation operator S_T and γ_is are dependent on the quantum of shift and the exponential parameters. In principle, since the annihilation is valid for all values of t, one could consider a generic annihilator $\prod_{i=1}^{p}(I - \exp(\alpha_i T_i) S_{T_i})$, but we shall not consider the generic one in the subsequent analysis.

To estimate the parameters a_i and α_i, we start with the annihilation equation using operator M:

$$\sum_{i=0}^{p} \gamma_i f(t - i T) = 0, \forall t, \tag{13.7}$$

and setup a system of equations

$$\sum_{i=0}^{p} \gamma_i f(t_\ell - i T) = 0, \;\; \ell = 1, 2, 3, \cdots, p, \tag{13.8}$$

corresponding to a set of nonuniform sampling instants. However, from the preceding equation, it is also clear that corresponding to each t_ℓ, the measurements are also required at $t_\ell - iT$, $i = 0, 1, 2, \cdots, p$, requiring a total of $p(p+1)$ measurements of f. The corresponding matrix equation for annihilation would be

$$\underbrace{\begin{pmatrix} f(t_1) & f(t_1 - T) & f(t_1 - 2T) & \cdots & f(t_1 - pT) \\ f(t_2) & f(t_2 - T) & f(t_2 - 2T) & \cdots & f(t_2 - pT) \\ f(t_3) & f(t_3 - T) & f(t_3 - 2T) & \cdots & f(t_3 - pT) \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ f(t_p) & f(t_p - T) & f(t_p - 2T) & \cdots & f(t_p - pT) \end{pmatrix}}_{\mathbb{F}} \underbrace{\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{pmatrix}}_{\boldsymbol{\gamma}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{0}}.$$

Once the γ vector is computed, the roots of the polynomial $\sum_{i=0}^{p} \gamma_i x^i$ are $\{e^{-\alpha_i T}, i = 1, 2, \cdots, p\}$, from which α_i can be calculated. Substituting the estimated α_i and the measurements in $f(t) = \sum_{i=1}^{p} a_i \exp(\alpha_i t)$, the coefficients a_i can be computed directly by solving a linear system of equations.

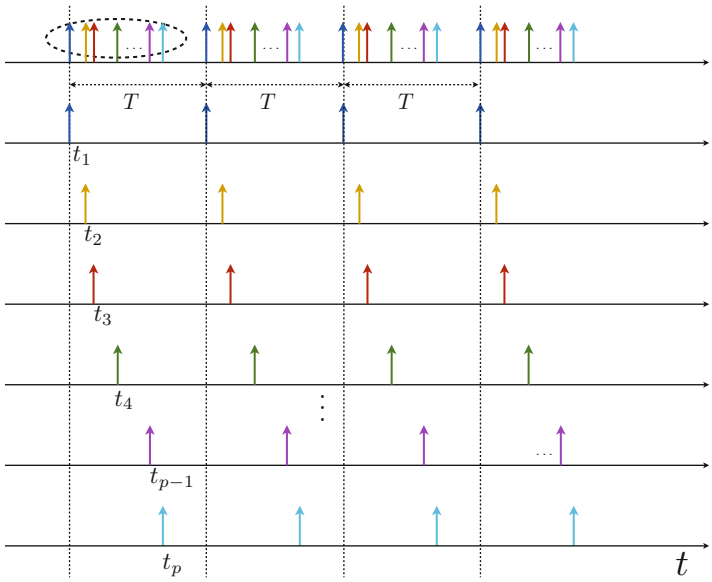


Fig. 13.6 Illustration of interleaved sampling grids. The nonuniform grid shown in the top most plot is actually comprised of various uniform sampling grids.

Hence, the annihilation methodology, in its generic form, works with nonuniformly spaced samples, except that the type of nonuniformity is structured. The sampling grid is doubly-indexed $\{t_\ell - iT, i = 0, 1, 2, \cdots, p; \ell = 1, 2, 3, \cdots, p\}$ and is actually *interleaved* in the sense that each of the grids corresponding to t_1, t_2, etc. is uniform, but across grids, the sampling instants are not necessarily uniformly spaced (cf. Figure 13.6).

We summarize the preceding developments in the form of the following proposition.

Proposition 4 (Annihilation on interleaved sampling grids). *Given a function* $f(t)$ *to be of the form of a linear combination of distinct exponentials:* $f(t) = \sum_{i=1}^{p} a_i \exp(\alpha_i t)$, $p(p+1)$ *measurements coming from the function and its* $T \times \mathbb{Z}$*-shifted versions at* p *distinct known instants on an interleaved sampling grid are sufficient to fully characterize the function* f *for all values of* t.

For illustration, consider a sum of five exponentials with poles selected randomly from the left-half s-plane. One particular configuration of the poles is shown in Figure 13.7. The amplitudes are chosen randomly from a uniform distribution over $[0, 1]$. The sampling instants are random as well, chosen according to a uniform distribution over $[0, 50]$. The value of T was set to unity. The reconstruction results are shown in Figure 13.8.

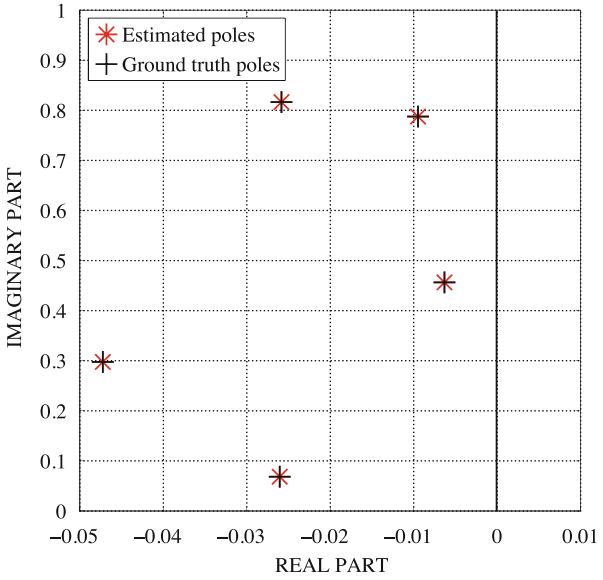


Fig. 13.7 The estimated pole locations in comparison with the ground truth poles.

13.3.1 Applications of interleaved sampling

Interleaved sampling may be viewed as a generalization of *multicoset sampling*, which has found applications mainly in multiband, wideband communication applications such as cognitive radio. Interleaved sampling has already been considered in the sampling literature and is known by the name periodic nonuniform sampling. For example, Lin and Vaidyanathan proposed periodic nonuniform sampling and reconstruction strategies for bandpass signals [35]. Lacaze proposed realizable circuits to perform higher-order periodic nonuniform sampling [36]. Cheung and Marks [37] showed that, when spectral holes exist within the support of two-dimensional bandlimited signals, samples of the signal can be periodically deleted and the deleted samples can be estimated from the retained ones using linear interpolation. Vaidyanathan and Liu [38] and Foster and Herely [39] showed that bandlimited signals can be reconstructed from nonuniformly decimated samples of the corresponding sequences. Feng and Bresler [40] proposed periodic nonuniform sampling or multicoset sampling for reconstruction of multiband signals at the Landau minimum rate [41]. Venkataramani and Bresler [42, 43] carried out a detailed analysis of the method and provided bounds on the aliasing error. More recently, Mishali and Eldar [44] proposed a multicoset sampling approach for multiband signals where the analog signal is reconstructed from interleaved samples within the framework of compressive sensing. Interleaved sampling grids are also used for sensing in capacitive touch-screen displays [45].

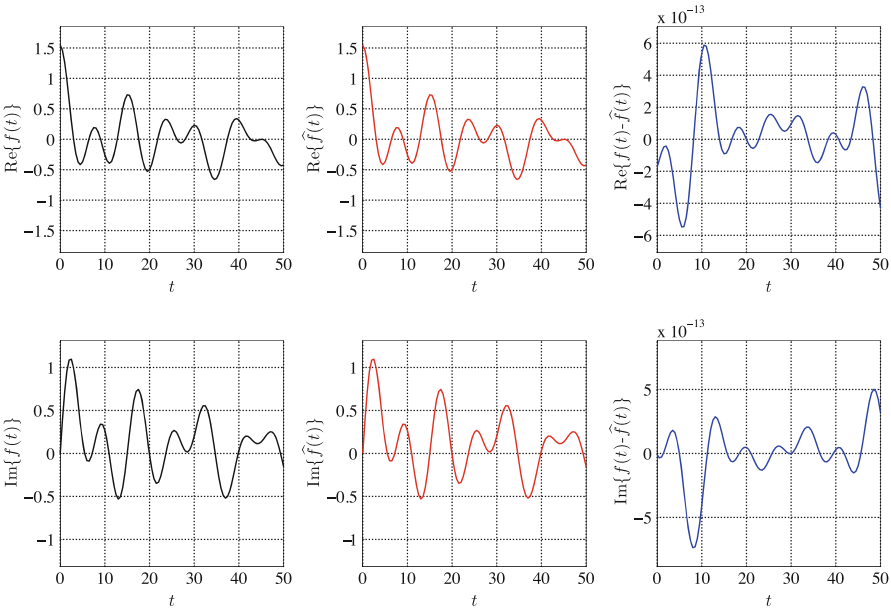


Fig. 13.8 Visual assessment of the reconstruction performance of the proposed technique. The first column shows the real and imaginary parts, respectively, of the ground truth signal, and the second column shows the counterparts for the reconstructed signal. The third column shows the instantaneous errors in estimating the real and imaginary parts. The error is of the order of 10^{-13}, which is quite small in comparison with the sampled signal amplitude.

13.3.2 Simulation results

Given uniformly sampled measurements, one can also obtain interleaved data using a random selection of a subset of samples. We illustrate the performance of the technique and its noise robustness in such a scenario. Consider two complex exponentials in white Gaussian noise:

$$f(n) = a_1 \exp(\mathrm{j}\alpha_1 n) + a_2 \exp(\mathrm{j}\alpha_2 n) + w(n), \quad n = 0, 1, 2, \cdots, N-1, \tag{13.9}$$

where $N = 128$ and $a_1 = 1.1, a_2 = 3.5, \alpha_1 = 0.2, \alpha_2 = 0.37$ are randomly selected parameters, but fixed throughout the subsequent experiment. $w(n)$ denotes samples of a zero-mean, white Gaussian noise process with variance σ^2. Interleaved subsets of this sequence are generated by selecting random integer n_i as $\{f(n_i + jT), i = 1, 2, \cdots, L; j = 0, 1, 2, \cdots, p\}$ such that $pT + \max(\{n_i, i = 1, 2, 3, \cdots, L\}) < N$, which is the given sequence length. In the present experiment, $L = 25$, $T = 6$, and $p = 2$. The interleaved sequences are stacked as

$$\mathbb{F} = \begin{pmatrix} f(n_1)\, f(n_1+T)\, f(n_1+2T) \\ f(n_2)\, f(n_2+T)\, f(n_2+2T) \\ f(n_3)\, f(n_3+T)\, f(n_3+2T) \\ \vdots \qquad\quad \vdots \qquad\quad \vdots \\ f(n_L)\, f(n_L+T)\, f(n_L+2T) \end{pmatrix}. \tag{13.10}$$

In the presence of noise, since exact annihilation cannot be achieved, we solve the minimization problem:

$$\min_{\boldsymbol{\gamma}} \|\mathbb{F}\boldsymbol{\gamma}\|^2 \quad \text{subject to} \quad \|\boldsymbol{\gamma}\|^2 = 1, \tag{13.11}$$

where $\boldsymbol{\gamma} = \begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \end{pmatrix}$. $\boldsymbol{\gamma}$ is the minimum eigenvector of $\mathbb{F}^H\mathbb{F}$. From the estimated $\boldsymbol{\gamma}$, the parameters α_1 and α_2 are obtained by computing the roots of a polynomial with coefficient vector $\boldsymbol{\gamma}$. The amplitude parameters a_1 and a_2 are estimated using a standard linear least-squares procedure. For every value of the signal-to-noise ratio, 1000 Monte Carlo trials are conducted and the parameters estimated using four different methods is carried out. The methods are the standard annihilating filter method applied directly on $f(n)$, with and without Cadzow's denoising method [46, 47], ESPRIT, and proposed interleaved sampling method with and without Cadzow's denoising on $f(n)$. The Cramér-Rao lower bound on the variance of α_1 and α_2 is given by $\text{CRLB}(\alpha_\ell) = \dfrac{6\sigma^2}{N^3 a_\ell^2}, \ell = 1, 2$ [27]. A comparison of the bias, variance, and mean-square errors for the location and amplitude parameters (α_1, α_2 and a_1 and a_2, respectively) for different techniques is shown in Figures 13.9 and 13.10. Some observations are in order:

1. The performance of the interleaved sampling based annihilation method is consistently better (in terms of bias, variance, and MSE) than the standard annihilating filter method operating on uniform samples.
2. Cadzow's denoising method significantly improves upon the performance of the annihilating filter method operating on uniform sampling or interleaved sampling grids.
3. The performance of the annihilation technique operating on interleaved sampling grid is on par with that of ESPRIT and slightly better than ESPRIT for SNR less than 0 dB.
4. Annihilation on interleaved or uniform sampling grids coupled with Cadzow's preprocessing step makes the performance meet the CRLB for SNR greater than 5 dB, making the estimators statistically efficient.

In the experiment, we considered integer values for $n_1, n_2, \cdots, n_L$, which means that the same set of measured samples, when rearranged in an interleaved form,

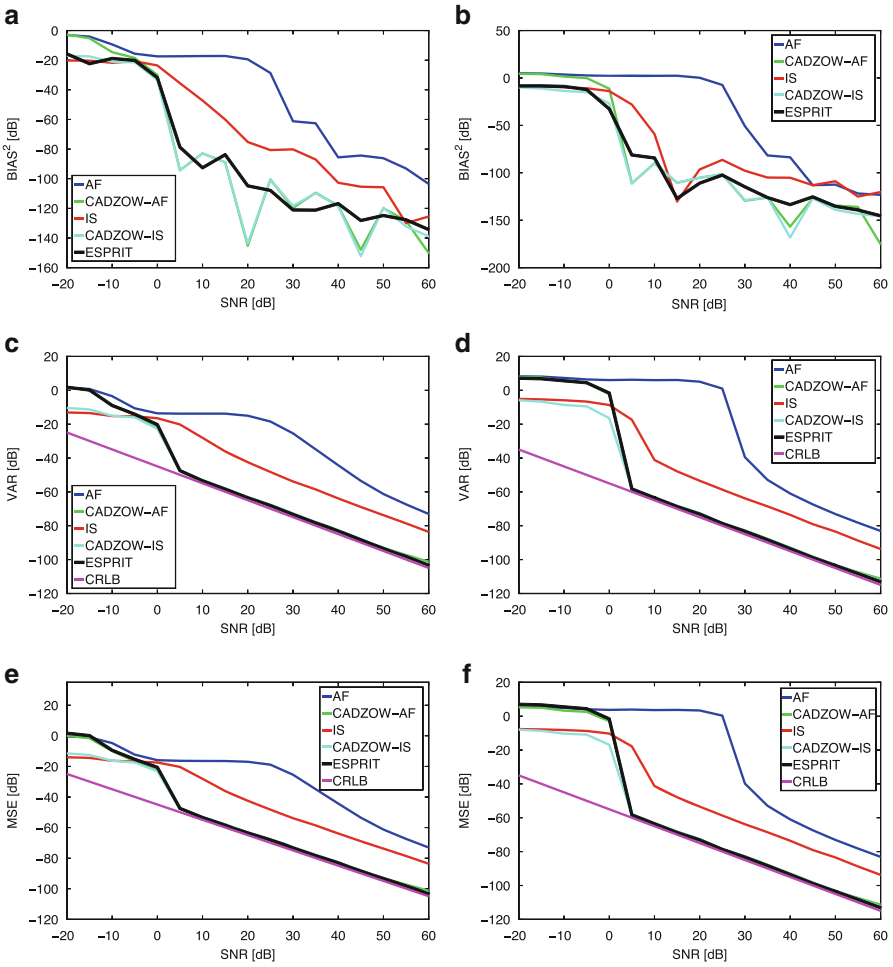


Fig. 13.9 (Color in electronic version) Squared bias (first row), variance (second row), and mean-square error (third row) performance in estimation of the location parameters α_1 (first column), and α_2 (second column). AF: standard annihilating filter (Prony's method); IS: interleaved sampling; Cadzow-AF: AF preceded by Cadzow's denoising method; Cadzow-IS: Interleaved sampling method preceded by Cadzow's denoising method.

lead to an improvement in performance. This is an interesting consequence in favor of the interleaved sampling method, and requires much detailed investigation. This property might be useful in practical imaging applications such as ultrasound [19, 30] or frequency-domain optical-coherence tomography [32].

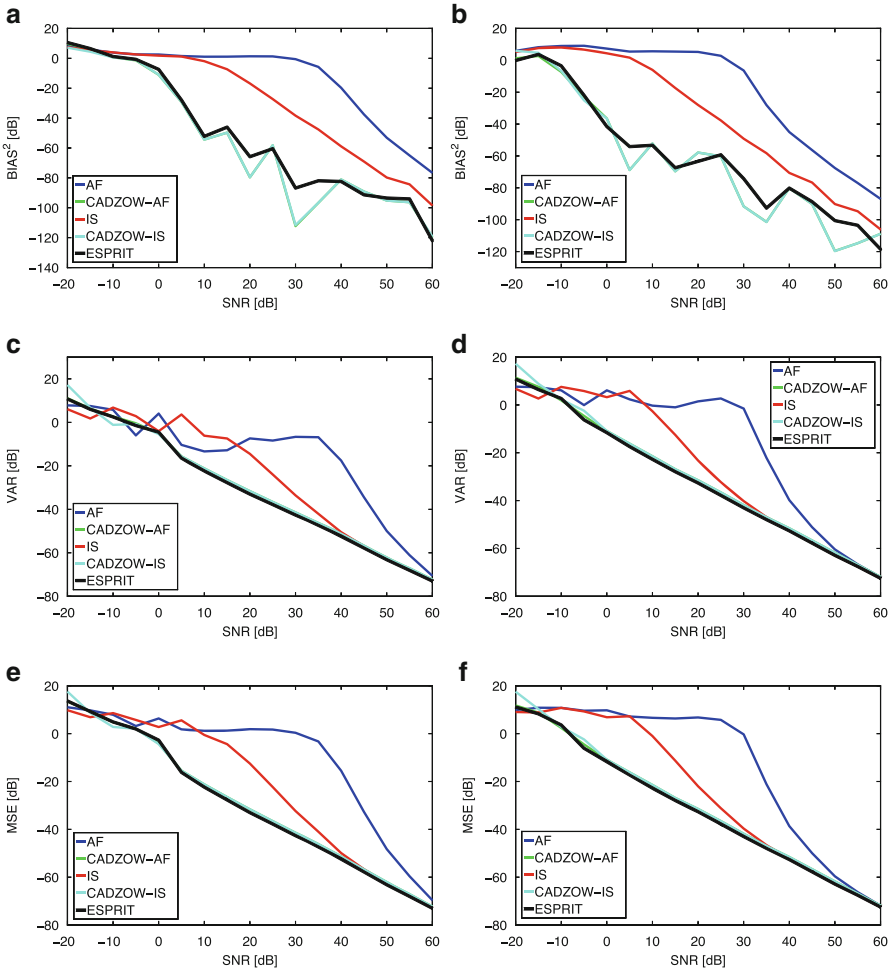


Fig. 13.10 (Color in electronic version) Squared bias (first row), variance (second row), and mean-square error (third row) performance in estimation of the amplitude parameters a_1 (first column), and a_2 (second column). AF: standard annihilating filter (Prony's method); IS: interleaved sampling; Cadzow-AF: AF preceded by Cadzow's denoising method; Cadzow-IS: Interleaved sampling method preceded by Cadzow's denoising method.

13.3.3 From an interleaved grid, to a uniform grid

If equispaced sampling instants are chosen, that is, $t_\ell = \ell T$, then we have the discrete-time convolution-based annihilation equation:

$$\sum_{i=0}^{p} \gamma_i f(\ell T - i T) = 0, \quad \ell = 1, 2, 3, \cdots, p, \tag{13.12}$$

and the corresponding matrix form would be

$$\underbrace{\begin{pmatrix} f(T) & f(0) & f(-T) & \cdots f((1-p)T) \\ f(2T) & f(T) & f(0) & \cdots f((2-p)T) \\ f(3T) & f(2T) & f(T) & \cdots f((3-p)T) \\ \vdots & \vdots & \vdots & \ddots \quad \vdots \\ f(pT) & f((p-1)T) & f((p-2)T) & \cdots \quad f(0) \end{pmatrix}}_{\mathbb{F}} \underbrace{\begin{pmatrix} \gamma_0 \\ \gamma_1 \\ \gamma_2 \\ \vdots \\ \gamma_p \end{pmatrix}}_{\boldsymbol{\gamma}} = \underbrace{\begin{pmatrix} 0 \\ 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}}_{\mathbf{0}},$$

where the matrix $\mathbb{F}$ is Toeplitz. The number of measurements required in the uniform sampling case is $2p$ as opposed to the nonuniform case where it is $p(p+1)$. Once $\boldsymbol{\gamma}$ is estimated, the exponential parameters α_i are estimated by root-finding and the weights a_i are estimated by solving a linear system of equations.

Proposition 5 (Annihilation on uniform sampling grids). *Given a function $f(t)$ to be of the form of a linear combination of distinct exponentials: $f(t) = \sum_{i=1}^{p} a_i \exp(\alpha_i t)$, $2p$ contiguous uniformly spaced measurements of the function are sufficient to fully characterize the function f for all values of t.*

13.3.4 Causal exponentials

Let us next consider the case of causal exponentials, which play a fundamental role in linear system theory and circuit analysis. For the exponential, $f(t) = \exp(\alpha t)u(t)$, the annihilation equation $(I - \exp(\alpha T)S_T)\{f\}(t) = 0$ does not hold for all values of t, but only for those values greater than T (Figure 13.11). For a p-component function

$$\sum_{i=0}^{p} \gamma_i f(t - iT) = 0, \quad t > pT. \tag{13.13}$$

Except for the difference in the interval over which annihilation takes place, the rest of the machinery for estimating the parameters remains the same as in the case of non-causal exponentials.

13.3.5 Translation-based annihilators and exponential splines

The translation-operator-based annihilator and causal exponentials have a close link with exponential splines. For example, consider $f(t) = \exp(\alpha t)u(t)$, and the annihilator $(I - \exp(\alpha T)S_T)\{f\}(t) = \exp(\alpha t)(u(t) - u(t-T)) = 0$ for $t > T$, which

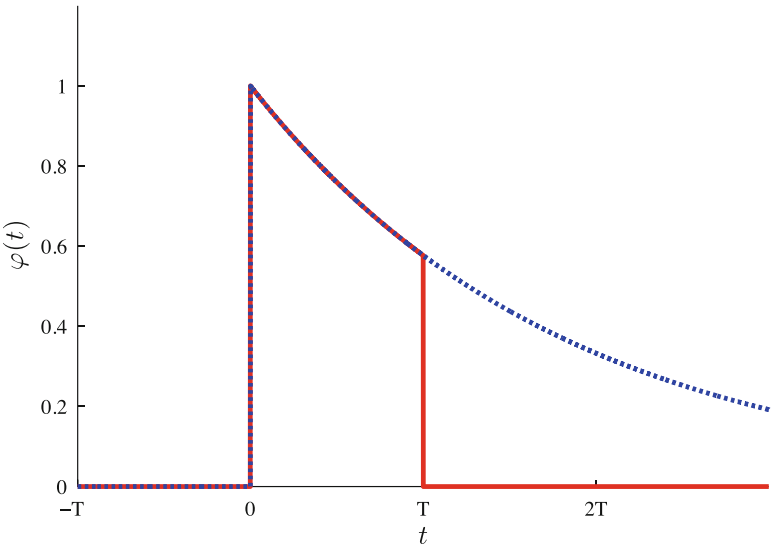


Fig. 13.11 A causal exponential (shown in blue) is annihilated over the interval $[T, \infty]$ using the translation-based localization operator, to generate an exponential spline of order zero (shown in red).

is actually the first-order exponential spline with parameter α. The operator $(I - \exp(\alpha\, T)S_T)$ is essentially the *spline localization operator*. Also, the Green function of the first-order operator $\mathrm{D} - \alpha\, I$ is the one-sided or causal exponential: $\exp(\alpha\, t)u(t)$. In signal processing terms, $f(t)$ is the impulse response of the inverse operator L^{-1}. The Green function of the cascaded operator $\prod_{i=1}^{p}(\mathrm{D} - \alpha_i\, I)$ is the convolution of the individual Green functions $\exp(\alpha_i\, t)\, u(t)$, which is also causal. The Green function of L can also be expressed as their linear combination: $f(t) = \sum_{i=1}^{p} a_i\, \exp(\alpha_i\, t)u(t)$. The corresponding annihilator is $\prod_{i=1}^{p}(I - \exp(\alpha_i\, T)S_T)$, which when acting on the Green function of L produces the exponential spline with parameters $\{\alpha_1, \alpha_2, \cdots, \alpha_p\}$. Hence, the process of annihilating a sum of causal exponentials over a semi-infinite line can be viewed as equivalent to generating the corresponding E-spline. When the parameters α_i are zero, then we get the corresponding polynomial B-spline, and the annihilator is actually the B-spline localization filter [6, 7]. The spline link is important because spline functions and their integer-shifted versions possess the Riesz bases property, which ensures stability of representation when transiting between continuous-domain functions and their discrete manifestations.

13.4 Conclusions

We investigated the aspect of annihilation in the context of FRI signal sampling and reconstruction, and showed that annihilation operators can be designed effectively using differential operators and translation operators. While most of the current FRI literature focuses on annihilation of sequences using discrete operators, we adopted a continuous-time approach, which resulted in many benefits. First, we realized annihilation using differential operators as well as translation operators. Second, the sampling geometry has been extended to include more generic versions such as nonuniform sampling and interleaved sampling. Third, the analysis established a direct link between the annihilator design and the localization filter in exponential spline theory. Although we have demonstrated results only on sum-of-exponential signals and linearly modulated exponentials, the key point is that the FRI signal measurements can be reduced to one of these forms and consequently, the proposed approaches become applicable to FRI signal sampling/reconstruction. Of particular interest is the annihilation on interleaved sampling grids, which can be constructed even from data measured on a uniform sampling grid. Monte Carlo performance analysis in the presence of noise showed that significant gain in estimation accuracy can be achieved by suitably interleaving the samples. We hope that such improvement in accuracy will also translate to superior quality reconstruction in ultrasound/optical imaging applications. The FRI sampling methods have been shown to be efficient for ultrasound signal reconstruction with less number of samples [19, 30]. Recently, we have also shown that FRI methods significantly improve the image reconstruction quality in frequency-domain optical-coherence tomography [32]. We hope that the approaches developed in this chapter will prove to be useful in further enhancing the reconstruction quality and resolution in such imaging modalities.

Acknowledgements I would like to thank my Ph.D. student Satish Mulleti for technical discussions and for generating the Monte Carlo simulation results reported in this chapter.

References

1. C.E. Shannon, A mathematical theory of communication. *Bell Syst. Tech. J.* **27**, 623–656 (1948)
2. A.J. Jerri, The Shannon sampling theorem - its various extensions and applications: a tutorial review. *Proc. IEEE* **65**(11), 1565–1596 (1977)
3. M. Unser, Sampling-50 years after Shannon. *Proc. IEEE* **88**(4), 569–587 (2000)
4. A. Papoulis, Generalized sampling expansion. *IEEE Trans. Circuits Syst.* **24**, 652–654 (1977)
5. M. Unser, J. Zerubia, A generalized sampling theory without band-limiting constraints. *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.* **45**(8), 959–969 (1998)
6. M. Unser, A. Aldroubi, M. Eden, B-spline signal processing. I - theory. *IEEE Trans. Signal Process.* **41**(2), 821–833 (1993)

7. M. Unser, A. Aldroubi, M. Eden, B-spline signal processing. II - efficient design and applications. *IEEE Trans. Signal Process.* **41**(2), 834–848 (1993)
8. M. Unser, Splines: a perfect fit for signal and image processing. *IEEE Signal Process. Mag.* **16**(6), 22–38 (1999)
9. M. Unser, T. Blu, Cardinal exponential splines: Part I - theory and filtering algorithms. *IEEE Trans. Signal Process.* **53**(4), 1425–1438 (2005)
10. M. Unser, Cardinal exponential splines: Part II - think analog, act digital. *IEEE Trans. Signal Process.* 53(4), 1439–1449 (2005)
11. D.L. Donoho, Compressed sensing. *IEEE Trans. Inf. Theory* **52**(4), 1289–1306 (2006)
12. D.L. Donoho, For most large underdetermined systems of linear equations the minimal ℓ_1-norm solution is also the sparsest solution. *Commun. Pure Appl. Math.* **59**(6), 797–829 (2006)
13. E.J. Candès, M.B. Wakin, An introduction to compressive sampling. *IEEE Signal Process. Mag.* **25**(2), 21–30 (2008)
14. E.J. Candès, J. Romberg, T. Tao, Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information. *IEEE Trans. Inf. Theory* **52**(2), 489–509 (2006)
15. E.J. Candès, T. Tao, Near-optimal signal recovery from random projections: universal encoding strategies? *IEEE Trans. Inf. Theory* **52**(12), 5406–5425 (2006)
16. M. Vetterli, P. Marziliano, T. Blu, Sampling signals with finite rate of innovation. *IEEE Trans. Signal Process.* **50**(6), 1417–1428 (2002)
17. J. Berent, P.L. Dragotti, T. Blu, Sampling piecewise sinusoidal signals with finite rate of innovation methods. *IEEE Trans. Signal Process.* **58**(2), 613–625 (2010)
18. J. Kusuma, V.K. Goyal, On the accuracy and resolution of powersum-based sampling methods. *IEEE Trans. Signal Process.* **57**(1), 182–193 (2009)
19. R. Tur, Y.C. Eldar, Z. Friedman, Innovation rate sampling of pulse streams with application to ultrasound imaging. *IEEE Trans. Signal Process.* **59**(4), 1827–1842 (2011)
20. P.L. Dragotti, M. Vetterli, T. Blu, Sampling moments and reconstructing signals of finite rate of innovation: Shannon meets Strang-Fix. *IEEE Trans. Signal Process.* **55**(5), 1741–1757 (2007)
21. C.S. Seelamantula, M. Unser, A generalized sampling method for finite-rate-of-innovation-signal reconstruction. *IEEE Signal Process. Lett.* 813–816 (2008)
22. H. Olkkonen, J.T. Olkkonen, Measurement and reconstruction of impulse train by parallel exponential filters. *IEEE Signal Process. Lett.* **15**, 241–244 (2008)
23. G.R. DeProny, Essai experimental et analytique: sur les lois de la dilatabilité de fluides élastiques et sur celles de la force expansive de la vapeur de l'eau et de la vapeur de l'alcool, à différentes températures. *J. de l'Ecole Polytechnique* **1**(2), 24–76 (1795)
24. R.O. Schmidt, Multiple emitter location and signal parameter estimation. *IEEE Trans. Antennas Propag.* **34**(3), 276–280 (1986)
25. A. Paulraj, R. Roy, T. Kailath, A subspace rotation approach to signal parameter estimation. *Proc. IEEE* **74**(7), 1044–1046 (1986)
26. D.W. Tufts, R. Kumaresan, Estimation of frequencies of multiple sinusoids: making linear prediction perform like maximum likelihood. *Proc. IEEE* **70**(9), 975–989 (1982)
27. P. Stoica, R.L. Moses, *Introduction to Spectral Analysis* (Prentice Hall, Upper Saddle River, 1997)
28. S.M. Kay, *Modern Spectral Estimation—Theory and Application* (Prentice Hall, Englewood Cliffs, 1988)
29. J.A. Uriguen, T. Blu, P.L. Dragotti, FRI sampling with arbitrary kernels. *IEEE Trans. Signal Process.* **61**(21), 5310–5323 (2013)
30. S. Mulleti, S. Nagesh, R. Langoju, A. Patil, C.S. Seelamantula, Ultrasound image reconstruction using the finite-rate-of-innovation principles, in *Proceedings of the IEEE International Conference on Image Processing (ICIP)*, October 2014
31. T. Blu, H. Bay, M. Unser, A new high-resolution processing method for the deconvolution of optical coherence tomography signals, in *Proceedings of the First IEEE International Symposium on Biomedical Imaging: Macro to Nano (ISBI '02)*, vol. III, 7–10 July 2002, pp. 777–780

32. C.S. Seelamantula, S. Mulleti, Super-resolution reconstruction in frequency-domain optical-coherence tomography using the finite-rate-of-innovation principle. *IEEE Trans. Signal Process.* **62**(19), 5020–5029 (2014)
33. C. Vonesch, T. Blu, M. Unser, Generalized Daubechies wavelet families. *IEEE Trans. Signal Process.* **55**(9), 4415–4429 (2007)
34. A. Nathan, On sampling a function and its derivatives. *Inf. Control* **22**(2), 172–182 (1973)
35. Y.P. Lin, P.P. Vaidyanathan, Periodically nonuniform sampling of bandpass signals. *IEEE Trans. Circuits Syst. II, Analog Digit. Signal Process.* **45**(3), 340–351 (1998)
36. B. Lacaze, Equivalent circuits for the PNS2 sampling scheme. *IEEE Trans. Circuits Syst. I Regul. Pap.* **57**(11), 2904–2914 (2010)
37. K.F. Cheung, R.J. Marks, Imaging sampling below the Nyquist density without aliasing. *J. Opt. Soc. Am. A* **7**(1), 92–105 (1990)
38. P.P. Vaidyanathan, V.C. Liu, Efficient reconstruction of band-limited sequences from nonuniformly decimated versions by use of polyphase filter banks. *IEEE Trans. Acoust. Speech Signal Process.* **38**(11), 1927–1936 (1990)
39. B. Foster, C. Herley, Exact reconstruction from periodic nonuniform samples, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 2, May 1995, pp. 1452–1455
40. P. Feng, Y. Bresler, Spectrum-blind minimum-rate sampling and reconstruction of multiband signals, in *Proceedings IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, May 1996, pp. 1688–1691
41. H.J. Landau, Necessary density conditions for sampling and interpolation of certain entire functions. *Acta Math.* **117**(1), 37–52 (1967)
42. R. Venkataramani, Y. Bresler, Perfect reconstruction formulas and bounds on aliasing error in sub-Nyquist nonuniform sampling of multiband signals. *IEEE Trans. Inf. Theory* **46**(6), 2173–2183 (2000)
43. R. Venkataramani, Y. Bresler, Optimal sub-Nyquist nonuniform sampling and reconstruction for multiband signals. *IEEE Trans. Signal Process.* **49**(10), 2301–2313 (2001)
44. M. Mishali, Y.C. Eldar, From theory to practice: sub-Nyquist sampling of sparse wideband analog signals. *IEEE J. Sel. Topics Signal Process.* **4**(2), 375–391 (2010)
45. H. Akhtar, R. Kakarala, A methodology for evaluating accuracy of capacitive touch sensing grid patterns. *J. Disp. Technol.* **10**(8), 672–682 (2014)
46. J.A. Cadzow, Signal enhancement—a composite property mapping algorithm. *IEEE Trans. Acoust. Speech Signal Process.* **36**, 49–62 (1988)
47. T. Blu, P.-L. Dragotti, M. Vetterli, P. Marziliano, L. Coulot, Sparse sampling of signal innovations. *IEEE Signal Process. Mag.* **25**(2), 31–40 (2008)

Chapter 14
Digital Adaptive Calibration of Data Converters Using Independent Component Analysis

Yun Chiu

Abstract The theory and practice of applying a neural network model and learning algorithm—*Independent Component Analysis* (ICA)—to the online adaptive calibration of analog-to-digital converters (ADCs) is covered in this chapter. Exploiting the independence between the input signal and an injected pseudorandom bit sequence (PRBS), the technique attempts to blindly separate the two in the digital conversion output, and while doing so, an equivalent model of the ADC non-idealities is identified, resulting in the subsequent linearization of the conversion process. The ICA framework offers new signal-processing insights into the widely used correlation-based error-parameter identification method for the background calibration of multistage ADCs. In addition, it provides a useful technique to minimize the analog overhead associated with the calibration by simultaneously identifying multiple model parameters using a single PRBS, improving the efficiency and potentially the application regime of the online calibration approach for data converters.

14.1 Background and Introduction

CMOS technology advancement has inspired a trend in mixed-signal IC design to exploit the abundantly available on-chip digital processing power to help compensate or improve the analog circuit performance. In analog-to-digital converters (ADCs), the output bits naturally provide a digital means to infer the non-idealities of the constituent, imperfect analog building blocks. While this statement is nearly always true for any ADC types, those employing a multistage or multistep conversion architecture probably benefited the most from the advocated digital assistance.

The fundamental reason that a multistage or multistep ADC is more amenable to the digital treatment is that, once the conversion is partitioned into multiple circuit

Y. Chiu (✉)
Analog and Mixed-Signal Lab, Texas Analog Center of Excellence,
University of Texas at Dallas, Richardson, TX 75080, USA
e-mail: chiu.yun@ieee.org

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4_14

stages, a succinct—often in closed form—relationship between the analog input samples and the partial as well as the final digital output codes can be derived without resorting to lookup tables, which are cumbersome to use for high-resolution converters of 10 bits and above. Multistage/multistep ADCs commonly encountered in practice are cyclic ADC, pipelined ADC, successive approximation register (SAR) ADC, and multistage sigma–delta modulator (MASH). A brief overview of some of these ADC architectures will be given in this section.

14.1.1 Overview of Multistage ADC Architectures

Analog-to-digital conversion is usually composed of a tandem of sampling and quantization operations. The sampling action discretizes a continuous-time analog signal into samples, stored as a voltage, charge, or current signal in a sample-and-hold (S/H) circuit. The analog samples are then quantized by another circuit, the quantizer, to obtain a fixed-point, digital representation of the analog value of the samples. ADCs are categorized usually by the architecture of the quantizer employed, as in practice the quantizer often occupies most of the silicon area and consumes most of the power of the ADC (albeit the S/H also sets a fundamental limit on the overall speed and accuracy performance).

The dominant analog tradeoff in ADC design is probably between the sample rate (or, loosely speaking, the ADC speed) and the resolution (or, more precisely, the signal-to-noise plus distortion ratio or SNDR). Fast ADC architectures such as the flash ADC often produce low resolution in the range of 4–8 bits. To achieve an accuracy of 10 bits and above, a multistage architecture is often the choice in practice, in which the quantization operation is divided into multiple circuit stages, each responsible for resolving a small number of bits. Once S/H circuits are inserted in between the stages, the operation can be pipelined, resulting in the so-called pipelined ADC. In the limiting case, a pipelined ADC resolves only one bit in each stage while maintaining a simple analog circuit structure such that the sample rate or conversion speed can be high.

14.1.1.1 A Two-Stage Pipelined ADC

The pipelined ADC architecture can be explained with the two-stage example shown in Fig. 14.1, in which a one-bit first stage is connected with a backend 11-bit ADC (assumed ideal for the sake of simplicity) through a feedback amplifier termed the residue amplifier (RA). In normal operation, the sub-ADC of the first stage resolves one bit ($d_1 = 0$ or 1), the sub-DAC converts it back into the analog form and subtracts it from the sampled-and-held input, and lastly the RA produces an amplified version of the difference, i.e., the *residue*, of the unresolved 11-bit information and passes it to the backend ADC for further processing. After the residue is sampled by the backend ADC, the first stage is free to accept another

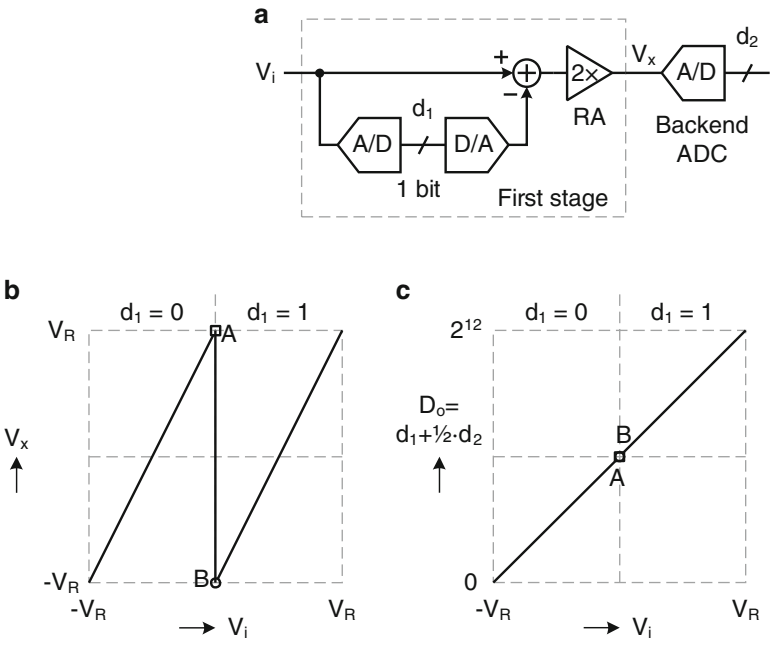


Fig. 14.1 A 12-bit, two-stage ADC example: (*a*) block diagram, (*b*) interstage residue transfer curve, and (*c*) overall ADC curve

input sample while the backend is working on resolving the 11-bit information from the residue. This is how the conversion throughput can be improved at the cost of latency, i.e., *pipeline*.

The residue transfer curve of the first stage is illustrated in Fig. 14.1*b*. The most-significant bit (MSB) transition point is set by the sub-ADC (in this case, just one comparator with a threshold of zero, i.e., the comparator output flips at $V_i = 0$). To maintain the same signal swing for both of the stages, the RA must produce a residue gain of 2× exactly. Note that this interstage gain set by the RA needs to be very accurate to guarantee the linearity of the ADC, often characterized as the integral nonlinearity (INL) and the differential nonlinearity (DNL), is commensurate with its resolution. For example, let us examine the points A and B, shown in Fig. 14.1*b*, on the opposite sides of the MSB transition point. While the two points are located on different segments of the residue transfer curve, since they correspond to the same analog input ($V_i = 0$), they must resolve to the same output code if the ADC is ideal. When the residue gain is exactly 2×, the two points will indeed rejoin each other at the midpoint of the overall conversion curve, illustrated in Fig. 14.1*c*, once the two partial decision codes (d_1 and d_2) from the two stages are weighted and combined. In this example, the weighting factor is unity for d_1 and ½ for d_2.

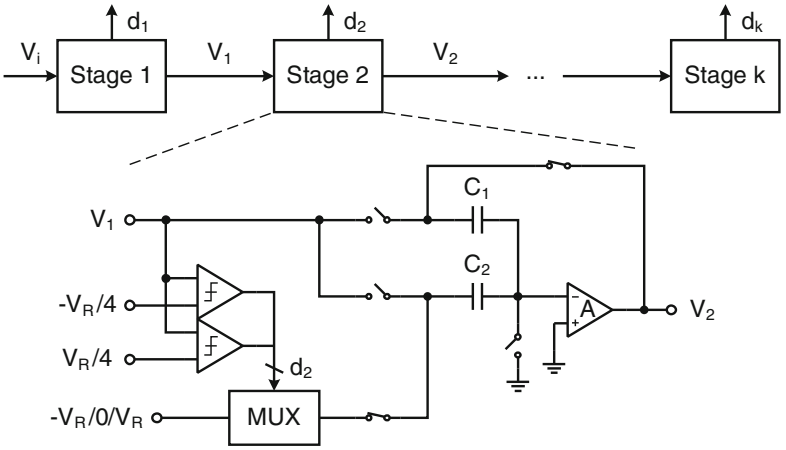


Fig. 14.2 A multistage pipelined ADC

14.1.1.2 Pipelined ADC and Cyclic ADC

Once we have the two-stage example laid out, it is not difficult to generalize the concept to a multistage pipelined ADC, which can be simply conceived by imaging that the 11-bit backend ADC is constructed in the same way with a two-stage architecture; and then the 10-bit backend can also be constructed the same way; and so on... At the end, we will end up with a pipelined multistage ADC that has N stages, a total resolution of N bits, and each stage yields one bit. This ADC is depicted in Fig. 14.2, wherein the switched-capacitor (SC) circuit realization of a typical pipeline stage is also rendered [1]. Note that the sub-ADC consists of two comparators with two threshold voltages, $\pm V_R/4$, respectively, instead of one comparator with a threshold of zero as introduced before. This has to do with the internal redundancy of the pipelined ADC for circuit particularly comparator offset tolerance. Redundancy will be introduced in Section 14.1.2.

In some applications where throughput is not the primary performance target, large savings on the hardware cost of a pipelined ADC can be obtained by removing all the conversion stages except the first one and iterating the conversion process around it. This architecture is termed the *cyclic* or *algorithmic* ADC [2]. A block diagram is shown in Fig. 14.3.

Notice that no matter for the pipelined or cyclic ADC the most critical circuit accuracy concern always lies with the first conversion stage and this concern diminishes toward the LSB stage as more bits are resolved. In practice, all pipelined stages are not designed with the same specs and power/area consumptions. Considering the relaxed accuracy requirement, the later stages are often trimmed down for cost reductions. However, we note that this is not possible for the cyclic ADC.

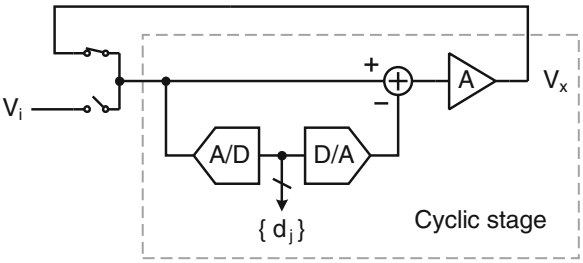


Fig. 14.3 A cyclic ADC

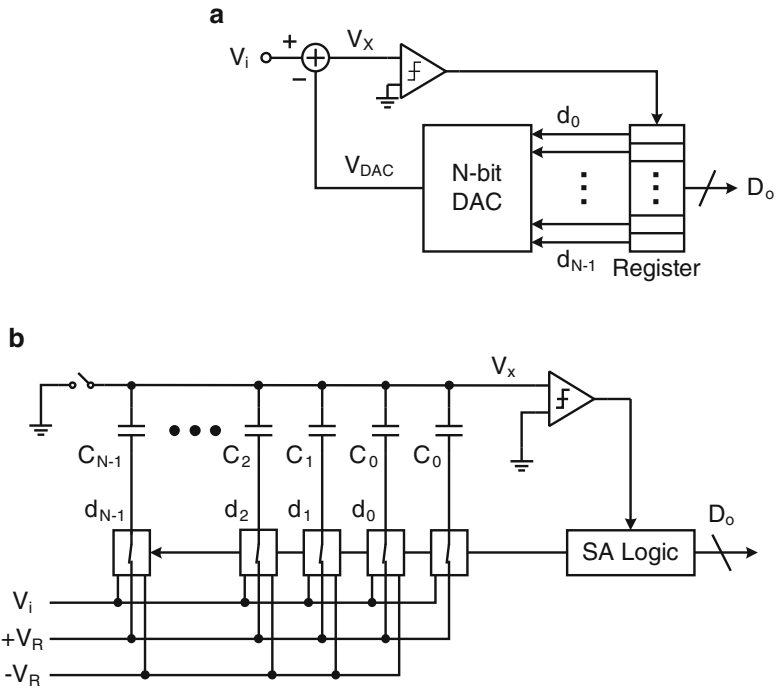


Fig. 14.4 A SAR ADC: (*a*) block diagram and (*b*) typical SC realization

14.1.1.3 SAR ADC

Successive-approximation ADC is another multistep converter that is very efficient in hardware construction, similar to the cyclic ADC. The operation principle of the SAR ADC can be explained as follows. As shown in Fig. 14.4*a*, once the input analog sample is acquired, the SAR ADC employs a *binary search* algorithm to determine the digital code that best approximates the analog sample. This is done sequentially for all the bits starting with the MSB. Taking the MSB cycle for example, the *N*-bit DAC first produces an analog level corresponding to the midpoint

of the ADC input range; the held analog sample is then compared to this DAC output by a comparator; the decision (1 or 0) indicates whether the input resides in the upper or lower half of the conversion range. Once the MSB is resolved, the procedure moves on to determine the second significant bit, i.e., the DAC will produce another analog level of ¼ or ¾ of the conversion range dependent on whether the MSB is 0 or 1, respectively; this DAC output is again compared to the analog sample and its relative location is encoded (0 for below and 1 for above); thus the second bit is determined. The procedure repeats itself until all the N bits are resolved.

Compared to the cyclic ADC, it is obvious that the SAR operates similarly except that the residue production or amplification is absent during its bit cycles. A typical SC circuit realization of the SAR ADC is rendered in Fig. 14.4*b*. The summing-node (i.e., node V_x) subtraction operation is realized by the SC DAC using a process called *charge redistribution*, which can be analyzed by noting that the total charge on the high-impedance summing node must remain constant during the SAR bit cycles. The threshold of the comparator is always zero, i.e., the comparator is simply a zero-crossing detector.

14.1.2 Built-in Redundancy

The one-bit residue curve depicted in Fig. 14.1*b* is probably easier for conceptual understanding than for practical circuit implementation. The problem can be appreciated by examining the same residue transfer curve, reproduced in Fig. 14.5*a*, representing the jth stage of a pipelined ADC (e.g., the one shown in Fig. 14.2) in general. This stage receives an input residue signal V_{j-1}, resolves a digital code d_j, and produces an output residue V_j. Using a bipolar representation, the comparator threshold is ideally placed at $V_{j-1} = 0$ and the residue curve consists of two parallel segments with an ideal slope of two. Now imagine that the comparator displays an offset V_{os} and the decision threshold shifts away from the center point; in this case, either an overflow (for a positive V_{os}) or an underflow (for a negative V_{os}) error will be experienced by the backend ADC because its resolvable range is limited to $[-V_R, +V_R]$.[1]

Because circuit offset is inevitable in practice, a multistage pipelined ADC is almost always realized with built-in redundancy. A well-known 1.5-bit-per-stage architecture is shown in Fig. 14.2 [1, 2]. The transformation from a 1-bit topology to a 1.5-bit one is illustrated with the diagrams shown in Fig. 14.5*a–d*. Note that in the 1.5-bit architecture a small comparator offset will not cause any overflow or

[1]One can imagine this by sliding the threshold at zero to the left or right while confining it in between the two parallel residue segments with a slope of two.

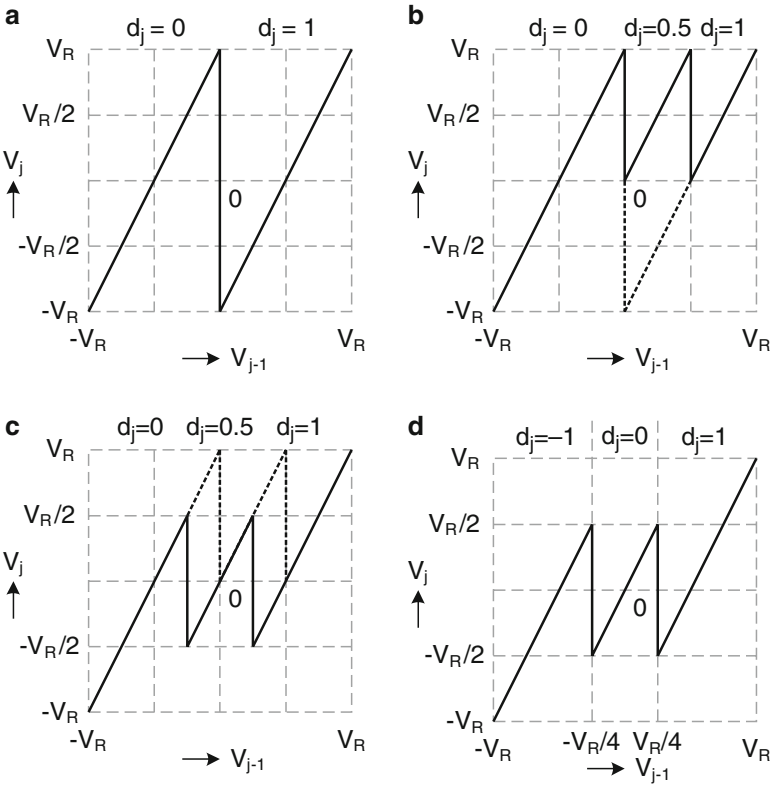


Fig. 14.5 Transformation of a 1-bit transfer curve without redundancy to a 1.5-bit one with redundancy: (*a*) original 1-bit residue curve, (*b*) with an extra comparator added at $V_R/2$, resolving three digital levels, (*c*) both comparator thresholds shifted by $-V_R/4$ to yield a symmetrical curve, and (*d*) the resulting 1.5-bit residue curve. Note that the digital bits representing the three levels in (*d*) are eventually all scaled by a factor of 2 and offset by -1

underflow error at either of the two MSB transition points (now located at $\pm V_R/4$), as the maximum values of the residue at these points are only half range ideally.[2]

Redundancy in SAR ADC works in a similar way, and it is also proven essential to the proper operation of the SAR in presence of circuit non-idealities. We will skip the discussion of SAR redundancy here. Interested readers are referred to [3–5] for more readings.

[2] As quantization can be understood as division, the 1.5-bit topology is actually a realization of the Sweeny–Robertson–Tocher (SRT) fast division algorithm well known in computer arithmetic.

14.1.3 Error Model of Multistage ADC

As mentioned earlier, it is often possible to derive a closed-form expression to relate the analog input and the digital output of a multistage ADC due to the explicit residue transfer in between the stages. When represented in terms of the circuit parameters of the ADC, the closed-form relationship thus describes a way by which the digital output can be remapped or corrected to compensate certain non-idealities of the analog-to-digital conversion process. Such a relationship is termed the error model of an ADC. In this section, we formulate this model using the examples of a simple two-stage ADC and a multistage pipelined ADC.

14.1.3.1 A Two-Stage Example of Interstage Gain Error

Let us reconsider the simple two-stage pipelined ADC example discussed in Section 14.1.1.1, but this time with the first stage replaced by a realistic 1.5-bit architecture. Again, the overall resolution is 12 bits and the backend ADC is assumed ideal. The circuit diagram is redrawn in Fig. 14.6*a*.

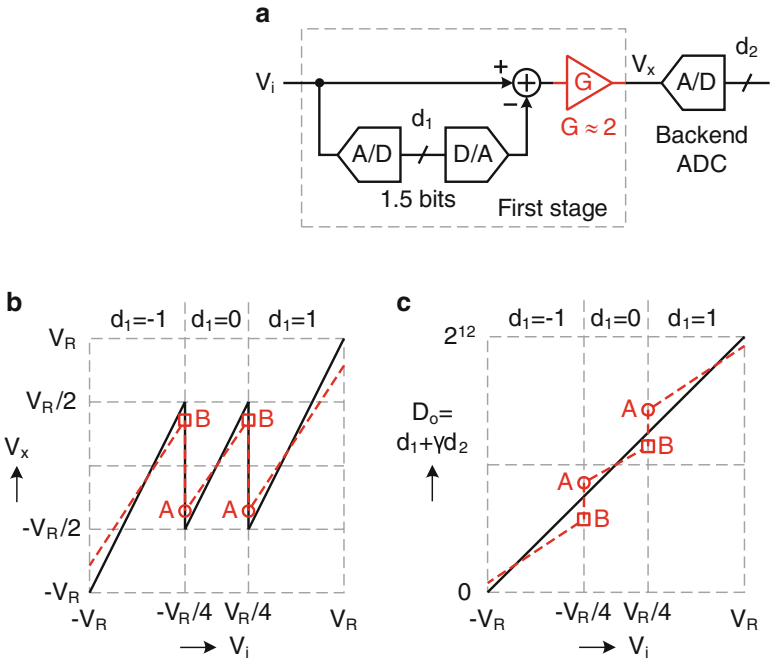


Fig. 14.6 A 12-bit, two-stage ADC example: (*a*) block diagram, (*b*) interstage residue transfer curve assuming a 1.5-bit architecture, and (*c*) the overall ADC curve

Due to the interstage built-in redundancy, the ADC operation can tolerate very large comparator offsets. For example, as indicated by the 1.5-bit residue curve in Fig. 14.6*b*, either comparator threshold can vary by as much as $\pm V_R/4$ without incurring an out-of-range residue. However, the offset tolerance does not relax the 2× interstage gain accuracy when we attempt to deliver a precision analog residue signal (V_x in Fig. 14.6*a*) to the backend ADC for further quantization. The residue transfer function in this case can be derived as [1]

$$V_x(n) = G \cdot [V_i(n) - V_R \cdot d_1(n)], \tag{14.1}$$

where n is the sample index, d_1 is the digital code resolved by the first stage, and ideally $G = 2$. In a 12-bit ADC, this residue signal needs to be as accurate as 11-bit, or $2^{-11} \approx 0.05$ %. Figure 14.6*b* displays examples of the residue curve when such an accuracy is satisfied (the solid line) as well as when it is not (the dashed line). Ideally, if V_x is resolved to a digital code d_2 by the backend stage,

$$V_x(n) = V_R \cdot [d_2(n) + QN], \tag{14.2}$$

where QN is the quantization noise of the backend stage, (14.1) can be reversed to assemble the final digitization outcome D_o,

$$D_o(n) = \left\lfloor \frac{V_i(n)}{V_R} \right\rfloor = d_1(n) + \gamma(G) \cdot d_2(n), \tag{14.3}$$

where $\gamma(\mathrm{G}) = G^{-1}$ is the radix (weight) for d_2. Note that the exact value of $\gamma(\mathrm{G})$ depends on the interstage gain G. The value of γ is ½ in this example when the ADC is ideal. The final output code D_o is plotted in Fig. 14.6*c* to illustrate the ideal conversion curve (the solid line) of the ADC in contrast to the case when an interstage gain error occurs (the dashed line)—the digital codes between points A and B in Fig. 14.6*c* never appear (i.e., the vertical gap between A and B). This type of conversion error is known as the missing code.

In solid-state technology, the RA is often realized by a feedback op-amp with a high open-loop gain and a few precisely matched capacitors or resistors. It can be shown that a 0.05 % residue accuracy leads to a 0.05 % matching accuracy of the passive elements setting the closed-loop gain and at least a 72-dB open-loop gain of the op-amp (accounting for a feedback factor of 0.5 for a closed-loop gain of 2), which are difficult to achieve in monolithic forms in fabrication technologies such as CMOS. Particularly, the simultaneous high-gain and wide-bandwidth requirements for the op-amp constitute a keen challenge for the realization of high-performance pipelined converters.

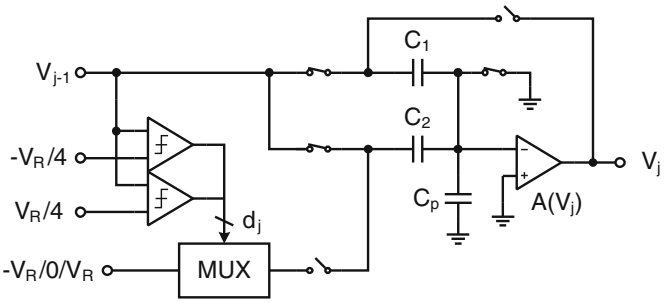


Fig. 14.7 Circuit diagram of a 1.5-bit SC pipelined ADC stage

14.1.3.2 A General Multistage Gain Error Model

We further examine the interstage residue transfer function in terms of a realistic 1.5-bit SC RA, with its circuit schematics reproduced in Fig. 14.7. Note that we also included a summing-node parasitic capacitor C_p and expressed the open-loop gain of the op-amp as a function of its output voltage (i.e., accounting for the static op-amp nonlinearity). The input and output residue signals can be related with the following expression [6]:

$$V_{j-1} = V_R \cdot d_j \cdot \left(\frac{C_2}{C_1 + C_2}\right) + V_j \cdot \left[\frac{C_1}{C_1 + C_2} + \frac{C_p + C_1 + C_2}{A\left(V_j\right) \cdot (C_1 + C_2)}\right], \quad (14.4)$$

where $d_j \in \{-1,\ 0,\ 1\}$ is the sub-ADC decision, $A(V_j)$ is the open-loop gain of the op-amp, and C_1 and C_2 are the sampling/DAC capacitors. In the ideal case, i.e., $C_1 = C_2$ and $A = \infty$, we have

$$V_{j-1} = V_R \cdot d_j \cdot \alpha_j + V_j \cdot \beta_j, \quad (14.5)$$

where $\alpha_j = \beta_j = ½$. If we divide both sides of (14.5) by V_R and apply it to the first three ADC stages, we have

$$\begin{aligned} D_o &= \left\lfloor \frac{V_i}{V_R} \right\rfloor \\ &= d_1 \cdot (\alpha_1) + d_2 \cdot (\alpha_2\beta_1) + d_3 \cdot (\alpha_3\beta_2\beta_1) + \ldots \\ &= d_1 \cdot \frac{1}{2} + d_2 \cdot \frac{1}{4} + d_3 \cdot \frac{1}{8} + \ldots \end{aligned} \quad (14.6)$$

Equation (14.6) essentially formulates the final ADC digital output code in the ideal case. When the capacitors are not well matched and/or the op-amp gain is finite, (14.6) is still valid as long as we neglect the nonlinearity in $A(V_j)$, and we have

$$D_o = d_1 \cdot \gamma_1 + d_2 \cdot \gamma_2 + d_3 \cdot \gamma_3 + \ldots, \tag{14.7}$$

i.e., once we know the exact values of γ_1, $\gamma_2 \ldots$, an ideal conversion still yields [6]. Note that the weighted sum of (14.7) is performed on d_1, $d_2 \ldots$ only, i.e., the operation is totally *digital*.

The formulation of (14.4) can also be easily generalized to multibit-per-stage pipelined ADC architectures [7, 8]. Furthermore, when the nonlinearity of $A(V_j)$ is included, (14.5) can be rewritten as

$$\begin{aligned} V_{j-1} &= V_R \cdot d_j \cdot \alpha_j + f\left(V_j\right) \\ &\approx V_R \cdot d_j \cdot \alpha_j + \sum_m V_j^m \cdot \beta_{j,m}, \end{aligned} \tag{14.8}$$

where in the last step of (14.8) we replaced the nonlinear function $f(V_j)$ by a power series approximation considering the fact that op-amp circuits are mostly weakly nonlinear. The value of m usually ranges from 3 to 5 in typical applications. Also, in modern fabrication processes, the metal–metal and poly–poly capacitors are usually linear up to 14 bits. Beyond this, capacitor voltage coefficients can also be included in the calibration [9].

The bit weights γ_1, $\gamma_2 \ldots$ in (14.7) are sometimes referred to as the conversion radices. In the 1.5-bit example, their ideal values are ½, ¼... Interstage residue gain error will result in nonideal bit weights, manifested as discontinuities in the conversion curve. In an ADC with built-in redundancy, the discontinuity can take one of two forms, a vertical nonoverlapping gap (i.e., missing code) or an overlapping gap (i.e., non-monotonic code). The gaps in the dashed conversion curve illustrated in Fig. 14.6*c* correspond to the missing-code case.

Once an error model is available, the remaining work is to identify the model parameters, i.e., γ_1, $\gamma_2 \ldots$ in (14.7), α_j and $\beta_{j,m}$ in (14.8), etc., to certain accuracy level to suit a particular application. While the error models employed do not vary much from work to work, the error-parameter identification process certainly embraces much more varieties—in fact, a digital calibration technique is often categorized according to the exact parameter identification procedure it employs.

14.2 Online ADC Calibration with PRBS Injection

Although digital calibration of data converters can be traced back to 1980s, utilizing simple, two-level test signals for online calibration was probably first reported in [10], in which a square-wave dither was injected into the first-stage quantizer input of a 2-1 MASH sigma–delta modulator to eliminate the quantization noise leakage. A square-wave dither was also later utilized in [11] to calibrate the multibit DAC mismatch errors in a sigma–delta ADC. The square wave was revised to a pseudorandom bit sequence (PRBS) in [12] and [13]. Around the same time, PRBS injection was also utilized in the gain-mismatch calibration of time-interleaved

pipelined ADCs [14]. Sub-DAC PRBS injection was reported in [15] and [16] to correct the interstage residue gain error in pipelined ADCs. The method in [16] was to seek the correct bit weight, and a mixed-signal correction was used in [15], where the ADC reference voltage was digitally trimmed. The technique was also applied to treating the capacitor mismatch errors in a multibit DAC of a pipelined ADC in [17] using a multi-PRBS injection scheme, with each PRBS responsible for identifying the mismatch coefficient of one capacitor. The technique was adapted and further improved in [18–23]. A sub-ADC dither was exploited to calibrate the interstage gain error and/or nonlinearity in pipelined [24, 25] and algorithmic [26] ADCs. Lately, the sub-DAC injection method was also augmented (with a multi-PRBS injection) to treating the residue-amplifier nonlinearity in pipelined ADCs [27, 28].

14.2.1 Test-Signal Injection and Dither

The purpose of test signal injection is to let it traverse the conversion path of the ADC and observe the digital outcome that may bear information of circuit non-idealities. However, when operating in the online (background) mode, the coexistence of the test signal and the normal input signal may result in some undesirable interference, restricting the choice of the test signal and its generation/injection method. In practice, a single PRBS is often employed for such a purpose as it can be realized, for example, by a small capacitor switching on and off randomly depending on the value of a digital bit. In addition, a pseudorandom, two-level test signal has a close-to-white spectrum, which reduces its chance of being cluttered by any narrowband input, resulting in improved robustness of treatment.

PRBS injection—a.k.a. dither—has been used as a dynamic-element matching (DEM) technique to improve the spectral performance of converters. It works by disrupting a deterministic, nonlinear conversion process with a PRBS injected into either the signal path [29] or the sub-ADC path [30] of a multistage ADC. While the PRBS needs to be removed subsequently in the digital domain in the former case, it can be treated as dynamic comparator offset and safely neglected in the latter one due to the built-in redundancy of the converter.

The sub-ADC dither can be explained as follows. A small PRBS is injected into the first-stage sub-ADC path (i.e., the comparators) of a 1.5-bit, SC pipelined ADC, as the one shown in Fig. 14.2. The net effect is that the two comparator thresholds will shift slightly to the left or right dependent on the PRBS value. The resulting residue curves of the first stage and the overall ADC curve are illustrated in Fig. 14.8*a* and Fig. 14.8*b*, respectively. Dependent on the PRBS value, the ADC will randomly pick one of the two (redundant) residue paths, i.e., the solid and dashed paths shown in Fig. 14.8, whenever an input falls into the shaded regions. When the ADC is ideal, the final digital outcome will be identical no matter which path the conversion takes place. In contrast, once the residue production suffers from a gain error, a gap δ_1 exists at each of the two MSB transition points, shown in

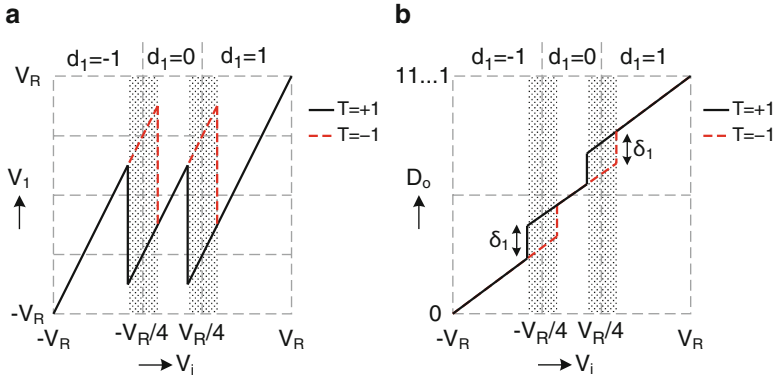


Fig. 14.8 Sub-ADC dither: (*a*) residue curve of the first stage with PRBS injection and (*b*) overall ADC curve with discontinuities (δ_1) at MSB transition points due to a residue gain error

Fig. 14.8*b*. The randomization essentially bounces the erroneous (and deterministic) conversion process between two trajectories, thus making the ADC conversion error also appear to be *random*. Although the absolute conversion error is not reduced at all, the deterministic error structure is randomized, leading to improved spectral performance of the ADC. This is the essence of dither.

14.2.2 Sub-ADC vs. Sub-DAC PRBS Injection

Albeit dither does not eliminate conversion errors, one can exploit the result of dither to identify the non-idealities of a conversion process. For example, in Fig. 14.8, we need to identify the value of the gap δ_1 at the MSB transition points. We can employ a PRBS injection to achieve this goal. Two injection methods are often encountered in practice, i.e., the sub-ADC injection method and the sub-DAC injection method, sketched in Fig. 14.9*a* and Fig. 14.9*b*, respectively.

14.2.2.1 Sub-ADC Injection

Suppose that an input sample falls into the shaded regions shown in Fig. 14.8*a*; in the sub-ADC injection case, depending on the value of the PRBS, the conversion will follow either the solid or the dashed curve and resolve to two digital codes differing by δ_1 on average, as indicated in Fig. 14.8*b*. Correlating the ADC output code with the PRBS (thus the name correlation-based calibration) can effectively measure the magnitude of δ_1,

$$\overline{D_o \cdot T} = \frac{1}{2}\delta_1 \cdot \Pr\left(V_i \text{ falls into the shaded regions}\right), \tag{14.9}$$

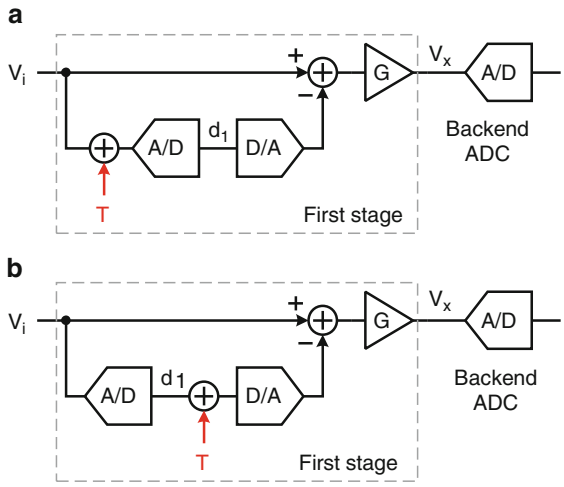


Fig. 14.9 Common PRBS injection methods in pipelined ADCs: (*a*) sub-ADC injection and (*b*) sub-DAC injection

where T is the injected PRBS and $\Pr(\cdot)$ is the probability of V_i falling into the shaded regions. As the input statistics is unknown, a gradient-descent (iterative) algorithm can be used to estimate δ_1,

$$\Delta\delta_1 = -\mu \cdot D_o \cdot T, \tag{14.10}$$

where $\Delta\delta_1$ is the incremental update for δ_1 and μ is a step size. Based on our analysis, when (14.10) converges after removing the discontinuities in between the segments of the ADC curve using the learned value of δ_1, the linearity of the conversion process can be restored.

The advantage of the sub-ADC injection method is that the PRBS does not need to be removed from D_o during the calibration. In addition, the exact size of injection (i.e., the width of the shaded regions in Fig. 14.8) bears no consequence to the learning accuracy. However, an increase of the sub-ADC resolution (usually a factor of two) is often necessary to accommodate the comparator threshold shift as a result of injection [24, 26].

14.2.2.2 Sub-DAC Injection

Alternatively, a PRBS can be injected into the sub-DAC to identify the error parameters, without needing the input to fall into the shaded regions [15, 16]. However, in this case, the PRBS needs to be accurately removed from D_o. Thus the exact size of the injection circuit element, e.g., a capacitor, must be known. In addition, to minimize the input dynamic range loss due to injection, a signal-dependent PRBS

injection method can be employed [23]. Lastly, while the injection element must be matched to the other DAC elements, it can be argued that the single (and small) injection element leads to a minimum hardware cost of accommodating the test signal relative to the sub-ADC case discussed above.

Lastly, compared to the online adaptive calibration approaches based on split-ADC [31–35] or offset double conversion [5, 36, 37], the runtime of the correlation in (14.9) tends to be much longer in order to obtain a clear observation of the error parameters, especially in the presence of a large, potentially busy input signal [31]. In spite of the slow convergence, the analog simplicity of the PRBS injection method (especially in the sub-DAC case) has popularized this calibration technique in recent years.

14.3 ICA-Based Digital Online Calibration

In our examination of the sub-ADC dither technique in Section 14.2.1, we observed that when the ADC is ideal, dither will produce no noticeable difference in the final output due to the built-in redundancy. This observation delivers a very useful intuition about online ADC calibration, i.e., one can potentially obtain information about the non-idealities of the conversion process by observing the correlation between the PRBS and the normal output of the ADC—when the ADC is ideal, the two are obviously independent; conversely, the ADC output cannot be free of residual PRBS information when the ADC is not fully linearized. Therefore, a technique can be devised to separate the two in the digital output, and while doing so, an error model of the ADC can be identified for digital calibration.

We will see in this section that a new approach for online digital calibration of multistage ADCs is derived based on the above intuition. In addition, the technique closely resonates with the neural network model and learning algorithm *Independent Component Analysis* (ICA), which offers new signal-processing insights into the widely used parameter identification methods based on PRBS injection. It also helps minimize the analog overhead associated with the calibration by simultaneously identifying multiple model parameters with a single PRBS.

14.3.1 Independent Component Analysis

ICA is a statistical signal-processing technique that finds a wide range of applications in signal processing, neural computing, statistics, communications, and finance. Imagine that two persons speak simultaneously at a cocktail party as shown in Fig. 14.10. There are two microphones at different locations in the room that record the speeches, which we can denote by $x_1(t)$ and $x_2(t)$. Each of the recorded signals is a weighted sum of the original speeches denoted by $s_1(t)$ and $s_2(t)$, i.e.,

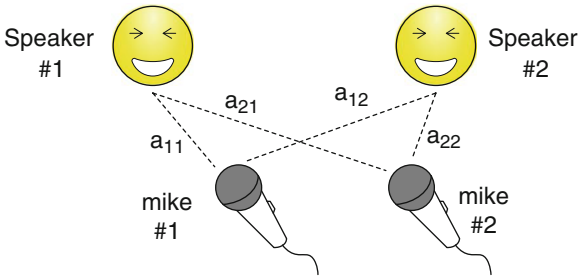


Fig. 14.10 Cocktail-party problem

$$\begin{aligned} x_1(t) &= a_{11} \cdot s_1(t) + a_{12} \cdot s_2(t), \\ x_2(t) &= a_{21} \cdot s_1(t) + a_{22} \cdot s_2(t), \end{aligned} \quad \text{or} \quad \begin{bmatrix} x_1(t) \\ x_2(t) \end{bmatrix} = \boldsymbol{A} \begin{bmatrix} s_1(t) \\ s_2(t) \end{bmatrix}, \tag{14.11}$$

where a_{11}, a_{12}, a_{21}, and a_{22} are the weighting parameters depending on the distances of the microphones to the speakers. It could be very useful if the original speech signals $s_1(t)$ and $s_2(t)$ can be estimated using only the recorded signals $x_1(t)$ and $x_2(t)$. When the mixing matrix $\boldsymbol{A} = [a_{11}\ a_{12};\ a_{21}\ a_{22}]$ is known, solving (14.11) is almost trivial; however, the problem becomes considerably more difficult without a priori knowledge of $\boldsymbol{A}$. This is a classic example of the *Blind Source Separation* (BSS) problem [38–50].

One approach to solving the BSS problem is to use some information on the statistical properties of $s_1(t)$ and $s_2(t)$, in which the signals are assumed statistically independent. This is not an unrealistic assumption in many cases, and it need not be exactly true in practice. One such technique is called the *Independent Component Analysis* [38–50], which can be defined as follows: ICA of the random vector $\boldsymbol{x}$ consists of finding a linear transform $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$ so that the components y_i are as independent as possible, in the sense of maximizing some objective function $C(y_1, y_2, \ldots)$ that measures independence. The first attempt in line with the ICA principle is the nonlinear de-correlation algorithm reported by Hérault–Jutten in 1986 [41], which can be illustrated for example by the two-cell structure, shown in Fig. 14.11, to separate the two speeches in the cocktail-party problem. In this case, $\boldsymbol{W} = [1\ -c_{12};\ -c_{21}\ 1]$. It can be shown that if a gradient-descent method is employed of the form,

$$\begin{aligned} \tfrac{dc_{12}}{dt} &= \mu \cdot g_1(y_1)\, g_2(y_2), \\ \tfrac{dc_{21}}{dt} &= \mu \cdot g_2(y_1)\, g_1(y_2), \end{aligned} \tag{14.12}$$

for updating the two synaptic weights c_{12} and c_{21}, the outcomes of the H-J network—$y_1(t)$ and $y_2(t)$—will converge within a scaling constant to $s_1(t)$ and $s_2(t)$. In (14.12), $g_1(\cdot)$ and $g_2(\cdot)$ are some odd nonlinear functions satisfying certain constraints [41] and μ is the learning rate of the weights.

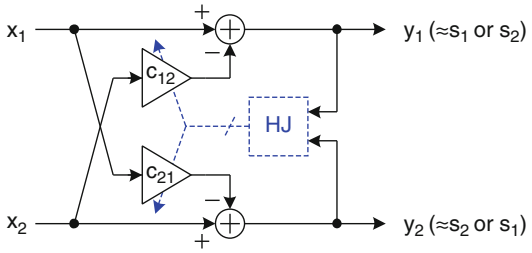


Fig. 14.11 Two-cell example of the H-J network

14.3.2 Two-Stage ADC Example Revisited

A question arises as to how ICA can be useful to the design of converter circuits. To identify this, let us revisit the interstage gain error problem of the two-stage pipelined ADC, previously studied in Section 14.1.3.1.

The block diagram of the ADC is shown in Fig. 14.6. Suppose that a -10 % residue gain error is the consequence of some poorly matched capacitors or an insufficient op-amp gain. The resulting gain $G = 1.8$ leads to roughly 10 % of the digital codes missing (the segments between A and B in Fig. 14.6*c*). However, if the imperfect G is known precisely, we see that a simple scaling of the digital code d_2 by $G_{ideal}/G_{actual} = 2/1.8$ can perfectly recover the ideal conversion curve, as illustrated by the solid line in Fig. 14.6*c*. In addition, this operation can be accomplished with a simple digital multiplier. The remaining question, obviously, is how to determine the value of G precisely, given that it varies from chip to chip and process to process—this is where the ICA technique enters the picture.

We note that although an interstage gain error stemming from capacitor mismatch may not drift over time, the op-amp gain error, in contrast, can be supply voltage, temperature, and circuit age dependent, thus necessitating an online treatment. Also, as introduced in Section 14.1.3.2, a low op-amp gain is often accompanied by nonlinearities that invalidate the simple piecewise-linear model of the ADC curves depicted in Fig. 14.6*c*. A nonlinear treatment is necessary and will be covered in Section 14.3.3.3.

14.3.2.1 ICA Calibration of Interstage Gain Error

Let us superpose the input signal V_i with an independent PRBS, T, at the ADC input as shown in Fig. 14.12. Now the summation of V_i and T traverses the analog signal path of the ADC and gets converted into digital code. Neglecting quantization noise, (14.3) can be rewritten as[3]

[3]For simplicity and better clarity, we will drop the constant scaling factor V_R in all the equations from this point onward.

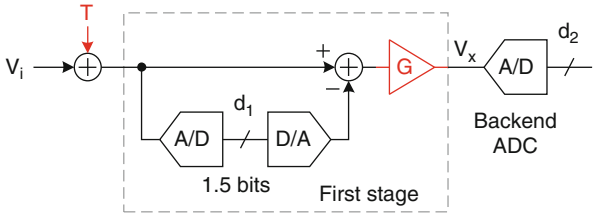


Fig. 14.12 Input PRBS injection in the two-stage ADC studied in Fig. 14.6

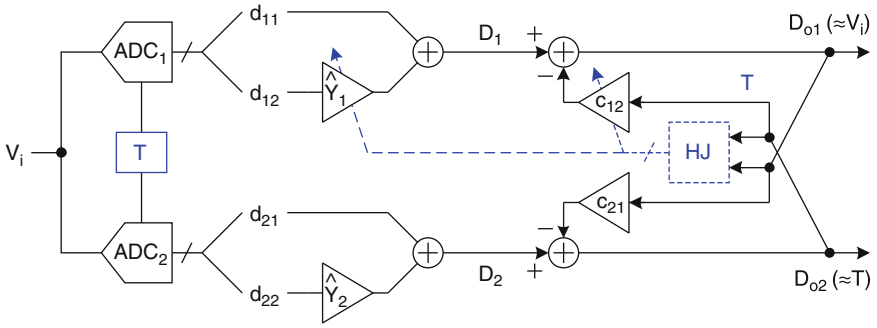


Fig. 14.13 H-J algorithm applied to the two-stage ADC calibration. The second ADC path (the lower half) to recover T does not need to be implemented as T is a known signal

$$V_i(n) + T(n) = d_1(n) + \gamma \cdot d_2(n). \tag{14.13}$$

Suppose that the ADC in Fig. 14.12 is duplicated to form a second path to digitize the sum of V_i and T as shown in Fig. 14.13. With the two ADCs and considering potential analog mismatch errors in between the two otherwise identical paths, we have

$$\begin{aligned} a_{11} \cdot V_i + a_{12} \cdot T &= d_{11} + \gamma_1 \cdot d_{12}, \\ a_{21} \cdot V_i + a_{22} \cdot T &= d_{21} + \gamma_2 \cdot d_{22}, \end{aligned} \tag{14.14}$$

where $[a_{11}\ a_{12}; a_{21}\ a_{22}]$ is the mismatch matrix for V_i and T, γ_1 and γ_2 are the radices, and $d_{11}, \ldots d_{22}$ are the stage decision codes of the two ADCs. To relate this to the cocktail-party problem, we note that

1. V_i and T are s_1 and s_2 of the original two speeches, respectively;
2. The mismatch matrix is essentially the mixing matrix;
3. V_i, A, γ_1, and γ_2 are unknown while T, d_{11}, d_{12}, d_{21}, and d_{22} are known;
4. The hypothesis is that V_i and T are independent.

One key difference between the two problems is that, while s_2 is unknown and to be recovered in the speech problem, T is a known signal in the ADC problem but the radices γ_1 and γ_2 are unknown. Thus our goal is to revise the ICA algorithm to identify γ_1 and γ_2 instead of to recover T.

Figure 14.13 depicts how the H-J algorithm can be adapted to recover V_i and T in a straightforward way. Under usual circumstances, the two outcomes D_{o1} and D_{o2} will converge within a scaling constant to V_i and T. However, a closer examination reveals that the duplicate ADC path in Fig. 14.13 does not need to be implemented at all since it is unnecessary to recover a known signal T. In addition, as this in turn means that c_{21} and γ_2 do not need to be identified anymore, the second gradient in (14.12) is reassigned to identify γ_1, which yields

$$\begin{aligned} c_{12}(n+1) &= c_{12}(n) - \mu_c \cdot g_1(D_{o1})\, g_2(T), \\ \widehat{\gamma}_1(n+1) &= \widehat{\gamma}_1(n) - \mu_\gamma \cdot g_2(D_{o1})\, g_1(T). \end{aligned} \tag{14.15}$$

Alternatively, we may scale d_{11} and d_{12} by a common factor in Fig. 14.13 while setting c_{12} to unity, resulting in a more structural but equivalent adaptation scheme shown in Fig. 14.14 (the path subscripts are also dropped in the notation after deleting the second ADC path). Thus, we have

$$\begin{aligned} \alpha(n+1) &= \alpha(n) - \mu_\alpha \cdot g_1(D_o)\, g_2(T), \\ \beta(n+1) &= \beta(n) - \mu_\beta \cdot g_2(D_o)\, g_1(T), \end{aligned} \tag{14.16}$$

where α and β are the effective conversion radices for d_1 and d_2, respectively. For example, Fig. 14.15 shows the simulation results of a 12-bit ADC example using (14.16), where a common choice of $g_1(x) = x$ and $g_2(x) = x^3$ was selected [40]. This results in the following update equations,

$$\begin{aligned} \alpha(n+1) &= \alpha(n) - \mu_\alpha \cdot D_o(n) \cdot T(n), \\ \beta(n+1) &= \beta(n) - \mu_\beta \cdot {D_o}^3(n) \cdot T(n), \end{aligned} \tag{14.17}$$

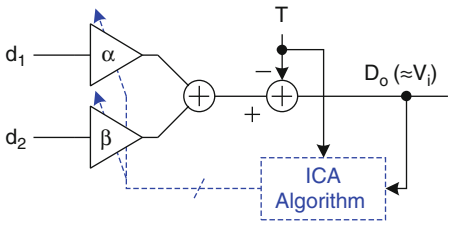


Fig. 14.14 ICA-based weight adaptation and PRBS removal for the ADC interstage gain calibration (evolved from the upper half of Fig. 14.13)

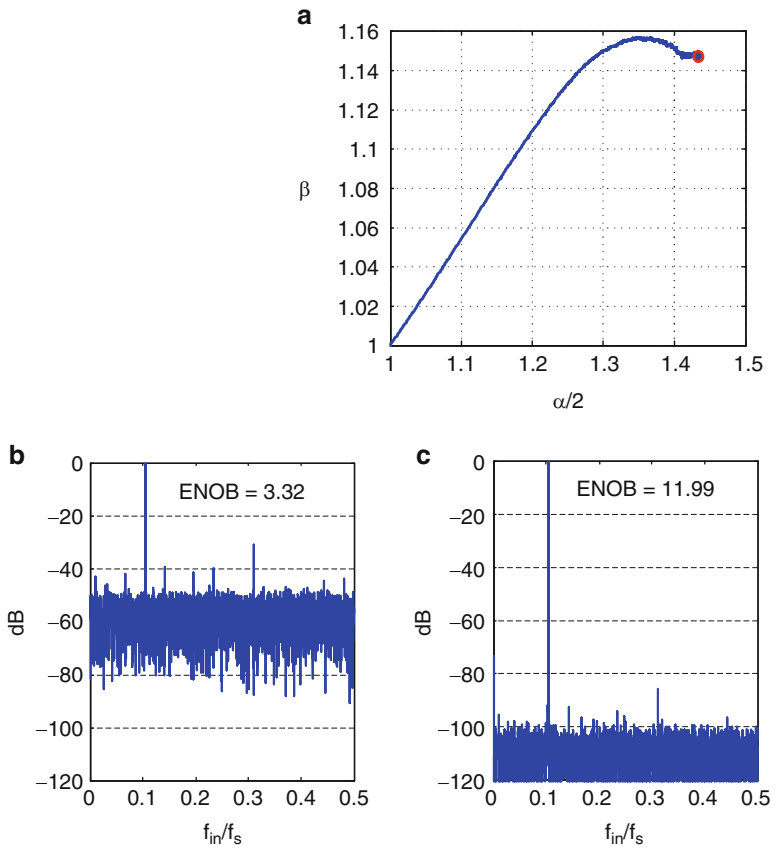


Fig. 14.15 Simulation results of the 12-bit ADC using H-J ICA algorithm: (*a*) learning trajectory of $\alpha/2$ and β, (b) FFT spectrum of the ADC output before ($\alpha/2 = \beta = 1$), and (*c*) after learning. The input has a frequency of 10 % of f_s and occupies 85 % of the conversion range

as T takes on a value of $+1$ or -1. The simulation setup also includes approximately a 10 % interstage gain error and a 30 % PRBS mismatch error, resulting in $\alpha = 2.8888$ and $\beta = 1.1556$ in this example.

As illustrated in Fig. 14.15*a*, starting from the default value of unity, the learning trajectory of $\alpha/2$ and β quickly approaches the locus of zero covariance before turning more slowly toward the target point marked by the circle, displaying a characteristic learning pattern of the H-J algorithm [40]. The step sizes in (14.17) were chosen in the simulation to minimize the steady-state coefficient fluctuations.

The effective number of bits (ENOB) of the 12-bit ADC before and after calibration is 3.32 and 11.99, respectively. The convergence time is approximately 300 million samples.[4]

In Fig. 14.14 two unknown parameters are identified with a single PRBS. This stems from the fact that an ADC input PRBS injection is chosen instead of either the sub-ADC or sub-DAC injection covered in Section 14.2.2. In the input-injection method, the size of the (single) PRBS capacitor can be of free choice and its exact value does not need to match to any other capacitors in the ADC. This compares to the sub-DAC injection method in which the PRBS capacitor needs to be matched to the sub-DAC capacitors [16, 17]. In contrast to the sub-ADC injection case wherein the calibration functions only within a small region around the comparator thresholds [26], every sample counts in the input-injection case—this usually means a shorter convergence time of the treatment.

Lastly, about the magnitude of T, obviously, the larger the faster the convergence of (14.17); however, a large T would also occupy too much of the input range, effectively limiting the signal-to-noise ratio (SNR) of the input signal. In practice, the amplitude of T can vary anywhere between 1 % and 10 % of the conversion range of the ADC.

14.3.3 Application to Other ADCs

In Section 14.3.2.1, the groundwork was established to apply ICA to the adaptive calibration of a two-stage ADC. The essence of the approach lies in the blind separation of the input signal and the injected PRBS, possible only when the ADC is a linear system, which is forced so upon the ICA training of the calibration engine. The ICA approach also exhibits a distinctive advantage—multiple error parameters can be identified simultaneously through the injection of a single, one-bit PRBS. As the injection can be realized by a small capacitor (without knowing its exact value) toggling between two voltage levels, the calibration overhead on the analog circuits is nearly negligible. In this section, we will generalize this approach to apply to some other types of data conversion circuits.

[4]The reported simulation results correspond to a sinusoidal input waveform in this example. The convergence time and learning accuracy do not seem to depend on the input waveform much, as long as it is busy and occupies most of the input range. The readers are referred to [6] for a more detailed discussion on this.

14.3.3.1 Cyclic ADC

The cyclic ADC shown in Fig. 14.16*a* can be considered as the simplest multistage ADC that loops the digitization process around itself, resolving one or a few bits during each iteration.

A 10-bit SC cyclic ADC employing the 1.5-bit architecture was modeled in the computer simulation. The circuit schematic (omitted) is similar to that of the 1.5-bit pipelined ADC shown in Fig. 14.2. A 15 % capacitor mismatch and a 20-dB op-amp gain were included in the modeling. An H-J algorithm was used to identify the bit weights resulting from the nonideal residue gain. Shown in Fig. 14.16*b*, the convergence time of the calibration was around 20 million samples. The ADC output spectra for a sinusoidal input before and after calibration are shown in Fig. 14.16*c* and Fig. 14.16*d*, respectively. The ENOB was improved from 3.28 to 9.79.

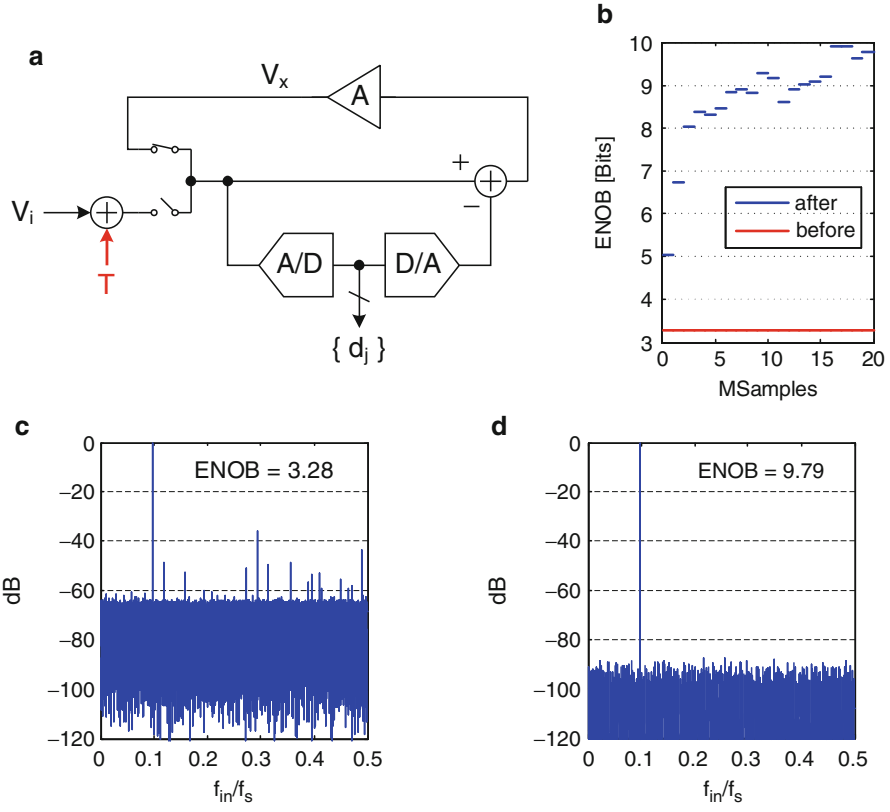


Fig. 14.16 ICA calibration of a 10-bit cyclic ADC: (*a*) block diagram and input PRBS injection, (*b*) ENOB learning curve, (*c*) output spectrum before calibration, and (*d*) the same spectrum after calibration

In this example, although the total number of bits resolved is 12, the number of independent bit weights is only two. In other words, the learning behavior is actually quite similar to the two-stage ADC example studied in Section 14.3.2.1. We will examine cases with more independent model parameters in the following sections.

14.3.3.2 SAR ADC

A conventional charge-redistribution SAR ADC employing an SC DAC is shown in Fig. 14.17. Because its operation does not require precision residue amplifiers, SAR ADC scales in a similar way as digital circuits—which has been witnessed by many recently reported SAR works [4, 5, 37, 51]. While the speed and bandwidth performance of SAR is benefiting significantly from scaling, its SNR and linearity performance are still largely limited by the decreasing supply voltage and the matching accuracy of the constituent DAC, i.e., the capacitors C_{N-1} through C_0 shown in Fig. 14.17. In an N-bit SAR ADC, the linearity of the DAC must be N-bit as well. For example, if a 14-bit linearity is desirable, the maximum mismatch error between any two of the $N+1$ capacitors must be smaller than 2^{-14} of the total capacitance of the array, which is very difficult to attain if no post-fabrication laser trimming is allowed. As a result, while many SAR works of superior power efficiency have been reported, few demonstrate an ENOB of 10 or above.

As the matching involves a total of $N+1$ capacitors, a direct application of the H-J algorithm would require $N+1$ nonlinear functions to extract the mismatch coefficients, dictating a large amount of digital computation. To resolve this problem, a bitwise ICA algorithm was introduced in [51]:

$$w_j(n+1) = w_j(n) - \mu_j \cdot \widehat{d}_j \cdot T, \quad j = 0, \ldots, N-1, \tag{14.18}$$

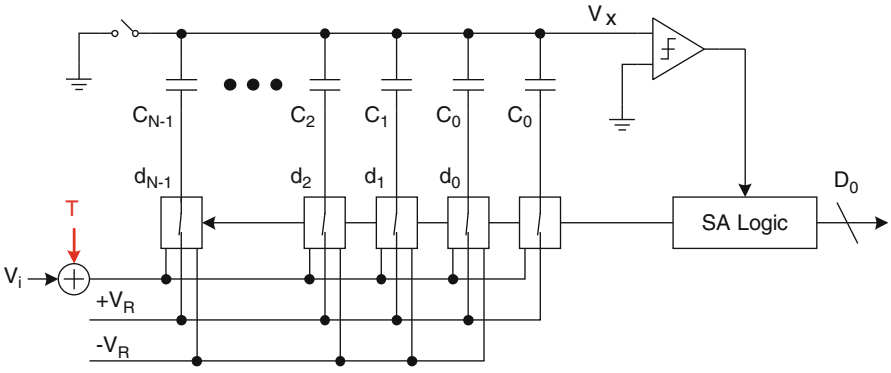


Fig. 14.17 PRBS injection in SAR ADC to identify the bit weights of DAC capacitors

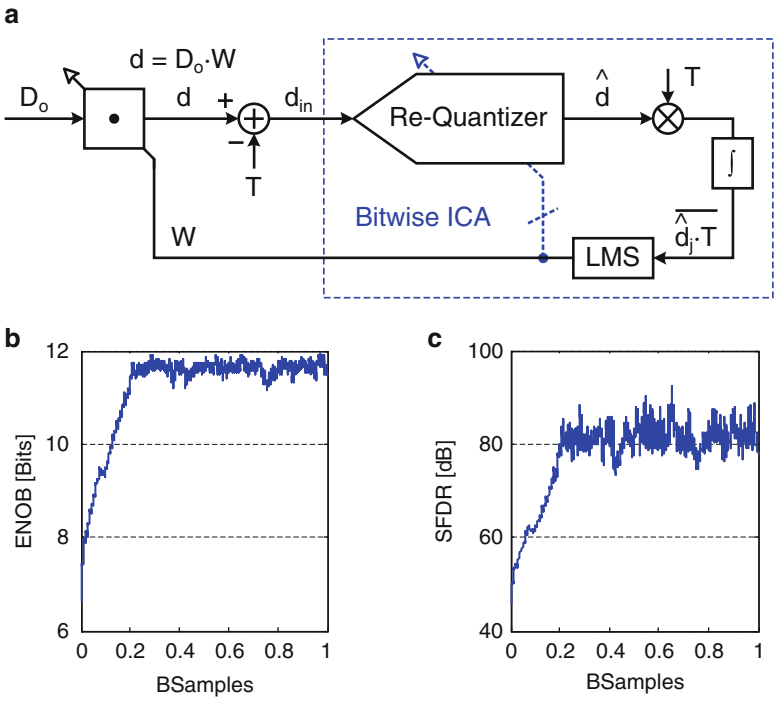


Fig. 14.18 SAR ADC with ICA calibration: (*a*) digital calibration engine, (*b*) ENOB learning curve, and (*c*) SFDR learning curve

where $w_j = C_j/\Sigma C$ is the normalized weight of the *j*th capacitor and $\widehat{d}_j$ is the *j*th bit obtained by re-quantizing d_{in}—the digital version of V_{in}—using the learned weights $\{w_j\}$. A block diagram of the calibration engine is illustrated in Fig. 14.18*a*. The computer simulation results are also shown in Fig. 14.18*b* and Fig. 14.18*c* for a 12-bit SAR ADC employing a sub-binary DAC using the bitwise ICA calibration algorithm. The convergence time is around 200 million samples.

To better apprehend the bitwise ICA technique, one can perceive the re-quantization intuitively as N nonlinear functions $\widehat{d}_j = g_j(d_{in})$ that are followed by the bitwise H-J de-correlation to identify all corresponding weights. Again, a single PRBS is sufficient to identify all weights. Since $\widehat{d}_j$ and T are both one-bit signals, the iteration of (14.18) can be efficiently executed in the digital domain.

In practice, only a few leading bit weights need to be calibrated. In a prototype 12-bit, 50-MS/s SAR ADC fabricated in a 90-nm CMOS process [51], the first 10 bit weights were calibrated. The chip measured a 66.5-dB SNDR and an 86.0-dB SFDR with calibration, while occupying 0.05 mm^2 and dissipating 3.3 mW from a 1.2-V supply. The calibration engine was estimated to occupy 0.07 mm^2 with a power consumption of 1.4 mW in the same process. Lastly, the analog overhead of the PRBS injection, as mentioned before, is nearly negligible.

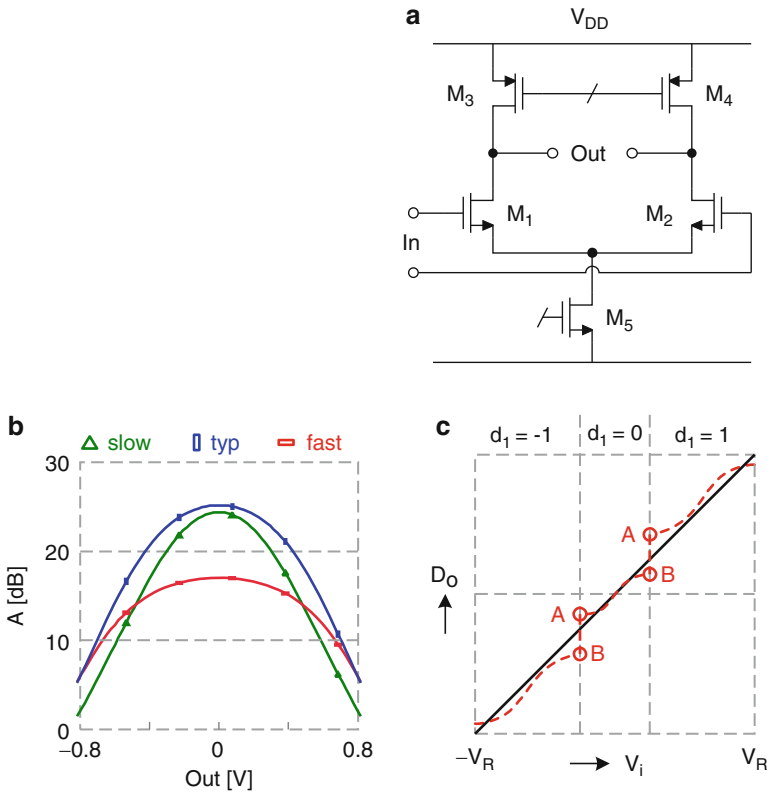


Fig. 14.19 A five-transistor amplifier: (*a*) circuit schematic, (*b*) simulated open-loop gain across process corners, and (*c*) the resulting ADC nonlinearity when the amplifier is used in the RA shown in Fig. 14.6

14.3.3.3 Nonlinear Amplifier

For simplicity, a piecewise-linear error model was adopted in Section 14.3.2 to capture the constant interstage gain error of a two-stage ADC. In practice, nonlinearities are always present in an amplifier due to the signal-dependent I–V characteristics of transistors. For example, a simple five-transistor amplifier is shown in Fig. 14.19*a*; its signal-dependent small-signal gain extracted from a transistor-level simulation is displayed in Fig. 14.19*b*. When this amplifier is used in a two-stage ADC similar to the one shown in Fig. 14.6, severe nonlinear distortion will result in addition to the missing codes, as indicated by the dashed conversion curve in Fig. 14.19*c*.

In general, the voltage transfer function of a weakly nonlinear amplifier can be approximated by a low-order power series,

$$V_o = f(V_i) \approx a_0 + a_1 V_i + a_2 {V_i}^2 + a_3 {V_i}^3 + \ldots, \tag{14.19}$$

where a_0 is the DC offset, a_1 is the small-signal gain, and so on. For differential circuits, the even-order coefficients are much smaller than the odd-order ones. Let us pick a third-order polynomial that models the nonlinearity of a voltage buffer with $a_0 = 0.1$, $a_1 = 1$, $a_2 = 0.01$, and $a_3 = -0.15$. A fifth-order power series trained by the H-J algorithm is used to treat the resulting nonlinearity,

$$b_j(n+1) = b_j(n) - \mu_j \cdot {D_o}^j \cdot T, \quad j = 1, \ldots, 5. \tag{14.20}$$

The setup of the calibration engine is sketched in Fig. 14.20*a*. Essentially, the algorithm attempts to adjust the coefficients $\{b_j\}$ to make sure that the various moment-correlation functions $E[{D_o}^j \cdot T]$ identically go to zero, thus achieving independence between D_o (i.e., the digital version of V_i) and T. From our previous discussions, this can be achieved only when the post-processing yields an optimum inverse of the buffer transfer function in the mean-square sense.

Note that offset cannot be corrected since its correlation with a zero-mean PRBS is always zero. However, this is in general not of concern since our focus is to treat nonlinearity.

In computer simulation, the untreated buffer exhibits a −34.1-dB total harmonic distortion (THD), which is reduced to −64.1 dB after 50 million iterations of (14.20). In contrast, an ideal fifth-order polynomial inverse of $f(V_i)$ yields a −65.5-dB THD. The learning curve of the first-, third-, and fifth-order coefficients is displayed in Fig. 14.20*b*. When a series of higher order is used, the linearity can be further improved.

The simplicity of the ICA approach also compares favorably to other reported works on amplifier linearization [27, 28], in which multiple PRBS injections are necessary, with one responsible for learning one b_j in (14.20). The resulting circuit implementation is much more complicated.

The ICA amplifier linearization technique has recently been applied to treat the RA nonlinear distortion in a two-step SAR ADC [52].

14.3.3.4 ΣΔ Modulator

Conventionally, circuit non-idealities in discrete-time ΣΔ modulators are modeled as linear effects, i.e., z-domain IIR-form signal transfer function (STF) and noise transfer function (NTF) [13]. In the first-order modulator shown in Fig. 14.21*a*, the quantizer output can be derived as

$$D(z) = \underbrace{\frac{\alpha\ z^{-1}}{1 + (\alpha\ \gamma - \beta)\ z^{-1}}}_{\text{STF}} X(z) + \underbrace{\frac{\left(1 - \beta\ z^{-1}\right)}{1 + (\alpha\ \gamma - \beta)\ z^{-1}}}_{\text{NTF}} E_q(z), \tag{14.21}$$

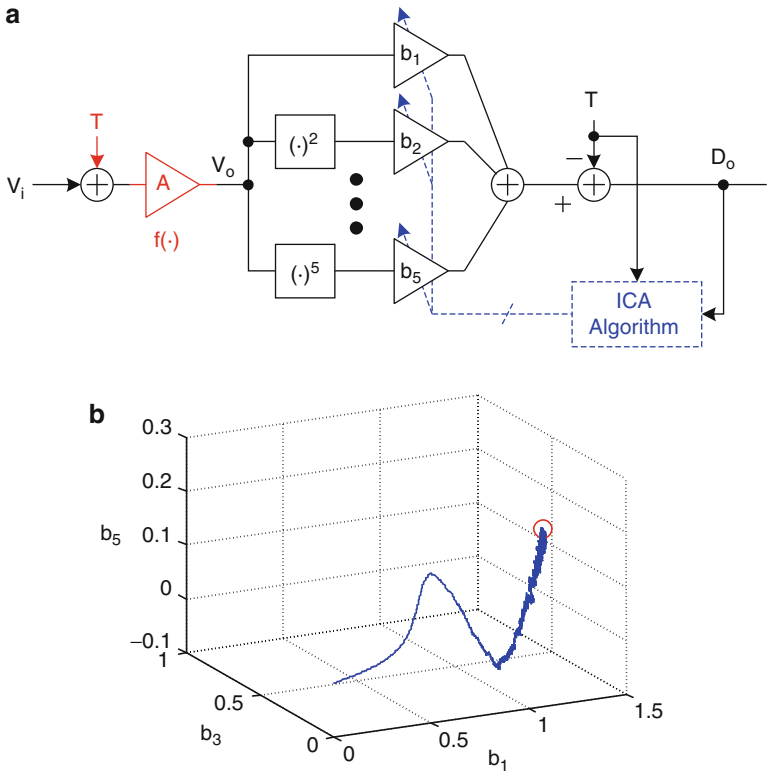


Fig. 14.20 ICA treatment of a nonlinear amplifier and simulation results: (*a*) adaptive polynomial inverse and (*b*) learning trajectory of the first-, third-, and fifth-order coefficients (the even-order terms are not shown)

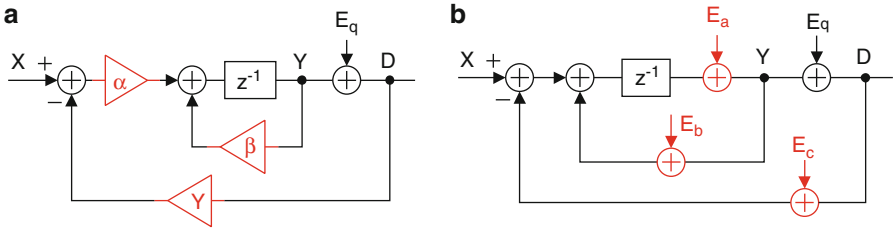


Fig. 14.21 (*a*) Conventional *z*-domain model and (*b*) output-referred nonlinear model of a first-order ΣΔ modulator. The signals *X*, *Y*, and *D* are the modulator input, integrator output, and quantizer output, respectively

where α, β, and γ capture the signal-path circuit non-idealities including capacitor mismatch, integrator leakage, and the DAC mismatch. In an ideal first-order modulator, $\alpha = \beta = \gamma = 1$ hold. A direct extension of the model of (14.21) to

account for the signal-path nonlinearities, i.e., the signal dependence of α, β, and γ, yields complicated analysis and no readily useful results.

In contrast, a recently reported nonlinear model [53] is shown in Fig. 14.21*b*, in which three additive error terms E_a, E_b, and E_c are used to represent the signal-path distortions that are all dependent on the modulator output signal D. The quantizer output can be derived as

$$x(n-1) = \underbrace{d(n) - e_q(n) + e_q(n-1)}_{\text{ideal modulator}} \underbrace{-e_a(n) - e_b(n-1) + e_c(n-1)}_{\text{additive error terms}}, \quad (14.22a)$$

where

$$\begin{aligned} e_a(n) &= \left(1 - \tfrac{1}{\alpha}\right)\, y(n) \approx \sum_i a_i [y(n)]^i \approx \sum_i a_i [d(n)]^i, \\ e_b(n-1) &= -\left(1 - \tfrac{\beta}{\alpha}\right)\, y(n-1) \approx \sum_i b_i [y(n-1)]^i \approx \sum_i b_i [d(n-1)]^i, \\ e_c(n-1) &= -(1-\gamma)\, d(n-1) = \sum_i c_i d_i(n-1). \end{aligned} \quad (14.22b)$$

A unique finding of [53] is that an FIR form, i.e., the two-tap model shown in (14.22) for a first-order modulator is sufficient and accurate in representing the long-term nonlinear memory errors of the modulator. Among the three terms, E_a and $z^{-1}E_b$ (dependent on $d(n)$ and $d(n-1)$, respectively) capture the integrator nonlinearity, whereas $z^{-1}E_c$ (dependent on the DAC thermometer code $d_i(n-1)$) expresses the component mismatch error of the feedback DAC [54].

For model parameter identification, conventionally, a one-bit PRBS is used to determine one parameter; an estimation of multiple nonlinear coefficients thus dictates multiple PRBS injections, potentially degrading the ADC dynamic range and complicating the analog circuitry involved for the injection. The learning algorithm reported in [53], based on the principle of ICA, trains multiple model parameters using a single PRBS injected into the input. Figure 14.22 illustrates the general setup of the nonlinear calibration. In a first-order modulator, if both E_a and E_b terms are properly removed by the adaptive polynomial transversal filter (APTF), the calibrated output $\widehat{X}$—a digitized version of the input X—will not contain any intermodulation products between $\widehat{X}$ and T. Therefore, the model parameters $\{a_i\}$ and $\{b_i\}$ can be updated iteratively until all moment-correlation terms are minimized, i.e.,

$$\begin{aligned} a_i(n) &= a_i(n-1) + \mu_{ai} \cdot \varepsilon(n) \cdot T(n-1) \cdot \widehat{x}^{i-1}(n-1), \\ b_i(n) &= b_i(n-1) + \mu_{bi} \cdot \varepsilon(n) \cdot T(n-2) \cdot \widehat{x}^{i-1}(n-2). \end{aligned} \quad (14.23)$$

The DAC distortion term E_c, which is not shown, is identified by a second PRBS injected into the DAC code [54].

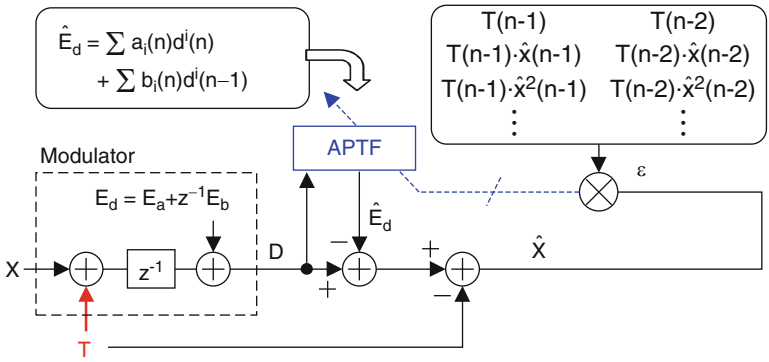


Fig. 14.22 Multiple nonlinear parameter identification with a single one-bit PRBS injection in a first-order ΣΔ modulator

A 1-0 MASH ΣΔ ADC (Fig. 14.23*a*) was prototyped [55], using an ICA-based digital calibration technique. The prototype employing 29-dB gain amplifiers measured an 85-dB SFDR and a 67-dB SNDR for a −1-dBFS, 5-MHz sinusoidal input at 240 MS/s. The core ADC consumes 37 mW from a 1.25-V supply and occupies 0.28 mm^2 in a 65-nm digital CMOS process. The measured ADC output spectra are shown in Fig. 14.23*b* (before calibration) and Fig. 14.23*c* (after calibration).

14.4 Summary and Remarks

In this chapter we developed a theoretical framework to apply ICA to the adaptive online calibration of multistage/multistep ADCs. In this approach, the statistical independence between an injected test signal (i.e., the PRBS) and the input signal is exploited to blindly separate the two in the conversion output, and while doing so an error model of the ADC circuit non-idealities is identified and subsequently applied to postprocessing the ADC output to linearize the conversion process.

In a signal-processing sense, the online calibration process can be considered as dynamic programming, to which many adaptive algorithms can be applied, including the H-J stochastic de-correlation algorithm, maximum likelihood, Bussgang method based on cumulants, negentropy method, and projection pursuit. In the treatment of sensitive, potentially time-varying analog circuits, critical performance parameters such as stability and convergence speed will need to be optimized. In addition, any algorithm employed must be amenable to VLSI implementation when operation speed and low power consumption are desirable.

Lastly, the ICA framework offers additional insights into the correlation-based calibration techniques that have been deployed widely in practice but lacked a rigorous theoretical treatment. For example, there is a well-known fact of ICA—the

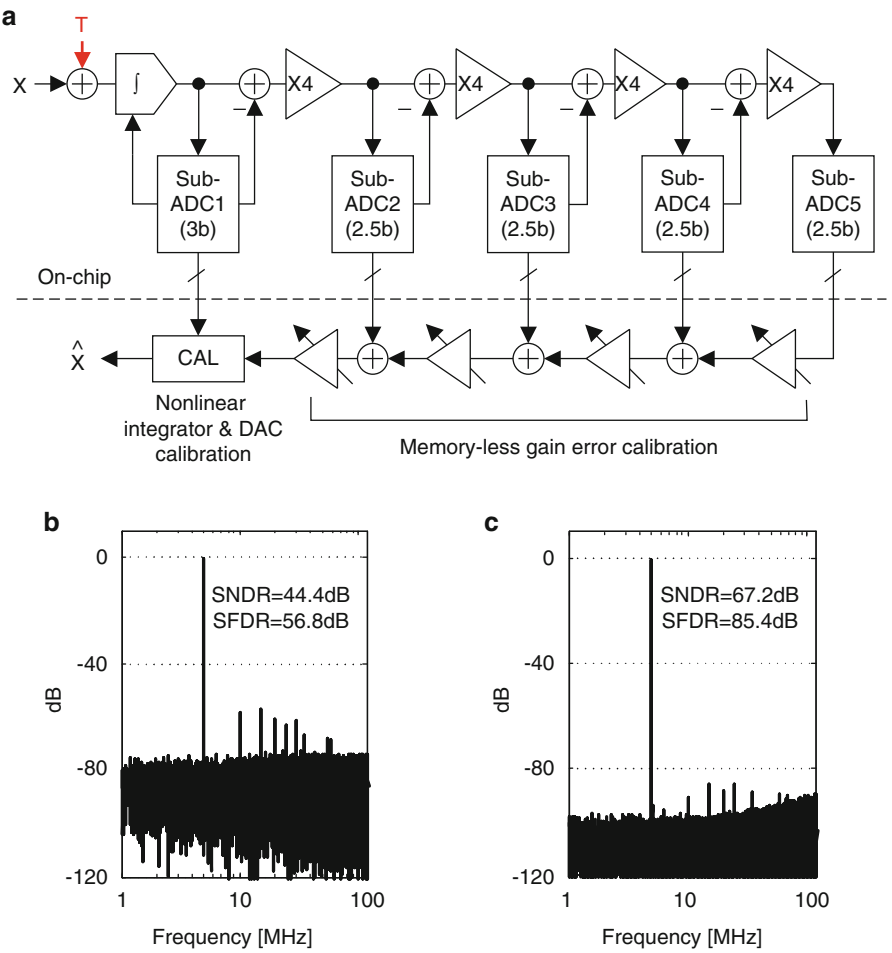


Fig. 14.23 Prototype 1-0 MASH ΣΔ ADC with ICA calibration: (a) block diagram, (b) measured ADC output spectrum before calibration, and (c) the same spectrum after calibration

key to estimating an ICA model is the *non-gaussianity*. The classic measure of non-gaussianity is *Kurtosis* or the fourth-order cumulant [50], which is defined as

$$Kurt(y) = E\left[y^4\right] - 3\left(E\left[y^2\right]\right)^2 \tag{14.24}$$

for a random variable y. In [25], a technique exhibiting little dependence on the input signal statistics was proposed to calibrate the interstage residue gain nonlinearity of a pipelined ADC. The key term used in [25] to extract the third-order coefficient of a polynomial inverse was derived from a heuristic function

$$K(y) = E\left[y^4\right] - \left(E\left[y^2\right]\right)^2. \tag{14.25}$$

It appears that the authors of [25] have essentially (incidentally) maximized a quantity of non-gaussianity alike in their work, implying that a more systematic treatment of the problem can perhaps be approached using the ICA framework.

References

1. S.H. Lewis, H.S. Fetterman, G.F. Gross Jr., R. Ramachandran, T.R. Viswanathan, A 10-b 20-MS/s analog-to-digital converter. IEEE J. Solid State Circuits **27**, 351–358 (1992)
2. B. Ginetti, P.G.A. Jespers, A. Vandemeulebroecke, A CMOS 13-b cyclic RSD A/D converter. IEEE J. Solid State Circuits **27**, 957–964 (1992)
3. F. Kuttner, A 1.2 V 10b 20 MS/s non-binary SAR ADC in 0.13 μm CMOS, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 176–177, 2002
4. C.-C. Liu et al., A 10b 100 MS/s 1.13 mW SAR ADC with binary-scaled error compensation, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 386–387, 2010
5. W. Liu et al., A 12b 22.5/45 MS/s 3.0 mW 0.059 mm^2 CMOS SAR ADC achieving over 90 dB SFDR, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 380–381, 2010
6. Y. Chiu et al., Least-mean-square adaptive digital background calibration of pipelined analog-to-digital converters. IEEE Trans. Circuits Syst. I **51**, 38–46 (2004)
7. C. Tsang et al., Background ADC calibration in digital domain, in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 301–304, 2008
8. Y. Chiu, Recent advances in digital-domain background calibration techniques for multistep analog-to-digital converters, in *IEEE Int. Conf. Solid-State and Integrated-Circuit Tech.*, pp. 1905−1908, 2008
9. M.K. Mayes, S.W. Chin, A 200-mW, 1-MS/s, 16-b pipelined A/D converter with on-chip 32-b microcontroller. IEEE J. Solid State Circuits **31**, 1862–1872 (1996)
10. A. Wiesbauer, G.C. Temes, Adaptive compensation of analog circuit imperfections for cascaded sigma-delta modulators, in *Proc. Asilomar Conf. Circuits, Systems and Computers*, vol. 2, pp. 1073–1077, 1996
11. C. Petrie, M. Miller, A background calibration technique for multibit delta-sigma modulators, in *Proc. IEEE Int. Sym. Circuits and Systems*, vol. 2, pp. 29–32, 2000
12. T. Sun, A. Wiesbauer, G.C. Temes, Adaptive compensation of analog circuit imperfections for cascaded delta-sigma ADCs, in *Proc. IEEE Int. Sym. Circuits and Systems*, vol. 1, pp. 405–407, 1998
13. P. Kiss et al., Adaptive digital correction of analog errors in MASH ADC's—Part II: correction using test-signal injection. IEEE Trans. Circuits Syst. II **47**, 629–638 (2000)
14. D. Fu, K.C. Dyer, S.H. Lewis, P.J. Hurst, A digital back-ground calibration technique for time-interleaved analog-to-digital converters. IEEE J. Solid State Circuits **33**, 1904–1911 (1998)
15. J. Ming, S.H. Lewis, An 8 b 80 Msample/s pipelined ADC with background calibration, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 42–43, 2000
16. E.J. Siragusa, I. Galton, Gain error correction technique for pipelined analogue-to-digital converters. Electron. Lett. **36**, 617–618 (2000)
17. I. Galton, Digital cancellation of D/A converter noise in pipelined A/D converters. IEEE Trans. Circuits Syst. II **47**, 185–196 (2000)
18. P.C. Yu et al., A 14b 40 MSample/s pipelined ADC with DFCA, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 136–137, 2001
19. E. Siragusa, I. Galton, A digitally enhanced 1.8-V 15-bit 40-MSample/s CMOS pipelined ADC. IEEE J. Solid State Circuits **39**, 2126–2138 (2004)

20. H.-C. Liu, Z.-M. Lee, J.-T. Wu, A 15b 20 MS/s CMOS pipelined ADC with digital background calibration, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 454–455, 2004
21. K. Nair, R. Harjani, A 96 dB SFDR 50 MS/s digitally enhanced CMOS pipeline A/D converter, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 456–457, 2004
22. R. Massolini, G. Cesura, R. Castello, A fully digital fast convergence algorithm for nonlinearity correction in multistage ADC. IEEE Trans. Circuits Syst. II **53**, 389–393 (2006)
23. Y.-S. Shu, B.-S. Song, A 15b linear, 20 MS/s, 1.5b/stage pipelined ADC digitally calibrated with signal-dependent dithering, in *IEEE Sym. VLSI Circuits*, *Dig. Tech. Papers*, pp. 218–219, 2006
24. B. Murmann, B. Boser, A 12b 75 MS/s pipelined ADC using open-loop residue amplification, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 328–329, 2003
25. J. Keane et al., Background interstage gain calibration technique for pipelined ADCs. IEEE Trans. Circuits Syst. I **52**, 32–43 (2005)
26. J. Li, U.-K. Moon, Background calibration techniques for multistage pipelined ADC's with digital redundancy. IEEE Trans. Circuits Syst. II **50**, 531–538 (2003)
27. A. Panigada, I. Galton, Digital background correction of harmonic distortion in pipelined ADCs. IEEE Trans. Circuits Syst. I **53**, 1885–1895 (2006)
28. A. Panigada, I. Galton, A 130 mW 100 MS/s pipelined ADC with 69 dB SNDR enabled by digital harmonic distortion correction, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 162–163, 2009
29. R. Jewett, K. Poulton, K.-C. Hsieh, J. Doernberg, A 12b 128 MS/s ADC with 0.05LSB DNL, in *IEEE Int. Solid-State Circuits Conf.*, *Dig. Tech. Papers*, pp. 138–139, 1997
30. H.S. Fetterman, D.G. Martin, D.A. Rich, CMOS pipelined ADC employing dither to improve linearity, in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 109–112, 1999
31. J. McNeill, M.C.W. Coln, B.J. Larivee, "Split ADC" architecture for deterministic digital background calibration of a 16-bit 1-MS/s ADC. IEEE J. Solid State Circuits **40**, 2437–2445 (2005)
32. J.A. McNeill, S. Goluguri, A. Nair, Split ADC digital background correction of open loop residue amplifier non linearity errors in a 14b pipelined ADC, in *Proc. IEEE Int. Symp. Circuits and Systems*, pp. 1237–1240, 2007
33. I. Ahmed, D. Johns, An 11-bit 45 MS/s pipelined ADC with rapid calibration of DAC errors in a multibit pipeline stage. IEEE J. Solid State Circuits **43**, 1626–1637 (2008)
34. L.-H. Hung, T.-C. Lee, A split-based digital background calibration technique in pipelined ADCs. IEEE Trans. Circuits Syst. II **56**, 855–859 (2009)
35. L. Ding et al., A 13-bit 60 MS/s split pipelined ADC with background gain and mismatch error calibration, in *IEEE Asian Solid-State Circuits Conf.*, pp. 77–80, 2013
36. B. Peng, H. Li, P. Lin, Y. Chiu, An offset double conversion technique for digital calibration of pipelined ADCs. IEEE Trans. Circuits Syst. II **57**, 961–965 (2010)
37. W. Liu, P. Huang, Y. Chiu, A 12-bit, 45-MS/s, 3-mW redundant successive-approximation-register analog-to-digital converter with digital calibration. IEEE J. Solid State Circuits **46**, 2661–2672 (2011)
38. J. Hérault, C. Jutten, Space or time adaptive signal processing by neural network models, in *Proc. Neural Networks for Computing*, vol. 151, pp. 206–211, 1986
39. J.-F. Cardoso, Source separation using higher order moments, in *Proc. IEEE Int. Conf. Acoustics, Speech and Signal Processing*, pp. 2109–2112, 1989
40. E.A. Vittoz, X. Arreguit, CMOS integration of Hérault-Jutten cells for separation of sources, in *Analog VLSI Implementation of Neural Systems*, ed. by C. Mead, M. Ismail (Springer, Berlin, 1989)
41. C. Jutten, J. Hérault, Blind separation of sources, part I: an adaptive algorithm based on neuromimetic architecture. Signal Process. **24**, 1–10 (1991)
42. P. Comon, C. Jutten, J. Hérault, Blind separation of sources, part II: problems statement. Signal Process. **24**, 11–20 (1991)
43. E. Sorouchyari, Blind separation of sources, part III: stability analysis. Signal Process. **24**, 21–29 (1991)

44. G. Burel, Blind separation of sources: a nonlinear neural algorithm. Neural Netw. **5**, 937–947 (1992)
45. A. Cichocki, R. Unbehauen, Robust neural networks with on-line learning for blind identification and blind separation of sources. IEEE Trans. Circuits Syst. **43**, 894–906 (1996)
46. S.-I. Amari, A. Cichocki, Adaptive blind signal processing—neural network approaches. Proc. IEEE **86**, 2026–2048 (1998)
47. H.B. Barlow, Unsupervised learning. Neural Comput. **1**, 295–311 (1989)
48. J.-F. Cardoso, P. Comon, Independent Component Analysis: a survey of some algebraic methods, in *Proc. IEEE Int. Sym. Circuits and Systems*, vol. 2, pp. 93–96, 1996
49. S. Haykin, *Adaptive Filter Theory*, 3rd edn. (Prentice Hall, Upper Saddle River, 1996)
50. A. Hyvärinen, J. Karhunen, E. Oja, *Independent Component Analysis* (Wiley, New York, 2001)
51. W. Liu, P. Huang, Y. Chiu, A 12-bit 50-MS/s 3.3-mW SAR ADC with background digital calibration, in *Proc. IEEE Custom Integrated Circuits Conf.*, pp. 1–4, 2012
52. Y. Zhou, Y. Chiu, Digital calibration of inter-stage nonlinear errors in pipelined SAR ADC, in *Proc. IEEE Int. Midwest Symp. Circuits and Systems*, pp. 677–680, 2013
53. S.-C. Lee, Y. Chiu, Digital calibration of nonlinear memory errors in sigma-delta modulators. IEEE Trans. Circuits Syst. I **57**, 2462–2475 (2010)
54. S.-C. Lee, Y. Chiu, Digital calibration of capacitor mismatch in sigma-delta modulators. IEEE Trans. Circuits Syst. I **58**, 690–698 (2011)
55. S.-C. Lee, Y. Chiu, A 15-MHz bandwidth 1-0 MASH $\Sigma\Delta$ ADC with nonlinear memory error calibration achieving 85-dBc SFDR. IEEE J. Solid State Circuits **49**, 695–707 (2014)

Index

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4

Applied and Numerical Harmonic Analysis (70 volumes)

A. Saichev and W.A. Woyczyñski: *Distributions in the Physical and Engineering Sciences* (ISBN 978-0-8176-3924-2)
C.E. D'Attellis and E.M. Fernandez-Berdaguer: *Wavelet Theory and Harmonic Analysis in Applied Sciences* (ISBN 978-0-8176-3953-2)
H.G. Feichtinger and T. Strohmer: *Gabor Analysis and Algorithms* (ISBN 978-0-8176-3959-4)
R. Tolimieri and M. An: *Time-Frequency Representations* (ISBN 978-0-8176-3918-1)
T.M. Peters and J.C. Williams: *The Fourier Transform in Biomedical Engineering* (ISBN 978-0-8176-3941-9)
G.T. Herman: *Geometry of Digital Spaces* (ISBN 978-0-8176-3897-9)
A. Teolis: *Computational Signal Processing with Wavelets* (ISBN 978-0-8176-3909-9)
J. Ramanathan: *Methods of Applied Fourier Analysis* (ISBN 978-0-8176-3963-1)
J.M. Cooper: *Introduction to Partial Differential Equations with MATLAB* (ISBN 978-0-8176-3967-9)
A. Procházka, N.G. Kingsbury, P.J. Payner, and J. Uhlir: *Signal Analysis and Prediction* (ISBN 978-0-8176-4042-2)
W. Bray and C. Stanojevic: *Analysis of Divergence* (ISBN 978-1-4612-7467-4)
G.T. Herman and A. Kuba: *Discrete Tomography* (ISBN 978-0-8176-4101-6)
K. Gröchenig: *Foundations of Time-Frequency Analysis* (ISBN 978-0-8176-4022-4)
L. Debnath: *Wavelet Transforms and Time-Frequency Signal Analysis* (ISBN 978-0-8176-4104-7)
J.J. Benedetto and P.J.S.G. Ferreira: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)
D.F. Walnut: *An Introduction to Wavelet Analysis* (ISBN 978-0-8176-3962-4)
A. Abbate, C. DeCusatis, and P.K. Das: *Wavelets and Subbands* (ISBN 978-0-8176-4136-8)

G.E. Pfander (ed.), *Sampling Theory, a Renaissance*, Applied and Numerical Harmonic Analysis, DOI 10.1007/978-3-319-19749-4

O. Bratteli, P. Jorgensen, and B. Treadway: *Wavelets Through a Looking Glass* (ISBN 978-0-8176-4280-80)
H.G. Feichtinger and T. Strohmer: *Advances in Gabor Analysis* (ISBN 978-0-8176-4239-6)
O. Christensen: *An Introduction to Frames and Riesz Bases* (ISBN 978-0-8176-4295-2)
L. Debnath: *Wavelets and Signal Processing* (ISBN 978-0-8176-4235-8)
G. Bi and Y. Zeng: *Transforms and Fast Algorithms for Signal Analysis and Representations* (ISBN 978-0-8176-4279-2)
J.H. Davis: *Methods of Applied Mathematics with a MATLAB Overview* (ISBN 978-0-8176-4331-7)
J.J. Benedetto and A.I. Zayed: *Modern Sampling Theory* (ISBN 978-0-8176-4023-1)
E. Prestini: *The Evolution of Applied Harmonic Analysis* (ISBN 978-0-8176-4125-2)
L. Brandolini, L. Colzani, A. Iosevich, and G. Travaglini: *Fourier Analysis and Convexity* (ISBN 978-0-8176-3263-2)
W. Freeden and V. Michel: *Multiscale Potential Theory* (ISBN 978-0-8176-4105-4)
O. Christensen and K.L. Christensen: *Approximation Theory* (ISBN 978-0-8176-3600-5)
O. Calin and D.-C. Chang: *Geometric Mechanics on Riemannian Manifolds* (ISBN 978-0-8176-4354-6)
J.A. Hogan: *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, M.I. Vai, and Y. Xu: *Wavelet Analysis and Applications* (ISBN 978-3-7643-7777-9)
G.T. Herman and A. Kuba: *Advances in Discrete Tomography and Its Applications* (ISBN 978-0-8176-3614-2)
M.C. Fu, R.A. Jarrow, J.-Y. Yen, and R.J. Elliott: *Advances in Mathematical Finance* (ISBN 978-0-8176-4544-1)
O. Christensen: *Frames and Bases* (ISBN 978-0-8176-4677-6)
P.E.T. Jorgensen, J.D. Merrill, and J.A. Packer: *Representations, Wavelets, and Frames* (ISBN 978-0-8176-4682-0)
M. An, A.K. Brodzik, and R. Tolimieri: *Ideal Sequence Design in Time-Frequency Space* (ISBN 978-0-8176-4737-7)
S.G. Krantz: *Explorations in Harmonic Analysis* (ISBN 978-0-8176-4668-4)
B. Luong: *Fourier Analysis on Finite Abelian Groups* (ISBN 978-0-8176-4915-9)
G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 1* (ISBN 978-0-8176-4802-2)
C. Cabrelli and J.L. Torrea: *Recent Developments in Real and Harmonic Analysis* (ISBN 978-0-8176-4531-1)
M.V. Wickerhauser: *Mathematics for Multimedia* (ISBN 978-0-8176-4879-4)

B. Forster, P. Massopust, O. Christensen, K. Gröchenig, D. Labate, P. Vandergheynst, G. Weiss, and Y. Wiaux: *Four Short Courses on Harmonic Analysis* (ISBN 978-0-8176-4890-9)
O. Christensen: *Functions, Spaces, and Expansions* (ISBN 978-0-8176-4979-1)
J. Barral and S. Seuret: *Recent Developments in Fractals and Related Fields* (ISBN 978-0-8176-4887-9)
O. Calin, D.-C. Chang, and K. Furutani, and C. Iwasaki: *Heat Kernels for Elliptic and Sub-elliptic Operators* (ISBN 978-0-8176-4994-4)
C. Heil: *A Basis Theory Primer* (ISBN 978-0-8176-4686-8)
J.R. Klauder: *A Modern Approach to Functional Integration* (ISBN 978-0-8176-4790-2)
J. Cohen and A.I. Zayed: *Wavelets and Multiscale Analysis* (ISBN 978-0-8176-8094-7)
D. Joyner and J.-L. Kim: *Selected Unsolved Problems in Coding Theory* (ISBN 978-0-8176-8255-2)
G.S. Chirikjian: *Stochastic Models, Information Theory, and Lie Groups, Volume 2* (ISBN 978-0-8176-4943-2)
J.A. Hogan and J.D. Lakey: *Duration and Bandwidth Limiting* (ISBN 978-0-8176-8306-1)
G. Kutyniok and D. Labate: *Shearlets* (ISBN 978-0-8176-8315-3)
P.G. Casazza and P. Kutyniok: *Finite Frames* (ISBN 978-0-8176-8372-6)
V. Michel: *Lectures on Constructive Approximation* (ISBN 978-0-8176-8402-0)
D. Mitrea, I. Mitrea, M. Mitrea, and S. Monniaux: *Groupoid Metrization Theory* (ISBN 978-0-8176-8396-2)
T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 1* (ISBN 978-0-8176-8375-7)
T.D. Andrews, R. Balan, J.J. Benedetto, W. Czaja, and K.A. Okoudjou: *Excursions in Harmonic Analysis, Volume 2* (ISBN 978-0-8176-8378-8)
D.V. Cruz-Uribe and A. Fiorenza: *Variable Lebesgue Spaces* (ISBN 978-3-0348-0547-6)
W. Freeden and M. Gutting: *Special Functions of Mathematical (Geo-)Physics* (ISBN 978-3-0348-0562-9)
A. Saichev and W.A. Woyczyñski: *Distributions in the Physical and Engineering Sciences, Volume 2: Linear and Nonlinear Dynamics of Continuous Media* (ISBN 978-0-8176-3942-6)
S. Foucart and H. Rauhut: *A Mathematical Introduction to Compressive Sensing* (ISBN 978-0-8176-4947-0)
G. Herman and J. Frank: *Computational Methods for Three-Dimensional Microscopy Reconstruction* (ISBN 978-1-4614-9520-8)
A. Paprotny and M. Thess: *Realtime Data Mining: Self-Learning Techniques for Recommendation Engines* (ISBN 978-3-319-01320-6)
A. Zayed and G. Schmeisser: *New Perspectives on Approximation and Sampling Theory: Festschrift in Honor of Paul Butzer's 85th Birthday* (978-3-319-08800-6)
R. Balan, M. Begue, J. Benedetto, W. Czaja, and K.A Okoudjou: *Excursions in Harmonic Analysis, Volume 3* (ISBN 978-3-319-13229-7)

H. Boche, R. Calderbank, G. Kutyniok, and J. Vybiral: *Compressed Sensing and its Applications: MATHEON Workshop 2013* (ISBN 978-3-319-16041-2)
S. Dahlke, F. De Mari, P. Grohs, and D. Labate: *Harmonic and Applied Analysis: From Groups to Signals* (ISBN 978-3-319-18862-1)
G. Pfander: *Sampling Theory, a Renaissance* (ISBN 978-3-319-19748-7)

For an up-to-date list of ANHA titles, please visit: http://www.springer.com/series/4968