ETH
Eidgenössische Technische Hochschule Zürich
Swiss Federal Institute of Technology Zurich

Chair for Mathematical Information Science

Prof. Dr. H. Bölcskei
Sternwartstrasse 7
CH-8092 Zürich

Lecture Notes
Mathematics of Information

Helmut Bölcskei

February 2022

The author gratefully acknowledges T. Allard's contributions to Section 12, Appendices A and B are entirely by T. Allard. Section 1 is joint work with V. Morgenshtern, Section 2 is joint work with E. Riegler, and R. Heckel contributed to Sections 8 and 9.

Contents

0.1 Notation

$\mathbf{a}, \mathbf{b}, \ldots$	vectors
$\mathbf{A}, \mathbf{B}, \ldots$	matrices
$\mathbf{a}^\mathsf{T},\ \mathbf{A}^\mathsf{T}$	transpose of the vector $\mathbf{a}$ and the matrix $\mathbf{A}$
$a^*,\ \mathbf{a}^*,\ \mathbf{A}^*$	complex conjugate of the scalar a, element-wise complex conjugate of the vector $\mathbf{a}$, and the matrix $\mathbf{A}$
$\mathbf{a}^\mathsf{H},\ \mathbf{A}^\mathsf{H}$	Hermitian transpose of the vector $\mathbf{a}$ and the matrix $\mathbf{A}$
$\mathbf{I}_N$	identity matrix of size $N \times N$
$\operatorname{rank}(\mathbf{A})$	rank of the matrix $\mathbf{A}$
$\lambda(\mathbf{A})$	eigenvalue of the matrix $\mathbf{A}$
$\lambda_{\min}(\mathbf{A}), \lambda_{\min}(\mathbb{A})$	smallest eigenvalue of the matrix $\mathbf{A}$, smallest spectral value of the self-adjoint operator $\mathbb{A}$
$\lambda_{\max}(\mathbf{A}), \lambda_{\max}(\mathbb{A})$	largest eigenvalue of the matrix $\mathbf{A}$, largest spectral value of the self-adjoint operator $\mathbb{A}$
i	$\sqrt{-1}$
$\triangleq$	definition
$\mathcal{A}, \mathcal{B}, \ldots$	sets
$\mathbb{R}, \mathbb{C}, \mathbb{Z}, \mathbb{N}$	real line, complex plane, set of all integers, set of natural numbers (including zero)
$\mathcal{L}^2$	Hilbert space of complex-valued finite-energy functions
$\mathcal{L}^2(B)$	space of square-integrable functions bandlimited to B Hz
$\mathcal{H}$	abstract Hilbert space
l^2	Hilbert space of square-summable sequences
$\langle \mathbf{a}, \mathbf{b} \rangle$	inner product of the vectors $\mathbf{a}$ and $\mathbf{b}$: $\langle \mathbf{a}, \mathbf{b} \rangle \triangleq \sum_i [\mathbf{a}]_i ([\mathbf{b}]_i)^*$
$\langle x, y \rangle$	depending on the context: inner product in the abstract Hilbert space $\mathcal{H}$ or inner product of the functions $x(t)$ and $y(t)$: $\langle x, y \rangle \triangleq \int_{-\infty}^{\infty} x(t) y^*(t) dt$
$\lVert \mathbf{a} \rVert^2$	squared ℓ^2-norm of the vector $\mathbf{a}$: $\lVert \mathbf{a} \rVert^2 \triangleq \sum_i \lvert [\mathbf{a}]_i \rvert^2$
$\lVert y \rVert^2$	depending on the context: squared norm in the abstract Hilbert space $\mathcal{H}$ or squared $\mathcal{L}^2$-norm of the function $y(t)$: $\lVert y \rVert^2 \triangleq \int_{-\infty}^{\infty} \lvert y(t) \rvert^2 dt$
$\mathbb{I}_{\mathcal{H}}, \mathbb{I}_{l^2}, \mathbb{I}_{\mathcal{L}^2}, \mathbb{I}_{\mathcal{L}^2(B)}$	identity operator in the corresponding space
$\mathcal{R}(\mathbb{A})$	range space of operator $\mathbb{A}$
$\mathbb{A}^*$	adjoint of operator $\mathbb{A}$
$\widehat{x}(f)$	Fourier transform of $x(t)$: $\widehat{x}(f) \triangleq \int_{-\infty}^{\infty} x(t) e^{-\mathrm{i}2\pi t f} dt$
$\widehat{x}_d(f)$	Discrete-time Fourier transform of $x[k]$: $\widehat{x}_d(f) \triangleq \sum_{k=-\infty}^{\infty} x[k] e^{-\mathrm{i}2\pi k f}$

Chapter 1

A Short Course on Frame Theory

Hilbert spaces [1, Def. 3.1-1] and the associated concept of orthonormal bases are of fundamental importance in signal processing, communications, control, and information theory. However, linear independence and orthonormality of the basis elements impose constraints that often make it difficult to have the basis elements satisfy additional desirable properties. This calls for a theory of signal decompositions that is flexible enough to accommodate decompositions into possibly nonorthogonal and redundant signal sets. The theory of frames provides such a tool.

This chapter is an introduction to the theory of frames, which was developed by Duffin and Schaeffer [2] and popularized mostly through [3–6]. Meanwhile frame theory, in particular the aspect of redundancy in signal expansions, has found numerous applications such as, e.g., denoising [7, 8], code division multiple access (CDMA) [9], orthogonal frequency division multiplexing (OFDM) systems [10], coding theory [11, 12], quantum information theory [13], analog-to-digital (A/D) converters [14–16], and compressive sensing [17–19]. A more extensive list of relevant references can be found in [20]. For a comprehensive treatment of frame theory we refer to the excellent textbook [21].

1.1 Examples of Signal Expansions

We start by considering some simple motivating examples.

Example 1.1 (Orthonormal basis in $\mathbb{R}^2$)**.** Consider the orthonormal basis (ONB)

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{e}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

in $\mathbb{R}^2$ (see Figure 1.1). We can represent every signal $\mathbf{x} \in \mathbb{R}^2$ as the following linear combination of the basis vectors $\mathbf{e}_1$ and $\mathbf{e}_2$:

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{e}_1 \rangle \, \mathbf{e}_1 + \langle \mathbf{x}, \mathbf{e}_2 \rangle \, \mathbf{e}_2. \tag{1.1}$$

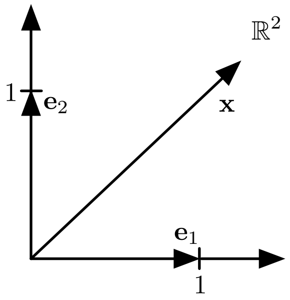


Figure 1.1: Orthonormal basis in $\mathbb{R}^2$.

To rewrite (1.1) in vector-matrix notation, we start by defining the vector of expansion coefficients as

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \triangleq \begin{bmatrix} \langle \mathbf{x}, \mathbf{e}_1 \rangle \\ \langle \mathbf{x}, \mathbf{e}_2 \rangle \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1^\mathsf{T} \\ \mathbf{e}_2^\mathsf{T} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x}.$$

It is convenient to define the matrix

$$\mathbf{T} \triangleq \begin{bmatrix} \mathbf{e}_1^\mathsf{T} \\ \mathbf{e}_2^\mathsf{T} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$

Henceforth we call $\mathbf{T}$ the *analysis matrix*; it multiplies the signal $\mathbf{x}$ to produce the expansion coefficients

$$\mathbf{c} = \mathbf{T}\mathbf{x}.$$

Following (1.1), we can reconstruct the signal $\mathbf{x}$ from the coefficient vector $\mathbf{c}$ according to

$$\mathbf{x} = \mathbf{T}^\mathsf{T}\mathbf{c} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix} \mathbf{c} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix} \begin{bmatrix} \langle \mathbf{x}, \mathbf{e}_1 \rangle \\ \langle \mathbf{x}, \mathbf{e}_2 \rangle \end{bmatrix} = \langle \mathbf{x}, \mathbf{e}_1 \rangle \mathbf{e}_1 + \langle \mathbf{x}, \mathbf{e}_2 \rangle \mathbf{e}_2. \tag{1.2}$$

We call

$$\mathbf{T}^\mathsf{T} = \begin{bmatrix} \mathbf{e}_1 & \mathbf{e}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \tag{1.3}$$

the *synthesis matrix*; it multiplies the coefficient vector $\mathbf{c}$ to recover the signal $\mathbf{x}$. It follows from (1.2) that (1.1) is equivalent to

$$\mathbf{x} = \mathbf{T}^\mathsf{T}\mathbf{T}\mathbf{x} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x}. \tag{1.4}$$

The introduction of the analysis and the synthesis matrix in the example above may seem artificial and may appear as complicating matters unnecessarily. After all, both $\mathbf{T}$ and $\mathbf{T}^\mathsf{T}$ are equal to the identity matrix in this example. We will, however, see shortly that this notation paves the way to developing a unified framework for nonorthogonal and redundant signal expansions. Let us now look at a somewhat more interesting example.

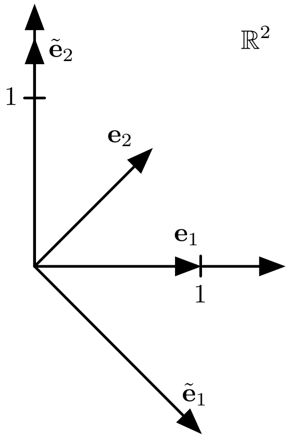


Figure 1.2: Biorthonormal bases in $\mathbb{R}^2$.

Example 1.2 (Biorthonormal bases in $\mathbb{R}^2$)**.** Consider two noncollinear unit norm vectors in $\mathbb{R}^2$. For concreteness, take (see Figure 1.2)

$$\mathbf{e}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \qquad \mathbf{e}_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$

For an arbitrary signal $\mathbf{x} \in \mathbb{R}^2$, we can compute the expansion coefficients

$$c_1 \triangleq \langle \mathbf{x}, \mathbf{e}_1 \rangle$$
$$c_2 \triangleq \langle \mathbf{x}, \mathbf{e}_2 \rangle .$$

As in Example 1.1 above, we stack the expansion coefficients into a vector so that

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \langle \mathbf{x}, \mathbf{e}_1 \rangle \\ \langle \mathbf{x}, \mathbf{e}_2 \rangle \end{bmatrix} = \begin{bmatrix} \mathbf{e}_1^{\mathsf{T}} \\ \mathbf{e}_2^{\mathsf{T}} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \mathbf{x}. \tag{1.5}$$

Analogously to Example 1.1, we can define the analysis matrix

$$\mathbf{T} \triangleq \begin{bmatrix} \mathbf{e}_1^{\mathsf{T}} \\ \mathbf{e}_2^{\mathsf{T}} \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix}$$

and rewrite (1.5) as

$$\mathbf{c} = \mathbf{T}\mathbf{x}.$$

Now, obviously, the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are not orthonormal (or, equivalently, $\mathbf{T}$ is not unitary) so that we cannot write $\mathbf{x}$ in the form (1.1). We could, however, try to find a decomposition of $\mathbf{x}$ of the form

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{e}_1 \rangle \tilde{\mathbf{e}}_1 + \langle \mathbf{x}, \mathbf{e}_2 \rangle \tilde{\mathbf{e}}_2 \tag{1.6}$$

with $\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2 \in \mathbb{R}^2$. That this is, indeed, possible is easily seen by rewriting (1.6) according to

$$\mathbf{x} = \begin{bmatrix} \tilde{\mathbf{e}}_1 & \tilde{\mathbf{e}}_2 \end{bmatrix} \mathbf{T}\mathbf{x} \tag{1.7}$$

and choosing the vectors $\tilde{\mathbf{e}}_1$ and $\tilde{\mathbf{e}}_2$ to be given by the columns of $\mathbf{T}^{-1}$ according to

$$\begin{bmatrix} \tilde{\mathbf{e}}_1 & \tilde{\mathbf{e}}_2 \end{bmatrix} = \mathbf{T}^{-1}. \tag{1.8}$$

Note that $\mathbf{T}$ is invertible as a consequence of $\mathbf{e}_1$ and $\mathbf{e}_2$ not being collinear. For the specific example at hand we find

$$\begin{bmatrix} \tilde{\mathbf{e}}_1 & \tilde{\mathbf{e}}_2 \end{bmatrix} = \mathbf{T}^{-1} = \begin{bmatrix} 1 & 0 \\ -1 & \sqrt{2} \end{bmatrix}$$

and therefore (see Figure 1.2)

$$\tilde{\mathbf{e}}_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}, \qquad \tilde{\mathbf{e}}_2 = \begin{bmatrix} 0 \\ \sqrt{2} \end{bmatrix}.$$

Note that (1.8) implies that $\mathbf{T} \begin{bmatrix} \tilde{\mathbf{e}}_1 & \tilde{\mathbf{e}}_2 \end{bmatrix} = \mathbf{I}_2$, which is equivalent to

$$\begin{bmatrix} \mathbf{e}_1^{\mathsf{T}} \\ \mathbf{e}_2^{\mathsf{T}} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}}_1 & \tilde{\mathbf{e}}_2 \end{bmatrix} = \mathbf{I}_2.$$

More directly the two sets of vectors $\{\mathbf{e}_1, \mathbf{e}_2\}$ and $\{\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2\}$ satisfy a “biorthonormality” property according to

$$\langle \mathbf{e}_j, \tilde{\mathbf{e}}_k \rangle = \begin{cases} 1, & j = k \\ 0, & \text{else} \end{cases}, \qquad j, k = 1, 2.$$

We say that $\{\mathbf{e}_1, \mathbf{e}_2\}$ and $\{\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2\}$ are biorthonormal bases. Analogously to (1.3), we can now define the synthesis matrix as follows:

$$\tilde{\mathbf{T}}^{\mathsf{T}} \triangleq \begin{bmatrix} \tilde{\mathbf{e}}_1 & \tilde{\mathbf{e}}_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ -1 & \sqrt{2} \end{bmatrix}.$$

Our observations can be summarized according to

$$\begin{aligned} \mathbf{x} &= \langle \mathbf{x}, \mathbf{e}_1 \rangle \tilde{\mathbf{e}}_1 + \langle \mathbf{x}, \mathbf{e}_2 \rangle \tilde{\mathbf{e}}_2 \\ &= \tilde{\mathbf{T}}^{\mathsf{T}} \mathbf{c} = \tilde{\mathbf{T}}^{\mathsf{T}} \mathbf{T} \mathbf{x} \\ &= \begin{bmatrix} 1 & 0 \\ -1 & \sqrt{2} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 1/\sqrt{2} & 1/\sqrt{2} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \mathbf{x}. \end{aligned} \tag{1.9}$$

Comparing (1.9) to (1.4), we observe the following: To synthesize $\mathbf{x}$ from the expansion coefficients $\mathbf{c}$ corresponding to the nonorthogonal set $\{\mathbf{e}_1, \mathbf{e}_2\}$, we need to use the synthesis matrix $\tilde{\mathbf{T}}^{\mathsf{T}}$ obtained from the set $\{\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2\}$, which forms a biorthonormal pair with $\{\mathbf{e}_1, \mathbf{e}_2\}$. In Example 1.1 $\{\mathbf{e}_1, \mathbf{e}_2\}$ is an orthonormal basis (ONB) and hence $\tilde{\mathbf{T}} = \mathbf{T}$, or, equivalently, $\{\mathbf{e}_1, \mathbf{e}_2\}$ forms a biorthonormal pair with itself.

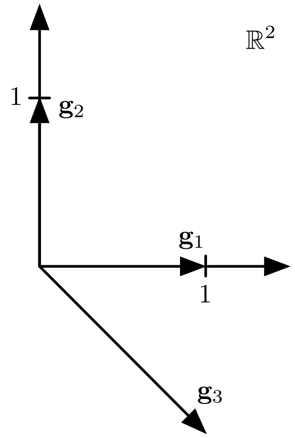


Figure 1.3: Overcomplete set of vectors in $\mathbb{R}^2$.

As the vectors $\mathbf{e}_1$ and $\mathbf{e}_2$ are linearly independent, the 2×2 analysis matrix $\mathbf{T}$ has full rank and is hence invertible, i.e., there is a *unique* matrix $\mathbf{T}^{-1}$ that satisfies $\mathbf{T}^{-1}\mathbf{T} = \mathbf{I}_2$. According to (1.7) this means that for each analysis set $\{\mathbf{e}_1, \mathbf{e}_2\}$ there is precisely one synthesis set $\{\tilde{\mathbf{e}}_1, \tilde{\mathbf{e}}_2\}$ such that (1.6) is satisfied for all $\mathbf{x} \in \mathbb{R}^2$.

So far we considered nonredundant signal expansions where the number of expansion coefficients is equal to the dimension of the Hilbert space. Often, however, redundancy in the expansion is desirable.

Example 1.3 (Overcomplete expansion in $\mathbb{R}^2$, [20, Ex. 3.1])**.** Consider the following three vectors in $\mathbb{R}^2$ (see Figure 1.3):

$$\mathbf{g}_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad \mathbf{g}_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{g}_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.$$

Three vectors in a two-dimensional space are always linearly dependent. In particular, in this example we have $\mathbf{g}_3 = \mathbf{g}_1 - \mathbf{g}_2$. Let us compute the expansion coefficients $\mathbf{c}$ corresponding to $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$:

$$\mathbf{c} = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \triangleq \begin{bmatrix} \langle \mathbf{x}, \mathbf{g}_1 \rangle \\ \langle \mathbf{x}, \mathbf{g}_2 \rangle \\ \langle \mathbf{x}, \mathbf{g}_3 \rangle \end{bmatrix} = \begin{bmatrix} \mathbf{g}_1^{\mathsf{T}} \\ \mathbf{g}_2^{\mathsf{T}} \\ \mathbf{g}_3^{\mathsf{T}} \end{bmatrix} \mathbf{x} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \end{bmatrix} \mathbf{x}. \tag{1.10}$$

Following Examples 1.1 and 1.2, we define the analysis matrix

$$\mathbf{T} \triangleq \begin{bmatrix} \mathbf{g}_1^{\mathsf{T}} \\ \mathbf{g}_2^{\mathsf{T}} \\ \mathbf{g}_3^{\mathsf{T}} \end{bmatrix} \triangleq \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \end{bmatrix}$$

and rewrite (1.10) as

$$\mathbf{c} = \mathbf{T}\mathbf{x}.$$

Note that here, unlike in Examples 1.1 and 1.2, $\mathbf{c}$ is a redundant representation of $\mathbf{x}$ as we have *three* expansion coefficients for a *two*-dimensional signal $\mathbf{x}$.

We next ask if $\mathbf{x}$ can be represented as a linear combination of the form

$$\mathbf{x} = \underbrace{\langle \mathbf{x}, \mathbf{g}_1 \rangle}_{c_1} \tilde{\mathbf{g}}_1 + \underbrace{\langle \mathbf{x}, \mathbf{g}_2 \rangle}_{c_2} \tilde{\mathbf{g}}_2 + \underbrace{\langle \mathbf{x}, \mathbf{g}_3 \rangle}_{c_3} \tilde{\mathbf{g}}_3 \tag{1.11}$$

with $\tilde{\mathbf{g}}_1, \tilde{\mathbf{g}}_2, \tilde{\mathbf{g}}_3 \in \mathbb{R}^2$? To answer this question (in the affirmative) we first note that the vectors $\mathbf{g}_1, \mathbf{g}_2$ form an orthonormal basis (ONB) for $\mathbb{R}^2$. We therefore know that the following is true:

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{g}_1 \rangle \mathbf{g}_1 + \langle \mathbf{x}, \mathbf{g}_2 \rangle \mathbf{g}_2. \tag{1.12}$$

Setting

$$\tilde{\mathbf{g}}_1 = \mathbf{g}_1, \ \tilde{\mathbf{g}}_2 = \mathbf{g}_2, \ \tilde{\mathbf{g}}_3 = \mathbf{0}$$

obviously yields a representation of the form (1.11). It turns out, however, that this representation is not unique and that an alternative representation of the form (1.11) can be obtained as follows. We start by adding zero to the right-hand side of (1.12):

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{g}_1 \rangle \mathbf{g}_1 + \langle \mathbf{x}, \mathbf{g}_2 \rangle \mathbf{g}_2 + \underbrace{\langle \mathbf{x}, \mathbf{g}_1 - \mathbf{g}_2 \rangle (\mathbf{g}_1 - \mathbf{g}_1)}_{\mathbf{0}}.$$

Rearranging terms in this expression, we obtain

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{g}_1 \rangle 2\mathbf{g}_1 + \langle \mathbf{x}, \mathbf{g}_2 \rangle (\mathbf{g}_2 - \mathbf{g}_1) - \langle \mathbf{x}, \mathbf{g}_1 - \mathbf{g}_2 \rangle \mathbf{g}_1. \tag{1.13}$$

We recognize that $\mathbf{g}_1 - \mathbf{g}_2 = \mathbf{g}_3$ and set

$$\tilde{\mathbf{g}}_1 = 2\mathbf{g}_1, \ \tilde{\mathbf{g}}_2 = \mathbf{g}_2 - \mathbf{g}_1, \ \tilde{\mathbf{g}}_3 = -\mathbf{g}_1. \tag{1.14}$$

This allows us to rewrite (1.13) as

$$\mathbf{x} = \langle \mathbf{x}, \mathbf{g}_1 \rangle \tilde{\mathbf{g}}_1 + \langle \mathbf{x}, \mathbf{g}_2 \rangle \tilde{\mathbf{g}}_2 + \langle \mathbf{x}, \mathbf{g}_3 \rangle \tilde{\mathbf{g}}_3.$$

The redundant set of vectors $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$ is called a *frame*. The set $\{\tilde{\mathbf{g}}_1, \tilde{\mathbf{g}}_2, \tilde{\mathbf{g}}_3\}$ in (1.14) is called a *dual frame* to the frame $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$. Obviously another dual frame is given by $\tilde{\mathbf{g}}_1 = \mathbf{g}_1$, $\tilde{\mathbf{g}}_2 = \mathbf{g}_2$, and $\tilde{\mathbf{g}}_3 = \mathbf{0}$. In fact, there are infinitely many dual frames. To see this, we first define the synthesis matrix corresponding to a dual frame $\{\tilde{\mathbf{g}}_1, \tilde{\mathbf{g}}_2, \tilde{\mathbf{g}}_3\}$ as

$$\tilde{\mathbf{T}}^{\mathsf{T}} \triangleq \begin{bmatrix} \tilde{\mathbf{g}}_1 & \tilde{\mathbf{g}}_2 & \tilde{\mathbf{g}}_3 \end{bmatrix}. \tag{1.15}$$

It then follows that we can write

$$\begin{aligned} \mathbf{x} &= \langle \mathbf{x}, \mathbf{g}_1 \rangle \tilde{\mathbf{g}}_1 + \langle \mathbf{x}, \mathbf{g}_2 \rangle \tilde{\mathbf{g}}_2 + \langle \mathbf{x}, \mathbf{g}_3 \rangle \tilde{\mathbf{g}}_3 \\ &= \tilde{\mathbf{T}}^{\mathsf{T}} \mathbf{c} = \tilde{\mathbf{T}}^{\mathsf{T}} \mathbf{T} \mathbf{x}, \end{aligned}$$

which implies that setting $\tilde{\mathbf{T}}^{\mathsf{T}} = [\tilde{\mathbf{g}}_1 \; \tilde{\mathbf{g}}_2 \; \tilde{\mathbf{g}}_3]$ to be a left-inverse of $\mathbf{T}$ yields a valid dual frame. Since $\mathbf{T}$ is a 3×2 ("tall") matrix, its left-inverse is not unique. In fact, $\mathbf{T}$ has infinitely many left-inverses (two of them were found above). Every left-inverse of $\mathbf{T}$ leads to a dual frame according to (1.15).

Thanks to the redundancy of the frame $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$, we obtain design freedom: In order to synthesize the signal $\mathbf{x}$ from its expansion coefficients $c_k = \langle \mathbf{x}, \mathbf{g}_k \rangle$, $k = 1, 2, 3$, in the frame $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$, we can choose between infinitely many dual frames $\{\tilde{\mathbf{g}}_1, \tilde{\mathbf{g}}_2, \tilde{\mathbf{g}}_3\}$. In practice the particular choice of the dual frame is usually dictated by the requirements of the specific problem at hand. We shall discuss this issue in detail in the context of sampling theory in Section 1.4.2.

1.2 Signal Expansions in Finite-Dimensional Hilbert Spaces

Motivated by the examples above, we now consider general signal expansions in finite-dimensional Hilbert spaces. As in the previous section, we first review the concept of an orthonormal basis (ONB), we then consider arbitrary (nonorthogonal) bases, and, finally, we discuss redundant vector sets — frames. While the discussion in this section is confined to the finite-dimensional case, we develop the general (possibly infinite-dimensional) case in Section 1.3.

1.2.1 Orthonormal Bases

We start by reviewing the concept of an ONB.

Definition 1.4. The set of vectors $\{\mathbf{e}_k\}_{k=1}^{M}$, $\mathbf{e}_k \in \mathbb{C}^M$, is called an ONB for $\mathbb{C}^M$ if

1. $\mathrm{span}\{\mathbf{e}_k\}_{k=1}^{M} = \{c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \ldots + c_M\mathbf{e}_M \,|\, c_1, c_2, \ldots, c_M \in \mathbb{C}\} = \mathbb{C}^M$

2. $$\langle \mathbf{e}_k, \mathbf{e}_j \rangle = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases} \qquad k, j = 1, \ldots, M.$$

When $\{\mathbf{e}_k\}_{k=1}^{M}$ is an ONB, thanks to the spanning property in Definition 1.4, every $\mathbf{x} \in \mathbb{C}^M$ can be decomposed as

$$\mathbf{x} = \sum_{k=1}^{M} c_k \mathbf{e}_k. \tag{1.16}$$

The expansion coefficients $\{c_k\}_{k=1}^{M}$ in (1.16) can be found through the following calculation:

$$\langle \mathbf{x}, \mathbf{e}_j \rangle = \left\langle \sum_{k=1}^{M} c_k \mathbf{e}_k, \mathbf{e}_j \right\rangle = \sum_{k=1}^{M} c_k \langle \mathbf{e}_k, \mathbf{e}_j \rangle = c_j.$$

In summary, we have the decomposition

$$\mathbf{x} = \sum_{k=1}^{M} \langle \mathbf{x}, \mathbf{e}_k \rangle \mathbf{e}_k.$$

Just like in Example 1.1, in the previous section, we define the analysis matrix

$$\mathbf{T} \triangleq \begin{bmatrix} \mathbf{e}_1^{\mathsf{H}} \\ \vdots \\ \mathbf{e}_M^{\mathsf{H}} \end{bmatrix}.$$

If we organize the inner products $\{\langle \mathbf{x}, \mathbf{e}_k \rangle\}_{k=1}^{M}$ into the vector $\mathbf{c}$, we have

$$\mathbf{c} \triangleq \begin{bmatrix} \langle \mathbf{x}, \mathbf{e}_1 \rangle \\ \vdots \\ \langle \mathbf{x}, \mathbf{e}_M \rangle \end{bmatrix} = \mathbf{T}\mathbf{x} = \begin{bmatrix} \mathbf{e}_1^{\mathsf{H}} \\ \vdots \\ \mathbf{e}_M^{\mathsf{H}} \end{bmatrix} \mathbf{x}.$$

Thanks to the orthonormality of the vectors $\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_M$ the matrix $\mathbf{T}$ is unitary, i.e., $\mathbf{T}^{\mathsf{H}} = \mathbf{T}^{-1}$ and hence

$$\mathbf{T}\mathbf{T}^{\mathsf{H}} = \begin{bmatrix} \mathbf{e}_1^{\mathsf{H}} \\ \vdots \\ \mathbf{e}_M^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} \mathbf{e}_1 & \dots & \mathbf{e}_M \end{bmatrix} = \begin{bmatrix} \langle \mathbf{e}_1, \mathbf{e}_1 \rangle & \cdots & \langle \mathbf{e}_M, \mathbf{e}_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \mathbf{e}_1, \mathbf{e}_M \rangle & \cdots & \langle \mathbf{e}_M, \mathbf{e}_M \rangle \end{bmatrix} = \mathbf{I}_M = \mathbf{T}^{\mathsf{H}}\mathbf{T}.$$

Thus, if we multiply the vector $\mathbf{c}$ by $\mathbf{T}^{\mathsf{H}}$, we synthesize $\mathbf{x}$ according to

$$\mathbf{T}^{\mathsf{H}}\mathbf{c} = \mathbf{T}^{\mathsf{H}}\mathbf{T}\mathbf{x} = \sum_{k=1}^{M} \langle \mathbf{x}, \mathbf{e}_k \rangle \, \mathbf{e}_k = \mathbf{I}_M \mathbf{x} = \mathbf{x}. \tag{1.17}$$

We shall therefore call the matrix $\mathbf{T}^{\mathsf{H}}$ the synthesis matrix, corresponding to the analysis matrix $\mathbf{T}$. In the ONB case considered here the synthesis matrix is simply the Hermitian adjoint of the analysis matrix.

1.2.2 General Bases

We next relax the orthonormality property, i.e., the second condition in Definition 1.4, and consider general bases.

Definition 1.5. The set of vectors $\{\mathbf{e}_k\}_{k=1}^{M}$, $\mathbf{e}_k \in \mathbb{C}^M$, is a basis for $\mathbb{C}^M$ if

1. $\mathrm{span}\{\mathbf{e}_k\}_{k=1}^{M} = \{c_1\mathbf{e}_1 + c_2\mathbf{e}_2 + \ldots + c_M\mathbf{e}_M \,|\, c_1, c_2, \dots, c_M \in \mathbb{C}\} = \mathbb{C}^M$

2. $\{\mathbf{e}_k\}_{k=1}^{M}$ is a linearly independent set, i.e., if $\sum_{k=1}^{M} c_k\mathbf{e}_k = \mathbf{0}$ for some scalar coefficients $\{c_k\}_{k=1}^{M}$, then necessarily $c_k = 0$ for all $k = 1, \dots, M$.

Now consider a signal $\mathbf{x} \in \mathbb{C}^M$ and compute the expansion coefficients

$$c_k \triangleq \langle \mathbf{x}, \mathbf{e}_k \rangle, \quad k = 1, \dots, M. \tag{1.18}$$

Again, it is convenient to introduce the analysis matrix

$$\mathbf{T} \triangleq \begin{bmatrix} \mathbf{e}_1^{\mathsf{H}} \\ \vdots \\ \mathbf{e}_M^{\mathsf{H}} \end{bmatrix}$$

and to stack the coefficients $\{c_k\}_{k=1}^{M}$ in the vector $\mathbf{c}$. Then (1.18) can be written as

$$\mathbf{c} = \mathbf{T}\mathbf{x}.$$

Next, let us ask how we can find a set of vectors $\{\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_M\}$, $\tilde{\mathbf{e}}_k \in \mathbb{C}^M$, $k = 1, \ldots, M$, that is dual to the set $\{\mathbf{e}_1, \ldots, \mathbf{e}_M\}$ in the sense that

$$\mathbf{x} = \sum_{k=1}^{M} c_k \tilde{\mathbf{e}}_k = \sum_{k=1}^{M} \langle \mathbf{x}, \mathbf{e}_k \rangle \, \tilde{\mathbf{e}}_k \tag{1.19}$$

for all $\mathbf{x} \in \mathbb{C}^M$. If we introduce the synthesis matrix

$$\tilde{\mathbf{T}}^{\mathsf{H}} \triangleq [\tilde{\mathbf{e}}_1 \, \cdots \, \tilde{\mathbf{e}}_M],$$

we can rewrite (1.19) in vector-matrix notation as follows

$$\mathbf{x} = \tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{c} = \tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T}\mathbf{x}.$$

This shows that finding vectors $\tilde{\mathbf{e}}_1, \ldots, \tilde{\mathbf{e}}_M$ that satisfy (1.19) is equivalent to finding the inverse of the analysis matrix $\mathbf{T}$ and setting $\tilde{\mathbf{T}}^{\mathsf{H}} = \mathbf{T}^{-1}$. Thanks to the linear independence of the vectors $\{\mathbf{e}_k\}_{k=1}^{M}$, the matrix $\mathbf{T}$ has full rank and is, therefore, invertible.

Summarizing our findings, we conclude that in the case of a basis $\{\mathbf{e}_k\}_{k=1}^{M}$, the analysis matrix and the synthesis matrix are inverses of each other, i.e., $\tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T} = \mathbf{T}\tilde{\mathbf{T}}^{\mathsf{H}} = \mathbf{I}_M$. Recall that in the case of an ONB the analysis matrix $\mathbf{T}$ is *unitary* and hence its inverse is simply given by $\mathbf{T}^{\mathsf{H}}$ [see (1.17)], so that in this case $\tilde{\mathbf{T}} = \mathbf{T}$.

Next, note that $\mathbf{T}\tilde{\mathbf{T}}^{\mathsf{H}} = \mathbf{I}_M$ is equivalent to

$$\begin{bmatrix} \mathbf{e}_1^{\mathsf{H}} \\ \vdots \\ \mathbf{e}_M^{\mathsf{H}} \end{bmatrix} \begin{bmatrix} \tilde{\mathbf{e}}_1 & \ldots & \tilde{\mathbf{e}}_M \end{bmatrix} = \begin{bmatrix} \langle \tilde{\mathbf{e}}_1, \mathbf{e}_1 \rangle & \cdots & \langle \tilde{\mathbf{e}}_M, \mathbf{e}_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle \tilde{\mathbf{e}}_1, \mathbf{e}_M \rangle & \cdots & \langle \tilde{\mathbf{e}}_M, \mathbf{e}_M \rangle \end{bmatrix} = \mathbf{I}_M$$

or equivalently

$$\langle \mathbf{e}_k, \tilde{\mathbf{e}}_j \rangle = \begin{cases} 1, & k = j \\ 0, & \text{else} \end{cases}, \qquad k, j = 1, \ldots, M. \tag{1.20}$$

The sets $\{\mathbf{e}_k\}_{k=1}^{M}$ and $\{\tilde{\mathbf{e}}_k\}_{k=1}^{M}$ are biorthonormal bases. ONBs are biorthonormal to themselves in this terminology, as already noted in Example 1.2. We emphasize that it is the fact that $\mathbf{T}$ and

$\tilde{\mathbf{T}}^{\mathsf{H}}$ are square and full-rank that allows us to conclude that $\tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T} = \mathbf{I}_M$ implies $\mathbf{T}\tilde{\mathbf{T}}^{\mathsf{H}} = \mathbf{I}_M$ and hence to conclude that (1.20) holds. We shall see below that for redundant expansions $\mathbf{T}$ is a tall matrix and $\tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T} \neq \mathbf{T}\tilde{\mathbf{T}}^{\mathsf{H}}$ ($\tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T}$ and $\mathbf{T}\tilde{\mathbf{T}}^{\mathsf{H}}$ have different dimensions) so that dual frames will not be biorthonormal.

As $\mathbf{T}$ is a square matrix and of full rank, its inverse is *unique*, which means that for a given analysis set $\{\mathbf{e}_k\}_{k=1}^{M}$, the synthesis set $\{\tilde{\mathbf{e}}_k\}_{k=1}^{M}$ is *unique*. Alternatively, for a given synthesis set $\{\tilde{\mathbf{e}}_k\}_{k=1}^{M}$, there is a *unique* analysis set $\{\mathbf{e}_k\}_{k=1}^{M}$. This uniqueness property is not always desirable. For example, one may want to impose certain structural properties on the synthesis set $\{\tilde{\mathbf{e}}_k\}_{k=1}^{M}$ in which case having freedom in choosing the synthesis set as in Example 1.2 is helpful.

An important property of ONBs is that they are norm-preserving: The norm of the coefficient vector $\mathbf{c}$ is equal to the norm of the signal $\mathbf{x}$. This can easily be seen by noting that

$$\|\mathbf{c}\|^2 = \mathbf{c}^{\mathsf{H}}\mathbf{c} = \mathbf{x}^{\mathsf{H}}\mathbf{T}^{\mathsf{H}}\mathbf{T}\mathbf{x} = \mathbf{x}^{\mathsf{H}}\mathbf{I}_M\mathbf{x} = \|\mathbf{x}\|^2, \tag{1.21}$$

where we used (1.17). Biorthonormal bases are *not* norm-preserving, in general. Rather, the equality in (1.21) is relaxed to a double-inequality, by application of the Rayleigh-Ritz theorem [22, Sec. 9.7.2.2] according to

$$\lambda_{\text{min}}\big(\mathbf{T}^{\mathsf{H}}\mathbf{T}\big)\,\|\mathbf{x}\|^2 \leq \|\mathbf{c}\|^2 = \mathbf{x}^{\mathsf{H}}\mathbf{T}^{\mathsf{H}}\mathbf{T}\mathbf{x} \leq \lambda_{\text{max}}\big(\mathbf{T}^{\mathsf{H}}\mathbf{T}\big)\,\|\mathbf{x}\|^2. \tag{1.22}$$

1.2.3 Redundant Signal Expansions

The signal expansions we considered so far are non-redundant in the sense that the number of expansion coefficients equals the dimension of the Hilbert space. Such signal expansions have a number of disadvantages. First, corruption or loss of expansion coefficients can result in significant reconstruction errors. Second, the reconstruction process is very rigid: As we have seen in Section 1.2.2, for each set of analysis vectors, there is a *unique* set of synthesis vectors. In practical applications it is often desirable to impose additional constraints on the reconstruction functions, such as smoothness properties or structural properties that allow for computationally efficient reconstruction.

Redundant expansions allow to overcome many of these problems as they offer design freedom and robustness to corruption or loss of expansion coefficients. We already saw in Example 1.3 that in the case of redundant expansions, for a given set of analysis vectors the set of synthesis vectors that allows perfect recovery of a signal from its expansion coefficients is not unique; in fact, there are infinitely many sets of synthesis vectors, in general. This results in design freedom and provides robustness. Suppose that the expansion coefficient $c_3 = \langle \mathbf{x}, \mathbf{g}_3 \rangle$ in Example 1.3 is corrupted or even completely lost. We can still reconstruct $\mathbf{x}$ *exactly* from (1.12).

Now, let us turn to developing the general theory of redundant signal expansions in finite-dimensional Hilbert spaces. Consider a set of N vectors $\{\mathbf{g}_1, \ldots, \mathbf{g}_N\}$, $\mathbf{g}_k \in \mathbb{C}^M$, $k = 1, \ldots, N$, with $N \geq M$. Clearly, when N is *strictly* greater than M, the vectors $\mathbf{g}_1, \ldots, \mathbf{g}_N$ must be linearly dependent. Next, consider a signal $\mathbf{x} \in \mathbb{C}^M$ and compute the expansion coefficients

$$c_k = \langle \mathbf{x}, \mathbf{g}_k \rangle, \quad k = 1, \ldots, N. \tag{1.23}$$

Just as before, it is convenient to introduce the analysis matrix

$$\mathbf{T} \triangleq \begin{bmatrix} \mathbf{g}_1^{\mathsf{H}} \\ \vdots \\ \mathbf{g}_N^{\mathsf{H}} \end{bmatrix} \tag{1.24}$$

and to stack the coefficients $\{c_k\}_{k=1}^N$ in the vector $\mathbf{c}$. Then (1.23) can be written as

$$\mathbf{c} = \mathbf{T}\mathbf{x}. \tag{1.25}$$

Note that $\mathbf{c} \in \mathbb{C}^N$ and $\mathbf{x} \in \mathbb{C}^M$. Differently from ONBs and biorthonormal bases considered in Sections 1.2.1 and 1.2.2, respectively, in the case of redundant expansions, the signal $\mathbf{x}$ and the expansion coefficient vector $\mathbf{c}$ will, in general, belong to different Hilbert spaces.

The question now is: How can we find a set of vectors $\{\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_N\}$, $\tilde{\mathbf{g}}_k \in \mathbb{C}^M$, $k = 1, \ldots, N$, such that

$$\mathbf{x} = \sum_{k=1}^{N} c_k \tilde{\mathbf{g}}_k = \sum_{k=1}^{N} \langle \mathbf{x}, \mathbf{g}_k \rangle \tilde{\mathbf{g}}_k \tag{1.26}$$

for all $\mathbf{x} \in \mathbb{C}^M$? If we introduce the synthesis matrix

$$\tilde{\mathbf{T}}^{\mathsf{H}} \triangleq [\tilde{\mathbf{g}}_1 \; \cdots \; \tilde{\mathbf{g}}_N],$$

we can rewrite (1.26) in vector-matrix notation as follows

$$\mathbf{x} = \tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{c} = \tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T}\mathbf{x}. \tag{1.27}$$

Finding vectors $\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_N$ that satisfy (1.26) for all $\mathbf{x} \in \mathbb{C}^M$ is therefore equivalent to finding a left-inverse $\tilde{\mathbf{T}}^{\mathsf{H}}$ of $\mathbf{T}$, i.e.,

$$\tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T} = \mathbf{I}_M.$$

First note that $\mathbf{T}$ is left-invertible if and only if $\mathbb{C}^M = \mathrm{span}\{\mathbf{g}_k\}_{k=1}^N$, i.e., if and only if the set of vectors $\{\mathbf{g}_k\}_{k=1}^N$ spans $\mathbb{C}^M$. Next observe that when $N > M$, the $N \times M$ matrix $\mathbf{T}$ is a "tall" matrix, and therefore its left-inverse is, in general, not unique. In fact, there are infinitely many left-inverses. The following theorem [23, Ch. 2, Th. 1] provides a convenient parametrization of all these left-inverses.

Theorem 1.6. *Let* $\mathbf{A} \in \mathbb{C}^{N\times M}$, $N \geq M$. *Assume that* $\mathrm{rank}(\mathbf{A}) = M$. *Then* $\mathbf{A}^\dagger \triangleq (\mathbf{A}^{\mathsf{H}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{H}}$ *is a left-inverse of* $\mathbf{A}$, *i.e.,* $\mathbf{A}^\dagger\mathbf{A} = \mathbf{I}_M$. *Moreover, the general solution* $\mathbf{L} \in \mathbb{C}^{M\times N}$ *of the equation* $\mathbf{L}\mathbf{A} = \mathbf{I}_M$ *is given by*

$$\mathbf{L} = \mathbf{A}^\dagger + \mathbf{M}(\mathbf{I}_N - \mathbf{A}\mathbf{A}^\dagger), \tag{1.28}$$

where $\mathbf{M} \in \mathbb{C}^{M\times N}$ *is an arbitrary matrix.*

Proof. Since $\operatorname{rank}(\mathbf{A}) = M$, the matrix $\mathbf{A}^{\mathsf{H}}\mathbf{A}$ is invertible and hence $\mathbf{A}^{\dagger}$ is well defined. Now, let us verify that $\mathbf{A}^{\dagger}$ is, indeed, a left-inverse of $\mathbf{A}$:

$$\mathbf{A}^{\dagger}\mathbf{A} = (\mathbf{A}^{\mathsf{H}}\mathbf{A})^{-1}\mathbf{A}^{\mathsf{H}}\mathbf{A} = \mathbf{I}_M. \tag{1.29}$$

The matrix $\mathbf{A}^{\dagger}$ is called the Moore-Penrose inverse of $\mathbf{A}$.

Next, we show that every matrix $\mathbf{L}$ of the form (1.28) is a valid left-inverse of $\mathbf{A}$:

$$\begin{aligned}\mathbf{L}\mathbf{A} &= \left(\mathbf{A}^{\dagger} + \mathbf{M}(\mathbf{I}_N - \mathbf{A}\mathbf{A}^{\dagger})\right)\mathbf{A} \\ &= \underbrace{\mathbf{A}^{\dagger}\mathbf{A}}_{\mathbf{I}_M} + \mathbf{M}\mathbf{A} - \mathbf{M}\mathbf{A}\underbrace{\mathbf{A}^{\dagger}\mathbf{A}}_{\mathbf{I}_M} \\ &= \mathbf{I}_M + \mathbf{M}\mathbf{A} - \mathbf{M}\mathbf{A} = \mathbf{I}_M,\end{aligned}$$

where we used (1.29) twice.

Finally, assume that $\mathbf{L}$ is a valid left-inverse of $\mathbf{A}$, i.e., $\mathbf{L}$ is a solution of the equation $\mathbf{L}\mathbf{A} = \mathbf{I}_M$. We show that $\mathbf{L}$ can be written in the form (1.28). Multiplying the equation $\mathbf{L}\mathbf{A} = \mathbf{I}_M$ by $\mathbf{A}^{\dagger}$ from the right, we have

$$\mathbf{L}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger}.$$

Adding $\mathbf{L}$ to both sides of this equation and rearranging terms yields

$$\mathbf{L} = \mathbf{A}^{\dagger} + \mathbf{L} - \mathbf{L}\mathbf{A}\mathbf{A}^{\dagger} = \mathbf{A}^{\dagger} + \mathbf{L}(\mathbf{I}_N - \mathbf{A}\mathbf{A}^{\dagger}),$$

which shows that $\mathbf{L}$ can be written in the form (1.28) (with $\mathbf{M} = \mathbf{L}$), as required. □

We conclude that for each redundant set of vectors $\{\mathbf{g}_1, \dots, \mathbf{g}_N\}$ that spans $\mathbb{C}^M$, there are infinitely many dual sets $\{\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_N\}$ such that the decomposition (1.26) holds for all $\mathbf{x} \in \mathbb{C}^M$. These dual sets are obtained by identifying $\{\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_N\}$ with the columns of $\mathbf{L}$ according to

$$[\tilde{\mathbf{g}}_1 \ \cdots \ \tilde{\mathbf{g}}_N] = \mathbf{L},$$

where $\mathbf{L}$ can be written as follows

$$\mathbf{L} = \mathbf{T}^{\dagger} + \mathbf{M}(\mathbf{I}_N - \mathbf{T}\mathbf{T}^{\dagger})$$

and $\mathbf{M} \in \mathbb{C}^{M \times N}$ is an arbitrary matrix.

The dual set $\{\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_N\}$ corresponding to the Moore-Penrose inverse $\mathbf{L} = \mathbf{T}^{\dagger}$ of the matrix $\mathbf{T}$, i.e.,

$$[\tilde{\mathbf{g}}_1 \ \cdots \ \tilde{\mathbf{g}}_N] = \mathbf{T}^{\dagger} = (\mathbf{T}^{\mathsf{H}}\mathbf{T})^{-1}\mathbf{T}^{\mathsf{H}}$$

is called the *canonical dual* of $\{\mathbf{g}_1, \dots, \mathbf{g}_N\}$. Using (1.24), we see that in this case

$$\tilde{\mathbf{g}}_k = (\mathbf{T}^{\mathsf{H}}\mathbf{T})^{-1}\mathbf{g}_k, \quad k = 1, \dots, N. \tag{1.30}$$

Note that *unlike* in the case of a basis, the equation $\tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T} = \mathbf{I}_M$ *does not* imply that the sets $\{\tilde{\mathbf{g}}_k\}_{k=1}^N$ and $\{\mathbf{g}_k\}_{k=1}^N$ are biorthonormal. This is because the matrix $\mathbf{T}$ is *not* a square matrix, and thus, $\tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T} \neq \mathbf{T}\tilde{\mathbf{T}}^{\mathsf{H}}$ ($\tilde{\mathbf{T}}^{\mathsf{H}}\mathbf{T}$ and $\mathbf{T}\tilde{\mathbf{T}}^{\mathsf{H}}$ have different dimensions).

Similar to biorthonormal bases, redundant sets of vectors are, in general, not norm-preserving. Indeed, from (1.25) we see that

$$\|\mathbf{c}\|^2 = \mathbf{x}^{\mathsf{H}}\mathbf{T}^{\mathsf{H}}\mathbf{T}\mathbf{x}$$

and thus, by the Rayleigh-Ritz theorem [22, Sec. 9.7.2.2], we have

$$\lambda_{\text{min}}\big(\mathbf{T}^{\mathsf{H}}\mathbf{T}\big)\, \|\mathbf{x}\|^2 \leq \|\mathbf{c}\|^2 \leq \lambda_{\text{max}}\big(\mathbf{T}^{\mathsf{H}}\mathbf{T}\big)\, \|\mathbf{x}\|^2 \tag{1.31}$$

as in the case of biorthonormal bases.

We already saw some of the basic issues that a theory of orthonormal, biorthonormal, and redundant signal expansions should address. It should account for the signals and the expansion coefficients belonging, potentially, to different Hilbert spaces; it should account for the fact that for a given analysis set, the synthesis set is not unique in the redundant case, it should prescribe how synthesis vectors can be obtained from the analysis vectors. Finally, it should apply not only to finite-dimensional Hilbert spaces, as considered so far, but also to infinite-dimensional Hilbert spaces. We now proceed to develop this general theory, known as the theory of frames.

1.3 Frames for General Hilbert Spaces

Let $\{g_k\}_{k\in\mathcal{K}}$ ($\mathcal{K}$ is a countable set) be a set of elements taken from the Hilbert space $\mathcal{H}$. Note that this set need not be orthogonal.

In developing a general theory of signal expansions in Hilbert spaces, as outlined at the end of the previous section, we start by noting that the central quantity in Section 1.2 was the analysis matrix $\mathbf{T}$ associated to the (possibly nonorthogonal or redundant) set of vectors $\{\mathbf{g}_1, \ldots, \mathbf{g}_N\}$. Now matrices are nothing but linear operators in finite-dimensional Hilbert spaces. In formulating frame theory for general (possibly infinite-dimensional) Hilbert spaces, it is therefore sensible to define the analysis operator $\mathbb{T}$ that assigns to each signal $x \in \mathcal{H}$ the sequence of inner products $\mathbb{T}x = \{\langle x, g_k\rangle\}_{k\in\mathcal{K}}$. Throughout this section, we assume that $\{g_k\}_{k\in\mathcal{K}}$ is a Bessel sequence, i.e., $\sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 < \infty$ for all $x \in \mathcal{H}$.

Definition 1.7. The linear operator $\mathbb{T}$ is defined as the operator that maps the Hilbert space $\mathcal{H}$ into the space l^2 of square-summable complex sequences[1], $\mathbb{T} : \mathcal{H} \to l^2$, by assigning to each signal $x \in \mathcal{H}$ the sequence of inner products $\langle x, g_k\rangle$ according to

$$\mathbb{T} : x \to \{\langle x, g_k\rangle\}_{k\in\mathcal{K}}.$$

[1] The fact that the range space of $\mathbb{T}$ is contained in l^2 is a consequence of $\{g_k\}_{k\in\mathcal{K}}$ being a Bessel sequence.

Note that $\|\mathbb{T}x\|^2 = \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2$, i.e., the energy $\|\mathbb{T}x\|^2$ of $\mathbb{T}x$ can be expressed as

$$\|\mathbb{T}x\|^2 = \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 . \tag{1.32}$$

We shall next formulate properties that the set $\{g_k\}_{k\in\mathcal{K}}$ and hence the operator $\mathbb{T}$ should satisfy if we have signal expansions in mind:

1. The signal x can be perfectly reconstructed from the coefficients $\{\langle x, g_k\rangle\}_{k\in\mathcal{K}}$. This means that we want $\langle x, g_k\rangle = \langle y, g_k\rangle$, for all $k \in \mathcal{K}$, (i.e., $\mathbb{T}x = \mathbb{T}y$) to imply that $x = y$, for all $x, y \in \mathcal{H}$. In other words, the operator $\mathbb{T}$ has to be left-invertible, which means that $\mathbb{T}$ is invertible on its range space $\mathcal{R}(\mathbb{T}) = \{y \in l^2 : y = \mathbb{T}x,\ x \in \mathcal{H}\}$.

   This requirement will clearly be satisfied if we demand that there exist a constant $A > 0$ such that for all $x, y \in \mathcal{H}$ we have

   $$A\|x - y\|^2 \leq \|\mathbb{T}x - \mathbb{T}y\|^2.$$

   Setting $z = x - y$ and using the linearity of $\mathbb{T}$, we see that this condition is equivalent to

   $$A\|z\|^2 \leq \|\mathbb{T}z\|^2 \tag{1.33}$$

   for all $z \in \mathcal{H}$ with $A > 0$.

2. The energy in the sequence of expansion coefficients $\mathbb{T}x = \{\langle x, g_k\rangle\}_{k\in\mathcal{K}}$ should be related to the energy in the signal x. For example, we saw in (1.21) that if $\{\mathbf{e}_k\}_{k=1}^M$ is an ONB for $\mathbb{C}^M$, then

   $$\|\mathbf{T}\mathbf{x}\|^2 = \sum_{k=1}^{M} |\langle \mathbf{x}, \mathbf{e}_k\rangle|^2 = \|\mathbf{x}\|^2, \text{ for all } \mathbf{x} \in \mathbb{C}^M. \tag{1.34}$$

   This property is a consequence of the unitarity of $\mathbb{T} = \mathbf{T}$ and it is clear that it will not hold for general sets $\{g_k\}_{k\in\mathcal{K}}$ (see the discussion around (1.22) and (1.31)). Instead, we will relax (1.34) to demand that for all $x \in \mathcal{H}$ there exist a finite constant B such that[2]

   $$\|\mathbb{T}x\|^2 = \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 \leq B\|x\|^2. \tag{1.35}$$

Together with (1.33) this "sandwiches" the quantity $\|\mathbb{T}x\|^2$ according to

$$A\|x\|^2 \leq \|\mathbb{T}x\|^2 \leq B\|x\|^2.$$

We are now ready to formally define a frame for the Hilbert space $\mathcal{H}$.

[2] Note that if (1.35) is satisfied with $B < \infty$, then $\{g_k\}_{k\in\mathcal{K}}$ is a Bessel sequence.

Definition 1.8. A set of elements $\{g_k\}_{k\in\mathcal{K}},\ g_k \in \mathcal{H},\ k \in \mathcal{K}$, is called a frame for the Hilbert space $\mathcal{H}$ if

$$A\|x\|^2 \leq \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 \leq B\|x\|^2, \quad \text{for all} \quad x \in \mathcal{H}, \tag{1.36}$$

with $A, B \in \mathbb{R}$ and $0 < A \leq B < \infty$. Valid constants A and B are called frame bounds. The largest valid constant A and the smallest valid constant B are called *the* (tightest possible) *frame bounds*.

Let us next consider some simple examples of frames.

Example 1.9 ([21])**.** Let $\{e_k\}_{k=1}^{\infty}$ be an ONB for an infinite-dimensional Hilbert space $\mathcal{H}$. By repeating each element in $\{e_k\}_{k=1}^{\infty}$ once, we obtain the redundant set

$$\{g_k\}_{k=1}^{\infty} = \{e_1, e_1, e_2, e_2, \ldots\}.$$

To see that this set is a frame for $\mathcal{H}$, we note that because $\{e_k\}_{k=1}^{\infty}$ is an ONB, for all $x \in \mathcal{H}$, we have

$$\sum_{k=1}^{\infty} |\langle x, e_k\rangle|^2 = \|x\|^2$$

and therefore

$$\sum_{k=1}^{\infty} |\langle x, g_k\rangle|^2 = \sum_{k=1}^{\infty} |\langle x, e_k\rangle|^2 + \sum_{k=1}^{\infty} |\langle x, e_k\rangle|^2 = 2\|x\|^2.$$

This verifies the frame condition (1.36) and shows that the frame bounds are given by $A = B = 2$.

Example 1.10 ([21])**.** Starting from the ONB $\{e_k\}_{k=1}^{\infty}$, we can construct another redundant set as follows

$$\{g_k\}_{k=1}^{\infty} = \left\{e_1, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{2}}e_2, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \frac{1}{\sqrt{3}}e_3, \ldots\right\}.$$

To see that the set $\{g_k\}_{k=1}^{\infty}$ is a frame for $\mathcal{H}$, take an arbitrary $x \in \mathcal{H}$ and note that

$$\sum_{k=1}^{\infty} |\langle x, g_k\rangle|^2 = \sum_{k=1}^{\infty} k \left|\left\langle x, \frac{1}{\sqrt{k}}e_k\right\rangle\right|^2 = \sum_{k=1}^{\infty} k\frac{1}{k} |\langle x, e_k\rangle|^2 = \sum_{k=1}^{\infty} |\langle x, e_k\rangle|^2 = \|x\|^2.$$

We conclude that $\{g_k\}_{k=1}^{\infty}$ is a frame with the frame bounds $A = B = 1$.

From (1.32) it follows that an equivalent formulation of (1.36) is

$$A\|x\|^2 \leq \|\mathbb{T}x\|^2 \leq B\|x\|^2, \quad \text{for all} \quad x \in \mathcal{H}.$$

This means that the energy in the coefficient sequence $\mathbb{T}x$ is bounded above and below by bounds that are proportional to the signal energy. The existence of a lower frame bound $A > 0$ guarantees that the linear operator $\mathbb{T}$ is left-invertible, i.e., our first requirement above is satisfied. Besides that it also guarantees completeness of the set $\{g_k\}_{k\in\mathcal{K}}$ for $\mathcal{H}$, as we shall see next. To this end, we first recall the following definition:

Definition 1.11. A set of elements $\{g_k\}_{k\in\mathcal{K}}$, $g_k \in \mathcal{H}$, $k \in \mathcal{K}$, is complete for the Hilbert space $\mathcal{H}$ if $\langle x, g_k\rangle = 0$ for all $k \in \mathcal{K}$ and with $x \in \mathcal{H}$ implies $x = 0$, i.e., the only element in $\mathcal{H}$ that is orthogonal to all g_k, is $x = 0$.

To see that the frame $\{g_k\}_{k\in\mathcal{K}}$ is complete for $\mathcal{H}$, take an arbitrary signal $x \in \mathcal{H}$ and assume that $\langle x, g_k\rangle = 0$ for all $k \in \mathcal{K}$. Due to the existence of a lower frame bound $A > 0$ we have

$$A\|x\|^2 \leq \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 = 0,$$

which implies $\|x\|^2 = 0$ and hence $x = 0$.

Finally, note that the existence of an upper frame bound $B < \infty$ guarantees that $\mathbb{T}$ is a bounded linear operator[3] (see [1, Def. 2.7-1]), and, therefore (see [1, Th. 2.7-9]), continuous[4] (see [1, Sec. 2.7]).

Recall that we would like to find a general method to reconstruct a signal $x \in \mathcal{H}$ from its expansion coefficients $\{\langle x, g_k\rangle\}_{k\in\mathcal{K}}$. In Section 1.2.3 we saw that in the finite-dimensional case, this can be accomplished according to

$$\mathbf{x} = \sum_{k=1}^{N} \langle \mathbf{x}, \mathbf{g}_k\rangle\, \tilde{\mathbf{g}}_k.$$

Here $\{\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_N\}$ can be chosen to be the canonical dual to the set $\{\mathbf{g}_1, \ldots, \mathbf{g}_N\}$ obtained as follows: $\tilde{\mathbf{g}}_k = (\mathbf{T}^{\mathsf{H}}\mathbf{T})^{-1}\mathbf{g}_k$, $k = 1, \ldots, N$. We already know that $\mathbb{T}$ is the generalization of $\mathbf{T}$ to the infinite-dimensional setting. Which operator will then correspond to $\mathbf{T}^{\mathsf{H}}$? To answer this question we start with a definition.

Definition 1.12. The linear operator $\mathbb{T}^{\times}$ is defined as

$$\begin{aligned} &\mathbb{T}^{\times} : l^2 \to \mathcal{H} \\ &\mathbb{T}^{\times} : \{c_k\}_{k\in\mathcal{K}} \to \sum_{k\in\mathcal{K}} c_k g_k. \end{aligned}$$

Next, we recall the definition of the adjoint of an operator.

Definition 1.13. Let $\mathbb{A} : \mathcal{H} \to \mathcal{H}'$ be a bounded linear operator between the Hilbert spaces $\mathcal{H}$ and $\mathcal{H}'$. The unique bounded linear operator $\mathbb{A}^* : \mathcal{H}' \to \mathcal{H}$ that satisfies

$$\langle \mathbb{A}x, y\rangle = \langle x, \mathbb{A}^* y\rangle \tag{1.37}$$

for all $x \in \mathcal{H}$ and all $y \in \mathcal{H}'$ is called the adjoint of $\mathbb{A}$.

[3] Let $\mathcal{H}$ and $\mathcal{H}'$ be Hilbert spaces and $\mathbb{A} : \mathcal{H} \to \mathcal{H}'$ a linear operator. The operator $\mathbb{A}$ is said to be *bounded* if there exists a finite number c such that for all $x \in \mathcal{H}$, $\|\mathbb{A}x\| \leq c\|x\|$.

[4] Let $\mathcal{H}$ and $\mathcal{H}'$ be Hilbert spaces and $\mathbb{A} : \mathcal{H} \to \mathcal{H}'$ a linear operator. The operator $\mathbb{A}$ is said to be *continuous* at a point $x_0 \in \mathcal{H}$ if for every $\epsilon > 0$ there is a $\delta > 0$ such that for all $x \in \mathcal{H}$ satisfying $\|x - x_0\| < \delta$ it follows that $\|\mathbb{A}x - \mathbb{A}x_0\| < \epsilon$. The operator $\mathbb{A}$ is said to be continuous on $\mathcal{H}$, if it is continuous at every point $x_0 \in \mathcal{H}$.

Note that the concept of the adjoint of an operator directly generalizes that of the Hermitian transpose of a matrix: if $\mathbf{A} \in \mathbb{C}^{N \times M}$, $\mathbf{x} \in \mathbb{C}^M$, $\mathbf{y} \in \mathbb{C}^N$, then

$$\langle \mathbf{A}\mathbf{x}, \mathbf{y} \rangle = \mathbf{y}^{\mathsf{H}}\mathbf{A}\mathbf{x} = (\mathbf{A}^{\mathsf{H}}\mathbf{y})^{\mathsf{H}}\mathbf{x} = \left\langle \mathbf{x}, \mathbf{A}^{\mathsf{H}}\mathbf{y} \right\rangle,$$

which, comparing to (1.37), shows that $\mathbf{A}^{\mathsf{H}}$ corresponds to $\mathbb{A}^*$.

We shall next show that the operator $\mathbb{T}^\times$ defined above is nothing but the adjoint $\mathbb{T}^*$ of the operator $\mathbb{T}$. To see this consider an arbitrary sequence $\{c_k\}_{k\in\mathcal{K}} \in l^2$ and an arbitrary signal $x \in \mathcal{H}$. We have to prove that

$$\langle \mathbb{T}x, \{c_k\}_{k\in\mathcal{K}} \rangle = \left\langle x, \mathbb{T}^\times \{c_k\}_{k\in\mathcal{K}} \right\rangle.$$

This can be established by noting that

$$\begin{aligned}\langle \mathbb{T}x, \{c_k\}_{k\in\mathcal{K}} \rangle &= \sum_{k\in\mathcal{K}} \langle x, g_k \rangle \, c_k^* \\ \left\langle x, \mathbb{T}^\times \{c_k\}_{k\in\mathcal{K}} \right\rangle &= \left\langle x, \sum_{k\in\mathcal{K}} c_k g_k \right\rangle = \sum_{k\in\mathcal{K}} c_k^* \langle x, g_k \rangle .\end{aligned}$$

We therefore showed that the adjoint operator of $\mathbb{T}$ is $\mathbb{T}^\times$, i.e.,

$$\mathbb{T}^\times = \mathbb{T}^*.$$

In what follows, we shall always write $\mathbb{T}^*$ instead of $\mathbb{T}^\times$. As pointed out above the concept of the adjoint of an operator generalizes the concept of the Hermitian transpose of a matrix to the infinite-dimensional case. Thus, $\mathbb{T}^*$ is the generalization of $\mathbf{T}^{\mathsf{H}}$ to the infinite-dimensional setting.

1.3.1 The Frame Operator

Let us return to the discussion we had immediately before Definition 1.12. We saw that in the finite-dimensional case, the canonical dual set $\{\tilde{\mathbf{g}}_1, \dots, \tilde{\mathbf{g}}_N\}$ to the set $\{\mathbf{g}_1, \dots, \mathbf{g}_N\}$ can be computed as follows: $\tilde{\mathbf{g}}_k = (\mathbf{T}^{\mathsf{H}}\mathbf{T})^{-1}\mathbf{g}_k$, $k = 1, \dots, N$. We know that $\mathbb{T}$ is the generalization of $\mathbf{T}$ to the infinite-dimensional case and we have just seen that $\mathbb{T}^*$ is the generalization of $\mathbf{T}^{\mathsf{H}}$. It is now obvious that the operator $\mathbb{T}^*\mathbb{T}$ must correspond to $\mathbf{T}^{\mathsf{H}}\mathbf{T}$. The operator $\mathbb{T}^*\mathbb{T}$ is of central importance in frame theory.

Definition 1.14. Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$. The operator $\mathbb{S} : \mathcal{H} \to \mathcal{H}$ defined as

$$\begin{aligned}\mathbb{S} &= \mathbb{T}^*\mathbb{T}, \qquad (1.38) \\ \mathbb{S}x &= \sum_{k\in\mathcal{K}} \langle x, g_k \rangle \, g_k\end{aligned}$$

is called the frame operator.

We note that

$$\sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 = \|\mathbb{T}x\|^2 = \langle \mathbb{T}x, \mathbb{T}x\rangle = \langle \mathbb{T}^*\mathbb{T}x, x\rangle = \langle \mathbb{S}x, x\rangle\,. \tag{1.39}$$

We are now able to formulate the frame condition in terms of the frame operator $\mathbb{S}$ by simply noting that (1.36) can be written as

$$A\|x\|^2 \le \langle \mathbb{S}x, x\rangle \le B\|x\|^2. \tag{1.40}$$

We shall next discuss the properties of $\mathbb{S}$.

Theorem 1.15. *The frame operator $\mathbb{S}$ satisfies the following properties:*

1. *$\mathbb{S}$ is linear and bounded;*

2. *$\mathbb{S}$ is self-adjoint, i.e., $\mathbb{S}^* = \mathbb{S}$;*

3. *$\mathbb{S}$ is positive definite, i.e., $\langle \mathbb{S}x, x\rangle > 0$ for all $x \in \mathcal{H}, x \neq 0$;*

4. *$\mathbb{S}$ has a unique self-adjoint positive definite square root (denoted as $\mathbb{S}^{1/2}$).*

Proof. 1. Linearity and boundedness of $\mathbb{S}$ follow from the fact that $\mathbb{S}$ is obtained by cascading a bounded linear operator and its adjoint (see (1.38)).

2. To see that $\mathbb{S}$ is self-adjoint simply note that

$$\mathbb{S}^* = (\mathbb{T}^*\mathbb{T})^* = \mathbb{T}^*\mathbb{T} = \mathbb{S}.$$

3. To see that $\mathbb{S}$ is positive definite note that, with (1.40)

$$\langle \mathbb{S}x, x\rangle \ge A\|x\|^2 > 0$$

for all $x \in \mathcal{H},\ x \neq 0$.

4. Recall the following basic fact from functional analysis [1, Th. 9.4-2].

Lemma 1.16. *Every self-adjoint positive definite bounded operator $\mathbb{A} : \mathcal{H} \to \mathcal{H}$ has a unique self-adjoint positive definite square root, i.e., there exists a unique self-adjoint positive-definite operator $\mathbb{B}$ such that $\mathbb{A} = \mathbb{B}\mathbb{B}$. The operator $\mathbb{B}$ commutes with the operator $\mathbb{A}$, i.e., $\mathbb{B}\mathbb{A} = \mathbb{A}\mathbb{B}$.*

Property 4 now follows directly form Property 2, Property 3, and Lemma 1.16.

□

We next show that the tightest possible frame bounds A and B are given by the smallest and the largest spectral value [1, Def. 7.2-1] of the frame operator $\mathbb{S}$, respectively.

Theorem 1.17. *Let A and B be the tightest possible frame bounds for a frame with frame operator $\mathbb{S}$. Then*

$$A = \lambda_{\min} \quad \textit{and} \quad B = \lambda_{\max}, \tag{1.41}$$

where $\lambda_{\min}$ and $\lambda_{\max}$ denote the smallest and the largest spectral value of $\mathbb{S}$, respectively.

Proof. By standard results on the spectrum of self-adjoint operators [1, Th. 9.2-1, Th. 9.2-3, Th. 9.2-4], we have

$$\lambda_{\min} = \inf_{x \in \mathcal{H}} \frac{\langle \mathbb{S}x, x\rangle}{\|x\|^2} \quad \text{and} \quad \lambda_{\max} = \sup_{x \in \mathcal{H}} \frac{\langle \mathbb{S}x, x\rangle}{\|x\|^2}. \tag{1.42}$$

This means that $\lambda_{\min}$ and $\lambda_{\max}$ are, respectively, the largest and the smallest constants such that

$$\lambda_{\min}\|x\|^2 \leq \langle \mathbb{S}x, x\rangle \leq \lambda_{\max}\|x\|^2 \tag{1.43}$$

is satisfied for every $x \in \mathcal{H}$. According to (1.40) this implies that $\lambda_{\min}$ and $\lambda_{\max}$ are the tightest possible frame bounds. □

It is instructive to compare (1.43) to (1.31). Remember that $\mathbb{S} = \mathbb{T}^*\mathbb{T}$ corresponds to the matrix $\mathbf{T}^{\mathsf{H}}\mathbf{T}$ in the finite-dimensional case considered in Section 1.2.3. Thus, $\|\mathbf{c}\|^2 = \mathbf{x}^{\mathsf{H}}\mathbf{T}^{\mathsf{H}}\mathbf{T}\mathbf{x} = \langle \mathbf{S}\mathbf{x}, \mathbf{x}\rangle$, which upon insertion into (1.31), shows that (1.43) is simply a generalization of (1.31) to the infinite-dimensional case.

1.3.2 The Canonical Dual Frame

Recall that in the finite-dimensional case considered in Section 1.2.3, the canonical dual frame $\{\tilde{\mathbf{g}}_k\}_{k=1}^N$ of the frame $\{\mathbf{g}_k\}_{k=1}^N$ can be used to reconstruct the signal $\mathbf{x}$ from the expansion coefficients $\{\langle \mathbf{x}, \mathbf{g}_k\rangle\}_{k=1}^N$ according to

$$\mathbf{x} = \sum_{k=1}^{N} \langle \mathbf{x}, \mathbf{g}_k\rangle\, \tilde{\mathbf{g}}_k.$$

In (1.30) we saw that the canonical dual frame can be computed as follows:

$$\tilde{\mathbf{g}}_k = (\mathbf{T}^{\mathsf{H}}\mathbf{T})^{-1}\mathbf{g}_k, \quad k = 1, \ldots, N. \tag{1.44}$$

We already pointed out that the frame operator $\mathbb{S} = \mathbb{T}^*\mathbb{T}$ is represented by the matrix $\mathbf{T}^{\mathsf{H}}\mathbf{T}$ in the finite-dimensional case. The matrix $(\mathbf{T}^{\mathsf{H}}\mathbf{T})^{-1}$ therefore corresponds to the operator $\mathbb{S}^{-1}$, which will be studied next.

From (1.41) it follows that $\lambda_{\min}$, the smallest spectral value of $\mathbb{S}$, satisfies $\lambda_{\min} > 0$ if $\{g_k\}_{k\in\mathcal{K}}$ is a frame. This implies that zero is a regular value [1, Def. 7.2-1] of $\mathbb{S}$ and hence $\mathbb{S}$ is invertible on $\mathcal{H}$, i.e., there exists a unique operator $\mathbb{S}^{-1}$ such that $\mathbb{S}\mathbb{S}^{-1} = \mathbb{S}^{-1}\mathbb{S} = \mathbb{I}_{\mathcal{H}}$. Next, we summarize the properties of $\mathbb{S}^{-1}$.

Theorem 1.18. *The following properties hold:*

1. *$\mathbb{S}^{-1}$ is self-adjoint, i.e., $(\mathbb{S}^{-1})^* = \mathbb{S}^{-1}$;*

2. *$\mathbb{S}^{-1}$ satisfies*
$$\frac{1}{B} = \inf_{x\in\mathcal{H}} \frac{\langle \mathbb{S}^{-1}x, x\rangle}{\|x\|^2} \quad \text{and} \quad \frac{1}{A} = \sup_{x\in\mathcal{H}} \frac{\langle \mathbb{S}^{-1}x, x\rangle}{\|x\|^2}, \tag{1.45}$$
where A and B are the tightest possible frame bounds of $\mathbb{S}$;

3. *$\mathbb{S}^{-1}$ is positive definite.*

Proof. 1. To prove that $\mathbb{S}^{-1}$ is self-adjoint we write
$$(\mathbb{S}\mathbb{S}^{-1})^* = (\mathbb{S}^{-1})^*\mathbb{S}^* = \mathbb{I}_{\mathcal{H}}.$$
Since $\mathbb{S}$ is self-adjoint, i.e., $\mathbb{S} = \mathbb{S}^*$, we conclude that
$$(\mathbb{S}^{-1})^*\mathbb{S} = \mathbb{I}_{\mathcal{H}}.$$
Multiplying by $\mathbb{S}^{-1}$ from the right, we finally obtain
$$(\mathbb{S}^{-1})^* = \mathbb{S}^{-1}.$$

2. To prove the first equation in (1.45) we write
$$B = \sup_{x\in\mathcal{H}} \frac{\langle \mathbb{S}x, x\rangle}{\|x\|^2} = \sup_{y\in\mathcal{H}} \frac{\langle \mathbb{S}\mathbb{S}^{1/2}\mathbb{S}^{-1}y, \mathbb{S}^{1/2}\mathbb{S}^{-1}y\rangle}{\langle \mathbb{S}^{1/2}\mathbb{S}^{-1}y, \mathbb{S}^{1/2}\mathbb{S}^{-1}y\rangle} = \sup_{y\in\mathcal{H}} \frac{\langle \mathbb{S}^{-1}\mathbb{S}^{1/2}\mathbb{S}\mathbb{S}^{1/2}\mathbb{S}^{-1}y, y\rangle}{\langle \mathbb{S}^{-1}\mathbb{S}^{1/2}\mathbb{S}^{1/2}\mathbb{S}^{-1}y, y\rangle} = \sup_{y\in\mathcal{H}} \frac{\langle y, y\rangle}{\langle \mathbb{S}^{-1}y, y\rangle} \tag{1.46}$$
where the first equality follows from (1.41) and (1.42); in the second equality we used the fact that the operator $\mathbb{S}^{1/2}\mathbb{S}^{-1}$ is one-to-one on $\mathcal{H}$ and changed variables according to $x = \mathbb{S}^{1/2}\mathbb{S}^{-1}y$; in the third equality we used the fact that $\mathbb{S}^{1/2}$ and $\mathbb{S}^{-1}$ are self-adjoint, and in the fourth equality we used $\mathbb{S} = \mathbb{S}^{1/2}\mathbb{S}^{1/2}$. The first equation in (1.45) is now obtained by noting that (1.46) implies
$$\frac{1}{B} = 1 \bigg/ \left(\sup_{y\in\mathcal{H}} \frac{\langle y, y\rangle}{\langle \mathbb{S}^{-1}y, y\rangle}\right) = \inf_{y\in\mathcal{H}} \frac{\langle \mathbb{S}^{-1}y, y\rangle}{\langle y, y\rangle}.$$
The second equation in (1.45) is proved analogously.

3. Positive-definiteness of $\mathbb{S}^{-1}$ follows from the first equation in (1.45) and the fact that $B < \infty$ so that $1/B > 0$.

□

We are now ready to generalize (1.44) and state the main result on canonical dual frames in the case of general (possibly infinite-dimensional) Hilbert spaces.

Theorem 1.19. *Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$ with the frame bounds A and B, and let $\mathbb{S}$ be the corresponding frame operator. Then, the set $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ given by*

$$\tilde{g}_k = \mathbb{S}^{-1} g_k, \quad k \in \mathcal{K}, \tag{1.47}$$

is a frame for $\mathcal{H}$ with the frame bounds $\tilde{A} = 1/B$ and $\tilde{B} = 1/A$.

The analysis operator associated with $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ defined as

$$\begin{aligned}&\tilde{\mathbb{T}} : \mathcal{H} \to l^2\\ &\tilde{\mathbb{T}} : x \to \{\langle x, \tilde{g}_k\rangle\}_{k\in\mathcal{K}}\end{aligned}$$

satisfies

$$\tilde{\mathbb{T}} = \mathbb{T}\mathbb{S}^{-1} = \mathbb{T}\left(\mathbb{T}^*\mathbb{T}\right)^{-1}. \tag{1.48}$$

Proof. Recall that $\mathbb{S}^{-1}$ is self-adjoint. Hence, we have $\langle x, \tilde{g}_k\rangle = \langle x, \mathbb{S}^{-1} g_k\rangle = \langle \mathbb{S}^{-1}x, g_k\rangle$ for all $x \in \mathcal{H}$. Thus, using (1.39), we obtain

$$\begin{aligned}\sum_{k\in\mathcal{K}} |\langle x, \tilde{g}_k\rangle|^2 &= \sum_{k\in\mathcal{K}} \left|\left\langle \mathbb{S}^{-1}x, g_k\right\rangle\right|^2\\ &= \left\langle \mathbb{S}(\mathbb{S}^{-1}x), \mathbb{S}^{-1}x\right\rangle = \left\langle x, \mathbb{S}^{-1}x\right\rangle = \left\langle \mathbb{S}^{-1}x, x\right\rangle.\end{aligned}$$

Therefore, we conclude from (1.45) that

$$\frac{1}{B}\|x\|^2 \le \sum_{k\in\mathcal{K}} |\langle x, \tilde{g}_k\rangle|^2 \le \frac{1}{A}\|x\|^2,$$

i.e., the set $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ constitutes a frame for $\mathcal{H}$ with frame bounds $\tilde{A} = 1/B$ and $\tilde{B} = 1/A$; moreover, it follows from (1.45) that $\tilde{A} = 1/B$ and $\tilde{B} = 1/A$ are the tightest possible frame bounds. It remains to show that $\tilde{\mathbb{T}} = \mathbb{T}\mathbb{S}^{-1}$:

$$\tilde{\mathbb{T}}x = \{\langle x, \tilde{g}_k\rangle\}_{k\in\mathcal{K}} = \left\{\left\langle x, \mathbb{S}^{-1} g_k\right\rangle\right\}_{k\in\mathcal{K}} = \left\{\left\langle \mathbb{S}^{-1}x, g_k\right\rangle\right\}_{k\in\mathcal{K}} = \mathbb{T}\mathbb{S}^{-1}x.$$

□

We call $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ the *canonical dual frame* associated to the frame $\{g_k\}_{k\in\mathcal{K}}$. It is convenient to introduce the *canonical dual frame operator*:

Definition 1.20. The frame operator associated to the canonical dual frame,

$$\tilde{\mathbb{S}} = \tilde{\mathbb{T}}^*\tilde{\mathbb{T}}, \quad \tilde{\mathbb{S}}x = \sum_{k\in\mathcal{K}} \langle x, \tilde{g}_k\rangle\, \tilde{g}_k \tag{1.49}$$

is called the canonical dual frame operator.

Theorem 1.21. *The canonical dual frame operator $\tilde{\mathbb{S}}$ satisfies $\tilde{\mathbb{S}} = \mathbb{S}^{-1}$.*

Proof. For every $x \in \mathcal{H}$, we have

$$\begin{aligned}\tilde{\mathbb{S}}x &= \sum_{k\in\mathcal{K}} \langle x, \tilde{g}_k\rangle \tilde{g}_k = \sum_{k\in\mathcal{K}} \left\langle x, \mathbb{S}^{-1}g_k\right\rangle \mathbb{S}^{-1}g_k \\ &= \mathbb{S}^{-1}\sum_{k\in\mathcal{K}} \left\langle \mathbb{S}^{-1}x, g_k\right\rangle g_k = \mathbb{S}^{-1}\mathbb{S}\mathbb{S}^{-1}x = \mathbb{S}^{-1}x,\end{aligned}$$

where in the first equality we used (1.49), in the second we used (1.47), in the third we made use of the fact that $\mathbb{S}^{-1}$ is self-adjoint, and in the fourth we used the definition of $\mathbb{S}$. □

Note that canonical duality is a reciprocity relation. If the frame $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ is the canonical dual of the frame $\{g_k\}_{k\in\mathcal{K}}$, then $\{g_k\}_{k\in\mathcal{K}}$ is the canonical dual of the frame $\{\tilde{g}_k\}_{k\in\mathcal{K}}$. This can be seen by noting that

$$\tilde{\mathbb{S}}^{-1}\tilde{g}_k = (\mathbb{S}^{-1})^{-1}\mathbb{S}^{-1}g_k = \mathbb{S}\mathbb{S}^{-1}g_k = g_k.$$

1.3.3 Signal Expansions

The following theorem can be considered as one of the *central results in frame theory.* It states that every signal $x \in \mathcal{H}$ can be expanded into a frame. The expansion coefficients can be chosen as the inner products of x with the canonical dual frame elements.

Theorem 1.22. *Let $\{g_k\}_{k\in\mathcal{K}}$ and $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ be canonical dual frames for the Hilbert space $\mathcal{H}$. Every signal $x \in \mathcal{H}$ can be decomposed as follows*

$$\begin{aligned}x &= \mathbb{T}^*\tilde{\mathbb{T}}x = \sum_{k\in\mathcal{K}} \langle x, \tilde{g}_k\rangle g_k \\ x &= \tilde{\mathbb{T}}^*\mathbb{T}x = \sum_{k\in\mathcal{K}} \langle x, g_k\rangle \tilde{g}_k. \end{aligned} \tag{1.50}$$

Note that, equivalently, we have

$$\mathbb{T}^*\tilde{\mathbb{T}} = \tilde{\mathbb{T}}^*\mathbb{T} = \mathbb{I}_{\mathcal{H}}.$$

Proof. We have

$$\begin{aligned}\mathbb{T}^*\tilde{\mathbb{T}}x &= \sum_{k\in\mathcal{K}} \langle x, \tilde{g}_k\rangle g_k = \sum_{k\in\mathcal{K}} \left\langle x, \mathbb{S}^{-1}g_k\right\rangle g_k \\ &= \sum_{k\in\mathcal{K}} \left\langle \mathbb{S}^{-1}x, g_k\right\rangle g_k = \mathbb{S}\mathbb{S}^{-1}x = x.\end{aligned}$$

This proves that $\mathbb{T}^*\tilde{\mathbb{T}} = \mathbb{I}_{\mathcal{H}}$. The proof of $\tilde{\mathbb{T}}^*\mathbb{T} = \mathbb{I}_{\mathcal{H}}$ is similar. □

Note that (1.50) corresponds to the decomposition (1.26) we found in the finite-dimensional case.

It is now natural to ask whether reconstruction of x from the coefficients $\langle x, g_k\rangle$, $k \in \mathcal{K}$, according to (1.50) is the only way of recovering x from $\langle x, g_k\rangle$, $k \in \mathcal{K}$. Recall that we showed in the finite-dimensional case (see Section 1.2.3) that for each complete and redundant set of vectors $\{\mathbf{g}_1, \ldots, \mathbf{g}_N\}$, there are infinitely many dual sets $\{\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_N\}$ that can be used to reconstruct a signal $\mathbf{x}$ from the coefficients $\langle \mathbf{x}, \mathbf{g}_k \rangle$, $k = 1, \ldots, N$, according to (1.26). These dual sets are obtained by identifying $\{\tilde{\mathbf{g}}_1, \ldots, \tilde{\mathbf{g}}_N\}$ with the columns of $\mathbf{L}$, where $\mathbf{L}$ is a left-inverse of the analysis matrix $\mathbf{T}$. In the infinite-dimensional case the question of finding all dual frames for a given frame boils down to finding, for a given analysis operator $\mathbb{T}$, all linear operators $\mathbb{L}$ that satisfy

$$\mathbb{L}\mathbb{T}x = x$$

for all $x \in \mathcal{H}$. In other words, we want to identify all left-inverses $\mathbb{L}$ of the analysis operator $\mathbb{T}$. The answer to this question is the infinite-dimensional version of Theorem 1.6 that we state here without proof.

Theorem 1.23. *Let $\mathbb{A} : \mathcal{H} \to l^2$ be a bounded linear operator. Assume that $\mathbb{A}^*\mathbb{A} : \mathcal{H} \to \mathcal{H}$ is invertible on $\mathcal{H}$. Then, the operator $\mathbb{A}^\dagger : l^2 \to \mathcal{H}$ defined as $\mathbb{A}^\dagger \triangleq (\mathbb{A}^*\mathbb{A})^{-1}\mathbb{A}^*$ is a left-inverse of $\mathbb{A}$, i.e., $\mathbb{A}^\dagger\mathbb{A} = \mathbb{I}_{\mathcal{H}}$, where $\mathbb{I}_{\mathcal{H}}$ is the identity operator on $\mathcal{H}$. Moreover, the general solution $\mathbb{L}$ of the equation $\mathbb{L}\mathbb{A} = \mathbb{I}_{\mathcal{H}}$ is given by*

$$\mathbb{L} = \mathbb{A}^\dagger + \mathbb{M}(\mathbb{I}_{l^2} - \mathbb{A}\mathbb{A}^\dagger)$$

where $\mathbb{M} : l^2 \to \mathcal{H}$ is an arbitrary bounded linear operator and $\mathbb{I}_{l^2}$ is the identity operator on l^2.

Applying this theorem to the operator $\mathbb{T}$ we see that all left-inverses of $\mathbb{T}$ can be written as

$$\mathbb{L} = \mathbb{T}^\dagger + \mathbb{M}(\mathbb{I}_{l^2} - \mathbb{T}\mathbb{T}^\dagger) \tag{1.51}$$

where $\mathbb{M} : l^2 \to \mathcal{H}$ is an arbitrary bounded linear operator and

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

Now, using (1.48), we obtain the following important identity:

$$\mathbb{T}^\dagger = (\mathbb{T}^*\mathbb{T})^{-1}\mathbb{T}^* = \mathbb{S}^{-1}\mathbb{T}^* = \tilde{\mathbb{T}}^*.$$

This shows that reconstruction according to (1.50), i.e., by applying the operator $\tilde{\mathbb{T}}^*$ to the coefficient sequence $\mathbb{T}x = \{\langle x, g_k \rangle\}_{k \in \mathcal{K}}$ is nothing but applying the infinite-dimensional analog of the Moore-Penrose inverse $\mathbf{T}^\dagger = (\mathbf{T}^{\mathsf{H}}\mathbf{T})^{-1}\mathbf{T}^{\mathsf{H}}$. As already noted in the finite-dimensional case the existence of infinitely many left-inverses of the operator $\mathbb{T}$ provides us with freedom in designing dual frames.

We close this discussion with a geometric interpretation of the parametrization (1.51). First observe the following.

Theorem 1.24. *The operator*

$$\mathbb{P} : l^2 \to \mathcal{R}(\mathbb{T}) \subseteq l^2$$

defined as

$$\mathbb{P} = \mathbb{T}\mathbb{S}^{-1}\mathbb{T}^*$$

satisfies the following properties:

1. *$\mathbb{P}$ is the identity operator $\mathbb{I}_{l^2}$ on $\mathcal{R}(\mathbb{T})$.*

2. *$\mathbb{P}$ is the zero operator on $\mathcal{R}(\mathbb{T})^{\perp}$, where $\mathcal{R}(\mathbb{T})^{\perp}$ denotes the orthogonal complement of the space $\mathcal{R}(\mathbb{T})$.*

In other words, $\mathbb{P}$ is the orthogonal projection operator onto $\mathcal{R}(\mathbb{T}) = \{\{c_k\}_{k\in\mathcal{K}} \,|\, \{c_k\}_{k\in\mathcal{K}} = \mathbb{T}x, x \in \mathcal{H}\}$, the range space of the operator $\mathbb{T}$.

Proof. 1. Take a sequence $\{c_k\}_{k\in\mathcal{K}} \in \mathcal{R}(\mathbb{T})$ and note that it can be written as $\{c_k\}_{k\in\mathcal{K}} = \mathbb{T}x$, where $x \in \mathcal{H}$. Then, we have

$$\mathbb{P}\{c_k\}_{k\in\mathcal{K}} = \mathbb{T}\mathbb{S}^{-1}\mathbb{T}^*\mathbb{T}x = \mathbb{T}\mathbb{S}^{-1}\mathbb{S}x = \mathbb{T}\mathbb{I}_{\mathcal{H}}x = \mathbb{T}x = \{c_k\}_{k\in\mathcal{K}}.$$

This proves that $\mathbb{P}$ is the identity operator on $\mathcal{R}(\mathbb{T})$.

2. Next, take a sequence $\{c_k\}_{k\in\mathcal{K}} \in \mathcal{R}(\mathbb{T})^{\perp}$. As the orthogonal complement of the range space of an operator is the null space of its adjoint, we have $\mathbb{T}^*\{c_k\}_{k\in\mathcal{K}} = 0$ and therefore

$$\mathbb{P}\{c_k\}_{k\in\mathcal{K}} = \mathbb{T}\mathbb{S}^{-1}\mathbb{T}^*\{c_k\}_{k\in\mathcal{K}} = 0.$$

This proves that $\mathbb{P}$ is the zero operator on $\mathcal{R}(\mathbb{T})^{\perp}$.

□

Now using that $\mathbb{T}\mathbb{T}^{\dagger} = \mathbb{T}\mathbb{S}^{-1}\mathbb{T}^* = \mathbb{P}$ and $\mathbb{T}^{\dagger} = \mathbb{S}^{-1}\mathbb{T}^* = \mathbb{S}^{-1}\mathbb{S}\mathbb{S}^{-1}\mathbb{T}^* = \mathbb{S}^{-1}\mathbb{T}^*\mathbb{T}\mathbb{S}^{-1}\mathbb{T}^* = \tilde{\mathbb{T}}^*\mathbb{P}$, we can rewrite (1.51) as follows

$$\mathbb{L} = \tilde{\mathbb{T}}^*\mathbb{P} + \mathbb{M}(\mathbb{I}_{l^2} - \mathbb{P}). \tag{1.52}$$

Next, we show that $(\mathbb{I}_{l^2} - \mathbb{P}) : l^2 \to l^2$ is the orthogonal projection onto $\mathcal{R}(\mathbb{T})^{\perp}$. Indeed, we can directly verify the following: For every $\{c_k\}_{k\in\mathcal{K}} \in \mathcal{R}(\mathbb{T})^{\perp}$, we have $(\mathbb{I}_{l^2} - \mathbb{P})\{c_k\}_{k\in\mathcal{K}} = \mathbb{I}_{l^2}\{c_k\}_{k\in\mathcal{K}} - 0 = \{c_k\}_{k\in\mathcal{K}}$, i.e., $\mathbb{I}_{l^2} - \mathbb{P}$ is the identity operator on $\mathcal{R}(\mathbb{T})^{\perp}$; for every $\{c_k\}_{k\in\mathcal{K}} \in (\mathcal{R}(\mathbb{T})^{\perp})^{\perp} = \mathcal{R}(\mathbb{T})$, we have $(\mathbb{I}_{l^2} - \mathbb{P})\{c_k\}_{k\in\mathcal{K}} = \mathbb{I}_{l^2}\{c_k\}_{k\in\mathcal{K}} - \{c_k\}_{k\in\mathcal{K}} = 0$, i.e., $\mathbb{I}_{l^2} - \mathbb{P}$ is the zero operator on $(\mathcal{R}(\mathbb{T})^{\perp})^{\perp}$.

We are now ready to re-interpret (1.52) as follows. Every left-inverse $\mathbb{L}$ of $\mathbb{T}$ acts as $\tilde{\mathbb{T}}^*$ (the synthesis operator of the canonical dual frame) on the range space of the analysis operator $\mathbb{T}$, and can act in an arbitrary linear and bounded fashion on the orthogonal complement of the range space of the analysis operator $\mathbb{T}$.

1.3.4 Tight Frames

The frames considered in Examples 1.9 and 1.10 above have an interesting property: In both cases the tightest possible frame bounds A and B are equal. Frames with this property are called tight frames.

Definition 1.25. A frame $\{g_k\}_{k\in\mathcal{K}}$ with tightest possible frame bounds $A = B$ is called a tight frame.

Tight frames are of significant practical interest because of the following central fact.

Theorem 1.26. *Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$. The frame $\{g_k\}_{k\in\mathcal{K}}$ is tight with frame bound A if and only if its corresponding frame operator satisfies $\mathbb{S} = A\mathbb{I}_{\mathcal{H}}$, or equivalently, if*

$$x = \frac{1}{A}\sum_{k\in\mathcal{K}} \langle x, g_k\rangle\, g_k \tag{1.53}$$

for all $x \in \mathcal{H}$.

Proof. First observe that $\mathbb{S} = A\mathbb{I}_{\mathcal{H}}$ is equivalent to $\mathbb{S}x = A\mathbb{I}_{\mathcal{H}}x = Ax$ for all $x \in \mathcal{H}$, which, in turn, is equivalent to (1.53) by definition of the frame operator.

To prove that tightness of $\{g_k\}_{k\in\mathcal{K}}$ implies $\mathbb{S} = A\mathbb{I}_{\mathcal{H}}$, note that by Definition 1.25, using (1.40) we can write

$$\langle \mathbb{S}x, x\rangle = A\,\langle x, x\rangle\,, \text{ for all } x \in \mathcal{H}.$$

Therefore

$$\langle (\mathbb{S} - A\mathbb{I}_{\mathcal{H}})x, x\rangle = 0, \text{ for all } x \in \mathcal{H},$$

which implies $\mathbb{S} = A\mathbb{I}_{\mathcal{H}}$.

To prove that $\mathbb{S} = A\mathbb{I}_{\mathcal{H}}$ implies tightness of $\{g_k\}_{k\in\mathcal{K}}$, we take the inner product with x on both sides of (1.53) to obtain

$$\langle x, x\rangle = \frac{1}{A}\sum_{k\in\mathcal{K}} \langle x, g_k\rangle\,\langle g_k, x\rangle\,.$$

This is equivalent to

$$A\|x\|^2 = \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2\,,$$

which shows that $\{g_k\}_{k\in\mathcal{K}}$ is a tight frame for $\mathcal{H}$ with frame bound equal to A. □

The practical importance of tight frames lies in the fact that they make the computation of the canonical dual frame, which in the general case requires inversion of an operator and application of this inverse to all frame elements, particularly simple. Specifically, we have:

$$\tilde{g}_k = \mathbb{S}^{-1}g_k = \frac{1}{A}\mathbb{I}_{\mathcal{H}}g_k = \frac{1}{A}g_k.$$

A well-known example of a tight frame for $\mathbb{R}^2$ is the following:

Example 1.27 (The Mercedes-Benz frame [20])**.** The Mercedes-Benz frame (see Figure 1.4) is given by the following three vectors in $\mathbb{R}^2$:

$$\mathbf{g}_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad \mathbf{g}_2 = \begin{bmatrix} -\sqrt{3}/2 \\ -1/2 \end{bmatrix}, \quad \mathbf{g}_3 = \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix}. \tag{1.54}$$

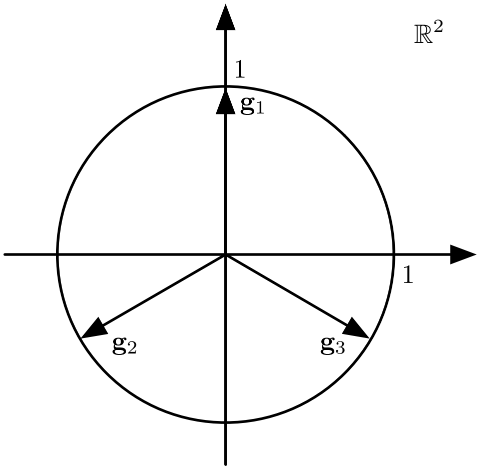


Figure 1.4: The Mercedes-Benz frame.

To see that this frame is indeed tight, note that its analysis operator $\mathbb{T}$ is given by the matrix

$$\mathbf{T} = \begin{bmatrix} 0 & 1 \\ -\sqrt{3}/2 & -1/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix}.$$

The adjoint $\mathbb{T}^*$ of the analysis operator is given by the matrix

$$\mathbf{T}^{\mathsf{H}} = \begin{bmatrix} 0 & -\sqrt{3}/2 & \sqrt{3}/2 \\ 1 & -1/2 & -1/2 \end{bmatrix}.$$

Therefore, the frame operator $\mathbb{S}$ is represented by the matrix

$$\mathbf{S} = \mathbf{T}^{\mathsf{H}}\mathbf{T} = \begin{bmatrix} 0 & -\sqrt{3}/2 & \sqrt{3}/2 \\ 1 & -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -\sqrt{3}/2 & -1/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix} = \frac{3}{2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = \frac{3}{2}\mathbf{I}_2,$$

and hence $\mathbb{S} = A\mathbb{I}_{\mathbb{R}^2}$ with $A = 3/2$, which implies, by Theorem 1.26, that $\{\mathbf{g}_1, \mathbf{g}_2, \mathbf{g}_3\}$ is a tight frame (for $\mathbb{R}^2$).

The design of tight frames is challenging in general. It is hence interesting to devise simple systematic methods for obtaining tight frames. The following theorem shows how we can obtain a tight frame from a given general frame.

Theorem 1.28. *Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$ with frame operator $\mathbb{S}$. Denote the positive definite square root of $\mathbb{S}^{-1}$ by $\mathbb{S}^{-1/2}$. Then $\{\mathbb{S}^{-1/2}g_k\}_{k\in\mathcal{K}}$ is a tight frame for $\mathcal{H}$ with frame bound $A = 1$, i.e.,*

$$x = \sum_{k\in\mathcal{K}} \left\langle x, \mathbb{S}^{-1/2}g_k \right\rangle \mathbb{S}^{-1/2}g_k, \quad \textit{for all } x \in \mathcal{H}.$$

Proof. Since $\mathbb{S}^{-1}$ is self-adjoint and positive definite by Theorem 1.18, it has, by Lemma 1.16, a unique self-adjoint positive definite square root $\mathbb{S}^{-1/2}$ that commutes with $\mathbb{S}^{-1}$. Moreover $\mathbb{S}^{-1/2}$ also commutes with $\mathbb{S}$, which can be seen as follows:

$$\begin{aligned}\mathbb{S}^{-1/2}\mathbb{S}^{-1} &= \mathbb{S}^{-1}\mathbb{S}^{-1/2}\\ \mathbb{S}\mathbb{S}^{-1/2}\mathbb{S}^{-1} &= \mathbb{S}^{-1/2}\\ \mathbb{S}\mathbb{S}^{-1/2} &= \mathbb{S}^{-1/2}\mathbb{S}.\end{aligned}$$

The proof is then effected by noting the following:

$$\begin{aligned}x = \mathbb{S}^{-1}\mathbb{S}x &= \mathbb{S}^{-1/2}\mathbb{S}^{-1/2}\mathbb{S}x\\ &= \mathbb{S}^{-1/2}\mathbb{S}\mathbb{S}^{-1/2}x\\ &= \sum_{k\in\mathcal{K}} \left\langle \mathbb{S}^{-1/2}x, g_k\right\rangle \mathbb{S}^{-1/2}g_k\\ &= \sum_{k\in\mathcal{K}} \left\langle x, \mathbb{S}^{-1/2}g_k\right\rangle \mathbb{S}^{-1/2}g_k.\end{aligned}$$

□

It is evident that every ONB is a tight frame with $A = 1$. Note, however, that conversely a tight frame (even with $A = 1$) need not be an orthonormal or orthogonal basis, as can be seen from Example 1.10. However, as the next theorem shows, a tight frame with $A = 1$ and $\|g_k\| = 1$, for all $k \in \mathcal{K}$, is necessarily an ONB.

Theorem 1.29. *A tight frame $\{g_k\}_{k\in\mathcal{K}}$ for the Hilbert space $\mathcal{H}$ with $A = 1$ and $\|g_k\| = 1$, for all $k \in \mathcal{K}$, is an ONB for $\mathcal{H}$.*

Proof. Combining

$$\langle \mathbb{S}g_k, g_k\rangle = A\|g_k\|^2 = \|g_k\|^2$$

with

$$\langle \mathbb{S}g_k, g_k\rangle = \sum_{j\in\mathcal{K}} |\langle g_k, g_j\rangle|^2 = \|g_k\|^4 + \sum_{j\neq k} |\langle g_k, g_j\rangle|^2$$

we obtain

$$\|g_k\|^4 + \sum_{j\neq k} |\langle g_k, g_j\rangle|^2 = \|g_k\|^2.$$

Since $\|g_k\|^2 = 1$, for all $k \in \mathcal{K}$, it follows that $\sum_{j\neq k} |\langle g_k, g_j\rangle|^2 = 0$, for all $k \in \mathcal{K}$. This implies that the elements of $\{g_j\}_{j\in\mathcal{K}}$ are necessarily orthogonal to each other. □

There is an elegant result that tells us that every tight frame with frame bound $A = 1$ can be realized as an orthogonal projection of an ONB from a space with larger dimension. This result is known as Naimark's theorem. Here we state the finite-dimensional version of this theorem, for the infinite-dimensional version see [24].

Theorem 1.30 (Naimark, [24, Prop. 1.1]). *Let $N > M$. Suppose that the set $\{\mathbf{g}_1, \ldots, \mathbf{g}_N\}$, $\mathbf{g}_k \in \mathcal{H}$, $k = 1, \ldots, N$, is a tight frame for an M-dimensional Hilbert space $\mathcal{H}$ with frame bound $A = 1$. Then, there exists an N-dimensional Hilbert space $\mathcal{K} \supset \mathcal{H}$ and an ONB $\{\mathbf{e}_1, \ldots, \mathbf{e}_N\}$ for $\mathcal{K}$ such that $\mathbb{P}\mathbf{e}_k = \mathbf{g}_k$, $k = 1, \ldots, N$, where $\mathbb{P} : \mathcal{K} \to \mathcal{K}$ is the orthogonal projection onto $\mathcal{H}$.*

We omit the proof and illustrate the theorem by an example instead.

Example 1.31. Consider the Hilbert space $\mathcal{K} = \mathbb{R}^3$, and assume that $\mathcal{H} \subset \mathcal{K}$ is the plane spanned by the vectors $[1\ 0\ 0]^\mathsf{T}$ and $[0\ 1\ 0]^\mathsf{T}$, i.e.,

$$\mathcal{H} = \operatorname{span}\left\{[1\ 0\ 0]^\mathsf{T}, [0\ 1\ 0]^\mathsf{T}\right\}.$$

We can construct a tight frame for $\mathcal{H}$ with three elements and frame bound $A = 1$ if we rescale the Mercedes-Benz frame from Example 1.27. Specifically, consider the vectors $\mathbf{g}_k,\ k = 1, 2, 3$, defined in (1.54) and let $\mathbf{g}'_k \triangleq \sqrt{2/3}\,\mathbf{g}_k,\ k = 1, 2, 3$. In the following, we think about the two-dimensional vectors $\mathbf{g}'_k$ as being embedded into the three-dimensional space $\mathcal{K}$ with the third coordinate (in the standard basis of $\mathcal{K}$) being equal to zero. Clearly, $\{\mathbf{g}'_k\}_{k=1}^3$ is a tight frame for $\mathcal{H}$ with frame bound $A = 1$. Now consider the following three vectors in $\mathcal{K}$:

$$\mathbf{e}_1 = \begin{bmatrix} 0 \\ \sqrt{2/3} \\ -1/\sqrt{3} \end{bmatrix}, \quad \mathbf{e}_2 = \begin{bmatrix} -1/\sqrt{2} \\ -1/\sqrt{6} \\ -1/\sqrt{3} \end{bmatrix}, \quad \mathbf{e}_3 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{6} \\ -1/\sqrt{3} \end{bmatrix}.$$

Direct calculation reveals that $\{\mathbf{e}_k\}_{k=1}^3$ is an ONB for $\mathcal{K}$. Observe that the frame vectors $\mathbf{g}'_k,\ k = 1, 2, 3$, can be obtained from the ONB vectors $\mathbf{e}_k,\ k = 1, 2, 3$, by applying the orthogonal projection from $\mathcal{K}$ onto $\mathcal{H}$:

$$\mathbf{P} \triangleq \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

according to $\mathbf{g}'_k = \mathbf{P}\mathbf{e}_k,\ k = 1, 2, 3$. This illustrates Naimark's theorem.

1.3.5 Exact Frames and Biorthonormality

In Section 1.2.2 we studied expansions of signals in $\mathbb{C}^M$ into (not necessarily orthogonal) bases. The main results we established in this context can be summarized as follows:

1. The number of vectors in a basis is always equal to the dimension of the Hilbert space under consideration. Every set of vectors that spans $\mathbb{C}^M$ and has more than M vectors is necessarily redundant, i.e., the vectors in this set are linearly dependent. Removal of an arbitrary vector from a basis for $\mathbb{C}^M$ leaves a set that no longer spans $\mathbb{C}^M$.

2. For a given basis $\{\mathbf{e}_k\}_{k=1}^M$ every signal $\mathbf{x} \in \mathbb{C}^M$ has a *unique* representation according to

$$\mathbf{x} = \sum_{k=1}^{M} \langle \mathbf{x}, \mathbf{e}_k \rangle \tilde{\mathbf{e}}_k. \tag{1.55}$$

The basis $\{\mathbf{e}_k\}_{k=1}^M$ and its dual basis $\{\tilde{\mathbf{e}}_k\}_{k=1}^M$ satisfy the biorthonormality relation (1.20).

The theory of ONBs in infinite-dimensional spaces is well-developed. In this section, we ask how the concept of general (i.e., not necessarily orthogonal) bases can be extended to infinite-dimensional spaces. Clearly, in the infinite-dimensional case, we can not simply say that the number of elements in a basis must be equal to the dimension of the Hilbert space. However, we can use the property that removing an element from a basis, leaves us with an incomplete set of vectors to motivate the following definition.

Definition 1.32. Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$. We call the frame $\{g_k\}_{k\in\mathcal{K}}$ *exact* if, for all $m \in \mathcal{K}$, the set $\{g_k\}_{k\neq m}$ is incomplete for $\mathcal{H}$; we call the frame $\{g_k\}_{k\in\mathcal{K}}$ *inexact* if there is at least one element g_m that can be removed from the frame, so that the set $\{g_k\}_{k\neq m}$ is again a frame for $\mathcal{H}$.

There are two more properties of general bases in finite-dimensional spaces that carry over to the infinite-dimensional case, namely uniqueness of representation in the sense of (1.55) and biorthonormality between the frame and its canonical dual. To show that representation of a signal in an exact frame is unique and that an exact frame is biorthonormal to its canonical dual frame, we will need the following two lemmas.

Let $\{g_k\}_{k\in\mathcal{K}}$ and $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ be canonical dual frames. The first lemma below states that for a fixed $x \in \mathcal{H}$, among all possible expansion coefficient sequences $\{c_k\}_{k\in\mathcal{K}}$ satisfying $x = \sum_{k\in\mathcal{K}} c_k g_k$, the coefficients $c_k = \langle x, \tilde{g}_k\rangle$ have minimum l^2-norm.

Lemma 1.33 ([5]). *Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$ and $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ its canonical dual frame. For a fixed $x \in \mathcal{H}$, let $c_k = \langle x, \tilde{g}_k\rangle$ so that $x = \sum_{k\in\mathcal{K}} c_k g_k$. If it is possible to find scalars $\{a_k\}_{k\in\mathcal{K}} \neq \{c_k\}_{k\in\mathcal{K}}$ such that $x = \sum_{k\in\mathcal{K}} a_k g_k$, then we must have*

$$\sum_{k\in\mathcal{K}} |a_k|^2 = \sum_{k\in\mathcal{K}} |c_k|^2 + \sum_{k\in\mathcal{K}} |c_k - a_k|^2 . \tag{1.56}$$

Proof. We have

$$c_k = \langle x, \tilde{g}_k\rangle = \left\langle x, \mathbb{S}^{-1} g_k\right\rangle = \left\langle \mathbb{S}^{-1} x, g_k\right\rangle = \langle \tilde{x}, g_k\rangle$$

with $\tilde{x} = \mathbb{S}^{-1}x$. Therefore,

$$\langle x, \tilde{x}\rangle = \left\langle \sum_{k\in\mathcal{K}} c_k g_k, \tilde{x} \right\rangle = \sum_{k\in\mathcal{K}} c_k \langle g_k, \tilde{x}\rangle = \sum_{k\in\mathcal{K}} c_k c_k^* = \sum_{k\in\mathcal{K}} |c_k|^2$$

and

$$\langle x, \tilde{x} \rangle = \left\langle \sum_{k \in \mathcal{K}} a_k g_k, \tilde{x} \right\rangle = \sum_{k \in \mathcal{K}} a_k \langle g_k, \tilde{x} \rangle = \sum_{k \in \mathcal{K}} a_k c_k^*.$$

We can therefore conclude that

$$\sum_{k \in \mathcal{K}} |c_k|^2 = \sum_{k \in \mathcal{K}} a_k c_k^* = \sum_{k \in \mathcal{K}} a_k^* c_k. \tag{1.57}$$

Hence,

$$\begin{aligned} \sum_{k \in \mathcal{K}} |c_k|^2 + \sum_{k \in \mathcal{K}} |c_k - a_k|^2 &= \sum_{k \in \mathcal{K}} |c_k|^2 + \sum_{k \in \mathcal{K}} (c_k - a_k)(c_k^* - a_k^*) \\ &= \sum_{k \in \mathcal{K}} |c_k|^2 + \sum_{k \in \mathcal{K}} |c_k|^2 - \sum_{k \in \mathcal{K}} c_k a_k^* - \sum_{k \in \mathcal{K}} c_k^* a_k + \sum_{k \in \mathcal{K}} |a_k|^2 . \end{aligned}$$

Using (1.57), we get

$$\sum_{k \in \mathcal{K}} |c_k|^2 + \sum_{k \in \mathcal{K}} |c_k - a_k|^2 = \sum_{k \in \mathcal{K}} |a_k|^2 .$$

□

Note that this lemma implies $\sum_{k \in \mathcal{K}} |a_k|^2 > \sum_{k \in \mathcal{K}} |c_k|^2$, i.e., the coefficient sequence $\{a_k\}_{k \in \mathcal{K}}$ has larger l^2-norm than the coefficient sequence $\{c_k = \langle x, \tilde{g}_k \rangle\}_{k \in \mathcal{K}}$.

Lemma 1.34 ([5]). *Let $\{g_k\}_{k \in \mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$ and $\{\tilde{g}_k\}_{k \in \mathcal{K}}$ its canonical dual frame. Then for each $m \in \mathcal{K}$, we have*

$$\sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 = \frac{1 - |\langle g_m, \tilde{g}_m \rangle|^2 - |1 - \langle g_m, \tilde{g}_m \rangle|^2}{2}.$$

Proof. We can represent g_m in two different ways. Obviously $g_m = \sum_{k \in \mathcal{K}} a_k g_k$ with $a_m = 1$ and $a_k = 0$ for $k \neq m$, so that $\sum_{k \in \mathcal{K}} |a_k|^2 = 1$. Furthermore, we can write $g_m = \sum_{k \in \mathcal{K}} c_k g_k$ with $c_k = \langle g_m, \tilde{g}_k \rangle$. From (1.56) it then follows that

$$\begin{aligned} 1 = \sum_{k \in \mathcal{K}} |a_k|^2 &= \sum_{k \in \mathcal{K}} |c_k|^2 + \sum_{k \in \mathcal{K}} |c_k - a_k|^2 \\ &= \sum_{k \in \mathcal{K}} |c_k|^2 + |c_m - a_m|^2 + \sum_{k \neq m} |c_k - a_k|^2 \\ &= \sum_{k \in \mathcal{K}} |\langle g_m, \tilde{g}_k \rangle|^2 + |\langle g_m, \tilde{g}_m \rangle - 1|^2 + \sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 \\ &= 2 \sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 + |\langle g_m, \tilde{g}_m \rangle|^2 + |1 - \langle g_m, \tilde{g}_m \rangle|^2 \end{aligned}$$

and hence

$$\sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 = \frac{1 - |\langle g_m, \tilde{g}_m \rangle|^2 - |1 - \langle g_m, \tilde{g}_m \rangle|^2}{2}.$$

□

We are now able to formulate an equivalent condition for a frame to be exact.

Theorem 1.35 ([5]). *Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$ and $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ its canonical dual frame. Then,*

1. *$\{g_k\}_{k\in\mathcal{K}}$ is exact if and only if $\langle g_m, \tilde{g}_m\rangle = 1$ for all $m \in \mathcal{K}$;*
2. *$\{g_k\}_{k\in\mathcal{K}}$ is inexact if and only if there exists at least one $m \in \mathcal{K}$ such that $\langle g_m, \tilde{g}_m\rangle \neq 1$.*

Proof. We first show that if $\langle g_m, \tilde{g}_m\rangle = 1$ for all $m \in \mathcal{K}$, then $\{g_k\}_{k\neq m}$ is incomplete for $\mathcal{H}$ (for all $m \in \mathcal{K}$) and hence $\{g_k\}_{k\in\mathcal{K}}$ is an exact frame for $\mathcal{H}$. Indeed, fix an arbitrary $m \in \mathcal{K}$. From Lemma 1.34 we have

$$\sum_{k\neq m} |\langle g_m, \tilde{g}_k\rangle|^2 = \frac{1 - |\langle g_m, \tilde{g}_m\rangle|^2 - |1 - \langle g_m, \tilde{g}_m\rangle|^2}{2}.$$

Since $\langle g_m, \tilde{g}_m\rangle = 1,$ we have $\sum_{k\neq m} |\langle g_m, \tilde{g}_k\rangle|^2 = 0$ so that $\langle g_m, \tilde{g}_k\rangle = \langle \tilde{g}_m, g_k\rangle = 0$ for all $k \neq m$. But $\tilde{g}_m \neq 0$ since $\langle g_m, \tilde{g}_m\rangle = 1$. Therefore, $\{g_k\}_{k\neq m}$ is incomplete for $\mathcal{H}$, because $\tilde{g}_m \neq 0$ is orthogonal to all elements of the set $\{g_k\}_{k\neq m}$.

Next, we show that if there exists at least one $m \in \mathcal{K}$ such that $\langle g_m, \tilde{g}_m\rangle \neq 1$, then $\{g_k\}_{k\in\mathcal{K}}$ is inexact. More specifically, we will show that $\{g_k\}_{k\neq m}$ is still a frame for $\mathcal{H}$ if $\langle g_m, \tilde{g}_m\rangle \neq 1$. We start by noting that

$$g_m = \sum_{k\in\mathcal{K}} \langle g_m, \tilde{g}_k\rangle g_k = \langle g_m, \tilde{g}_m\rangle g_m + \sum_{k\neq m} \langle g_m, \tilde{g}_k\rangle g_k. \tag{1.58}$$

If $\langle g_m, \tilde{g}_m\rangle \neq 1$, (1.58) can be rewritten as

$$g_m = \frac{1}{1 - \langle g_m, \tilde{g}_m\rangle} \sum_{k\neq m} \langle g_m, \tilde{g}_k\rangle\, g_k,$$

and for every $x \in \mathcal{H}$ we have

$$\begin{aligned} |\langle x, g_m\rangle|^2 &= \left|\frac{1}{1 - \langle g_m, \tilde{g}_m\rangle}\right|^2 \left|\sum_{k\neq m} \langle g_m, \tilde{g}_k\rangle\, \langle x, g_k\rangle\right|^2 \\ &\leq \frac{1}{|1 - \langle g_m, \tilde{g}_m\rangle|^2} \left[\sum_{k\neq m} |\langle g_m, \tilde{g}_k\rangle|^2\right] \left[\sum_{k\neq m} |\langle x, g_k\rangle|^2\right]. \end{aligned}$$

Therefore

$$\begin{aligned}
\sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 &= |\langle x, g_m\rangle|^2 + \sum_{k\neq m} |\langle x, g_k\rangle|^2 \\
&\leq \frac{1}{|1-\langle g_m, \tilde{g}_m\rangle|^2} \left[\sum_{k\neq m} |\langle g_m, \tilde{g}_k\rangle|^2\right] \left[\sum_{k\neq m} |\langle x, g_k\rangle|^2\right] + \sum_{k\neq m} |\langle x, g_k\rangle|^2 \\
&= \sum_{k\neq m} |\langle x, g_k\rangle|^2 \underbrace{\left[1 + \frac{1}{|1-\langle g_m, \tilde{g}_m\rangle|^2} \sum_{k\neq m} |\langle g_m, \tilde{g}_k\rangle|^2\right]}_{C} \\
&= C \sum_{k\neq m} |\langle x, g_k\rangle|^2
\end{aligned}$$

or equivalently

$$\frac{1}{C} \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 \leq \sum_{k\neq m} |\langle x, g_k\rangle|^2 .$$

With (1.36) it follows that

$$\frac{A}{C}\|x\|^2 \leq \frac{1}{C} \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 \leq \sum_{k\neq m} |\langle x, g_k\rangle|^2 \leq \sum_{k\in\mathcal{K}} |\langle x, g_k\rangle|^2 \leq B\|x\|^2, \tag{1.59}$$

where A and B are the frame bounds of the frame $\{g_k\}_{k\in\mathcal{K}}$. Note that (trivially) $C > 0$; moreover $C < \infty$ since $\langle g_m, \tilde{g}_m\rangle \neq 1$ and $\sum_{k\neq m} |\langle g_m, \tilde{g}_k\rangle|^2 < \infty$ as a consequence of $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ being a frame for $\mathcal{H}$. This implies that $A/C > 0$, and, therefore, (1.59) shows that $\{g_k\}_{k\neq m}$ is a frame with frame bounds A/C and B.

To see that, conversely, exactness of $\{g_k\}_{k\in\mathcal{K}}$ implies that $\langle g_m, \tilde{g}_m\rangle = 1$ for all $m \in \mathcal{K}$, we suppose that $\{g_k\}_{k\in\mathcal{K}}$ is exact and $\langle g_m, \tilde{g}_m\rangle \neq 1$ for at least one $m \in \mathcal{K}$. But the condition $\langle g_m, \tilde{g}_m\rangle \neq 1$ for at least one $m \in \mathcal{K}$ implies that $\{g_k\}_{k\in\mathcal{K}}$ is inexact, which results in a contradiction. It remains to show that $\{g_k\}_{k\in\mathcal{K}}$ inexact implies $\langle g_m, \tilde{g}_m\rangle \neq 1$ for at least one $m \in \mathcal{K}$. Suppose that $\{g_k\}_{k\in\mathcal{K}}$ is inexact and $\langle g_m, \tilde{g}_m\rangle = 1$ for all $m \in \mathcal{K}$. But the condition $\langle g_m, \tilde{g}_m\rangle = 1$ for all $m \in \mathcal{K}$ implies that $\{g_k\}_{k\in\mathcal{K}}$ is exact, which again results in a contradiction. □

Now we are ready to state the two main results of this section. The first result generalizes the biorthonormality relation (1.20) to the infinite-dimensional setting.

Corollary 1.36 ([5]). *Let $\{g_k\}_{k\in\mathcal{K}}$ be a frame for the Hilbert space $\mathcal{H}$. If $\{g_k\}_{k\in\mathcal{K}}$ is exact, then $\{g_k\}_{k\in\mathcal{K}}$ and its canonical dual $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ are biorthonormal, i.e.,*

$$\langle g_m, \tilde{g}_k\rangle = \begin{cases} 1, & \text{if } k = m \\ 0, & \text{if } k \neq m. \end{cases}$$

Conversely, if $\{g_k\}_{k\in\mathcal{K}}$ and $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ are biorthonormal, then $\{g_k\}_{k\in\mathcal{K}}$ is exact.

Proof. If $\{g_k\}_{k\in\mathcal{K}}$ is exact, then biorthonormality follows by noting that Theorem 1.35 implies $\langle g_m, \tilde{g}_m\rangle = 1$ for all $m \in \mathcal{K}$, and Lemma 1.34 implies $\sum_{k\neq m} |\langle g_m, \tilde{g}_k\rangle|^2 = 0$ for all $m \in \mathcal{K}$ and thus $\langle g_m, \tilde{g}_k\rangle = 0$ for all $k \neq m$. To show that, conversely, biorthonormality of $\{g_k\}_{k\in\mathcal{K}}$ and $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ implies that the frame $\{g_k\}_{k\in\mathcal{K}}$ is exact, we simply note that $\langle g_m, \tilde{g}_m\rangle = 1$ for all $m \in \mathcal{K}$, by Theorem 1.35, implies that $\{g_k\}_{k\in\mathcal{K}}$ is exact. □

The second main result in this section states that the expansion into an exact frame is unique and, therefore, the concept of an exact frame generalizes that of a basis to infinite-dimensional spaces.

Theorem 1.37 ([5]). *If $\{g_k\}_{k\in\mathcal{K}}$ is an exact frame for the Hilbert space $\mathcal{H}$ and $x = \sum_{k\in\mathcal{K}} c_k g_k$ with $x \in \mathcal{H}$, then the coefficients $\{c_k\}_{k\in\mathcal{K}}$ are unique and are given by*

$$c_k = \langle x, \tilde{g}_k\rangle\,,$$

where $\{\tilde{g}_k\}_{k\in\mathcal{K}}$ is the canonical dual frame to $\{g_k\}_{k\in\mathcal{K}}$.

Proof. We know from (1.50) that x can be written as $x = \sum_{k\in\mathcal{K}} \langle x, \tilde{g}_k\rangle\, g_k$. Now assume that there is another set of coefficients $\{c_k\}_{k\in\mathcal{K}}$ such that

$$x = \sum_{k\in\mathcal{K}} c_k g_k. \tag{1.60}$$

Taking the inner product of both sides of (1.60) with $\tilde{g}_m$ and using the biorthonormality relation

$$\langle g_k, \tilde{g}_m\rangle = \begin{cases} 1, & k = m \\ 0, & k \neq m \end{cases}$$

we obtain

$$\langle x, \tilde{g}_m\rangle = \sum_{k\in\mathcal{K}} c_k \langle g_k, \tilde{g}_m\rangle = c_m.$$

Thus, $c_m = \langle x, \tilde{g}_m\rangle$ for all $m \in \mathcal{K}$ and the proof is completed. □

1.4 The Sampling Theorem

We now discuss one of the most important results in signal processing—the sampling theorem. We will then show how the sampling theorem can be interpreted as a frame decomposition.

Consider a signal $x(t)$ in the space of square-integrable functions $\mathscr{L}^2$. In general, we can not expect this signal to be uniquely specified by its samples $\{x(kT)\}_{k\in\mathbb{Z}}$, where T is the sampling period. The sampling theorem tells us, however, that if a signal is strictly bandlimited, i.e., its Fourier transform vanishes outside a certain finite interval, and if T is chosen small enough (relative to the signal's bandwidth), then the samples $\{x(kT)\}_{k\in\mathbb{Z}}$ do uniquely specify the signal and we can reconstruct $x(t)$ from $\{x(kT)\}_{k\in\mathbb{Z}}$ perfectly. The process of obtaining the samples $\{x(kT)\}_{k\in\mathbb{Z}}$

from the continuous-time signal $x(t)$ is called A/D conversion[5]; the process of reconstruction of the signal $x(t)$ from its samples is called digital-to-analog (D/A) conversion. We shall now formally state and prove the sampling theorem.

Let $\widehat{x}(f)$ denote the Fourier transform of the signal $x(t)$, i.e.,

$$\widehat{x}(f) = \int_{-\infty}^{\infty} x(t) e^{-\mathrm{i}2\pi t f} dt.$$

We say that $x(t)$ is bandlimited to $B\,\mathrm{Hz}$ if $\widehat{x}(f) = 0$ for $|f| > B$. Note that this implies that the total bandwidth of $x(t)$, counting positive and negative frequencies, is $2B$. The Hilbert space of $\mathscr{L}^2$ functions that are bandlimited to $B\,\mathrm{Hz}$ is denoted as $\mathcal{L}^2(B)$.

Next, consider the sequence of samples $\bigl\{x[k] \triangleq x(kT)\bigr\}_{k\in\mathbb{Z}}$ of the signal $x(t) \in \mathcal{L}^2(B)$ and compute its discrete-time Fourier transform (DTFT):

$$\begin{aligned}\widehat{x}_d(f) &\triangleq \sum_{k=-\infty}^{\infty} x[k] e^{-\mathrm{i}2\pi k f}\\ &= \sum_{k=-\infty}^{\infty} x(kT) e^{-\mathrm{i}2\pi k f}\\ &= \frac{1}{T}\sum_{k=-\infty}^{\infty} \widehat{x}\!\left(\frac{f+k}{T}\right),\end{aligned} \tag{1.61}$$

where in the last step we used the Poisson summation formula[6] [25, Cor. 2.6].

We can see that $\widehat{x}_d(f)$ is simply a periodized version of $\widehat{x}(f)$. Now, it follows that for $1/T \geq 2B$ there is no overlap between the shifted replica of $\widehat{x}(f/T)$, whereas for $1/T < 2B$, we do get the different shifted versions to overlap (see Figure 1.5). We can therefore conclude that for $1/T \geq 2B$, $\widehat{x}(f)$ can be recovered exactly from $\widehat{x}_d(f)$ by means of applying an ideal lowpass filter with gain T and cutoff frequency BT to $\widehat{x}_d(f)$. Specifically, we find that

$$\widehat{x}(f/T) = \widehat{x}_d(f)\, T\, \widehat{h}_{\mathrm{LP}}(f) \tag{1.62}$$

with

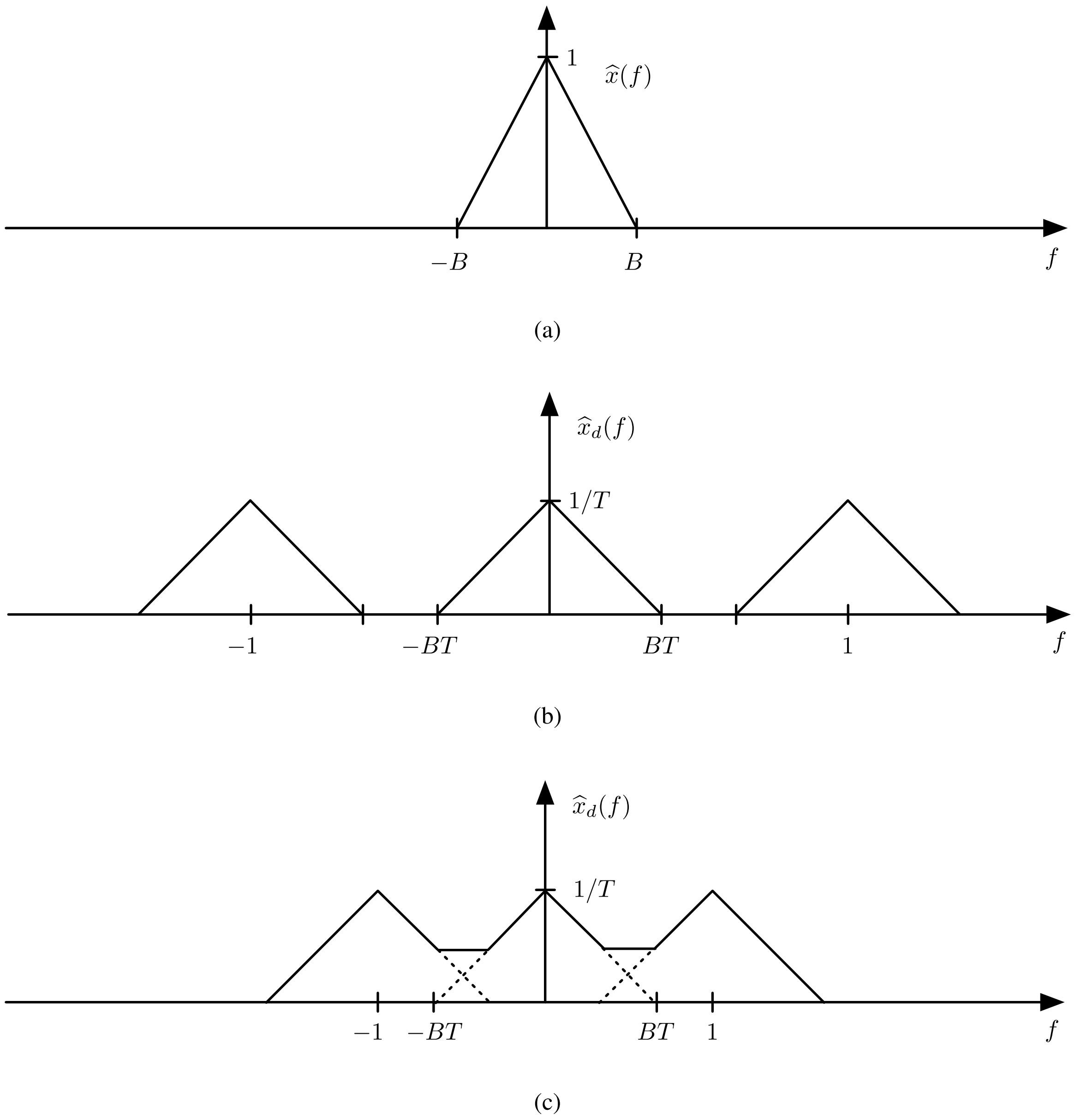
$$\widehat{h}_{\mathrm{LP}}(f) = \begin{cases} 1, & |f| \leq BT \\ 0, & \text{otherwise.}\end{cases} \tag{1.63}$$

From (1.62), using (1.61), we immediately see that we can recover the Fourier transform of $x(t)$ from the sequence of samples $\{x[k]\}_{k\in\mathbb{Z}}$ according to

$$\widehat{x}(f) = T\,\widehat{h}_{\mathrm{LP}}(fT) \sum_{k=-\infty}^{\infty} x[k] e^{-\mathrm{i}2\pi k f T}. \tag{1.64}$$

[5] Strictly speaking A/D conversion also involves quantization of the samples.

[6] Let $x(t) \in \mathcal{L}^2$ with Fourier transform $\widehat{x}(f) = \int_{-\infty}^{\infty} x(t) e^{-\mathrm{i}2\pi t f} dt$. The Poisson summation formula states that $\sum_{k=-\infty}^{\infty} x(k) = \sum_{k=-\infty}^{\infty} \widehat{x}(k)$.



(a)

(b)

(c)

Figure 1.5: Sampling of a signal that is band-limited to B Hz: (a) spectrum of the original signal; (b) spectrum of the sampled signal for $1/T > 2B$; (c) spectrum of the sampled signal for $1/T < 2B$, where aliasing occurs.

We can therefore recover $x(t)$ as follows:

$$\begin{aligned} x(t) &= \int_{-\infty}^{\infty} \widehat{x}(f) e^{\mathrm{i}2\pi t f} df \\ &= \int_{-\infty}^{\infty} T\widehat{h}_{\mathrm{LP}}(fT) \sum_{k=-\infty}^{\infty} x[k] e^{-\mathrm{i}2\pi k f T} e^{\mathrm{i}2\pi f t} df \\ &= \sum_{k=-\infty}^{\infty} x[k] \int_{-\infty}^{\infty} \widehat{h}_{\mathrm{LP}}(fT) e^{\mathrm{i}2\pi f T(t/T-k)} d(fT) \\ &= \sum_{k=-\infty}^{\infty} x[k] h_{\mathrm{LP}}\left(\frac{t}{T} - k\right) \qquad (1.65) \\ &= 2BT \sum_{k=-\infty}^{\infty} x[k] \operatorname{sinc}(2B(t-kT)), \end{aligned}$$

where $h_{\mathrm{LP}}(t)$ is the inverse Fourier transform of $\widehat{h}_{\mathrm{LP}}(f)$, i.e,

$$h_{\mathrm{LP}}(t) = \int_{-\infty}^{\infty} \widehat{h}_{\mathrm{LP}}(f) e^{\mathrm{i}2\pi t f} df,$$

and

$$\operatorname{sinc}(x) \triangleq \frac{\sin(\pi x)}{\pi x}.$$

Summarizing our findings, we obtain the following theorem.

Theorem 1.38 (Sampling theorem [26, Sec. 7.2])**.** *Let $x(t) \in \mathcal{L}^2(B)$. Then $x(t)$ is uniquely specified by its samples $x(kT)$, $k \in \mathbb{Z}$, if $1/T \geq 2B$. Specifically, we can reconstruct $x(t)$ from $x(kT)$, $k \in \mathbb{Z}$, according to*

$$x(t) = 2BT \sum_{k=-\infty}^{\infty} x(kT) \operatorname{sinc}(2B(t-kT)). \qquad (1.66)$$

1.4.1 Sampling Theorem as a Frame Expansion

We shall next show how the representation (1.66) can be interpreted as a frame expansion. The samples $x(kT)$ can be written as the inner product of the signal $x(t)$ with the functions

$$g_k(t) = 2B \operatorname{sinc}(2B(t-kT)), \quad k \in \mathbb{Z}. \qquad (1.67)$$

Indeed, using the fact that the signal $x(t)$ is band-limited to B Hz, we get

$$x(kT) = \int_{-B}^{B} \widehat{x}(f) e^{\mathrm{i}2\pi k f T} df = \langle \widehat{x}, \widehat{g}_k \rangle,$$

where

$$\widehat{g}_k(f) = \begin{cases} e^{-\mathrm{i}2\pi k f T}, \ |f| \le B \\ 0, \ \text{otherwise} \end{cases}$$

is the Fourier transform of $g_k(t)$. We can thus rewrite (1.66) as

$$x(t) = T \sum_{k=-\infty}^{\infty} \langle x, g_k \rangle \, g_k(t).$$

Therefore, the interpolation of an analog signal from its samples $\{x(kT)\}_{k\in\mathbb{Z}}$ can be interpreted as the reconstruction of $x(t)$ from its expansion coefficients $x(kT) = \langle x, g_k \rangle$ in the function set $\{g_k(t)\}_{k\in\mathbb{Z}}$. We shall next prove that $\{g_k(t)\}_{k\in\mathbb{Z}}$ is a frame for the space $\mathcal{L}^2(B)$. Simply note that for $x(t) \in \mathcal{L}^2(B)$, we have

$$\|x\|^2 = \langle x, x \rangle = \left\langle T \sum_{k=-\infty}^{\infty} \langle x, g_k \rangle \, g_k(t), x \right\rangle = T \sum_{k=-\infty}^{\infty} |\langle x, g_k \rangle|^2$$

and therefore

$$\frac{1}{T}\|x\|^2 = \sum_{k=-\infty}^{\infty} |\langle x, g_k \rangle|^2 .$$

This shows that $\{g_k(t)\}_{k\in\mathbb{Z}}$ is, in fact, a tight frame for $\mathcal{L}^2(B)$ with frame bound $A = 1/T$. We emphasize that the frame is tight irrespective of the sampling rate (of course, as long as $1/T > 2B$).

The analysis operator corresponding to this frame is given by $\mathbb{T} : \mathcal{L}^2(B) \to l^2$ as

$$\mathbb{T} : x \to \{\langle x, g_k \rangle\}_{k\in\mathbb{Z}}, \tag{1.68}$$

i.e., $\mathbb{T}$ maps the signal $x(t)$ to the sequence of samples $\{x(kT)\}_{k\in\mathbb{Z}}$.

The action of the adjoint of the analysis operator $\mathbb{T}^* : l^2 \to \mathcal{L}^2(B)$ is to perform interpolation according to

$$\mathbb{T}^* : \{c_k\}_{k\in\mathbb{Z}} \to \sum_{k=-\infty}^{\infty} c_k g_k.$$

The frame operator $\mathbb{S} : \mathcal{L}^2(B) \to \mathcal{L}^2(B)$ is given by $\mathbb{S} = \mathbb{T}^*\mathbb{T}$ and acts as follows

$$\mathbb{S} : x(t) \to \sum_{k=-\infty}^{\infty} \langle x, g_k \rangle \, g_k(t).$$

Since $\{g_k(t)\}_{k\in\mathbb{Z}}$ is a tight frame for $\mathcal{L}^2(B)$ with frame bound $A = 1/T$, as already shown, it follows that $\mathbb{S} = (1/T)\mathbb{I}_{\mathcal{L}^2(B)}$.

The canonical dual frame can be computed easily by applying the inverse of the frame operator to the frame functions $\{g_k(t)\}_{k\in\mathbb{Z}}$ according to

$$\tilde{g}_k(t) = \mathbb{S}^{-1} g_k(t) = T\mathbb{I}_{\mathcal{L}^2(B)} g_k(t) = T g_k(t), \quad k \in \mathbb{Z}.$$

Recall that exact frames have a minimality property in the following sense: If we remove anyone element from an exact frame, the resulting set will be incomplete. In the case of sampling, we have an analogous situation: In the proof of the sampling theorem we saw that if we sample at a rate smaller than the *critical sampling rate* $1/T = 2B$, we cannot recover the signal $x(t)$ from its samples $\{x(kT)\}_{k \in \mathbb{Z}}$. In other words, the set $\{g_k(t)\}_{k \in \mathbb{Z}}$ in (1.67) is *not* complete for $\mathcal{L}^2(B)$ when $1/T < 2B$. This suggests that critical sampling $1/T = 2B$ could implement an exact frame decomposition. We show now that this is, indeed, the case. Simply note that

$$\langle g_k, \tilde{g}_k \rangle = T \langle g_k, g_k \rangle = T\|g_k\|^2 = T\|\widehat{g}_k\|^2 = 2BT, \quad \text{for all } k \in \mathbb{Z}.$$

For critical sampling $2BT = 1$ and, hence, $\langle g_k, \tilde{g}_k \rangle = 1$, for all $k \in \mathbb{Z}$. Theorem 1.35 therefore allows us to conclude that $\{g_k(t)\}_{k \in \mathbb{Z}}$ is an exact frame for $\mathcal{L}^2(B)$.

Next, we show that $\{g_k(t)\}_{k \in \mathbb{Z}}$ is not only an exact frame, but, when properly normalized, even an ONB for $\mathcal{L}^2(B)$, a fact well-known in sampling theory. To this end, we first renormalize the frame functions $g_k(t)$ according to

$$g'_k(t) = \sqrt{T} g_k(t)$$

so that

$$x(t) = \sum_{k=-\infty}^{\infty} \langle x, g'_k \rangle \, g'_k(t).$$

We see that $\{g'_k(t)\}_{k \in \mathbb{Z}}$ is a tight frame for $\mathcal{L}^2(B)$ with $A = 1$. Moreover, we have

$$\|g'_k\|^2 = T\|g_k\|^2 = 2BT.$$

Thus, in the case of critical sampling, $\|g'_k\|^2 = 1$, for all $k \in \mathbb{Z}$, and Theorem 1.29 allows us to conclude that $\{g'_k(t)\}_{k \in \mathbb{Z}}$ is an ONB for $\mathcal{L}^2(B)$.

In contrast to exact frames, inexact frames are redundant, in the sense that there is at least one element that can be removed with the resulting set still being complete. The situation is similar in the *oversampled* case, i.e., when the sampling rate satisfies $1/T > 2B$. In this case, we collect more samples than actually needed for perfect reconstruction of $x(t)$ from its samples. This suggests that $\{g_k(t)\}_{k \in \mathbb{Z}}$ could be an inexact frame for $\mathcal{L}^2(B)$ in the oversampled case. Indeed, according to Theorem 1.35 the condition

$$\langle g_m, \tilde{g}_m \rangle = 2BT < 1, \quad \text{for all } m \in \mathbb{Z}, \tag{1.69}$$

implies that the frame $\{g_k(t)\}_{k \in \mathbb{Z}}$ is inexact for $1/T > 2B$. In fact, as can be seen from the proof of Theorem 1.35, (1.69) guarantees even more: for *every* $m \in \mathbb{Z}$, the set $\{g_k(t)\}_{k \neq m}$ is complete for $\mathcal{L}^2(B)$. Hence, the removal of *any* sample $x(mT)$ from the set of samples $\{x(kT)\}_{k \in \mathbb{Z}}$ still leaves us with a frame decomposition so that $x(t)$ can, in theory, be recovered from the samples $\{x(kT)\}_{k \neq m}$. The resulting frame $\{g_k(t)\}_{k \neq m}$ will, however, no longer be tight, which makes the computation of the canonical dual frame complicated, in general.

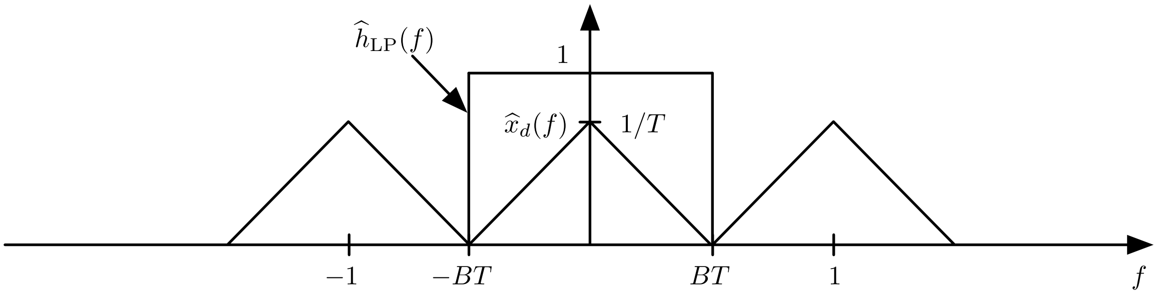


Figure 1.6: Reconstruction filter in the critically sampled case.

1.4.2 Design Freedom in Oversampled A/D Conversion

In the critically sampled case, $1/T = 2B$, the ideal lowpass filter of bandwidth BT with the transfer function specified in (1.63) is the only filter that provides perfect reconstruction of the spectrum $\widehat{x}(f)$ of $x(t)$ according to (1.62) (see Figure 1.6). In the oversampled case, there is, in general, an infinite number of reconstruction filters that provide perfect reconstruction. The only requirement the reconstruction filter has to satisfy is that its transfer function be constant within the frequency range $-BT \leq f \leq BT$ (see Figure 1.7). Therefore, in the oversampled case one has more freedom in designing the reconstruction filter. In A/D converter practice this design freedom is exploited to design reconstruction filters with desirable filter characteristics, like, e.g., rolloff in the transfer function.

Specifically, repeating the steps leading from (1.62) to (1.65), we see that

$$x(t) = \sum_{k=-\infty}^{\infty} x[k] h\left(\frac{t}{T} - k\right), \tag{1.70}$$

where the Fourier transform of $h(t)$ is given by

$$\widehat{h}(f) = \begin{cases} 1, & |f| \leq BT \\ \mathrm{arb}(f), & BT < |f| \leq \frac{1}{2} \\ 0, & |f| > \frac{1}{2} \end{cases} . \tag{1.71}$$

Here and in what follows $\mathrm{arb}(\cdot)$ denotes an arbitrary bounded function. In other words, every set $\{h(t/T - k)\}_{k\in\mathbb{Z}}$ with the Fourier transform of $h(t)$ satisfying (1.71) is a valid dual frame for the frame $\{g_k(t) = 2B\,\mathrm{sinc}(2B(t - kT))\}_{k\in\mathbb{Z}}$. Obviously, there are infinitely many dual frames in the oversampled case.

We next show how the freedom in the design of the reconstruction filter with transfer function specified in (1.71) can be interpreted in terms of the freedom in choosing the left-inverse $\mathbb{L}$ of the analysis operator $\mathbb{T}$ as discussed in Section 1.3.3. Recall the parametrization (1.52) of all left-inverses of the operator $\mathbb{T}$:

$$\mathbb{L} = \tilde{\mathbb{T}}^{*}\mathbb{P} + \mathbb{M}(\mathbb{I}_{l^2} - \mathbb{P}), \tag{1.72}$$

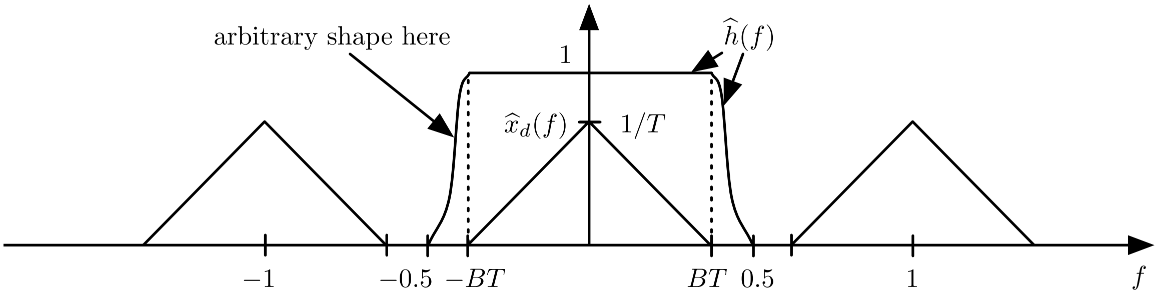


Figure 1.7: Freedom in the design of the reconstruction filter.

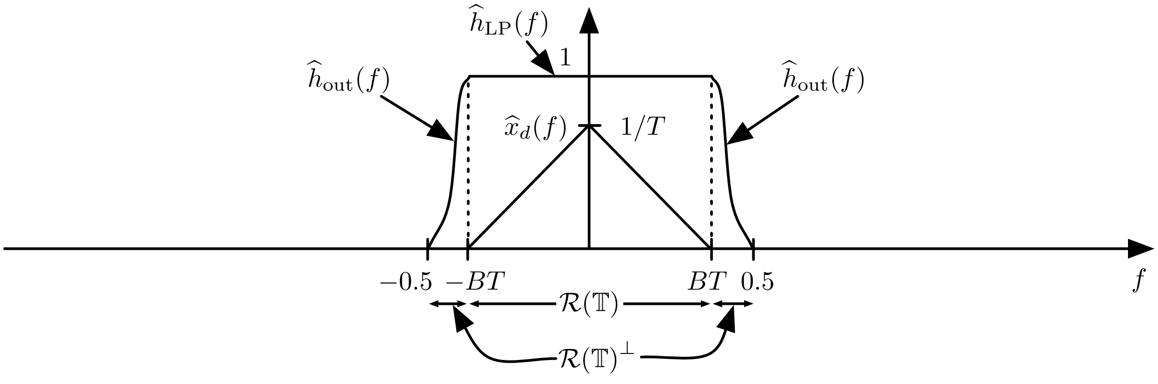

$\widehat{h}_{\mathrm{LP}}(f)$
1
$\widehat{h}_{\mathrm{out}}(f)$
$\widehat{h}_{\mathrm{out}}(f)$
$\widehat{x}_d(f)$
$1/T$
-0.5
$-BT$
BT
0.5
f
$\mathcal{R}(\mathbb{T})$
$\mathcal{R}(\mathbb{T})^{\perp}$

Figure 1.8: The reconstruction filter as a parametrized left-inverse of the analysis operator.

where $\mathbb{M} : l^2 \to \mathcal{H}$ is an arbitrary bounded linear operator and $\mathbb{P} : l^2 \to l^2$ is the orthogonal projection operator onto the range space of $\mathbb{T}$. In (1.61) we saw that the DTFT[7] of the sequence $\{x[k] = x(kT)\}_{k\in\mathbb{Z}}$ is compactly supported on the frequency interval $[-BT, BT]$ (see Figure 1.8). In other words, the range space of the analysis operator $\mathbb{T}$ defined in (1.68) is the space of l^2-sequences with DTFT supported on the interval $[-BT, BT]$ (see Figure 1.8). It is left as an exercise to the reader to verify, using Parseval's theorem,[8] that the orthogonal complement of the range space of $\mathbb{T}$ is the space of l^2-sequences with DTFT supported on the set $[-1/2, -BT] \cup [BT, 1/2]$ (see Figure 1.8). Thus, in the case of oversampled A/D conversion, the operator $\mathbb{P} : l^2 \to l^2$ is the orthogonal projection operator onto the subspace of l^2-sequences with DTFT supported on the interval $[-BT, BT]$; the operator $(\mathbb{I}_{l^2} - \mathbb{P}) : l^2 \to l^2$ is the orthogonal projection operator onto the subspace of l^2-sequences with DTFT supported on the set $[-1/2, -BT] \cup [BT, 1/2]$.

[7] The DTFT is a periodic function with period one. From here on, we consider the DTFT as a function supported on its fundamental period $[-1/2, 1/2]$.

[8] Let $\{a_k\}_{k\in\mathbb{Z}}, \{b_k\}_{k\in\mathbb{Z}} \in l^2$ with DTFT $\widehat{a}(f) = \sum_{k=-\infty}^{\infty} a_k e^{-\mathrm{i}2\pi kf}$ and $\widehat{b}(f) = \sum_{k=-\infty}^{\infty} b_k e^{-\mathrm{i}2\pi kf}$, respectively. Parseval's theorem states that $\sum_{k=-\infty}^{\infty} a_k b_k^* = \int_{-1/2}^{1/2} \widehat{a}(f)\widehat{b}^*(f)df$. In particular, $\sum_{k=-\infty}^{\infty} |a_k|^2 = \int_{-1/2}^{1/2} |\widehat{a}(f)|^2\, df$.

To see the parallels between (1.70) and (1.72), let us decompose $h(t)$ as follows (see Figure 1.8)

$$h(t) = h_{\text{LP}}(t) + h_{\text{out}}(t)\,, \tag{1.73}$$

where the Fourier transform of $h_{\text{LP}}(t)$ is given by (1.63) and the Fourier transform of $h_{\text{out}}(t)$ is

$$\widehat{h}_{\text{out}}(f) = \begin{cases} \operatorname{arb}(f), & BT \le |f| \le \frac{1}{2} \\ 0,\ \text{otherwise.} & \end{cases} \tag{1.74}$$

Now it is clear, and it is left to the reader to verify formally, that the operator $\mathbb{A} : l^2 \to \mathcal{L}^2(B)$ defined as

$$\mathbb{A} : \{c_k\}_{k\in\mathbb{Z}} \to \sum_{k=-\infty}^{\infty} c_k h_{\text{LP}}\left(\frac{t}{T} - k\right) \tag{1.75}$$

acts by first projecting the sequence $\{c_k\}_{k\in\mathbb{Z}}$ onto the subspace of l^2-sequences with DTFT supported on the interval $[-BT, BT]$ and then performs interpolation using the canonical dual frame elements $\tilde{g}_k(t) = h_{\text{LP}}(t/T - k)$. In other words $\mathbb{A} = \tilde{\mathbb{T}}^*\mathbb{P}$. Similarly, it is left to the reader to verify formally, that the operator $\mathbb{B} : l^2 \to \mathcal{L}^2(B)$ defined as

$$\mathbb{B} : \{c_k\}_{k\in\mathbb{Z}} \to \sum_{k=-\infty}^{\infty} c_k h_{\text{out}}\left(\frac{t}{T} - k\right) \tag{1.76}$$

can be written as $\mathbb{B} = \mathbb{M}(\mathbb{I}_{l^2} - \mathbb{P})$. Here, $(\mathbb{I}_{l^2} - \mathbb{P}) : l^2 \to l^2$ is the projection operator onto the subspace of l^2-sequences with DTFT supported on the set $[-1/2, -BT] \cup [BT, 1/2]$; the operator $\mathbb{M} : l^2 \to \mathcal{L}^2$ is defined as

$$\mathbb{M} : \{c_k\}_{k\in\mathbb{Z}} \to \sum_{k=-\infty}^{\infty} c_k h_M\left(\frac{t}{T} - k\right) \tag{1.77}$$

with the Fourier transform of $h_M(t)$ given by

$$\widehat{h}_M(f) = \begin{cases} \operatorname{arb}_2(f), & -\frac{1}{2} \le |f| \le \frac{1}{2} \\ 0, & \text{otherwise,} \end{cases} \tag{1.78}$$

where $\operatorname{arb}_2(f)$ is an arbitrary bounded function that equals $\operatorname{arb}(f)$ for $BT \le |f| \le \frac{1}{2}$. To summarize, we note that the operator $\mathbb{B}$ corresponds to the second term on the right-hand side of (1.72).

We can therefore write the decomposition (1.70) as

$$\begin{aligned} x(t) &= \sum_{k=-\infty}^{\infty} x[k] h\left(\frac{t}{T} - k\right) \\ &= \underbrace{\sum_{k=-\infty}^{\infty} x[k] h_{\text{LP}}\left(\frac{t}{T} - k\right)}_{\tilde{\mathbb{T}}^*\mathbb{P}\mathbb{T}x(t)} + \underbrace{\sum_{k=-\infty}^{\infty} x[k] h_{\text{out}}\left(\frac{t}{T} - k\right)}_{\mathbb{M}(\mathbb{I}_{l^2} - \mathbb{P})\mathbb{T}x(t)} \\ &= \mathbb{L}\mathbb{T}x(t). \end{aligned}$$

1.4.3 Noise Reduction in Oversampled A/D Conversion

Consider again the bandlimited signal $x(t) \in \mathcal{L}^2(B)$. Assume, as before, that the signal is sampled at a rate $1/T \geq 2B$. Now assume that the corresponding samples $x[k] = x(kT),\ k \in \mathbb{Z}$, are subject to noise, i.e., we observe

$$x'[k] = x[k] + w[k],\ k \in \mathbb{Z},$$

where the $w[k]$ are independent identically distributed zero-mean random variables, with variance $\mathbb{E}\,|w[k]|^2 = \sigma^2$. Assume that reconstruction is performed from the noisy samples $x'[k],\ k \in \mathbb{Z}$, using the ideal lowpass filter with transfer function $\widehat{h}_{\text{LP}}(f)$ of bandwidth BT specified in (1.63), i.e., we reconstruct using the canonical dual frame according to

$$x'(t) = \sum_{k=-\infty}^{\infty} x'[k] h_{\text{LP}}\!\left(\frac{t}{T} - k\right).$$

Obviously, the presence of noise precludes perfect reconstruction. It is, however, interesting to assess the impact of oversampling on the variance of the reconstruction error defined as

$$\sigma^2_{\text{oversampling}} \triangleq \mathbb{E}_w\,|x(t) - x'(t)|^2, \tag{1.79}$$

where the expectation is with respect to the random variables $w[k],\ k \in \mathbb{Z}$, and the right-hand side of (1.79) does not depend on t, as we shall see below. If we decompose $x(t)$ as in (1.65), we see that

$$\begin{aligned}\sigma^2_{\text{oversampling}} &= \mathbb{E}_w\,|x(t) - x'(t)|^2 \qquad (1.80)\\ &= \mathbb{E}_w \left| \sum_{k=-\infty}^{\infty} w[k] h_{\text{LP}}\!\left(\frac{t}{T} - k\right)\right|^2 \\ &= \sum_{k=-\infty}^{\infty} \sum_{k'=-\infty}^{\infty} \mathbb{E}_w\{w[k]w^*[k']\}\, h_{\text{LP}}\!\left(\frac{t}{T} - k\right) h^*_{\text{LP}}\!\left(\frac{t}{T} - k'\right) \\ &= \sigma^2 \sum_{k=-\infty}^{\infty} \left| h_{\text{LP}}\!\left(\frac{t}{T} - k\right)\right|^2. \qquad (1.81)\end{aligned}$$

Next applying the Poisson summation formula (as stated in Footnote 6) to the function $l(t') \triangleq h_{\text{LP}}\big(\frac{t}{T} - t'\big)\, e^{-2\pi i t' f}$ with Fourier transform $\widehat{l}(f') = \widehat{h}_{\text{LP}}(-f - f')e^{-2\pi i (t/T)(f+f')}$, we have

$$\sum_{k=-\infty}^{\infty} h_{\text{LP}}\!\left(\frac{t}{T} - k\right) e^{-2\pi i k f} = \sum_{k=-\infty}^{\infty} l(k) = \sum_{k=-\infty}^{\infty} \widehat{l}(k) = \sum_{k=-\infty}^{\infty} \widehat{h}_{\text{LP}}(-f - k)e^{-2\pi i (t/T)(f+k)}. \tag{1.82}$$

Since $\widehat{h}_{\text{LP}}(f)$ is zero outside the interval $-1/2 \leq f \leq 1/2$, it follows that

$$\sum_{k=-\infty}^{\infty} \widehat{h}_{\text{LP}}(-f - k)e^{-2\pi i (t/T)(f+k)} = \widehat{h}_{\text{LP}}(-f)e^{-2\pi i (t/T) f}, \quad \text{for } f \in [-1/2, 1/2]. \tag{1.83}$$

We conclude from (1.82) and (1.83) that the DTFT of the sequence $\left\{a_k \triangleq h_{\text{LP}}(t/T-k)\right\}_{k\in\mathbb{Z}}$ is given (in the fundamental interval $f \in [-1/2, 1/2]$) by $\widehat{h}_{\text{LP}}(-f)e^{-2\pi \mathrm{i}(t/T)f}$ and hence we can apply Parseval's theorem (as stated in Footnote 8) and rewrite (1.81) according to

$$\sigma^2_{\text{oversampling}} = \sigma^2 \sum_{k=-\infty}^{\infty} \left| h_{\text{LP}}\left(\frac{t}{T}-k\right)\right|^2 = \sigma^2 \int_{-1/2}^{1/2} \left|\widehat{h}_{\text{LP}}(-f)e^{-2\pi \mathrm{i}(t/T)f}\right|^2 df = \sigma^2 \int_{-1/2}^{1/2} \left|\widehat{h}_{\text{LP}}(f)\right|^2 df = \sigma^2 2BT. \tag{1.84}$$

We see that the average mean squared reconstruction error is inversely proportional to the oversampling factor $1/(2BT)$. Therefore, each doubling of the oversampling factor decreases the mean squared error by $3\,\mathrm{dB}$.

Consider now reconstruction performed using a general filter that provides perfect reconstruction in the noiseless case. Specifically, we have

$$x'(t) = \sum_{k=-\infty}^{\infty} x'[k] h\left(\frac{t}{T}-k\right),$$

where $h(t)$ is given by (1.73). In this case, the average mean squared reconstruction error can be computed repeating the steps leading from (1.80) to (1.84) and is given by

$$\sigma^2_{\text{oversampling}} = \sigma^2 \int_{-1/2}^{1/2} \left|\widehat{h}(f)\right|^2 df \tag{1.85}$$

where $\widehat{h}(f)$ is the Fourier transform of $h(t)$ and is specified in (1.71). Using (1.73), we can now decompose $\sigma^2_{\text{oversampling}}$ in (1.85) into two terms according to

$$\sigma^2_{\text{oversampling}} = \sigma^2 \underbrace{\int_{-BT}^{BT} \left|\widehat{h}_{\text{LP}}(f)\right|^2 df}_{2BT} + \sigma^2 \int_{BT\le|f|\le 1/2} \left|\widehat{h}_{\text{out}}(f)\right|^2 df. \tag{1.86}$$

We see that two components contribute to the reconstruction error. Comparing (1.86) to (1.84), we conclude that the first term in (1.86) corresponds to the error due to noise in the signal-band $|f| \le BT$ picked up by the ideal lowpass filter with transfer function $\widehat{h}_{\text{LP}}(f)$. The second term in (1.86) is due to the fact that a generalized inverse passes some of the noise in the out-of-band region $BT \le |f| \le 1/2$. The amount of additional noise in the reconstructed signal is determined by the bandwidth and the shape of the reconstruction filter's transfer function in the out-of-band region. In this sense, there exists a tradeoff between noise reduction and design freedom in oversampled A/D conversion. Practically desirable (or realizable) reconstruction filters (i.e., filters with rolloff) lead to additional reconstruction error.

1.5 Important Classes of Frames

There are two important classes of structured signal expansions that have found widespread use in practical applications, namely Weyl-Heisenberg (or Gabor) expansions and affine (or wavelet) expansions. Weyl-Heisenberg expansions provide a decomposition into time-shifted and modulated versions of a "window function" $g(t)$. Wavelet expansions realize decompositions into time-shifted and dilated versions of a mother wavelet $g(t)$. Thanks to the strong structural properties of Weyl-Heisenberg and wavelet expansions, there are efficient algorithms for applying the corresponding analysis and synthesis operators. Weyl-Heisenberg and wavelet expansions have been successfully used in signal detection, image representation, object recognition, and wireless communications. We shall next show that these signal expansions can be cast into the language of frame theory. For a detailed analysis of these classes of frames, we refer the interested reader to [4].

1.5.1 Weyl-Heisenberg Frames

We start by defining a linear operator that realizes time-frequency shifts when applied to a given function.

Definition 1.39. The Weyl operator $\mathbb{W}_{m,n}^{(T,F)} : \mathcal{L}^2 \to \mathcal{L}^2$ is defined as

$$\mathbb{W}_{m,n}^{(T,F)} : x(t) \to e^{\mathrm{i}2\pi nFt} x(t - mT),$$

where $m, n \in \mathbb{Z}$, and $T > 0$ and $F > 0$ are fixed time and frequency shift parameters, respectively.

Now consider some prototype (or window) function $g(t) \in \mathcal{L}^2$. Fix the parameters $T > 0$ and $F > 0$. By shifting the window function $g(t)$ in time by integer multiples of T and in frequency by integer multiples of F, we get a highly-structured set of functions according to

$$g_{m,n}(t) \triangleq \mathbb{W}_{m,n}^{(T,F)} g(t) = e^{\mathrm{i}2\pi nFt} g(t - mT), \quad m \in \mathbb{Z}, \ n \in \mathbb{Z}.$$

The set $\left\{g_{m,n}(t) = e^{\mathrm{i}2\pi nFt} g(t - mT)\right\}_{m\in\mathbb{Z},\, n\in\mathbb{Z}}$ is referred to as a *Weyl-Heisenberg (WH) set* and is denoted by (g, T, F). When the Weyl-Heisenberg (WH) set (g, T, F) is a frame for $\mathcal{L}^2$, it is called a *WH frame* for $\mathcal{L}^2$.

Whether or not a WH set (g, T, F) is a frame for $\mathcal{L}^2$ is, in general, difficult to answer. The answer depends on the window function $g(t)$ as well as on the shift parameters T and F. Intuitively, if the parameters T and F are "too large" for a given window function $g(t)$, the WH set (g, T, F) cannot be a frame for $\mathcal{L}^2$. This is because a WH set (g, T, F) with "large" parameters T and F "leaves holes in the time-frequency plane" or equivalently in the Hilbert space $\mathcal{L}^2$. Indeed, this intuition is correct and the following fundamental result formalizes it:

Theorem 1.40 ([21, Thm. 8.3.1]). *Let $g(t) \in \mathcal{L}^2$ and $T, F > 0$ be given. Then the following holds:*

- *If $TF > 1$, then (g, T, F) is* not *a frame for $\mathcal{L}^2$.*
- *If (g, T, F) is a frame for $\mathscr{L}^2$, then (g, T, F) is an exact frame if and only if $TF = 1$.*

We see that (g, T, F) can be a frame for $\mathscr{L}^2$ only if $TF \leq 1$, i.e., when the shift parameters T and F are such that the grid they induce in the time-frequency plane is sufficiently dense. Whether or not a WH set (g, T, F) with $TF \leq 1$ is a frame for $\mathcal{L}^2$ depends on the window function $g(t)$ and on the values of T and F. There is an important special case where a simple answer can be given.

Example 1.41 (Gaussian, [21, Thm. 8.6.1])**.** Let $T, F > 0$ and take $g(t) = e^{-t^2}$. Then the WH set

$$\left\{\mathbb{W}_{m,n}^{(T,F)} g(t)\right\}_{m\in\mathbb{Z},\, n\in\mathbb{Z}}$$

is a frame for $\mathcal{L}^2$ if and only if $TF < 1$.

1.5.2 Wavelets

Both for wavelet frames and WH frames we deal with function sets that are obtained by letting a special class of parametrized operators act on a fixed function. In the case of WH frames the underlying operator realizes time and frequency shifts. In the case of wavelets, the generating operator realizes time-shifts and scaling. Specifically, we have the following definition.

Definition 1.42. The operator $\mathbb{V}_{m,n}^{(T,S)} : \mathcal{L}^2 \to \mathcal{L}^2$ is defined as

$$\mathbb{V}_{m,n}^{(T,S)} : x(t) \to S^{n/2} x(S^n t - mT),$$

where $m, n \in \mathbb{Z}$, and $T > 0$ and $S > 0$ are fixed time and scaling parameters, respectively.

Now, just as in the case of WH expansions, consider a prototype function (or mother wavelet) $g(t) \in \mathcal{L}^2$. Fix the parameters $T > 0$ and $S > 0$ and consider the set of functions

$$g_{m,n}(t) \triangleq \mathbb{V}_{m,n}^{(T,S)} g(t) = S^{n/2} g(S^n t - mT), \quad m \in \mathbb{Z},\ n \in \mathbb{Z}.$$

This set is referred to as a *wavelet set*. When the wavelet set $\left\{g_{m,n}(t) = S^{n/2} g(S^n t - mT)\right\}_{m\in\mathbb{Z},\, n\in\mathbb{Z}}$ with parameters $T, S > 0$ is a frame for $\mathcal{L}^2$, it is called a *wavelet frame*.

Similar to the case of Weyl-Heisenberg sets it is hard to say, in general, whether a given wavelet set forms a frame for $\mathcal{L}^2$ or not. The answer depends on the window function $g(t)$ and on the parameters T and S and explicit results are known only in certain cases. We conclude this section by detailing such a case.

Example 1.43 (Mexican hat, [21, Ex. 11.2.7])**.** Take $S = 2$ and consider the mother wavelet

$$g(t) = \frac{2}{\sqrt{3}} \pi^{-1/4} (1 - t^2) e^{-\frac{1}{2}t^2}.$$

Due to its shape, $g(t)$ is called the Mexican hat function. It turns out that for each $T < 1.97$, the wavelet set

$$\left\{\mathbb{V}_{m,n}^{(T,S)} g(t)\right\}_{m\in\mathbb{Z},\, n\in\mathbb{Z}}$$

is a frame for $\mathcal{L}^2$ [21, Ex. 11.2.7].

Chapter 2

Uncertainty Relations and Sparse Signal Recovery

2.1 Introduction

The uncertainty principle in quantum mechanics says that certain pairs of physical properties of a particle, such as position and momentum, can be known to within a limited precision only [27]. Uncertainty relations in signal analysis [28–31] state that a signal and its Fourier transform can not both be arbitrarily well concentrated; corresponding mathematical formulations exist for square-integrable or integrable functions [32, 33], for vectors in $(\mathbb{C}^m, \|\cdot\|_2)$ or $(\mathbb{C}^m, \|\cdot\|_1)$ [32–36], and for finite abelian groups [37, 38]. These results feature prominently in many areas of the mathematical data sciences. Specifically, in compressed sensing [32–35, 39, 40] uncertainty relations lead to sparse signal recovery thresholds, in Gabor and Wilson frame theory [41] they characterize limits on the time-frequency localization of frame elements, in communications [42] they play a fundamental role in the design of pulse shapes for orthogonal frequency division multiplexing (OFDM) systems [43], in the theory of partial differential equations they serve to characterize existence and smoothness properties of solutions [44], and in coding theory they help to understand questions around the existence of good cyclic codes [45].

This chapter provides a principled introduction to uncertainty relations underlying sparse signal recovery, starting with the seminal work by Donoho and Stark [32], ranging over the Elad-Bruckstein coherence-based uncertainty relation for general pairs of orthonormal bases [34], to uncertainty relations for general pairs of dictionaries [36]. We also elaborate on the remarkable connection [33] between uncertainty relations for signals and their Fourier transforms—with concentration measured in terms of support—and the “large sieve”, a family of inequalities involving trigonometric polynomials, originally developed in the field of analytic number theory [46, 47]. While the flavor of these results is that beyond certain thresholds something is not possible, for example a nonzero vector can not be concentrated with respect to two different orthonormal bases beyond a certain limit, uncertainty relations can also reveal that something unexpected is possible.

Specifically, we demonstrate that signals that are sparse in certain bases can be recovered in a stable fashion from partial and noisy observations.

To keep the exposition simple and to elucidate the main conceptual aspects, we restrict ourselves to the finite-dimensional cases $(\mathbb{C}^m, \|\cdot\|_2)$ and $(\mathbb{C}^m, \|\cdot\|_1)$ throughout. References to uncertainty relations for the infinite-dimensional case will be given wherever possible and appropriate. Some of the results in this chapter have not been reported before in the literature. Detailed proofs will be provided for most of the statements with the goal of allowing the reader to acquire a technical working knowledge that can serve as a basis for further own research.

The chapter is organized as follows. In Sections 2.2 and 2.3, we derive uncertainty relations for vectors in $(\mathbb{C}^m, \|\cdot\|_2)$ and $(\mathbb{C}^m, \|\cdot\|_1)$, respectively, discuss the connection to the large sieve, present applications to noisy signal recovery problems, and establish a fundamental relation between uncertainty relations for sparse vectors and null-space properties of the accompanying dictionary matrices. Section 2.4 presents a large sieve inequality in $(\mathbb{C}^m, \|\cdot\|_2)$ one of our results in Section 2.2 is based on. Section 2.5 lists infinite-dimensional counterparts—available in the literature—to some of the results in this chapter. Finally, Section 2.6 contains results on operator norms used frequently in this chapter.

Notation. For $\mathcal{A} \subseteq \{1, \dots, m\}$, $\mathbf{D}_\mathcal{A}$ denotes the $m \times m$ diagonal matrix with diagonal entries $(\mathbf{D}_\mathcal{A})_{i,i} = 1$ for $i \in \mathcal{A}$, and $(\mathbf{D}_\mathcal{A})_{i,i} = 0$ else. With $\mathbf{U} \in \mathbb{C}^{m\times m}$ unitary and $\mathcal{A} \subseteq \{1, \dots, m\}$, we define the orthogonal projection $\mathbf{P}_\mathcal{A}(\mathbf{U}) = \mathbf{U}\mathbf{D}_\mathcal{A}\mathbf{U}^\mathsf{H}$ and set $\mathcal{W}^{\mathbf{U},\mathcal{A}} = \mathcal{R}(\mathbf{P}_\mathcal{A}(\mathbf{U}))$. For $\mathbf{x} \in \mathbb{C}^m$ and $\mathcal{A} \subseteq \{1, \dots, m\}$, we let $\mathbf{x}_\mathcal{A} = \mathbf{D}_\mathcal{A}\mathbf{x}$. With $\mathbf{A} \in \mathbb{C}^{m\times m}$, $|||\mathbf{A}|||_1 = \max_{\mathbf{x}\colon \|\mathbf{x}\|_1=1} \|\mathbf{A}\mathbf{x}\|_1$ refers to the operator 1-norm, $|||\mathbf{A}|||_2 = \max_{\mathbf{x}\colon \|\mathbf{x}\|_2=1} \|\mathbf{A}\mathbf{x}\|_2$ designates the operator 2-norm, $\|\mathbf{A}\|_2 = \sqrt{\operatorname{tr}(\mathbf{A}\mathbf{A}^\mathsf{H})}$ is the Frobenius norm, and $\|\mathbf{A}\|_1 = \sum_{i,j=1}^m |A_{i,j}|$. The vector $\mathbf{x} \in \mathbb{C}^m$ is said to be s-sparse if it has at most s nonzero entries. We use the convention $0 \cdot \infty = 0$.

2.2 Uncertainty Relations in $(\mathbb{C}^m, \|\cdot\|_2)$

Donoho and Stark [32] define uncertainty relations as upper bounds on the operator norm of the band-limitation operator followed by the time-limitation operator. We adopt this elegant concept and extend it to refer to an upper bound on the operator norm of a general orthogonal projection operator (replacing the band-limitation operator) followed by the "time-limitation operator" $\mathbf{D}_\mathcal{P}$ as an uncertainty relation. More specifically, let $\mathbf{U} \in \mathbb{C}^{m\times m}$ be a unitary matrix, $\mathcal{P}, \mathcal{Q} \subseteq \{1, \dots, m\}$, and consider the orthogonal projection $\mathbf{P}_\mathcal{Q}(\mathbf{U})$ onto the subspace $\mathcal{W}^{\mathbf{U},\mathcal{Q}}$ which is spanned by $\{\mathbf{u}_i : i \in \mathcal{Q}\}$. Let[1] $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) = |||\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U})|||_2$. In the setting of [32] $\mathbf{U}$ would correspond to the DFT matrix $\mathbf{F}$ and $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F})$ is the operator 2-norm of the band-limitation operator followed by the

[1] We note that, for general unitary $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times m}$, unitary invariance of $\|\cdot\|_2$ yields $|||\mathbf{P}_\mathcal{P}(\mathbf{A})\mathbf{P}_\mathcal{Q}(\mathbf{B})|||_2 = |||\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U})|||_2$ with $\mathbf{U} = \mathbf{A}^\mathsf{H}\mathbf{B}$. The situation where both the band-limitation and the time-limitation operator are replaced by general orthogonal projection operators can hence be reduced to the case considered here.

time-limitation operator, both in finite dimensions. By Lemma 2.20 we have

$$\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) = \max_{\mathbf{x}\in\mathcal{W}^{\mathbf{U},\mathcal{Q}}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{x}_{\mathcal{P}}\|_2}{\|\mathbf{x}\|_2}. \tag{2.1}$$

An uncertainty relation in $(\mathbb{C}^m, \|\cdot\|_2)$ is an upper bound of the form $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \leq c$ with $c \geq 0$, and states that $\|\mathbf{x}_{\mathcal{P}}\|_2 \leq c\|\mathbf{x}\|_2$ for all $\mathbf{x} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$. $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U})$ hence quantifies how well a vector supported on $\mathcal{Q}$ in the basis $\mathbf{U}$ can be concentrated on $\mathcal{P}$. Note that an uncertainty relation in $(\mathbb{C}^m, \|\cdot\|_2)$ is nontrivial only if $c < 1$. Application of Lemma 2.21 now yields

$$\frac{\|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\|_2}{\sqrt{\operatorname{rank}(\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))}} \leq \Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \leq \|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\|_2, \tag{2.2}$$

where the upper bound constitutes an uncertainty relation and the lower bound will allow us to assess its tightness. Next, note that

$$\|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\|_2 = \sqrt{\operatorname{tr}(\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))} \tag{2.3}$$

and

$$\begin{align}\operatorname{rank}(\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})) &= \operatorname{rank}(\mathbf{D}_{\mathcal{P}}\mathbf{U}\mathbf{D}_{\mathcal{Q}}\mathbf{U}^{\mathsf{H}}) \tag{2.4}\\ &\leq \min(|\mathcal{P}|, |\mathcal{Q}|), \tag{2.5}\end{align}$$

where (2.5) follows from $\operatorname{rank}(\mathbf{D}_{\mathcal{P}}\mathbf{U}\mathbf{D}_{\mathcal{Q}}) \leq \min(|\mathcal{P}|, |\mathcal{Q}|)$ and [48, Property (c), Chapter 0.4.5]. When used in (2.2) this implies

$$\sqrt{\frac{\operatorname{tr}(\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))}{\min(|\mathcal{P}|, |\mathcal{Q}|)}} \leq \Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \leq \sqrt{\operatorname{tr}(\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))}. \tag{2.6}$$

Particularizing to $\mathbf{U} = \mathbf{F}$, we obtain

$$\begin{align}\sqrt{\operatorname{tr}(\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{F}))} &= \sqrt{\operatorname{tr}(\mathbf{D}_{\mathcal{P}}\mathbf{F}\mathbf{D}_{\mathcal{Q}}\mathbf{F}^{\mathsf{H}})} \tag{2.7}\\ &= \sqrt{\sum_{i\in\mathcal{P}}\sum_{j\in\mathcal{Q}} |\mathbf{F}_{i,j}|^2} \tag{2.8}\\ &= \sqrt{\frac{|\mathcal{P}||\mathcal{Q}|}{m}}, \tag{2.9}\end{align}$$

so that (2.6) reduces to

$$\sqrt{\frac{\max(|\mathcal{P}|, |\mathcal{Q}|)}{m}} \leq \Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \leq \sqrt{\frac{|\mathcal{P}||\mathcal{Q}|}{m}}. \tag{2.10}$$

There exist sets $\mathcal{P}, \mathcal{Q} \subseteq \{1, \ldots, m\}$ that saturate both bounds in (2.10), e.g., $\mathcal{P} = \{1\}$ and $\mathcal{Q} = \{1, \ldots, m\}$, which yields $\sqrt{\max(|\mathcal{P}|, |\mathcal{Q}|)/m} = \sqrt{|\mathcal{P}||\mathcal{Q}|/m} = 1$ and therefore $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) = 1$.

An example of sets $\mathcal{P}, \mathcal{Q} \subseteq \{1, \ldots, m\}$ saturating only the lower bound in (2.10) is as follows. Take n to divide m and set

$$\mathcal{P} = \left\{ \frac{m}{n}, \frac{2m}{n}, \ldots, \frac{(n-1)m}{n}, m \right\} \tag{2.11}$$

and

$$\mathcal{Q} = \{l+1, \ldots, l+n\} \tag{2.12}$$

with $l \in \{1, \ldots, m\}$ and $\mathcal{Q}$ interpreted circularly in $\{1, \ldots, m\}$. Then, the upper bound in (2.10) is

$$\sqrt{\frac{|\mathcal{P}||\mathcal{Q}|}{m}} = \frac{n}{\sqrt{m}}, \tag{2.13}$$

whereas the lower bound becomes

$$\sqrt{\frac{\max(|\mathcal{P}|, |\mathcal{Q}|)}{m}} = \sqrt{\frac{n}{m}}. \tag{2.14}$$

Thus, for $m \to \infty$ with fixed ratio m/n, the upper bound in (2.10) tends to infinity whereas the corresponding lower bound remains constant. The following result states that the lower bound in (2.10) is tight for $\mathcal{P}$ and $\mathcal{Q}$ as in (2.11) and (2.12), respectively. This implies a lack of tightness of the uncertainty relation $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \leq \sqrt{|\mathcal{P}||\mathcal{Q}|/m}$ by a factor of $\sqrt{n}$. The large sieve-based uncertainty relation developed in the next section will be seen to remedy this problem.

Lemma 2.1. *[32, Theorem 11] Let n divide m and consider*

$$\mathcal{P} = \left\{ \frac{m}{n}, \frac{2m}{n}, \ldots, \frac{(n-1)m}{n}, m \right\} \tag{2.15}$$

and

$$\mathcal{Q} = \{l+1, \ldots, l+n\} \tag{2.16}$$

with $l \in \{1, \ldots, m\}$ and $\mathcal{Q}$ interpreted circularly in $\{1, \ldots, m\}$. Then, $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) = \sqrt{n/m}$.

Proof. We have

$$\begin{aligned}
\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) &= |||\mathbf{P}_{\mathcal{Q}}(\mathbf{F})\mathbf{D}_{\mathcal{P}}|||_2 &(2.17)\\
&= |||\mathbf{D}_{\mathcal{Q}}\mathbf{F}^{\mathsf{H}}\mathbf{D}_{\mathcal{P}}|||_2 &(2.18)\\
&= \max_{\mathbf{x}:\, \|\mathbf{x}\|_2=1} \|\mathbf{D}_{\mathcal{Q}}\mathbf{F}^{\mathsf{H}}\mathbf{D}_{\mathcal{P}}\mathbf{x}\|_2 &(2.19)\\
&= \max_{\mathbf{x}:\, \mathbf{x}\neq\mathbf{0}} \frac{\|\mathbf{D}_{\mathcal{Q}}\mathbf{F}^{\mathsf{H}}\mathbf{x}_{\mathcal{P}}\|_2}{\|\mathbf{x}\|_2} &(2.20)\\
&= \max_{\substack{\mathbf{x}:\, \mathbf{x}=\mathbf{x}_{\mathcal{P}}\\ \mathbf{x}\neq\mathbf{0}}} \frac{\|\mathbf{D}_{\mathcal{Q}}\mathbf{F}^{\mathsf{H}}\mathbf{x}\|_2}{\|\mathbf{x}\|_2}, &(2.21)
\end{aligned}$$

where in (2.17) we applied Lemma 2.20 and in (2.18) we used unitary invariance of $\|\cdot\|_2$. Next, consider an arbitrary but fixed $\mathbf{x} \in \mathbb{C}^m$ with $\mathbf{x} = \mathbf{x}_\mathcal{P}$ and define $\mathbf{y} \in \mathbb{C}^n$ according to $y_s = x_{ms/n}$ for $s = 1, \dots, n$. It follows that

$$\|\mathbf{D}_\mathcal{Q}\mathbf{F}^\mathsf{H}\mathbf{x}\|_2^2 = \frac{1}{m}\sum_{q\in\mathcal{Q}}\Big|\sum_{p\in\mathcal{P}} x_p\, e^{\frac{2\pi i p q}{m}}\Big|^2 \tag{2.22}$$

$$= \frac{1}{m}\sum_{q\in\mathcal{Q}}\Big|\sum_{s=1}^{n} x_{ms/n}\, e^{\frac{2\pi i s q}{n}}\Big|^2 \tag{2.23}$$

$$= \frac{1}{m}\sum_{q\in\mathcal{Q}}\Big|\sum_{s=1}^{n} y_s\, e^{\frac{2\pi i s q}{n}}\Big|^2 \tag{2.24}$$

$$= \frac{n}{m}\|\mathbf{F}^\mathsf{H}\mathbf{y}\|_2^2 \tag{2.25}$$

$$= \frac{n}{m}\|\mathbf{y}\|_2^2, \tag{2.26}$$

where $\mathbf{F}$ in (2.25) is the $n \times n$ DFT matrix and in (2.26) we used unitary invariance of $\|\cdot\|_2$. With (2.22)–(2.26) and $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2$ in (2.21), we get $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) = \sqrt{n/m}$. □

2.2.1 Uncertainty Relations Based on the Large Sieve

The uncertainty relation in (2.6) is very crude as it simply upper-bounds the operator 2-norm by the Frobenius norm. For $\mathbf{U} = \mathbf{F}$ a more sophisticated upper bound on $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F})$ was reported in [33, Theorem 12]. The proof of this result establishes a remarkable connection to the so-called "large sieve", a family of inequalities involving trigonometric polynomials originally developed in the field of analytic number theory [46, 47]. We next present a slightly improved and generalized version of [33, Theorem 12].

Theorem 2.2. *Let $\mathcal{P} \subseteq \{1, \dots, m\}$, $l, n \in \{1, \dots, m\}$, and*

$$\mathcal{Q} = \{l+1, \dots, l+n\} \tag{2.27}$$

with $\mathcal{Q}$ interpreted circularly in $\{1, \dots, m\}$. For $\lambda \in (0, m]$, we define the circular Nyquist density $\rho(\mathcal{P}, \lambda)$ according to

$$\rho(\mathcal{P}, \lambda) = \frac{1}{\lambda}\max_{r\in[0,m)} |\widetilde{\mathcal{P}} \cap (r, r+\lambda)|, \tag{2.28}$$

where $\widetilde{\mathcal{P}} = \mathcal{P} \cup \{m + p : p \in \mathcal{P}\}$. Then,

$$\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \le \sqrt{\left(\frac{\lambda(n-1)}{m} + 1\right)\rho(\mathcal{P}, \lambda)} \tag{2.29}$$

for all $\lambda \in (0, m]$.

Proof. If $\mathcal{P} = \emptyset$, then $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) = 0$ as a consequence of $\mathbf{P}_\emptyset(\mathbf{F}) = \mathbf{0}$ and (2.29) holds trivially. Suppose now that $\mathcal{P} \neq \emptyset$, consider an arbitrary but fixed $\mathbf{x} \in \mathcal{W}^{\mathbf{F},\mathcal{Q}}$ with $\|\mathbf{x}\|_2 = 1$, and set $\mathbf{a} = \mathbf{F}^{\mathsf{H}}\mathbf{x}$. Then, $\mathbf{a} = \mathbf{a}_\mathcal{Q}$ and, by unitarity of $\mathbf{F}$, $\|\mathbf{a}\|_2 = 1$. We have

$$\begin{aligned} |x_p|^2 &= |(\mathbf{F}\mathbf{a})_p|^2 && (2.30)\\ &= \frac{1}{m}\left|\sum_{q\in\mathcal{Q}} a_q e^{-\frac{2\pi i p q}{m}}\right|^2 && (2.31)\\ &= \frac{1}{m}\left|\sum_{k=1}^{n} a_k e^{-\frac{2\pi i p k}{m}}\right|^2 && (2.32)\\ &= \frac{1}{m}\left|\psi\left(\frac{p}{m}\right)\right|^2 \quad \text{for } p \in \{1,\dots,m\}, && (2.33) \end{aligned}$$

where we defined the 1-periodic trigonometric polynomial $\psi(s)$ according to

$$\psi(s) = \sum_{k=1}^{n} a_k e^{-2\pi i k s}. \tag{2.34}$$

Next, let ν_t denote the unit Dirac measure centered at $t \in \mathbb{R}$ and set $\mu = \sum_{p\in\mathcal{P}} \nu_{p/m}$ with 1-periodic extension outside $[0,1)$. Then,

$$\begin{aligned} \|\mathbf{x}_\mathcal{P}\|_2^2 &= \frac{1}{m}\sum_{p\in\mathcal{P}}\left|\psi\left(\frac{p}{m}\right)\right|^2 && (2.35)\\ &= \frac{1}{m}\int_{[0,1)} |\psi(s)|^2 \mathrm{d}\mu(s) && (2.36)\\ &\leq \left(\frac{n-1}{m} + \frac{1}{\lambda}\right) \sup_{r\in[0,1)} \mu\left(\left(r, r+\frac{\lambda}{m}\right)\right) && (2.37) \end{aligned}$$

for all $\lambda \in (0, m]$, where (2.35) is by (2.30)–(2.33) and in (2.37) we applied the large sieve inequality Lemma 2.19 with $\delta = \lambda/m$ and $\|\mathbf{a}\|_2 = 1$. Now,

$$\begin{aligned} &\sup_{r\in[0,1)} \mu\left(\left(r, r+\frac{\lambda}{m}\right)\right) && (2.38)\\ &= \sup_{r\in[0,m)} \sum_{p\in\mathcal{P}} (\nu_p((r, r+\lambda)) + \nu_{m+p}((r, r+\lambda))) && (2.39)\\ &= \max_{r\in[0,m)} |\widetilde{\mathcal{P}} \cap (r, r+\lambda)| && (2.40)\\ &= \lambda\, \rho(\mathcal{P}, \lambda) \quad \text{for all } \lambda \in (0, m], && (2.41) \end{aligned}$$

where in (2.39) we used the 1-periodicity of μ. Using (2.38)–(2.41) in (2.37) yields

$$\|\mathbf{x}_\mathcal{P}\|_2^2 \leq \left(\frac{\lambda(n-1)}{m} + 1\right)\rho(\mathcal{P}, \lambda) \quad \text{for all } \lambda \in (0, m]. \tag{2.42}$$

As $\mathbf{x} \in \mathcal{W}^{\mathbf{F},\mathcal{Q}}$ with $\|\mathbf{x}\|_2 = 1$ was arbitrary, we conclude that

$$\Delta^2_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) = \max_{\mathbf{x}\in\mathcal{W}^{\mathbf{F},\mathcal{Q}}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{x}_{\mathcal{P}}\|_2^2}{\|\mathbf{x}\|_2^2} \tag{2.43}$$

$$\leq \left(\frac{\lambda(n-1)}{m} + 1\right)\rho(\mathcal{P},\lambda) \quad \text{for all } \lambda \in (0,m], \tag{2.44}$$

thereby finishing the proof. □

Theorem 2.2 slightly improves upon [33, Theorem 12] by virtue of applying to more general sets $\mathcal{Q}$ and defining the circular Nyquist density in (2.28) in terms of open intervals $(r, r+\lambda)$.

We next apply Theorem 2.2 to specific choices of $\mathcal{P}$ and $\mathcal{Q}$. First, consider $\mathcal{P} = \{1\}$ and $\mathcal{Q} = \{1,\dots,m\}$, which were shown to saturate the upper and the lower bound in (2.10) leading to $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) = 1$. Since $\mathcal{P}$ consists of a single point, $\rho(\mathcal{P},\lambda) = 1/\lambda$ for all $\lambda \in (0,m]$. Thus, Theorem 2.2 with $n = m$ yields

$$\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \leq \sqrt{\frac{m-1}{m} + \frac{1}{\lambda}} \quad \text{for all } \lambda \in (0,m]. \tag{2.45}$$

Setting $\lambda = m$ in (2.45) yields $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \leq 1$.

Next, consider $\mathcal{P}$ and $\mathcal{Q}$ as in (2.11) and (2.12), respectively, which, as already mentioned, have the uncertainty relation in (2.10) lacking tightness by a factor of $\sqrt{n}$. Since $\mathcal{P}$ consists of points spaced m/n apart, we get $\rho(\mathcal{P},\lambda) = 1/\lambda$ for all $\lambda \in (0, m/n]$. The upper bound (2.29) now becomes

$$\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \leq \sqrt{\frac{n-1}{m} + \frac{1}{\lambda}} \quad \text{for all } \lambda \in \left(0, \frac{m}{n}\right]. \tag{2.46}$$

Setting $\lambda = m/n$ in (2.46) yields

$$\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \leq \sqrt{(2n-1)/m} \leq \sqrt{2}\sqrt{n/m}, \tag{2.47}$$

which is tight up to a factor of $\sqrt{2}$ (cf. Lemma 2.1). We hasten to add, however, that the large sieve technique applies to $\mathbf{U} = \mathbf{F}$ only.

2.2.2 Coherence-based Uncertainty Relation

We next present an uncertainty relation that is of simple form and applies to general unitary $\mathbf{U}$. To this end, we first introduce the concept of coherence of a matrix.

Definition 2.3. For $\mathbf{A} = (\mathbf{a}_1 \dots \mathbf{a}_n) \in \mathbb{C}^{m\times n}$ with columns $\|\cdot\|_2$-normalized to 1, the coherence is defined as $\mu(\mathbf{A}) = \max_{i\neq j} |\mathbf{a}_i^{\mathsf{H}}\mathbf{a}_j|$.

We have the following coherence-based uncertainty relation valid for general unitary $\mathbf{U}$.

Lemma 2.4. *Let* $\mathbf{U} \in \mathbb{C}^{m\times m}$ *be unitary and* $\mathcal{P}, \mathcal{Q} \subseteq \{1, \ldots, m\}$. *Then,*

$$\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \leq \sqrt{|\mathcal{P}||\mathcal{Q}|}\,\mu([\mathbf{I}\ \ \mathbf{U}]). \tag{2.48}$$

Proof. The claim follows from

$$\begin{aligned}\Delta^2_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) &\leq \operatorname{tr}(\mathbf{D}_{\mathcal{P}}\mathbf{U}\mathbf{D}_{\mathcal{Q}}\mathbf{U}^{\mathsf{H}}) && (2.49)\\ &= \sum_{k\in\mathcal{P}}\sum_{l\in\mathcal{Q}} |\mathbf{U}_{k,l}|^2 && (2.50)\\ &\leq |\mathcal{P}||\mathcal{Q}| \max_{k,l} |\mathbf{U}_{k,l}|^2 && (2.51)\\ &= |\mathcal{P}||\mathcal{Q}|\,\mu^2([\mathbf{I}\ \ \mathbf{U}]), && (2.52)\end{aligned}$$

where (2.49) is by (2.6) and in (2.52) we used the definition of coherence. □

Since $\mu([\mathbf{I}\ \ \mathbf{F}]) = 1/\sqrt{m}$, Lemma 2.4 particularized to $\mathbf{U} = \mathbf{F}$ recovers the upper bound in (2.10).

2.2.3 Concentration Inequalities

As mentioned at the beginning of this chapter, the classical uncertainty relation in signal analysis quantifies how well concentrated a signal can be in time and frequency. In the finite-dimensional setting considered here this amounts to characterizing the concentration of $\mathbf{p}$ and $\mathbf{q}$ in $\mathbf{p} = \mathbf{F}\mathbf{q}$. We will actually study the more general case obtained by replacing $\mathbf{I}$ and $\mathbf{F}$ by unitary $\mathbf{A} \in \mathbb{C}^{m\times m}$ and $\mathbf{B} \in \mathbb{C}^{m\times m}$, respectively, and will ask ourselves how well concentrated $\mathbf{p}$ and $\mathbf{q}$ in $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$ can be. Rewriting $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$ according to $\mathbf{p} = \mathbf{U}\mathbf{q}$ with $\mathbf{U} = \mathbf{A}^{\mathsf{H}}\mathbf{B}$, we now show how the uncertainty relation in Lemma 2.4 can be used to answer this question. Let us start by introducing a measure for concentration in $(\mathbb{C}^m, \|\cdot\|_2)$.

Definition 2.5. Let $\mathcal{P} \subseteq \{1, \ldots, m\}$ and $\varepsilon_{\mathcal{P}} \in [0, 1]$. The vector $\mathbf{x} \in \mathbb{C}^m$ is said to be $\varepsilon_{\mathcal{P}}$-concentrated if $\|\mathbf{x} - \mathbf{x}_{\mathcal{P}}\|_2 \leq \varepsilon_{\mathcal{P}}\|\mathbf{x}\|_2$.

The fraction of 2-norm an $\varepsilon_{\mathcal{P}}$-concentrated vector exhibits outside $\mathcal{P}$ is therefore no more than $\varepsilon_{\mathcal{P}}$. In particular, if $\mathbf{x}$ is $\varepsilon_{\mathcal{P}}$-concentrated with $\varepsilon_{\mathcal{P}} = 0$, then $\mathbf{x} = \mathbf{x}_{\mathcal{P}}$ and $\mathbf{x}$ is $|\mathcal{P}|$-sparse. The zero vector is trivially $\varepsilon_{\mathcal{P}}$-concentrated for all $\mathcal{P} \subseteq \{1, \ldots, m\}$ and $\varepsilon_{\mathcal{P}} \in [0, 1]$.

We next derive a lower bound on $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U})$ for unitary matrices $\mathbf{U}$ that relate $\varepsilon_{\mathcal{P}}$-concentrated vectors $\mathbf{p}$ to $\varepsilon_{\mathcal{Q}}$-concentrated vectors $\mathbf{q}$ through $\mathbf{p} = \mathbf{U}\mathbf{q}$. The formal statement is as follows.

Lemma 2.6. *Let* $\mathbf{U} \in \mathbb{C}^{m\times m}$ *be unitary and* $\mathcal{P}, \mathcal{Q} \subseteq \{1, \ldots, m\}$. *Suppose that there exist a nonzero* $\varepsilon_{\mathcal{P}}$*-concentrated* $\mathbf{p} \in \mathbb{C}^m$ *and a nonzero* $\varepsilon_{\mathcal{Q}}$*-concentrated* $\mathbf{q} \in \mathbb{C}^m$ *such that* $\mathbf{p} = \mathbf{U}\mathbf{q}$. *Then,*

$$\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \geq [1 - \varepsilon_{\mathcal{P}} - \varepsilon_{\mathcal{Q}}]_+. \tag{2.53}$$

Proof. We have

$$\begin{aligned}
\|\mathbf{p} - \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{p}_{\mathcal{P}}\|_2 &\leq \|\mathbf{p} - \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{p}\|_2 + \|\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{p}_{\mathcal{P}} - \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{p}\|_2 && (2.54)\\
&\leq \|\mathbf{p} - \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{p}\|_2 + |||\mathbf{P}_{\mathcal{Q}}(\mathbf{U})|||_2\|\mathbf{p}_{\mathcal{P}} - \mathbf{p}\|_2 && (2.55)\\
&\leq \|\mathbf{p} - \mathbf{U}\mathbf{D}_{\mathcal{Q}}\mathbf{U}^{\mathsf{H}}\mathbf{p}\|_2 + \|\mathbf{p}_{\mathcal{P}} - \mathbf{p}\|_2 && (2.56)\\
&= \|\mathbf{q} - \mathbf{q}_{\mathcal{Q}}\|_2 + \|\mathbf{p}_{\mathcal{P}} - \mathbf{p}\|_2, && (2.57)\\
&\leq \varepsilon_{\mathcal{Q}}\|\mathbf{q}\|_2 + \varepsilon_{\mathcal{P}}\|\mathbf{p}\|_2 && (2.58)\\
&= (\varepsilon_{\mathcal{P}} + \varepsilon_{\mathcal{Q}})\|\mathbf{p}\|_2, && (2.59)
\end{aligned}$$

where in (2.57) we made use of the unitary invariance of $\|\cdot\|_2$. It follows that

$$\begin{aligned}
\|\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{p}_{\mathcal{P}}\|_2 &\geq [\|\mathbf{p}\|_2 - \|\mathbf{p} - \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{p}_{\mathcal{P}}\|_2]_+ && (2.60)\\
&\geq \|\mathbf{p}\|_2[1 - \varepsilon_{\mathcal{P}} - \varepsilon_{\mathcal{Q}}]_+, && (2.61)
\end{aligned}$$

where (2.60) is by the reverse triangle inequality and in (2.61) we used (2.54)–(2.59). Since $\mathbf{p} \neq \mathbf{0}$ by assumption, (2.60)–(2.61) implies

$$\left\|\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{D}_{\mathcal{P}}\frac{\mathbf{p}}{\|\mathbf{p}\|_2}\right\|_2 \geq [1 - \varepsilon_{\mathcal{P}} - \varepsilon_{\mathcal{Q}}]_+, \tag{2.62}$$

which in turn yields $|||\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{D}_{\mathcal{P}}|||_2 \geq [1 - \varepsilon_{\mathcal{P}} - \varepsilon_{\mathcal{Q}}]_+$. This concludes the proof as $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) = |||\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{D}_{\mathcal{P}}|||_2$ by Lemma 2.20. □

Combining Lemma 2.6 with the uncertainty relation Lemma 2.4 yields the announced result stating that a nonzero vector can not be arbitrarily well concentrated with respect to two different orthonormal bases.

Corollary 2.7. *Let* $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times m}$ *be unitary and* $\mathcal{P}, \mathcal{Q} \subseteq \{1, \dots, m\}$. *Suppose that there exist a nonzero* $\varepsilon_{\mathcal{P}}$*-concentrated* $\mathbf{p} \in \mathbb{C}^m$ *and a nonzero* $\varepsilon_{\mathcal{Q}}$*-concentrated* $\mathbf{q} \in \mathbb{C}^m$ *such that* $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$. *Then,*

$$|\mathcal{P}||\mathcal{Q}| \geq \frac{[1 - \varepsilon_{\mathcal{P}} - \varepsilon_{\mathcal{Q}}]_+^2}{\mu^2([\mathbf{A}\ \ \mathbf{B}])}. \tag{2.63}$$

Proof. Let $\mathbf{U} = \mathbf{A}^{\mathsf{H}}\mathbf{B}$. Then, by Lemmata 2.4 and 2.6, we have

$$[1 - \varepsilon_{\mathcal{P}} - \varepsilon_{\mathcal{Q}}]_+ \leq \Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \leq \sqrt{|\mathcal{P}||\mathcal{Q}|}\,\mu([\mathbf{I}\ \ \mathbf{U}]). \tag{2.64}$$

The claim now follows by noting that $\mu([\mathbf{I}\ \ \mathbf{U}]) = \mu([\mathbf{A}\ \ \mathbf{B}])$. □

For $\varepsilon_{\mathcal{P}} = \varepsilon_{\mathcal{Q}} = 0$, we recover the well-known Elad-Bruckstein result.

Corollary 2.8. *[34, Theorem 1] Let* $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times m}$ *be unitary. If* $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$*for nonzero* $\mathbf{p}, \mathbf{q} \in \mathbb{C}^m$, *then* $\|\mathbf{p}\|_0\|\mathbf{q}\|_0 \geq 1/\mu^2([\mathbf{A}\ \ \mathbf{B}])$.

2.2.4 Noisy Recovery in $(\mathbb{C}^m, \|\cdot\|_2)$

Uncertainty relations are typically employed to prove that something is not possible. For example, by Corollary 2.7 there is a limit on how well a nonzero vector can be concentrated with respect to two different orthonormal bases. Donoho and Stark [32] noticed that uncertainty relations can also be used to show that something unexpected is possible. Specifically, [32, Section 4] considers a noisy signal recovery problem, which we now translate to the finite-dimensional setting. Let $\mathbf{p}, \mathbf{n} \in \mathbb{C}^m$ and $\mathcal{P} \subseteq \{1, \ldots, m\}$, set $\mathcal{P}^c = \{1, \ldots, m\} \backslash \mathcal{P}$, and suppose that we observe $\mathbf{y} = \mathbf{p}_{\mathcal{P}^c} + \mathbf{n}$. Note that the information contained in $\mathbf{p}_{\mathcal{P}}$ is completely lost in the observation. Without structural assumptions on $\mathbf{p}$, it is therefore not possible to recover information on $\mathbf{p}_{\mathcal{P}}$ from $\mathbf{y}$. However, if $\mathbf{p}$ is sufficiently sparse with respect to an orthonormal basis and $|\mathcal{P}|$ is sufficiently small, it turns out that all entries of $\mathbf{p}$ can be recovered in a linear fashion to within a precision determined by the noise level. This is often referred to in the literature as stable recovery [32]. The corresponding formal statement is as follows.

Lemma 2.9. *Let $\mathbf{U} \in \mathbb{C}^{m\times m}$ be unitary, $\mathcal{Q} \subseteq \{1, \ldots, m\}$, $\mathbf{p} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$, and consider*

$$\mathbf{y} = \mathbf{p}_{\mathcal{P}^c} + \mathbf{n}, \tag{2.65}$$

where $\mathbf{n} \in \mathbb{C}^m$ and $\mathcal{P}^c = \{1, \ldots, m\} \backslash \mathcal{P}$ with $\mathcal{P} \subseteq \{1, \ldots, m\}$. If $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1$, then there exists a matrix $\mathbf{L} \in \mathbb{C}^{m\times m}$ such that

$$\|\mathbf{L}\mathbf{y} - \mathbf{p}\|_2 \leq C\|\mathbf{n}_{\mathcal{P}^c}\|_2 \tag{2.66}$$

with $C = 1/(1 - \Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}))$. In particular,

$$|\mathcal{P}||\mathcal{Q}| < \frac{1}{\mu^2([\mathbf{I}\ \ \mathbf{U}])} \tag{2.67}$$

is sufficient for $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1$.

Proof. For $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1$, it follows that (cf. [48, p. 301]) $(\mathbf{I} - \mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))$ is invertible with

$$|||(\mathbf{I} - \mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))^{-1}|||_2 \leq \frac{1}{1 - |||\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})|||_2} \tag{2.68}$$

$$= \frac{1}{1 - \Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U})}. \tag{2.69}$$

We now set $\mathbf{L} = (\mathbf{I} - \mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))^{-1}\mathbf{D}_{\mathcal{P}^c}$ and note that

$$\mathbf{L}\mathbf{p}_{\mathcal{P}^c} = (\mathbf{I} - \mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))^{-1}\mathbf{p}_{\mathcal{P}^c} \tag{2.70}$$

$$= (\mathbf{I} - \mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))^{-1}(\mathbf{I} - \mathbf{D}_{\mathcal{P}})\mathbf{p} \tag{2.71}$$

$$= (\mathbf{I} - \mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))^{-1}(\mathbf{I} - \mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U}))\mathbf{p} \tag{2.72}$$

$$= \mathbf{p}, \tag{2.73}$$

where in (2.72) we used $\mathbf{P}_\mathcal{Q}(\mathbf{U})\mathbf{p} = \mathbf{p}$, which is by assumption. Next, we upper-bound $\|\mathbf{Ly} - \mathbf{p}\|_2$ according to

$$\|\mathbf{Ly} - \mathbf{p}\|_2 = \|\mathbf{L}\mathbf{p}_{\mathcal{P}^c} + \mathbf{Ln} - \mathbf{p}\|_2 \tag{2.74}$$

$$= \|\mathbf{Ln}\|_2 \tag{2.75}$$

$$\leq |||(\mathbf{I} - \mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U}))^{-1}|||_2 \|\mathbf{n}_{\mathcal{P}^c}\|_2 \tag{2.76}$$

$$\leq \frac{1}{1 - \Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U})} \|\mathbf{n}_{\mathcal{P}^c}\|_2, \tag{2.77}$$

where in (2.75) we used (2.70)–(2.73). Finally, Lemma 2.4 implies that (2.67) is sufficient for $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1$. □

Note that in the noise free case $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1$ is a sufficient condition for perfect recovery of $\mathbf{p}$ in Lemma 2.9. We next particularize Lemma 2.9 for $\mathbf{U} = \mathbf{F}$,

$$\mathcal{P} = \left\{ \frac{m}{n}, \frac{2m}{n}, \dots, \frac{(n-1)m}{n}, m \right\} \tag{2.78}$$

with n dividing m, and

$$\mathcal{Q} = \{l + 1, \dots, l + n\} \tag{2.79}$$

with $l \in \{1, \dots, m\}$ and $\mathcal{Q}$ interpreted circularly in $\{1, \dots, m\}$. This means that $\mathbf{p}$ is n-sparse in $\mathbf{F}$ and we are missing n entries in the noisy observation $\mathbf{y}$. From Lemma 2.1 we know that $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) = \sqrt{n/m}$. Since n divides m by assumption, stable recovery of $\mathbf{p}$ is possible for $n \leq m/2$. In contrast, the coherence-basedcoherence uncertainty relation in Lemma 2.4 yields $\Delta_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \leq \frac{n}{\sqrt{m}}$, and would hence suggest that $n^2 < m$ is needed for stable recovery.

2.3 Uncertainty Relations in $(\mathbb{C}^m, \|\cdot\|_1)$

We introduce uncertainty relations in $(\mathbb{C}^m, \|\cdot\|_1)$ following the same story line as in Section 2.2. Specifically, let $\mathbf{U} = (\mathbf{u}_1 \dots \mathbf{u}_m) \in \mathbb{C}^{m \times m}$ be a unitary matrix, $\mathcal{P}, \mathcal{Q} \subseteq \{1, \dots, m\}$, and consider the orthogonal projection $\mathbf{P}_\mathcal{Q}(\mathbf{U})$ onto the subspace $\mathcal{W}^{\mathbf{U},\mathcal{Q}}$, which is spanned by $\{\mathbf{u}_i : i \in \mathcal{Q}\}$. Let[2] $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) = |||\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U})|||_1$. By Lemma 2.20 we have

$$\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) = \max_{\mathbf{x} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}} \setminus \{\mathbf{0}\}} \frac{\|\mathbf{x}_\mathcal{P}\|_1}{\|\mathbf{x}\|_1}. \tag{2.80}$$

[2] In contrast to the operator 2-norm, the operator 1-norm is not invariant under unitary transformations so that we do not have $|||\mathbf{P}_\mathcal{P}(\mathbf{A})\mathbf{P}_\mathcal{Q}(\mathbf{B})|||_1 \neq |||\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{A}^\mathsf{H}\mathbf{B})|||_1$ for general unitary $\mathbf{A}, \mathbf{B}$. This, however, does not constitute a problem as whenever we apply uncertainty relations in $(\mathbb{C}^m, \|\cdot\|_1)$, the case of general unitary $\mathbf{A}, \mathbf{B}$ can always be reduced directly to $\mathbf{P}_\mathcal{P}(\mathbf{I}) = \mathbf{D}_\mathcal{P}$ and $\mathbf{P}_\mathcal{Q}(\mathbf{A}^\mathsf{H}\mathbf{B})$, simply by rewriting $\mathbf{Ap} = \mathbf{Bq}$ according to $\mathbf{p} = \mathbf{A}^\mathsf{H}\mathbf{Bq}$.

An uncertainty relation in $(\mathbb{C}^m, \|\cdot\|_1)$ is an upper bound of the form $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \leq c$ with $c \geq 0$ and states that $\|\mathbf{x}_{\mathcal{P}}\|_1 \leq c\|\mathbf{x}\|_1$ for all $\mathbf{x} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$. $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U})$ hence quantifies how well a vector supported on $\mathcal{Q}$ in the basis $\mathbf{U}$ can be concentrated on $\mathcal{P}$, where now concentration is measured in terms of 1-norm. Again, an uncertainty relation in $(\mathbb{C}^m, \|\cdot\|_1)$ is nontrivial only if $c < 1$. Application of Lemma 2.22 yields

$$\frac{1}{m}\|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\|_1 \leq \Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \leq \|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\|_1, \tag{2.81}$$

which constitutes the 1-norm equivalent of (2.2).

2.3.1 Coherence-based Uncertainty Relation

We next derive a coherence-based uncertainty relation for $(\mathbb{C}^m, \|\cdot\|_1)$, which comes with the same advantages and disadvantages as its 2-norm counterpart.

Lemma 2.10. *Let* $\mathbf{U} \in \mathbb{C}^{m\times m}$ *be a unitary matrix and* $\mathcal{P}, \mathcal{Q} \subseteq \{1, \dots, m\}$. *Then,*

$$\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \leq |\mathcal{P}||\mathcal{Q}|\mu^2([\mathbf{I}\ \ \mathbf{U}]). \tag{2.82}$$

Proof. Let $\tilde{\mathbf{u}}_i$ denote the column vectors of $\mathbf{U}^{\mathsf{H}}$. It follows from Lemma 2.22 that

$$\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) = \max_{j\in\{1,\dots,m\}} \|\mathbf{D}_{\mathcal{P}}\mathbf{U}\mathbf{D}_{\mathcal{Q}}\tilde{\mathbf{u}}_j\|_1. \tag{2.83}$$

With

$$\begin{aligned} \max_{j\in\{1,\dots,m\}} \|\mathbf{D}_{\mathcal{P}}\mathbf{U}\mathbf{D}_{\mathcal{Q}}\tilde{\mathbf{u}}_j\|_1 &\leq |\mathcal{P}| \max_{i,j\in\{1,\dots,m\}} |\tilde{\mathbf{u}}_i^{\mathsf{H}}\mathbf{D}_{\mathcal{Q}}\tilde{\mathbf{u}}_j| & (2.84)\\ &\leq |\mathcal{P}||\mathcal{Q}| \max_{i,j,k\in\{1,\dots,m\}} |\mathbf{U}_{i,k}||\mathbf{U}_{j,k}| & (2.85)\\ &\leq |\mathcal{P}||\mathcal{Q}|\mu^2([\mathbf{I}\ \ \mathbf{U}]), & (2.86)\end{aligned}$$

this establishes the proof. □

For $\mathcal{P} = \{1\}$, $\mathcal{Q} = \{1, \dots, m\}$, and $\mathbf{U} = \mathbf{F}$, the upper bounds on $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{F})$ in (2.81) and (2.82) coincide and equal 1. We next present an example where (2.82) is sharper than (2.81). Let m be even, $\mathcal{P} = \{m\}$, $\mathcal{Q} = \{1, \dots, m/2\}$, and $\mathbf{U} = \mathbf{F}$. Then, (2.82) becomes $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \le 1/2$, whereas

$$\|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{F})\|_1 = \frac{1}{m}\sum_{l=1}^{m}\left|\sum_{k=1}^{m/2} e^{\frac{2\pi i l k}{m}}\right| \tag{2.87}$$

$$= \frac{1}{2} + \frac{1}{m}\sum_{l=1}^{m-1}\left|\frac{1-e^{\pi i l}}{1-e^{\frac{2\pi i l}{m}}}\right| \tag{2.88}$$

$$= \frac{1}{2} + \frac{2}{m}\sum_{l=1}^{m/2}\frac{1}{\left|1-e^{\frac{2\pi i(2l-1)}{m}}\right|} \tag{2.89}$$

$$= \frac{1}{2} + \frac{1}{m}\sum_{l=1}^{m/2}\frac{1}{\sin\!\left(\frac{\pi(2l-1)}{m}\right)}. \tag{2.90}$$

Applying Jensen's inequality [49, Theorem 2.6.2] to (2.90) and using $\sum_{l=1}^{\frac{m}{2}}(2l-1) = (m/2)^2$ then yields $\|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{F})\|_1 \ge 1$, which shows that (2.81) is trivial.

For $\mathcal{P}$ and $\mathcal{Q}$ as in (2.11) and (2.12), respectively, (2.82) becomes $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{F}) \le n^2/m$, which for fixed ratio n/m increases linearly in m and becomes trivial for $m \ge (m/n)^2$. A more sophisticated uncertainty relation based on a large sieve inequality exists for strictly band-limited (infinite) ℓ_1-sequences [33, Theorem 14]; a corresponding finite-dimensional result does not seem to be available.

2.3.2 Concentration Inequalities

Analogously to Section 2.2.3, we next ask how well concentrated a given signal vector can be in two different orthonormal bases. Here we, however, consider a different measure of concentration accounting for the fact that we deal with the 1-norm.

Definition 2.11. Let $\mathcal{P} \subseteq \{1, \dots, m\}$ and $\varepsilon_{\mathcal{P}} \in [0,1]$. The vector $\mathbf{x} \in \mathbb{C}^m$ is said to be $\varepsilon_{\mathcal{P}}$-concentrated if $\|\mathbf{x} - \mathbf{x}_{\mathcal{P}}\|_1 \le \varepsilon_{\mathcal{P}}\|\mathbf{x}\|_1$.

The fraction of 1-norm an $\varepsilon_{\mathcal{P}}$-concentrated vector exhibits outside $\mathcal{P}$ is therefore no more than $\varepsilon_{\mathcal{P}}$. In particular, if $\mathbf{x}$ is $\varepsilon_{\mathcal{P}}$-concentrated for $\varepsilon_{\mathcal{P}} = 0$, then $\mathbf{x} = \mathbf{x}_{\mathcal{P}}$ and $\mathbf{x}$ is $|\mathcal{P}|$-sparse. The zero vector is trivially $\varepsilon_{\mathcal{P}}$-concentrated for all $\mathcal{P} \subseteq \{1, \dots, m\}$ and $\varepsilon_{\mathcal{P}} \in [0,1]$. In the remainder of Section 2.3, concentration is with respect to the 1-norm according to Definition 2.11.

We are now ready to state the announced result on the concentration of a vector in two different orthonormal bases.

Lemma 2.12. *Let $\mathbf{A}, \mathbf{B} \in \mathbb{C}^{m\times m}$ be unitary and $\mathcal{P}, \mathcal{Q} \subseteq \{1, \dots, m\}$. Suppose that there exist a nonzero $\varepsilon_{\mathcal{P}}$-concentrated $\mathbf{p} \in \mathbb{C}^m$ and a nonzero $\mathbf{q} \in \mathbb{C}^m$ with $\mathbf{q} = \mathbf{q}_{\mathcal{Q}}$ such that $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$. Then,*

$$|\mathcal{P}||\mathcal{Q}| \ge \frac{1-\varepsilon_{\mathcal{P}}}{\mu^2([\mathbf{A}\ \ \mathbf{B}])}. \tag{2.91}$$

Proof. Rewriting $\mathbf{Ap} = \mathbf{Bq}$ according to $\mathbf{p} = \mathbf{A}^{\mathsf{H}}\mathbf{Bq}$, it follows that $\mathbf{p} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$ with $\mathbf{U} = \mathbf{A}^{\mathsf{H}}\mathbf{B}$. We have

$$1 - \varepsilon_{\mathcal{P}} \leq \frac{\|\mathbf{p}_{\mathcal{P}}\|_1}{\|\mathbf{p}\|_1} \tag{2.92}$$

$$\leq \Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) \tag{2.93}$$

$$\leq |\mathcal{P}||\mathcal{Q}|\mu^2([\mathbf{I}\ \ \mathbf{U}]), \tag{2.94}$$

where (2.92) is by $\varepsilon_{\mathcal{P}}$-concentration of $\mathbf{p}$, (2.93) follows from (2.80) and $\mathbf{p} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$, and in (2.94) we applied Lemma 2.10. The proof is concluded by noting that $\mu([\mathbf{I}\ \ \mathbf{U}]) = \mu([\mathbf{A}\ \ \mathbf{B}])$.

□

For $\varepsilon_{\mathcal{P}} = 0$, Lemma 2.12 recovers Corollary 2.8.

2.3.3 Noisy Recovery in $(\mathbb{C}^m, \|\cdot\|_1)$

We next consider a noisy signal recovery problem akin to that in Section 2.2.4. Specifically, we investigate recovery—through 1-norm minimization—of a sparse signal corrupted by $\varepsilon_{\mathcal{P}}$-concentrated noise.

Lemma 2.13. *Let*

$$\mathbf{y} = \mathbf{p} + \mathbf{n}, \tag{2.95}$$

where $\mathbf{n} \in \mathbb{C}^m$ is $\varepsilon_{\mathcal{P}}$-concentrated to $\mathcal{P} \subseteq \{1, \dots, m\}$ and $\mathbf{p} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$ for $\mathbf{U} \in \mathbb{C}^{m\times m}$ unitary and $\mathcal{Q} \subseteq \{1, \dots, m\}$. Denote

$$\mathbf{z} = \underset{\mathbf{w}\in\mathcal{W}^{\mathbf{U},\mathcal{Q}}}{\operatorname{argmin}}(\|\mathbf{y} - \mathbf{w}\|_1). \tag{2.96}$$

If $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1/2$, then $\|\mathbf{z} - \mathbf{p}\|_1 \leq C\varepsilon_{\mathcal{P}}\|\mathbf{n}\|_1$ with $C = 2/(1 - 2\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}))$. In particular,

$$|\mathcal{P}||\mathcal{Q}| < \frac{1}{2\mu^2([\mathbf{I}\ \ \mathbf{U}])} \tag{2.97}$$

is sufficient for $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1/2$.

Proof. Set $\mathcal{P}^c = \{1, \dots, m\}\backslash\mathcal{P}$ and let $\mathbf{q} = \mathbf{U}^{\mathsf{H}}\mathbf{p}$. Note that $\mathbf{q}_{\mathcal{Q}} = \mathbf{q}$ as a consequence of $\mathbf{p} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$, which is by assumption. We have

$$\|\mathbf{n}\|_1 = \|\mathbf{y} - \mathbf{p}\|_1 \tag{2.98}$$

$$\geq \|\mathbf{y} - \mathbf{z}\|_1 \tag{2.99}$$

$$= \|\mathbf{n} - \tilde{\mathbf{z}}\|_1 \tag{2.100}$$

$$= \|(\mathbf{n} - \tilde{\mathbf{z}})_{\mathcal{P}}\|_1 + \|(\mathbf{n} - \tilde{\mathbf{z}})_{\mathcal{P}^c}\|_1 \tag{2.101}$$

$$\geq \|\mathbf{n}_{\mathcal{P}}\|_1 - \|\mathbf{n}_{\mathcal{P}^c}\|_1 + \|\tilde{\mathbf{z}}_{\mathcal{P}^c}\|_1 - \|\tilde{\mathbf{z}}_{\mathcal{P}}\|_1 \tag{2.102}$$

$$= \|\mathbf{n}\|_1 - 2\|\mathbf{n}_{\mathcal{P}^c}\|_1 + \|\tilde{\mathbf{z}}\|_1 - 2\|\tilde{\mathbf{z}}_{\mathcal{P}}\|_1 \tag{2.103}$$

$$\geq \|\mathbf{n}\|_1(1 - 2\varepsilon_{\mathcal{P}}) + \|\tilde{\mathbf{z}}\|_1(1 - 2\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U})), \tag{2.104}$$

where in (2.100) we set $\tilde{\mathbf{z}} = \mathbf{z} - \mathbf{p}$, in (2.102) we applied the reverse triangle inequality, and in (2.104) we used that $\mathbf{n}$ is $\varepsilon_{\mathcal{P}}$-concentrated and $\tilde{\mathbf{z}} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$, owing to $\mathbf{z} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$ and $\mathbf{p} \in \mathcal{W}^{\mathbf{U},\mathcal{Q}}$, together with (2.80). This yields

$$\begin{aligned}\|\mathbf{z} - \mathbf{p}\|_1 &= \|\tilde{\mathbf{z}}\|_1 && (2.105)\\ &\leq \frac{2\varepsilon_{\mathcal{P}}}{1 - 2\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U})}\|\mathbf{n}\|_1. && (2.106)\end{aligned}$$

Finally, (2.97) implies $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1/2$ thanks to (2.82). □

Note that for $\varepsilon_{\mathcal{P}} = 0$, i.e., the noise vector is supported on $\mathcal{P}$, we can recover $\mathbf{p}$ from $\mathbf{y} = \mathbf{p} + \mathbf{n}$ perfectly provided that $\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{U}) < 1/2$. For the special case $\mathbf{U} = \mathbf{F}$, this is guaranteed by

$$|\mathcal{P}||\mathcal{Q}| < \frac{m}{2}, \tag{2.107}$$

and perfect recovery of $\mathbf{p}$ from $\mathbf{y} = \mathbf{p} + \mathbf{n}$ amounts to the finite-dimensional version of what is known as Logan's phenomenon [32, Section 6.2].

2.3.4 Coherence-based Uncertainty Relation for Pairs of General Matrices

In practice, one is often interested in sparse signal representations with respect to general (i.e., possibly redundant or incomplete) dictionaries. The purpose of this section is to provide a corresponding general uncertainty relation. Specifically, we consider representations of a given signal vector s according to $\mathbf{s} = \mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$, where $\mathbf{A} \in \mathbb{C}^{m\times p}$ and $\mathbf{B} \in \mathbb{C}^{m\times q}$ are general matrices, $\mathbf{p} \in \mathbb{C}^p$, and $\mathbf{q} \in \mathbb{C}^q$. We start by introducing the notion of mutual coherence for pairs of matrices.

Definition 2.14. For $\mathbf{A} = (\mathbf{a}_1 \dots \mathbf{a}_p) \in \mathbb{C}^{m\times p}$ and $\mathbf{B} = (\mathbf{b}_1 \dots \mathbf{b}_q) \in \mathbb{C}^{m\times q}$, both with columns $\|\cdot\|_2$-normalized to 1, the mutual coherence $\bar{\mu}(\mathbf{A},\mathbf{B})$ is defined as $\bar{\mu}(\mathbf{A},\mathbf{B}) = \max_{i,j} |\mathbf{a}_i^{\mathsf{H}}\mathbf{b}_j|$.

The general uncertainty relation we are now ready to state is in terms of a pair of upper bounds on $\|\mathbf{p}_{\mathcal{P}}\|_1$ and $\|\mathbf{q}_{\mathcal{Q}}\|_1$ for $\mathcal{P} \subseteq \{1,\dots,p\}$ and $\mathcal{Q} \subseteq \{1,\dots,q\}$.

Theorem 2.15. *Let $\mathbf{A} \in \mathbb{C}^{m\times p}$ and $\mathbf{B} \in \mathbb{C}^{m\times q}$, both with column vectors $\|\cdot\|_2$-normalized to 1, and consider $\mathbf{p} \in \mathbb{C}^p$ and $\mathbf{q} \in \mathbb{C}^q$. Suppose that $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$. Then, we have*

$$\|\mathbf{p}_{\mathcal{P}}\|_1 \leq |\mathcal{P}|\left(\frac{\mu(\mathbf{A})\|\mathbf{p}\|_1 + \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{q}\|_1}{1 + \mu(\mathbf{A})}\right) \tag{2.108}$$

for all $\mathcal{P} \subseteq \{1,\dots,p\}$ and, by symmetry,

$$\|\mathbf{q}_{\mathcal{Q}}\|_1 \leq |\mathcal{Q}|\left(\frac{\mu(\mathbf{B})\|\mathbf{q}\|_1 + \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{p}\|_1}{1 + \mu(\mathbf{B})}\right) \tag{2.109}$$

for all $\mathcal{Q} \subseteq \{1,\dots,q\}$.

Proof. Since (2.109) follows from (2.108) simply by replacing $\mathbf{A}$ by $\mathbf{B}$, $\mathbf{p}$ by $\mathbf{q}$, $\mathcal{P}$ by $\mathcal{Q}$, and noting that $\bar{\mu}(\mathbf{A},\mathbf{B}) = \bar{\mu}(\mathbf{B},\mathbf{A})$, it suffices to prove (2.108). Let $\mathcal{P} \subseteq \{1,\dots,p\}$ and consider an arbitrary but fixed $i \in \{1,\dots,p\}$. Multiplying $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$ from the left by $\mathbf{a}_i^{\mathsf{H}}$ and taking absolute values results in

$$|\mathbf{a}_i^{\mathsf{H}}\mathbf{A}\mathbf{p}| = |\mathbf{a}_i^{\mathsf{H}}\mathbf{B}\mathbf{q}|. \tag{2.110}$$

The left-hand side of (2.110) can be lower-bounded according to

$$\begin{aligned}
|\mathbf{a}_i^{\mathsf{H}}\mathbf{A}\mathbf{p}| &= \Bigg|p_i + \sum_{\substack{k=1\\k\neq i}}^{p} \mathbf{a}_i^{\mathsf{H}}\mathbf{a}_k p_k\Bigg| && (2.111)\\
&\geq |p_i| - \Bigg|\sum_{\substack{k=1\\k\neq i}}^{p} \mathbf{a}_i^{\mathsf{H}}\mathbf{a}_k p_k\Bigg| && (2.112)\\
&\geq |p_i| - \sum_{\substack{k=1\\k\neq i}}^{p} |\mathbf{a}_i^{\mathsf{H}}\mathbf{a}_k||p_k| && (2.113)\\
&\geq |p_i| - \mu(\mathbf{A})\sum_{\substack{k=1\\k\neq i}}^{p} |p_k| && (2.114)\\
&= (1+\mu(\mathbf{A}))|p_i| - \mu(\mathbf{A})\|\mathbf{p}\|_1, && (2.115)
\end{aligned}$$

where (2.112) is by the reverse triangle inequality and in (2.114) we used Definition 2.3. Next, we upper-bound the right-hand side of (2.110) according to

$$\begin{aligned}
|\mathbf{a}_i^{\mathsf{H}}\mathbf{B}\mathbf{q}| &= \Bigg|\sum_{k=1}^{q} \mathbf{a}_i^{\mathsf{H}}\mathbf{b}_k q_k\Bigg| && (2.116)\\
&\leq \sum_{k=1}^{q} |\mathbf{a}_i^{\mathsf{H}}\mathbf{b}_k||q_k| && (2.117)\\
&\leq \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{q}\|_1, && (2.118)
\end{aligned}$$

where the last step is by Definition 2.14. Combining the lower bound (2.111)–(2.115) and the upper bound (2.116)–(2.118) yields

$$(1+\mu(\mathbf{A}))|p_i| - \mu(\mathbf{A})\|\mathbf{p}\|_1 \leq \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{q}\|_1. \tag{2.119}$$

Since (2.119) holds for arbitrary $i \in \{1,\dots,p\}$, we can sum over all $i \in \mathcal{P}$ and get

$$\|\mathbf{p}_{\mathcal{P}}\|_1 \leq |\mathcal{P}|\left(\frac{\mu(\mathbf{A})\|\mathbf{p}\|_1 + \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{q}\|_1}{1+\mu(\mathbf{A})}\right). \tag{2.120}$$

□

For the special case $\mathbf{A} = \mathbf{I} \in \mathbb{C}^{m \times m}$ and $\mathbf{B} \in \mathbb{C}^{m \times m}$ with $\mathbf{B}$ unitary, we have $\mu(\mathbf{A}) = \mu(\mathbf{B}) = 0$ and $\bar{\mu}(\mathbf{I}, \mathbf{B}) = \mu([\mathbf{I}\ \ \mathbf{B}])$, so that (2.108) and (2.109) simplify to

$$\|\mathbf{p}_{\mathcal{P}}\|_1 \leq |\mathcal{P}|\, \mu([\mathbf{I}\ \ \mathbf{B}])\, \|\mathbf{q}\|_1 \tag{2.121}$$

and

$$\|\mathbf{q}_{\mathcal{Q}}\|_1 \leq |\mathcal{Q}|\, \mu([\mathbf{I}\ \ \mathbf{B}])\, \|\mathbf{p}\|_1, \tag{2.122}$$

respectively. Thus, for arbitrary but fixed $\mathbf{p} \in \mathcal{W}^{\mathbf{B}, \mathcal{Q}}$ and $\mathbf{q} = \mathbf{B}^{\mathsf{H}}\mathbf{p}$, we have $\mathbf{q}_{\mathcal{Q}} = \mathbf{q}$ so that (2.121) and (2.122) taken together yield

$$\|\mathbf{p}_{\mathcal{P}}\|_1 \leq |\mathcal{P}||\mathcal{Q}|\, \mu^2([\mathbf{I}\ \ \mathbf{B}])\, \|\mathbf{p}\|_1. \tag{2.123}$$

As $\mathbf{p}$ was assumed to be arbitrary, by (2.80) this recovers the uncertainty relation

$$\Sigma_{\mathcal{P},\mathcal{Q}}(\mathbf{B}) \leq |\mathcal{P}||\mathcal{Q}|\mu^2([\mathbf{I}\ \ \mathbf{B}]) \tag{2.124}$$

in Lemma 2.10.

2.3.5 Concentration Inequalities for Pairs of General Matrices

We next refine the result in Theorem 2.15 to vectors that are concentrated in 1-norm according to Definition 2.11. The formal statement is as follows.

Corollary 2.16. *Let $\mathbf{A} \in \mathbb{C}^{m \times p}$ and $\mathbf{B} \in \mathbb{C}^{m \times q}$, both with column vectors $\|\cdot\|_2$-normalized to 1, $\mathcal{P} \subseteq \{1, \ldots, p\}$, $\mathcal{Q} \subseteq \{1, \ldots, q\}$, $\mathbf{p} \in \mathbb{C}^p$, and $\mathbf{q} \in \mathbb{C}^q$. Suppose that $\mathbf{Ap} = \mathbf{Bq}$. Then, the following statements hold.*

1. *If $\mathbf{q}$ is $\varepsilon_{\mathcal{Q}}$-concentrated, then,*

$$\|\mathbf{p}_{\mathcal{P}}\|_1 \leq \frac{|\mathcal{P}|}{1 + \mu(\mathbf{A})} \left(\mu(\mathbf{A}) + \frac{\bar{\mu}^2(\mathbf{A}, \mathbf{B})|\mathcal{Q}|}{[(1 + \mu(\mathbf{B}))(1 - \varepsilon_{\mathcal{Q}}) - \mu(\mathbf{B})|\mathcal{Q}|]_+} \right) \|\mathbf{p}\|_1. \tag{2.125}$$

2. *If $\mathbf{p}$ is $\varepsilon_{\mathcal{P}}$-concentrated, then,*

$$\|\mathbf{q}_{\mathcal{Q}}\|_1 \leq \frac{|\mathcal{Q}|}{1 + \mu(\mathbf{B})} \left(\mu(\mathbf{B}) + \frac{\bar{\mu}^2(\mathbf{A}, \mathbf{B})|\mathcal{P}|}{[(1 + \mu(\mathbf{A}))(1 - \varepsilon_{\mathcal{P}}) - \mu(\mathbf{A})|\mathcal{P}|]_+} \right) \|\mathbf{q}\|_1. \tag{2.126}$$

3. *If $\mathbf{p}$ is $\varepsilon_{\mathcal{P}}$-concentrated, $\mathbf{q}$ is $\varepsilon_{\mathcal{Q}}$-concentrated, $\bar{\mu}(\mathbf{A}, \mathbf{B}) > 0$, and $(\mathbf{p}^{\mathsf{T}}\ \mathbf{q}^{\mathsf{T}})^{\mathsf{T}} \neq \mathbf{0}$, then,*

$$|\mathcal{P}||\mathcal{Q}| \geq \frac{[(1 + \mu(\mathbf{A}))(1 - \varepsilon_{\mathcal{P}}) - \mu(\mathbf{A})|\mathcal{P}|]_+ [(1 + \mu(\mathbf{B}))(1 - \varepsilon_{\mathcal{Q}}) - \mu(\mathbf{B})|\mathcal{Q}|]_+}{\bar{\mu}^2(\mathbf{A}, \mathbf{B})}. \tag{2.127}$$

Proof. By Theorem 2.15, we have

$$\|\mathbf{p}_{\mathcal{P}}\|_1 \leq |\mathcal{P}|\left(\frac{\mu(\mathbf{A})\|\mathbf{p}\|_1 + \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{q}\|_1}{1+\mu(\mathbf{A})}\right) \tag{2.128}$$

and

$$\|\mathbf{q}_{\mathcal{Q}}\|_1 \leq |\mathcal{Q}|\left(\frac{\mu(\mathbf{B})\|\mathbf{q}\|_1 + \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{p}\|_1}{1+\mu(\mathbf{B})}\right). \tag{2.129}$$

Suppose now that $\mathbf{q}$ is $\varepsilon_{\mathcal{Q}}$-concentrated, i.e., $\|\mathbf{q}_{\mathcal{Q}}\|_1 \geq (1-\varepsilon_{\mathcal{Q}})\|\mathbf{q}\|_1$. Then, (2.129) implies that

$$\|\mathbf{q}\|_1 \leq \frac{|\mathcal{Q}|\bar{\mu}(\mathbf{A},\mathbf{B})}{[(1+\mu(\mathbf{B}))(1-\varepsilon_{\mathcal{Q}}) - \mu(\mathbf{B})|\mathcal{Q}|]_+}\|\mathbf{p}\|_1. \tag{2.130}$$

Using (2.130) in (2.128) yields (2.125). The relation (2.126) follows from (2.125) by swapping the roles of $\mathbf{A}$ and $\mathbf{B}$, $\mathbf{p}$ and $\mathbf{q}$, and $\mathcal{P}$ and $\mathcal{Q}$, and upon noting that $\bar{\mu}(\mathbf{A},\mathbf{B}) = \bar{\mu}(\mathbf{B},\mathbf{A})$. It remains to establish (2.127). Using $\|\mathbf{p}_{\mathcal{P}}\|_1 \geq (1-\varepsilon_{\mathcal{P}})\|\mathbf{p}\|_1$ in (2.128) and $\|\mathbf{q}_{\mathcal{Q}}\|_1 \geq (1-\varepsilon_{\mathcal{Q}})\|\mathbf{q}\|_1$ in (2.129) yields

$$\|\mathbf{p}\|_1[(1+\mu(\mathbf{A}))(1-\varepsilon_{\mathcal{P}}) - \mu(\mathbf{A})|\mathcal{P}|]_+ \leq \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{q}\|_1|\mathcal{P}| \tag{2.131}$$

and

$$\|\mathbf{q}\|_1[(1+\mu(\mathbf{B}))(1-\varepsilon_{\mathcal{Q}}) - \mu(\mathbf{B})|\mathcal{Q}|]_+ \leq \bar{\mu}(\mathbf{A},\mathbf{B})\|\mathbf{p}\|_1|\mathcal{Q}|, \tag{2.132}$$

respectively. Suppose first that $\mathbf{p} = \mathbf{0}$. Then, $\mathbf{q} \neq \mathbf{0}$ by assumption, and (2.132) becomes

$$[(1+\mu(\mathbf{B}))(1-\varepsilon_{\mathcal{Q}}) - \mu(\mathbf{B})|\mathcal{Q}|]_+ = 0. \tag{2.133}$$

In this case (2.127) holds trivially. Similarly, if $\mathbf{q} = \mathbf{0}$, then $\mathbf{p} \neq \mathbf{0}$ again by assumption, and (2.131) becomes

$$[(1+\mu(\mathbf{A}))(1-\varepsilon_{\mathcal{P}}) - \mu(\mathbf{A})|\mathcal{P}|]_+ = 0. \tag{2.134}$$

As before, (2.127) holds trivially. Finally, if $\mathbf{p} \neq \mathbf{0}$ and $\mathbf{q} \neq \mathbf{0}$, then we multiply (2.131) by (2.132) and divide the result by $\bar{\mu}^2(\mathbf{A},\mathbf{B})\|\mathbf{p}\|_1\|\mathbf{q}\|_1$ which yields (2.127). □

The lower bound on $|\mathcal{P}||\mathcal{Q}|$ in (2.127) is [35, Theorem 1] and states that a nonzero vector can not be arbitrarily well concentrated with respect to two different general matrices $\mathbf{A}$ and $\mathbf{B}$. For the special case $\varepsilon_{\mathcal{Q}} = 0$ and $\mathbf{A}$ and $\mathbf{B}$ unitary, and hence $\mu(\mathbf{A}) = \mu(\mathbf{B}) = 0$ and $\bar{\mu}(\mathbf{A},\mathbf{B}) = \mu([\mathbf{A}\ \ \mathbf{B}])$, (2.127) recovers Lemma 2.12.

Particularizing (2.127) to $\varepsilon_{\mathcal{P}} = \varepsilon_{\mathcal{Q}} = 0$ yields the following result.

Corollary 2.17. *[36, Lemma 33] Let* $\mathbf{A} \in \mathbb{C}^{m\times p}$ *and* $\mathbf{B} \in \mathbb{C}^{m\times q}$*, both with column vectors* $\|\cdot\|_2$*-normalized to* 1*, and consider* $\mathbf{p} \in \mathbb{C}^p$ *and* $\mathbf{q} \in \mathbb{C}^q$ *with* $(\mathbf{p}^\mathsf{T}\ \mathbf{q}^\mathsf{T})^\mathsf{T} \neq \mathbf{0}$*. Suppose that* $\mathbf{Ap} = \mathbf{Bq}$*. Then,* $\|\mathbf{p}\|_0\|\mathbf{q}\|_0 \geq f_{\mathbf{A},\mathbf{B}}(\|\mathbf{p}\|_0, \|\mathbf{q}\|_0)$*, where*

$$f_{\mathbf{A},\mathbf{B}}(u,v) = \frac{[1+\mu(\mathbf{A})(1-u)]_+[1+\mu(\mathbf{B})(1-v)]_+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})}. \tag{2.135}$$

Proof. Let $\mathcal{P} = \{i \in \{1,\dots,p\} : p_i \neq 0\}$ and $\mathcal{Q} = \{i \in \{1,\dots,q\} : q_i \neq 0\}$, so that $\mathbf{p}_\mathcal{P} = \mathbf{p}$, $\mathbf{q}_\mathcal{Q} = \mathbf{q}$, $|\mathcal{P}| = \|\mathbf{p}\|_0$, and $|\mathcal{Q}| = \|\mathbf{q}\|_0$. The claim now follows directly from (2.127) with $\varepsilon_\mathcal{P} = \varepsilon_\mathcal{Q} = 0$. □

If $\mathbf{A}$ and $\mathbf{B}$ are both unitary, then $\mu(\mathbf{A}) = \mu(\mathbf{B}) = 0$ and $\bar{\mu}(\mathbf{A},\mathbf{B}) = \mu([\mathbf{A}\ \ \mathbf{B}])$, and Corollary 2.17 recovers the Elad-Bruckstein result in Corollary 2.8.

Corollary 2.17 admits the following appealing geometric interpretation in terms of a null-space property.

Lemma 2.18. *Let $\mathbf{A} \in \mathbb{C}^{m\times p}$ and $\mathbf{B} \in \mathbb{C}^{m\times q}$, both with column vectors $\|\cdot\|_2$-normalized to 1. Then, the set (which actually is a finite union of subspaces)*

$$\mathcal{S} = \left\{ \begin{pmatrix} \mathbf{p} \\ \mathbf{q} \end{pmatrix} : \mathbf{p} \in \mathbb{C}^p,\ \mathbf{q} \in \mathbb{C}^q,\ \|\mathbf{p}\|_0\|\mathbf{q}\|_0 < f_{\mathbf{A},\mathbf{B}}(\|\mathbf{p}\|_0, \|\mathbf{q}\|_0) \right\} \tag{2.136}$$

with $f_{\mathbf{A},\mathbf{B}}$ defined in (2.135) intersects the kernel of $[\mathbf{A}\ \ \mathbf{B}]$ trivially, i.e.,

$$\ker([\mathbf{A}\ \ \mathbf{B}]) \cap \mathcal{S} = \{\mathbf{0}\}. \tag{2.137}$$

Proof. The statement of this lemma is equivalent to the statement of Corollary 2.17 through a chain of equivalences between the following statements:

1. $\ker([\mathbf{A}\ \ \mathbf{B}]) \cap \mathcal{S} = \{\mathbf{0}\}$;
2. if $(\mathbf{p}^\mathsf{T}\ -\mathbf{q}^\mathsf{T})^\mathsf{T} \in \ker([\mathbf{A}\ \ \mathbf{B}])\backslash\{\mathbf{0}\}$, then $\|\mathbf{p}\|_0\|\mathbf{q}\|_0 \geq f_{\mathbf{A},\mathbf{B}}(\|\mathbf{p}\|_0, \|\mathbf{q}\|_0)$;
3. if $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$ with $(\mathbf{p}^\mathsf{T}\ \mathbf{q}^\mathsf{T})^\mathsf{T} \neq \mathbf{0}$, then $\|\mathbf{p}\|_0\|\mathbf{q}\|_0 \geq f_{\mathbf{A},\mathbf{B}}(\|\mathbf{p}\|_0, \|\mathbf{q}\|_0)$,

where 1 ⇔ 2 is by definition of $\mathcal{S}$, 2 ⇔ 3 follows from the fact that $\mathbf{A}\mathbf{p} = \mathbf{B}\mathbf{q}$ with $(\mathbf{p}^\mathsf{T}\ \mathbf{q}^\mathsf{T})^\mathsf{T} \neq \mathbf{0}$ is equivalent to $(\mathbf{p}^\mathsf{T}\ \ -\mathbf{q}^\mathsf{T})^\mathsf{T} \in \ker([\mathbf{A}\ \ \mathbf{B}])\backslash\{\mathbf{0}\}$, and 3 is the statement in Corollary 2.17. □

2.4 A Large Sieve Inequality in $(\mathbb{C}^m, \|\cdot\|_2)$

We present a slightly improved and generalized version of the large sieve inequality stated in [33, Equation (32)].

Lemma 2.19. *Let μ be a 1-periodic, σ-finite measure on $\mathbb{R}$, $n \in \mathbb{N}$, $\varphi \in [0,1)$, $\mathbf{a} \in \mathbb{C}^n$, and consider the 1-periodic trigonometric polynomial*

$$\psi(s) = e^{\mathrm{i}2\pi\varphi} \sum_{k=1}^{n} a_k e^{-2\pi \mathrm{i} k s}. \tag{2.138}$$

Then,

$$\int_{[0,1)} |\psi(s)|^2 \mathrm{d}\mu(s) \leq \left(n - 1 + \frac{1}{\delta}\right) \sup_{r\in[0,1)} \mu((r, r+\delta))\|\mathbf{a}\|_2^2 \tag{2.139}$$

for all $\delta \in (0,1]$.

Proof. Since

$$|\psi(s)| = \left|\sum_{k=1}^{n} a_k e^{-2\pi \mathrm{i} k s}\right|, \tag{2.140}$$

we can assume, without loss of generality, that $\varphi = 0$. The proof now follows closely the line of argumentation in [50, pp. 185–186] and in the proof of [33, Lemma 5]. Specifically, we make use of the result in [50, p. 185] saying that, for every $\delta > 0$, there exists a function $g \in L^2(\mathbb{R})$ with Fourier transform

$$G(s) = \int_{-\infty}^{\infty} g(t) e^{-2\pi \mathrm{i} s t} \mathrm{d}t \tag{2.141}$$

such that $\|G\|_2^2 = n - 1 + 1/\delta$, $|g(t)|^2 \geq 1$ for all $t \in [1, n]$, and $G(s) = 0$ for all $s \notin [-\delta/2, \delta/2]$. With this g, consider the 1-periodic trigonometric polynomial

$$\theta(s) = \sum_{k=1}^{n} \frac{a_k}{g(k)} e^{-2\pi \mathrm{i} k s} \tag{2.142}$$

and note that

$$\int_{-\delta/2}^{\delta/2} G(r)\theta(s-r)\mathrm{d}r = \sum_{k=1}^{n} \frac{a_k}{g(k)} e^{-2\pi \mathrm{i} k s} \int_{-\infty}^{\infty} G(r) e^{2\pi \mathrm{i} k r} \mathrm{d}r \tag{2.143}$$

$$= \sum_{k=1}^{n} a_k e^{-2\pi \mathrm{i} k s} \tag{2.144}$$

$$= \psi(s) \quad \text{for all } s \in \mathbb{R}. \tag{2.145}$$

We now have

$$\int_{[0,1)} |\psi(s)|^2 \mathrm{d}\mu(s) = \int_{[0,1)} \left| \int_{-\delta/2}^{\delta/2} G(r)\theta(s-r)\mathrm{d}r \right|^2 \mathrm{d}\mu(s) \tag{2.146}$$

$$\le \|G\|_2^2 \int_{[0,1)} \left(\int_{-\delta/2}^{\delta/2} |\theta(s-r)|^2 \mathrm{d}r \right) \mathrm{d}\mu(s) \tag{2.147}$$

$$= \|G\|_2^2 \int_{[0,1)} \left(\int_{s-\delta/2}^{s+\delta/2} |\theta(r)|^2 \mathrm{d}r \right) \mathrm{d}\mu(s) \tag{2.148}$$

$$= \|G\|_2^2 \int_{-1}^{2} \mu\big((r-\delta/2, r+\delta/2) \cap [0,1)\big) |\theta(r)|^2 \mathrm{d}r \tag{2.149}$$

$$= \|G\|_2^2 \sum_{i=-1}^{1} \int_{0+i}^{1+i} \mu\big((r-\delta/2, r+\delta/2) \cap [0,1)\big) |\theta(r)|^2 \mathrm{d}r \tag{2.150}$$

$$= \|G\|_2^2 \sum_{i=-1}^{1} \int_{0}^{1} \mu\big((r-\delta/2, r+\delta/2) \cap [i,1+i)\big) |\theta(r)|^2 \mathrm{d}r \tag{2.151}$$

$$= \|G\|_2^2 \int_{0}^{1} \mu\big((r-\delta/2, r+\delta/2) \cap [-1,2)\big) |\theta(r)|^2 \mathrm{d}r \tag{2.152}$$

$$= \|G\|_2^2 \int_{0}^{1} \mu\big((r-\delta/2, r+\delta/2)\big) |\theta(r)|^2 \mathrm{d}r \tag{2.153}$$

for all $\delta \in (0,1]$, where (2.146) follows from (2.143)–(2.145), in (2.147) we applied the Cauchy-Schwartz inequality [51, Theorem 1.37], (2.149) is by Fubini's theorem [52, Theorem 1.14] (recall that μ is σ-finite by assumption) upon noting that

$$\{(r,s) : s \in [0,1), r \in (s-\delta/2, s+\delta/2)\} \tag{2.154}$$

$$= \{(r,s) : r \in [-1,2), s \in (r-\delta/2, r+\delta/2) \cap [0,1)\} \tag{2.155}$$

for all $\delta \in (0,1]$, in (2.151) we used the 1-periodicity of μ and θ, and (2.152) is by σ-additivity of μ. Now,

$$\int_0^1 \mu\big((r-\delta/2, r+\delta/2)\big) |\theta(r)|^2 \mathrm{d}r \le \sup_{r\in[0,1)} \mu((r, r+\delta)) \int_0^1 |\theta(r)|^2 \mathrm{d}r \tag{2.156}$$

$$= \sup_{r\in[0,1)} \mu((r, r+\delta)) \sum_{k=1}^{n} \frac{|a_k|^2}{|g(k)|^2} \tag{2.157}$$

$$\le \sup_{r\in[0,1)} \mu((r, r+\delta)) \|\mathbf{a}\|_2^2 \tag{2.158}$$

for all $\delta > 0$, where (2.158) follows from $|g(t)|^2 \ge 1$ for all $t \in [1,n]$. Using (2.156)–(2.158) and $\|G\|_2^2 = n - 1 + 1/\delta$ in (2.153) establishes (2.139). □

Lemma 2.19 is a slightly strengthened version of the large sieve inequality [33, Equation (32)]. Specifically, in (2.139) it is sufficient to consider open intervals $(r, r+\delta)$, whereas [33, Equation

(32)] requires closed intervals $[r, r+\delta]$. Thus, the upper bound in [33, Equation (32)] can be strictly larger than that in (2.139) whenever μ has mass points.

2.5 Uncertainty Relations in L_1 and L_2

The following table contains a list of infinite-dimensional counterparts—available in the literature—to results in this chapter. Specifically, these results apply to band-limited L_1- and L_2-functions and correspond to $\mathbf{A} = \mathbf{I}$ and $\mathbf{B} = \mathbf{F}$ in our setting.

	L_2 analog	L_1 analog
Upper bound in (2.10)	[32, Lemma 2]	
Corollary 2.7	[32, Theorem 2]	
Lemma 2.9	[32, Theorem 4]	
Lemma 2.10		[32, Lemma 3]
Lemma 2.13		[33, Lemma 2]
Lemma 2.19		[33, Theorem 4]

2.6 Results for $|||\cdot|||_1$ and $|||\cdot|||_2$

Lemma 2.20. *Let $\mathbf{U} \in \mathbb{C}^{m\times m}$ be unitary, $\mathcal{P}, \mathcal{Q} \subseteq \{1, \dots, m\}$, and consider the orthogonal projection $\mathbf{P}_\mathcal{Q}(\mathbf{U}) = \mathbf{U}\mathbf{D}_\mathcal{Q}\mathbf{U}^\mathsf{H}$ onto the subspace $\mathcal{W}^{\mathbf{U},\mathcal{Q}}$. Then,*

$$|||\mathbf{P}_\mathcal{Q}(\mathbf{U})\mathbf{D}_\mathcal{P}|||_2 = |||\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U})|||_2. \tag{2.159}$$

Moreover, we have

$$|||\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U})|||_2 = \max_{\mathbf{x}\in\mathcal{W}^{\mathbf{U},\mathcal{Q}}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{x}_\mathcal{P}\|_2}{\|\mathbf{x}\|_2} \tag{2.160}$$

and

$$|||\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U})|||_1 = \max_{\mathbf{x}\in\mathcal{W}^{\mathbf{U},\mathcal{Q}}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{x}_\mathcal{P}\|_1}{\|\mathbf{x}\|_1}. \tag{2.161}$$

Proof. The identity (2.159) follows from

$$\begin{aligned} |||\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U})|||_2 &= |||(\mathbf{D}_\mathcal{P}\mathbf{P}_\mathcal{Q}(\mathbf{U}))^*|||_2 && (2.162)\\ &= |||\mathbf{P}^*_\mathcal{Q}(\mathbf{U})\mathbf{D}^*_\mathcal{P}|||_2 && (2.163)\\ &= |||\mathbf{P}_\mathcal{Q}(\mathbf{U})\mathbf{D}_\mathcal{P}|||_2, && (2.164)\end{aligned}$$

where in (2.162) we used that $|||\cdot|||_2$ is self-adjoint [48, p. 309], $\mathbf{P}_{\mathcal{Q}}(\mathbf{U})^* = \mathbf{P}_{\mathcal{Q}}(\mathbf{U})$, and $\mathbf{D}_{\mathcal{P}}^* = \mathbf{D}_{\mathcal{P}}$. To establish (2.160), we note that

$$\begin{aligned}
|||\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})|||_2 &= \max_{\mathbf{x}\colon \|\mathbf{x}\|_2=1} \|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\|_2 && (2.165)\\
&= \max_{\substack{\mathbf{x}\colon \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\neq\mathbf{0}\\ \|\mathbf{x}\|_2=1}} \|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\|_2 && (2.166)\\
&\leq \max_{\substack{\mathbf{x}\colon \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\neq\mathbf{0}\\ \|\mathbf{x}\|_2=1}} \left\|\mathbf{D}_{\mathcal{P}}\frac{\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}}{\|\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\|_2}\right\|_2 && (2.167)\\
&\leq \max_{\mathbf{x}\colon \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\neq\mathbf{0}} \left\|\mathbf{D}_{\mathcal{P}}\frac{\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}}{\|\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\|_2}\right\|_2 && (2.168)\\
&= \max_{\mathbf{x}\in\mathcal{W}^{\mathbf{U},\mathcal{Q}}\setminus\{\mathbf{0}\}} \frac{\|\mathbf{x}_{\mathcal{P}}\|_2}{\|\mathbf{x}\|_2} && (2.169)\\
&= \max_{\mathbf{x}\colon \mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\neq\mathbf{0}} \left\|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\frac{\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}}{\|\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\|_2}\right\|_2 && (2.170)\\
&\leq \max_{\mathbf{x}\colon \|\mathbf{x}\|_2=1} \|\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\|_2 && (2.171)\\
&= |||\mathbf{D}_{\mathcal{P}}\mathbf{P}_{\mathcal{Q}}(\mathbf{U})|||_2, && (2.172)
\end{aligned}$$

where in (2.167) we used $\|\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\|_2 \leq \|\mathbf{x}\|_2$, which implies $\|\mathbf{P}_{\mathcal{Q}}(\mathbf{U})\mathbf{x}\|_2 \leq 1$ for all $\mathbf{x}$ with $\|\mathbf{x}\|_2 = 1$. Finally, (2.161) follows by repeating the steps in (2.165)–(2.172) with $\|\cdot\|_2$ replaced by $\|\cdot\|_1$ at all occurrences. □

Lemma 2.21. *Let* $\mathbf{A} \in \mathbb{C}^{m\times n}$. *Then,*

$$\frac{\|\mathbf{A}\|_2}{\sqrt{\operatorname{rank}(\mathbf{A})}} \leq |||\mathbf{A}|||_2 \leq \|\mathbf{A}\|_2. \tag{2.173}$$

Proof. The proof is trivial for $\mathbf{A} = \mathbf{0}$. If $\mathbf{A} \neq \mathbf{0}$, set $r = \operatorname{rank}(\mathbf{A})$ and let $\sigma_1, \dots, \sigma_r$ denote the nonzero singular values of $\mathbf{A}$ organized in decreasing order. Unitary invariance of $|||\cdot|||_2$ and $\|\cdot\|_2$ (cf. [48, Problem 5, p. 311]) yields $|||\mathbf{A}|||_2 = \sigma_1$ and $\|\mathbf{A}\|_2 = \sqrt{\sum_{i=1}^r \sigma_i^2}$. The claim now follows from

$$\sigma_1 \leq \sqrt{\sum_{i=1}^r \sigma_i^2} \leq \sqrt{r}\sigma_1. \tag{2.174}$$

□

Lemma 2.22. *For* $\mathbf{A} = (\mathbf{a}_1 \dots \mathbf{a}_n) \in \mathbb{C}^{m\times n}$*, we have*

$$|||\mathbf{A}|||_1 = \max_{j\in\{1,\dots,n\}} \|\mathbf{a}_j\|_1 \tag{2.175}$$

and

$$\frac{1}{n}\|\mathbf{A}\|_1 \leq |||\mathbf{A}|||_1 \leq \|\mathbf{A}\|_1. \tag{2.176}$$

Proof. The identity (2.175) is established in [48, p.294], and (2.176) follows directly from (2.175). □

Chapter 3

Compressive Sensing

3.1 Discrete Fourier Transform

Signals that we process in practice have finite length. We thus have

$$\hat{x}(\theta) = \sum_{n=-\infty}^{\infty} x[n]e^{-2\pi i\theta n} = \sum_{n=0}^{N-1} x[n]e^{-2\pi i\theta n}.$$

Do we really need to know $\hat{x}(\theta)$ for $\theta \in [0,1)$ to specify the finite length signal $x[n]$? Since $x[n]$ has length N, N samples of $\hat{x}(\theta)$ should suffice to uniquely specify $x[n]$. We sample $\hat{x}(\theta)$ uniformly, i.e., we compute

$$\begin{aligned} \widehat{x}[k] &:= \frac{1}{\sqrt{N}}\widehat{x}\left(\frac{k}{N}\right) \\ &= \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} x[n]e^{-2\pi \mathrm{i}kn/N} \\ &= \frac{1}{\sqrt{N}}\sum_{n=0}^{N-1} x[n]\omega_N^{kn}, \quad \text{for } k = 0,1,\dots,N-1, \end{aligned}$$

where $\omega_N = e^{-2\pi\mathrm{i}/N}$. It holds that $\widehat{x}[k+N] = \widehat{x}[k]$, for all $k \in \mathbb{Z}$. Now we write this relationship in the vector-matrix form:

$$\underbrace{\begin{bmatrix} \widehat{x}[0] \\ \widehat{x}[1] \\ \vdots \\ \widehat{x}[N-1] \end{bmatrix}}_{=:\,\widehat{\mathbf{x}}} = \underbrace{\frac{1}{\sqrt{N}}\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_N & \omega_N^2 & \cdots & \omega_N^{N-1} \\ 1 & \omega_N^2 & \omega_N^4 & \cdots & \omega_N^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \cdots & \omega_N^{(N-1)^2} \end{bmatrix}}_{\mathbf{F}_N} \underbrace{\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}}_{=:\,\mathbf{x}}.$$

The matrix $\mathbf{F}_N$ is the $N \times N$ DFT-Matrix and is unitary, i.e., $\mathbf{F}_N \mathbf{F}_N^H = \mathbf{F}_N^H \mathbf{F}_N = \mathbf{I}_N$, where $\mathbf{I}_N$ denotes the $N \times N$ identity matrix. We can thus directly specify how the signal $\mathbf{x}$ of length N can be recovered from $\hat{\mathbf{x}}$. From $\widehat{\mathbf{x}} = \mathbf{F}_N \mathbf{x}$ it follows after multiplication with $\mathbf{F}_N^H$ from left that

$$\mathbf{F}_N^H \widehat{\mathbf{x}} = \mathbf{F}_N^H \mathbf{F}_N \mathbf{x} = \mathbf{I}_N \mathbf{x} = \mathbf{x} \quad \Rightarrow \quad \mathbf{x} = \mathbf{F}_N^H \widehat{\mathbf{x}}$$

and

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \widehat{x}[k] e^{2\pi \mathrm{i} nk/N}.$$

We thus have the following transformation pair for the discrete Fourier transform:

$$\begin{aligned} \widehat{x}[k] &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-2\pi \mathrm{i} kn/N}, \quad \text{with } \widehat{x}[k+N] = \widehat{x}[k] \\ x[n] &= \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \widehat{x}[k] e^{2\pi \mathrm{i} nk/N}, \quad \text{with } x[n+N] = x[n]. \end{aligned}$$

In vector matrix form this transformation pair can be written as

$$\begin{aligned} \widehat{\mathbf{x}} &= \mathbf{F}_N \mathbf{x} \\ \mathbf{x} &= \mathbf{F}_N^H \widehat{\mathbf{x}}. \end{aligned}$$

With $\mathbf{F}_N^H = \begin{bmatrix} \mathbf{f}_0 & \mathbf{f}_1 & \cdots & \mathbf{f}_{N-1} \end{bmatrix}$ we obtain

$$\mathbf{x} = \mathbf{F}_N^H \mathbf{F}_N \mathbf{x} = \begin{bmatrix} \mathbf{f}_0 & \mathbf{f}_1 & \cdots & \mathbf{f}_{N-1} \end{bmatrix} \begin{bmatrix} \mathbf{f}_0^H \\ \mathbf{f}_1^H \\ \vdots \\ \mathbf{f}_{N-1}^H \end{bmatrix} \mathbf{x} = \sum_{\ell=0}^{N-1} \langle \mathbf{x}, \mathbf{f}_\ell \rangle \mathbf{f}_\ell.$$

Thus we have shown that the DFT is nothing else than the expansion of the vector $\mathbf{x}$ into an orthonormal basis (ONB) for $\mathbb{C}^N$.

3.1.1 Oversampling

$$\begin{aligned} \widehat{x}[k] &= \frac{1}{\sqrt{M}} \widehat{x}\left(\frac{k}{M}\right) \\ &= \frac{1}{\sqrt{M}} \sum_{n=0}^{N-1} x[n] e^{-2\pi \mathrm{i} kn/M}, \quad k = 0, 1, \ldots, M-1, \text{ with } M > N. \end{aligned}$$

In vector-matrix form we obtain

$$\underbrace{\begin{bmatrix} \widehat{x}(0) \\ \widehat{x}[1/M] \\ \vdots \\ \widehat{x}[(M-1)/M] \end{bmatrix}}_{\widehat{\mathbf{x}}} = \underbrace{\frac{1}{\sqrt{M}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_M & \omega_M^2 & \cdots & \omega_M^{N-1} \\ 1 & \omega_M^2 & \omega_M^4 & \cdots & \omega_M^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_M^{M-1} & \omega_M^{2(M-1)} & \cdots & \omega_M^{(N-1)(M-1)} \end{bmatrix}}_{\mathbf{F}_{\mathrm{o}} \in \mathbb{C}^{M\times N} \square} \underbrace{\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}}_{\mathbf{x}}$$

The columns of F_{o} are orthonormal since these are the first N columns of the DFT matrix $\mathbf{F}_M$. This implies that we can recover $\mathbf{x}$ from $\widehat{\mathbf{x}}$, but there are infinitely many inverses. A selected inverse is the Moore-Penrose pseudo-inverse given by $(\mathbf{F}_{\mathrm{o}}^H\mathbf{F}_{\mathrm{o}})^{-1}\mathbf{F}_{\mathrm{o}}^H$. First, we observe that

$$(\mathbf{F}_{\mathrm{o}}^H\mathbf{F}_{\mathrm{o}})^{-1}\mathbf{F}_{\mathrm{o}}^H\widehat{\mathbf{x}} = \underbrace{(\mathbf{F}_{\mathrm{o}}^H\mathbf{F}_{\mathrm{o}})^{-1}\mathbf{F}_{\mathrm{o}}^H\mathbf{F}_{\mathrm{o}}}_{\mathbf{I}_N}\mathbf{x} = \mathbf{x}. \tag{3.1}$$

Because of the orthonormality of the columns of $\mathbf{F}_{\mathrm{o}}$, we have $\mathbf{F}_{\mathrm{o}}^H\mathbf{F}_{\mathrm{o}} = \mathbf{I}_N$ so that

$$\mathbf{x} = \mathbf{F}_{\mathrm{o}}^H\widehat{\mathbf{x}}. \tag{3.2}$$

It is noteworthy that, despite oversampling, the inverse transform (corresponding to the pseudo-inverse) is given by

$$x[n] = \frac{1}{\sqrt{M}}\sum_{k=0}^{M-1}\widehat{x}[k]\omega_M^{-kn}.$$

3.1.2 Undersampling

$$\begin{aligned} \widehat{x}[k] &= \frac{1}{\sqrt{M}}\widehat{x}\left(\frac{k}{M}\right) \\ &= \frac{1}{\sqrt{M}}\sum_{n=0}^{N-1}x[n]e^{-2\pi\mathrm{i}kn/M}, \quad k = 0, 1, \ldots, M-1, \text{ with } M < N. \end{aligned}$$

In vector-matrix form we obtain

$$\underbrace{\begin{bmatrix} \widehat{x}(0) \\ \widehat{x}(1/M) \\ \vdots \\ \widehat{x}((M-1)/M) \end{bmatrix}}_{\widehat{\mathbf{x}}} = \underbrace{\frac{1}{\sqrt{M}} \begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & \omega_M & \omega_M^2 & \cdots & \omega_M^{N-1} \\ 1 & \omega_M^2 & \omega_M^4 & \cdots & \omega_M^{2(N-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & \omega_M^{M-1} & \omega_M^{2(M-1)} & \cdots & \omega_M^{(N-1)(M-1)} \end{bmatrix}}_{\mathbf{F}_{\mathrm{u}} \in \mathbb{C}^{M\times N} \square} \underbrace{\begin{bmatrix} x[0] \\ x[1] \\ \vdots \\ x[N-1] \end{bmatrix}}_{\mathbf{x}}$$

This system of equations is not invertible because we have $M < N$ equations and N unknowns. The case $M = N$ corresponds to critical sampling.

3.2 Compressive Sensing

We start by considering undersampling in the finite-dimensional case. Given the measurement vector $\mathbf{y}$ of dimension $\dim \mathbf{y} = m \times 1$ obtained when compressing the signal $\mathbf{x}$ of dimension $\dim \mathbf{x} = n \times 1$ by means of the matrix $\mathbf{D}$ of dimension $\dim \mathbf{D} = m \times n$ according to

$$\mathbf{y} = \mathbf{D}\mathbf{x},$$

we would like to recover the original vector $\mathbf{x}$. Clearly, this problem has infinitely many solutions when the set of signals is not constrained. We consider the case where $\mathbf{x}$ is s-sparse, i.e., $\|\mathbf{x}\|_0 \leq s$ or, in words, the signal vector $\mathbf{x}$ has at most s nonzero entries.

Where do such problems arise?
Whenever we are given fewer observations than unknowns; this situation is, e.g., important in the following two cases

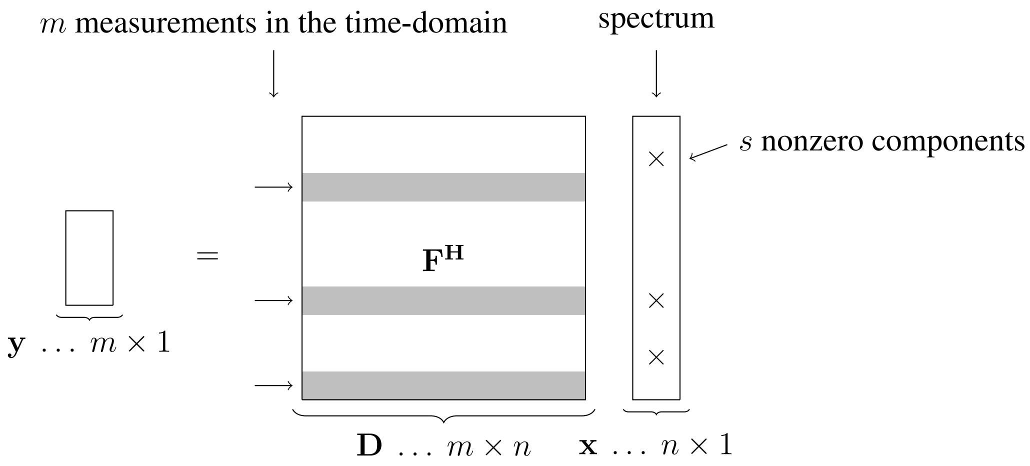
- signal recovery from partial (incomplete) measurements
- difficult and/or costly data acquisition

Specific examples of practical interest are

- geophysics
- medical imaging
- earth observations
- A/D conversion

Another important application is finding sparse representations of signals in a given (highly redundant) dictionary $\mathbf{D}$.

We go back to the case of spectrally sparse signals, this time in the finite-dimensional case.


m measurements in the time-domain
spectrum
s nonzero components
=
$\mathbf{F^H}$
$\mathbf{y}\ \ldots\ m \times 1$
$\mathbf{D}\ \ldots\ m \times n$
$\mathbf{x}\ \ldots\ n \times 1$

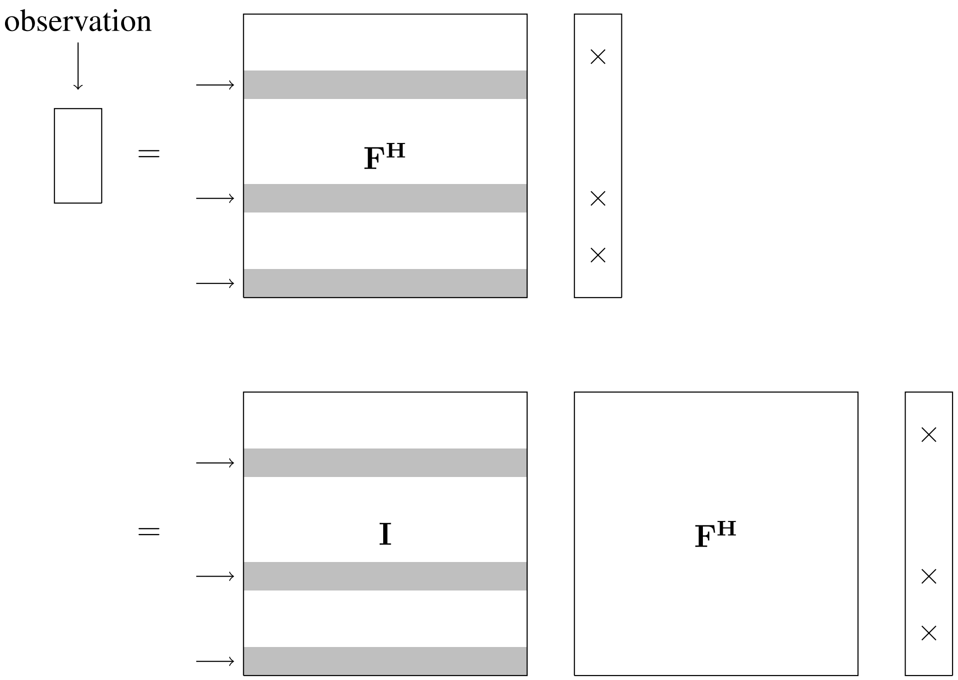
$m = s$: sampling at Landau rate
$m > s$: oversampling

The sampling instants (rows) must be chosen such that $\mathbf{D}$ is well-conditioned.

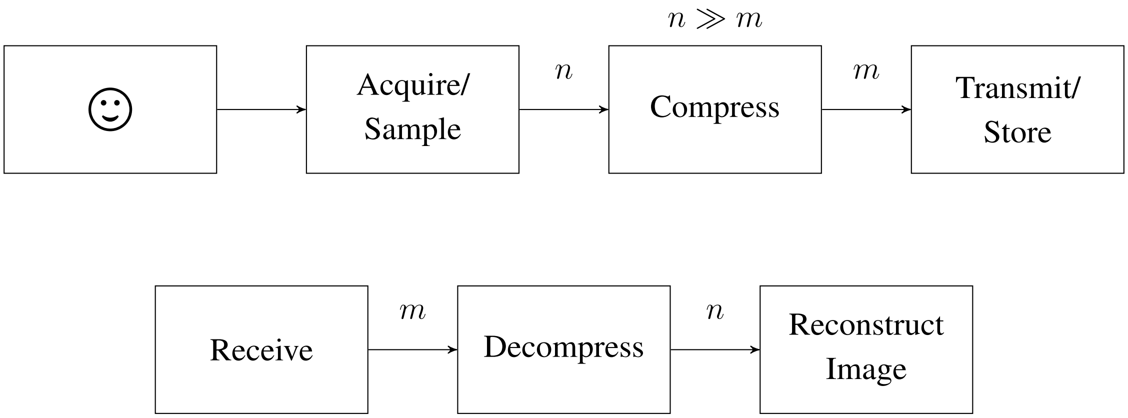
- If the support set of $\mathbf{x}$ is known, we can apply the following universal sampling pattern. Choose the first $m = s$ rows of $\mathbf{F}$. The resulting $\mathbf{D}$-matrix is Vandermonde and hence full-rank, irrespectively of the support set.

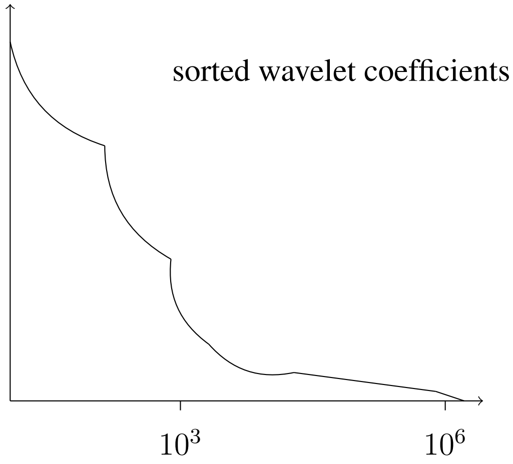
- If the support set of $\mathbf{x}$ is unknown, we can apply the following universal sampling pattern. Choose the first $m = 2s$ rows of $\mathbf{F}$. Any $\mathbf{x}_1 - \mathbf{x}_2$ with $\mathbf{x}_1, \mathbf{x}_2$ s-sparse and $\mathbf{x}_1 \neq \mathbf{x}_2$ satisfies $\|\mathbf{D}\,(\mathbf{x}_1 - \mathbf{x}_2)\|^2 > 0$ since $\mathbf{x}_1 - \mathbf{x}_2$ is $2s$-sparse and the resulting $\mathbf{D}$-matrix is Vandermonde and hence of rank $2s$.

Is there anything special about sparsity in the frequency-domain and sampling in the time-domain?

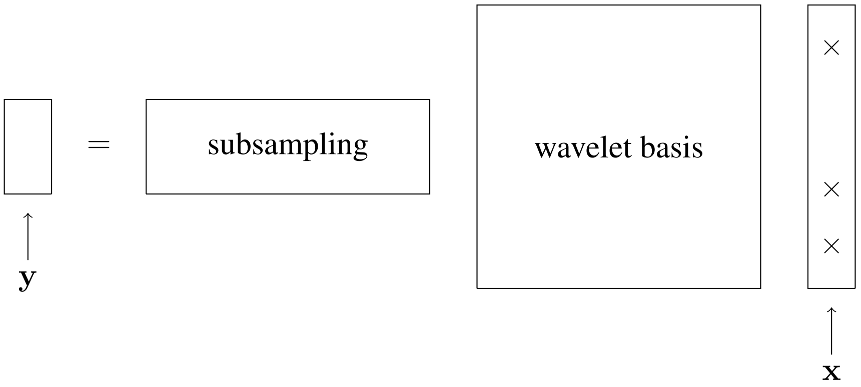


Can we do this for more general sparsity bases?
If yes, what would the required minimum sampling rate be?
Why would this be interesting?



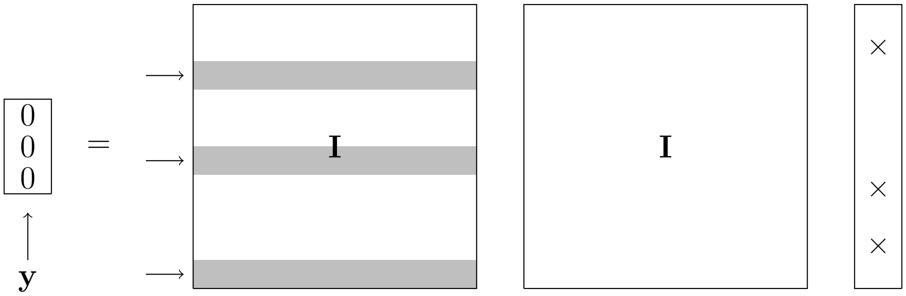


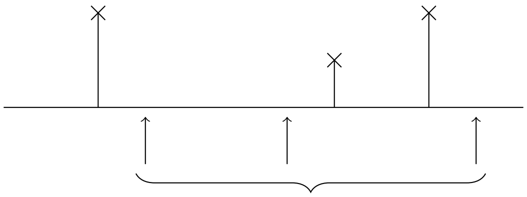
Typically, we obtain a graph similar to the figure above when plotting the amplitude of the sorted wavelet coefficients. Hence, $10^3 - 10^6$ of the costly acquired wavelet coefficients are thrown away in the process of compression. It is therefore important to ask whether we cannot just acquire the information that will not end up being thrown away. Analogously to the case where we subsample a spectrally sparse signal, we would like to reconstruct an image which is sparse in the wavelet domain according to a universal scheme given the measurement vector $\mathbf{y}$ depicted in the figure below.



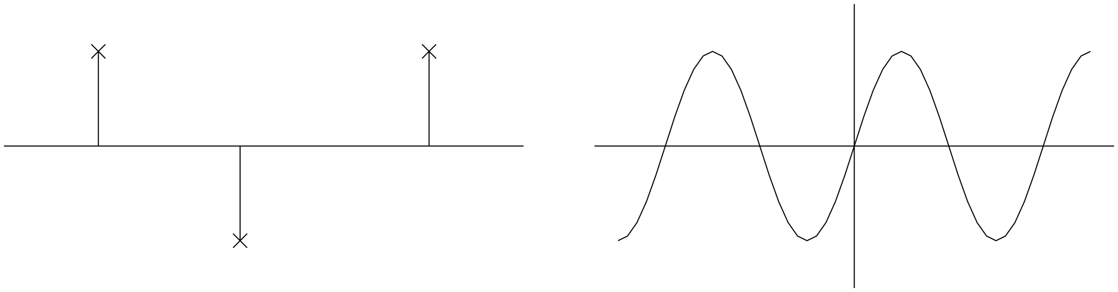
3.2.1 Incoherence

Since we require the compressive sensing scheme to be *universal*, recovery must be possible independently of the s-sparse signal vector $\mathbf{x}$. In the example depicted below, this is clearly not the case.

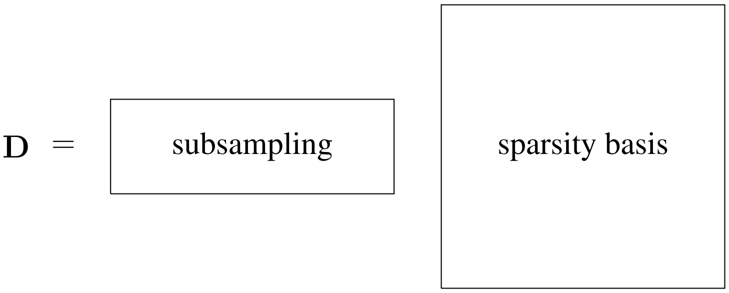




Spikes and sinusoids are, e.g., ’incoherent’.



3.2.2 The General Problem



For a given sparsity basis (e.g., wavelets), find a sampling basis such that s-sparse vectors are distinguishable, i.e., for all $\mathbf{x}_1, \mathbf{x}_2$ that are s-sparse with $\mathbf{x}_1 \neq \mathbf{x}_2$

$$\|\mathbf{D}\left(\mathbf{x}_1 - \mathbf{x}_2\right)\|_2^2 > 0.$$

Hence, all collections of $2s$ columns of $\mathbf{D}$ have to be linearly independent. Clearly, this is possible only if "we sample at least at twice the Landau rate", i.e., $m \geq 2s$.
In the following, we assume that every column of a dictionary $\mathbf{D}$ is normalized to unit 2-norm.

Definition 3.1. The spark of a matrix $\mathbf{A}$ denoted by $\operatorname{spark}(\mathbf{A})$ is defined as the cardinality of the smallest set of linearly dependent columns.

For a given matrix $\mathbf{D}$ of dimension $m \times n$, uniqueness of recovery of s-sparse vectors $\mathbf{x}$ from the observation $\mathbf{y} = \mathbf{D}\mathbf{x}$ is guaranteed for

$$s < \frac{\operatorname{spark}(\mathbf{D})}{2}.$$

3.3 The Recovery Problem (P0)

If $\mathbf{D}$ is a (known) ONB, recovering $\mathbf{x}$ from $\mathbf{y}$ is simply

$$\mathbf{D}^{\mathsf{H}}\mathbf{y} = \mathbf{D}^{\mathsf{H}}\mathbf{D}\mathbf{x} = \mathbf{x}.$$

If $\mathbf{D}$ is a (known) basis, we have

$$\mathbf{D}^{-1}\mathbf{y} = \mathbf{D}^{-1}\mathbf{D}\mathbf{x} = \mathbf{x}.$$

If $\mathbf{D} = [\mathbf{A} \;\; d]$ where $\mathbf{A}$ is an ONB and d is an extra column, then we cannot uniquely determine $\mathbf{x}$ from $\mathbf{y} = \mathbf{D}\mathbf{x}$.

However, if $\mathbf{x}$ is s-sparse and

$$s < \frac{\operatorname{spark}(\mathbf{D})}{2}$$

we can recover $\mathbf{x}$ through a combinatorial search:

$$\text{(P0)} \quad \text{find } \arg\min \|\hat{\mathbf{x}}\|_0 \text{ subject to } \mathbf{y} = \mathbf{D}\hat{\mathbf{x}}$$

For any vector $\mathbf{x}$, the quasi-norm $\|\mathbf{x}\|_0$ denotes the number of nonzero entries.

Suppose that $\|\mathbf{x}\|_0 \leq s$ and $s < \frac{\operatorname{spark}(\mathbf{D})}{2}$. Let $\tilde{\mathbf{x}} \neq \mathbf{x}$ with $\|\tilde{\mathbf{x}}\|_0 \leq s$ and $\mathbf{y} = \mathbf{D}\tilde{\mathbf{x}}$, then

$$0 = \mathbf{y} - \mathbf{y} = \mathbf{D}\mathbf{x} - \mathbf{D}\tilde{\mathbf{x}} = \mathbf{D}\underbrace{(\mathbf{x} - \tilde{\mathbf{x}})}_{\|\mathbf{x}-\tilde{\mathbf{x}}\|_0 \leq 2s}.$$

Since $2s < \operatorname{spark}(\mathbf{D})$, we know, however, that

$$\|\mathbf{D}(\mathbf{x} - \tilde{\mathbf{x}})\| > 0, \mathbf{x} - \tilde{\mathbf{x}} \neq 0$$

as any set of $2s$ columns of $\mathbf{D}$ is linearly independent. Therefore (P0) recovers $\mathbf{x}$ uniquely.

Determining the spark of a dictionary is a combinatorial problem and leads to huge computational complexity even for small problem size. Specifically, every set of a columns out of the $\binom{n}{a}$ possible sets has to be checked for linear independence and the parameter a has to be increased starting from two.

We next derive a lower bound on the spark in terms of the dictionary's coherence $\mu(\mathbf{D})$ defined according to (see Definition 2.3)

$$\mu(\mathbf{D}) = \max_{r \neq l} |\langle d_l, d_r \rangle|.$$

Theorem 3.2. [19]; [53] *(P0) applied to* $\mathbf{y} = \mathbf{D}\mathbf{x}$ *recovers* $\mathbf{x}$ *if*

$$\|\mathbf{x}\|_0 \leq s < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right).$$

Proof. We will show that $\operatorname{spark}(\mathbf{D}) \geq 1 + 1/\mu(\mathbf{D})$. Consider $\mathbf{x} \in \mathbb{C}^n$ with $\|\mathbf{x}\|_0 = \operatorname{spark}(\mathbf{D})$ and $\mathbf{D}\mathbf{x} = \mathbf{0}$. Then, we have

$$d_l \mathbf{x}_l = -\sum_{r \neq l} d_r \mathbf{x}_r, \quad \text{for all } l \in \{1, \ldots, n\}.$$

Left-multiplying both sides by d_l^{H} and using $\|d_l\|_2 = 1$ yields

$$\mathbf{x}_l = -\sum_{r \neq l} d_l^{\mathsf{H}} d_r \mathbf{x}_r,$$

which implies

$$|\mathbf{x}_l| = \left|\sum_{r \neq l} d_l^{\mathsf{H}} d_r \mathbf{x}_r\right| \leq \sum_{r \neq l} \left|d_l^{\mathsf{H}} d_r\right| |\mathbf{x}_r| \leq \mu(\mathbf{D}) \sum_{r \neq l} |\mathbf{x}_r| \quad \text{for } l \in \{1, \ldots, n\}.$$

Adding $\mu(\mathbf{D})|\mathbf{x}_l|$ on both sides results in

$$(1 + \mu(\mathbf{D}))\,|\mathbf{x}_l| \leq \mu(\mathbf{D})\|\mathbf{x}\|_1 \quad \text{for } l \in \{1, \dots, n\}.$$

Summing over all l for which $\mathbf{x}_l \neq 0$ finally leads to

$$\begin{aligned}&(1 + \mu(\mathbf{D}))\,\|\mathbf{x}\|_1 \leq \mu(\mathbf{D})\|\mathbf{x}\|_1 \operatorname{spark}(\mathbf{D})\\ &\Rightarrow \operatorname{spark}(\mathbf{D}) \geq 1 + \frac{1}{\mu(\mathbf{D})}.\end{aligned}$$

□

Notice that determining $\mu(\mathbf{D})$ has the complexity of doing the first step in the computation of $\operatorname{spark}(\mathbf{D})$, i.e., checking whether any two columns are linearly independent.

3.4 Basis Pursuit (BP)

In this section, we consider the recovery problem below

$$(\text{P1}) \quad \text{find } \arg\min \|\hat{\mathbf{x}}\|_1 \text{ subject to } \mathbf{y} = \mathbf{D}\hat{\mathbf{x}}$$

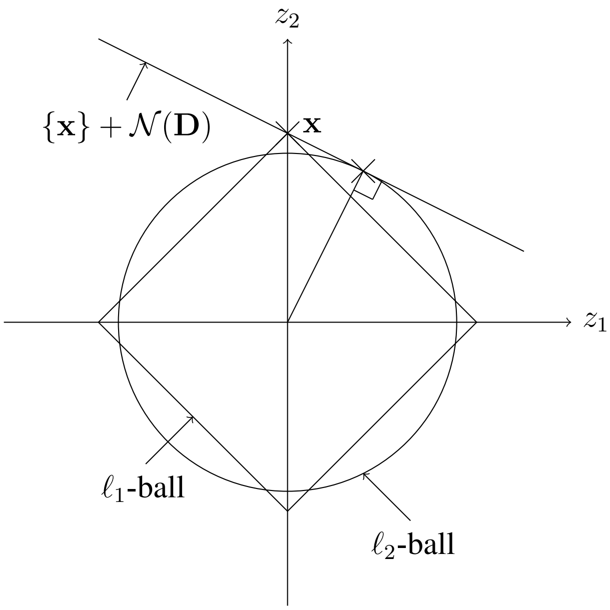
(P1) – often referred to as basis pursuit (BP) – can be cast as a linear program and is therefore more efficiently solvable than (P0) discussed in the previous section.

Early results on ℓ_1-reconstruction:

- Logan, 1965
- Donoho & Logan, 1992

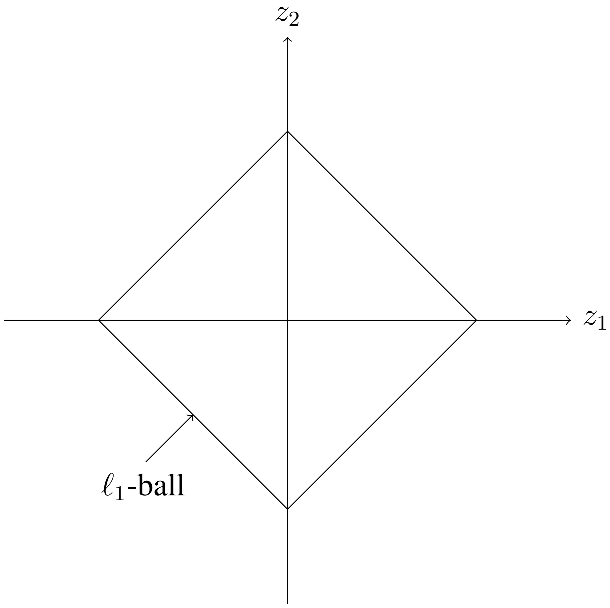
Why does ℓ_1-reconstruction work?

$$\begin{gathered}\arg\min \|\hat{\mathbf{x}}\|_1 \text{ subject to } \mathbf{y} = \mathbf{D}\hat{\mathbf{x}}\\ \Updownarrow\\ \arg\min \|\hat{\mathbf{x}}\|_1 \text{ subject to } \hat{\mathbf{x}} \in (\{\mathbf{x}\} + \mathcal{N}(\mathbf{D}))\end{gathered}$$

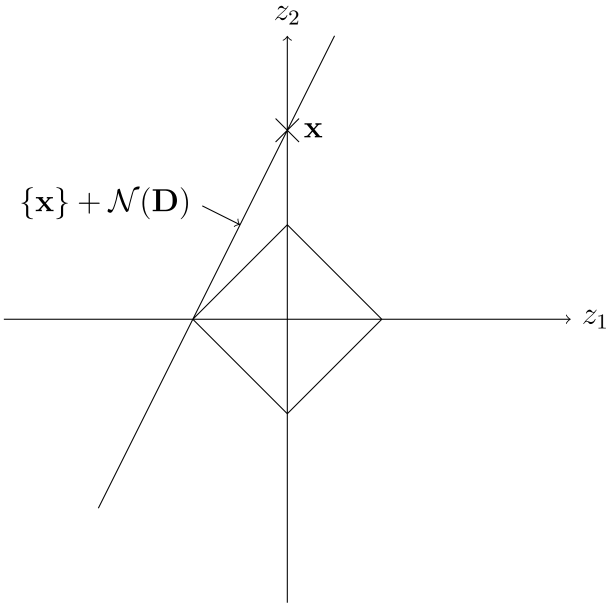


ℓ_1-ball: $|z_1| + |z_2| =$ const.

Case $z_1, z_2 > 0$: $z_1 + z_2 =$ const. $\Rightarrow z_2 =$ const. $- z_1$
By symmetry, the ℓ_1-ball must look as depicted below.



Clearly, (P1) cannot always recover the correct solutions, e.g., consider the scenario in the figure ahead.



Can we characterize analytically under which conditions $(\mathrm{P1})$ finds the correct solution?

We will need a result on the concentration of ℓ_1 norms.

Definition 3.3. We denote

$$P_1(\mathcal{S}, \mathbf{D}) \triangleq \max_{\mathbf{x} \in \mathcal{N}(\mathbf{D}), \mathbf{x} \neq 0} \frac{\sum_{k \in \mathcal{S}} |\mathbf{x}_k|}{\sum_k |\mathbf{x}_k|}.$$

Theorem 3.4. *Arbitrarily fix* $\mathbf{x}$ *with support set* $\mathcal{S}$ *and let* $\mathbf{y} = \mathbf{D}\mathbf{x}$. *If* $P_1(\mathcal{S}, \mathbf{D}) < 1/2$, *then* $\mathbf{x}$ *is the unique solution to*

$$(\textit{P1}) \quad \textit{find } \arg\min \|\hat{\mathbf{x}}\|_1 \textit{ subject to } \mathbf{y} = \mathbf{D}\hat{\mathbf{x}}$$

Proof. We need to prove that for all $\alpha \in \mathcal{N}(\mathbf{D})$

$$\sum_k |\mathbf{x}_k + \alpha_k| > \sum_k |\mathbf{x}_k|.$$

Application of the reverse triangle inequality

$$|a + b| \geq |a| - |b|$$

to the LHS of the above equation yields

$$\begin{aligned}\sum_k |\mathbf{x}_k + \alpha_k| &= \sum_{k \notin \mathcal{S}} |\mathbf{x}_k + \alpha_k| + \sum_{k \in \mathcal{S}} |\mathbf{x}_k + \alpha_k| \\ &= \sum_{k \notin \mathcal{S}} |\alpha_k| + \sum_{k \in \mathcal{S}} |\mathbf{x}_k + \alpha_k| \\ &\geq \sum_{k \notin \mathcal{S}} |\alpha_k| + \sum_{k \in \mathcal{S}} |\mathbf{x}_k| - \sum_{k \in \mathcal{S}} |\alpha_k|.\end{aligned}$$

Therefore, the theorem follows from

$$\sum_{k \notin \mathcal{S}} |\alpha_k| > \sum_{k \in \mathcal{S}} |\alpha_k| .$$

Adding $\sum_{k \in \mathcal{S}} |\alpha_k|$ to both sides of the above equation results in

$$\begin{aligned} \sum_{k} |\alpha_k| &> 2 \sum_{k \in \mathcal{S}} |\alpha_k| \\ \Rightarrow \underbrace{\frac{\sum_{k \in \mathcal{S}} |\alpha_k|}{\sum_k |\alpha_k|}}_{P_1(\mathcal{S},\mathbf{D})} &< \frac{1}{2} \end{aligned}$$

but this is satisfied for all $\alpha \in \mathcal{N}(\mathbf{D})$ since $P_1(\mathcal{S}, \mathbf{D}) < 1/2$ by assumption. □

Next, we find a sufficient condition for $P_1(\mathcal{S}, \mathbf{D}) < \frac{1}{2}$ in terms of the cardinality of the support set $|\mathcal{S}|$ and the dictionary coherence μ.

Consider $\alpha \in \mathcal{N}(\mathbf{D})$. Due to the proof of Theorem 3.2 we know that

$$(1 + \mu(\mathbf{D}))\, |\alpha_l| \leq \mu(\mathbf{D}) \|\alpha\|_1, \quad \text{for all } l = 1, \ldots, n.$$

Summing over all $l \in \mathcal{S}$, we get

$$\begin{aligned} (1 + \mu(\mathbf{D})) \sum_{l \in \mathcal{S}} |\alpha_l| &\leq \mu(\mathbf{D}) \|\alpha\|_1 |\mathcal{S}| \\ (1 + \mu(\mathbf{D}))\, \frac{\sum_{l \in \mathcal{S}} |\alpha_l|}{\sum_l |\alpha_l|} &\leq \mu(\mathbf{D}) |\mathcal{S}| \\ (1 + \mu(\mathbf{D}))\, P_1(\mathcal{S}, \mathbf{D}) &\leq \mu(\mathbf{D}) |\mathcal{S}| \\ P_1(\mathcal{S}, \mathbf{D}) &\leq \frac{1}{1 + 1/\mu(\mathbf{D})} |\mathcal{S}|. \end{aligned}$$

Therefore if $\frac{1}{1+1/\mu} |\mathcal{S}| < 1/2$, then $P_1(\mathcal{S}, \mathbf{D}) < 1/2$, i.e., the desired upper bound on the cardinality of the support set is given by

$$|\mathcal{S}| < \frac{1}{2} \left(1 + \frac{1}{\mu(\mathbf{D})} \right)$$

just as in the case of recovery via (P0).

We therefore proved

Theorem 3.5. [19]; [53] *(P1) applied to* $\mathbf{y} = \mathbf{D}\mathbf{x}$ *recovers* $\mathbf{x}$ *if*

$$\|\mathbf{x}\|_0 < \frac{1}{2} \left(1 + \frac{1}{\mu(\mathbf{D})} \right).$$

3.5 Signal Separation, Super-Resolution, Recovery of Corrupted Signals

Important signal recovery problems and applications

- signal separation
- super-resolution
- inpainting
- recovery of clipped signals
- recovery of signals subject to impulse noise
- recovery of signals subject to narrowband interference

3.5.1 Inpainting and Super-resolution

In the case of inpainting and super-resolution, we consider the following setting

- the signal $\mathbf{x}$ is sparse with unknown support set
- we observe $\mathbf{y} = \mathbf{A}\mathbf{x}$
- only a subset of the entries of $\mathbf{y} = \mathbf{A}\mathbf{x}$ is available
- inpainting and super-resolution amount to filling in the missing entries
- we account for the missing entries by taking the observation to be

  $$\mathbf{z} = \underbrace{\mathbf{A}\mathbf{x}}_{\mathbf{y}} + \mathbf{e} = \mathbf{A}\mathbf{x} + \mathbf{I}\mathbf{e}$$

  and choose $\mathbf{e}$ such that the entries of $\mathbf{z} = \mathbf{y} + \mathbf{e}$ corresponding to the missing entries in $\mathbf{y}$ are set to some arbitrary value, e.g., zero

If there are not too many entries missing or the area to be inpainted is not too big, $\mathbf{e}$ will be sparse.

Hence, we observe $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$ and know $\mathbf{A}, \mathbf{B}$ and that $\mathbf{x}, \mathbf{e}$ are sparse. Based on the observation $\mathbf{z}$, we want to recover $\mathbf{x}$ and $\mathbf{e}$.

3.5.2 Clipping

Instead of $\mathbf{y} = \mathbf{D}\mathbf{x}$ we observe $\mathbf{z} = g_a(\mathbf{y})$, where the function $g_a(\mathbf{y})$ realizes entry-wise signal clipping to the interval $[-a, a]$.

Clipping can equivalently be modeled as

$$\mathbf{z} = \mathbf{y} + \mathbf{e}$$

with $\mathbf{e} = g_a(\mathbf{y}) - \mathbf{y}$. Notice that the error locations can be determined by comparing the entries of $\mathbf{z}$ to the clipping threshold a.

Consequently, we observe

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \underbrace{\mathbf{B}}_{\mathbf{I}}\mathbf{e}$$

where $\mathbf{e}$ can depend on $\mathbf{A}$ and $\mathbf{x}$.

3.5.3 Signal Separation

We consider the superposition of $\mathbf{A}\mathbf{x}$ and $\mathbf{B}\mathbf{e}$, i.e., based on the observation

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$$

we want to recover the sparse signals $\mathbf{x}$ and $\mathbf{e}$.

3.5.4 Recovery of Signals Subject to Impulse Noise

In this scenario, a spectrally sparse signal with unknown spectrum is (sparsely) corrupted by impulses with unknown locations, i.e., we observe

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$$

where $\mathbf{A} = \mathbf{F}$ and $\mathbf{B} = \mathbf{I}$.

3.5.5 Recovery of Signals Subject to Narrowband Interference

We observe a sparse signal corrupted by spectrally sparse noise, i.e.,

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$$

where $\mathbf{A} = \mathbf{I}$ and $\mathbf{B} = \mathbf{F}$.

3.5.6 The General Problem

All signal recovery problems and applications discussed above can be embedded into a single model where we observe

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e} = \underbrace{[\mathbf{A}\ \ \mathbf{B}]}_{\mathbf{D}} \underbrace{\begin{bmatrix} \mathbf{x} \\ \mathbf{e} \end{bmatrix}}_{\check{\mathbf{x}}} = \mathbf{D}\check{\mathbf{x}}$$

and we want to recover $\check{\mathbf{x}}$ given $\mathbf{z} = \mathbf{D}\check{\mathbf{x}}$. To account for all possible scenarios, we cannot assume $\mathbf{e}$ to be independent of $\mathbf{A}$ and $\mathbf{x}$.

In particular, we are interested in dictionaries $\mathbf{D}$ that are the concatenation of two ONBs, e.g., $\mathbf{D} = [\mathbf{I}\ \ \mathbf{F}]$. This is useful if we want to sparsely represent signals that consist of two distinct features, e.g., spikes and sines.

3.5.7 Concatenation of ONBs (or Frames)

When a dictionary $\mathbf{D}$ is the concatenation of two ONBs $\mathbf{A}$ and $\mathbf{B}$, refined bounds on $\mathrm{spark}(\mathbf{D}) = \mathrm{spark}([\mathbf{A}\ \ \mathbf{B}])$ exist.

For a vector in the null-space of $\mathbf{D} = [\mathbf{A}\ \ \mathbf{B}]$, we have

$$[\mathbf{A}\ \ \mathbf{B}] \underbrace{\begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix}}_{\mathbf{v}} = 0.$$

Hence,

$$\mathbf{A}\mathbf{p} + \mathbf{B}\mathbf{q} = \mathbf{0} \Rightarrow \mathbf{A}\mathbf{p} = \mathbf{B}\,(-\mathbf{q}) \triangleq \mathbf{s}.$$

The signal $\mathbf{s}$ is represented in two different ways, namely in the dictionary $\mathbf{A}$ and in the dictionary $\mathbf{B}$.

Finding the vector $\mathbf{v}$ with minimum 0-norm among all vectors that satisfy

$$[\mathbf{A}\ \ \mathbf{B}]\mathbf{v} = \mathbf{0}$$

amounts to answering the question: How sparse can $\mathbf{p}$ and $\mathbf{q}$ concurrently be? The uncertainty relation in Corollary 2.8 states that $\mathbf{A}\mathbf{p} + \mathbf{B}\mathbf{q} = \mathbf{0}$ is only possible if $\|\mathbf{p}\|_0 \|\mathbf{q}\|_0 \geq 1/\mu^2([\mathbf{A}\ \ \mathbf{B}])$.

In the case where $\mathbf{D} = [\mathbf{A}\ \ \mathbf{B}]$ is a concatenation of two ONBs, the dictionary coherence evaluates to

$$\mu(\mathbf{D}) = \sup_{i \neq j} |\langle \mathbf{d}_i, \mathbf{d}_j \rangle| = \sup_{i,j} |\langle \mathbf{a}_i, \mathbf{b}_j \rangle|\,.$$

How can these uncertainty relations be used to obtain recovery thresholds?
By assumption, we have

$$\operatorname{spark}(\mathbf{D}) = \|\mathbf{p}\|_0 + \|\mathbf{q}\|_0 = \left\| \begin{bmatrix} \mathbf{p} \\ \mathbf{q} \end{bmatrix} \right\|_0 .$$

The arithmetic-mean – geometric-mean (AM-GM) inequality implies

$$\|\mathbf{p}\|_0 + \|\mathbf{q}\|_0 \geq 2\sqrt{\|\mathbf{p}\|_0 \|\mathbf{q}\|_0} \geq \frac{2}{\mu(\mathbf{D})}.$$

This yields a lower bound on the number of nonzero entries a vector in the null-space of $\mathbf{D} = [\mathbf{A}\ \ \mathbf{B}]$ can have and hence we get a lower bound on the spark.
Recall that the solution to (P0) is unique if the sparsity of the signal s satisfies

$$s < \frac{\operatorname{spark}(\mathbf{D})}{2}.$$

Together with $\|\mathbf{p}\|_0 + \|\mathbf{q}\|_0 \geq 2/\mu(\mathbf{D})$, the last inequality results in the following threshold for (P0)-uniqueness:

$$s < \frac{1}{\mu(\mathbf{D})}.$$

Compare this to the old threshold

$$s < \frac{1}{2}\left(1 + \frac{1}{\mu(\mathbf{D})}\right) \approx \frac{1}{2\mu(\mathbf{D})}$$

to observe that for dictionaries which are the concatenation of two ONBs we get an improvement in the recovery threshold by a factor of two.

3.5.8 General Matrices and Different Scenarios

There are various applications for the problem of recovering the vectors $\mathbf{x}$ and $\mathbf{e}$ from the observation

$$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$$

knowing that $\mathbf{x}$ and $\mathbf{e}$ are sparse. Let $\mathcal{X} = \operatorname{supp}(\mathbf{x})$ and $\mathcal{E} = \operatorname{supp}(\mathbf{e})$.

1. no knowledge about $\mathcal{X}$ and $\mathcal{E}$
   e.g. a spectrally sparse signal with unknown spectrum corrupted by impulses with unknown locations, i.e., $\mathbf{A} = \mathbf{F}$, $\mathbf{B} = \mathbf{I}$

2. cardinality of $\mathcal{X}$ or $\mathcal{E}$ known
   e.g. recovery of a sparse pulse-stream with known number of pulses per unit time corrupted by an electric hum with unknown base-frequency but known number of harmonics, i.e., $\mathbf{A} = \mathbf{I}$, $\mathbf{B} = \mathbf{F}$

3. knowledge of $\mathcal{X}$ or $\mathcal{E}$

   $$\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}_{\mathcal{E}}\mathbf{e}_{\mathcal{E}}$$

   e.g. a clipped spectrally sparse signal with unknown spectrum, i.e., $\mathbf{A} = \mathbf{F}$ and $\mathbf{B} = \mathbf{I}$, or inpainting (or super-resolution) where the unknown signal has a sparse representation in $\mathbf{A}$ (e.g. 2D-DCT or wavelet transform)

   $$\mathbf{z} = \mathbf{A}_{\mathcal{X}}\mathbf{x}_{\mathcal{X}} + \mathbf{B}\mathbf{e}$$

   e.g. recovery of spectrally sparse signals with known support $\mathcal{X}$ impaired by impulse noise at unknown locations

4. knowledge of both $\mathcal{X}$ and $\mathcal{E}$

   $$\mathbf{z} = \mathbf{A}_{\mathcal{X}}\mathbf{x}_{\mathcal{X}} + \mathbf{B}_{\mathcal{E}}\mathbf{e}_{\mathcal{E}}$$

   e.g. recovery of clipped band-limited signals with known spectral support, i.e., $\mathbf{A} = \mathbf{F}$, $\mathbf{B} = \mathbf{I}$.

We start by considering recovery via (P0) in case 3 when $\mathcal{E}$ is known.

Theorem 3.6. [35] *Let* $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$*, where* $\mathcal{E} = \operatorname{supp}(\mathbf{e})$ *is known. Consider the problem*

$$(P0,\, \mathcal{E}) \quad \begin{cases} \text{minimize } \|\hat{\mathbf{x}}\|_0 \\ \text{subject to } \mathbf{A}\hat{\mathbf{x}} \in (\{\mathbf{z}\} + \mathcal{R}(\mathbf{B}_{\mathcal{E}}))\,. \end{cases}$$

If $n_{\mathbf{x}} = \|\mathbf{x}\|_0$ *and* $n_{\mathbf{e}} = \|\mathbf{e}\|_0$ *satisfy*

$$2n_{\mathbf{x}}n_{\mathbf{e}} < \frac{[1-\mu(\mathbf{A})\,(2n_{\mathbf{x}}-1)]^+\,[1-\mu(\mathbf{B})\,(n_{\mathbf{e}}-1)]^+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})},$$

then the unique solution of $(P0,\, \mathcal{E})$ *applied to* $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$ *is given by* $\mathbf{x}$.

Proof. Assume that there exists an alternative vector $\tilde{\mathbf{x}}$ that satisfies $\mathbf{A}\tilde{\mathbf{x}} \in (\{\mathbf{z}\} + \mathcal{R}(\mathbf{B}_{\mathcal{E}}))$ with $\|\tilde{\mathbf{x}}\|_0 \le n_{\mathbf{x}}$. Then, there must exist an $\tilde{\mathbf{e}}$ with $\operatorname{supp}(\tilde{\mathbf{e}}) \subseteq \mathcal{E}$ such that

$$\mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e} = \mathbf{A}\tilde{\mathbf{x}} + \mathbf{B}\tilde{\mathbf{e}},$$

which implies

$$\mathbf{A}\,(\mathbf{x} - \tilde{\mathbf{x}}) = \mathbf{B}\,(\tilde{\mathbf{e}} - \mathbf{e})\,.$$

We now apply the uncertainty principle Corollary 2.17 to the vectors $(\mathbf{x}-\tilde{\mathbf{x}})$ and $(\mathbf{e}-\tilde{\mathbf{e}})$. If both $\mathbf{x}$ and $\tilde{\mathbf{x}}$ have at most $n_{\mathbf{x}}$ nonzero entires, they can differ in at most $2n_{\mathbf{x}}$ positions, i.e., $\|\mathbf{x}-\tilde{\mathbf{x}}\|_0 \leq 2n_{\mathbf{x}}$. The vectors $\mathbf{e}$ and $\tilde{\mathbf{e}}$ are supported on the same set $\mathcal{E}$ of cardinality $n_{\mathbf{e}}$. Hence $\mathbf{e}$ and $\tilde{\mathbf{e}}$ can differ in at most $n_{\mathbf{e}}$ positions, i.e., $\|\mathbf{e}-\tilde{\mathbf{e}}\|_0 \leq n_{\mathbf{e}}$.
The uncertainty relation Corollary 2.17 sets a limit on how sparse $(\mathbf{x}-\tilde{\mathbf{x}})$ and $(\tilde{\mathbf{e}}-\mathbf{e})$ can be. In particular, $(\mathbf{x}-\tilde{\mathbf{x}})$ and $(\tilde{\mathbf{e}}-\mathbf{e})$ cannot both be arbitrarily sparse. Substituting

$$\begin{aligned} \mathbf{p} = \mathbf{x}-\tilde{\mathbf{x}}, \quad \mathcal{P} = \operatorname{supp}(\mathbf{x}-\tilde{\mathbf{x}})\,, \quad |\mathcal{P}| \leq 2n_{\mathbf{x}} \\ \mathbf{q} = \tilde{\mathbf{e}}-\mathbf{e}, \quad \mathcal{Q} = \operatorname{supp}(\tilde{\mathbf{e}}-\mathbf{e})\,, \quad |\mathcal{Q}| \leq n_{\mathbf{e}} \end{aligned}$$

into the uncertainty relation results in

$$\begin{aligned} \|\mathbf{p}\|_0\|\mathbf{q}\|_0 = |\mathcal{P}||\mathcal{Q}| &\geq \frac{[1-\mu(\mathbf{A})\,(|\mathcal{P}|-1)]^+\,[1-\mu(\mathbf{B})\,(|\mathcal{Q}|-1)]^+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})} \\ &\geq \frac{[1-\mu(\mathbf{A})\,(2n_{\mathbf{x}}-1)]^+\,[1-\mu(\mathbf{B})\,(n_{\mathbf{e}}-1)]^+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})}. \end{aligned}$$

However, since $\|\mathbf{p}\|_0\|\mathbf{q}\|_0 \leq 2n_{\mathbf{x}}n_{\mathbf{e}}$, this contradicts the original assumption. □

In the theorem below, we state a sufficient condition for recovery via (BP) in case 3 when $\mathcal{E}$ is known.

Theorem 3.7. [35] *Let $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$, where $\mathcal{E} = \operatorname{supp}(\mathbf{e})$ is known. Consider the convex problem*

$$(BP,\, \mathcal{E}) \quad \begin{cases} \text{minimize } \|\hat{\mathbf{x}}\|_1 \\ \text{subject to } \mathbf{A}\hat{\mathbf{x}} \in (\{\mathbf{z}\} + \mathcal{R}(\mathbf{B}_{\mathcal{E}}))\,. \end{cases}$$

If $n_{\mathbf{x}} = \|\mathbf{x}\|_0$ and $n_{\mathbf{e}} = \|\mathbf{e}\|_0$ satisfy

$$2n_{\mathbf{x}}n_{\mathbf{e}} < \frac{[1-\mu(\mathbf{A})\,(2n_{\mathbf{x}}-1)]^+\,[1-\mu(\mathbf{B})\,(n_{\mathbf{e}}-1)]^+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})},$$

then the unique solution of $(BP,\, \mathcal{E})$ applied to $\mathbf{z} = \mathbf{A}\mathbf{x} + \mathbf{B}\mathbf{e}$ is given by $\mathbf{x}$.

Proof. Assume that there exists an alternative solution $\tilde{\mathbf{x}}$ with $\|\tilde{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1$. This would imply

$$\mathbf{A}\,(\mathbf{x}-\tilde{\mathbf{x}}) = \mathbf{B}\,(\tilde{\mathbf{e}}-\mathbf{e})\,.$$

Set $\mathbf{p} = \mathbf{x}-\tilde{\mathbf{x}}$ and let $(\cdot)_{\mathcal{A}}$ denote the projection onto the space of vectors with support $\mathcal{A}$, i.e., $(\mathbf{h}_{\mathcal{A}})_l = h_l I_{l\in\mathcal{A}}$. We find the following lower bound on the 1-norm of $\tilde{\mathbf{x}}$:

$$\begin{aligned} \|\tilde{\mathbf{x}}\|_1 = \|\mathbf{x}-\mathbf{p}\|_1 &= \|(\mathbf{x}-\mathbf{p})_{\mathcal{X}}\|_1 + \|\mathbf{p}_{\mathcal{X}^c}\|_1 \\ &\geq \|\mathbf{x}_{\mathcal{X}}\|_1 - \|\mathbf{p}_{\mathcal{X}}\|_1 + \|\mathbf{p}_{\mathcal{X}^c}\|_1 \\ &= \|\mathbf{x}\|_1 - \|\mathbf{p}_{\mathcal{X}}\|_1 + \|\mathbf{p}_{\mathbf{x}^c}\|_1. \end{aligned}$$

Therefore, $\|\tilde{\mathbf{x}}\|_1 \leq \|\mathbf{x}\|_1$ is possible only if

$$\|\mathbf{p}_{\mathcal{X}}\|_1 \geq \|\mathbf{p}_{\mathcal{X}^c}\|_1,$$

i.e., if $\mathbf{p}$ is $(1/2)_{\mathcal{X}}$-concentrated (see Definition 2.11). Let $\mathbf{q} = \tilde{\mathbf{e}} - \mathbf{e}$, $\mathcal{Q} = \operatorname{supp}(\tilde{\mathbf{e}} - \mathbf{e}) \subseteq \mathcal{E} \Rightarrow |\mathcal{Q}| \leq n_{\mathbf{e}}$. Applying the uncertainty principle stated in Part 3 of Corollary 2.16 to $\mathbf{p}$ and $\mathbf{q}$ yields

$$\begin{aligned} n_{\mathbf{x}} n_{\mathbf{e}} \geq |\mathcal{X}||\mathcal{Q}| &\geq \frac{[(1+\mu(\mathbf{A}))(1-\epsilon_{\mathcal{X}}) - |\mathcal{X}|\mu(\mathbf{A})]^+ [1-\mu(\mathbf{B})(|\mathcal{Q}|-1)]^+}{\mu^2} \\ &\geq \frac{[(1+\mu(\mathbf{A}))/2 - |\mathcal{X}|\mu(\mathbf{A})]^+ [1-\mu(\mathbf{B})(|\mathcal{Q}|-1)]^+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})} \\ &\geq \frac{1}{2} \frac{[1-\mu(\mathbf{A})(2n_{\mathbf{x}}-1)]^+ [1-\mu(\mathbf{B})(n_{\mathbf{e}}-1)]^+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})}. \end{aligned}$$

Multiplying both sides of the above inequality by a factor of two gives

$$2n_{\mathbf{x}} n_{\mathbf{e}} \geq \frac{[1-\mu(\mathbf{A})(2n_{\mathbf{x}}-1)]^+ [1-\mu(\mathbf{B})(n_{\mathbf{e}}-1)]^+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})},$$

which contradicts the original assumption

$$2n_{\mathbf{x}} n_{\mathbf{e}} < \frac{[1-\mu(\mathbf{A})(2n_{\mathbf{x}}-1)]^+ [1-\mu(\mathbf{B})(n_{\mathbf{e}}-1)]^+}{\bar{\mu}^2(\mathbf{A},\mathbf{B})}.$$

□

We next present an example saturating the threshold derived in the previous two theorems. For $\mathbf{A} = \mathbf{F}_m$ and $\mathbf{B} = \mathbf{I}_m$, this threshold becomes

$$2n_{\mathbf{x}} n_{\mathbf{e}} < \frac{1}{\bar{\mu}^2(\mathbf{I},\mathbf{F})} = m.$$

Take

$$\begin{aligned} \mathbf{x} &= \delta_{2\sqrt{m}} - \delta_{\sqrt{m}} \\ \mathbf{e} &= \delta_{\sqrt{m}} \end{aligned}$$

where δ_t denotes the vector whose l-th entry satisfies

$$[\delta_t]_l = \begin{cases} 1, \text{ if } (l-1) \bmod t = 0 \\ 0, \text{ else.} \end{cases}$$

Then, we have $2n_{\mathbf{x}} n_{\mathbf{e}} = 2\frac{\sqrt{m}}{2}\sqrt{m} = m$. One can verify by straightforward calculation that

$$\mathbf{F}_m \delta_t = \frac{\sqrt{m}}{t} \delta_{m/t}.$$

We therefore have

$$\mathbf{z} = \mathbf{F}_m\mathbf{x} + \mathbf{e} = \underbrace{\mathbf{F}_m\delta_{2\sqrt{m}} + \mathbf{F}_m\delta_{\sqrt{m}}}_{\mathbf{F}_m\mathbf{x}} - \underbrace{\delta_{\sqrt{m}}}_{\mathbf{e}} = \frac{1}{2}\delta_{\sqrt{m}/2} - \delta_{\sqrt{m}} + \delta_{\sqrt{m}} = \frac{1}{2}\delta_{\sqrt{m}/2}.$$

Next, consider the vectors $\tilde{\mathbf{x}} = \delta_{2\sqrt{m}}$ and $\tilde{\mathbf{e}} = \mathbf{0}$ with

$$\mathbf{F}_m\tilde{\mathbf{x}} + \mathbf{I}_m\tilde{\mathbf{e}} = \frac{1}{2}\delta_{\sqrt{m}/2} = \mathbf{F}_m\mathbf{x} + \mathbf{e}.$$

We thus have

$$\begin{aligned}(\mathbf{F}_m\mathbf{x}) &\in (\{\mathbf{z}\} + \mathcal{R}((\mathbf{I}_m)_{\mathcal{E}}))\\ (\mathbf{F}_m\tilde{\mathbf{x}}) &\in (\{\mathbf{z}\} + \mathcal{R}((\mathbf{I}_m)_{\mathcal{E}}))\,.\end{aligned}$$

In addition, we have $\|\mathbf{x}\|_0 = \|\tilde{\mathbf{x}}\|_0$ and $\|\mathbf{x}\|_1 = \|\tilde{\mathbf{x}}\|_1$. Therefore, (P0, $\mathcal{E}$) and (BP, $\mathcal{E}$) cannot distinguish between $\mathbf{x}$ and $\tilde{\mathbf{x}}$.

All sparsity thresholds we obtained so far are proportional to $1/\mu(\mathbf{D})$. What can we say about the dictionary coherence $\mu(\mathbf{D})$?

Theorem 3.8. [54] *Let $\mathbf{D} \in \mathbb{C}^{m\times n}$ be a dictionary with coherence $\mu(\mathbf{D})$. Then,*

$$\mu(\mathbf{D}) \geq \sqrt{\frac{n-m}{m\,(n-1)}},$$

where $m \leq n$.

Proof. Set $\mathbf{G} = \mathbf{D}^{\mathsf{H}}\mathbf{D} \in \mathbb{C}^{n\times n}$. Then, $\mathbf{G}$ has the following properties:

1. $\mathbf{G}$ has ones along its diagonal (since all dictionary columns have unit ℓ_2 norm);

2. $\mathbf{G}$ is positive semi-definite with rank (at most) m.

Let $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_m)^{\mathsf{T}}$ denote the vector of nonzero eigenvalues λ_i of $\mathbf{G}$. Then, we have

$$\operatorname{tr}\mathbf{G} = \sum_{i=1}^{m}\lambda_i = \|\boldsymbol{\lambda}\|_1 = n \tag{3.3}$$

$$\|\mathbf{G}\|_2^2 = \sum_{i=1}^{m}\lambda_1^2 = \|\boldsymbol{\lambda}\|_2^2. \tag{3.4}$$

Since

$$\left(\frac{1}{m}\sum_{i=1}^{m}\lambda_i\right)^2 \leq \frac{1}{m}\sum_{i=1}^{m}\lambda_i^2 \tag{3.5}$$

by Jensen's inequality, it follow that $\|\boldsymbol{\lambda}\|_1^2 \leq m\|\boldsymbol{\lambda}\|_2^2$, which implies in turn that

$$\|\mathbf{G}\|_2^2 \geq \frac{n^2}{m}.$$

We thus have

$$\begin{aligned}\|\mathbf{G}\|_2^2 &= n + \sum_{i=1}^{n}\sum_{j\neq i} |\langle \mathbf{d}_i, \mathbf{d}_j\rangle|^2 \\ &\geq \frac{n^2}{m},\end{aligned}$$

which finally yields

$$\mu(\mathbf{D})^2 \geq \frac{1}{n(n-1)}\sum_{i=1}^{n}\sum_{j\neq i} |\langle \mathbf{d}_i, \mathbf{d}_j\rangle|^2 \tag{3.6}$$

$$\geq \frac{1}{n(n-1)}\left(\frac{n^2}{m} - n\right) \tag{3.7}$$

$$= \frac{n-m}{m\,(n-1)}. \tag{3.8}$$

□

For $n \gg m$ the Welch lower bound implies

$$\mu(\mathbf{D}) \geq \sqrt{\frac{n-m}{m\,(n-1)}} \approx \frac{1}{\sqrt{m}}$$

and hence all sparsity thresholds obtained so far obey the fundamental upper bound

$$s \lesssim \sqrt{m} \quad \Rightarrow \quad m \gtrsim s^2.$$

We therefore say that the derived sparsity thresholds are bounded by the *square-root bottleneck* meaning that recovery is only guaranteed if we take $m \approx s^2$ samples for an s-sparse signal.

Take, e.g., $s = 30$ and $n = 1000$. The square-root bottleneck implies that we would need ≈ 900 samples to get recovery through $(\mathrm{P1})$ or OMP. This is very disappointing.

Hence, the question: Can we improve upon the scaling behavior $m \gtrsim s^2$?
Since we have proved that the derived thresholds are tight, the answer to this question is of course negative if we insist upon successful recovery of each signal. The way out is randomized sampling that essentially asks for recoverability only in almost all cases, i.e., we exclude the (few) signals for which recovery fails when the threshold is saturated or only slightly exceeded.

Chapter 4

Finite Rate of Innovation

Consider a finite-length sequence consisting of K Dirac impulses of unknown locations and with unknown weights

$$x(t) = \sum_{k=0}^{K-1} c_k \delta(t - t_k) \quad , \quad 0 \leq t_k \leq \tau$$

This signal has bandwidth ∞ and according to the classical sampling theorem, we would have to sample it at rate ∞ if we wanted to reconstruct the signal from its samples. However, we realize that the signal has only $2K$ unknown parameters, namely $\{t_k, c_k\}_{k=0}^{K-1}$. It is therefore conceivable that this signal can be recovered from a finite number of measurements. Specifically, we will consider lowpass measurements in the form of Fourier series coefficients.

$$d_n = \frac{1}{\tau} \int_0^\tau \sum_{k=0}^{K-1} c_k \delta(t - t_k) e^{-i2\pi n \frac{t}{\tau}} dt = \frac{1}{\tau} \sum_{k=0}^{K-1} c_k e^{-i2\pi n \frac{t_k}{\tau}}.$$

The periodized version of $x(t)$ can hence be written as

$$x(t) = \sum_{n \in \mathbb{Z}} \left(\frac{1}{\tau} \sum_{k=0}^{K-1} c_k e^{-i2\pi n \frac{t_k}{\tau}} \right) e^{i2\pi n \frac{t}{\tau}}.$$

and the question we want to answer is: How can we recover $x(t)$ from a finite number of Fourier series coefficients d_n, how many Fourier series coefficients would we need at least, and how would a corresponding recovery algorithm look like?

The algorithm we consider is the so-called ”annihilating filter method”, which is a Berlekamp-Massy algorithm over the complex numbers. The method starts with the construction of a filter

$$A(z) = \sum_{m=0}^{K} a_m z^{-m},$$

which has K zeros, namely at the locations $u_k = e^{-i2\pi \frac{t_k}{\tau}}$, that is

$$A(z) = \prod_{k=0}^{K-1} \left(1 - e^{-i2\pi \frac{t_k}{\tau}} z^{-1} \right).$$

Note that this $A(z)$ is the convolution of K elementary filters with impulse responses $(\delta[n] - e^{-i2\pi\frac{t_k}{\tau}}\delta[n-1])_{n\in\mathbb{Z}}$, $k = 0, 1, \dots, K-1$. The convolution of such an elementary filter with the sequence $e^{-i2\pi\frac{t_k}{\tau}n}$ equals zero. This can be seen as follows:

$$
\begin{aligned}
&\left(e^{-i2\pi\frac{t_k}{\tau}\cdot} \star \left(\delta[\cdot] - e^{-i2\pi\frac{t_k}{\tau}}\delta[\cdot-1]\right)\right)[n]\\
=&e^{-i2\pi\frac{t_k}{\tau}n} - e^{-i2\pi\frac{t_k}{\tau}}e^{-i2\pi\frac{t_k}{\tau}(n-1)} =\\
=&e^{-i2\pi\frac{t_k}{\tau}n} - \underbrace{e^{-i2\pi\frac{t_k}{\tau}}e^{i2\pi\frac{t_k}{\tau}}}_{=1}e^{-i2\pi\frac{t_k}{\tau}n} = 0, \quad \forall n \in \mathbb{Z}
\end{aligned}
$$

As the Fourier series coefficients d_n are linear combinations of exponentials $e^{-i2\pi\frac{t_k}{\tau}n}$, it follows that

$$(d_l)_{l\in\mathbb{Z}} \star (a_l)_{l\in\mathbb{Z}} = 0.$$

Specifically, each of the exponentials in this linear combination is annihilated by one of the factors $(1 - e^{-i2\pi\frac{t_k}{\tau}}z^{-1})$.

In summary, for a given Fourier series coefficient sequence d_n, if we can find the correspondig annihilating filter impulse response a_n, the zeros of $A(z) = \sum_{n=0}^{K} a_n z^{-n}$ yield the locations t_k through $A(e^{-i2\pi\frac{t_k}{\tau}}) = 0$, as $t_k = -\frac{\tau}{2\pi}\arg(e^{-i2\pi\frac{t_k}{\tau}})$. Once we have the t_k, the corresponding weights c_k can be obtained by solving a linear system of equations given by $d_n = \frac{1}{\tau}\sum_{k=0}^{K-1} c_k u_k^n$.

4.1 Finding the annihilating filter

The condition

$$\sum_{l=0}^{K} a_l d_{n-l} = 0, \quad \forall n \in \mathbb{Z}$$

in matrix-vector form reads

$$
\begin{array}{r}
\\
n=0 \rightarrow\\
n=1 \rightarrow\\
\vdots\\
n=K \rightarrow\\
\\
\end{array}
\left[\begin{array}{|cccc|}
\vdots & & := \mathbf{S} & \vdots\\
\hline
d_0 & d_{-1} & \dots & d_{-K}\\
d_1 & d_0 & \dots & d_{-K+1}\\
\vdots & & & \vdots\\
d_K & d_{K-1} & \dots & d_0\\
\hline
\vdots & & & \vdots
\end{array}\right]
\begin{bmatrix} a_0\\ a_1\\ \vdots\\ a_K \end{bmatrix} = 0. \tag{4.1}
$$

To solve this linear system of equations, we need at least $2K+1$ Fourier series coefficients, namely $\{d_{-K}, \dots, d_0, \dots, d_K\}$. In practice, this linear system of equations is solved by identifying the singular vector of $\mathbf{S}$ corresponding to the smallest singular value.

4.1.1 Finding the a_k

Once the filter impulse response coefficients $a_0, \ldots, a_K$ are found, we write

$$A(z) = \sum_{m=0}^{K} a_m z^{-m} = \prod_{k=0}^{K-1} \left(1 - \alpha_k z^{-1}\right)$$

and identify the zeros α_k which yield the numbers $u_k = e^{-i2\pi \frac{t_k}{\tau}}$.

4.1.2 Finding the c_k

To determine the weights c_k, it suffices to take K equations among

$$d_n = \frac{1}{\tau} \sum_{k=0}^{K-1} c_k u_k^n$$

which, in matrix-vector form, reads

$$\frac{1}{\tau} \begin{bmatrix} 1 & 1 & \ldots & 1 \\ u_0 & u_1 & \ldots & u_{K-1} \\ \vdots & \vdots & & \vdots \\ u_0^{K-1} & u_1^{K-1} & \ldots & u_{K-1}^{K-1} \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{K-1} \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{K-1} \end{bmatrix}$$

and has a unique solution when $u_k \neq u_l, \forall k \neq l$ (since the system matrix is a Vandermonde matrix). We hence have a method that retrieves the $2K$ unknowns $\{t_k, c_k\}$ from $\geq 2K+1$ Fourier series coefficients.

4.1.3 Uniqueness

We now deal with the question of uniqueness of the solution to (4.1). First rewrite S as a linear combination of rank-1 matrices:

$$\mathbf{S} = \frac{1}{\tau} \sum_{k=0}^{K-1} c_k \begin{bmatrix} 1 & u_k^{-1} & \ldots & u_k^{-K} \\ u_k & 1 & \ldots & u_k^{-K+1} \\ u_k^2 & u_k & \ldots & u_k^{-K+2} \\ \vdots & \vdots & & \vdots \\ u_k^K & u_k^{K-1} & \ldots & 1 \end{bmatrix}$$

$$= \frac{1}{\tau} \sum_{k=0}^{K-1} c_k \underbrace{\begin{bmatrix} 1 \\ u_k \\ u_k^2 \\ \vdots \\ u_k^K \end{bmatrix} \begin{bmatrix} 1 & u_k^{-1} & u_k^{-2} & \ldots & u_k^{-K} \end{bmatrix}}_{(K+1)\times(K+1) \text{ matrix of rank 1}}$$

The individual rank-one matrices are linearly independent as the Vandermonde matrices

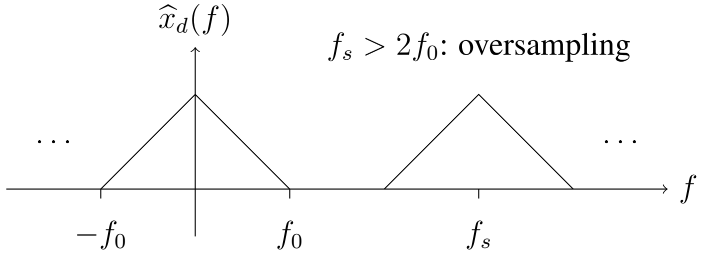
$$\begin{bmatrix} 1 & 1 & \dots & 1 \\ u_0 & u_1 & \dots & u_K^1 \\ u_0^2 & u_1^2 & \dots & u_K^2 \\ \vdots & \vdots & & \vdots \\ u_0^K & u_1^K & \dots & u_K^K \end{bmatrix}, \quad \begin{bmatrix} 1 & 1 & \dots & 1 \\ u_0^{-1} & u_1^{-1} & \dots & u_K^{-1} \\ u_0^{-2} & u_1^{-2} & \dots & u_K^{-2} \\ \vdots & \vdots & & \vdots \\ u_0^{-K} & u_1^{-K} & \dots & u_K^{-K} \end{bmatrix}$$

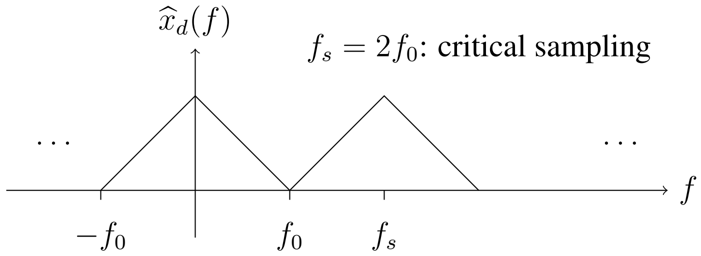
are full-rank, provided all the u_k are different. Therefore, provided all $c_k \neq 0$, the $(K+1) \times (K+1)$ matrix **S** is of rank K and hence the system (4.1) has a unique solution.

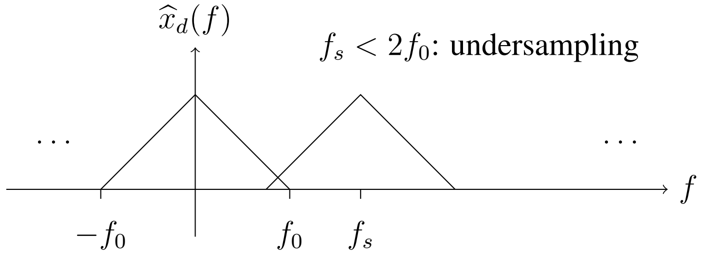
Chapter 5

Sampling of Multi-Band Signals

5.1 Classical Sampling Theorem

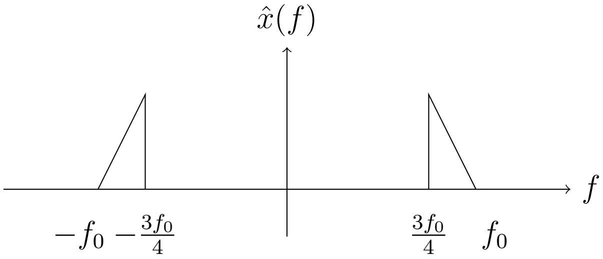




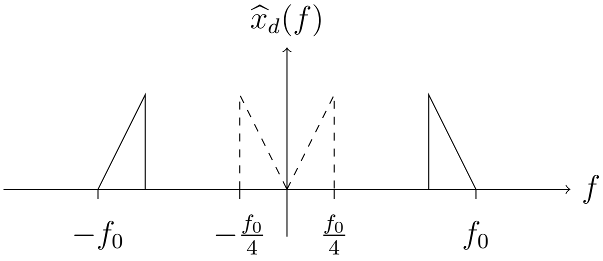


5.2 Sampling Spectrally Sparse Signals

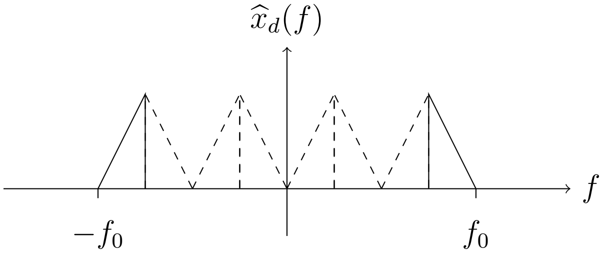
Assume that the spectrum has sparse support in $[-f_0, f_0]$, e.g.



Then two-fold undersampling, i.e., $f_s = f_0$, results in the periodized spectrum below.



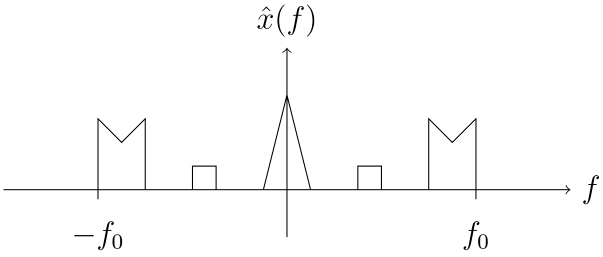
Four-fold undersampling, i.e., $f_s = f_0/2$, yields the following periodized spectrum.



The original signal $x(t)$ can be perfectly reconstructed even though we undersample it by a factor of four. For the example above, the minimum sampling rate required in order for exact recovery to be possible equals the support set size of the nonzero spectral components.

Can we do this in general? Yes, Landau's multi-band sampling theorem suggests that this is, indeed, the case.

Consider a signal with spectral occupancy $\mathcal{I} \subset [-f_0, f_0]$.



Assume the sampling set $\mathcal{P} = \{t_n\}$, i.e., we are given the signal values $\{x(t_n)\}$.

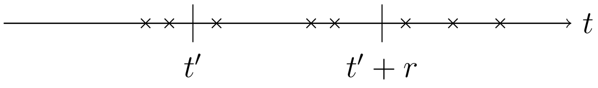
Theorem 5.1 (Landau, 1967)**.** *To reconstruct* stably*, we need*

$$\mathcal{D}^-(\mathcal{P}) = \lim_{r\to\infty} \inf_{t\in\mathbb{R}} \frac{|\mathcal{P} \cap [t, t+r]|}{r} \geq |\mathcal{I}|$$

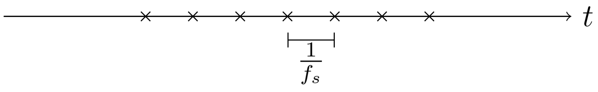
where $\mathcal{D}^-(\mathcal{P})$ denotes the lower Beurling density.

5.2.1 Interpretation of the Lower Beurling Density

- Fix r.
- Slide a window of length r across the t-axis, find the smallest number of sampling points in any of these intervals and divide by r. Note that the result depends on r.
- Take the window length to infinity and compute the limit of the function of r specified in the previous point.



Lower Beurling density for regular sampling at rate f_s:



The number of samples in an interval $[t, t+r]$ of length r is given by an integer $N_{t,r}$ satisfying

$$\left| N_{t,r} - \frac{r}{T_s} \right| = |N_{t,r} - rf_s| \leq 1.$$

Consequently, it holds that

$$\inf_{t\in\mathbb{R}} \frac{|\mathcal{P}\cap[t,t+r]|}{r} \in \left[f_s - \frac{1}{r}, f_s + \frac{1}{r}\right]$$

and the lower Beurling density is given by

$$\lim_{r\to\infty} \inf_{t\in\mathbb{R}} \frac{|\mathcal{P}\cap[t,t+r]|}{r} = f_s \geq |\mathcal{I}|.$$

5.2.2 Stable Sampling

Definition 5.2. A set of points $\mathcal{P} = \{t_n\}$ is called a *stable sampling set* if for all $x_1, x_2 \in \mathcal{H}$

$$A\|x_1 - x_2\|_{\mathcal{H}}^2 \leq \|x_1(\mathcal{P}) - x_2(\mathcal{P})\|_2^2 \leq B\|x_1 - x_2\|_{\mathcal{H}}^2$$

for some $A > 0$ and $B < \infty$.

If $\mathcal{H}$ is a vector space (and therefore satisfies the linearity property), we have $x_1 - x_2 \in \mathcal{H}$ and hence for all $x \in \mathcal{H}$

$$A\|x\|_{\mathcal{H}}^2 \leq \|\mathbb{T}x\|_2^2 \leq B\|x\|_{\mathcal{H}}^2$$

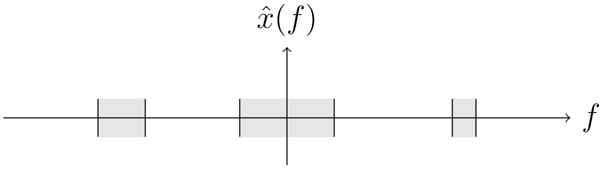
where $\mathbb{T} : x(t) \to \{x(\mathcal{P})\}$ denotes the sampling operator. This is nothing but a frame condition with the frame operator given by $\mathbb{S} = \mathbb{T}^*\mathbb{T}$, i.e., for all $x \in \mathcal{H}$

$$A\|x\|_{\mathcal{H}}^2 \leq \|\mathbb{T}x\|_2^2 = \langle \mathbb{T}x, \mathbb{T}x\rangle = \left\langle \underbrace{\mathbb{T}^*\mathbb{T}}_{\mathbb{S}} x, x \right\rangle \leq B\|x\|_{\mathcal{H}}^2.$$

We go back to sampling of multi-band signals and consider the set

$$\mathcal{B}(\mathcal{I}) \triangleq \left\{x(t) \in \mathscr{L}^2(\mathbb{R}) : \hat{x}(f) = 0, \forall f \notin \mathcal{I}\right\}.$$

Is the space $\mathcal{B}(\mathcal{I})$ of signals with Fourier transform supported on a given interval $\mathcal{I}$ a vector space?



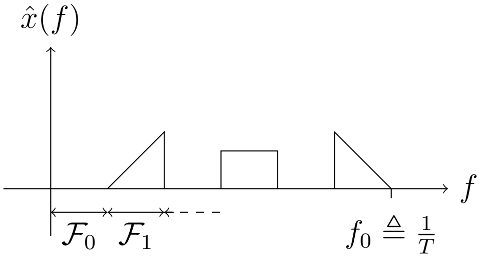
Since every linear combination of signals in $\mathcal{B}(\mathcal{I})$ is in $\mathcal{B}(\mathcal{I})$, the above question can be answered in the affirmative, i.e., $\mathcal{B}(\mathcal{I})$ is a vector space.

Theorem 5.1 states that sampling at or above the Landau rate is necessary, is it also sufficient? In other words, can we identify a universal sampling pattern with rate equal to the Landau rate so that any signal in $\mathcal{B}(\mathcal{I})$ can be stably reconstructed from these samples?

5.3 Multicoset Sampling

We partition the overall spectral support region into L cells $\mathcal{F}_i$ of equal length $\frac{f_0}{L}$, i.e.,

$$\mathcal{F}_i = \left[i\frac{f_0}{L}, (i+1)\frac{f_0}{L} \right), \quad \text{for } i \in \{0, \dots, L-1\}.$$



For $L \to \infty$ this setup becomes the general setup considered previously. For L finite, we approximate $\mathcal{I}$ by s intervals of length $\frac{f_0}{L}$, i.e., $|\mathcal{I}| \approx \frac{sf_0}{L}$.

The signal $x(t)$ is sampled on a periodic nonuniform grid

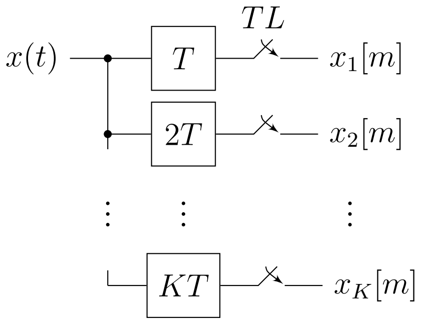
$$\Psi = \Psi_1 \cup \dots \cup \Psi_K,$$

which is the union of K subgrids

$$\Psi_k = \{(mL+k)T : m \in \mathbb{Z}\}, \quad \text{for } k = 1, \dots, K.$$

The samples corresponding to Ψ_k are

$$x_k[m] \triangleq x((mL+k)T), m \in \mathbb{Z}.$$



The overall sampling rate is given by (For every subgrid Ψ_k the sampling rate is $1/(LT) = f_0/L$.)

$$\mathcal{D}^{-}(\mathcal{P}) = \frac{K}{LT} = \frac{K}{L} f_0.$$

For every coset $\{x_k[m]\}_{m\in\mathbb{Z}}$, we compute the discrete-time Fourier transform

$$\begin{aligned}
x_d^{(k)}(f) &= \sum_{m\in\mathbb{Z}} x_k[m]\, e^{-i2\pi fmTL} \\
&= \sum_{m\in\mathbb{Z}} x((mL+k)\,T)\, e^{-i2\pi fmTL} \\
&= e^{i2\pi fkT} \sum_{m\in\mathbb{Z}} x\Bigg(mTL + \underbrace{kT}_{t}\Bigg)\, e^{-i2\pi f(mTL+kT)} \\
&= e^{i2\pi fkT} \frac{1}{TL} \sum_{m\in\mathbb{Z}} \widehat{x}\Big(f + \frac{m}{TL}\Big)\, e^{i2\pi\frac{mk}{L}}, \quad f\in[0,1),
\end{aligned}$$

where the last equality follows from the Poisson summation formula

$$\sum_{l\in\mathbb{Z}} s(t+lT) = \frac{1}{T}\sum_{l\in\mathbb{Z}} \widehat{s}\Big(\frac{l}{T}\Big)\, e^{i2\pi\frac{l}{T}t}.$$

Next, we introduce the functions

$$\begin{aligned}
v_k(f) &:= x_d^{(k)}(f) e^{-i2\pi fkT} TL \\
&= \sum_{m\in\mathbb{Z}} \widehat{x}\Big(f + \frac{m}{TL}\Big)\, e^{i2\pi\frac{mk}{L}}, \quad f\in[0,1/(LT)),
\end{aligned}$$

and write

$$\underbrace{\begin{bmatrix} v_1(f) \\ v_2(f) \\ \vdots \\ v_K(f) \end{bmatrix}}_{=:\mathbf{v}(f)\,\in\,\mathbb{C}^K} = \underbrace{\begin{bmatrix} 1 & e^{i2\pi\frac{1}{L}} & \cdots & e^{i2\pi\frac{L-1}{L}} \\ 1 & e^{i2\pi\frac{2}{L}} & \cdots & e^{i2\pi\frac{2(L-1)}{L}} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & e^{i2\pi\frac{K}{L}} & \cdots & e^{i2\pi\frac{K(L-1)}{L}} \end{bmatrix}}_{=:\mathbf{A}\,\in\,\mathbb{C}^{K\times L}} \underbrace{\begin{bmatrix} \widehat{x}(f) \\ \widehat{x}\big(f+\frac{1}{TL}\big) \\ \vdots \\ \widehat{x}\big(f+\frac{L-1}{TL}\big) \end{bmatrix}}_{=:\hat{\mathbf{x}}(f)\in\mathbb{C}^L}, \quad f\in[0,1/(LT)).$$

It holds that $K \leq L$; ideally, we would like to choose $K = s$ so that sampling is performed at the Landau rate. Clearly, the original signal $x(t)$ can be reconstructed as soon as the vector $\hat{\mathbf{x}}(f)$ has been recovered for all $f \in \big[0, \frac{1}{TL}\big]$. Noting that $[\mathbf{A}]_{k,m} = e^{i2\pi\frac{mk}{L}}$ for $m \in \{0,\dots,L-1\}$ and $k\in\{1,\dots,K\}$ and setting $a_j \triangleq e^{i2\pi\frac{j}{L}}$ we can write

$$\mathbf{A} = \begin{bmatrix} a_0 & a_1 & a_2 & \dots & a_{L-1} \\ a_0^2 & a_1^2 & a_2^2 & \dots & a_{L-1}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_0^K & a_1^K & a_2^K & \dots & a_{L-1}^K \end{bmatrix}.$$

A *Vandermonde* matrix $\mathbf{V}(z_0, z_1, \ldots, z_{K-1})$ has full rank when all z_i are distinct, where

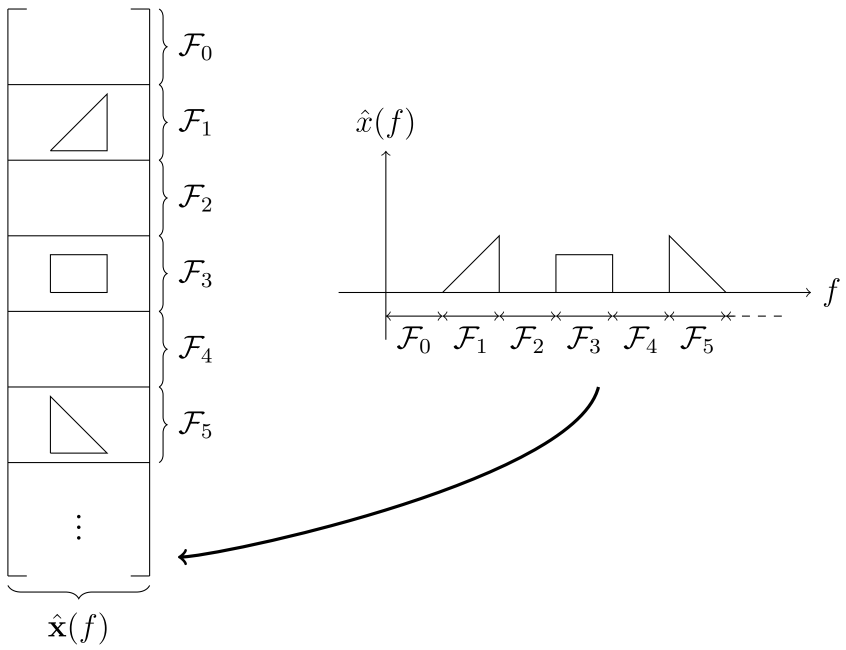
$$\mathbf{V}(z_0, z_1, \ldots, z_{K-1}) = \begin{bmatrix} 1 & z_0 & z_0^2 & \ldots & z_0^{K-1} \\ 1 & z_1 & z_1^2 & \ldots & z_1^{K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{K-1} & z_{K-1}^2 & \ldots & z_{K-1}^{K-1} \end{bmatrix}.$$

For every set $\mathcal{S} = \{s_0, s_1, \ldots, s_{K-1}\} \subset \{1, \ldots, L\}$ of cardinality $|\mathcal{S}| = K$, let $\mathbf{A}_\mathcal{S}$ denote the submatrix which contains the columns of $\mathbf{A}$ indexed by $\mathcal{S}$. Since the rank of a Vandermonde matrix is easy to find and for every matrix $\mathbf{B}$ we have $\operatorname{rank} \mathbf{B} = \operatorname{rank} \mathbf{B}^\mathsf{T}$, it is helpful to express $\mathbf{A}_\mathcal{S}^\mathsf{T}$ in terms of a Vandermonde matrix. Take $z_i \triangleq a_{s_i}$ for $i \in \{0, \ldots, K-1\}$ and note that

$$\begin{bmatrix} z_0 & z_1 & z_2 & \ldots & z_{K-1} \\ z_0^2 & z_1^2 & z_2^2 & \ldots & z_{K-1}^2 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_0^K & z_1^K & z_2^K & \ldots & z_{K-1}^K \end{bmatrix}^\mathsf{T} = \begin{bmatrix} z_0 & z_0^2 & z_0^3 & \ldots & z_0^K \\ z_1 & z_1^2 & z_1^3 & \ldots & z_1^K \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ z_{K-1}^1 & z_{K-1}^2 & z_{K-1}^3 & \ldots & z_{K-1}^K \end{bmatrix}$$

$$= \underbrace{\operatorname{diag}\left((z_0, z_1, \ldots, z_{K-1})\right)}_{\text{full-rank}} \underbrace{\begin{bmatrix} 1 & z_0 & z_0^2 & \ldots & z_0^{K-1} \\ 1 & z_1 & z_1^2 & \ldots & z_1^{K-1} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & z_{K-1} & z_{K-1}^2 & \ldots & z_{K-1}^{K-1} \end{bmatrix}}_{\mathbf{V}(z_0, z_1, \ldots, z_{K-1})}.$$

Consequently, every set of $s \leq K$ columns of the matrix $\mathbf{A}$ is linearly independent as the transpose of the resulting matrix is obtained when multiplying a matrix of full-rank by a Vandermonde matrix $\mathbf{V}(z_0, z_1, \ldots, z_{K-1})$ with distinct z_i.



Note that the vector $\hat{\mathbf{x}}(f)$ is sparse as illustrated by the figure above. Since we know the spectral support $\mathcal{I}$ of the original signal $x(t)$ and hence the support set of the vector $\hat{\mathbf{x}}(f)$, we can recover the vector $\hat{\mathbf{x}}(f)$ (or, more precisely, the entries corresponding to nonzero components) according to

$$\mathbf{v}(f) = \mathbf{A}\hat{\mathbf{x}}(f) = \mathbf{A}_\gamma \hat{\mathbf{x}}_\gamma(f) \Rightarrow \hat{\mathbf{x}}_\gamma(f) = \mathbf{A}_\gamma^\dagger \mathbf{v}(f),$$

where γ denotes the set of indices corresponding to nonzero entries of $\hat{\mathbf{x}}(f)$ and $\mathbf{A}_\gamma$ denotes the subset of columns of $\mathbf{A}$ with indices in γ.
Given the support set of $\hat{\mathbf{x}}(f)$, the minimum K we need is s. This corresponds to

$$\mathcal{D}^-(\mathcal{P}) = \frac{K}{LT} \geq \frac{s}{LT} \approx |\mathcal{I}|.$$

Multicoset sampling therefore allows recovery from samples taken at the Landau rate and is universal in the sense that it is applicable irrespective of the spectral occupancy $\mathcal{I}$ provided that the number of occupied cells $\mathcal{F}_i$ is at most s.

5.4 Spectrum-Blind Sampling

Let us now consider the case where the support set γ is not known a priori, but we know that $|\gamma| \leq s$. This amounts to considering the set

$$\mathcal{X}(C) = \bigcup_{|\mathcal{I}| \leq C} \mathcal{B}(\mathcal{I})\,.$$

Recall the definition of stable sampling

$$A\|x_1 - x_2\|^2 \leq \|\mathbb{T}x_1 - \mathbb{T}x_2\|_2^2 \leq B\|x_1 - x_2\|^2, \quad \forall x_1, x_2 \in \mathcal{X}(C)\,.$$

$x_1 - x_2 \notin \mathcal{X}(C)$ in general but $x_1 - x_2 \in \mathcal{X}(2C)$ so that the stable sampling condition reduces to

$$A\|x\|_{\mathcal{X}(2C)}^2 \leq \|\mathbb{T}x\|_2^2 \leq B\|x\|_{\mathcal{X}(2C)}^2, \quad \forall x \in \mathcal{X}(2C)\,.$$

To satisfy

$$\|\mathbb{T}x\|_2^2 \geq A\|x\|_{\mathcal{X}(2C)}^2, \quad \forall x \in \mathcal{X}(2C)$$

we obtain the necessary condition

$$\mathcal{D}^-(\mathcal{P}) \geq 2C$$

from Landau's Theorem 5.1.

How about sufficiency? If $\|\hat{\mathbf{x}}_1(f) - \hat{\mathbf{x}}_2(f)\|_0 \leq 2s \leq K, \forall f \in \left[0, \frac{1}{LT}\right]$, the Vandermonde structure implies that for all $x_1 - x_2$ multicoset sampling is stable. (Recall that $\operatorname{spark}(\mathbf{A}) = K + 1$, where the spark of a matrix denotes the minimal number of linearly dependent columns. Hence, a vector with positive ℓ_2 norm can only be mapped to zero if it has at least $\operatorname{spark}(\mathbf{A})$ nonzero entries.)

$$\|\hat{\mathbf{x}}(f)\|_0 = s \leq \frac{K}{2}$$

implies that the cardinality of the spectral occupancy satisfies

$$|\mathcal{I}| = \frac{s}{LT} \leq \frac{1}{LT}\frac{K}{2} = \frac{1}{2}\underbrace{\frac{K}{LT}}_{\mathcal{D}^-(\mathcal{P})} = \frac{\mathcal{D}^-(\mathcal{P})}{2}.$$

Since $|\mathcal{I}| \leq C$ holds for all $x \in \mathcal{X}(C)$ we find that $\mathcal{D}^-(\mathcal{P}) \geq 2C$ is also sufficient for stable sampling.

Chapter 6

The ESPRIT Algorithm

6.1 Introduction

The foundations of high-resolution methods for estimating the parameters of a sum of complex exponentials (cisoids) were laid in 1795 by [55]. Prony's method is, however, very sensitive to additive noise [56, 57]. Modern high resolution methods rely on signal subspace concepts and exhibit better noise robustness properties. Prominent subspace methods include the MUltiple SIgnal Classification (MUSIC) algorithm [58], including Pisarenko's method [59] as a special case, the Estimation of Signal Parameters via Rotational Invariance Techniques (ESPRIT) algorithm [60, 61], and the Toeplitz Approximation Method (TAM) [62, 63]. Originally developed for the estimation of the parameters of sums of undamped cisoids in noise, all these techniques were later found to apply to exponentially damped cisoids as well. Other subspace methods—specifically devised for damped cisoids in noise—include the Kumaresan-Tufts (KT) algorithm [64], the modified KT (MKT) algorithm [65], and the Matrix Pencil (MP) method [66, 67]. An excellent survey of subspace methods can be found in [68].

The problem of estimating the parameters of a sum of (damped or undamped) cisoids arises in numerous practical applications such as direction finding in array signal processing [69], velocity and acceleration estimation from lidar or radar echoes [70, 71], resolution of overlapping echoes [72], super-resolution [73], signal sampling theory [74, 75], line spectral estimation [76], spectral analysis of musical signals [77], speech signal analysis and synthesis [78], and musical signal modification [79].

Formally, the problem we consider is as follows. Recover the complex numbers $z_1, z_2, \ldots, z_K$, with $|z_k| \leq 1$, $k \in \{1, 2, \ldots, K\}$, henceforth referred to as "nodes", and the corresponding complex weights $\alpha_1, \alpha_2, \ldots, \alpha_K$ from the measurements

$$x_n = \sum_{k=1}^{K} \alpha_k z_k^n, \quad n \in \{0, 1, \ldots, N-1\}, \tag{6.1}$$

where the number of samples, N, satisfies $N \geq 2K$. The nodes $z_1, z_2, \ldots, z_K$ can be written as

$z_k = e^{-d_k} e^{2\pi i \xi_k}$, $k \in \{1, 2, \ldots, K\}$, where $d_k \geq 0$ is referred to as the damping factor and ξ_k as the normalized frequency of the k-th cisoid. In the remainder of this chapter, we assume that the nodes $z_1, z_2, \ldots, z_K$ as well as the corresponding weights $\alpha_1, \alpha_2, \ldots, \alpha_K$ are all non-zero. We furthermore take the $z_1, z_2, \ldots, z_K$ to be pairwise distinct, i.e., $z_{k_1} \neq z_{k_2}$, for $k_1 \neq k_2$.

Notation. The complex conjugate of $z \in \mathbb{C}$ is denoted by $\overline{z}$. Lowercase boldface letters stand for column vectors and uppercase boldface letters for matrices. The superscripts ^{T} and ^{H} designate transposition and Hermitian transposition, respectively. The ℓ^2-norm of the vector $\boldsymbol{x} \triangleq \{x_k\}_{k=1}^K \in \mathbb{C}^K$ is $\|\boldsymbol{x}\|_2 \triangleq \left(\sum_{k=1}^K |x_k|^2\right)^{1/2}$. For the matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$, $\|\mathbf{A}\|_{2,2} \triangleq \max\{\|\mathbf{A}\boldsymbol{x}\|_2 \colon \boldsymbol{x} \in \mathbb{C}^N, \|\boldsymbol{x}\|_2 = 1\}$ denotes its spectral norm and $\|\mathbf{A}\|_{\mathrm{F}} \triangleq \left(\sum_{m=1}^M \sum_{n=1}^N |a_{m,n}|^2\right)^{1/2}$ its Frobenius norm. The Moore-Penrose pseudo-inverse of the full-rank matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$, $M \geq N$, is given by $\mathbf{A}^\dagger = (\mathbf{A}^H \mathbf{A})^{-1} \mathbf{A}^H \in \mathbb{C}^{N \times M}$. We write $\sigma_k(\mathbf{A})$ for the kth singular value of the matrix $\mathbf{A} \in \mathbb{C}^{M \times N}$, sorted in descending order, i.e., $\sigma_1(\mathbf{A}) \geq \sigma_2(\mathbf{A}) \geq \ldots \geq \sigma_{\min\{M,N\}}(\mathbf{A})$, and denote its smallest and largest singular values by $\sigma_{\min}(\mathbf{A})$ and $\sigma_{\max}(\mathbf{A})$, respectively, its spectral condition number is $\kappa(\mathbf{A}) \triangleq \sigma_{\max}(\mathbf{A}) / \sigma_{\min}(\mathbf{A})$, and $\mathrm{colsp}(\mathbf{A})$ and $\mathrm{rowsp}(\mathbf{A})$ designate its column span and row span, respectively. For $z_1, z_2, \ldots, z_K \in \mathbb{C}$ and $L \in \mathbb{N}$ such that $L \geq K$, $\mathscr{V}_{L \times K}(z_1, z_2, \ldots, z_K) \in \mathbb{C}^{L \times K}$ is the Vandermonde matrix

$$\mathscr{V}_{L \times K}(z_1, z_2, \ldots, z_K) \triangleq \begin{pmatrix} 1 & 1 & \ldots & 1 & 1 \\ z_1 & z_2 & \ldots & z_{K-1} & z_K \\ z_1^2 & z_2^2 & \ldots & z_{K-1}^2 & z_K^2 \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ z_1^{L-1} & z_2^{L-1} & \ldots & z_{K-1}^{L-1} & z_K^{L-1} \end{pmatrix},$$

with nodes z_k, $k \in \{1, 2, \ldots, K\}$. $\mathrm{diag}\,(a_1, a_2, \ldots, a_L) \in \mathbb{C}^{L \times L}$ denotes the diagonal matrix with $a_1, a_2, \ldots, a_L \in \mathbb{C}$ on its main diagonal. A matrix $\mathbf{A} \triangleq (\mathbf{\Lambda}\ \mathbf{0}) \in \mathbb{C}^{M \times N}$, where $M < N$ and $\mathbf{\Lambda} \triangleq \mathrm{diag}\,(\lambda_1, \lambda_2, \ldots, \lambda_M) \in \mathbb{C}^{M \times M}$ is referred to as a wide diagonal matrix and its transpose would be called a tall diagonal matrix. We say that a matrix is rectangular diagonal if it is either wide or tall diagonal. For $x_0, x_1, \ldots, x_{N-1} \in \mathbb{C}$ and $L \in \mathbb{N}$ such that $1 \leq L \leq N$, $\mathscr{H}_L(x_0, x_1, \ldots, x_{N-1}) \in \mathbb{C}^{L \times (N-L+1)}$ designates the Hankel matrix

$$\mathscr{H}_L(x_0, x_1, \ldots, x_{N-1}) \triangleq \begin{pmatrix} x_0 & x_1 & \cdots & x_{N-L-1} & x_{N-L} \\ x_1 & x_2 & \cdots & x_{N-L} & x_{N-L+1} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ x_{L-2} & x_{L-1} & \cdots & x_{N-3} & x_{N-2} \\ x_{L-1} & x_L & \cdots & x_{N-2} & x_{N-1} \end{pmatrix}.$$

For a Hermitian positive semi-definite matrix $\mathbf{A} \triangleq \mathbf{U}\mathbf{\Sigma}\mathbf{U}^H \in \mathbb{C}^{M \times M}$, where $\mathbf{U} \in \mathbb{C}^{M \times M}$ is unitary and $\mathbf{\Sigma} \in \mathbb{C}^{M \times M}$ is a diagonal matrix with the singular values $\sigma_1(\mathbf{A}), \sigma_2(\mathbf{A}), \ldots, \sigma_M(\mathbf{A})$ on the diagonal, the square root of $\mathbf{A}$ is given by $\mathbf{A}^{1/2} \triangleq \mathbf{U}\mathbf{\Sigma}^{1/2}\mathbf{U}^H$, where $\mathbf{\Sigma}^{1/2}$ is diagonal with $\sqrt{\sigma_1(\mathbf{A})}, \sqrt{\sigma_2(\mathbf{A})}, \ldots, \sqrt{\sigma_M(\mathbf{A})}$ on its diagonal. Finally, $\overline{\mathbf{A}}$ denotes the matrix obtained by element-wise complex conjugation of $\mathbf{A}$. For $\mathbf{A} \in \mathbb{C}^{M\times N}$, $\mathbf{B} \in \mathbb{C}^{M\times P}$, and $\mathbf{C} \in \mathbb{C}^{N\times P}$, $\begin{pmatrix}\mathbf{A} & \mathbf{B}\end{pmatrix} \in \mathbb{C}^{M\times(N+P)}$ and $\begin{pmatrix}\mathbf{B}\\ \mathbf{C}\end{pmatrix} \in \mathbb{C}^{(M+N)\times P}$ are the horizontal and vertical concatenation of $\mathbf{A}$ and $\mathbf{B}$, and $\mathbf{B}$ and $\mathbf{C}$, respectively.

6.2 Signal and Noise Subspaces

We start by describing the subspace concepts ESPRIT relies on. To this end, we first construct the data matrix $\mathbf{X} \triangleq \mathscr{H}_L(x_0, x_1, \dots, x_{N-1}) \in \mathbb{C}^{L\times(N-L+1)}$ from the signal samples $\{\sigma_n\}_{n=0}^{N-1}$ given by (6.1), and note that $\mathbf{X}$ can be factorized according to

$$\mathbf{X} = \mathbf{V}_L \mathbf{D}_{\boldsymbol{\alpha}} \mathbf{V}_{N-L+1}^T \in \mathbb{C}^{L\times(N-L+1)}, \tag{6.2}$$

where

$$\begin{aligned}
&\mathbf{V}_L \triangleq \mathscr{V}_{L\times K}(z_1, z_2, \dots, z_K) \in \mathbb{C}^{L\times K},\\
&\mathbf{D}_{\boldsymbol{\alpha}} \triangleq \operatorname{diag}(\alpha_1, \alpha_2, \dots, \alpha_K),\\
&\mathbf{V}_{N-L+1} \triangleq \mathscr{V}_{(N-L+1)\times K}(z_1, z_2, \dots, z_K) \in \mathbb{C}^{(N-L+1)\times K}.
\end{aligned}$$

Here, $L \in \mathbb{N}$ is a parameter, satisfying $K+1 \leq L \leq N-K-1$, that controls the aspect ratio of the data matrix $\mathbf{X}$. Since the nodes $z_1, z_2, \dots, z_K$ are non-zero and pairwise distinct and $K+1 \leq L \leq N-K-1$, the Vandermonde matrices $\mathbf{V}_L$ and $\mathbf{V}_{N-L+1}$ both have full rank K. As the weights $\alpha_1, \alpha_2, \dots, \alpha_K$ are non-zero, $\mathbf{D}_{\boldsymbol{\alpha}}$ is invertible, and hence by (6.2) the data matrix $\mathbf{X}$ has rank K. Therefore, $\mathbf{X}$ has K non-zero singular values $\sigma_1, \sigma_2, \dots, \sigma_K$ and can be decomposed according to

$$\mathbf{X} \triangleq \underbrace{\begin{pmatrix}\mathbf{S} & \mathbf{S}_\perp\end{pmatrix}}_{=:\,\mathbf{U}\in\mathbb{C}^{L\times L}} \underbrace{\begin{pmatrix}\boldsymbol{\Lambda} & \mathbf{0}\\ \mathbf{0} & \mathbf{0}\end{pmatrix}}_{=:\,\boldsymbol{\Sigma}\in\mathbb{R}^{L\times(N-L+1)}} \underbrace{\begin{pmatrix}\mathbf{R}^H\\ \mathbf{R}_\perp^H\end{pmatrix}}_{=:\,\mathbf{W}^H\in\mathbb{C}^{(N-L+1)\times(N-L+1)}} = \mathbf{S}\boldsymbol{\Lambda}\mathbf{R}^H, \tag{6.3}$$

where $\mathbf{U}$ and $\mathbf{W}$ are unitary, $\boldsymbol{\Sigma}$ is a rectangular diagonal matrix, and $\boldsymbol{\Lambda} \triangleq \operatorname{diag}(\sigma_1, \sigma_2, \dots, \sigma_K)$. The columns of $\mathbf{S} \in \mathbb{C}^{L\times K}$ and $\mathbf{R} \in \mathbb{C}^{(N-L+1)\times K}$ are the left-singular and the right-singular vectors, respectively, corresponding to the K non-zero singular values $\sigma_1, \sigma_2, \dots, \sigma_K$ of $\mathbf{X}$. The matrices $\mathbf{S}_\perp \in \mathbb{C}^{L\times(L-K)}$ and $\mathbf{R}_\perp \in \mathbb{C}^{(N-L+1)\times(N-L-K+1)}$ contain the remaining $L-K$ left-singular and the remaining $N-L-K+1$ right-singular vectors of $\mathbf{X}$, respectively, that is, the singular vectors corresponding to the singular values of $\mathbf{X}$ that equal zero. Since $\mathbf{U} = \begin{pmatrix}\mathbf{S} & \mathbf{S}_\perp\end{pmatrix}$ is unitary, the columns of $\mathbf{S}$ constitute an orthonormal basis for the column span of $\mathbf{S}$, henceforth denoted by $\mathcal{S}$, and the columns of $\mathbf{S}_\perp$ form an orthonormal basis for the orthogonal complement $\mathcal{S}^\perp$ of $\mathcal{S}$ in $\mathbb{C}^L$. Owing to (6.3) and the fact that $\mathbf{X}$ and $\mathbf{S}$ both have rank K, we can conclude that $\mathcal{S}$ coincides with the column span of $\mathbf{X}$.

6.3 The ESPRIT Algorithm

Starting from the decomposition (6.2) of the data matrix $\mathbf{X}$, we let $\mathbf{V}_\downarrow \in \mathbb{C}^{(L-1)\times K}$ be the matrix consisting of the top $L-1$ rows of $\mathbf{V}_L$ and $\mathbf{V}_\uparrow \in \mathbb{C}^{(L-1)\times K}$ the matrix consisting of the bottom $L-1$ rows of $\mathbf{V}_L$. We have

$$\mathbf{V}_\uparrow = \mathbf{V}_\downarrow \mathbf{D}_{\boldsymbol{z}}, \qquad \text{where} \quad \mathbf{D}_{\boldsymbol{z}} \triangleq \operatorname{diag}(z_1, z_2, \ldots, z_K).$$

Since the columns of both $\mathbf{S}$ and $\mathbf{V}_L$ form bases for the signal subspace $\mathcal{S}$, there exists an invertible matrix $\mathbf{P} \in \mathbb{C}^{K\times K}$ such that $\mathbf{S} = \mathbf{V}_L \mathbf{P}$. Next, letting $\mathbf{S}_\downarrow \in \mathbb{C}^{(L-1)\times K}$ be the matrix consisting of the top $L-1$ rows of $\mathbf{S}$ and $\mathbf{S}_\uparrow \in \mathbb{C}^{(L-1)\times K}$ the matrix consisting of the bottom $L-1$ rows of $\mathbf{S}$, it follows from $\mathbf{V}_\uparrow = \mathbf{V}_\downarrow \mathbf{D}_{\boldsymbol{z}}$ that

$$\mathbf{S}_\uparrow = \mathbf{S}_\downarrow \mathbf{\Phi},$$

where $\mathbf{\Phi} \triangleq \mathbf{P}^{-1}\mathbf{D}_{\boldsymbol{z}}\mathbf{P}$. As $\mathbf{D}_{\boldsymbol{z}} = \operatorname{diag}(z_1, z_2, \ldots, z_K)$ and $\mathbf{P} \in \mathbb{C}^{K\times K}$ is invertible, $z_1, z_2, \ldots, z_K$ are the eigenvalues of the matrix $\mathbf{\Phi} = \mathbf{S}_\downarrow^\dagger \mathbf{S}_\uparrow$. This is a direct consequence of the similarity principle reviewed at the end of this section.

We next show that $\mathbf{\Phi} = \mathbf{S}_\downarrow^\dagger \mathbf{S}_\uparrow$ is indeed a solution to $\mathbf{S}_\uparrow = \mathbf{S}_\downarrow \mathbf{Y}$. To this end, we first note that $\mathbf{S}_\downarrow = \mathbf{V}_{L-1}\mathbf{P}$ and $\mathbf{S}_\uparrow = \mathbf{V}_{L-1}\mathbf{D}_{\boldsymbol{z}}\mathbf{P}$. Taken together this yields

$$\mathbf{S}_\downarrow \mathbf{S}_\downarrow^\dagger \mathbf{S}_\uparrow = \mathbf{V}_{L-1}\mathbf{P}\mathbf{P}^{-1}\mathbf{V}_{L-1}^\dagger \mathbf{V}_{L-1}\mathbf{D}_{\boldsymbol{z}}\mathbf{P} = \mathbf{V}_{L-1}\mathbf{D}_{\boldsymbol{z}}\mathbf{P} = \mathbf{S}_\uparrow,$$

as $L \geq K+1$ ensures that $\mathbf{V}_{L-1}$ has full rank and hence $\mathbf{V}_{L-1}^\dagger \mathbf{V}_{L-1} = \mathbf{I}_K$. This implies that $\mathbf{\Phi} = \mathbf{S}_\downarrow^\dagger \mathbf{S}_\uparrow$ is a solution to $\mathbf{S}_\downarrow \mathbf{Y} = \mathbf{S}_\uparrow$. As $\mathbf{S}_\downarrow = \mathbf{V}_{L-1}\mathbf{P}$, $\operatorname{rank}(\mathbf{V}_{L-1}) = K$, and $\mathbf{P}$ is invertible, it follows that $\mathbf{S}_\downarrow$ has full rank, and hence $\mathbf{\Phi} = \mathbf{S}_\downarrow^\dagger \mathbf{S}_\uparrow$ is the unique solution of $\mathbf{S}_\downarrow \mathbf{Y} = \mathbf{S}_\uparrow$.

Review of the similarity principle. We start with the definition of the similarity of matrices.

Definition 6.1. The matrices $\mathbf{X} \in \mathbb{C}^{n\times n}$ and $\mathbf{Y} \in \mathbb{C}^{n\times n}$ are similar if there exists an invertible $n \times n$ matrix $\mathbf{P}$ such that $\mathbf{X} = \mathbf{P}^{-1}\mathbf{Y}\mathbf{P}$.

Theorem 6.2. *Let $\mathbf{A}$ and $\mathbf{B}$ be similar matrices. Then, $\mathbf{A}$ and $\mathbf{B}$ have the same eigenvalues with the same geometric multiplicities.*

Proof.

$$\mathbf{A} = \mathbf{P}^{-1}\mathbf{B}\mathbf{P} \implies \mathbf{B} = \mathbf{P}\mathbf{A}\mathbf{P}^{-1}.$$

If $\mathbf{A}\mathbf{u} = \lambda\mathbf{u}$, then $\mathbf{P}^{-1}\mathbf{B}\mathbf{P} = \lambda\mathbf{u}$ and therefore $\mathbf{B}\mathbf{P}\mathbf{u} = \lambda\mathbf{P}\mathbf{u}$. So, if $\mathbf{u}$ is an eigenvector of $\mathbf{A}$, then $\mathbf{P}\mathbf{u}$ is an eigenvector of $\mathbf{B}$ with eigenvalue λ, i.e., with the same eigenvalue. Every eigenvalue of $\mathbf{A}$ is hence an eigenvalue of $\mathbf{B}$.

Conversely, since

$$\begin{aligned} \mathbf{B}\mathbf{u} = \lambda\mathbf{u} &\implies \mathbf{P}\mathbf{A}\mathbf{P}^{-1}\mathbf{u} = \lambda\mathbf{u} \\ &\implies \mathbf{A}\mathbf{P}^{-1}\mathbf{u} = \lambda\mathbf{P}^{-1}\mathbf{u}, \end{aligned}$$

every eigenvalue of $\mathbf{B}$ is an eigenvalue of $\mathbf{A}$. □

6.4 Finding the Zeros of a Polynomial

We now show how the ideas developed in the previous section can be used to devise an algorithm for finding the zeros of arbitrary polynomials through the SVD.

Consider the polynomial

$$p(z) = \alpha_0 + \alpha_1 z + \cdots + \alpha_K z^K.$$

For simplicity of exposition, we assume that all zeros of $p(z)$ have multiplicity one. First, we write

$$p(z) = (\alpha_0\ \alpha_1\ \dots\ \alpha_K) \begin{pmatrix} 1 \\ z \\ \vdots \\ z^K \end{pmatrix},$$

and note that for the zeros $z_0, \dots, z_{K-1}$ (to be found), we have

$$\underbrace{(\alpha_0\ \alpha_1\ \dots\ \alpha_K)}_{=:\boldsymbol{\alpha}\in\mathbb{C}^{1\times(K+1)}} \underbrace{\begin{pmatrix} 1 & 1 & \dots & 1 \\ z_0 & z_1 & \dots & z_{K-1} \\ \vdots & \vdots & & \vdots \\ z_0^K & z_1^K & \dots & z_{K-1}^K \end{pmatrix}}_{=:\mathbf{V}\in\mathbb{C}^{(K+1)\times K}} = \mathbf{0}^T,$$

where $\mathbf{V}$ is a Vandermonde matrix with nodes given by the zeros of the polynomial $p(z)$. As the zeros z_i are all of multiplicity one, by assumption, it follows that the Vandermonde matrix $\mathbf{V}$ is of rank K, i.e., of full rank.

Next, we note that $\mathbf{V}$ is a basis for the right null-space of $\boldsymbol{\alpha}$. This null-space has dimension K (the ambient space has dimension $K + 1$). We determine a basis for this null-space according to

$$\boldsymbol{\alpha} = \underbrace{\begin{bmatrix} u_1\ \mathbf{u}_2 \end{bmatrix}}_{1\times(K+1)} \underbrace{\begin{bmatrix} \sigma & \mathbf{0}_{1\times K} \\ \mathbf{0}_{K\times 1} & \mathbf{0}_{K\times K} \end{bmatrix}}_{(K+1)\times(K+1)} \underbrace{\begin{bmatrix} \mathbf{w}^T \\ \mathbf{Z}^T \end{bmatrix}}_{(K+1)\times(K+1)},$$

where u_1 and σ are scalars, $\mathbf{u}_2$ is a $1 \times K$ vector, $\mathbf{w}$ is a $(K+1) \times 1$ vector, and $\mathbf{Z}$ is a $(K+1) \times K$ matrix containing the basis for the null-space. It is easily verified that $\boldsymbol{\alpha}\mathbf{Z} = \mathbf{0}_{1\times K}$. As both $\mathbf{V}$ and $\mathbf{Z}$ are bases for the right null-space of $\boldsymbol{\alpha}$, they must be related through a full-rank matrix, denoted as $\mathbf{T}$, according to

$$\mathbf{V} = \mathbf{Z}\,\mathbf{T}.$$

Finally, we again exploit the shift structure of the Vandermonde matrix $\mathbf{V}$ according to

$$\mathbf{V}_{\uparrow} = \mathbf{V}_{\downarrow}\mathbf{D}_{\mathbf{z}}$$

and note that

$$\left.\begin{aligned}\mathbf{V}_{\uparrow} &= \mathbf{Z}_{\uparrow}\mathbf{T}\\ \mathbf{V}_{\downarrow} &= \mathbf{Z}_{\downarrow}\mathbf{T}\end{aligned}\right\} \implies \mathbf{Z}_{\uparrow}\mathbf{T} = \mathbf{Z}_{\downarrow}\mathbf{T}\mathbf{D}_{\mathbf{z}}$$

$$\implies \mathbf{Z}_{\uparrow} = \mathbf{Z}_{\downarrow}\underbrace{\mathbf{T}\mathbf{D}_{\mathbf{z}}\mathbf{T}^{-1}}_{\mathbf{\Phi}}.$$

Just like in the ESPRIT algorithm, by similarity, the zeros z_i are given by the eigenvalues of $\mathbf{\Phi} = \mathbf{Z}_{\downarrow}^{\dagger}\mathbf{Z}_{\uparrow}$.

Chapter 7

The restricted isometry property

Assume that we observe $\mathbf{y} = \mathbf{\Phi}\mathbf{x} \in \mathbb{C}^m$, where $\mathbf{x} \in \mathbb{C}^n$ is a signal, unknown to us and that we want to reconstruct, and $\mathbf{\Phi} \in \mathbb{C}^{m\times n}$ is a known measurement matrix. Here, we consider the underdetermined case with fewer equations than unknowns, i.e., $m < n$. In the noisy case, the observation is given by $\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{n} \in \mathbb{C}^m$.

Definition 7.1. For each integer $s = 1, 2, \ldots$, define the isometry constant δ_s of a matrix $\mathbf{\Phi}$ as the smallest number such that

$$(1-\delta_s)\|\mathbf{x}\|_2^2 \le \|\mathbf{\Phi}\mathbf{x}\|_2^2 \le (1+\delta_s)\|\mathbf{x}\|_2^2$$

holds for all s-sparse vectors $\mathbf{x}$. A vector is said to be s-sparse if it has at most s nonzero entries.

Theorem 7.2. *Let $\mathbf{y} = \mathbf{\Phi}\mathbf{x}$. Assume that $\delta_{2s} < \sqrt{2} - 1$. Then, the solution $\mathbf{x}^*$ to*

$$\underset{\widetilde{\mathbf{x}}\in\mathbb{R}^n}{\text{minimize}}\ \|\widetilde{\mathbf{x}}\|_1 \quad \text{subject to } \mathbf{\Phi}\widetilde{\mathbf{x}} = \mathbf{y} \tag{7.1}$$

obeys

$$\|\mathbf{x}^* - \mathbf{x}\|_1 \le C_0\|\mathbf{x} - \mathbf{x}_s\|_1 \tag{7.2}$$

and

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \le C_0 s^{-1/2}\|\mathbf{x} - \mathbf{x}_s\|_1 \tag{7.3}$$

for some constant C_0 specified in the proof. Here, $\mathbf{x}_s$ is the vector obtained by setting all but the s largest entries of $\mathbf{x}$ equal to zero. In particular, if $\mathbf{x}$ is s-sparse, recovery is exact.

Theorem 7.3. *Let $\mathbf{y} = \mathbf{\Phi}\mathbf{x} + \mathbf{n}$. Assume that $\delta_{2s} < \sqrt{2} - 1$ and $\|\mathbf{n}\|_2 \le \varepsilon$. Then, the solution $\mathbf{x}^*$ to*

$$\underset{\widetilde{\mathbf{x}}\in\mathbb{R}^n}{\text{minimize}}\ \|\widetilde{\mathbf{x}}\|_1 \quad \text{subject to } \|\mathbf{y} - \mathbf{\Phi}\widetilde{\mathbf{x}}\|_2 \le \varepsilon$$

obeys

$$\|\mathbf{x}^* - \mathbf{x}\|_2 \le C_0 s^{-1/2}\|\mathbf{x} - \mathbf{x}_s\|_1 + C_1\varepsilon$$

with the same constant C_0 as before and some constant C_1 given explicitly in the proof.

We note that the bound (7.3) in Theorem 7.2 follows directly by setting $\mathbf{n} = \mathbf{0}$ and hence $\varepsilon = 0$ in Theorem 7.3. The bound (7.2) will be proved separately. Before embarking on the proof of Theorem 7.3, we need the following auxiliary result.

Lemma 7.4. *We have*

$$|\langle \mathbf{\Phi v}, \mathbf{\Phi v}' \rangle| \leq \delta_{s+s'} \|\mathbf{v}\|_2 \|\mathbf{v}'\|_2$$

for all $\mathbf{v}$*,* $\mathbf{v}'$ *supported on disjoint subsets* $\mathcal{Q}, \mathcal{Q}' \subseteq \{1, \ldots, n\}$ *with* $|\mathcal{Q}| \leq s$ *and* $|\mathcal{Q}'| \leq s'$*.*

Proof. Without loss of generality, let us assume that $\|\mathbf{v}\|_2 = \|\mathbf{v}'\|_2 = 1$. By definition of the restricted isometry constant $\delta_{s+s'}$, it holds that

$$\|\mathbf{v} \pm \mathbf{v}'\|_2^2 (1 - \delta_{s+s'}) \leq \|\mathbf{\Phi}(\mathbf{v} \pm \mathbf{v}')\|_2^2 \leq \|\mathbf{v} \pm \mathbf{v}'\|_2^2 (1 + \delta_{s+s'}).$$

Since $\mathbf{v}$ and $\mathbf{v}'$ are disjointly supported and $\|\mathbf{v}\|_2 = \|\mathbf{v}'\|_2 = 1$ by assumption, we have

$$\|\mathbf{v} \pm \mathbf{v}'\|_2^2 = \|\mathbf{v}\|_2^2 + \|\mathbf{v}'\|_2^2 = 2.$$

It follows that

$$2(1 - \delta_{s+s'}) \leq \|\mathbf{\Phi}(\mathbf{v} \pm \mathbf{v}')\|_2^2 \leq 2(1 + \delta_{s+s'}).$$

Applying the polarization identity

$$\langle \mathbf{u}, \mathbf{u}' \rangle = \frac{1}{4} \left(\|\mathbf{u} + \mathbf{u}'\|_2^2 - \|\mathbf{u} - \mathbf{u}'\|_2^2 \right)$$

to $\mathbf{u} = \mathbf{\Phi v}$ and $\mathbf{u}' = \mathbf{\Phi v}'$, we obtain

$$|\langle \mathbf{\Phi v}, \mathbf{\Phi v}' \rangle| = \frac{1}{4} \left| \|\mathbf{\Phi v} + \mathbf{\Phi v}'\|_2^2 - \|\mathbf{\Phi v} - \mathbf{\Phi v}'\|_2^2 \right|.$$

Resolving $| \cdot |$ such that $\arg > 0$ yields

$$\begin{aligned} |\langle \mathbf{\Phi v}, \mathbf{\Phi v}' \rangle| &= \frac{1}{4} \left(\|\mathbf{\Phi v} + \mathbf{\Phi v}'\|_2^2 - \|\mathbf{\Phi v} - \mathbf{\Phi v}'\|_2^2 \right) \\ &\leq \frac{1}{4} \cdot 2(1 + \delta_{s+s'}) - \frac{1}{4} \cdot 2(1 - \delta_{s+s'})) \\ &= \frac{1}{2}(1 + \delta_{s+s'} - 1 + \delta_{s+s'}) = \delta_{s+s'}. \end{aligned}$$

Resolving $| \cdot |$ such that $\arg < 0$ yields

$$\begin{aligned} |\langle \mathbf{\Phi v}, \mathbf{\Phi v}' \rangle| &= \frac{1}{4} \left(\|\mathbf{\Phi v} - \mathbf{\Phi v}'\|_2^2 - \|\mathbf{\Phi v} + \mathbf{\Phi v}'\|_2^2 \right) \\ &\leq \frac{1}{4} \cdot 2(1 + \delta_{s+s'}) - \frac{1}{4} \cdot 2(1 - \delta_{s+s'})) \\ &= \frac{1}{2}(1 + \delta_{s+s'} - 1 + \delta_{s+s'}) = \delta_{s+s'}. \end{aligned}$$

□

Proof of Theorem 7.3. Let us denote by $\mathbf{x}_\mathcal{Q}$ the vector equal to $\mathbf{x}$ on the index set $\mathcal{Q}$ and zero elsewhere. We start with the basic observation:

$$\|\mathbf{\Phi}(\mathbf{x}^* - \mathbf{x})\|_2 \leq \underbrace{\|\mathbf{\Phi}\mathbf{x}^* - \mathbf{y}\|_2}_{\leq \varepsilon \,(\text{as } \mathbf{x}^* \text{ is feasible})} + \underbrace{\|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2}_{= \|\mathbf{n}\|_2 \leq \varepsilon} \leq 2\varepsilon$$

Write $\mathbf{x}^* = \mathbf{x} + \mathbf{h}$ and decompose $\mathbf{h}$ into a sum of vectors $\mathbf{h}_{\mathcal{Q}_0}, \mathbf{h}_{\mathcal{Q}_1}, \ldots$, each of sparsity s. $\mathcal{Q}_0$ corresponds to the locations of the s largest—in magnitude—entries of $\mathbf{x}$, $\mathcal{Q}_1$ to the locations of the s largest entries of $\mathbf{h}_{\mathcal{Q}_0^c}$, $\mathcal{Q}_2$ to the locations of the s next largest entries of $\mathbf{h}_{\mathcal{Q}_0^c}$ and so on. The proof proceeds in two steps:

1. the size of $\mathbf{h}$ outside $\mathcal{Q}_0 \cup \mathcal{Q}_1$ is essentially bounded by that of $\mathbf{h}$ on $\mathcal{Q}_0 \cup \mathcal{Q}_1$,
2. $\|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2$ is approximately small.

For the first step, we note that for each $j \geq 2$, we have

$$\left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_2 \leq s^{1/2}\left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_\infty \leq s^{-1/2}\left\|\mathbf{h}_{\mathcal{Q}_{j-1}}\right\|_1$$

because $s\left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_\infty \leq \left\|\mathbf{h}_{\mathcal{Q}_{j-1}}\right\|_1$, for $j \geq 2$. We therefore get

$$\sum_{j\geq 2} \left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_2 \leq s^{-1/2}\left(\|\mathbf{h}_{\mathcal{Q}_1}\|_1 + \|\mathbf{h}_{\mathcal{Q}_2}\|_1 + \ldots\right) \tag{7.4}$$

$$= s^{-1/2}\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1. \tag{7.5}$$

This yields the estimate

$$\left\|\mathbf{h}_{(\mathcal{Q}_0 \cup \mathcal{Q}_1)^c}\right\|_2 = \left\|\sum_{j\geq 2} \mathbf{h}_{\mathcal{Q}_j}\right\|_2 \leq \sum_{j\geq 2} \left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_2 \leq s^{-1/2}\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1. \tag{7.6}$$

The key point now is that $\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1$ cannot be very large as $\|\mathbf{x} + \mathbf{h}\|_1 = \|\mathbf{x}^*\|_1$ is minimal. By applying the reverse triangle inequality twice, we obtain, as both $\mathbf{x}$ and $\mathbf{x}^* = \mathbf{x} + \mathbf{h}$ are consistent, that

$$\|\mathbf{x}\|_1 \geq \|\mathbf{x} + \mathbf{h}\|_1 = \sum_{j \in \mathcal{Q}_0} |x_j + h_j| + \sum_{j \in \mathcal{Q}_0^c} |x_j + h_j| \tag{7.7}$$

$$\geq \|\mathbf{x}_{\mathcal{Q}_0}\|_1 - \|\mathbf{h}_{\mathcal{Q}_0}\|_1 + \left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1 - \left\|\mathbf{x}_{\mathcal{Q}_0^c}\right\|_1. \tag{7.8}$$

This yields the following chain of inequalities

$$\|\mathbf{x}\|_1 - \|\mathbf{x}_{\mathcal{Q}_0}\|_1 + \left\|\mathbf{x}_{\mathcal{Q}_0^c}\right\|_1 \geq -\|\mathbf{h}_{\mathcal{Q}_0}\|_1 + \left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1 \tag{7.9}$$

$$\|\mathbf{x}_{\mathcal{Q}_0}\|_1 + \left\|\mathbf{x}_{\mathcal{Q}_0^c}\right\|_1 - \|\mathbf{x}_{\mathcal{Q}_0}\|_1 + \left\|\mathbf{x}_{\mathcal{Q}_0^c}\right\|_1 \geq -\|\mathbf{h}_{\mathcal{Q}_0}\|_1 + \left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1 \tag{7.10}$$

$$2\left\|\mathbf{x}_{\mathcal{Q}_0^c}\right\|_1 + \|\mathbf{h}_{\mathcal{Q}_0}\|_1 \geq \left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1 \tag{7.11}$$

$$2\left\|\mathbf{x}_{\mathcal{Q}_0^c}\right\|_1 + \|\mathbf{h}_{\mathcal{Q}_0}\|_1 \geq s^{1/2}\left\|\mathbf{h}_{(\mathcal{Q}_0 \cup \mathcal{Q}_1)^c}\right\|_2, \tag{7.12}$$

where the last inequality follows from (7.6). Using the fact that $\|\mathbf{h}_{\mathcal{Q}_0}\|_1 \leq s^{1/2}\|\mathbf{h}_{\mathcal{Q}_0}\|_2$, this becomes

$$\begin{aligned} 2\big\|\mathbf{x}_{\mathcal{Q}_0^c}\big\|_1 + s^{1/2}\|\mathbf{h}_{\mathcal{Q}_0}\|_2 &\geq s^{1/2}\big\|\mathbf{h}_{(\mathcal{Q}_0\cup\mathcal{Q}_1)^c}\big\|_2 \\ 2s^{-1/2}\big\|\mathbf{x}_{\mathcal{Q}_0^c}\big\|_1 + \|\mathbf{h}_{\mathcal{Q}_0}\|_2 &\geq \big\|\mathbf{h}_{(\mathcal{Q}_0\cup\mathcal{Q}_1)^c}\big\|_2. \end{aligned}$$

By definition, $\mathbf{x}_{\mathcal{Q}_0^c} = \mathbf{x} - \mathbf{x}_s$. Therefore,

$$2\,\underbrace{s^{-1/2}\|\mathbf{x}-\mathbf{x}_s\|_1}_{=e_0} + \|\mathbf{h}_{\mathcal{Q}_0}\|_2 \geq \big\|\mathbf{h}_{(\mathcal{Q}_0\cup\mathcal{Q}_1)^c}\big\|_2. \tag{7.13}$$

Next, we bound $\big\|\mathbf{h}_{(\mathcal{Q}_0\cup\mathcal{Q}_1)^c}\big\|_2$. We have the following

$$\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1} = \mathbf{\Phi}\left(\mathbf{h} - \sum_{j\geq 2}\mathbf{h}_{\mathcal{Q}_j}\right) = \mathbf{\Phi}\mathbf{h} - \sum_{j\geq 2}\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_j},$$

which implies

$$\|\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2^2 = \langle\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}, \mathbf{\Phi}\mathbf{h}\rangle - \left\langle \mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}, \sum_{j\geq 2}\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_j}\right\rangle. \tag{7.14}$$

By the Cauchy-Schwarz inequality, we get

$$|\langle\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}, \mathbf{\Phi}\mathbf{h}\rangle| \leq \|\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2\|\mathbf{\Phi}\mathbf{h}\|_2.$$

Moreover, it holds that

$$\|\mathbf{\Phi}\underbrace{(\mathbf{x}^*-\mathbf{x})}_{=\mathbf{h}}\|_2 \leq \|\mathbf{\Phi}\mathbf{x}^* - \mathbf{y}\|_2 + \|\mathbf{y} - \mathbf{\Phi}\mathbf{x}\|_2 \leq 2\varepsilon, \tag{7.15}$$

which, combined with the definition of the restricted isometry constant, yields

$$\begin{aligned} |\langle\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}, \mathbf{\Phi}\mathbf{h}\rangle| &\leq \|\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2 \cdot 2\varepsilon \\ &\leq 2\varepsilon\sqrt{1+\delta_{2s}}\|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2. \end{aligned} \tag{7.16}$$

It follows from Lemma 7.4 that for all $j \geq 1$,

$$\left|\left\langle\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0}, \mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_j}\right\rangle\right| \leq \delta_{2s}\|\mathbf{h}_{\mathcal{Q}_0}\|_2\big\|\mathbf{h}_{\mathcal{Q}_j}\big\|_2, \tag{7.17}$$

and for all $j \geq 2$,

$$\left|\left\langle\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_1}, \mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_j}\right\rangle\right| \leq \delta_{2s}\|\mathbf{h}_{\mathcal{Q}_1}\|_2\big\|\mathbf{h}_{\mathcal{Q}_j}\big\|_2. \tag{7.18}$$

The sets $\mathcal{Q}_0$ and $\mathcal{Q}_1$ are disjoint, and hence,

$$\|\mathbf{h}_{\mathcal{Q}_0}\|_2 + \|\mathbf{h}_{\mathcal{Q}_1}\|_2 \leq \sqrt{2}\|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2. \tag{7.19}$$

This can be seen as follows:

$$\|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2^2 = \underbrace{\|\mathbf{h}_{\mathcal{Q}_0}\|_2^2}_{=a^2} + \underbrace{\|\mathbf{h}_{\mathcal{Q}_1}\|_2^2}_{=b^2},$$

and we have

$$\begin{aligned}\sqrt{2(a^2+b^2)} &\geq a+b\\ 2(a^2+b^2) &\geq a^2+b^2+2ab\\ a^2+b^2-2ab &\geq 0\\ (a-b)^2 &\geq 0.\end{aligned}$$

Using the triangle inequality, (7.17), (7.18), and (7.19), we obtain

$$\begin{aligned}\left|\left\langle \mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}, \sum_{j\geq 2}\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_j}\right\rangle\right| &\leq \left|\left\langle \mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0}, \sum_{j\geq 2}\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_j}\right\rangle\right| + \left|\left\langle \mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_1}, \sum_{j\geq 2}\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_j}\right\rangle\right|\\ &\leq \sum_{j\geq 2}\delta_{2s}\left(\|\mathbf{h}_{\mathcal{Q}_0}\|_2 + \|\mathbf{h}_{\mathcal{Q}_1}\|_2\right)\left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_2\\ &\leq \sqrt{2}\delta_{2s}\|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2\sum_{j\geq 2}\left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_2. \end{aligned} \tag{7.20}$$

Combining (7.14), (7.16), and (7.20), and using again the definition of the restricted isometry constant, we get

$$(1-\delta_{2s})\|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2^2 \leq \|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2\left(2\varepsilon\sqrt{1+\delta_{2s}} + \sqrt{2}\delta_{2s}\sum_{j\geq 2}\left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_2\right).$$

Furthermore, we have by (7.5),

$$\sum_{j\geq 2}\left\|\mathbf{h}_{\mathcal{Q}_j}\right\|_2 \leq s^{-1/2}\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1,$$

which implies

$$\|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2 \leq \underbrace{\frac{2\sqrt{1+\delta_{2s}}}{1-\delta_{2s}}}_{=\alpha}\varepsilon + \underbrace{\frac{\sqrt{2}\delta_{2s}}{1-\delta_{2s}}}_{=\rho}s^{-1/2}\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1 = \alpha\varepsilon + \rho s^{-1/2}\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1. \tag{7.21}$$

Note that we can divide by $1-\delta_{2s}$ because $\delta_{2s} < 1$. Using

$$\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1 \leq \|\mathbf{h}_{\mathcal{Q}_0}\|_1 + \underbrace{2\left\|\mathbf{x}_{\mathcal{Q}_0^c}\right\|_1}_{=2\|\mathbf{x}-\mathbf{x}_s\|_1}, \tag{7.22}$$

which follows from (7.8), this becomes

$$\|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\|_2 \leq \alpha\varepsilon + \rho\underbrace{s^{-1/2}\|\mathbf{h}_{\mathcal{Q}_0}\|_1}_{\leq\left\|\mathbf{h}_{\mathcal{Q}_0\cup\mathcal{Q}_1}\right\|_2} + \underbrace{\rho s^{-1/2}2\|\mathbf{x}-\mathbf{x}_s\|_1}_{2\,\rho\,e_0}.$$

Here, we used $\|\mathbf{x}\|_1 \leq \sqrt{s}\,\|\mathbf{x}\|_2$, for $\mathbf{x} \in \mathbb{C}^s$, applied to $\mathbf{h}_{\mathcal{Q}_0}$ and combined the result thereof with $\|\mathbf{h}_{\mathcal{Q}_0}\|_2 \leq \|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2$. This yields

$$\|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2 \leq \alpha\varepsilon + \rho\|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2 + 2\rho\, e_0,$$

and therefore, as $\delta_{2s} < \sqrt{2} - 1$ by assumption, it holds that $\rho < 1$, and hence

$$\|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2 \leq \frac{\alpha\varepsilon + 2\rho e_0}{1-\rho}.$$

Finally, applying the triangle inequality to $\mathbf{h} = \mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1} + \mathbf{h}_{(\mathcal{Q}_0 \cup \mathcal{Q}_1)^c}$, we get

$$\begin{aligned}
\|\mathbf{h}\|_2 &\leq \|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2 + \left\|\mathbf{h}_{(\mathcal{Q}_0 \cup \mathcal{Q}_1)^c}\right\|_2 \\
&\overset{(a)}{\leq} \|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2 + \|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2 + 2\, e_0 \\
&= 2\,\|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2 + 2\, e_0 \leq 2\,\frac{\alpha\varepsilon + 2\,\rho\, e_0}{1-\rho} + 2\, e_0 \\
&= 2\,\frac{\alpha\varepsilon + 2\rho e_0 + e_o - e_0\rho}{1-\rho} = 2\,\frac{\alpha\varepsilon + e_0\rho + e_0}{1-\rho} \\
&= 2\,\frac{\alpha\varepsilon + e_0(1+\rho)}{1-\rho} = \underbrace{2\,\frac{\alpha\varepsilon}{1-\rho}}_{=C_1\varepsilon} + \underbrace{2\,\frac{1+\rho}{1-\rho}}_{=C_0}\, s^{-1/2}\|\mathbf{x} - \mathbf{x}_s\|_1,
\end{aligned}$$

where (a) follows from (7.13). This concludes the proof. □

We next state a result that will turn out useful later.

Lemma 7.5. *Let* $\mathbf{h}$ *be a vector in the null space of* $\mathbf{\Phi}$ *and let* $\mathcal{Q}_0$ *be any set of cardinality* s. *Then,*

$$\|\mathbf{h}_{\mathcal{Q}_0}\|_1 \leq \rho\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1,$$

with $\rho = \sqrt{2}\delta_{2s}/(1-\delta_{2s})$.

Proof. We have

$$\begin{aligned}
\|\mathbf{h}_{\mathcal{Q}_0}\|_1 \leq s^{1/2}\|\mathbf{h}_{\mathcal{Q}_0}\|_2 &\leq s^{1/2}\|\mathbf{h}_{\mathcal{Q}_0 \cup \mathcal{Q}_1}\|_2 \\
&\leq s^{1/2}(\rho s^{-1/2}\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1) \\
&= \rho\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1,
\end{aligned}$$

where we used (7.21) with $\varepsilon = 0$ as $\mathbf{h}$ is in the null space of $\mathbf{\Phi}$ and hence (7.15) is satisfied with $\varepsilon = 0$. □

We are now in a position to prove Theorem 7.2.

Proof of Theorem 7.2. The bound (7.3) follows by setting $\mathbf{h} = \mathbf{0}$ and $\varepsilon = 0$ in Theorem 7.3. To get the bound in (7.2), we first note that, by (7.11), we have

$$\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1 \leq 2\left\|\mathbf{x}_{\mathcal{Q}_0^c}\right\|_1 + \|\mathbf{h}_{\mathcal{Q}_0}\|_1. \tag{7.23}$$

Using the fact that $\mathbf{h}$ is in the null space of $\mathbf{\Phi}$, we can employ Lemma 7.5 in (7.23) to conclude that

$$\left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1 \leq \frac{2}{1-\rho}\|\mathbf{x} - \mathbf{x}_s\|_1. \tag{7.24}$$

The proof is finalized by noting that $\|\mathbf{h}\|_1 = \|\mathbf{h}_{\mathcal{Q}_0}\|_1 + \left\|\mathbf{h}_{\mathcal{Q}_0^c}\right\|_1$ and using Lemma 7.5 again combined with 7.24. □

We now interpret the RIP property and start by considering the case $s = 1$. Here,

$$(1-\delta_1)\|\mathbf{x}\|_2^2 \leq \|\mathbf{\Phi}\mathbf{x}\|_2^2 \leq (1+\delta_1)\|\mathbf{x}\|_2^2$$

for all 1-sparse vectors $\mathbf{x}$. Setting $\mathbf{x} = (0, \ldots, 0, \mathbf{x}_i, 0, \ldots, 0)^T$, this yields

$$(1-\delta_1)|\mathbf{x}_i|^2 \leq \|\mathbf{\Phi}_i\|_2^2|\mathbf{x}_i|^2 \leq (1+\delta_1)|\mathbf{x}_i|^2,$$

where $\mathbf{\Phi}_i$ denotes the i-th column of $\mathbf{\Phi}$. We hence have

$$(1-\delta_1) \leq \|\mathbf{\Phi}_i\|_2^2 \leq (1+\delta_1),$$

i.e., the individual columns of $\mathbf{\Phi}$ have near unit-norm.

In the case $s = 2$, the isometry constant δ_s equals the coherence of the dictionary $\mathbf{\Phi}$. To see this, first recall the definition of dictionary coherence: for $\mathbf{\Phi}$ with columns $\|\cdot\|_2$-normalized to 1, the coherence is defined as $\mu(\mathbf{\Phi}) := \max_{i\neq j} |\mathbf{\Phi}_i^H\mathbf{\Phi}_j|$. Next, we note that the RIP can equivalently be expressed as

$$-\delta_s\|\mathbf{x}\|_2^2 \leq \mathbf{x}^H(\mathbf{\Phi}^H\mathbf{\Phi} - \mathbf{I})\mathbf{x} \leq \delta_s\|\mathbf{x}\|_2^2, \quad \text{for all } s\text{-sparse } \mathbf{x}.$$

This implies that

$$\delta_s = \max_{\mathcal{S}:|\mathcal{S}|\leq s} \lambda_{\max}\left(\mathbf{\Phi}_{\mathcal{S}}^H\mathbf{\Phi}_{\mathcal{S}} - \mathbf{I}\right),$$

where $\mathbf{\Phi}_{\mathcal{S}}$ denotes the matrix obtained from $\mathbf{\Phi}$ by retaining the columns indexed by $\mathcal{S}$. Particularizing to $s = 2$, and considering two arbitrary columns of $\mathbf{\Phi}$ denoted as $\mathbf{\Phi}_i$ and $\mathbf{\Phi}_j$, we obtain

$$\begin{aligned}
&\lambda_{\max}\left(\begin{pmatrix} \|\mathbf{\Phi}_i\|_2^2 & \langle\mathbf{\Phi}_i, \mathbf{\Phi}_j\rangle^* \\ \langle\mathbf{\Phi}_i, \mathbf{\Phi}_j\rangle & \|\mathbf{\Phi}_j\|_2^2 \end{pmatrix} - \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right) \\
&= \lambda_{\max}\begin{pmatrix} 0 & \langle\mathbf{\Phi}_i, \mathbf{\Phi}_j\rangle^* \\ \langle\mathbf{\Phi}_i, \mathbf{\Phi}_j\rangle & 0 \end{pmatrix} \\
&= |\langle\mathbf{\Phi}_i, \mathbf{\Phi}_j\rangle|.
\end{aligned}$$

This establishes that $\delta_2 = \mu(\mathbf{\Phi})$.

Next, we relate the RIP property to the restricted null space property $P_1(\mathcal{S}, \mathbf{\Phi}) < 1/2$ in Theorem 3.4. We start by noting that the condition

$$P_1(\mathcal{S}, \mathbf{\Phi}) = \max_{\substack{\mathbf{x} \in \mathcal{N}(\mathbf{\Phi}) \\ \mathbf{x} \neq \mathbf{0}}} \frac{\sum_{r \in \mathcal{S}} |\mathbf{x}_r|}{\sum_r |\mathbf{x}_r|} < \frac{1}{2}$$

is equivalent to

$$\frac{\sum_{r \in \mathcal{S}} |\mathbf{x}_r|}{\sum_{r \in \mathcal{S}} |\mathbf{x}_r| + \sum_{r \in \mathcal{S}^c} |\mathbf{x}_r|} < \frac{1}{2}, \quad \forall \mathbf{x} \in \mathcal{N}(\mathbf{\Phi}), \mathbf{x} \neq \mathbf{0} \tag{7.25}$$

and hence

$$\begin{aligned} \sum_{r \in \mathcal{S}} |\mathbf{x}_r| &< \frac{1}{2} \sum_{r \in \mathcal{S}} |\mathbf{x}_r| + \frac{1}{2} \sum_{r \in \mathcal{S}^c} |\mathbf{x}_r| \\ \sum_{r \in \mathcal{S}} |\mathbf{x}_r| &< \sum_{r \in \mathcal{S}^c} |\mathbf{x}_r|, \quad \forall \mathbf{x} \in \mathcal{N}(\mathbf{\Phi}). \end{aligned}$$

Defining

$$C(\mathcal{S}) := \left\{ \mathbf{x} \in \mathbb{C}^n \;\middle|\; \sum_{r \in \mathcal{S}^c} |\mathbf{x}_r| \leq \sum_{r \in \mathcal{S}} |\mathbf{x}_r| \right\},$$

(7.25) hence says that (P1) provides perfect recovery if the following property is satisfied.

Definition 7.6 (Restricted null-space property)**.** The dictionary $\mathbf{\Phi}$ satisfies the restricted null space property with respect to $\mathcal{S}$ if $C(\mathcal{S}) \cap \mathcal{N}(\mathbf{\Phi}) = \{\mathbf{0}\}$.

The coherence-based recovery threshold for (P1) given by

$$|\mathcal{S}| \leq \frac{1}{2} \left(1 + \frac{1}{\mu(\mathbf{\Phi})} \right)$$

constitutes a sufficient condition for the restricted null space property. Note that this condition depends on the cardinality of $\mathcal{S}$ only. We proceed to showing that the RIP provides a more sophisticated sufficient condition for the restricted null space property to hold. Roughly speaking, it can be understood as a generalization of the pairwise incoherence condition, based on the conditioning of larger subsets of columns. Specifically, we start by noting that Lemma 7.5 constitutes a restricted null space property if $\rho < 1$, which is guaranteed for $\delta_{2s} < \sqrt{2} - 1$. We next present a result that is slightly weaker, but has a less elaborate proof than that of Lemma 7.5, which, in turn, is based heavily on the proof of Theorem 7.3. Yet, this proof contains all the key ideas underlying that of Lemma 7.5.

Theorem 7.7. *If the RIP constant of order $2s$ satisfies $\delta_{2s}(\mathbf{\Phi}) < 1/3$, then the restricted null space property holds for any subset $\mathcal{S}$ of cardinality $|\mathcal{S}| \leq s$.*

Proof. Let $\mathbf{h}$ be a vector in the null space of $\mathbf{\Phi}$, i.e., $\mathbf{\Phi h} = \mathbf{0}$. We decompose $\mathbf{h}$ as follows

$$\mathbf{h} = \mathbf{h}_{\mathcal{Q}_0} + \sum_{j\geq 1} \mathbf{h}_{\mathcal{Q}_j},$$

where $\mathcal{Q}_0$ denotes the support set of the s entries of $\mathbf{h}$ that are largest in absolute value. We write $\mathcal{Q}_0^c = \bigcup\limits_{j\geq 1} \mathcal{Q}_j$, where $\mathcal{Q}_1$ is the subset of indices corresponding to the s largest entries of $\mathbf{h}_{\mathcal{Q}_0^c}$ (here $\mathbf{h}_{\mathcal{Q}_0^c}$ denotes the vector that equals $\mathbf{h}$ on $\mathcal{Q}_0^c$ and has zero entries else). The subset $\mathcal{Q}_2$ is the set corresponding to the s largest entries in the set $\mathcal{Q}_0^c \setminus \mathcal{Q}_1$ and so on. Note that the last set may contain fewer than s entries. We need to establish that

$$\|\mathbf{h}_{\mathcal{Q}_0}\|_1 < \|\mathbf{h}_{\mathcal{Q}_0^c}\|_1, \tag{7.26}$$

which implies

$$\|\mathbf{h}_{\mathcal{S}}\|_1 < \|\mathbf{h}_{\mathcal{S}^c}\|_1, \quad |\mathcal{S}| \leq s, \tag{7.27}$$

as $\mathcal{Q}_0$ was chosen to contain the s largest entries of $\mathbf{h}$. We proceed by noting that $\mathbf{\Phi h} = \mathbf{0}$ implies

$$\mathbf{\Phi h}_{\mathcal{Q}_0} = -\sum_{j\geq 1} \mathbf{\Phi h}_{\mathcal{Q}_j}.$$

By the RIP property for $\mathbf{h}_{\mathcal{Q}_0}$ (note that $\mathbf{h}_{\mathcal{Q}_0}$ is s-sparse and the RIP property is assumed to hold for $2s$-sparse vectors), we get

$$\begin{aligned}(1-\delta_{2s})\|\mathbf{h}_{\mathcal{Q}_0}\|_2^2 &\leq \langle \mathbf{\Phi h}_{\mathcal{Q}_0}, \mathbf{\Phi h}_{\mathcal{Q}_0}\rangle \\ &= \Big|\Big\langle \mathbf{\Phi h}_{\mathcal{Q}_0}, -\sum_{j\geq 1} \mathbf{\Phi h}_{\mathcal{Q}_j}\Big\rangle\Big| \\ &\leq \sum_{j\geq 1} |\langle \mathbf{\Phi h}_{\mathcal{Q}_0}, \mathbf{\Phi h}_{\mathcal{Q}_j}\rangle| \\ &\overset{(a)}{\leq} \delta_{2s}\|\mathbf{h}_{\mathcal{Q}_0}\|_2 \sum_{j\geq 1} \|\mathbf{h}_{\mathcal{Q}_j}\|_2\end{aligned}$$

and hence

$$\|\mathbf{h}_{\mathcal{Q}_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{j\geq 1} \|\mathbf{h}_{\mathcal{Q}_j}\|_2. \tag{7.28}$$

Here, we used Lemma 7.4 in (a). Next, we note that, thanks to the basic inequality

$$\|\mathbf{x}\|_2 \leq \sqrt{s}\|\mathbf{x}\|_\infty, \quad \mathbf{x} \in \mathbb{C}^s,$$

and the definition of the sets $\mathcal{Q}_j$, we get

$$\begin{aligned}\|\mathbf{h}_{\mathcal{Q}_j}\|_2 &\leq \sqrt{s}\|\mathbf{h}_{\mathcal{Q}_j}\|_\infty \leq \sqrt{s}\,\frac{1}{s}\|\mathbf{h}_{\mathcal{Q}_{j-1}}\|_1 \\ &= \frac{1}{\sqrt{s}}\|\mathbf{h}_{\mathcal{Q}_{j-1}}\|_1, \quad j = 1, 2, \dots\ .\end{aligned}$$

Using this in (7.28) yields

$$\|\mathbf{h}_{\mathcal{Q}_0}\|_2 \leq \frac{\delta_{2s}}{1-\delta_{2s}} \frac{1}{\sqrt{s}} \sum_{j\geq 1} \|\mathbf{h}_{\mathcal{Q}_{j-1}}\|_1$$

and by the basic inequality

$$\|\mathbf{x}\|_1 \leq \sqrt{s}\|\mathbf{x}\|_2, \quad \mathbf{x} \in \mathbb{C}^s,$$

this results in

$$\begin{aligned}\|\mathbf{h}_{\mathcal{Q}_0}\|_1 &\leq \frac{\delta_{2s}}{1-\delta_{2s}} \sum_{j\geq 1} \|\mathbf{h}_{\mathcal{Q}_{j-1}}\|_1 \\ &= \frac{\delta_{2s}}{1-\delta_{2s}} \left(\|\mathbf{h}_{\mathcal{Q}_0}\|_1 + \|\mathbf{h}_{\mathcal{Q}_0^c}\|_1\right).\end{aligned}$$

Rearranging terms yields

$$\begin{aligned}\|\mathbf{h}_{\mathcal{Q}_0}\|_1 \left(1 - \frac{\delta_{2s}}{1-\delta_{2s}}\right) &\leq \frac{\delta_{2s}}{1-\delta_{2s}} \|\mathbf{h}_{\mathcal{Q}_0^c}\|_1 \\ \|\mathbf{h}_{\mathcal{Q}_0}\|_1 &\leq \frac{1}{1-\frac{\delta_{2s}}{1-\delta_{2s}}} \frac{\delta_{2s}}{1-\delta_{2s}} \|\mathbf{h}_{\mathcal{Q}_0^c}\|_1 \\ &= \frac{\delta_{2s}}{1-2\delta_{2s}} \|\mathbf{h}_{\mathcal{Q}_0^c}\|_1.\end{aligned}$$

Finally, thanks to $\delta_{2s} < 1/3$, it follows that

$$\frac{\delta_{2s}}{1-2\delta_{2s}} < 1,$$

which establishes the desired result.

□

Chapter 8

The Johnson-Lindenstrauss Lemma

Suppose we are given a set of $\mathcal{U}$ of m points in $\mathbb{R}^n$. We would like to embed these points into a lower dimensional Euclidean space (i.e., in $\mathbb{R}^k$ with $k < n$), while approximately preserving the distances between the points in $\mathcal{U}$. The Johnson-Lindenstrauss (JL) Lemma, stated below, shows that any set of m points can be embedded in $k = \mathcal{O}(\log(m)/\epsilon^2)$ dimensions while the distances between any two points change by at most a factor of $1 \pm \epsilon$. The JL Lemma, in particular the concentration of measure inequality from which the JL Lemma follows (as shown later), will turn out to be an essential ingredient in proving the restricted isometry property (RIP) for random matrices. As a reference for these notes, see [80, 81].

Lemma 8.1 (Johnson-Lindenstrauss Lemma)**.** *Choose ϵ with $0 < \epsilon < 1$ and suppose k satisfies*

$$k \geq \frac{8}{\epsilon^2 - \epsilon^3} \log(2m). \tag{8.1}$$

Then, for every set $\mathcal{U}$ of m points, there exists a map $f\colon \mathbb{R}^n \to \mathbb{R}^k$ such that for all $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$, we have

$$(1-\epsilon)\|\mathbf{u} - \mathbf{u}'\|^2 \leq \|f(\mathbf{u}) - f(\mathbf{u}')\|^2 \leq (1+\epsilon)\|\mathbf{u} - \mathbf{u}'\|^2. \tag{8.2}$$

The JL Lemma is essentially tight according to [82, Thm. 9.3]. The original proof of the JL Lemma, as well as the proof discussed here, is based on random projections. The JL Lemma will follow directly from the following concentration inequality. This concentration inequality will be an essential ingredient for verifying the RIP for random matrices.

Lemma 8.2. *Let $\mathbf{A} \in \mathbb{R}^{k \times n}$ be a random matrix with i.i.d. $\mathcal{N}(0, 1/k)$ entries. Then, for ϵ with $0 < \epsilon < 1$ and fixed $\mathbf{u} \in \mathbb{R}^n$,*

$$\mathbb{P}\left(\left|\|\mathbf{Au}\|^2 - \mathbb{E}\left[\|\mathbf{Au}\|^2\right]\right| \geq \epsilon\|\mathbf{u}\|^2\right) < 2\, e^{-k\frac{\epsilon^2-\epsilon^3}{4}} \tag{8.3}$$

with

$$\mathbb{E}\left[\|\mathbf{Au}\|^2\right] = \|\mathbf{u}\|^2. \tag{8.4}$$

In words, Lemma 8.2 states that the random variable $\|\mathbf{Au}\|^2$ is concentrated around its mean. A relation of the form (8.3) is called a "concentration of measure inequality" or simply "concentration inequality" in the literature. Note that Lemma 8.2 is not restricted to Gaussian random matrices, but generalizes to other random matrices. E.g., essentially the same inequality holds if each entry $a_{i,j}$ of $\mathbf{A}$ is i.i.d. sub-Gaussian, i.e., its tail probability satisfies $\mathbb{P}(|a_{i,j}| > t) \leq c_1 e^{-c_2 t^2}$ for constants c_1, c_2.

Before proving Lemma 8.2, we will show how it implies the JL Lemma.

Proof of the JL Lemma. The proof is effected by showing that the (linear) map $f(\mathbf{u}) = \mathbf{Au}$ with $\mathbf{A} \in \mathbb{R}^{k\times n}$ a random matrix with i.i.d. $\mathcal{N}(0, 1/k)$ entries, satisfies (8.2) for all $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$ with non-zero probability.

Applying the union bound over all $m(m-1)/2 < m^2$ pairs of points in $\mathcal{U}$, it follows from Lemma 8.2 that (8.2) is violated for one or more pairs of points $(\mathbf{u}, \mathbf{u}')$ with $\mathbf{u}, \mathbf{u}' \in \mathcal{U}$ with probability less than $m^2\, 2\, e^{-k\frac{\epsilon^2-\epsilon^3}{4}}$. The proof is concluded by showing that $m^2\, 2\, e^{-k\frac{\epsilon^2-\epsilon^3}{4}} \leq 1/2$ is implied by (8.1), since this ensures that $f(\mathbf{u}) = \mathbf{Au}$ satisfies (8.2) with probability at least $1/2$:

$$m^2\, 2\, e^{-k\frac{\epsilon^2-\epsilon^3}{4}} \leq 1/2 \Leftrightarrow -k\frac{\epsilon^2-\epsilon^3}{4} \leq \log(1/(4m^2)) \Leftrightarrow k \geq \frac{4}{\epsilon^2-\epsilon^3}\, 2\log(2m).$$

□

Proof of Lemma 8.2. First observe that

$$\mathbb{E}\left[\|\mathbf{Au}\|^2\right] = \mathbb{E}\left[\mathbf{u}^T\mathbf{A}^T\mathbf{Au}\right] = \mathbf{u}^T\mathbb{E}\left[\mathbf{A}^T\mathbf{A}\right]\mathbf{u} = \mathbf{u}^T\mathbf{I}\mathbf{u} = \|\mathbf{u}\|^2$$

which proves (8.4).

Next, let $\mathbf{a}_j^T$ be the j-th row of $\mathbf{A}$, and set $X_j := \frac{\sqrt{k}}{\|\mathbf{u}\|}\mathbf{a}_j^T\mathbf{u}$. Note that $\mathbf{a}_j^T\mathbf{u}$ is the sum of independent Gaussians and is therefore $\mathcal{N}(0, \|\mathbf{u}\|^2/k)$ distributed. It follows that the X_j are i.i.d. $\mathcal{N}(0,1)$ distributed. Next, set $X = \sum_{j=1}^k X_j^2$. With this notation, we have

$$X = \sum_{j=1}^k X_j^2 = \frac{k}{\|\mathbf{u}\|^2}\sum_{j=1}^k \left|\mathbf{a}_j^T\mathbf{u}\right|^2 = \frac{k}{\|\mathbf{u}\|^2}\|\mathbf{Au}\|^2.$$

Thus, for $\lambda \geq 0$,

$$\begin{aligned}
\mathbb{P}\left(\|\mathbf{Au}\|^2 \geq (1+\epsilon)\|\mathbf{u}\|^2\right) &= \mathbb{P}\left(X \geq (1+\epsilon)k\right) \\
&= \mathbb{P}\left(e^{\lambda X} \geq e^{\lambda(1+\epsilon)k}\right) \\
&\leq \frac{1}{e^{(1+\epsilon)k\lambda}}\mathbb{E}\left[e^{\lambda X}\right] && (8.5) \\
&= \frac{1}{e^{(1+\epsilon)k\lambda}}\prod_{j=1}^k \mathbb{E}\left[e^{\lambda X_j^2}\right] && (8.6) \\
&= \frac{1}{e^{(1+\epsilon)k\lambda}}\left(\mathbb{E}\left[e^{\lambda X_1^2}\right]\right)^k, && (8.7)
\end{aligned}$$

where we used Markov's inequality for a nonnegative random variable in (8.5), independence of the X_j for (8.6) and that all X_j have the same distribution for (8.7).

It remains to evaluate the moment generating function $\mathbb{E}\left[e^{\lambda X_1^2}\right]$. Since X_1 is $\mathcal{N}(0,1)$ distributed,

$$\begin{aligned}\mathbb{E}\left[e^{\lambda X_1^2}\right] &= \int_{-\infty}^{\infty} e^{\lambda x^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} dx \\ &= \frac{1}{\sqrt{1-2\lambda}} \int_{-\infty}^{\infty} \frac{\sqrt{1-2\lambda}}{\sqrt{2\pi}} e^{-\frac{x^2}{2}(1-2\lambda)} dx \\ &= \frac{1}{\sqrt{1-2\lambda}}, \end{aligned} \tag{8.8}$$

where we used that the integrand is the normal density with standard deviation $\frac{1}{\sqrt{1-2\lambda}}$. The conclusion above holds for any $\lambda < 1/2$. Using (8.8) in (8.7) yields

$$\mathbb{P}\left(\|\mathbf{Au}\|^2 \geq (1+\epsilon)\|\mathbf{u}\|^2\right) \leq \left(\frac{e^{-2(1+\epsilon)\lambda}}{1-2\lambda}\right)^{\frac{k}{2}}. \tag{8.9}$$

We next minimize the right hand side (RHS) of (8.9). To this end, we choose λ such that the term $\frac{e^{-2(1+\epsilon)\lambda}}{1-2\lambda}$ is minimal. It is easily verified (by setting the derivative to zero) that the optimal choice is $\lambda = \frac{\epsilon}{2(1+\epsilon)}$. With this choice,

$$\mathbb{P}\left(\|\mathbf{Au}\|^2 \geq (1+\epsilon)\|\mathbf{u}\|^2\right) \leq \left((1+\epsilon)\, e^{-\epsilon}\right)^{\frac{k}{2}} < e^{-(\epsilon^2-\epsilon^3)\frac{k}{4}} \tag{8.10}$$

where for the last inequality we used

$$1+\epsilon < e^{\epsilon - \frac{\epsilon^2-\epsilon^3}{2}}$$

which is a consequence of the Taylor series expansion of $\exp(\cdot)$.

Similarly, we obtain

$$\mathbb{P}\left(\|\mathbf{Au}\|^2 \leq (1-\epsilon)\|\mathbf{u}\|^2\right) < e^{-(\epsilon^2-\epsilon^3)\frac{k}{4}}. \tag{8.11}$$

Combining (8.10) and (8.11) via the union bound concludes the proof. □

Chapter 9

Verifying the RIP through the JL Lemma

We show how to prove the RIP for random matrices. Given $\mathcal{S}$ with $|\mathcal{S}| \leq k$, denote by $\mathcal{X}_\mathcal{S}$ the set of all vectors in $\mathbb{R}^n$ that are zero outside $\mathcal{S}$. This is a k-dimensional linear space.

Our general approach will be to construct nets of points in each k-dimensional subspace, then apply the concentration inequality to all of these points, through a union bound, and then extend the result from our finite set of points to all possible k-dimensional signals.

Lemma 9.1. *Let $\mathbf{\Phi} \in \mathbb{R}^{m \times n}$ be a random matrix with i.i.d. $\mathcal{N}(0, 1/m)$ entries. Then, for any set $\mathcal{S}$ with $|\mathcal{S}| = k < m$ and any $0 < \delta < 1$, we have*

$$(1 - \delta)\|\mathbf{x}\| \leq \|\mathbf{\Phi x}\| \leq (1 + \delta)\|\mathbf{x}\|, \quad \forall \mathbf{x} \in \mathcal{X}_\mathcal{S} \tag{9.1}$$

with probability

$$\geq 1 - 2(12/\delta)^k e^{-c_0(\delta/2)m},$$

where $c_0(x) = \frac{1}{4}(x^2 - x^3)$.

Proof. It suffices to prove (9.1) for $\|\mathbf{x}\| = 1$ as $\mathbf{\Phi x}$ is a linear map. Choose a finite set of points $\mathcal{Q}_\mathcal{S}$ such that i) $\mathcal{Q}_\mathcal{S} \subseteq \mathcal{X}_\mathcal{S}$, $\|\mathbf{q}\| = 1$ for all $\mathbf{q} \in \mathcal{Q}_\mathcal{S}$, and ii) for all $\mathbf{x} \in \mathcal{X}_\mathcal{S}$ with $\|\mathbf{x}\| = 1$, we have

$$\min_{\mathbf{q} \in \mathcal{Q}_\mathcal{S}} \|\mathbf{x} - \mathbf{q}\| \leq \delta/4.$$

From the theory of covering numbers, it is known that we can choose such a set $\mathcal{Q}_\mathcal{S}$ with $|\mathcal{Q}_\mathcal{S}| \leq (12/\delta)^k$. We use the union bound to apply Lemma 8.2 to this set of points with $\varepsilon = \delta/2$, which yields that

$$(1 - \delta/2)\|\mathbf{q}\|^2 \leq \|\mathbf{\Phi q}\|^2 \leq (1 + \delta/2)\|\mathbf{q}\|^2, \quad \forall \mathbf{q} \in \mathcal{Q}_\mathcal{S}$$

holds with probability

$$\geq 1 - 2\,(12/\delta)^k e^{-c_0(\delta/2)m},$$

where

$$c_0(x) = \frac{x^2 - x^3}{4}.$$

We next define A as the smallest number such that

$$\|\mathbf{\Phi x}\| \leq (1+A)\|\mathbf{x}\|, \quad \forall \mathbf{x} \in \mathcal{X}_{\mathcal{S}}.$$

Our goal is to show that $A \leq \delta$. To this end, recall that for any $\mathbf{x} \in \mathcal{X}_{\mathcal{S}}$ with $\|\mathbf{x}\| = 1$, we can find a $\mathbf{q} \in \mathcal{Q}_{\mathcal{S}}$ such that $\|\mathbf{x} - \mathbf{q}\| \leq \delta/4$. Hence, we have

$$\|\mathbf{\Phi x}\| \leq \|\mathbf{\Phi q}\| + \|\mathbf{\Phi}(\mathbf{x} - \mathbf{q})\| \leq 1 + \delta/2 + (1+A)\delta/4,$$

given that $\|\mathbf{q}\| = 1$. Since by definition, A is the smallest number for which $\|\mathbf{\Phi x}\| \leq (1+A)\|\mathbf{x}\|$, we have

$$\begin{aligned} A &\leq \delta/2 + (1+A)\delta/4 \\ A(1-\delta/4) &\leq \delta/2 + \delta/4 \\ A &\leq \frac{\delta/2 + \delta/4}{1-\delta/4} = \frac{2\delta+\delta}{4-\delta} \leq \frac{3\delta}{3} = \delta \end{aligned}$$

as desired. We therefore proved that

$$\|\mathbf{\Phi x}\| \leq (1+\delta)\|\mathbf{x}\|.$$

The inequality $\|\mathbf{\Phi x}\| \geq (1-\delta)\|\mathbf{x}\|$ follows since

$$\begin{aligned} \|\mathbf{\Phi x}\| \geq \|\mathbf{\Phi q}\| - \|\mathbf{\Phi}(\mathbf{x} - \mathbf{q})\| &\geq (1-\delta/2) - (1+\delta)\delta/4 \\ &= 1 - \delta/2 - \delta/4 - \delta^2/4 \\ &\geq 1 - \delta/2 - \delta/4 - \delta/4 = 1 - \delta, \end{aligned}$$

which completes the proof.

□

Theorem 9.2. *Suppose that m, n, and $0 < \delta < 1$ are given. If the pdf generating $\mathbf{\Phi}$ satisfies the concentration inequality in Lemma 8.2, then there exist constants $c_1, c_2 > 0$ depending only on δ such that the RIP holds for $\mathbf{\Phi}$ with the prescribed δ and any $k \leq c_1 m / \log(n/k)$ with probability $\geq 1 - 2e^{-c_2 m}$.*

Proof. We know that for each of the k-dimensional spaces $\mathcal{X}_{\mathcal{S}}$, the matrix $\mathbf{\Phi}$ will fail to satisfy

$$(1-\delta)\|\mathbf{x}\| \leq \|\mathbf{\Phi x}\| \leq (1+\delta)\|\mathbf{x}\|, \quad \forall \mathbf{x} \in \mathcal{X}_{\mathcal{S}} \tag{9.2}$$

with probability

$$\leq 2(12/\delta)^k e^{-c_0(\delta/2)m}. \tag{9.3}$$

There are $\binom{n}{k} \leq (en/k)^k$ such subspaces. Hence, by a union bound argument, (9.2) will fail to hold with probability

$$\leq 2(en/k)^k (12/\delta)^k e^{-c_0(\delta/2)m} = 2e^{-c_o(\delta/2)m + k[\log(en/k) + \log(12/\delta)]}.$$

We conclude the proof by showing that, for a fixed $c_1 > 0$, whenever

$$k \leq \frac{c_1 m}{\log(n/k)},$$

we will have that the exponent in (9.3) is smaller than $-c_2 m$, provided that

$$c_2 \leq c_0(\delta/2) - c_1 \left(1 + \frac{1 + \log(12/\delta)}{\log(n/k)}\right). \tag{9.4}$$

Noting that we can always choose $c_1 > 0$ sufficiently small to ensure $c_2 > 0$, finalizes the proof. To establish (9.4), we note that

$$e^{-c_o(\delta/2)m + k[\log(en/k) + \log(12/\delta)]} \leq e^{-c_2 m},$$

yields, under the assumption $k \leq \frac{c_1 m}{\log(n/k)}$, that

$$e^{-m\left(c_0(\delta/2) - c_1 \frac{\log(en/k) + \log(12/\delta)}{\log(n/k)}\right)} \leq e^{-c_2 m}$$

and hence

$$c_2 \leq c_0(\delta/2) - c_1 \left(1 + \frac{1 + \log(12/\delta)}{\log(n/k)}\right).$$

□

Chapter 10

Approximation Theory

10.1 Min-Max (Kolmogorov) Rate Distortion Theory

Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and consider the function class $\mathcal{C} \subset L^2(\Omega)$. Then, for each $\ell \in \mathbb{N}$, we denote by

$$\mathfrak{E}^\ell := \left\{ E : \mathcal{C} \to \{0,1\}^\ell \right\}$$

the set of *binary encoders of* $\mathcal{C}$ *of length* ℓ, and we let

$$\mathfrak{D}^\ell := \left\{ D : \{0,1\}^\ell \to L^2(\Omega) \right\}$$

be the set of *binary decoders of length* ℓ. An encoder-decoder pair $(E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell$ is said to *achieve uniform error* ε *over the function class* $\mathcal{C}$, if

$$\sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon.$$

A quantity of central interest is the minimal length $\ell \in \mathbb{N}$ for which there exists an encoder-decoder pair $(E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell$ that achieves uniform error ε over the function class $\mathcal{C}$, along with its asymptotic behavior as made precise in the following definition.

Definition 10.1. Let $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, and $\mathcal{C} \subset L^2(\Omega)$. Then, for $\varepsilon > 0$, the *minimax code length* $L(\varepsilon, \mathcal{C})$ is

$$L(\varepsilon, \mathcal{C}) := \min \left\{ \ell \in \mathbb{N} : \exists (E, D) \in \mathfrak{E}^\ell \times \mathfrak{D}^\ell : \sup_{f \in \mathcal{C}} \|D(E(f)) - f\|_{L^2(\Omega)} \leq \varepsilon \right\}.$$

Moreover, the *optimal exponent* $\gamma^*(\mathcal{C})$ is defined as

$$\gamma^*(\mathcal{C}) := \sup \left\{ \gamma \in \mathbb{R} : L(\varepsilon, \mathcal{C}) \in \mathcal{O}\big(\varepsilon^{-1/\gamma}\big),\ \varepsilon \to 0 \right\}. \tag{10.1}$$

The optimal exponent $\gamma^*(\mathcal{C})$ determines the minimum growth rate of $L(\varepsilon, \mathcal{C})$ as the error ε tends to zero and can hence be seen as quantifying the "description complexity" of the function class $\mathcal{C}$. Larger $\gamma^*(\mathcal{C})$ results in smaller growth rate and hence smaller memory requirements for storing signals $f \in \mathcal{C}$ such that reconstruction with uniformly bounded error is possible. The quantity $\gamma^*(\mathcal{C})$ is closely related to the concept of Kolmogorov-Tikhomirov entropy a.k.a. metric entropy [83]. This connection will be made explicit later in the context of optimal approximation in dictionaries.

10.2 Metric Entropy, Covering, and Packing

The discussion in this subsection is largely adopted from [84].

We will need the notions of covering and packing in metric spaces. Let $(\mathcal{X}, \rho)$ be a metric space. Recall that a metric space consists of a non-empty set $\mathcal{X}$ and a distance function $\rho : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ satisfying the following properties:

- Non-negativity: $\rho(x, x') \geq 0$, for all x, x', with equality iff $x = x'$
- Symmetry: $\rho(x, x') = \rho(x', x)$, for all x, x'
- Triangle inequality: $\rho(x, \tilde{x}) \leq \rho(x, x') + \rho(x', \tilde{x})$, for all $x, x', \tilde{x}$.

Examples of metric spaces include $\mathbb{R}^d$ with the Euclidean metric

$$\rho(x, x') = \sqrt{\sum_{j=1}^{d} (x_j - x'_j)^2}$$

and the discrete cube $\{0,1\}^d$ with the normalized Hamming metric

$$\rho_H(x, x') = \frac{1}{d} \sum_{j=1}^{d} I(x_j \neq x'_j).$$

We shall also consider metric spaces of functions, such as $L^2([0,1])$ equipped with

$$\|f - g\|_2 = \left[\int_0^1 (f(x) - g(x))^2 dx \right]^{1/2},$$

as well as the space $C([0,1])$ of continuous functions on $[0,1]$ equipped with the sup-norm metric

$$\|f - g\|_\infty = \text{sup}_{x \in [0,1]} |f(x) - g(x)|.$$

Given a metric space $(\mathcal{X}, \rho)$, a natural way of measuring the size of a compact subset $\mathcal{C}$ of $\mathcal{X}$ is in terms of the number of balls of a fixed radius ϵ required to cover $\mathcal{C}$, a quantity known as the covering number.

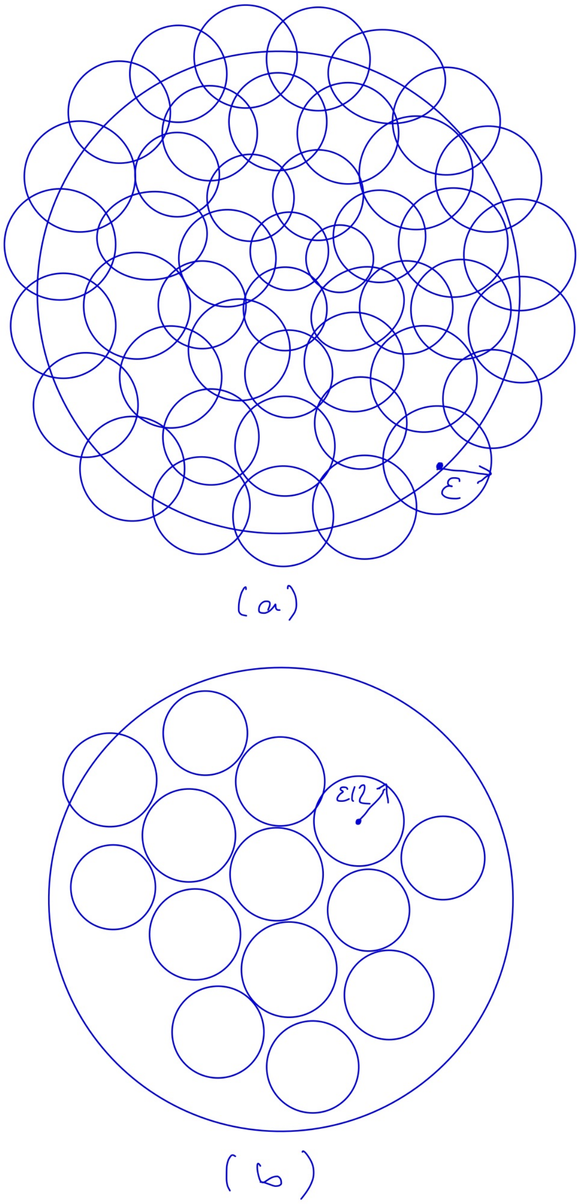


Figure 10.1: Illustration of covering and packing sets. (a) An ϵ-covering, the union of the balls covers the set. (b) An ϵ-packing is a collection of balls of radius $\epsilon/2$ with centers inside the set to be covered and such that no pair of balls has a non-empty intersection.

Definition 10.2. Let $(\mathcal{X}, \rho)$ be a metric space. An ϵ-covering of a compact set $\mathcal{C} \subseteq \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, ..., x_N\} \subset \mathcal{C}$ such that for each $x \in \mathcal{C}$, there exists an $i \in [1, N]$ such that $\rho(x, x_i) \leq \epsilon$. The ϵ-covering number $N(\epsilon; \mathcal{C}, \rho)$ is the cardinality of the smallest ϵ-covering.

As illustrated in Fig. 10.1, an ϵ-covering can be visualized as a collection of balls of radius ϵ that cover the set $\mathcal{C}$, i.e.,

$$\mathcal{C} \subset \bigcup_{i=1}^{N} B(x_i, \epsilon),$$

where $B(x_i, \epsilon)$ is a ball—in the metric ρ—of radius ϵ centered at x_i. The covering number is non-increasing in ϵ, i.e., $N(\epsilon) \geq N(\epsilon')$, for all $\epsilon \leq \epsilon'$. When the set $\mathcal{C}$ is not finite, the covering number goes to infinity as ϵ goes to zero. We shall be interested in the corresponding rate of growth, more specifically in the quantity $\log_2 N(\epsilon; \mathcal{C}, \rho)$ (in bits) known as the metric entropy of $\mathcal{C}$ with respect to ρ. The operational significance of metric entropy follows from the question: What is the minimum number of bits needed to represent any element $x \in \mathcal{C}$ with error—quantified in terms of the distance measure ρ—of at most ϵ? By what was just developed, the answer to this question is $\lceil \log_2 N(\epsilon; \mathcal{C}, \rho) \rceil$. Specifically, for a given $x \in \mathcal{X}$, the corresponding encoder $E(x)$ simply identifies the closest ball center x_i and encodes the index i using $\lceil \log_2 N(\epsilon; \mathcal{C}, \rho) \rceil$ bits. The corresponding decoder D delivers the ball center x_i, which guarantees that the resulting error satisfies $\|D(E(x)) - x\| \leq \epsilon$.

We proceed with a simple example computing an upper bound on the metric entropy of the interval $\mathcal{C} = [-1, 1]$ in $\mathbb{R}$ with respect to the metric $\rho(x, x') = |x - x'|$. To this end, we divide $\mathcal{C}$ up into intervals of length 2ϵ by setting $x_i = -1 + 2(i-1)\epsilon$, for $i \in [1, L]$, where $L = \lfloor \frac{1}{\epsilon} \rfloor + 1$. This guarantees that, for every point $x \in [-1, 1]$, there is an $i \in [1, L]$ such that $|x - x_i| \leq \epsilon$, which, in turn, establishes

$$N(\epsilon; \mathcal{C}, \rho) \leq \left\lfloor \frac{1}{\epsilon} \right\rfloor + 1 \leq \frac{1}{\epsilon} + 1$$

and hence yields an upper bound on metric entropy according to

$$\log_2 N(\epsilon; \mathcal{C}, \rho) \leq \log_2 \left(\frac{1}{\epsilon} + 1 \right) \asymp \log_2(\epsilon^{-1}). \tag{10.2}$$

This result can be generalized to the d-dimensional unit cube to yield $\log_2(N(\epsilon; \mathcal{C}, \rho)) \leq d \log_2(1/\epsilon + 1) \asymp d \log_2(\epsilon^{-1})$. In order to show that the upper bound (10.2) correctly reflects metric entropy scaling for $\mathcal{C} = [-1, 1]$, we would need a lower bound on $N(\epsilon; \mathcal{C}, \rho)$ that exhibits the same scaling (in ϵ) behavior. A systematic approach to establishing lower bounds on metric entropy is through the concept of packing, which will be introduced next.

We start with the definition of the packing number of a compact set $\mathcal{C}$ in a metric space $(\mathcal{X}, \rho)$.

Definition 10.3. Let $(\mathcal{X}, \rho)$ be a metric space. An ϵ-packing of a compact set $\mathcal{C} \subset \mathcal{X}$ with respect to the metric ρ is a set $\{x_1, ..., x_N\} \subset \mathcal{C}$ such that $\rho(x_i, x_j) > \epsilon$, for all distinct i, j. The ϵ-packing number $M(\epsilon; \mathcal{X}, \rho)$ is the cardinality of the largest ϵ-packing.

As illustrated in Fig. 10.1, an ϵ-packing can be viewed as a collection of balls of radius $\epsilon/2$, each centered at an element in $\mathcal{X}$, such that no two balls intersect. Although different, the covering number and the packing number provide essentially the same measure of massiveness of a set as formalized in the next result.

Lemma 10.4. *Let $(\mathcal{X}, \rho)$ be a metric space and $\mathcal{C}$ a compact set in $\mathcal{X}$. For all $\epsilon > 0$, the packing and the covering number are related according to*

$$M(2\epsilon; \mathcal{C}, \rho) \leq N(\epsilon; \mathcal{C}, \rho) \leq M(\epsilon; \mathcal{C}, \rho).$$

Proof. First, choose a minimal ϵ-covering and a maximal 2ϵ-packing of $\mathcal{C}$. Since no two centers of the 2ϵ-packing can lie in the same ball of the ϵ-covering, it follows that $M(2\epsilon; \mathcal{C}, \rho) \leq N(\epsilon; \mathcal{C}, \rho)$. To establish $N(\epsilon; \mathcal{C}, \rho) \leq M(\epsilon; \mathcal{C}, \rho)$, we note that, given a maximum packing $M(\epsilon; \mathcal{C}, \rho)$, for any $x \in \mathcal{C}$, we have the center of at least one of the balls in the packing within distance less than ϵ. If this were not the case, we could add another ball to the packing thereby violating its maximality. This maximal packing hence also provides an ϵ-covering and since $N(\epsilon; \mathcal{C}, \rho)$ is a minimal covering, we must have $N(\epsilon; \mathcal{C}, \rho) \leq M(\epsilon; \mathcal{C}, \rho)$. □

The result we just established shows that the scaling behavior, as $\epsilon \to 0$, of the covering number and the packing number must be identical up to constant factors. We now return to the example in which we computed an upper bound on the metric entropy of $\mathcal{C} = [-1, 1]$ and show how Lemma 10.4 can be employed to establish the scaling behavior of metric entropy. To this end, we simply note that the points $x_i = -1 + 2(i-1)\epsilon$, $i \in [1, L]$, are separated according to $|x_i - x_j| = 2\epsilon > \epsilon$, for all $i \neq j$, which implies that $M(\epsilon; \mathcal{C}, |\cdot|) \geq L = \lfloor 1/\epsilon \rfloor + 1 \geq \frac{1}{\epsilon}$. Combining this with the upper bound (10.2) and Lemma 10.4, we obtain $\log_2 N(\epsilon; \mathcal{C}, |\cdot|) \asymp \log_2(\epsilon^{-1})$. Likewise, it can be established that $\log_2 N(\epsilon; \mathcal{C}, |\cdot|) \asymp d \log_2(\epsilon^{-1})$ for the d-dimensional unit cube. This illustrates how an explicit construction of a packing set can be used to determine the scaling behavior of metric entropy.

We now seek a more general understanding of the geometrical properties that govern metric entropy. Specifically, we expect to see connections between covering numbers and the volume of the corresponding covering balls. The following result provides a precise statement for closed unit balls in $\mathcal{X} = \mathbb{R}^d$ and with respect to the norms

$$\|x\|_q = \begin{cases} \left(\sum_{i=1}^d |x_i|^q\right)^{1/q}, & q \in [1, \infty) \\ \max_{i=1,\ldots,d} |x_i|, & q = \infty. \end{cases}$$

The following lemma relates the so-called volume ratio to metric entropy.

Lemma 10.5. *Consider a pair of norms $\|\cdot\|$ and $\|\cdot\|'$ on $\mathbb{R}^d$, and let $\mathcal{B}$ and $\mathcal{B}'$ be their corresponding unit balls, i.e., $\mathcal{B} = \{x \in \mathbb{R}^d | \|x\| \leq 1\}$ and $\mathcal{B}' = \{x \in \mathbb{R}^d | \|x\|' \leq 1\}$. Then, the ϵ-covering number of $\mathcal{B}$ in the $\|\cdot\|'$-norm satisfies*

$$\left(\frac{1}{\epsilon}\right)^d \frac{vol(\mathcal{B})}{vol(\mathcal{B}')} \leq N(\epsilon; \mathcal{B}, \|\cdot\|') \leq \frac{vol(\frac{2}{\epsilon}\mathcal{B} + \mathcal{B}')}{vol(\mathcal{B}')}. \tag{10.3}$$

Proof. Let $\{x_1, ..., x_{N(\epsilon;\mathcal{B},\|\cdot\|')}\}$ be an ϵ-covering of $\mathcal{B}$ in $\|\cdot\|'$-norm. Then, we have

$$\mathcal{B} \subseteq \bigcup_{j=1}^{N(\epsilon;\mathcal{B},\|\cdot\|')} \{x_j + \epsilon\mathcal{B}'\},$$

which implies $vol(\mathcal{B}) \leq N(\epsilon; \mathcal{B}, \|\cdot\|')\, \epsilon^d\, vol(\mathcal{B}')$, thus establishing the lower bound in (10.3). The upper bound is obtained by starting with a maximal ϵ-packing $\{x_1, ..., x_{M(\epsilon;\mathcal{B},\|\cdot\|')}\}$ of $\mathcal{B}$ in the $\|\cdot\|'$-norm. The balls $\{x_j + \frac{\epsilon}{2}\mathcal{B}', j = 1, ..., M(\epsilon; \mathcal{B}, \|\cdot\|')\}$ are all disjoint and contained within $\mathcal{B} + \frac{\epsilon}{2}\mathcal{B}'$. Taking volumes, we can therefore conclude that

$$\sum_{j=1}^{M(\epsilon;\mathcal{B},\|\cdot\|')} vol\Big(x_j + \frac{\epsilon}{2}\mathcal{B}'\Big) \leq vol\Big(\mathcal{B} + \frac{\epsilon}{2}\mathcal{B}'\Big),$$

and hence

$$M(\epsilon; \mathcal{B}, \|\cdot\|')\, vol\Big(\frac{\epsilon}{2}\mathcal{B}'\Big) \leq vol\Big(\mathcal{B} + \frac{\epsilon}{2}\mathcal{B}'\Big).$$

Finally, we have $vol(\frac{\epsilon}{2}\mathcal{B}') = (\frac{\epsilon}{2})^d vol(\mathcal{B}')$ and $vol(\mathcal{B} + \frac{\epsilon}{2}\mathcal{B}') = (\frac{\epsilon}{2})^d vol(\frac{2}{\epsilon}\mathcal{B} + \mathcal{B}')$, which, together with $M(\epsilon; \mathcal{B}, \|\cdot\|') \geq N(\epsilon; \mathcal{B}, \|\cdot\|')$ thanks to Lemma 10.4, yields the upper bound in (10.3). □

This result now allows us to establish the scaling of the metric entropy of unit balls in terms of their own norm, thus yielding a measure of the massiveness of unit balls in d-dimensional spaces. Specifically, we set $\mathcal{B}' = \mathcal{B}$ in Lemma 10.5 and get

$$vol\left(\frac{2}{\epsilon}\mathcal{B} + \mathcal{B}'\right) = vol\left(\left(\frac{2}{\epsilon} + 1\right)\mathcal{B}\right) = \left(\frac{2}{\epsilon} + 1\right)^d vol(\mathcal{B}),$$

which when used in (10.3) yields $N(\epsilon; \mathcal{B}, \|\cdot\|) \asymp \epsilon^{-d}$ and hence results in metric entropy scaling according to $\log_2(N(\epsilon; \mathcal{B}, \|\cdot\|)) \asymp d\log_2(\epsilon^{-1})$. Particularizing this result to the unit ball $\mathcal{B}_\infty^d = [-1, 1]^d$ and the metric $\|\cdot\|_\infty$, we recover the result of our direct analysis in the example above.

So far we have studied the metric entropy of various subsets of $\mathbb{R}^d$. We now turn to the metric entropy of function classes, beginning with a simple one-parameter class. For a fixed θ, define the real-valued function $f_\theta(x) = 1 - e^{-\theta x}$, and consider the class

$$\mathcal{P} = \{f_\theta : [0, 1] \to \mathbb{R} \,|\, \theta \in [0, 1]\}.$$

The set $\mathcal{P}$ constitutes a metric space under the sup-norm given by $\|f - g\|_\infty = \sup_{x\in[0,1]} |f(x) - g(x)|$. We show that the covering number of $\mathcal{P}$ satisfies

$$1 + \left\lfloor \frac{1 - 1/e}{2\epsilon} \right\rfloor \leq N(\epsilon; \mathcal{P}, \|\cdot\|_\infty) \leq \frac{1}{2\epsilon} + 2,$$

which leads to the scaling behavior $N(\epsilon; \mathcal{P}, \|\cdot\|_\infty) \asymp \epsilon^{-1}$ and hence metric entropy scaling according to $\log_2(N(\epsilon; \mathcal{P}, \|\cdot\|_\infty)) = \log_2(\epsilon^{-1})$. We start by establishing the upper bound. For given $\epsilon \in [0,1]$, set $T = \lfloor \frac{1}{2\epsilon} \rfloor$, and define the points $\theta_i = 2\epsilon i$, for $i = 0, 1, ..., T$. By also adding the point $\theta_{T+1} = 1$, we obtain a collection of $T + 2$ points $\{\theta_0, \theta_1, ..., \theta_{T+1}\}$ contained within $[0, 1]$. We show that the associated functions $\{f_{\theta_0}, f_{\theta_1}, ..., f_{\theta_{T+1}}\}$ form an ϵ-cover for $\mathcal{P}$. Indeed, for any $f_\theta \in \mathcal{P}$, we can find some θ_i in the cover such that $|\theta - \theta_i| \leq \epsilon$. We then have

$$\|f_\theta - f_{\theta_i}\|_\infty = \max_{x \in [0,1]} |e^{-\theta x} - e^{-\theta_i x}| \leq |\theta - \theta_i|,$$

where we used, for $\theta < \theta_i$,

$$\begin{aligned} \max_{x\in[0,1]} |e^{-\theta x} - e^{-\theta_i x}| &= \max_{x\in[0,1]} (e^{-\theta x} - e^{-\theta_i x}) = \max_{x\in[0,1]} e^{-\theta x}(1 - e^{-(\theta_i - \theta)x}) \leq \max_{x\in[0,1]} (1 - e^{-(\theta_i - \theta)x}) \\ &\leq \max_{x\in[0,1]} (\theta_i - \theta)x \leq \theta_i - \theta = |\theta - \theta_i|, \end{aligned}$$

as a consequence of $1 - e^{-x} \leq x$, for $x \in [0, 1]$, which is easily verified by noting that the function $g(x) = 1 - e^{-x} - x$ satisfies $g(0) = 0$ and $g'(x) \leq 0$, for $x \in [0, 1]$. The case $\theta > \theta_i$ follows similarly. In summary, we have shown that $N(\epsilon; \mathcal{P}, \|\cdot\|_\infty) \leq T + 2 \leq \frac{1}{2\epsilon} + 2$.

In order to prove the lower bound, we first lower-bound the packing number and then use Lemma 10.4. We start by constructing an explicit packing as follows. Set $\theta_0 = 0$ and define $\theta_i = -\log(1 - \epsilon i)$, for all i such that $\theta_i \leq 1$. The largest index T such that this holds is given by $T = \lfloor \frac{1-e^{-1}}{\epsilon} \rfloor$. Moreover, note that for any $i \neq j$ in the resulting set of functions, we have $\|f_{\theta_i} - f_{\theta_j}\|_\infty \geq |f_{\theta_i}(1) - f_{\theta_j}(1)| = |\epsilon(i - j)| \geq \epsilon$. We can therefore conclude that $M(\epsilon; \mathcal{P}, \|\cdot\|_\infty) \geq \lfloor \frac{1-1/e}{\epsilon} \rfloor + 1$, and hence, thanks to the lower bound in Lemma 10.4 that

$$N(\epsilon; \mathcal{P}, \|\cdot\|_\infty) \geq M(2\epsilon; \mathcal{P}, \|\cdot\|_\infty) \geq \left\lfloor \frac{1 - 1/e}{2\epsilon} \right\rfloor + 1,$$

as claimed. We have thus established that the function class $\mathcal{P}$ has metric entropy scaling according to $\log_2(N(\epsilon; \mathcal{P}, \|\cdot\|_\infty)) \asymp \log_2(1/\epsilon)$, as $\epsilon \to 0$. This rate is typical for one-parameter function classes.

We now turn our attention to richer function classes and start by considering the class of Lipschitz functions on the unit interval

$$\mathcal{F}_L := \{g : [0,1] \to \mathbb{R} \,|\, g(0) = 0, \quad \text{and} \quad |g(x) - g(x')| \leq L|x - x'|, \; \forall x, x' \in [0,1]\}.$$

The metric entropy scaling of this function class is given by

$$\log_2 N(\epsilon; \mathcal{F}_L, \|\cdot\|_\infty) \asymp L/\epsilon,$$

which shows that this function class is much richer than the one-parameter family from the previous example whose metric entropy grows according to $\log_2(1/\epsilon)$. The preceding example can be extended to Lipschitz functions on the d-dimensional unit cube, meaning real-valued functions on $[0, 1]^d$ such that

$$|f(x) - f(y)| \leq L\|x - y\|_\infty, \qquad \text{for all} \quad x, y \in [0,1]^d,$$

a class that we denote by $\mathcal{F}_L([0,1]^d)$. This class has metric entropy scaling

$$\log_2 N(\epsilon; \mathcal{F}_L, \|\cdot\|_\infty) \asymp (L/\epsilon)^d. \tag{10.4}$$

It is interesting to contrast the exponential dependence of metric entropy in (10.4) on the ambient dimension d, as opposed to the linear dependence we saw earlier for simpler sets such as unit balls in $\mathbb{R}^d$, where we had

$$\log_2 N(\epsilon; \mathcal{B}, \|\cdot\|_\infty) \asymp d \log_2(\epsilon^{-1}).$$

Let us now relate the optimal exponent $\gamma^*(\mathcal{C})$ as defined in (10.1) to metric entropy scaling. All the examples of metric entropy scaling we have seen exhibit a behavior that fits the law $\log_2(N(\epsilon; \mathcal{C}, \|\cdot\|)) \asymp \epsilon^{-1/\gamma}$ or $\log_2(N(\epsilon; \mathcal{C}, \|\cdot\|)) \asymp \epsilon^{-1/\gamma} \log(\epsilon^{-1})^\beta$. The optimal exponent is hence a crude measure of growth insensitive to $\log$-factors or similar factors that are dominated by the growth of $\epsilon^{-1/\gamma}$.

10.3 Approximation with Representation Systems

We now show how Kolmogorov rate-distortion theory can be put to work in the context of optimal approximation with dictionaries. To this end, we start with a brief discussion of basics on optimal approximation in Hilbert spaces. Specifically, we shall consider two types of approximation, namely linear and nonlinear.

Let $\mathcal{H}$ be a Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$ and induced norm $\|\cdot\|_{\mathcal{H}}$ and let e_k, $k = 1, 2, ...$ be an orthonormal basis for $\mathcal{H}$. For linear approximation, we use the linear space $\mathcal{H}_M := \text{span}\{e_k : 1 \leq k \leq M\}$ to approximate a given element $f \in \mathcal{H}$. We measure the approximation error by

$$E_M(f)_{\mathcal{H}} := \inf_{g \in \mathcal{H}_M} \|f - g\|_{\mathcal{H}}.$$

In nonlinear approximation, we consider M-term approximation, which replaces $\mathcal{H}_M$ by the space Σ_M consisting of all elements $g \in \mathcal{H}$ that can be expressed as

$$g = \sum_{k \in \Lambda} c_k e_k,$$

where $\Lambda \subset \mathbb{N}$ is a set of indices with $|\Lambda| \leq M$. Note that, in contrast to $\mathcal{H}_M$, the space Σ_M is not linear. A linear combination of two elements in Σ_M will, in general, need $2M$ terms in its representation by the e_k. Analogous to E_M, we define the error of M-term approximation

$$\sigma_M(f)_{\mathcal{H}} := \inf_{g \in \Sigma_M} \|f - g\|_{\mathcal{H}}.$$

It is immediate that nonlinear M-term approximation can achieve smaller approximation error than linear M-term approximation.

We now proceed to M-term approximation in general dictionaries, i.e., we replace the orthonormal basis $\{e_k\}$ by a general, possibly redundant, set of functions. The specific setup we consider is as follows. Fix $\Omega \subset \mathbb{R}^d$. Let $\mathcal{C}$ be a compact set of functions in $L^2(\Omega)$, henceforth referred to as *function class*, and consider a corresponding system (dictionary) $\mathcal{D} := (\varphi_i)_{i\in I} \subset L^2(\Omega)$ with I countable, termed *representation system*. We study the *best M-term approximation error* of $f \in \mathcal{C}$ in $\mathcal{D}$ defined as follows.

Definition 10.6. [85] Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, a function class $\mathcal{C} \subset L^2(\Omega)$, and a representation system $\mathcal{D} = (\varphi_i)_{i\in I} \subset L^2(\Omega)$, we define, for $f \in \mathcal{C}$ and $M \in \mathbb{N}$,

$$\Gamma_M^{\mathcal{D}}(f) := \inf_{\substack{I_M \subseteq I, \\ \#I_M = M, (c_i)_{i\in I_M}}} \left\| f - \sum_{i\in I_M} c_i\varphi_i \right\|_{L^2(\Omega)} . \tag{10.5}$$

We call $\Gamma_M^{\mathcal{D}}(f)$ the *best M-term approximation error of f in $\mathcal{D}$*. Every $f_M = \sum_{i\in I_M} c_i\varphi_i$ attaining the infimum in (10.5) is referred to as a *best M-term approximation* of f in $\mathcal{D}$. The supremal $\gamma > 0$ such that

$$\sup_{f\in\mathcal{C}} \Gamma_M^{\mathcal{D}}(f) \in \mathcal{O}(M^{-\gamma}),\ \ M \to \infty,$$

will be denoted by $\gamma^*(\mathcal{C}, \mathcal{D})$. We say that the *best M-term approximation rate of $\mathcal{C}$ in the representation system $\mathcal{D}$* is $\gamma^*(\mathcal{C}, \mathcal{D})$.

Function classes $\mathcal{C}$ widely studied in the approximation theory literature include unit balls in Lebesgue, Sobolev, or Besov spaces [86], as well as α-cartoon-like functions [87]. A wealth of structured representation systems $\mathcal{D}$ is provided by the area of applied harmonic analysis, starting with wavelets [88], followed by ridgelets [89], curvelets [90], shearlets [91], parabolic molecules [92], and most generally α-molecules [87], which include all previously named systems as special cases. Further examples are Gabor frames [93], local cosine bases [94], and wave atoms [95].

The best M-term approximation rate $\gamma^*(\mathcal{C}, \mathcal{D})$ according to Definition 10.6 quantifies how difficult it is to approximate a given function class $\mathcal{C}$ in a fixed representation system $\mathcal{D}$. It is sensible to ask whether for given $\mathcal{C}$, there is a fundamental limit on $\gamma^*(\mathcal{C}, \mathcal{D})$ when one is allowed to vary over $\mathcal{D}$. As shown in [96, 97], every dense (and countable) $\mathcal{D} \subset L^2(\Omega)$, $\Omega \subset \mathbb{R}^d$, results in $\gamma^*(\mathcal{C}, \mathcal{D}) = \infty$ for all function classes $\mathcal{C} \subset L^2(\Omega)$. However, identifying the elements in $\mathcal{D}$ participating in the best M-term approximation is practically infeasible as it entails searching through the infinite set $\mathcal{D}$ and requires, in general, an infinite number of bits to describe the indices of the participating elements. This insight leads to the concept of "best M-term approximation subject to polynomial-depth search" as introduced by Donoho in [96]. Here, the basic idea is to restrict the search for the elements in $\mathcal{D}$ participating in the best M-term approximation to the first $\pi(M)$ elements of $\mathcal{D}$, with π a polynomial. We formalize this under the name of effective best M-term approximation as follows.

Definition 10.7. Given $d \in \mathbb{N}$, $\Omega \subset \mathbb{R}^d$, a function class $\mathcal{C} \subset L^2(\Omega)$, and a representation system $\mathcal{D} = (\varphi_i)_{i \in I} \subset L^2(\Omega)$, the supremal $\gamma > 0$ so that there exists a polynomial π

$$\sup_{f \in \mathcal{C}} \inf_{\substack{I_M \subset \{1,2,\ldots,\pi(M)\},\\ \#I_M = M,\, (c_i)_{i \in I_M}}} \left\| f - \sum_{i \in I_M} c_i \varphi_i \right\|_{L^2(\Omega)} \in \mathcal{O}(M^{-\gamma}),\ M \to \infty, \tag{10.6}$$

will be denoted by $\gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D})$ and referred to as *effective best M-term approximation rate of $\mathcal{C}$ in the representation system $\mathcal{D}$.*

We next show that $\sup_{\mathcal{D} \subset L^2(\Omega)} \gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D})$ is, indeed, finite under quite general conditions on $\mathcal{C}$; more specifically, it is upper-bounded by $\gamma^*(\mathcal{C})$ and hence limited by the "description complexity" of $\mathcal{C}$. This endows $\gamma^*(\mathcal{C})$ with operational meaning.

Theorem 10.8. *[96, 97] Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$. The effective best M-term approximation rate of the function class $\mathcal{C} \subset L^2(\Omega)$ in the representation system $\mathcal{D} \subset L^2(\Omega)$ satisfies*

$$\gamma^{*,\mathit{eff}}(\mathcal{C}, \mathcal{D}) \leq \gamma^*(\mathcal{C}).$$

Proof. Consider an effective best M-term approximation of f in $\mathcal{C}$, which by Definition 10.7 satisfies

$$\left\| f - \sum_{i \in I_M} c_i \varphi_i \right\|_{L^2(\Omega)} \leq C M^{-\gamma} \tag{10.7}$$

for all $\gamma < \gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D})$ and with some constant C. Moreover, thanks to the polynomial depth search constraint, we have $|I_M| \leq \pi(M)$.

The proof is based on an explicit construction of a bitstring encoding the indices of the dictionary elements participating in the best M-term approximation along with the corresponding coefficients c_i, or more precisely quantized versions of the coefficients. Moreover, this bitstring has to be uniquely decodable, i.e., it has to be possible to uniquely read off the indices in I_M and the (quantized) coefficients c_i from the bitstring. We start by encoding the indices of the dictionary elements participating in the best M-term approximation. As $|I_M| \leq \pi(M)$, each index can be encoded by at most $\log(\pi(M)) = C \log(M)$ bits, which results in a total of $CM \log(M)$ bits needed to encode the indices of all participating dictionary elements. The encoder and the decoder are assumed to know $\pi(M)$, which allows to determine a monomial M^P such that $|\pi(M)| \leq M^P$ and, in particular, to choose C in $\log(\pi(M)) = C \log(M)$ large enough so that stacking the binary representations of the indices does not lead to overlap between the individual indices' binary representations. As the decoder knows this constant C as well, it can uniquely read off the indices from the sequence of their binary representations.

Next, we encode the coefficients. As the coefficients are real numbers, we would, in principle, need an infinite number of bits to encode each one of them. What comes to our rescue is, however, the insight that we are not seeking a perfect, i.e., zero-error representation, but are, in fact, allowed an error according to (10.7). The goal is to quantize the coefficients such that a minimum number of bits is needed to represent them while the overall error incurred by quantization is commensurate with the error allowed by (10.7). To this end, we first perform a Gram-Schmidt orthogonalization on the elements $\{\varphi_i\}_{i\in I_M}$ participating in the best M-term approximation. This yields an orthonormal set of functions $\{\tilde{\varphi}_i\}_{i\in I_{\tilde{M}}}$ which has the same span as $\{\varphi_i\}_{i\in I_M}$. Next, we define (implicitly) the coefficients $\tilde{c}_i$ according to

$$\sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i = \sum_{i\in I_M} c_i\varphi_i, \tag{10.8}$$

and note that $\tilde{M} \leq M$. The error in approximating f by the modified sequence is given by

$$e = f - \sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i.$$

Thanks to (10.8), we get

$$e = f - \sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i = f - \sum_{i\in I_M} c_i\varphi_i$$

which implies

$$\left\|\sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i\right\| = \|f-e\| \leq \|f\| + \|e\|. \tag{10.9}$$

Now, exploiting the orthonormality of the $\tilde{\varphi}_i$, we obtain

$$\left\|\sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i\right\|^2 = \sum_{i\in I_{\tilde{M}}} |\tilde{c}_i|^2,$$

which when used in (10.9) yields

$$\sqrt{\sum_{i\in I_{\tilde{M}}} |\tilde{c}_i|^2} \leq \sup_{f\in\mathcal{C}} \|f\| + CM^{-\gamma}.$$

As $\mathcal{C}$ is compact, by assumption, it is bounded and hence $\sup_{f\in\mathcal{C}} \|f\| < \infty$, which establishes that the modified coefficients $\tilde{c}_i$ are uniformly bounded. This, in turn, allows us to quantize the $\tilde{c}_i$ as follows. We perform uniform quantization by rounding the coefficients $\tilde{c}_i$ to integer multiples of $M^{-(\gamma+1/2)}$, let us denote these rounded coefficients by $\hat{c}_i$. As the $\tilde{c}_i$ are uniformly bounded, this results in a number of quantization levels that is proportional to $M^{(\gamma+1/2)}$. We hence need a total of $\mathcal{O}(M\log(M))$ bits (recall that $\tilde{M} \leq M$) to store the binary representations of the quantized coefficients. Again, the proportionality constant is known to encoder and decoder, which allows us to stack the binary representations of the quantized versions of the $\tilde{c}_i$ in a uniquely decodable manner. We finally note that the specific choice of the exponent $\gamma + 1/2$ is informed by the upper bound on the reconstruction error we are allowed, this will be made explicit below in the description of the decoder.

In summary, we have mapped the signal f to a bitstring of length $\mathcal{O}(M\log(M))$. The decoder is presented with this bitstring and reconstructs an approximation to f as follows. It first reads out the indices of the set I_M and the $\hat{c}_i$. Recall that this is uniquely possible. Next, the decoder performs a Gram-Schmidt orthogonalization on the dictionary elements corresponding to the indices in I_M. The error resulting from reconstructing the signal f from the quantized coefficients $\hat{c}_i$ rather than from the exact coefficients $\tilde{c}_i$ can be bounded according to

$$\begin{aligned}
\left\| f - \sum_{i\in I_{\tilde{M}}} \hat{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} &= \left\| f - \sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i + \sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i - \sum_{i\in I_{\tilde{M}}} \hat{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} && (10.10)\\
&\le \left\| f - \sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} + \left\| \sum_{i\in I_{\tilde{M}}} (\tilde{c}_i - \hat{c}_i)\tilde{\varphi}_i \right\|_{L^2(\Omega)} && (10.11)\\
&= \left\| f - \sum_{i\in I_{\tilde{M}}} \tilde{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} + \left(\sum_{i\in I_{\tilde{M}}} |\tilde{c}_i - \hat{c}_i|^2 \right)^{1/2}, && (10.12)
\end{aligned}$$

where in the last step we again exploited the orthonormality of the $\tilde{\varphi}_i$. Next, we note that thanks to the choice of the quantizer resolution, we have $|\tilde{c}_i - \hat{c}_i|^2 \le C'M^{-2\gamma-1}$ for some constant C'. Together with $|I_M| = M$ this yields

$$\sum_{i\in I_{\tilde{M}}} |\tilde{c}_i - \hat{c}_i|^2 \le C''M^{-2\gamma},$$

for some constant C''. Together with (10.7) and (10.8), this finally yields

$$\left\| f - \sum_{i\in I_{\tilde{M}}} \hat{c}_i\tilde{\varphi}_i \right\|_{L^2(\Omega)} \le CM^{-\gamma}$$

for some constant C. Noting that all our arguments are valid for $\gamma < \gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})$, we have thereby established the existence of an encoder-decoder pair with a bitstring length scaling according to $\varepsilon^{-1/\gamma}\log(\varepsilon^{-1/\gamma}) \in \mathcal{O}(\varepsilon^{-1/(\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})-\delta)})$, $\varepsilon \to 0$, for every $\delta \in (0, \gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}))$. By Definition 10.1 this leads to $\gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}) - \delta \le \gamma^*(\mathcal{C})$, for every $\delta \in (0, \gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D}))$, which upon noting that δ can be chosen arbitrarily small, establishes the achievability part of the statement in the theorem, i.e., all exponents γ with $\gamma < \gamma^{*,\text{eff}}(\mathcal{C},\mathcal{D})$ can be achieved. Note that we have not only established achievability, but have also identified encoder-decoder pairs that achieve the optimal exponent.

A matching strong converse is obtained simply by noting that taking $\gamma > \gamma^*(\mathcal{C})$ in the arguments above, we would get an encoder-decoder pair with $L(\epsilon,\mathcal{C}) \in \mathcal{O}(\epsilon^{-1/\gamma})$, where $\gamma > \gamma^*(\mathcal{C})$, which owing to Definition 10.1 can not exist. More constructively speaking such an encoder-decoder pair would violate the fundamental limit imposed by metric entropy scaling of the function class under consideration. □

In light of the result just established, the following definition is natural (see also [97]).

Definition 10.9. Let $d \in \mathbb{N}$ and $\Omega \subset \mathbb{R}^d$. If the effective best M-term approximation rate of the function class $\mathcal{C} \subset L^2(\Omega)$ in the representation system $\mathcal{D} \subset L^2(\Omega)$ satisfies

$$\gamma^{*,\text{eff}}(\mathcal{C}, \mathcal{D}) = \gamma^*(\mathcal{C}),$$

we say that the function class $\mathcal{C}$ is *optimally representable* by $\mathcal{D}$.

Chapter 11

Sparsity in Redundant Dictionaries

11.1 Sparsity in Redundant Dictionaries

- Complex signals such as audio recordings or images often include structures that are not well represented by a few elements of a single basis.
- Small dictionaries have a limited capability of sparse expression.
- Large dictionaries incorporating more patterns can increase sparsity and thus improve performance in e.g. compression, denoising, inverse problems, and pattern recognition.
- Finding the set of M dictionary vectors that approximate a signal with a minimum error is NP-hard.

11.1.1 Linear Approximation Error

Let $\mathcal{B} = \{g_j\}_{j\in\mathbb{N}_0}$ be an ONB for $\mathcal{H}$. Any $\mathbf{x} \in \mathcal{H}$ can be written as

$$\mathbf{x} = \sum_{j=0}^{\infty} \langle \mathbf{x}, g_j \rangle \, g_j$$

If we approximate $\mathbf{x}$ by the first M inner products, we get

$$\mathbf{x}_m = \sum_{j=0}^{M-1} \langle \mathbf{x}, g_j \rangle \, g_j$$

This approximation is the orthogonal projection of $\mathbf{x}$ onto the space spanned by $\{g_j\}_{j=0}^{M-1}$.
The corresponding approximation error is

$$\mathbf{x} - \mathbf{x}_M = \sum_{j=M}^{\infty} \langle \mathbf{x}, g_j \rangle \, g_j$$

Approximation error

$$\mathcal{E}_l[M] = \|\mathbf{x} - \mathbf{x}_M\|^2 = \sum_{j=M}^{\infty} |\langle \mathbf{x}, g_j \rangle|^2$$

Since

$$\|\mathbf{x}\|^2 = \sum_{j=0}^{\infty} |\langle \mathbf{x}, g_j \rangle|^2 < \infty,$$

it follows that $\lim_{M\to\infty} \mathcal{E}_l[M] = 0$. The decay rate of the error depends on the decay of $|\langle \mathbf{x}, g_j \rangle|$ as j increases. The following theorem gives equivalent conditions on the decay of $\mathcal{E}_l[M]$ and $|\langle \mathbf{x}, g_j \rangle|$.

Theorem 11.1. *Let $s > 1/2$ and $\sum_{j=0}^{\infty} j^{2s} |\langle \mathbf{x}, g_j \rangle|^2 < \infty$. There exist constants $A, B > 0$ such that*

$$A \sum_{j=0}^{\infty} j^{2s} |\langle \mathbf{x}, g_j \rangle|^2 \leq \sum_{M=0}^{\infty} M^{2s-1} \mathcal{E}_l[M] \leq B \sum_{j=0}^{\infty} j^{2s} |\langle \mathbf{x}, g_j \rangle|^2$$

and hence $\mathcal{E}_l[M] = o(M^{-2s})$, i.e., $\lim_{M\to\infty} \mathcal{E}_l[M] M^{2s} = 0$.

Proof.

$$\begin{aligned}
&\sum_{M=0}^{\infty} M^{2s-1} \sum_{j=M}^{\infty} |\langle \mathbf{x}, g_j \rangle|^2 \\
&= 1 \cdot \sum_{j=1}^{\infty} |\langle \mathbf{x}, g_j \rangle|^2 + 2^{2s-1} \sum_{j=2}^{\infty} |\langle \mathbf{x}, g_j \rangle|^2 + 3^{2s-1} \sum_{j=3}^{\infty} |\langle \mathbf{x}, g_j \rangle|^2 + ... \\
&= \sum_{j=0}^{\infty} |\langle \mathbf{x}, g_j \rangle|^2 \sum_{M=0}^{j} M^{2s-1}
\end{aligned}$$

For any $s > 1/2$, we have

$$\underbrace{\int_0^j y^{2s-1} dy}_{\frac{1}{2s} j^{2s} = A j^{2s}} \leq \sum_{M=0}^{j} M^{2s-1} \leq \underbrace{\int_0^{j+1} y^{2s-1} dy}_{\frac{1}{2s}(j+1)^{2s} \leq B j^{2s}}$$

The area under the curve in figure 11.1 is $\sum_{M=0}^{j} M^{2s-1}$ and is upper bounded by the area under y^{2s-1} between 0 and $j+1$ and lower bounded by the area under $(y-1)^{2s-1}$ between 1 and $j+1$. The lower bound then follows from substitution of variables:

$$\int_1^{j+1} (y-1)^{2s-1} dy \underset{y'=y-1}{=} \int_0^j y^{2s-1} dy$$

Putting the pieces together, we get

$$A \sum_{j=0}^{\infty} j^{2s} |\langle \mathbf{x}, g_j \rangle|^2 \leq \sum_{M=0}^{\infty} M^{2s-1} \mathcal{E}_l[M] \leq B \sum_{j=0}^{\infty} j^{2s} |\langle \mathbf{x}, g_j \rangle|^2$$

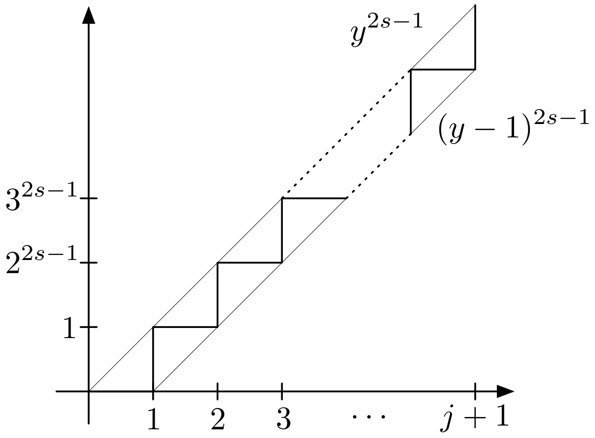


Figure 11.1:

which is the first part of the theorem. To verify that $\mathcal{E}_l[M] = o(M^{-2s})$, observe that $\mathcal{E}_l[m] \geq \mathcal{E}_l[M]$ for $m \leq M$, so that

$$\mathcal{E}_l[M] \sum_{j=M/2}^{M-1} j^{2s-1} \leq \sum_{j=M/2}^{M-1} j^{2s-1}\mathcal{E}_l[j] \leq \sum_{j=M/2}^{\infty} j^{2s-1}\mathcal{E}_l[j]$$

Since

$$\sum_{j=1}^{\infty} j^{2s-1}\mathcal{E}_l[j] < \infty,$$

by

$$\sum_{M=0}^{\infty} M^{2s-1}\mathcal{E}_l[M] \leq B \sum_{j=0}^{\infty} j^{2s} \left|\langle \mathbf{x}, g_j\rangle\right|^2 < \infty$$

where the last inequality follows by assumption, it follows that

$$\lim_{M\to\infty} \sum_{j=M/2}^{\infty} j^{2s-1}\mathcal{E}_l[j] = 0.$$

Moreover, there exists $C > 0$ such that

$$\sum_{j=M/2}^{M-1} j^{2s-1} \geq CM^{2s}$$

$$\left(\int_{M/2}^{M} y^{2s-1}dy = \frac{1}{2s}\left(M^{2s} - \left(\frac{M}{2}\right)^{2s}\right) \geq \frac{1}{2s}M^{2s} = C'M^{2s}\right)$$

$\Rightarrow \lim_{M\to\infty} \mathcal{E}_l[M]M^{2s} = 0.$ □

The theorem proves that the linear approximation error of $\mathbf{x}$ in the basis $\mathcal{B}$ decays faster than M^{-2s} if $\mathbf{x}$ belongs to the space

$$W_{\mathcal{B},s} = \{\mathbf{x} \in \mathcal{H} : \sum_{j=0}^{\infty} j^{2s} \left|\langle \mathbf{x}, g_j \rangle\right|^2 < \infty\}.$$

If $\mathcal{B}$ is a Fourier or wavelet basis, then $W_{\mathcal{B},s}$ is a so-called Sobolev space. Note that the linear approximation of $\mathbf{x}$ by the first M elements in $\mathcal{B}$ is not always the best in terms of minimizing the approximation error. It would be better to choose the M elements adaptively as a function of the signal $\mathbf{x}$ we wish to approximate.

We next have a closer look at $W_{\mathcal{B},s}$ if $\mathcal{B}$ is the Fourier basis.

11.1.2 Regularity and Decay

Theorem 11.2. *If* $\mathbf{x}$ *is integrable, then its Fourier transform is a uniformly continuous function satisfying*

$$|\hat{x}(f)| \leq \int_{-\infty}^{\infty} |x(t)|\, dt < \infty, \quad f \in \mathbb{R}$$

and

$$\lim_{|f| \to \infty} \hat{x}(f) = 0$$

From the theorem we can immediately conclude that if $\hat{x}$ is integrable, then $\mathbf{x}$ is uniformly continuous and bounded and satisfies $\lim_{|t| \to \infty} x(t) = 0$.

Proposition 11.3. A function $\mathbf{x}$ is bounded and p times continuously differentiable with bounded derivative if

$$\int_{-\infty}^{\infty} |\hat{x}(f)|\, (1 + |f|)^p df < \infty$$

Proof. The Fourier transform of $x^{(k)}(t)$ is $(2\pi \mathrm{i} f)^k \hat{x}(f)$. Therefore

$$\left|x^{(k)}(t)\right| \leq C \int_{-\infty}^{\infty} |\hat{x}(f)|\, |f|^k\, df \leq C \int_{-\infty}^{\infty} |\hat{x}(f)|\, (1 + |f|)^k df < \infty$$

The assumption implies that $\int_{-\infty}^{\infty} |\hat{x}(f)|\, |f|^k\, df < \infty$ for any $k \leq p$, so that $x^{(k)}(t)$ is continuous and bounded. □

The proposition implies that if there exist constants k and $\epsilon > 0$ such that

$$|\hat{x}(f)| \leq \frac{k}{(1 + |f|)^{p+1+\epsilon}}$$

then $\mathbf{x} \in C^p$.

The class C^p: The continuous function $\mathbf{x}$ is said to be of class C^p if the derivatives $x^{(1)}, x^{(2)}, \ldots, x^{(p)}$ exist and are continuous. The class C^0 consists of all continuous functions, the class C^1 of all functions that are continuous and differentiable with the derivative being continuous.

Back to $W_{\mathcal{B},s}$ if $\mathcal{B}$ is the Fourier basis.
Approximating a function with few coefficients in a Fourier basis with a small approximation error means that these functions are effectively lowpass functions and hence smooth in the time-domain.

11.1.3 Non-Linear Approximations

Linear approximations project the signal onto M basis elements selected a priori. This approximation can be improved by choosing the M basis elements to depend on the signal.

11.1.4 Approximation Error

A signal $\mathbf{x} \in \mathcal{H}$ is approximated with M elements in an ONB $\mathcal{B} = \{g_j\}_{j\in\mathbb{N}}$ of $\mathcal{H}$. Let $\mathbf{x}_M$ be the projection of $\mathbf{x}$ over M elements whose indices are collected in $\mathcal{I}_M$:

$$\mathbf{x}_M = \sum_{j\in\mathcal{I}_M} \langle \mathbf{x}, g_j \rangle \, g_j$$

$$\mathcal{E}_n[M] = \|\mathbf{x} - \mathbf{x}_M\|^2 = \sum_{j\notin\mathcal{I}_M} |\langle \mathbf{x}, g_j \rangle|^2$$

To minimize this error, the indices in $\mathcal{I}_M$ must correspond to the M elements having the largest inner product amplitudes. This is a consequence of

$$\|\mathbf{x}\|^2 = \sum_j |\langle \mathbf{x}, g_j \rangle|^2 = \sum_{j\in\mathcal{I}_M} |\langle \mathbf{x}, g_j \rangle|^2 + \sum_{j\notin\mathcal{I}_M} |\langle \mathbf{x}, g_j \rangle|^2$$

These are the elements in $\{g_j\}$ that describe the main ”features” of $\mathbf{x}$. The resulting error $\mathcal{E}_n[M]$ is necessarily smaller than or equal to the error obtained in the case of a linear approximation.

We sort $\{|\langle \mathbf{x}, g_j \rangle|\}_{j\in\mathbb{N}}$ in descending order and denote $x_{\mathcal{B}}^r[k] = \langle \mathbf{x}, g_{j_k} \rangle$ as the coefficient of rank k. Obviously, we have

$$|x_{\mathcal{B}}^r[k]| \geq |x_{\mathcal{B}}^r[k+1]| \,, \quad k \geq 1$$

The best non-linear approximation is:

$$\mathbf{x}_M = \sum_{k=1}^{M} x_{\mathcal{B}}^r[k] g_{j_k}$$

The corresponding minimum non-linear approximation error is

$$\mathcal{E}_n[M] = \|\mathbf{x} - \mathbf{x}_M\|^2 = \sum_{k=M+1}^{\infty} |x_{\mathcal{B}}^r[k]|^2 \,.$$

Theorem 11.4. *Let $s > 1/2$. If there exists $C > 0$ such that $|x_{\mathcal{B}}^r[k]| \le Ck^{-s}$, then*

$$\mathcal{E}_n[M] \le \frac{C^2}{2s-1} M^{1-2s}. \tag{11.1}$$

Conversely, if $\mathcal{E}_n[M]$ satisfies (11.1)*, then*

$$|x_{\mathcal{B}}^r[k]| \le (1 - \frac{1}{2s})^{-s} C k^{-s}. \tag{11.2}$$

Proof. Since

$$\mathcal{E}_n[M] = \sum_{k=M+1}^{\infty} |x_{\mathcal{B}}^r[k]|^2 \le C^2 \sum_{k=M+1}^{\infty} k^{-2s}$$

and

$$\sum_{k=M+1}^{\infty} k^{-2s} \le \int_M^{\infty} y^{-2s} dy = \frac{M^{1-2s}}{2s-1}$$

we get

$$\mathcal{E}_n[M] \le \frac{C^2}{2s-1} M^{1-2s},$$

which proves the first part of the theorem. Conversely, let $\alpha < 1$,

$$\mathcal{E}_n[\alpha M] = \sum_{k=\alpha M+1}^{\infty} |x_{\mathcal{B}}^r[k]|^2 \ge \sum_{k=\alpha M+1}^{M} |x_{\mathcal{B}}^r[k]|^2 \ge (1-\alpha) M \, |x_{\mathcal{B}}^r[M]|^2$$

$$\Rightarrow |x_{\mathcal{B}}^r[M]|^2 \le \frac{\mathcal{E}_n[\alpha M]}{(1-\alpha)M} \le \frac{1}{(1-\alpha)M} \frac{C^2}{2s-1} \alpha^{1-2s} M^{1-2s} = \frac{C^2}{2s-1} \frac{\alpha^{1-2s}}{(1-\alpha)} M^{-2s}$$

For $\alpha = 1 - \frac{1}{2s}$, we get $2s - 1 = \frac{\alpha}{1-\alpha}$, and hence

$$|x_{\mathcal{B}}^r[M]|^2 \le \frac{C^2}{\frac{\alpha}{1-\alpha}} \frac{\alpha^{1-2s}}{1-\alpha} M^{-2s} = C^2 \alpha^{-2s} M^{-2s} = C^2 (1 - \frac{1}{2s})^{-2s} M^{-2s}$$

$\Rightarrow |x_{\mathcal{B}}^r[k]| \le C(1 - \frac{1}{2s})^{-s} k^{-s}$. □

The decay of the sorted inner products can be related to the l^p-norm of the coefficients. Define

$$\|\mathbf{x}\|_{\mathcal{B},p} \triangleq \left(\sum_{j=0}^{\infty} |\langle \mathbf{x}, g_j \rangle|^p \right)^{1/p}$$

Theorem 11.5. *Let $p < 2$. If $\|\mathbf{x}\|_{\mathcal{B},p} < \infty$, then*

$$|x_{\mathcal{B}}^r[k]| \le \|\mathbf{x}\|_{\mathcal{B},p} k^{-1/p}$$

and

$$\mathcal{E}_n[M] = o(M^{1-2/p}),$$

i.e., $\lim_{M \to \infty} \mathcal{E}_n[M] M^{-1+2/p} = 0$.

Proof.

$$\|\mathbf{x}\|_{\mathcal{B},p}^p = \sum_{n=1}^{\infty} |x_{\mathcal{B}}^r[n]|^p \geq \sum_{n=1}^{k} |x_{\mathcal{B}}^r[n]|^p \geq k\,|x_{\mathcal{B}}^r[k]|^p$$
$$\Rightarrow |x_{\mathcal{B}}^r[k]| \leq \|\mathbf{x}\|_{\mathcal{B},p} k^{-1/p}$$

To show that $\mathcal{E}_n[M] = o(M^{1-2/p})$, we set

$$S[k] = \sum_{n=k}^{2k-1} |x_{\mathcal{B}}^r[n]|^p \geq k\,|x_{\mathcal{B}}^r[2k]|^p$$
$$\Rightarrow |x_{\mathcal{B}}^r[k]| \leq \left(\frac{S[k/2]}{k/2}\right)^{1/p}$$

Hence

$$\begin{aligned}\mathcal{E}_n[M] &= \sum_{k=M+1}^{\infty} |x_{\mathcal{B}}^r[k]|^2 \leq \sum_{k=M+1}^{\infty} S[k/2]^{2/p}(k/2)^{-2/p} \\ &\leq \sup_{k>M/2} (S[k])^{2/p} \sum_{k=M+1}^{\infty} (k/2)^{-2/p}\end{aligned}$$

$$\begin{aligned}\sum_{k=M+1}^{\infty} \left(\frac{k}{2}\right)^{-2/p} &\approx \int_M^{\infty} \left(\frac{y}{2}\right)^{-2/p} dy \\ &= \frac{2}{1-2/p}\left(\frac{y}{2}\right)^{1-2/p}\Bigg|_M^{\infty} = \frac{2}{2/p-1}\left(\frac{M}{2}\right)^{1-2/p} = C_p M^{1-2/p}\end{aligned}$$

$$\Rightarrow \mathcal{E}_n[M]\frac{1}{C_p}M^{-(1-2/p)} \leq \sup_{k>M/2} (S[k])^{2/p}$$

$$\lim_{M\to\infty} \sup_{k>M/2} (S[k])^{2/p} = \lim_{M\to\infty} \sup_{k>M/2} \underbrace{\left(\sum_{n=k}^{2k-1} |x_{\mathcal{B}}^r[n]|^p\right)^{2/p}}_{\substack{\to 0 (k\to\infty) \\ \text{as a consequence} \\ \text{of } \|x\|_{\mathcal{B}}^p<\infty}} = 0$$

$$\Rightarrow \lim_{M\to\infty} \mathcal{E}_n[M]M^{-(1-2/p)} = 0$$
$$\Rightarrow \mathcal{E}_n[M] = o(M^{1-2/p})$$

□

The theorem specifies spaces of functions that are well approximated by a few elements of an orthonormal basis $\mathcal{B}$. We denote

$$\mathcal{B}_{\mathcal{B},p} = \{\mathbf{x} \in \mathcal{H} : \|\mathbf{x}\|_{\mathcal{B},p} < \infty\}.$$

If $\mathbf{x} \in \mathcal{B}_{\mathcal{B},p}$, then the theorem above proves that $\mathcal{E}_n[M] = o(M^{1-2/p})$. This is called a ”Jackson inequality”. Conversely, if $\mathcal{E}_n[M] = o(M^{1-2/p})$ then the ”Bernstein inequality” (11.2) for $s = 1/p$ shows that $\mathbf{x} \in \mathcal{B}_{\mathcal{B},q}$ for any $q > p$. For wavelet bases the spaces $\mathbf{x} \in \mathcal{B}_{\mathcal{B},p}$ are Besov spaces.

11.2 Best M-term Approximation in Redundant Dictionaries

Let $\mathcal{D} = \{g_j\}_{j\in\Gamma}$ be a dictionary of unit norm vectors $\|g_j\| = 1$ in $\mathbb{C}^N$, with $|\Gamma| \geq N$. We study sparse approximations of $x \in \mathbb{C}^N$ with vectors selected in $\mathcal{D}$. Let $\{g_j\}_{j\in\Delta}$ be a subset of vectors in $\mathcal{D}$ with $|\Delta| \leq N$. Once we have selected the set $\{g_j\}_{j\in\Delta}$ the best (sparse in the case of $|\Delta| < N$) approximation of x is obtained by projecting (orthogonal projection) x onto the subspace spanned by $\{g_j\}_{j\in\Delta}$, i.e.,

$$x_\Delta = \sum_{j\in\Delta} a[j] g_j$$

The support set $\Delta \in \Gamma$ carries geometrical information about x relative to $\mathcal{D}$.
The corresponding approximation error is

$$\begin{aligned}\mathcal{E}_n(x,\Delta) = \|x - x_\Delta\|^2 = \|x - \sum_{j\in\Delta} a[j]g_j\|^2 &= \|\sum_{j\in\Gamma} \langle x, h_j\rangle\, g_j - \sum_{j\in\Delta} a[j]g_j\|^2 \\ &= \|\sum_{j\in\Delta} (\langle x, h_j\rangle - a[j])\, g_j + \sum_{j\in\Gamma\setminus\Delta} \langle x, h_j\rangle\, g_j\|^2\end{aligned}$$

where $\{h_j\}$ denotes the dual frame of $\{g_j\}$.
If $\{g_j\}_{j\in\Gamma}$ is an ONB, we get $(|\Gamma| = N)$ $h_j = g_j$ and hence

$$\begin{aligned}\mathcal{E}_n(x,\Delta) &= \|\sum_{j\in\Delta} (\langle x, g_j\rangle - a[j])\, g_j\|^2 + \|\sum_{j\in\Gamma\setminus\Delta} \langle x, g_j\rangle\, g_j\|^2 \\ &= \sum_{j\in\Delta} |\langle x, g_j\rangle - a[j]|^2 + \sum_{j\in\Gamma\setminus\Delta} |\langle x, g_j\rangle|^2\end{aligned}$$

For a given support set Δ, $\mathcal{E}_n(x,\Delta)$ is minimized if

$$a[j] = \langle x, g_j\rangle \Rightarrow \mathcal{E}_n(x,\Delta) = \sum_{j\in\Gamma\setminus\Delta} |\langle x, g_j\rangle|^2$$

From Parseval it follows that

$$\|x\|^2 = \sum_{j\in\Delta} |\langle x, g_j\rangle|^2 + \sum_{j\in\Gamma\setminus\Delta} |\langle x, g_j\rangle|^2$$

and hence the optimum support set of a given cardinality, say M, is obtained by maximizing $\sum_{j\in\Delta} |\langle x, g_j\rangle|^2$ over all support sets of $|\Delta| = M$. This corresponds to simply taking the M basis vectors that have the largest $|\langle x, g_j\rangle|^2$ and can be realized by thresholding (hard-thresholding) with the treshold T being a function of $|\Delta| = M$.

In the absence of orthonormality of the vectors g_j, we have

$$\mathcal{E}_n(x,\Delta) = \|\sum_{j\in\Delta} \left(\langle x, h_j\rangle - a[j]\right) g_j + \sum_{j\in\Gamma\setminus\Delta} \langle x, h_j\rangle\, g_j\|^2$$

Unless there is structure in the dictionary, minimizing $\mathcal{E}_n(x,\Delta)$ requires enumeration of all possibilities. This is hard, most of the time impossible, in practise. More specifically, this is an NP-hard problem.

11.3 Matching Pursuit

Computing an optimal M-term approximation is NP-hard. The matching pursuit algorithm computes non-optimal yet good approximations in a computationally efficient way. The algorithm was introduced by Mallat and Zhang in 1993 and is related to pursuit algorithms used in statistics and to shape-gain vector quantization.

Let $\mathcal{D} = \{g_j\}_{j\in\Gamma}$ be a dictionary of $|\Gamma| \geq N$ vectors having unit norm. The dictionary is assumed to be complete, i.e., it contains N linearly independent elements.
We consider the *orthogonal* matching pursuit algorithm.
First step of the algorithm:

$$x = \langle x, g_{j_0}\rangle\, g_{j_0} + \underbrace{Rx}_{\text{residue}}$$

$$\begin{aligned}
&Rx = x - \langle x, g_{j_0}\rangle\, g_{j_0}\\
&\langle Rx, g_{j_0}\rangle = \langle x, g_{j_0}\rangle - \langle x, g_{j_0}\rangle \underbrace{\|g_{j_0}\|^2}_{1} = 0\\
&\Rightarrow \|x\|^2 = |\langle x, g_{j_0}\rangle|^2 + \|Rx\|^2
\end{aligned}$$

To minimize $\|Rx\|$ we must choose $g_{j_0} \in \mathcal{D}$ such that $|\langle x, g_{j_0}\rangle|$ is maximum. We set $u_0 = g_{j_0}$.

Iteration: Let $R^0x = x$. Suppose that the m-th order residue R^mx is already computed for $m \geq 1$. The algorithm stops if $R^mx = 0$. Otherwise, the algorithm computes $g_{j_m} \in \mathcal{D}$ according to

$$j_m = \arg\max_{j\in\Gamma} |\langle R^m x, g_j\rangle|\,.$$

A Gram-Schmidt step orthogonalizes g_{j_m} with respect to $\{g_{j_e}\}_{0\leq e\leq m-1}$ and computes

$$u_m = g_{j_m} - \sum_{l=0}^{m-1} \frac{\langle g_{j_m}, u_l\rangle}{\|u_l\|^2} u_l.$$

The residue $R^m x$ is projected onto the orthogonal complement of the space spanned by u_m according to

$$R^{m+1}x = R^m x - \frac{\langle R^m x, u_m\rangle}{\|u_m\|^2}u_m, \quad m \geq 0$$

and so on.

$$\begin{aligned} u_m &= g_{j_m} - \sum_{l=0}^{m-1}\frac{\langle g_{j_m}, u_l\rangle}{\|u_l\|^2}u_l \\ &= \underbrace{(\mathbf{I} - \underbrace{\sum_{l=0}^{m-1}\frac{u_l u_l^H}{\|u_l\|^2}}_{\mathbb{P}_{V_{m-1}}})}_{\substack{\text{projection onto} \\ \text{orthogonal complement} \\ \text{of } V_{m-1} = \operatorname{span}\{u_0,\ldots,u_{m-1}\}}} \quad g_{j_m} \end{aligned}$$

This immediately implies that the u_i are orthogonal to each other.

$$\begin{aligned} & R^m x - \frac{\langle R^m x, u_m\rangle}{\|u_m\|^2} = R^{m+1}x \\ & (\mathbf{I} - \frac{u_m u_m^H}{\|u_m\|^2})R^m x = R^{m+1}x, \quad m \geq 0 \\ & (\mathbf{I} - \frac{u_0 u_0^H}{\|u_0\|^2})x = R^1 x \\ & (\mathbf{I} - \frac{u_1 u_1^H}{\|u_1\|^2})R^1 x = R^2 x = (\mathbf{I} - \frac{u_1 u_1^H}{\|u_1\|^2})(\mathbf{I} - \frac{u_0 u_0^H}{\|u_0\|^2})x \\ & R^m x = \underbrace{(\mathbf{I} - \frac{u_{m-1} u_{m-1}^H}{\|u_{m-1}\|^2})\ldots(\mathbf{I} - \frac{u_0 u_0^H}{\|u_0\|^2})}_{\substack{=(\mathbf{I}-\mathbb{P}_{V_{m-1}}) \\ \text{due to orthogonality of the } u_i}} x \\ & \Rightarrow R^m x = x - \mathbb{P}_{V_{m-1}}x \\ & \Rightarrow x = \mathbb{P}_{V_{m-1}}x + R^m x = \sum_{l=0}^{m-1}\frac{\langle x, u_l\rangle}{\|u_l\|^2}u_l + R^m x \end{aligned}$$

The algorithm stops after $M \leq N$ iterations and yields

$$x = \sum_{l=0}^{M-1}\frac{\langle x, u_l\rangle}{\|u_l\|^2}u_l$$

and hence

$$\|x\|^2 = \sum_{l=0}^{M-1}\frac{|\langle x, u_l\rangle|^2}{\|u_l\|^2}$$

We are guaranteed to add a linearly independent vector in each step. This can be seen as follows:

$$R^m x = \left(\mathbf{I} - \mathbb{P}_{V_{m-1}}\right) x$$
$$j_m = \arg\max_{j\in\Gamma} |\langle R^m x, g_j\rangle|$$

Notice that $R^m x$ lies in the orthogonal complement of

$$V_{m-1} = \operatorname{span}\{u_0, u_1, \ldots, u_{m-1}\} = \operatorname{span}\{g_{j_0}, g_{j_1}, \ldots, g_{j_{m-1}}\}.$$

This follows from

$$\begin{aligned}
u_0 &= g_{j_0} \\
u_1 &= g_{j_1} - \frac{\langle g_{j_1}, u_0\rangle}{\|u_0\|^2} g_{j_0} \quad \text{linear combination of } g_{j_0} \text{ and } g_{j_1} \\
u_2 &= g_{j_2} - \frac{\langle g_{j_2}, u_0\rangle}{\|u_0\|^2} \underbrace{u_0}_{g_{j_0}} - \frac{\langle g_{j_2}, u_1\rangle}{\|u_1\|^2} \underbrace{u_1}_{g_{j_1}-\alpha g_{j_0}} \quad \text{linear combination of } \{g_{j_0}, g_{j_1}, g_{j_2}\}
\end{aligned}$$

If g_j is linearly dependent from the previously chosen vectors $\{g_{j_0}, g_{j_1}, \ldots, g_{j_{m-1}}\}$, then $g_j \subset \operatorname{span}\{g_{j_0}, g_{j_1}, \ldots, g_{j_{m-1}}\}$ and hence $|\langle R^m x, g_j\rangle| = 0$. However, since the dictionary contains N linearly independent vectors, there exists a vector g_j which leads to $|\langle R^m x, g_j\rangle| > 0$ so that the algorithm will choose a vector that is linearly independent of the previously chosen vectors.
To expand x over the original dictionary vectors $\{g_j\}$, we need to perform a change of basis.

$$x = \sum_{l=0}^{M-1} \frac{\langle x, u_l\rangle}{\|u_l\|^2} u_l$$

$$\begin{aligned}
u_0 &= g_{j_0} \\
u_1 &= g_{j_1} - \frac{\langle x, u_0\rangle}{\|u_0\|^2} g_{j_0} \\
&\vdots \\
u_l &= \text{ linear combination of } \{g_{j_0}, g_{j_1}, \ldots, g_{j_l}\}.
\end{aligned}$$

Inserting these expressions for the u_i into

$$x = \sum_{l=0}^{M-1} \frac{\langle x, u_l\rangle}{\|u_l\|^2} u_l$$

yields an expansion of x in the set $\{g_{j_0}, g_{j_1}, \ldots, g_{j_{M-1}}\}$, i.e.,

$$x = \sum_{l=0}^{M-1} \alpha_l g_{j_l}. \tag{11.3}$$

Note that if $M < N$, the expansion (11.3) is an expansion into a subset of $\mathcal{D}$ of cardinality M.

11.3.1 Pursuit Recovery

We next study the approximation error of the orthogonal matching pursuit algorithm.

To simplify matters, we assume that the signals of interest have an exact representation given by a linear combination of (a few) elements of the dictionary $\mathcal{D} = \{g_j\}_{j\in\Gamma}$, i.e.,

$$x = \sum_{j\in\Lambda} a[j] g_j.$$

We want to establish a sufficient condition for OMP to recover x exactly, i.e., to identify the set $\{g_j\}_{j\in\Lambda}$. Moreover, we will find a sufficient condition that not only guarantees exact recovery, but also guarantees that the algorithm provides the correct answer with at most $|\Lambda|$ iterations.
Recall that in step m the algortithm selects an element in Λ if and only if

$$C(R^m x, \Lambda^c) = \frac{\max_{k\in\Lambda^c} |\langle R^m x, g_k\rangle|}{\max_{j\in\Lambda} |\langle R^m x, g_j\rangle|} < 1$$

where Λ^c is the complement of Λ in Γ. Generally, we define

$$C(h, \Lambda^c) = \frac{\max_{k\in\Lambda^c} |\langle h, g_k\rangle|}{\max_{j\in\Lambda} |\langle h, g_j\rangle|}.$$

Theorem 11.6. *If $\{\tilde{g}_{j,\Lambda}\}_{j\in\Lambda}$ is the dual basis (frame) of $\{g_{j,\Lambda}\}_{j\in\Lambda}$ in the space V_Λ, then the exact recovery condition (*ERC*) is given by*

$$\mathrm{ERC}(\Lambda) \triangleq \max_{k\in\Lambda^c} \sum_{j\in\Lambda} |\langle \tilde{g}_{j,\Lambda}, g_k\rangle| = \sup_{h\in V_\Lambda} C(h, \Lambda^c).$$

If $x \in V_\Lambda$ and $\mathrm{ERC}(\Lambda) < 1$, then the orthogonal matching pursuit algorithm recovers x exactly with at most $|\Lambda|$ iterations.

Proof. We first prove that

$$\sup_{h\in V_\Lambda} C(h, \Lambda^c) \leq \mathrm{ERC}(\Lambda).$$

Let Φ_Λ be the analysis operator corresponding to $\{g_{j,\Lambda}\}_{j\in\Lambda}$. It then follows that $\Phi_\Lambda^\dagger \Phi_\Lambda$ is an orthogonal projection onto V_Λ, and hence for $h \in V_\Lambda$ and $k \in \Lambda^c$, we have

$$|\langle h, g_k\rangle| = \left|\left\langle \Phi_\Lambda^\dagger \Phi_\Lambda h, g_k \right\rangle\right| = \left|\left\langle \Phi_\Lambda h, (\Phi_\Lambda^\dagger)^H g_k \right\rangle\right| \leq \|\Phi_\Lambda h\|_\infty \max_{k\in\Lambda^c} \|(\Phi_\Lambda^\dagger)^H g_k\|_1.$$

The analysis operator corresponding to the dual frame satisfies $\tilde{\Phi}_\Lambda^H = \Phi_\Lambda^\dagger$ and thus

$$(\Phi_\Lambda^\dagger)^H x = \begin{pmatrix} \langle x, \tilde{g}_{j_1,\Lambda}\rangle \\ \vdots \end{pmatrix}.$$

We therefore get

$$\mathrm{ERC}(\Lambda) = \max_{k\in\Lambda^c} \sum_{j\in\Lambda} |\langle \tilde{g}_{j,\Lambda}, g_k\rangle| = \max_{k\in\Lambda^c} \|(\Phi_\Lambda^\dagger)^H g_k\|_1$$

Frame-Theoretic Viewpoint

The vectors $\{g_j\}_{j\in\Lambda}$ form a frame for their span as they are linearly independent.

Define the corresponding analysis operator as

$$(\Phi_\Lambda x)_j = \begin{pmatrix} \langle x, g_1\rangle \\ \langle x, g_2\rangle \\ \vdots \\ \langle x, g_{|\Lambda|}\rangle \end{pmatrix} \qquad \Phi_\Lambda = \begin{pmatrix} g_1^H \\ g_2^H \\ \vdots \\ g_{|\Lambda|}^H \end{pmatrix}.$$

The right-inverse (pseudo-inverse) of Φ_Λ is given by

$$\Phi_\Lambda^\dagger = \Phi_\Lambda^H(\Phi_\Lambda\Phi_\Lambda^H)^{-1}.$$

From a basic result of pseudo-inverses it follows that $\Phi_\Lambda^\dagger\Phi_\Lambda$ is the orthogonal projector onto $\mathcal{R}(\Phi_\Lambda^H) = V_\Lambda = \operatorname{span}\{g_j\}_{j\in\Lambda}$.
The dual frame elements are given by

$$\Phi_\Lambda^H(\Phi_\Lambda\Phi_\Lambda^H)^{-1} = \tilde{\Phi}_\Lambda^H = \Phi_\Lambda^\dagger.$$

Proof continued:
...and hence by

$$|\langle h, g_k\rangle| \leq \|\Phi_\Lambda h\|_\infty \max_{k\in\Lambda^c}\|(\Phi_\Lambda^\dagger)^H g_k\|_1$$

we obtain

$$\max_{k\in\Lambda^c}|\langle h, g_k\rangle| \leq \|\Phi_\Lambda h\|_\infty \mathrm{ERC}(\Lambda)$$

and hence

$$\mathrm{ERC}(\Lambda) \geq \frac{\max_{k\in\Lambda^c}|\langle h, g_k\rangle|}{\|\Phi_\Lambda h\|_\infty} = \frac{\max_{k\in\Lambda^c}|\langle h, g_k\rangle|}{\max_{j\in\Lambda}|\langle h, g_j\rangle|} = C(h,\Lambda^c)$$
$$\Rightarrow \mathrm{ERC}(\Lambda) \geq \sup_{h\in V_\Lambda} C(h,\Lambda^c).$$

We now prove the reverse inequality, i.e.,

$$\mathrm{ERC}(\Lambda) \leq \sup_{\mathbf{h}\in V_\Lambda} C(h,\Lambda^c)$$

Let $k_0 \in \Lambda^c$ be the index such that

$$\sum_{j\in\Lambda}|\langle \tilde{g}_{j,\Lambda}, g_{k_0}\rangle| = \max_{k\in\Lambda^c}\sum_{j\in\Lambda}|\langle \tilde{g}_{j,\Lambda}, g_k\rangle|.$$

Introducing

$$h = \sum_{j\in\Lambda} \mathrm{sign}(\langle \tilde{g}_{j,\Lambda}, g_{k_0}\rangle)\tilde{g}_{j,\Lambda} \in V_\Lambda$$

leads to

$$\begin{aligned}\mathrm{ERC}(\Lambda) &= \max_{k\in\Lambda^c} \sum_{j\in\Lambda} |\langle \tilde{g}_{j,\Lambda}, g_k\rangle| = |\langle h, g_{k_0}\rangle| \\ &\leq \max_{k\in\Lambda^c} |\langle h, g_k\rangle| = C(h,\Lambda^c) \max_{j\in\Lambda} |\langle h, g_j\rangle| \,.\end{aligned}$$

Since

$$|\langle h, g_l\rangle| = \left| \left\langle \sum_{j\in\Lambda} \mathrm{sign}(\langle \tilde{g}_{j,\Lambda}, g_{k_0}\rangle)\tilde{g}_{j,\Lambda}, g_l \right\rangle \right| = |\mathrm{sign}(\langle \tilde{g}_{l,\Lambda}, g_{k_0}\rangle)| \leq 1$$

where the second equality follows from biorthogonality of $\{g_j\}$ and $\{\tilde{g}_j\}$, it follows that

$$C(h,\Lambda^c) \geq \mathrm{ERC}(\Lambda)$$

and thus

$$\mathrm{ERC}(\Lambda) = \sup_{\mathbf{h}\in V_\Lambda} C(h,\Lambda^c).$$

To prove the second part of the theorem, we suppose that $x = R^0 x \in V_\Lambda$ and $\mathrm{ERC}(\Lambda) < 1$. We prove by induction that the OMP algorithm selects only vectors in $\{g_j\}_{j\in\Lambda}$. Suppose that the first m matching pursuit vectors are in $\{g_j\}_{j\in\Lambda}$. Therefore, since

$$R^m x = (\mathbf{I} - \mathbb{P}_{V_{m-1}})x = \left(\mathbf{I} - \frac{u_{m-1}u_{m-1}^H}{\|u_{m-1}\|^2} \ldots - \frac{u_0 u_0^H}{\|u_0\|^2}\right) x$$

it follows that $R^m x \in V_\Lambda$ as well. If $R^m x \neq 0$, then $\mathrm{ERC}(\Lambda) < 1$ implies that $C(R^m x, \Lambda^c) < 1$ and thus that the next vectors selected is in Λ. Since $\dim(V_\Lambda) \leq |\Lambda|$ and the OMP is guaranteed to select a linearly independent vector in each step, it follows that the OMP converges in at most $|\Lambda|$ iterations. In the $|\Lambda|$-th step we are left with $R^{|\Lambda|}x = (\mathbf{I} - \mathbb{P}_{V_{|\Lambda|-1}})x = x - x = 0$ and hence the algorithm stops. □

The next result relates the ERC to the coherence of the dictionary $\mathcal{D}$.

Definition 11.7. The dictionary mutual coherence is defined as

$$\mu(\mathcal{D}) = \max_{\substack{(j,k)\in\Gamma^2 \\ j\neq k}} |\langle g_j, g_k\rangle| \,.$$

Theorem 11.8.

$$\mathrm{ERC}(\Lambda) \leq \frac{|\Lambda|\mu(\mathcal{D})}{1 - (|\Lambda| - 1)\mu(\mathcal{D})}$$

if $|\Lambda| < 1 + \frac{1}{\mu(\mathcal{D})}$.

Proof. It is shown in the proof of the previous theorem that

$$\begin{aligned}\mathrm{ERC}(\Lambda) &= \max_{k\in\Lambda^c}\|(\Phi_\Lambda^\dagger)^H g_k\|_1\\ \Phi_\Lambda^\dagger &= \Phi_\Lambda^H(\Phi_\Lambda\Phi_\Lambda^H)^{-1}\end{aligned}$$

$\Phi_\Lambda^\dagger\Phi_\Lambda$ is the orthogonal projector onto the range of Φ_Λ^H and hence onto V_Λ (by the definition of Φ_Λ).

$$\Rightarrow \mathrm{ERC}(\Lambda) = \max_{k\in\Lambda^c}\|(\Phi_\Lambda\Phi_\Lambda^H)^{-1}\Phi_\Lambda g_k\|_1.$$

Defining the l_1-operator-norm as

$$\|\mathbb{A}\|_{1,1} = \max_j \sum_i |a_{i,j}| = \max_{h\neq 0}\frac{\|\mathbb{A}h\|_1}{\|h\|_1}$$

leads to

$$\mathrm{ERC}(\Lambda) = \max_{k\in\Lambda^c}\|(\Phi_\Lambda\Phi_\Lambda^H)^{-1}\Phi_\Lambda g_k\|_1 \leq \|(\Phi_\Lambda\Phi_\Lambda^H)^{-1}\|_{1,1}\max_{k\in\Lambda^c}\|\Phi_\Lambda g_k\|_1$$

where we used that $\|\mathbb{A}x\|_1 \leq \|\mathbb{A}\|_{1,1}\|x\|_1$.

$$\max_{k\in\Lambda^c}\|\Phi_\Lambda g_k\|_1 = \max_{k\in\Lambda^c}\sum_{j\in\Lambda}|\langle g_k, g_j\rangle|$$

Neumann series expansion of $\Phi_\Lambda\Phi_\Lambda^H = \mathbb{I} + \mathbb{H}$ gives

$$\|(\Phi_\Lambda\Phi_\Lambda^H)^{-1}\|_{1,1} = \|\sum_{k=0}^{\infty}(-1)^k\mathbb{H}^k\|_{1,1} \leq \sum_{k=0}^{\infty}(\|\mathbb{H}\|_{1,1})^k \leq \frac{1}{1-\|\mathbb{H}\|_{1,1}}$$

with

$$\|\mathbb{H}\|_{1,1} = \max_{k\in\Lambda}\sum_{j\in\Lambda, j\neq k}|\langle g_j, g_k\rangle| \leq (|\Lambda|-1)\mu(\mathcal{D})$$

$$\mathrm{ERC}(\Lambda) \leq \frac{1}{1-(|\Lambda|-1)\mu(\mathcal{D})}\underbrace{\max_{k\in\Lambda^c}\sum_{j\in\Lambda}|\langle g_j, g_k\rangle|}_{\leq|\Lambda|\mu(\mathcal{D})} = \frac{|\Lambda|\mu(\mathcal{D})}{1-(|\Lambda|-1)\mu(\mathcal{D})}.$$

□

Combination of the two last theorems therefore yields the recovery condition

$$\begin{aligned}&\mathrm{ERC}(\Lambda) \leq \frac{|\Lambda|\mu(\mathcal{D})}{1-(|\Lambda|-1)\mu(\mathcal{D})} < 1\\ &|\Lambda|\mu(\mathcal{D}) < 1 - |\Lambda|\mu(\mathcal{D}) + \mu(\mathcal{D})\\ &|\Lambda|\mu(\mathcal{D}) < \frac{1}{2}(1+\mu(\mathcal{D}))\\ &|\Lambda| < \frac{1}{2}\left(1+\frac{1}{\mu(\mathcal{D})}\right)\end{aligned}$$

Chapter 12

Uniform laws of large numbers

This chapter is taken verbatimly from [84] with minor additional explanations and modifications.

Uniform laws of large numbers play an important role in the understanding of different types of estimators, are an essential tool in statistical learning theory, and constitute an entry point into the field of empirical process theory.

12.1 Uniform convergence of cumulative distributions functions

The law of any scalar random variable X can be fully specified by its cumulative distribution function (CDF), specified by $F(t) = \mathbb{P}[X \leq t], t \in \mathbb{R}$. Now, suppose we want to estimate $F(t)$ from a given collection $\{X_i\}_{i=1}^n$ of i.i.d. samples, each drawn according to the law specified by F. A natural estimate of F is the empirical CDF given by

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{(-\infty;t]}[X_i], \tag{12.1}$$

where $\mathbb{1}_{(-\infty;t]}[x]$ is the indicator function for the event $\{x \leq t\}$. This estimator is inspired by the fact that we can write

$$\begin{aligned} F(t) &= \mathbb{E}\big[\mathbb{1}_{(-\infty;t]}[X]\big] \\ &= \int_{-\infty}^{\infty} \mathbb{1}_{(-\infty;t]}[x] f(x)\, dx \\ &= \int_{-\infty}^{t} f(x)\, dx, \end{aligned} \tag{12.2}$$

hence (12.1) is simply an estimator for the expectation in (12.2). Note that $\hat{F}_n(t)$ is a random function, which for $n \to \infty$ approaches $F(t)$. In particular, we have

$$\begin{aligned} \mathbb{E}\left[\hat{F}_n(t)\right] &= \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}\left[\mathbb{1}_{(-\infty;t]}[X_i]\right] \\ &= \frac{1}{n}\sum_{i=1}^{n} F(t) = F(t). \end{aligned}$$

In fact, one can show that $\hat{F}_n(t)$ converges in an almost sure sense to $F(t)$, namely

$$\mathbb{P}\left[\lim_{n\to\infty} \hat{F}_n(t) = F(t)\right] = 1, \quad \forall t \in \mathbb{R}.$$

This says that events for which $\hat{F}_n(t)$ does not converge to $F(t)$ have probability 0.

We will however be interested in a more general result, termed uniform convergence. To develop this concept, we start by noting that in statistical settings, a typical use of the empirical CDF is often to construct estimators of various quantities associated with the population CDF $F(t)$. Many such estimation problems can be formulated in terms of a functional γ that maps a CDF F to a real number $\gamma(F)$, i.e., $F \mapsto \gamma(F)$. Given a set of samples distributed according to F, the plug-in principle suggests replacing the unknown F with the empirical CDF $\hat{F}_n$, thereby obtaining $\gamma(\hat{F}_n)$ as an estimate of $\gamma(F)$. The question is now under which conditions do we get convergence of $\gamma(\hat{F}_n)$ to $\gamma(F)$ as $n \to \infty$ and in what sense this convergence occurs. Before attempting to answer this question in generality, let us consider some examples illustrating the plug-in principle.

Example 12.1 (Expectation functionals)**.** Given some integrable function g, we define the expectation functional γ_g via

$$\gamma_g(F) = \mathbb{E}\left[g(X)\right] = \int g(x)f(x)\,dx = \int g(x)\,dF(x).$$

For instance, for $g(x) = x$, the functional $\gamma_g(F)$ maps F to $\mathbb{E}[X]$, where X is a random variable with CDF F. For any g, the plug-in-estimate is given by $\gamma_g(\hat{F}_n) = \frac{1}{n}\sum_{i=1}^{n} g(X_i)$, corresponding to the sample mean of $g(X)$. In the special case $g(x) = x$, we recover the usual sample mean $\frac{1}{n}\sum_{i=1}^{n} X_i$ as an estimate of the mean $\mu = \mathbb{E}[X]$.

Example 12.2 (Quantile functionals)**.** For any $\alpha \in [0, 1]$, the quantile functional Q_α is given by

$$Q_\alpha(F) = \inf\{t \in \mathbb{R} \mid F(t) \geq \alpha\}.$$

The median corresponds to $\alpha = 0.5$. The plug-in-estimate is given by

$$Q_\alpha(\hat{F}_n) = \inf\left\{t \in \mathbb{R} \,\middle|\, \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty;t]}[X_i] \geq \alpha\right\}$$

and corresponds to estimating the α-th quantile of the distribution by the α-th sample quantile. Again, it is of interest to determine in what sense (if any) the random variable $Q_\alpha(\hat{F}_n)$ converges to $Q_\alpha(F)$ as n becomes large. We note that here $Q_\alpha(\hat{F}_n)$ is a fairly complicated non linear function of all the variables.

Example 12.3 (Goodness-of-fit functionals)**.** It is frequently of interest to test the hypothesis of whether or not a given set of data has been drawn from a known distribution F_0. Such tests can be performed using functionals that measure the distance between F and the target CDF F_0, including the sup-norm distance $\|F - F_0\|_\infty$ or other distances such as Cramér-von Mises criterion based on the functional $\gamma(F) = \int_{-\infty}^{\infty} [F(x) - F_0(x)]^2 \, dF_0(x)$.

For any plug-in estimator $\gamma(\hat{F}_n)$, an important question is to understand when $\gamma(\hat{F}_n)$ converges to $\gamma(F)$ in probability (or almost surely). We recall the notion of convergence in probability.

Definition 12.4. We say that a sequence of random variables X_n converges in probability to the random variable X if

$$\lim_{n\to\infty} \mathbb{P}\left\{|X - X_n| > \varepsilon\right\} = 0, \quad \forall \varepsilon > 0.$$

The question of convergence of $\gamma(\hat{F}_n)$ to $\gamma(F)$ can be addressed in a unified manner for many functionals through a suitable notion of continuity. Given a pair of CDFs F and G, let us measure the distance between them using the sup-norm

$$\|G - F\|_\infty = \sup_{t\in\mathbb{R}} |G(t) - F(t)|\,.$$

We can then define the continuity of a functional γ with respect to this norm as follows.

Definition 12.5. The functional γ is continuous at F in the sup-norm if, for all $\varepsilon > 0$, there exists a $\delta > 0$ such that $\|G - F\|_\infty \leq \delta$ implies $|\gamma(G) - \gamma(F)| \leq \varepsilon$.

This notion is useful as for any functional that is continuous w.r.t. the sup-norm, the plug-in estimator is in probability consistent as an estimator of $\gamma(F)$. This statement follows by combining continuity of the functional with the Glivenko-Cantelli Theorem.

We need to show that

$$\lim_{n\to\infty} \mathbb{P}\left[|\gamma(\hat{F}_n) - \gamma(F)| > \varepsilon\right] = 0, \quad \forall \varepsilon > 0.$$

Now, by continuity of $\gamma(\cdot)$ w.r.t. $\|\cdot\|_\infty$-norm, we get that for each $\varepsilon > 0$, there exists a $\delta > 0$ such that $\|\hat{F}_n - F\|_\infty \leq \delta$ implies that $|\gamma(\hat{F}_n) - \gamma(F)| \leq \varepsilon$, i.e.,

$$\|\hat{F}_n - F\|_\infty \leq \delta \implies |\gamma(\hat{F}_n) - \gamma(F)| \leq \varepsilon.$$

Hence, it follows that

$$|\gamma(\hat{F}_n) - \gamma(F)| > \varepsilon \implies \|\hat{F}_n - F\|_\infty > \delta,$$

which implies

$$\mathbb{P}\left[|\gamma(\hat{F}_n) - \gamma(F)| > \varepsilon\right] \leq \mathbb{P}\left[\|\hat{F}_n - F\|_\infty > \delta\right]. \tag{12.3}$$

Now, by the classical Glivenko-Cantelli theorem, we get $\|\hat{F}_n - F\|_\infty \overset{a.s.}{\to} 0$ and hence, as a.s. convergence implies convergence in probability, $\|\hat{F}_n - F\|_\infty \overset{p.}{\to} 0$, i.e.,

$$\lim_{n\to\infty} \mathbb{P}\left[\|\hat{F}_n - F\|_\infty > \varepsilon'\right] = 0, \quad \forall \varepsilon' > 0.$$

This establishes that

$$\lim_{n\to\infty} \mathbb{P}\left[|\gamma(\hat{F}_n) - \gamma(F)| > \varepsilon\right] \leq \lim_{n\to\infty} \mathbb{P}\left[\|\hat{F}_n - F\|_\infty > \delta\right] = 0$$

and hence

$$\lim_{n\to\infty} \mathbb{P}\left[|\gamma(\hat{F}_n) - \gamma(F)| > \varepsilon\right] = 0, \quad \forall \varepsilon > 0,$$

as desired.

Theorem 12.6 (Glivenko-Cantelli)**.** *For any distribution, the empirical CDF $\hat{F}_n$ satisfies*

$$\|\hat{F}_n - F\|_\infty \overset{a.s.}{\to} 0.$$

12.2 Uniform laws for more general function classes

We now consider uniform laws for more general function classes. Let $\mathcal{F}$ be a class of integrable real-valued functions with domain $\mathcal{X}$, and let $\{X_i\}_{i=1}^n$ be a collection of i.i.d. samples from some distribution $\mathbb{P}$ over $\mathcal{X}$. Consider the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} = \sup_{f\in\mathcal{F}} \left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}[f(X)]\right|.$$

Definition 12.7. We say that $\mathcal{F}$ is a Glivenko-Cantelli class for $\mathbb{P}$ if $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ converges to zero in probability as $n \to \infty$.

A stronger notion of this concept requires almost sure convergence and leads to what is referred to as the strong Glivenko-Cantelli property. Recall that almost sure convergence implies convergence in probability.

Example 12.8 (Empirical CDFs and indicator functions)**.** Consider the function class

$$\mathcal{F} = \left\{\mathbb{1}_{(-\infty;t]}(\cdot) \mid t \in \mathbb{R}\right\}. \tag{12.4}$$

For each fixed $t \in \mathbb{R}$, we have

$$\mathbb{E}\left[\mathbb{1}_{(-\infty;t]}(X)\right] = \mathbb{P}[X \leq t] = F(t).$$

Hence, the classical Glivenko-Cantelli result $\|\hat{F}_n - F\|_\infty \overset{a.s.}{\to} 0$, i.e.,

$$\mathbb{P}\left[\lim_{n\to\infty} \sup_{t\in\mathbb{R}} |\hat{F}_n(t) - F(t)| = 0\right] = 1 \tag{12.5}$$

can be interpreted as a strong uniform law for the class $\mathcal{F}$. Specifically, this is seen by noting that (12.5) can be rewritten as

$$\mathbb{P}\Bigg[\lim_{n\to\infty} \sup_{t\in\mathbb{R}} \Bigg| \underbrace{\frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{(-\infty;t]}(X_i)}_{\hat{F}_n(t)} - \mathbb{E}\left[\mathbb{1}_{(-\infty;t]}(X)\right] \Bigg| = 0\Bigg] = 1,$$

which, upon using (12.4) is nothing but

$$\mathbb{P}\Bigg[\lim_{n\to\infty} \sup_{f\in\mathcal{F}} \Bigg|\frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}[f(X)]\Bigg| = 0\Bigg] = 1.$$

We note that not all function classes are Glivenko-Cantelli. Variables of the form $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$ are prevalent in statistics, in fact, they lie at the heart of empirical risk minimization. In order to describe these ideas, let us consider an indexed family of probability distributions $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$, and suppose that we are given n samples $\{X_i\}_{i=1}^n$, each sample lying in some space $\mathcal{X}$, and with the samples drawn i.i.d. according to a distribution $\mathbb{P}_{\theta^*}$, for some fixed but unknown $\theta^* \in \Omega$. Here, the index θ^* could lie within a finite-dimensional space, such as $\Omega = \mathbb{R}^d$ in a vector estimation problem, or it could lie within a function class $\Omega = \mathcal{G}$, in which case the problem is of the non-parametric variety.

A standard approach to estimating θ^* is based on minimizing a cost function $\mathcal{L}_\theta(X)$, which measures the "fit" between a parameter $\theta \in \Omega$ and the sample $X \in \mathcal{X}$. Given the collection of n samples $\{X_i\}_{i=1}^n$, the principle of empirical risk minimization is based on the objective function

$$\hat{R}_n(\theta, \theta^*) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_\theta(X_i).$$

This quantity is known as the empirical risk. We will also need the population risk defined as

$$R(\theta, \theta^*) = \mathbb{E}_{\theta^*}[\mathcal{L}_\theta(X)],$$

where the expectation $\mathbb{E}_{\theta^*}$ is taken over a sample $X \sim \mathbb{P}_{\theta^*}$. In practice, one minimizes the empirical risk over some subset Ω_0 of the full space Ω, thereby obtaining some estimate $\hat{\theta}$. The question is now how to bound the excess risk, measured in terms of population quantities—namely the difference

$$E(\hat{\theta}, \theta^*) = R(\hat{\theta}, \theta^*) - \inf_{\theta\in\Omega_0} R(\theta, \theta^*).$$

Example 12.9 (Maximum likelihood)**.** Consider a parametrized family of distributions, say $\{\mathbb{P}_\theta \mid \theta \in \Omega\}$, each with a strictly positive density p_θ defined with respect to a common underlying measure.

We want to estimate the unknown parameter θ^* from n i.i.d. samples of distribution p_{θ^*}. Consider the cost function

$$\mathcal{L}_\theta(x) = \log\left[\frac{p_{\theta^*}(x)}{p_\theta(x)}\right].$$

The term $p_{\theta^*}(x)$ is included for later convenience and has no effect on the minimization over θ. The maximum likelihood estimate is given by

$$\begin{aligned}\hat{\theta} &= \arg\min_{\theta\in\Omega_0}\left\{\frac{1}{n}\sum_{i=1}^n \log\left[\frac{p_{\theta^*}(X_i)}{p_\theta(X_i)}\right]\right\}\\ &= \arg\min_{\theta\in\Omega_0}\left\{\frac{1}{n}\sum_{i=1}^n \log\left[\frac{1}{p_\theta(X_i)}\right]\right\}.\end{aligned}$$

The population risk is given by $R(\theta,\theta^*) = \mathbb{E}_{\theta^*}\left[\log\left[\frac{p_{\theta^*}(X)}{p_\theta(X)}\right]\right]$, a quantity known as the Kullback-Leibler divergence between p_{θ^*} and p_θ, typically denoted as $D(p_{\theta^*}\,\|\,p_\theta)$. Next, we note that

$$\inf_{\theta\in\Omega_0} R(\theta,\theta^*) = \inf_{\theta\in\Omega_0}\mathbb{E}_{\theta^*}\left[\log\left[\frac{p_{\theta^*}(X)}{p_\theta(X)}\right]\right].$$

We now use a basic property of Kullback-Leibler divergence, namely $D(p\,\|\,q)\geq 0$ with equality if and only if $p(x) = q(x),\ \forall x$. For $\theta^*\in\Omega_0$, this implies

$$\inf_{\theta\in\Omega_0}\mathbb{E}_{\theta^*}\left[\log\left[\frac{p_{\theta^*}(X)}{p_\theta(X)}\right]\right] = \mathbb{E}_{\theta^*}\left[\log\left[\frac{p_{\theta^*}(X)}{p_{\theta^*}(X)}\right]\right] = 0.$$

The excess risk hence becomes

$$\begin{aligned}E(\hat{\theta},\theta^*) &= R(\hat{\theta},\theta^*) - \underbrace{\inf_{\theta\in\Omega_0} R(\theta,\theta^*)}_{=0}\\ &= R(\hat{\theta},\theta^*) = D(p_{\theta^*}\,\|\,p_{\hat{\theta}}),\end{aligned}$$

which, again by $D(p\,\|\,q)\geq 0$ with equality iff $p=q$, says that the excess risk is positive for $\hat{\theta}\neq\theta^*$ and equals zero in the case of a perfect estimate $\hat{\theta}=\theta^*$.

Returning to our main goal, we proceed to decompose the excess risk as follows. First, we assume that there exists some $\theta_0\in\Omega_0$ such that $R(\theta_0,\theta^*) = \inf_{\theta\in\Omega_0} R(\theta,\theta^*)$. The excess risk can now be decomposed as

$$E(\hat{\theta},\theta^*) = \underbrace{R(\hat{\theta},\theta^*) - \hat{R}_n(\hat{\theta},\theta^*)}_{T_1} + \underbrace{\hat{R}_n(\hat{\theta},\theta^*) - \hat{R}_n(\theta_0,\theta^*)}_{T_2} + \underbrace{\hat{R}_n(\theta_0,\theta^*) - R(\theta_0,\theta^*)}_{T_3}.$$

Note that $T_2\leq 0$ as $\hat{\theta}$ minimizes the empirical risk over Ω_0. The term

$$T_3 = \frac{1}{n}\sum_{i=1}^n \mathcal{L}_{\theta_0}(X_i) - \mathbb{E}_X[\mathcal{L}_{\theta_0}(X)]$$

corresponds to the deviation of a sample mean from its expectation and can, in principle, be bounded through concentration inequalities as θ_0 is a deterministic quantity.

Finally, we note that $T_1 = \mathbb{E}_X[\mathcal{L}_{\hat{\theta}}(X)] - \frac{1}{n}\sum_{i=1}^n \mathcal{L}_{\hat{\theta}}(X_i)$. This quantity is more challenging to deal with as $\hat{\theta}$ is a random variable and by virtue of being the minimizer of the empirical risk, depends on the samples $\{X_i\}_{i=1}^n$. It can be bounded by identifying the cost function $\mathcal{L}_\theta(X)$, parametrized by θ, with the function class $\mathcal{F}$, specifically by setting $\mathcal{F} = \{\mathcal{L}_\theta(\cdot) \mid \theta \in \Omega_0\}$ and noting that

$$T_1 \leq \sup_{\theta\in\Omega_0} \left| \frac{1}{n}\sum_{i=1}^n \mathcal{L}_\theta(X_i) - \mathbb{E}_X[\mathcal{L}_\theta(X)] \right| = \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}.$$

As $T_3 \leq \|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, it follows that the excess risk is upper-bounded by $2\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$, which demonstrates that the central challenge in analyzing estimates based on empirical risk minimization is to establish uniform laws of large numbers.

12.3 Uniform laws via Rademacher complexity

We can now turn to the technical details of deriving uniform laws of large numbers. An important quantity in this context is the Rademacher complexity of a function class $\mathcal{F}$. For any fixed collection $x_1^n = (x_1, \ldots, x_n)$ of points, consider the subset of $\mathbb{R}^n$ given by

$$\mathcal{F}(x_1^n) = \{(f(x_1), f(x_2), \ldots, f(x_n)) \mid f \in \mathcal{F}\},$$

i.e., the set of vectors in $\mathbb{R}^n$ that can be realized by applying a function $f \in \mathcal{F}$ to the collection $(x_1, \ldots, x_n)$. We define the empirical Rademacher complexity as

$$\mathcal{R}(\mathcal{F}(x_1^n)/n) = \mathbb{E}_\varepsilon \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i) \right| \right],$$

where $\{\varepsilon_k\}_{k=1}^n$ is an i.i.d. sequence of Rademacher random variables (i.e., taking values $\{-1, +1\}$ equiprobably). Given a collection $X_1^n = \{X_i\}_{i=1}^n$ of random variables, the corresponding empirical Rademacher complexity $\mathcal{R}(\mathcal{F}(X_1^n)/n)$ is a random variable. The expectation of $\mathcal{R}(\mathcal{F}(X_1^n)/n)$ yields the Rademacher complexity of the function class, namely the deterministic quantity

$$\begin{aligned} \mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_X[\mathcal{R}(\mathcal{F}(X_1^n)/n)] \\ &= \mathbb{E}_{\varepsilon,X} \left[\sup_{f\in\mathcal{F}} \left| \frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i) \right| \right]. \end{aligned}$$

Note that the Rademacher complexity is the average of the maximum correlation between the vector $(f(X_1), \ldots, f(X_n))$ and the "noise vector" $(\varepsilon_1, \ldots, \varepsilon_n)$ with the maximum taken over all functions $f \in \mathcal{F}$. The intuition is as follows: a function class is large—potentially "too large" for certain statistical purposes— if we can always find a function that has high correlation with a randomly drawn “noise” vector. Conversely, when the Rademacher complexity decays as a function of sample size n, then it is impossible to find a function that correlates highly with a randomly drawn noise vector.

The significance of Rademacher complexity for the Glivenko-Cantelli property is expressed through the following result.

Theorem 12.10. *For any b-uniformly bounded function class, i.e., $\|f\|_\infty \leq b$, for all $f \in \mathcal{F}$, any positive integer $n \geq 1$ and any scalar $\delta \geq 0$, we have*

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq 2\mathcal{R}_n(\mathcal{F}) + \delta$$

with $\mathbb{P}$-probability at least $1 - e^{-\frac{n\delta^2}{2b^2}}$. Consequently, as long as $\mathcal{R}_n(\mathcal{F}) = o(1)$, we have $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \overset{a.s.}{\to} 0$.

Proof. The proof of Theorem 12.10 is effected in two steps. The first step consists of showing that, for a sequence $\{X_i\}_{i=1}^n$ of i.i.d. random variables, the random variable

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n f(X_i) - \mathbb{E}\big[f(X)\big] \right|$$

is sharply concentrated around its mean. The second step consists of showing that the mean can be upper bounded by the Rademacher complexity up to a constant pre-factor.

First step: In order to simplify the notation, it is convenient to define the recentered functions $\bar{f}(x) := f(x) - \mathbb{E}[f(X)]$. Thinking of the samples as fixed for the moment, consider the function

$$G(x_1, \dots, x_n) := \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right|.$$

We claim that G satisfies the bounded difference property required to apply Lemma 2.3. Since the function G is invariant to permutation of its coordinates, it suffices to bound the difference when the first coordinate x_1 is perturbed. Accordingly, we define the vector $y \in \mathbb{R}$ with $y_i = x_i$ for all $i = 2, \dots, n$, and seek to bound the difference $|G(x) - G(y)|$. For any function f, we have

$$\begin{aligned}
\left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right| - \sup_{h \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^n \bar{h}(y_i) \right| &\leq \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) \right| - \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(y_i) \right| \\
&\overset{(*)}{\leq} \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(x_i) - \bar{f}(y_i) \right| + \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(y_i) \right| - \left| \frac{1}{n} \sum_{i=1}^n \bar{f}(y_i) \right| \\
&= \frac{1}{n} \left| \bar{f}(x_1) - \bar{f}(y_1) \right| \\
&\overset{(**)}{\leq} \frac{2b}{n},
\end{aligned}$$

where $(*)$ holds by the triangle inequality and the final inequality $(**)$ comes from the uniform boundedness assumption $\|f\|_\infty \leq b$. Since the inequality holds for any function f, we may take the supremum over $f \in \mathcal{F}$ on both sides to obtain

$$G(x) - G(y) \leq \frac{2b}{n}.$$

Repeating the same arguments with the roles of x and y reversed, we conclude that

$$|G(x) - G(y)| \leq \frac{2b}{n}.$$

Therefore, replacing the x_i with the X_i and taking the expectation, we have, from a one sided version of Lemma 2.3, with probability at least $1 - \exp\{-\frac{n\delta^2}{2b^2}\}$, that

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \leq \mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big] + \delta, \tag{12.6}$$

for all $\delta \geq 0$.

Second step: It remains to show that $\mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big]$ is upper-bounded by $2\mathcal{R}_n(\mathcal{F})$, which will be accomplished through a *symmetrization* argument. Specifically, let $\{Y_i\}_{i=1}^n$ be a second i.i.d. sequence sampled from $\mathbb{P}$, independent of $\{X_i\}_{i=1}^n$ and note that

$$\begin{aligned}
\mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \Big(f(X_i) - \mathbb{E}_{Y_i}\big[f(Y_i)\big]\Big)\right|\right] \\
&= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\mathbb{E}_Y\left[\frac{1}{n}\sum_{i=1}^n \big(f(X_i) - f(Y_i)\big)\right]\right|\right] \\
&\leq \mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \big(f(X_i) - f(Y_i)\big)\right|\right].
\end{aligned}$$

Now, let $(\varepsilon_1, \dots, \varepsilon_n)$ be an i.i.d. sequence of Rademacher random variables, i.e. such that ε_i takes values in $\{-1, 1\}$ equiprobably, independent of X and Y. For any function $f \in \mathcal{F}$, given that ε_i, X_i and Y_i are independent and that X_i and Y_i have the same distribution, the random vector with components $\varepsilon_i(f(X_i) - f(Y_i))$ has the same joint distribution as the random vector with components $f(X_i) - f(Y_i)$, hence

$$\begin{aligned}
\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \big(f(X_i) - f(Y_i)\big)\right|\right] &= \mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i(f(X_i) - f(Y_i))\right|\right] \\
&\leq 2\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] = 2\mathcal{R}_n(\mathcal{F}).
\end{aligned} \tag{12.7}$$

Combining (12.6) and (12.7) establishes the proof. □

For this result to be useful, we need to develop techniques to upper-bound the Rademacher complexity. Before doing that, we establish a lower bound on the random variable $\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}$.

Proposition 12.11. For any b-uniformly bounded function class $\mathcal{F}$, any integer $n \geq 1$ and any scalar $\delta \geq 0$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq \frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}} |\mathbb{E}_{\mathbb{P}}[f]|}{2\sqrt{n}} - \delta$$

with $\mathbb{P}$-probability at least $1 - e^{-\frac{n\delta^2}{2b^2}}$.

Proof. By a similar argument as in the first step of the proof of Theorem 12.10, we show that, with probability at least $1 - \exp\{-\frac{n\delta^2}{2b^2}\}$, we have

$$\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}} \geq \mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big] - \delta,$$

for all $\delta \geq 0$.

It hence remains to show that $\mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big]$ is lower-bounded by $\frac{1}{2}\mathcal{R}_n(\mathcal{F}) - \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f]|}{2\sqrt{n}}$. Once again, we use a symmetrization argument.

$$\begin{aligned}
\mathbb{E}\big[\|\mathbb{P}_n - \mathbb{P}\|_{\mathcal{F}}\big] &= \mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\big[f(X)\big]\right|\right] \\
&= \frac{1}{2}\mathbb{E}_X\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\big[f(X)\big]\right|\right] + \frac{1}{2}\mathbb{E}_Y\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}\big[f(Y)\big]\right|\right] \\
&\overset{(i)}{\geq} \frac{1}{2}\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left\{\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - \mathbb{E}\big[f(X)\big]\right| + \left|\frac{1}{n}\sum_{i=1}^n f(Y_i) - \mathbb{E}\big[f(Y)\big]\right|\right\}\right] \\
&\overset{(ii)}{\geq} \frac{1}{2}\mathbb{E}_{X,Y}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(X_i) - f(Y_i)\right|\right] \\
&\overset{(iii)}{=} \frac{1}{2}\mathbb{E}_{X,Y,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f(X_i) - f(Y_i)\big)\right|\right] \\
&\overset{(iv)}{\geq} \frac{1}{2}\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f(X_i) - \mathbb{E}[f(Y)]\big)\right|\right],
\end{aligned}$$

where (i) is due to the sum of $\sup$ being greater or equal to the $\sup$ of the sum, (ii) is by the triangle inequality, (iii) comes from the observation that the variables $\varepsilon_i\big(f(X_i) - f(Y_i)\big)$ and $f(X_i) - f(Y_i)$ are identically distributed and (iv) is obtained by interchanging the expectation and the supremum. We further note that, combining the triangle inequality with the fact that a sum of $\sup$ is greater or equal to the $\sup$ of the sum, we get

$$\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\big(f(X_i) - \mathbb{E}[f(X)]\big)\right|\right] \geq \mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)\right|\right] - \mathbb{E}_{\varepsilon}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}[f(X)]\right|\right].$$

The first term $\mathbb{E}_{X,\varepsilon}\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(X_i)|\right]$ equals the Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ and the second term $\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}|\frac{1}{n}\sum_{i=1}^n \varepsilon_i \mathbb{E}[f(X)]|\right]$ can be upper-bounded as follows:

$$\begin{aligned}
\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i\mathbb{E}[f(X)]\right|\right] &= \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{n}\mathbb{E}_\varepsilon\left[\left|\sum_{i=1}^n\varepsilon_i\right|\right] \\
&= \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{n}\mathbb{E}_\varepsilon\left[\sqrt{\left(\sum_{i=1}^n\varepsilon_i-\mathbb{E}_\varepsilon\left[\sum_{i=1}^n\varepsilon_i\right]\right)^2}\right] \\
&\overset{(*)}{\leq} \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{n}\sqrt{Var_\varepsilon\left[\sum_{i=1}^n\varepsilon_i\right]} \\
&\overset{(**)}{=} \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{n}\sqrt{\sum_{i=1}^n Var_{\varepsilon_i}[\varepsilon_i]} = \frac{\sup_{f\in\mathcal{F}}|\mathbb{E}[f(X)]|}{\sqrt{n}},
\end{aligned}$$

where $(*)$ is Jensen's inequality and $(**)$ holds because the variance of the sum of independent random variables is the sum of the variances of these random variable. Combining all the results yields the claim.

□

Combining Theorem12.10 and Proposition 12.11, we can now conclude that, for a uniformly bounded function class $\mathcal{F}$, the Rademacher complexity provides a necessary and sufficient condition for the function class to be Glivenko-Cantelli. Specifically, the function class $\mathcal{F}$ is Glivenko-Cantelli if and only if its Rademacher complexity $\mathcal{R}_n(\mathcal{F})$ goes to zero as $n\to\infty$.

Obtaining concrete results using Theorem12.10 requires methods for upper bounding the Rademacher complexity. Such methods range from simple union bounds, suitable for finite function classes, over techniques involving polynomial discrimination and Vapnik-Chervonenkis (VC) dimension, to be pursued in the next chapter, to very advanced techniques involving the notion of metric entropy (studied earlier in class in a different context) and chaining arguments.

12.4 Function classes with polynomial discrimination

Recall that

$$\mathcal{F}(x_1^n) = \{(f(x_1), f(x_2), \ldots, f(x_n)) \mid f\in\mathcal{F}\},$$

i.e., the set $\mathcal{F}(x_1^n)$ contains all the vectors in $\mathbb{R}^n$ that can be realized by applying a function $f\in\mathcal{F}$ to the collection $(x_1,\ldots,x_n)$. The Rademacher complexity of the function class $\mathcal{F}$ given by

$$\begin{aligned}
\mathcal{R}_n(\mathcal{F}) &= \mathbb{E}_X[\mathcal{R}(\mathcal{F}(X_1^n)/n)] \\
&= \mathbb{E}_{\varepsilon,X}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n\varepsilon_i f(X_i)\right|\right] \qquad (12.8)
\end{aligned}$$

measures the size of $\mathcal{F}$ by quantifying the average of the maximum correlation between the vector $(f(x_1), \ldots, f(x_n))$ and the "noise vector" $(\varepsilon_1, \ldots, \varepsilon_n)$, with the maximum taken over all functions $f \in \mathcal{F}$. The expectation in 12.8 and the sup are often difficult to compute analytically. We now look at techniques for upper-bounding $\mathcal{R}_n(\mathcal{F})$ that are more accessible, yet preserve the growth behavior needed to infer the Glivenko-Cantelli property. It is natural that such techniques will apply only to certain function classes. In particular, we will be concerned with classes with polynomial discrimination.

For a given collection of points $x_1^n = (x_1, \ldots, x_n)$, the "size" of the set $\mathcal{F}(x_1^n)$ provides a sample-dependent measure of the complexity of $\mathcal{F}$. For example, if the set $\mathcal{F}(x_1^n)$ contains only a finite number of vectors for all sample sizes, its "size" can be measured via its cardinality. This occurs naturally for function classes of finite cardinality. Such $\mathcal{F}$ are however, in general, of limited interest. For general function classes $\mathcal{F}$, the next best thing we can, however, be interested in is the growth rate of the cardinality of $\mathcal{F}(x_1^n)$ as a function of n. For instance, if $\mathcal{F}$ consists of a family of functions taking binary values (e.g. decision rules), then $\mathcal{F}(x_1^n)$ can contain at most 2^n elements. In the following, we shall be a bit more conservative, and shall consider function classes for which the cardinality of $\mathcal{F}(x_1^n)$ grows only polynomially in n.

Definition 12.12 (Polynomial discrimination)**.** A class $\mathcal{F}$ of functions with domain $\mathcal{X}$ has polynomial discrimination of order $\nu \geq 1$ if, for each positive integer n and collection $x_1^n = (x_1, \ldots, x_n)$ of n points in $\mathcal{X}$, the set $\mathcal{F}(x_1^n)$ has cardinality upper-bounded according to

$$|\mathcal{F}(x_1^n)| \leq (n+1)^\nu.$$

The significance of this property is that it provides a straightforward approach to control the Rademacher complexity such that we can infer the Glivenko-Cantelli property.

Lemma 12.13. *Suppose that $\mathcal{F}$ has polynomial discrimination of order ν. Then, for all positive integers n and any collection of points $x_1^n = (x_1, \ldots, x_n)$, we have*

$$\underbrace{\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right]}_{\mathcal{R}(\mathcal{F}(x_1^n)/n)} \leq 4D(x_1^n)\sqrt{\frac{\nu\log(n+1)}{n}}, \tag{12.9}$$

where $D(x_1^n) = \sup_{f\in\mathcal{F}}\sqrt{\frac{\sum\limits_{i=1}^n f^2(x_i)}{n}}$ is the ℓ_2-radius of the set $\mathcal{F}(x_1^n)/\sqrt{n}$.

Proof. Rewriting the LHS in (12.9) as an inner product, we have

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right] = \frac{1}{n}\mathbb{E}_\varepsilon\left[\max_{\theta\in\mathcal{F}(x_1^n)}|\langle\varepsilon, \theta\rangle|\right],$$

where, by assumption, the set $\mathcal{F}(x_1^n)$ contains at most $(n+1)^\nu$ elements. To be able to apply Lemma 2.4, we need to first show that, for any $\theta \in \mathcal{F}(x_1^n)$, the random variable $\langle \varepsilon, \theta \rangle$ is sub-Gaussian with parameter $\sigma := \sup_{\theta \in \mathcal{F}(x_1^n)} \sqrt{\sum_{i=1}^n \theta_i^2}$:

$$\begin{aligned}
\mathbb{E}_\varepsilon \left[e^{\lambda \langle \varepsilon, \theta \rangle} \right] &= \prod_{i=1}^n \mathbb{E}_{\varepsilon_i} \left[e^{\lambda \varepsilon_i \theta_i} \right] \\
&= \prod_{i=1}^n \frac{e^{\lambda \theta_i} + e^{-\lambda \theta_i}}{2} \\
&\leq e^{\frac{\lambda^2}{2} \sum_{i=1}^n \theta_i^2} \\
&\leq e^{\frac{\lambda^2}{2} \left\{ \sup_{\theta \in \mathcal{F}(x_1^n)} \sum_{i=1}^n \theta_i^2 \right\}},
\end{aligned}$$

where we used the identity $e^x + e^{-x} \leq 2e^{\frac{x^2}{2}}$ for all $x \in \mathbb{R}$. Applying Lemma 2.4 provides the desired result

$$\mathcal{R}_n(\mathcal{F}(x_1^n)) \leq 2D_n \sqrt{\frac{\nu \log(n+1)}{n}},$$

with $D_n := \sup_{f \in \mathcal{F}} \sqrt{\frac{\sum_{i=1}^n f(x_i)^2}{n}}$. □

Even though this lemma provides us with an upper bound on the empirical Rademacher complexity only, taking the expectation w.r.t. X, yields an upper-bound on $\mathcal{R}_n(\mathcal{F})$ according to

$$\mathcal{R}_n(\mathcal{F}) = \mathbb{E}_X[\mathcal{R}(\mathcal{F}(X_1^n)/n)] \leq 4\mathbb{E}_{X_1^n}[D(X_1^n)] \sqrt{\frac{\nu \log(n+1)}{n}}.$$

An especially simple case is now obtained when the function class is b-uniformly bounded, so that $D(x_1^n) \leq b$ for all samples. In this case, we get

$$\mathcal{R}_n(\mathcal{F}) \leq 4b \sqrt{\frac{\nu \log(n+1)}{n}}, \quad \text{for all } n \geq 1.$$

Thanks to Theorem12.10, we can now conclude that every bounded function class with polynomial discrimination is Glivenko-Cantelli.

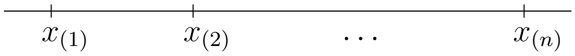
We next consider a specific example of a bounded function class with polynomial discrimination property, namely the left-sided indicator functions $\mathbb{1}_{(-\infty, t]}$ that lead to the classical Glivenko-Cantelli law. These functions are uniformly bounded with $b = 1$ and, as shown in the next result, satisfy the polynomial discrimination property with order $\nu = 1$.

Corollary 12.14 (Classical Glivenko-Cantelli)**.** *Let $F(t) = \mathbb{P}[X \leq t]$ be the CDF of a random variable $X \sim \mathbb{P}$, and let $\hat{F}_n$ be the empirical CDF based on n i.i.d. samples $X_i \sim \mathbb{P}$. Then,*

$$\mathbb{P}\left[\|\hat{F}_n - F\|_\infty \geq 8\sqrt{\frac{\log(n+1)}{n}} + \delta \right] \leq e^{-\frac{n\delta^2}{2}}, \quad \delta \geq 0$$

and hence $\|\hat{F}_n - F\|_\infty \overset{a.s.}{\to} 0$.

Proof. For a given sample $x_1^n = (x_1, \dots, x_n) \in \mathbb{R}^n$, consider the set $\mathcal{F}(x_1^n)$, where $\mathcal{F}$ is the set of all $\{0,1\}$-valued indicator functions of the half-intervals $(-\infty, t]$, for $t \in \mathbb{R}$. If we order the samples as $x_{(1)} \le x_{(2)} \le \cdots \le x_{(n)}$, then they split the real line into at most $n+1$ intervals (including the two end-intervals $(-\infty, x_{(1)}]$ and $[x_{(n)}, \infty)$). For a given t, the indicator function $\mathbb{1}_{(-\infty,t]}$ takes the value 1 for all $x_{(i)} \le t$, and the value 0 for all other samples. This shows that, for a given sample x_1^n, we have $|\mathcal{F}(x_1^n)| \le n+1$. This can be illustrated as follows. Consider $f \in \mathcal{F} = \left\{\mathbb{1}_{(-\infty,t]}\right\}_{t\in\mathbb{R}}$ and fix the sample x_1^n. The vector $f(x_1^n)$ obtained by applying f to the ordered version of x_1^n is now given by $\left(\mathbb{1}_{(-\infty,t]}(x_{(1)}), \mathbb{1}_{(-\infty,t]}(x_{(2)}), \dots, \mathbb{1}_{(-\infty,t]}(x_{(n)})\right)$. Based on the following picture



we can conclude that $f(x_1^n) = (0, 0, \dots, 0)$, for $t < x_{(1)}$, $f(x_1^n) = (1, 0, \dots, 0)$ for $t < x_{(2)}$, $f(x_1^n) = (1, 1, 0, \dots, 0)$ for $t < x_{(3)}$, and so on.

In summary, for a given sample x_1^n, we get no more than $n+1$ different vectors (there could be fewer vectors, namely when entries in x_1^n are identical). Hence, $\mathcal{F}$ satisfies the polynomial discrimination property with $\nu = 1$ and, by Lemma 12.13, we get,

$$\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right] \le 4\sqrt{\frac{\log(n+1)}{n}},$$

which, upon taking the expectation w.r.t. X, yields

$$\mathcal{R}_n(\mathcal{F}) \le 4\sqrt{\frac{\log(n+1)}{n}}.$$

The proof is finalized by applying Theorem 12.10. □

12.5 Vapnik-Chervonenkis dimension

We now turn to developing a theory for certifying the polynomial discrimination property, namely the theory of Vapnik-Chervonenkis (VC) dimension. We start by defining the notions of shattering and VC dimension. Let us consider a function class in which each function is binary-valued, taking the values $\{0,1\}$ for concreteness. In this case, the set

$$\mathcal{F}(x_1^n) = \{f(x_1), \dots, f(x_n) \mid f \in \mathcal{F}\}$$

can have at most 2^n elements.

Definition 12.15 (Shattering and VC dimension)**.** Given a class $\mathcal{F}$ of binary-valued functions, we say that the set $x_1^n = (x_1, \dots, x_n)$ is shattered by $\mathcal{F}$ if $|\mathcal{F}(x_1^n)| = 2^n$. The VC dimension $\nu(\mathcal{F})$ is the largest integer n for which there is some collection $x_1^n = (x_1, \dots, x_n)$ of n points that is shattered by $\mathcal{F}$.

When $\nu(\mathcal{F})$ is finite, the function class $\mathcal{F}$ is said to be a VC class.

Example 12.16 (Intervals in $\mathbb{R}$)**.** Consider the class of all indicator functions for left-sided half-intervals on the real line—namely, the class $\mathcal{S}_{\text{left}} := \{(-\infty, a] \mid a \in \mathbb{R}\}$. We note that for any single point, the set $\mathcal{F}(x_1)$ is given by $\mathcal{F} = \{0, 1\}$, while for $n = 2$, we get $\mathcal{F}(x_1^2) = \{(0, 0), (1, 0), (1, 1)\}$ so that $\nu(\mathcal{F}) = 1$. Now, consider the class of all two-sided intervals over the real line, namely the class $\mathcal{S}_{\text{two}} := \{(a, b] \mid a, b \in \mathbb{R}$ such that $a < b\}$. This class can shatter any two-point set. However, given three distinct points $x_1 < x_2 < x_3$, it cannot pick out the subset $\{x_1, x_3\}$, showing that $\nu(\mathcal{S}_{\text{two}}) = 2$. We now proceed to show that both the class of indicator functions on left-sided intervals and the class of two-sided intervals have polynomial discrimination property. For the former, we have already established in the proof of Corollary 12.14 that $|\mathcal{F}(x_1^n)| \leq n + 1$ and hence we have polynomial discrimination of order 1. For the class of indicator functions on two-sided intervals, we first note that any collection of n distinct points $x_1 < x_2 < \cdots < x_n$ divides up the real line into $(n + 1)$ intervals. Thus, any set of the form $(a, b]$ can be specified by choosing one of $(n + 1)$ intervals for b, and a second interval for a. Taken together, this shows that we can have at most $(n + 1)^2$ intervals. For shattering, we would need, however, 2^n intervals, each corresponding to a combination of the n points. We therefore have $|\mathcal{F}(x_1^n)| \leq (n + 1)^2$ and hence get polynomial discrimination of order 2.

Both example function classes we just considered are of finite VC dimension and also turn out to have polynomial discrimination. It is thus sensible to ask whether any finite VC class has polynomial discrimination. A deep result due to Vapnik and Chervonenkis, Sauer and Shelah states that, indeed, any finite VC class has polynomial discrimination with degree at most the VC dimension. To understand why this result is surprising, we note that, for a given set class $\mathcal{S}$, by definition of the VC dimension, we have that for all $n > \nu(\mathcal{S})$, $|\mathcal{S}(x_1^n)| < 2^n$ for all x_1^n. However, at least in principle, there could exist some subset with $|\mathcal{S}(x_1^n)| = 2^n - 1$, which is not significantly different from 2^n. The following result shows that this can not happen. Indeed, for any finite VC class, the cardinality of $\mathcal{S}(x_1^n)$ can grow no faster than polynomially in n.

Theorem 12.17 (Vapnik-Chervonenkis, Sauer-Shelah)**.** *Consider a set class $\mathcal{S}$ with $\nu(\mathcal{S}) < \infty$. Then, for any collection of points $P = (x_1, \dots, x_n)$ with $n \geq \nu(\mathcal{S})$, we have*

$$|\mathcal{S}(P)| \leq \sum_{i=0}^{\nu(\mathcal{S})} \binom{n}{i} \leq (n + 1)^{\nu(\mathcal{S})}.$$

Proof. Given a subset of points Q and a set class $\mathcal{T}$, we let $\nu(\mathcal{T}; Q)$ denote the VC dimension of $\mathcal{T}$ when considering only whether or not subsets of Q can be shattered. Note that $\nu(\mathcal{T}) \leq k$ implies that $\nu(\mathcal{T}; Q) \leq k$, for all point sets Q. For positive integers (n, k), define the functions

$$\Phi_k(n) := \sup_{\substack{\text{sets } Q \\ |Q| \leq n}} \; \sup_{\substack{\text{set classes} \mathcal{T} \\ \nu(\mathcal{T}; Q) \leq k}} |\mathcal{T}(Q)|$$

and

$$\Psi_k(n) := \sum_{i=0}^{k} \binom{n}{i}.$$

In the following, we shall use the convention $\binom{n}{i} = 0$ whenever $i > n$. We now claim that it suffices to prove that

$$\Phi_k(n) \leq \Psi_k(n). \tag{12.10}$$

Indeed, suppose there were some set class $\mathcal{S}$ with $\nu(\mathcal{S}) = k$ and collection $P = \{x_1, \ldots, x_n\}$ of n distinct points for which $|\mathcal{S}(P)| > \Psi_k(n)$. By definition of $\Phi_k(n)$, we would then have

$$\Phi_k(n) \overset{(i)}{\geq} \sup_{\substack{\text{set classes}\mathcal{T} \\ \nu(\mathcal{T};P)\leq k}} |\mathcal{T}(P)| \overset{(ii)}{\geq} |\mathcal{S}(P)| > \Psi_k(n),$$

which contradicts the claim (12.10). Here, inequality (i) follows because P is feasible for the supremum over Q that defines $\Phi_k(n)$, simply because $|P| = n$; and inequality (ii) follows because $\nu(\mathcal{S}) = k$ implies that $\nu(\mathcal{S}; P) \leq k$. We now prove the claim (12.10) by induction on the sum $n + k$ of the pair (n, k).

Base case. To start, we prove that (12.10) holds for all pairs (n, k) with $n + k = 2$. For $n = 0$, we have $\Phi_k(0) = 0 \leq \sum_{i=0}^{k} \binom{0}{i} = \binom{0}{0} + \binom{0}{1} = 1$, regardless of whether $k = 0$ or $k = 1$. For $k = 0$ and $n = 1$, we get $\Phi_k(n) = 1 \leq \sum_{i=0}^{0} \binom{1}{i} = 1$. Finally, for $k = n = 1$, we have $\Phi_k(n) = 2$ and $\Psi_1(1) = \sum_{i=0}^{1} \binom{1}{i} = \binom{1}{0} + \binom{1}{1} = 2$, and hence $\Phi_k(n) \leq \Psi_k(n)$.

We proceed to the induction step. Assume that for some integer $\ell > 2$, the inequality (12.10) holds for all pairs with $n + k < \ell$. We claim that it then holds for all pairs with $n + k = \ell$. Fix an arbitrary pair (n, k) such that $n + k = \ell$, a point set $P = \{x_1, \ldots, x_n\}$ and a set class $\mathcal{S}$ such that $\nu(\mathcal{S}; P) = k$. Define the point set $P' = P \setminus \{x_1\}$, and let $\mathcal{S}_0 \subseteq \mathcal{S}$ be the smallest collection of subsets that labels the point set P' in the maximal number of different ways. Let $\mathcal{S}_1$ be the smallest collection of subsets inside $\mathcal{S} \setminus \mathcal{S}_0$ that produce binary labelings of the point set P that are not in $\mathcal{S}_0(P)$. (The choices of $\mathcal{S}_0$ and $\mathcal{S}_1$ need not be unique.)

As a concrete example, given a set class $\mathcal{S} = \{s_1, s_2, s_3, s_4\}$ and a point set $P = \{x_1, x_2, x_3\}$, which we assume generate the binary labelings

$$s_1 \leftrightarrow (0, 1, 1), \quad s_2 \leftrightarrow (1, 1, 1), \quad s_3 \leftrightarrow (0, 1, 0), \quad s_4 \leftrightarrow (0, 1, 1).$$

In this particular case, we have $\mathcal{S}(P) = \{(0, 1, 1), (1, 1, 1), (0, 1, 0)\}$. One valid choice of the pair $(\mathcal{S}_0, \mathcal{S}_1)$ would be $\mathcal{S}_0 = \{s_1, s_3\}$ and $\mathcal{S}_1 = \{s_2\}$, generating the labelings $\mathcal{S}_0(P) = \{(0, 1, 1), (0, 1, 0)\}$ and $\mathcal{S}_1(P) = \{(1, 1, 1)\}$.

Using this decomposition, we claim that

$$|\mathcal{S}(P)| = |\mathcal{S}_0(P')| + |\mathcal{S}_1(P')|.$$

Indeed, any binary labeling in $\mathcal{S}(P)$ is either mapped to a member of $\mathcal{S}_0(P')$, or in the case that its labeling on P' corresponds to a duplicate, it can be uniquely identified with a member of $\mathcal{S}_1(P')$. This can be verified explicitly in the special case above.

Now since P' is a subset of P and $\mathcal{S}_0$ is a subset of $\mathcal{S}$, we have

$$\nu(\mathcal{S}_0; P') \leq \nu(\mathcal{S}_0; P) \leq k.$$

Since the cardinality of P' is equal to $n-1$, the induction hypothesis implies that $|\mathcal{S}_0(P')| \leq \Psi_k(n-1)$. On the other hand, we claim that the set $\mathcal{S}_1$ satisfies the upper bound $\nu(\mathcal{S}_1; P') \leq k-1$. Suppose that $\mathcal{S}_1$ shatters some subset $Q' \subseteq P'$ of cardinality m; it suffices to show that $m \leq k-1$. If $\mathcal{S}_1$ shatters such a set Q', then $\mathcal{S}$ would shatter the set $Q = Q' \cup \{x_1\} \subseteq P$. This fact follows by construction of $\mathcal{S}_1$, for every binary vector in the set $\mathcal{S}_1(P)$, the set $\mathcal{S}(P)$ must contain a binary vector with the label for x_1 flipped, this follows as $\mathcal{S}_0 \subseteq \mathcal{S}$ is the smallest collection of subsets in $\mathcal{S} \setminus \mathcal{S}_0$ that label P' in the maximal number of different ways and hence the binary vector with the flipped label for x_1 is not included in $\mathcal{S}_0$ as otherwise $\mathcal{S}_0$ would not be the smallest subset. Now, as subsets with both labels, 0 and 1, for x_1 are included in $\mathcal{S}$, it follows that $\mathcal{S}$ shatters $Q = Q' \cup \{x_1\}$. Next, since $\nu(\mathcal{S}; P) \leq k$, we must have $|Q| = m + 1 \leq k$, which implies that $\nu(\mathcal{S}_1; P) \leq k-1$. Consequently, the induction hypothesis implies that $|\mathcal{S}_1(P')| \leq \Psi_{k-1}(n-1)$.

Putting the pieces together, we have shown that

$$|\mathcal{S}(P)| \leq \Psi_k(n-1) + \Psi_{k-1}(n-1) = \Psi_k(n),$$

where the equality is obtained by employing $\binom{a}{b} + \binom{a}{b+1} = \binom{a+1}{b+1}$ in

$$\begin{aligned}
\Psi_k(n-1) + \Psi_{k-1}(n-1) &= \sum_{i=0}^{k} \binom{n-1}{i} + \sum_{i=0}^{k-1} \binom{n-1}{i} \\
&= \binom{n-1}{0} + \sum_{i=1}^{k} \binom{n-1}{i} + \sum_{i=0}^{k-1} \binom{n-1}{i} \\
&= \binom{n-1}{0} + \sum_{i=0}^{k-1} \left[\binom{n-1}{i+1} + \binom{n-1}{i} \right] \\
&= \binom{n-1}{0} + \sum_{i=0}^{k-1} \binom{n}{i+1} \\
&= \binom{n-1}{0} + \sum_{i=1}^{k} \binom{n}{i} \\
&= \sum_{i=0}^{k} \binom{n}{i} = \Psi_k(n).
\end{aligned}$$

It remains to establish

$$\sum_{i=0}^{\nu(\mathcal{S})} \binom{n}{i} \leq (n+1)^{\nu(\mathcal{S})},$$

which follows by expressing $(n+1)^{\nu(\mathcal{S})}$ as a binomial series according to

$$\sum_{i=0}^{\nu(\mathcal{S})} \binom{n}{i} \overset{(i)}{\leq} \sum_{i=0}^{\nu(\mathcal{S})} n^i = \sum_{i=0}^{\nu(\mathcal{S})} n^i 1^{\nu(\mathcal{S})-i} \leq \sum_{i=0}^{\nu(\mathcal{S})} \binom{\nu(\mathcal{S})}{i} n^i 1^{\nu(\mathcal{S})-i} \overset{(ii)}{=} (1+n)^{\nu(\mathcal{S})},$$

where we used the following upper bound

$$\binom{n}{i} \leq \frac{n!}{i!(n-i)!} \leq \frac{n(n-1)\dots(n-i+1)\cancel{(n-i)!}}{i!\cancel{(n-i)!}} \leq \frac{n^i}{i!} \leq n^i$$

in (i) and the binomial formula

$$(x+y)^n = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} y^k$$

in (ii). □

12.6 VC dimension and vector spaces

Any class $\mathcal{G}$ of real-valued functions defines a class of sets by the operation of taking subgraphs. In particular, given a real-valued function $g\colon \mathcal{X} \to \mathbb{R}$, its subgraph at level zero $S_g \coloneqq \{x \in \mathcal{X} \mid g(x) \leq 0\}$. In this way, we can associate to $\mathcal{G}$ the collection of subsets $\mathcal{S}(\mathcal{G}) \coloneqq \{S_g,\, g \in \mathcal{G}\}$, which we refer to as the subgraph class of $\mathcal{G}$. Many interesting classes of sets such as half-spaces and ellipsoids are naturally defined in this way. In many cases, the underlying function class $\mathcal{G}$ is a vector space and the following result relates its dimension to the VC dimension of the associated set class $\mathcal{S}(\mathcal{G})$.

Proposition 12.18. Let $\mathcal{G}$ be a vector space of functions $g\colon \mathbb{R}^d \to \mathbb{R}$ with dimension $dim(\mathcal{G}) < \infty$. Then, the subgraph class $\mathcal{S}(\mathcal{G})$ has VC dimension at most $dim(\mathcal{G})$.

Proof. By the definition of VC dimension, we need to show that no collection of $n = dim(\mathcal{G}) + 1$ points in $\mathbb{R}^d$ can be shattered by $\mathcal{S}(\mathcal{G})$. Fix an arbitrary collection $x_1^n = \{x_1, \dots, x_n\}$ of n points in $\mathbb{R}^d$, and consider the linear map $L\colon \mathcal{G} \to \mathbb{R}^n$ given by $L(g) = (g(x_1), \dots, g(x_n))$. By construction, the range of the mapping L is a linear subspace of $\mathbb{R}^n$ with dimension at most $dim(\mathcal{G}) = n - 1 < n$. Therefore, there must exist a non-zero vector $\gamma \in \mathbb{R}^n$ such that $\langle \gamma, L(g) \rangle = 0$, $\forall g \in \mathcal{G}$. We assume w.l.o.g. that at least one coordinate is positive, and then write

$$\langle \gamma, L(g) \rangle = \sum_{i=1}^{n} \gamma_i g(x_i) = \sum_{\{i \mid \gamma_i \leq 0\}} \gamma_i g(x_i) + \sum_{\{i \mid \gamma_i > 0\}} \gamma_i g(x_i) = 0$$

and hence

$$\sum_{\{i \mid \gamma_i \leq 0\}} (-\gamma_i) g(x_i) = \sum_{\{i \mid \gamma_i > 0\}} \gamma_i g(x_i), \quad \forall g \in \mathcal{G}. \tag{12.11}$$

Proceeding via proof by contradiction, suppose that the set $\{x_1, \dots, x_n\}$ is shattered by $\mathcal{S}(\mathcal{G})$. Then, for every combination of points in $\{x_1, \dots, x_n\}$, there must exist a set in $\mathcal{S}(\mathcal{G})$ that contains only this combination of points. By construction of $\mathcal{S}(\mathcal{G})$, there must consequently be a $g \in \mathcal{G}$ that generates this set, i.e., there must exist a corresponding $S_g = \{x \in \mathbb{R}^d \mid g(x) \leq 0\}$. Now, we take the pertinent subset to be given by $\{x_i \mid \gamma_i \leq 0\}$. This would, however, imply that the RHS of (12.11) would be strictly positive while the LHS would be non-positive, which constitutes a contradiction. We therefore conclude that $\mathcal{S}(\mathcal{G})$ fails to shatter the set $\{x_1, \ldots, x_n\}$, as claimed. □

We next illustrate Proposition 12.18 with two examples.

Example 12.19 (Linear functions in $\mathbb{R}^d$)**.** For a pair $(a, b) \in \mathbb{R}^d \times \mathbb{R}$, define the function $f_{a,b} := \langle a, x\rangle + b$, and consider the family $\mathcal{L}^d := \{f_{a,b} \mid (a, b) \in \mathbb{R}^d \times \mathbb{R}\}$ of all such functions. The associated subgraph class $\mathcal{S}(\mathcal{L}^d)$ corresponds to the collection of all half-spaces of the form $H_{a,b} := \{x \in \mathbb{R}^d \mid \langle a, x\rangle + b \leq 0\}$. Since the family $\mathcal{L}^d$ forms a vector space of dimension $d + 1$, we obtain as an immediate consequence of Proposition 12.18 that $\mathcal{S}(\mathcal{L}^d)$ has VC dimension at most $d + 1$.

For the special case $d = 1$, let us verify this statement by a more direct calculation. In this case, the class $\mathcal{S}(\mathcal{L}^1)$ corresponds to the collection of all left-sided and right-sided intervarls, i.e.,

$$\mathcal{S}(\mathcal{L}^1) = \{(-\infty, t] \mid t \in \mathbb{R}\} \cup \{[t, \infty) \mid t \in \mathbb{R}\}.$$

Given any two distinct points $x_1 < x_2$, the collection of all such intervals can pick out all possible subsets. However, given any three points $x_1 < x_2 < x_3$, there is no interval contained in $\mathcal{S}(\mathcal{L}^1)$ that contains x_2 while excluding both x_1 and x_3. This shows that $\nu(\mathcal{S}(\mathcal{L}^1)) = 2$ and hence our upper bound is tight in the case $d = 1$. More generally, it can be shown that the VC dimension of $\mathcal{S}(\mathcal{L}^d)$ is $d + 1$ so that Proposition 12.18 yields a tight upper bound in all dimensions.

Example 12.20 (Spheres in $\mathbb{R}^d$)**.** Consider the sphere $S_{a,b} := \{x \in \mathbb{R}^d \mid \|x - a\|_2 \leq b\}$, where $(a, b) \in \mathbb{R}^d \times \mathbb{R}_+$, specify its center and radius, respectively, and let $\mathcal{S}^d_{\text{sphere}}$ denote the collection of all such spheres. If we define the function

$$f_{a,b}(x) = \|x\|_2^2 - 2\sum_{j=1}^{d} a_j x_j + \|a\|_2^2 - b^2,$$

then we have $S_{a,b} = \{x \in \mathbb{R}^d \mid f_{a,b} \leq 0\}$, so that the sphere $S_{a,b}$ is a subgraph of the function $f_{a,b}$. In order to leverage Proposition 12.18, we first define a feature map $\phi\colon \mathbb{R}^d \to \mathbb{R}^{d+2}$ via

$$\phi(x) := (1, x_1, \ldots, x_d, \|x\|_2^2),$$

and then consider functions of the form

$$g_c(x) := \langle c, \phi(x)\rangle, \quad c \in \mathbb{R}^{d+2}.$$

The family of functions $\{g_c,\, c \in \mathbb{R}^{d+2}\}$ is a vector space of dimension $d + 2$, and it contains the function class $\{f_{a,b},\, (a, b) \in \mathbb{R}^d \times \mathbb{R}_+\}$. Consequently, by applying Proposition 12.18 to this larger vector space, we conclude that $\nu(\mathcal{S}^d_{\text{sphere}}) \leq d + 2$. A more refined analysis shows that the VC dimension of spheres in $\mathbb{R}^d$ is actually $d + 1$.

Appendix A

Martingales

Definition 1.1 (Probability space)**.** A probability space $(\Omega, \mathcal{G}, \mathbb{P})$ is a triple consisting of a sample space Ω, a σ-algebra $\mathcal{G}$ containing all the events and a probability measure $\mathbb{P}\colon \mathcal{G} \to [0,1]$.

A probability space is a special case of a measured space, for which the total mass of the measure is 1.

Definition 1.2 (Conditional expectation)**.** Let $(\Omega, \mathcal{G}, \mathbb{P})$ be a probability space. Let $\mathcal{G}' \subset \mathcal{G}$ be a sub-σ-algebra of $\mathcal{G}$ and X an integrable random variable. Then there exists a random variable Z, $\mathcal{G}'$ measurable and integrable, such that, for all bounded random variable U that is $\mathcal{G}'$-measurable

$$\mathbb{E}[XU] = \mathbb{E}[ZU].$$

Such a random variable Z is called the conditional expectation of X with respect to $\mathcal{G}'$ and is written

$$Z = \mathbb{E}[X \mid \mathcal{G}'].$$

When $\mathcal{G}' = \sigma(X_1, \dots, X_n)$ is the σ-algebra generated by a sequence of random variables $\{X_k\}_{k=1}^n$ and X a random variable, then we write $\mathbb{E}[X \mid X_1, \dots, X_n]$ for $\mathbb{E}[X \mid \mathcal{G}']$.

Lemma 1.3. *The conditional expectation satisfies the following properties:*

- *Linearity:*
$$\mathbb{E}[aX + bY \mid \mathcal{G}'] = a\mathbb{E}[X \mid \mathcal{G}'] + b\mathbb{E}[Y \mid \mathcal{G}']. \tag{A.1}$$

- *Pulling out known factors:*
$$\mathbb{E}[XY \mid \mathcal{G}'] = X\mathbb{E}[Y \mid \mathcal{G}'], \tag{A.2}$$
if X is $\mathcal{G}'$-measurable.

- *Tower property:*
$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{G}_2] \mid \mathcal{G}_1] = \mathbb{E}[X \mid \mathcal{G}_1], \quad \textit{for } \mathcal{G}_1 \subset \mathcal{G}_2. \tag{A.3}$$

- *Law of total expectation:*
$$\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{G}']] = \mathbb{E}[X]. \tag{A.4}$$

Definition 1.4 (Filtration)**.** A filtration $\{\mathcal{G}_k\}_{k=1}^{\infty}$ is a sequence of σ-algebras satisfying $\mathcal{G}_k \subseteq \mathcal{G}_{k+1}$, for all $k \geq 1$.

A sequence of random variables $\{X_k\}_{k=1}^{\infty}$ is said to be *adapted* to the filtration $\{\mathcal{G}_k\}_{k=1}^{\infty}$ if X_k is measurable with respect to $\mathcal{G}_k$, for all $k \geq 1$.

Definition 1.5 (Martingale)**.** Given a sequence of random variables $\{X_k\}_{k=1}^{\infty}$ adapted to the filtration $\{\mathcal{G}_k\}_{k=1}^{\infty}$, the sequence of pairs $\{(X_k, \mathcal{G}_k)\}_{k=1}^{\infty}$ is a martingale if, for all $k \geq 1$,

$$\mathbb{E}[|X_k|] < \infty \quad \text{and} \quad \mathbb{E}[X_{k+1} \,|\, \mathcal{G}_k] = X_k.$$

When the sequence is a martingale with respect to itself (i.e., with $\mathcal{G}_k = \sigma(X_1, \dots, X_k)$), then we simply write that $\{X_k\}_{k=1}^{\infty}$ is a martingale.

Example 1.6 (Doob Martingales)**.** Consider a sequence of real-valued independent random variables $\{X_k\}_{k=1}^{n}$ as well as the random variable $f(X) := f(X_1, \dots, X_n)$, for some function $f\colon \mathbb{R}^n \to \mathbb{R}$ such that $\mathbb{E}[|f(X)|] < \infty$. We construct a sequence of random variables $\{Y_k\}_{k=0}^{n}$ defined as

$$Y_k = \begin{cases} \mathbb{E}[f(X)] & \text{for } k = 0, \\ f(X) & \text{for } k = n, \\ \mathbb{E}[f(X) \,|\, X_1, \dots, X_k] & \text{for } k = 1 \dots, n-1. \end{cases}$$

Such a sequence of random variables is known as a *Doob martingale*. Considering the filtration

$$\mathcal{G}_k = \begin{cases} \{\emptyset, \Omega\} & \text{for } k = 0, \\ \sigma(X_1, \dots, X_k) & \text{for } k = 1, \dots, n, \end{cases}$$

we verify that $\{(Y_k, \mathcal{G}_k)\}_{k=0}^{n}$ actually defines a martingale sequence:

$$\mathbb{E}[|Y_k|] = \mathbb{E}[|\mathbb{E}[f(X) \,|\, \mathcal{G}_k]|] \leq \mathbb{E}[\mathbb{E}[|f(X)| \,|\, \mathcal{G}_k]] = \mathbb{E}[|f(X)|] < \infty$$

and
$$\mathbb{E}[Y_{k+1} \,|\, \mathcal{G}_k] = \mathbb{E}[\mathbb{E}[f(X) \,|\, \mathcal{G}_{k+1}] \,|\, \mathcal{G}_k] = \mathbb{E}[f(X) \,|\, \mathcal{G}_k] = Y_k$$
by the tower property (A.3).

A closely related notion is that of *martingale difference sequence*.

Definition 1.7 (Martingale difference sequence)**.** An adapted sequence $\{(D_k, \mathcal{G}_k)\}_{k=1}^{\infty}$ is a martingale difference sequence if, for all $k \geq 1$,

$$\mathbb{E}[|D_k|] < \infty \quad \text{and} \quad \mathbb{E}[D_{k+1} \,|\, \mathcal{G}_k] = 0.$$

As suggested by their name, such difference sequences arise in a natural way from martingales. In particular, given a martingale $\{(X_k, \mathcal{G}_k)\}_{k=0}^{\infty}$, let us define $D_k := X_k - X_{k-1}$, for $k \geq 1$. We then have

$$\begin{aligned} \mathbb{E}[D_{k+1} \,|\, \mathcal{G}_k] &= \mathbb{E}[X_{k+1} \,|\, \mathcal{G}_k] - \mathbb{E}[X_k \,|\, \mathcal{G}_k] \\ &= X_k - X_k = 0, \end{aligned}$$

where we have used the definition of a martingale (Definition 1.5), the linearity of conditional expectation (A.1) and the fact that X_k is measurable with respect to $\mathcal{G}_k$.

Appendix B

Concentration of Measure

The term *concentration inequality* refers to a variety of inequalities upper-bounding the probability that a random variable deviates from its mean by a given amount. The most classical concentration bounds are the laws of large numbers, which ensure that, with high probability, the average of independent random variables is close to their expectation. This section provides a brief introduction to the theory of concentration of measure; we refer the interested reader to [98] for a deeper analysis.

The most elementary tail bound is *Markov's inequality*. Let X be a non-negative random variable with finite mean. Markov's inequality states that

$$\mathbb{P}[X \geq \delta] \leq \frac{\mathbb{E}[X]}{\delta}, \quad \forall \delta > 0. \tag{B.1}$$

For a random variable X of finite variance $Var(X)$, applying Markov's inequality to the random variable $(X - \mathbb{E}[X])^2$ gives *Chebyshev's inequality*:

$$\mathbb{P}\big[\big|X - \mathbb{E}[X]\big| \geq \delta\big] \leq \frac{Var(X)}{\delta^2}, \quad \forall \delta > 0.$$

This is a simple form of concentration inequality, guaranteeing that X is close to its mean with high probability when the variance is small.

A wide variety of concentration bounds can be obtained from Markov's inequality (B.1). In what follows, we are particularly interested in bounding random variables X for which the moment generating function $\mathbb{E}\big[e^{\lambda X}\big]$ is defined for all $\lambda \in \mathbb{R}$. Markov's inequality then gives

$$\mathbb{P}[X - \mu \geq \delta] = \mathbb{P}\big[e^{\lambda(X-\mu)} \geq e^{\lambda\delta}\big] \leq \frac{\mathbb{E}\big[e^{\lambda(X-\mu)}\big]}{e^{\lambda\delta}}, \quad \forall \delta > 0. \tag{B.2}$$

Minimizing the RHS of (B.2) over λ so as to obtain the tightest possible result yields the *Chernoff bound*:

$$\log \mathbb{P}[X - \mu \geq \delta] \leq \inf_{\lambda \in \mathbb{R}} \Big\{ \log \mathbb{E}\big[e^{\lambda(X-\mu)}\big] - \lambda\delta \Big\}. \tag{B.3}$$

Example 2.1 (Sub-Gaussian random variables)**.** A random variable X with finite expectation μ is said to be *sub-Gaussian* if there is a positive number σ such that

$$\mathbb{E}\big[e^{\lambda(X-\mu)}\big] \leq e^{\frac{\sigma^2\lambda^2}{2}}, \quad \forall \lambda \in \mathbb{R}.$$

In particular, any Gaussian variable with variance σ^2 is sub-Gaussian with parameter σ. Combining the Chernoff bound (B.3) with the definition of sub-Gaussian random variables yields the following upper-deviation inequality

$$\log \mathbb{P}[X \geq \mu + \delta] \leq \inf_{\lambda \in \mathbb{R}} \left\{ \frac{\sigma^2\lambda^2}{2} - \lambda\delta \right\} = -\frac{\delta^2}{2\sigma^2}, \tag{B.4}$$

where the optimum of the quadratic function in λ has been found by differentiation. A similar argument applied to the sub-Gaussian random variable $-X$ yields the following lower-deviation inequality

$$\log \mathbb{P}[X \leq \mu - \delta] \leq -\frac{\delta^2}{2\sigma^2}. \tag{B.5}$$

Combining B.4 and B.5 through a union bound argument, we conclude that any sub-Gaussian variable satisfies the concentration inequality

$$\mathbb{P}[|X - \mu| \geq \delta] \leq 2e^{-\frac{\delta^2}{2\sigma^2}}. \tag{B.6}$$

We finalize this example by noting that sub-Gaussian random variables are very useful in practice as they extend concentration results of Gaussian variables to a broader class of random variables. For instance, it can be shown that every random variable $X_{a,b}$ taking value in $[a, b]$ almost surely, with $-\infty < a < b < \infty$, is sub-Gaussian with parameter upper-bounded by $\frac{b-a}{2}$. In particular, we obtain from (B.6) that every bounded random variable $X_{a,b}$ satisfies the following concentration inequality:

$$\mathbb{P}[|X_{a,b} - \mu| \geq \delta] \leq 2e^{-\frac{2\delta^2}{(b-a)^2}}. \tag{B.7}$$

We now turn our attention to concentration inequalities for functions of independent random variables. Let $\{X_i\}_{i=1}^n$ be a sequence of independent random variables, and consider the random variable $f(X) = f(X_1, \ldots, X_n)$ for some function $f \colon \mathbb{R}^n \to \mathbb{R}$. We are interested in obtaining bounds on the deviation of f from its mean. To this end, we consider the sequence of random variables $\{Y_k\}_{k=0}^n$ given by $Y_0 = \mathbb{E}[f(X)]$, $Y_n = f(X)$ and

$$Y_k = \mathbb{E}[f(X)|X_1, \ldots, X_k], \quad \forall k = 1, \ldots, n-1,$$

where we assumed that the conditional expectations exist. Note that Y_0 is deterministic and the random variables Y_k will tend to exhibit more “fluctuations” as we move along the sequence from Y_0 to Y_n. Based on this intuition, the tail bounds are based on the following telescoping decomposition:

$$f(X) - \mathbb{E}[f(X)] = \sum_{k=1}^{n} \underbrace{(Y_k - Y_{k-1})}_{=:D_k},$$

in which the deviation $f(X) - \mathbb{E}[f(X)]$ is written as a sum of increments $D_k := Y_k - Y_{k-1}$. The sequence $\{Y_k\}_{k=0}^n$ is an example of a martingale sequence, known as a *Doob martingale*, whereas the sequence $\{D_k\}_{k=1}^n$ is an example of a martingale difference sequence. Appendix A recalls basic definitions and properties about martingales.

The following lemma provides a concentration inequality for bounded martingale difference sequences.

Lemma 2.2 (Azuma-Hoeffding inequality). *Let* $\{(D_k, \mathcal{F}_k)\}_{k=1}^n$ *be a martingale difference sequence for which there exist constants* $\{(a_k, b_k)\}_{k=1}^n$ *such that* $D_k \in [a_k, b_k]$ *almost surely, for all* $k = 1, \ldots, n$*. Then*

$$\mathbb{P}\left[\left|\sum_{k=1}^n D_k\right| \geq \delta\right] \leq 2\exp\left\{-\frac{2\delta^2}{\sum_{k=1}^n (b_k - a_k)^2}\right\} \quad \forall \delta \geq 0.$$

Proof. Since $D_k \in [a_k, b_k]$ almost surely, the conditional random variable $(D_k | \mathcal{F}_{k-1})$ also belongs to this interval almost surely, and hence is sub-Gaussian with parameter $\sigma_k = \frac{b_k - a_k}{2}$ (cf. Example 2.1). We therefore have

$$\begin{aligned}\mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}\right] &\overset{(*)}{=} \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k} \mathbb{E}\left[e^{\lambda D_n} \mid \mathcal{F}_{n-1}\right]\right] \\ &\leq \mathbb{E}\left[e^{\lambda \sum_{k=1}^{n-1} D_k}\right] e^{\lambda^2 \frac{(b_n - a_n)^2}{8}},\end{aligned}$$

where $(*)$ comes from the definition of conditional expectation (cf. Definition 1.2 in Appendix A). Iterating this argument, we get

$$\mathbb{E}\left[e^{\lambda \sum_{k=1}^n D_k}\right] \leq e^{\lambda^2 \sum_{k=1}^n \frac{(b_k - a_k)^2}{8}},$$

which means that the random variable $\sum_{k=1}^n D_k$ is sub-Gaussian with parameter

$$\sigma := \sqrt{\sum_{k=1}^n \frac{(b_k - a_k)^2}{4}}.$$

We conclude using the same technique as to obtain (B.7):

$$\mathbb{P}\left[\left|\sum_{k=1}^n D_k\right| \geq \delta\right] \leq 2\exp\left\{-\frac{\delta^2}{2\sigma^2}\right\} = 2\exp\left\{-\frac{2\delta^2}{\sum_{k=1}^n (b_k - a_k)^2}\right\}, \quad \forall \delta \geq 0.$$

□

An important application of Lemma 2.2 concerns functions that satisfy the bounded difference property. Let us first introduce some notation. Given vectors $x, x' \in \mathbb{R}^n$ and an index $k \in \{1, \ldots, n\}$, we define a new vector $x^{\backslash k} \in \mathbb{R}^n$ via

$$x_j^{\backslash k} := \begin{cases} x_j, & \text{if } j \neq k, \\ x'_k, & \text{if } j = k. \end{cases} \tag{B.8}$$

With this notation in place, we say that $f\colon \mathbb{R}^n \to \mathbb{R}$ satisfies the *bounded difference property* with parameter L if, for each index $k \in \{1, \ldots, n\}$,

$$|f(x) - f(x^{\setminus k})| \leq L, \quad x, x' \in \mathbb{R}^n.$$

The following lemma gives a concentration inequality for functions satisfying the bounded difference property.

Lemma 2.3 (Bounded differences inequality)**.** *Suppose that $f\colon \mathbb{R}^n \to \mathbb{R}$ satisfies the bounded difference property with parameter L and that the random vector $X = (X_1, \ldots, X_n)$ has independent components. Then,*

$$\mathbb{P}\big[\big|f(X) - \mathbb{E}[f(X)]\big| \geq \delta\big] \leq 2e^{-\frac{2\delta^2}{nL^2}}, \quad \forall \delta \geq 0.$$

Proof. The proof of this lemma uses the Azuma-Hoeffding inequality. In order to use this result, we define the following martingale difference sequence

$$D_k := \mathbb{E}\big[f(X)|X_1, \ldots, X_k\big] - \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}\big].$$

We claim that D_k lies in an interval of length at most L almost surely. In order to prove this claim, we introduce the random variables

$$A_k := \inf_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big] - \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}\big]$$

and

$$B_k := \sup_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big] - \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}\big].$$

On one hand, we have

$$D_k - A_k = \mathbb{E}\big[f(X)|X_1, \ldots, X_k\big] - \inf_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big],$$

so that $D_k \geq A_k$ almost surely. A similar argument shows that $D_k \leq B_k$ almost surely.

We now need to show that $B_k - A_k \leq L$ almost surely. Observe that by the independence of the $\{X_k\}_{k=1}^n$, we have for any $(x_1, \ldots, x_k)$

$$\mathbb{E}\big[f(X)|x_1, \ldots, x_k\big] = \mathbb{E}_{k+1}\big[f(x_1, \ldots, x_k, X_{k+1}^n)\big],$$

where $\mathbb{E}_{k+1}$ denotes the expectation over $X_{k+1}^n := (X_{k+1}, \ldots, X_n)$. Consequently, we have

$$\begin{aligned}
B_k - A_k &= \sup_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big] - \inf_x \mathbb{E}\big[f(X)|X_1, \ldots, X_{k-1}, x\big] \\
&= \sup_x \mathbb{E}_{k+1}\big[f(X_1, \ldots, X_{k-1}, x, X_{k+1}^n)\big] - \inf_y \mathbb{E}_{k+1}\big[f(X_1, \ldots, X_{k-1}, y, X_{k+1}^n)\big] \\
&\leq \sup_{x,y} \big\{\mathbb{E}_{k+1}\big[f(X_1, \ldots, X_{k-1}, x, X_{k+1}^n) - f(X_1, \ldots, X_{k-1}, y, X_{k+1}^n)\big]\big\} \\
&\leq L,
\end{aligned}$$

using the bounded differences assumption. Thus the variable D_k lies within an interval of length at most L almost surely.

The proof is concluded by applying the Azuma-Hoeffding inequality (Lemma 2.2).

□

Under the same assumptions as in Lemma 2.3, a slight adaptation of the proof yields a one sided version of the result:

$$\mathbb{P}\big[f(X) - \mathbb{E}[f(X)] \geq \delta\big] \leq e^{-\frac{2\delta^2}{nL^2}}, \quad \forall \delta \geq 0,$$

or

$$\mathbb{P}\big[f(X) - \mathbb{E}[f(X)] \leq -\delta\big] \leq e^{-\frac{2\delta^2}{nL^2}}, \quad \forall \delta \geq 0.$$

We conclude this appendix by a concentration result on the maximum of sub-Gaussian random variables.

Lemma 2.4. *For $n \geq 2$, let $\{X_i\}_{i=1}^n$ be a set of zero-mean random variables, each sub-Gaussian with parameter σ. Then*

$$\mathbb{E}\left[\max_{i=1,\ldots,n} |X_i|\right] \leq 2\sigma\sqrt{\log n}.$$

Note that the random variables X_i are not assumed to be independent.

Proof. By Jensen's inequality inequality and using the sub-Gaussian assumption on the X_i, we obtain, for all $\lambda > 0$,

$$\begin{aligned}\mathbb{E}\left[\max_{i=1,\ldots,n} X_i\right] &\leq \frac{1}{\lambda}\log\mathbb{E}\left[\exp\left\{\lambda\max_{i=1,\ldots,n} X_i\right\}\right]\\ &\leq \frac{1}{\lambda}\log\sum_{i=1}^n \mathbb{E}\left[\exp\left\{\lambda X_i\right\}\right]\\ &\leq \frac{\log n}{\lambda} + \frac{\lambda\sigma^2}{2}.\end{aligned}$$

Optimizing over the choice of λ, we obtain

$$\mathbb{E}\left[\max_{i=1,\ldots,n} X_i\right] \leq \sigma\sqrt{2\log n}.$$

We apply this result to the set $\{Y_j\}_{j=1}^{2n}$ of zero-mean random variables, each sub-Gaussian with parameter σ defined in the following way:

$$Y_j := \begin{cases} X_i & \text{if } j = 2i,\\ -X_i & \text{if } j = 2i-1,\end{cases}$$

which gives the desired result

$$\mathbb{E}\left[\max_{i=1,\ldots,n} |X_i|\right] = \mathbb{E}\left[\max_{j=1,\ldots,2n} Y_j\right] \leq \sigma\sqrt{2\log 2n} \leq 2\sigma\sqrt{\log n}.$$

□

Bibliography

[1] E. Kreyszig, *Introductory Functional Analysis with Applications*. Wiley, 1989.

[2] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic fourier series," *Trans. Amer. Math. Soc.*, vol. 73, pp. 341–366, 1952.

[3] I. Daubechies, "The wavelet transform, time-frequency localization and signal analysis," *IEEE Trans. Inform. Theory*, vol. 36, pp. 961–1005, Sept. 1990.

[4] I. Daubechies, *Ten Lectures on Wavelets*. CBMS-NSF Regional Conference Series in Applied Mathematics, 1992.

[5] C. E. Heil and D. F. Walnut, "Continuous and discrete wavelet transforms," *SIAM Rev.*, vol. 31, pp. 628–666, Dec. 1989.

[6] R. M. Young, *An Introduction to Nonharmonic Fourier Series*. New York: Academic Press, 1980.

[7] D. L. Donoho, "De-noising by soft-thresholding," *IEEE Trans. Inform. Theory*, vol. 41, pp. 613–627, Mar. 1995.

[8] D. L. Donoho and I. M. Johnstone, "Ideal spatial adaptation via wavelet shrinkage," *Biometrika*, vol. 81, pp. 425–455, Aug. 1994.

[9] M. Rupf and J. L. Massey, "Optimum sequence multisets for synchronous code-division multiple-access channels," *IEEE Trans. Inform. Theory*, vol. 40, pp. 1261–1266, July 1994.

[10] M. Sandell, *Design and analysis of estimators for multicarrier modulation and ultrasonic imaging*. PhD thesis, Luleå Univ. Technol., Luleå, Sweden, Sept. 1996.

[11] R. W. Heath, Jr. and A. J. Paulraj, "Linear dispersion codes for MIMO systems based on frame theory," *IEEE Trans. Signal Processing*, vol. 50, pp. 2429–2441, Oct. 2002.

[12] M. Rudelson and R. Vershynin, "Geometric approach to error correcting codes and reconstruction of signals," *International Mathematics Research Notices*, no. 64, pp. 4019–4041, 2005.

[13] Y. C. Eldar and G. D. Forney, Jr., "Optimal tight frames and quantum measurement," *IEEE Trans. Inform. Theory*, vol. 48, pp. 599–610, Mar. 2002.

[14] H. Bölcskei and F. Hlawatsch, "Noise reduction in oversampled filter banks using predictive quantization," *IEEE Trans. Inform. Theory*, vol. 47, pp. 155–172, Jan. 2001.

[15] H. Bölcskei, *Oversampled Filter Banks and Predictive Subband Coders*. PhD thesis, Technische Universität Wien, Nov. 1997.

[16] J. J. Benedetto, A. M. Powell, and Ö. Özgür Yılmaz, "Sigma-delta ($\Sigma\Delta$) quantization and finite frames," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1990–2005, May 2006.

[17] D. L. Donoho, "Compressed sensing," *IEEE Trans. Inform. Theory*, vol. 52, pp. 1289–1306, Apr. 2006.

[18] E. J. Candès and T. Tao, "Near-optimal signal recovery from random projections: Universal encoding strategies?," *IEEE Trans. Inform. Theory*, vol. 52, pp. 5406–5425, Dec. 2006.

[19] D. L. Donoho and M. Elad, "Optimally sparse representation in general (nonorthogonal) dictionaries via l^1 minimization," *Proc. National Academy of Sciences of the US*, vol. 100, Mar. 2003.

[20] J. Kovačević and A. Chebira, *An Introduction to Frames*. Foundations and Trends in Signal Processing, NOW Publishers, 2008.

[21] O. Christensen, *An Introduction to Frames and Riesz Bases*. Boston, MA, U.S.A.: Birkhäuser, 2003.

[22] H. Lütkepohl, *Handbook of Matrices*. Chichester, U.K.: Wiley, 1996.

[23] A. Ben-Israel and T. N. Greville, *Generalized Inverses: Theory and Applications*. Canadian Mathematical Society, 2nd ed., 2002.

[24] D. Han and D. R. Larson, "Frames, bases and group representations," *Memoirs of the American Mathematical Society*, vol. 147, 2000.

[25] E. M. Stein and G. Weiss, *Introduction to Fourier Analysis on Euclidean Spaces*. Princeton Univ. Press, 1971.

[26] A. W. Oppenheim, A. S. Willsky, and S. Hamid, *Signals and Systems*. Prentice Hall, 2nd ed., 1996.

[27] W. Heisenberg, *The Physical Principles of the Quantum Theory*. Chicago, IL: Chicago Press, 1930.

[28] W. G. Faris, "Inequalities and uncertainty principles," *J. Math. Phys.*, vol. 19, pp. 461–466, Jan. 1978.

[29] M. G. Cowling and J. F. Price, "Bandwidth versus time concentration: The Heisenberg-Pauli-Weyl inequality," *SIAM J. Math. Anal.*, vol. 15, pp. 151–165, Jan. 1984.

[30] J. J. Benedetto, *Wavelets: Mathematics and Applications*, ch. Frame decompositions, sampling, and uncertainty principle inequalities. New York, NY: CRC Press, 1994.

[31] G. B. Folland and A. Sitaram, "The uncertainty principle: A mathematical survey," *J. Fourier Anal. Appl.*, vol. 3, no. 3, pp. 207–238, 1997.

[32] D. L. Donoho and P. B. Stark, "Uncertainty principles and signal recovery," *SIAM J. Appl. Math.*, vol. 49, pp. 906–931, June 1989.

[33] D. L. Donoho and B. F. Logan, "Signal recovery and the large sieve," *SIAM J. Appl. Math.*, vol. 52, pp. 577–591, Apr. 1992.

[34] M. Elad and A. M. Bruckstein, "A generalized uncertainty principle and sparse representation in pairs of bases," *IEEE Trans. Inform. Theory*, vol. 48, pp. 2558–2567, Sep. 2002.

[35] C. Studer, P. Kuppinger, G. Pope, and H. Bölcskei, "Recovery of sparsely corrupted signals," *IEEE Trans. Inform. Theory*, vol. 58, pp. 3115–3130, May 2012.

[36] P. Kuppinger, G. Durisi, and H. Bölcskei, "Uncertainty relations and sparse signal recovery for pairs of general signal sets," *IEEE Trans. Inform. Theory*, vol. 58, pp. 263–277, Jan. 2012.

[37] A. Terras, *Fourier Analysis on Finite Groups and Applications*. New York, NY: Cambridge Univ. Press, 1999.

[38] T. Tao, "An uncertainty principle for cyclic groups of prime order," *Math. Res. Lett.*, vol. 12, no. 1, pp. 121–127, 2005.

[39] D. Stotz, E. Riegler, E. Agustsson, and H. Bölcskei, "Almost lossless analog signal separation and probabilistic uncertainty relations," *IEEE Trans. Inform. Theory*, vol. 63, pp. 5445–5460, Sep. 2017.

[40] S. Foucart and H. Rauhut, *A Mathematical Introduction to Compressive Sensing*. Basel, Switzerland: Birkhäuser, 2013.

[41] K. Gröchenig, *Foundations of Time-Frequency Analysis*. Birkhäuser, 2001.

[42] D. Gabor, "Theory of communications," *J. Inst. Elec. Eng.*, vol. 96, pp. 429–457, 1946.

[43] H. Bölcskei, *Advances in Gabor analysis*, ch. Orthogonal frequency division multiplexing based on offset QAM. Birkhäuser, 2003.

[44] C. Fefferman, “The uncertainty principle,” *Bull. Amer. Math. Soc.*, vol. 9, pp. 129–206, Sep. 1983.

[45] S. Evra, E. Kowalski, and A. Lubotzky, “Good cyclic codes and the uncertainty principle,” *arXiv:1703.01080*, 2017.

[46] E. Bombieri, *Le grand crible dans la théorie analytique des nombres*. Paris, France: Soc. Math. de France, 1974.

[47] H. L. Montgomery, *Twentieth Century Harmonic Analysis – A Celebration*, ch. Harmonic analysis as found in analytic number theory. Dordrecht, Netherlands: Springer, 2001.

[48] R. A. Horn and C. R. Johnson, *Matrix Analysis*. Cambridge, U.K.: Cambridge Univ. Press, 1990.

[49] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. John Wiley & Sons, 1 ed., 1991.

[50] J. D. Vaaler, “Some extremal functions in Fourier analysis,” *Bull. Amer. Math. Soc.*, vol. 2, no. 12, pp. 183–216, 1985.

[51] C. Heil, *A Basis Theory Primer*. New York, NY: Springer, 2011.

[52] P. Mattila, *Geometry of Sets and Measures in Euclidean Space: Fractals and Rectifiability*. Cambridge, UK: Cambridge Univ. Press, 1999.

[53] R. Gribonval and M. Nielsen, “Highly sparse representations from dictionaries are unique and independent of the sparseness measure,” tech. rep., INRIA, 2003.

[54] L. Welch, “Lower bounds on the maximum cross correlation of signals (corresp.),” *IEEE Trans. Inform. Theory*, vol. 20, no. 3, pp. 397–399, 1974.

[55] G. R. de Prony, “Essai expérimental et analytique: Sur les lois de la dilatabilité de fluides élastiques et sur celles de la force expansive de la vapeur d’eau et de la vapeur de l’alcool à différentes températures,” *Journal de l’école polytechnique*, vol. 1, no. 22, pp. 24–76, 1795.

[56] A. C. Kot, S. Parthasarathyt, D. W. Tufts, and R. J. Vaccaro, “The statistical performance of state-variable balancing and Prony’s method in parameter estimation,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 12, (Dallas, TX, USA), pp. 1549–1552, Apr. 1987.

[57] P. Stoica and A. Nehorai, “Study of the statistical performance of the Pisarenko harmonic decomposition method,” *IEE Proceedings–Radar and Signal Processing*, vol. 135, pp. 161–168, Apr. 1988.

[58] R. O. Schmidt, "Multiple emitter location and signal parameter estimation," *IEEE Transactions on Antennas and Propagation*, vol. 34, pp. 276–280, Mar. 1986.

[59] V. F. Pisarenko, "The retrieval of harmonics from a covariance function," *Geophysical Journal of the Royal Astronomical Society*, vol. 33, pp. 347–366, Feb. 1973.

[60] A. Paulraj, R. Roy, and T. Kailath, "Estimation of signal parameters via rotational invariance techniques – ESPRIT," in *Proceedings of the 19th Asilomar Conference on Circuits, Systems, and Computers*, (Pacific Grove, CA, USA), pp. 83–89, Nov. 1985.

[61] R. Roy, A. Paulraj, and T. Kailath, "ESPRIT – A subspace rotation approach to estimation of parameters of cisoids in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 1340–1342, Oct. 1986.

[62] S. Y. Kung, "A Toeplitz approximation method and some applications," in *Proceedings of the International Symposium on the Mathematical Theory of Networks and Systems*, (Santa Monica, CA, USA), pp. 262–266, Aug. 1981.

[63] S. Kung, K. Arun, and B. Rao, "State-space and singular value decomposition based approximation methods for harmonic retrieval problem," *Journal of the Optical Society of America*, vol. 73, pp. 1799–1811, Dec. 1983.

[64] R. Kumaresan and D. W. Tufts, "Estimating the parameters of exponentially damped sinusoids and pole-zero modeling in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 30, pp. 833–840, Dec. 1982.

[65] Y. Li, K. J. R. Liu, and J. Razavilar, "A parameter estimation scheme for damped sinusoidal signals based on low-rank Hankel approximation," *IEEE Transactions on Signal Processing*, vol. 45, pp. 481–486, Feb. 1997.

[66] Y. Hua and T. K. Sarkar, "Matrix pencil method for estimating parameters of exponentially damped/undamped sinusoids in noise," *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 38, pp. 814–824, May 1990.

[67] Y. Hua and T. K. Sarkar, "On SVD for estimating generalized eigenvalues of singular matrix pencil in noise," *IEEE Transactions on Signal Processing*, vol. 39, pp. 892–900, Apr. 1991.

[68] P. Stoica and R. Moses, *Spectral Analysis of Signals*. Upper Saddle River, NJ, USA: Prentice Hall, 2005.

[69] S. V. Schell and W. A. Gardner, "High-resolution direction finding," in *Handbook of Statistics 10: Signal Processing and its Applications* (N. K. Bose and C. R. Rao, eds.), pp. 755–817, Amsterdam, Netherlands: Elsevier Science Publishers B.V., 1993.

[70] O. Besson and F. Castanié, “On estimating the frequency of a sinusoid in autoregressive multiplicative noise,” *Signal Processing*, vol. 30, pp. 65–83, Jan. 1993.

[71] W. Chen, G. Zhou, and G. B. Giannakis, “Velocity and acceleration estimation of Doppler weather radar/lidar signals in colored noise,” in *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP)*, vol. 3, (Detroit, MI, USA), pp. 2052–2055, May 1995.

[72] A. M. Bruckstein, T.-J. Shan, and T. Kailath, “The resolution of overlapping echos,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 33, pp. 1357–1367, Dec. 1985.

[73] E. J. Candès and C. Fernandez-Granda, “Towards a mathematical theory of super-resolution,” *Comm. Pure and Appl. Math.*, vol. 67, pp. 906–956, June 2014.

[74] M. Vetterli, P. Marziliano, and T. Blu, “Sampling signals with finite rate of innovation,” *IEEE Transactions on Signal Processing*, vol. 50, pp. 1417–1428, June 2002.

[75] K. Gröchenig, “Irregular sampling, Toeplitz matrices, and the approximation of entire functions of exponential type,” *Mathematics of Computation*, vol. 68, pp. 749–765, Apr. 1999.

[76] G. Tang, B. N. Bhaskar, and B. Recht, “Near minimax line spectral estimation,” *IEEE Transactions on Information Theory*, vol. 61, pp. 5987–5999, Dec. 2013.

[77] J. Laroche, “The use of the matrix pencil method for the spectrum analysis of musical signals,” *Journal of the Acoustical Society of America*, vol. 94, pp. 1958–1965, Oct. 1993.

[78] R. J. McAulay and T. F. Quatieri, “Speech analysis and synthesis based on a sinusoidal representation,” *IEEE Transactions on Acoustics, Speech, and Signal Processing*, vol. 34, pp. 744–754, Aug. 1986.

[79] D. David, G. Richard, and R. Badeau, “An EDS modeling tool for tracking and modifying musical signals,” in *Proceedings of the Stockholm Music Acoustic Conference (SMAC)*, (Stockholm, Sweden), pp. 715–718, Aug. 2003.

[80] D. Sanjoy and G. Anupam, “An elementary proof of a theorem of Johnson and Lindenstrauss,” *Random Structures & Algorithms*, vol. 22, no. 1, pp. 60–65, 2003.

[81] R. Baraniuk, M. Davenport, R. DeVore, and M. Wakin, “A simple proof of the restricted isometry property for random matrices,” *Constructive Approximation*, vol. 28, no. 3, pp. 253–263, 2008.

[82] N. Alon, “Problems and results in extremal combinatorics—i,” *Discrete Mathematics*, vol. 273, no. 1, pp. 31 – 53, 2003.

[83] E. Ott, *Chaos in Dynamical Systems*. Cambridge Univ. Press, 2002.

[84] M. Wainwright, *High-dimensional statistics*. Cambridge Series in Statistical and Probabilistic Mathematics, Cambridge University Press, 2019.

[85] R. A. DeVore and G. G. Lorentz, *Constructive Approximation*. Springer, 1993.

[86] R. A. DeVore, "Nonlinear approximation," *Acta Numerica*, vol. 7, pp. 51–150, 1998.

[87] P. Grohs, S. Keiper, G. Kutyniok, and M. Schäfer, "α-molecules," *Appl. Comput. Harmon. Anal.*, vol. 41, no. 1, pp. 297–336, 2016.

[88] I. Daubechies, *Ten Lectures on Wavelets*. SIAM, 1992.

[89] E. J. Candès, "Ridgelets: Theory and Applications," 1998. Ph.D. thesis, Stanford University.

[90] E. J. Candès and D. L. Donoho, "New tight frames of curvelets and optimal representations of objects with piecewise C2 singularities," *Comm. Pure Appl. Math.*, vol. 57, pp. 219–266, 2002.

[91] K. Guo, G. Kutyniok, and D. Labate, "Sparse multidimensional representations using anisotropic dilation and shear operators," in *Wavelets and Splines (Athens, GA, 2005)*, pp. 189–201, Nashboro Press, Nashville, TN, 2006.

[92] P. Grohs and G. Kutyniok, "Parabolic molecules," *Found. Comput. Math.*, vol. 14, pp. 299–337, 2014.

[93] K. Gröchenig, *Foundations of time-frequency analysis*. Springer Science & Business Media, 2013.

[94] K. Gröchenig and S. Samarah, "Nonlinear approximation with local Fourier bases," *Constructive Approximation*, vol. 16, pp. 317–331, Jul 2000.

[95] L. Demanet and L. Ying, "Wave atoms and sparsity of oscillatory patterns," *Appl. Comput. Harmon. Anal.*, vol. 23, no. 3, pp. 368–387, 2007.

[96] D. L. Donoho, "Unconditional bases are optimal bases for data compression and for statistical estimation," *Appl. Comput. Harmon. Anal.*, vol. 1, no. 1, pp. 100 – 115, 1993.

[97] P. Grohs, "Optimally sparse data representations," in *Harmonic and Applied Analysis*, pp. 199–248, Springer, 2015.

[98] S. Boucheron, G. Lugosi, and P. Massart, *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.