Lecture Notes
Mathematics of Information

Helmut Bölcskei

March 2020
## 2.3 Concentration Inequalities

2.3.2 Concentration Inequalities ........................................... 63
2.3.3 Noisy Recovery in \( (\mathbb{C}^m, \|\cdot\|_1) \) ....................... 64
2.3.4 Coherence-based Uncertainty Relation for Pairs of General Matrices ...................................................... 65
2.3.5 Concentration Inequalities for Pairs of General Matrices .......................................................... 67

2.4 A Large Sieve Inequality in \( (\mathbb{C}^m, \|\cdot\|_2) \) ........ 69

2.5 Uncertainty Relations in \( L_1 \) and \( L_2 \) .................. 72

2.6 Results for \( \|\|\cdot\|_1 \) and \( \|\|\cdot\|_2 \) .................. 72

## 3 Compressive Sensing 75

3.1 Discrete Fourier Transform ........................................... 75

3.1.1 Oversampling .......................................................... 76
3.1.2 Undersampling ......................................................... 77

3.2 Compressive Sensing ..................................................... 78

3.2.1 Incoherence .............................................................. 81
3.2.2 The General Problem ................................................. 82

3.3 The Recovery Problem \( (P_0) \) ........................................... 83

3.4 Basis Pursuit \( (BP) \) ....................................................... 85

3.5 Signal Separation, Super-Resolution, Recovery of Corrupted Signals .......................................................... 89

3.5.1 Inpainting and Super-resolution .................................... 89
3.5.2 Clipping ................................................................. 90
3.5.3 Signal Separation ...................................................... 90
3.5.4 Recovery of Signals Subject to Impulse Noise .................. 90
3.5.5 Recovery of Signals Subject to Narrowband Interference .. 90
3.5.6 The General Problem .................................................. 91
3.5.7 Concatenation of ONBs (or Frames) ............................ 91
3.5.8 General Matrices and Different Scenarios ...................... 92

## 4 Finite Rate of Innovation 98

4.1 Finding the annihilating filter ......................................... 99

4.1.1 Finding the \( a_k \) ......................................................... 100
4.1.2 Finding the \( c_k \) ......................................................... 100
4.1.3 Uniqueness .............................................................. 100

## 5 Sampling of Multi-Band Signals 102

5.1 Classical Sampling Theorem ........................................... 102

5.2 Sampling Spectrally Sparse Signals ............................... 103

5.2.1 Interpretation of the Lower Beurling Density ................. 104
5.2.2 Stable Sampling ....................................................... 105

5.3 Multicoset Sampling ..................................................... 106

5.4 Spectrum-Blind Sampling .............................................. 109
6  The restricted isometry property and its implications  111

7  The Johnson-Lindenstrauss Lemma  118

8  Verifying the RIP from concentration inequalities  121
0.1 Notation

\( a, b, \ldots \) vectors
\( A, B, \ldots \) matrices
\( a^T, A^T \) transpose of the vector \( a \) and the matrix \( A \)
\( a^*, a^*, A^* \) complex conjugate of the scalar \( a \), element-wise complex conjugate of the vector \( a \), and the matrix \( A \)
\( a^H, A^H \) Hermitian transpose of the vector \( a \) and the matrix \( A \)
\( I_N \) identity matrix of size \( N \times N \)
\( \text{rank}(A) \) rank of the matrix \( A \)
\( \lambda(A) \) eigenvalue of the matrix \( A \)
\( \lambda_{\text{min}}(A), \lambda_{\text{min}}(A) \) smallest eigenvalue of the matrix \( A \), smallest spectral value of the self-adjoint operator \( A \)
\( \lambda_{\text{max}}(A), \lambda_{\text{max}}(A) \) largest eigenvalue of the matrix \( A \), largest spectral value of the self-adjoint operator \( A \)
\( i \) \( \sqrt{-1} \)
\( \triangleq \) 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
\( \ell^2 \) Hilbert space of square-summable sequences
\( \langle a, b \rangle \) inner product of the vectors \( a \) and \( b \): \( \langle a, b \rangle \triangleq \sum_i [a]_i (b)_i^* \)
\( \langle x, y \rangle \) depending on the context: inner product in the abstract Hilbert space \( \mathcal{H} \) or inner product of the functions \( x(t) \) and \( y(t) \): \( \langle x, y \rangle \triangleq \int_{-\infty}^{\infty} x(t)y^*(t)dt \)
\( \|a\|_2^2 \) squared \( \ell^2 \)-norm of the vector \( a \): \( \|a\|_2^2 \triangleq \sum_i |[a]_i|^2 \)
\( \|y\|_2^2 \) depending on the context: squared norm in the abstract Hilbert space \( \mathcal{H} \) or squared \( \mathcal{L}^2 \)-norm of the function \( y(t) \): \( \|y\|_2^2 \triangleq \int_{-\infty}^{\infty} |y(t)|^2 dt \)
\( \mathbb{I}_\mathcal{H}, \mathbb{I}_{\ell^2}, \mathbb{I}_{\mathcal{L}^2}, \mathbb{I}_{\mathcal{L}^2(B)} \) identity operator in the corresponding space
\( \mathcal{R}(A) \) range space of operator \( A \)
\( A^* \) adjoint of operator \( A \)
\( \hat{x}(f) \) Fourier transform of \( x(t) \): \( \hat{x}(f) \triangleq \int_{-\infty}^{\infty} x(t)e^{-2\pi tif}dt \)
\( \hat{x}_d(f) \) Discrete-time Fourier transform of \( x[k] \): \( \hat{x}_d(f) \triangleq \sum_{k=-\infty}^{\infty} x[k]e^{-2\pi kf} \)
Chapter 1

A Short Course on Frame Theory

Hilbert spaces \([1, \text{Def. 3.1-1}]\) 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 1.1.1** (Orthonormal basis in \(\mathbb{R}^2\)). Consider the orthonormal basis (ONB)

\[
e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}
\]

in \(\mathbb{R}^2\) (see Figure 1.1). We can represent every signal \(x \in \mathbb{R}^2\) as the following linear combination of the basis vectors \(e_1\) and \(e_2\):

\[
x = \langle x, e_1 \rangle e_1 + \langle x, e_2 \rangle e_2.
\]

(1.1)
To rewrite (1.1) in vector-matrix notation, we start by defining the vector of expansion coefficients as
\[
c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} \triangleq \begin{bmatrix} (x, e_1) \\ (x, e_2) \end{bmatrix} = \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix} x = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x.
\]
It is convenient to define the matrix
\[
T \triangleq \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.
\]
Henceforth we call \( T \) the analysis matrix; it multiplies the signal \( x \) to produce the expansion coefficients
\[
c = Tx.
\]
Following (1.1), we can reconstruct the signal \( x \) from the coefficient vector \( c \) according to
\[
x = T^T c = \begin{bmatrix} e_1 & e_2 \end{bmatrix} c = \begin{bmatrix} e_1 & e_2 \end{bmatrix} \begin{bmatrix} (x, e_1) \\ (x, e_2) \end{bmatrix} = (x, e_1) e_1 + (x, e_2) e_2.
\]
We call
\[
T^T = \begin{bmatrix} e_1 & e_2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}
\]
the synthesis matrix; it multiplies the coefficient vector \( c \) to recover the signal \( x \). It follows from (1.2) that (1.1) is equivalent to
\[
x = T^T Tx = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} x.
\]

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 \( T \) and \( T^T \) 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.
Example 1.1.2 (Biorthonormal bases in $\mathbb{R}^2$). Consider two noncollinear unit norm vectors in $\mathbb{R}^2$. For concreteness, take (see Figure 1.2)

$$e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad e_2 = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 \\ 1 \end{bmatrix}.$$  

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

$$c_1 \triangleq \langle x, e_1 \rangle, \quad c_2 \triangleq \langle x, e_2 \rangle.$$  

As in Example 1.1.1 above, we stack the expansion coefficients into a vector so that

$$c = \begin{bmatrix} c_1 \\ c_2 \end{bmatrix} = \begin{bmatrix} \langle x, e_1 \rangle \\ \langle x, e_2 \rangle \end{bmatrix} = \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix} x = \begin{bmatrix} 1/\sqrt{2} \\ 0 \end{bmatrix} x = \begin{bmatrix} 1 \\ 1/\sqrt{2} \end{bmatrix} x.$$  

(1.5)

Analogously to Example 1.1.1, we can define the analysis matrix

$$T \triangleq \begin{bmatrix} e_1^T \\ e_2^T \end{bmatrix} = \begin{bmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{bmatrix}$$  

and rewrite (1.5) as

$$c = Tx.$$  

Now, obviously, the vectors $e_1$ and $e_2$ are not orthonormal (or, equivalently, $T$ is not unitary) so that we cannot write $x$ in the form (1.1). We could, however, try to find a decomposition of $x$ of the form

$$x = \langle x, e_1 \rangle \tilde{e}_1 + \langle x, e_2 \rangle \tilde{e}_2$$  

(1.6)

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

$$x = \begin{bmatrix} \tilde{e}_1 \\ \tilde{e}_2 \end{bmatrix} T x$$  

(1.7)
and choosing the vectors $\tilde{e}_1$ and $\tilde{e}_2$ to be given by the columns of $T^{-1}$ according to

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

(1.8)

Note that $T$ is invertible as a consequence of $e_1$ and $e_2$ not being collinear. For the specific example at hand we find

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

and therefore (see Figure 1.2)

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

Note that (1.8) implies that $T \begin{bmatrix}
\tilde{e}_1 & \tilde{e}_2
\end{bmatrix} = I_2$, which is equivalent to

$$
\begin{bmatrix}
e_1^T \\
e_2^T
\end{bmatrix} \begin{bmatrix}
\tilde{e}_1 & \tilde{e}_2
\end{bmatrix} = I_2.
$$

More directly the two sets of vectors $\{e_1, e_2\}$ and $\{\tilde{e}_1, \tilde{e}_2\}$ satisfy a “biorthonormality” property according to

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

We say that $\{e_1, e_2\}$ and $\{\tilde{e}_1, \tilde{e}_2\}$ are biorthonormal bases. Analogously to (1.3), we can now define the synthesis matrix as follows:

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

Our observations can be summarized according to

$$
x = \langle x, e_1 \rangle \tilde{e}_1 + \langle x, e_2 \rangle \tilde{e}_2 \\
= T^T c = \tilde{T}^T T x \\
= \begin{bmatrix}
1 & 0 \\
-1 & \sqrt{2}
\end{bmatrix} \begin{bmatrix}
1 & 0 \\
1/\sqrt{2} & 1/\sqrt{2}
\end{bmatrix} x = \begin{bmatrix}
1 & 0 \\
0 & 1
\end{bmatrix} x.
$$

(1.9)

Comparing (1.9) to (1.4), we observe the following: To synthesize $x$ from the expansion coefficients $c$ corresponding to the nonorthogonal set $\{e_1, e_2\}$, we need to use the synthesis matrix $\tilde{T}^T$ obtained from the set $\{\tilde{e}_1, \tilde{e}_2\}$, which forms a biorthonormal pair with $\{e_1, e_2\}$. In Example 1.1.1 $\{e_1, e_2\}$ is an orthonormal basis (ONB) and hence $\tilde{T} = T$, or, equivalently, $\{e_1, e_2\}$ forms a biorthonormal pair with itself.
As the vectors \( e_1 \) and \( e_2 \) are linearly independent, the \( 2 \times 2 \) analysis matrix \( T \) has full rank and is hence invertible, i.e., there is a unique matrix \( T^{-1} \) that satisfies \( T^{-1}T = I_2 \). According to (1.7) this means that for each analysis set \( \{e_1, e_2\} \) there is precisely one synthesis set \( \{\tilde{e}_1, \tilde{e}_2\} \) such that (1.6) is satisfied for all \( x \in \mathbb{R}^2 \).

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

\[
g_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}, \quad g_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad g_3 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}.
\]

Three vectors in a two-dimensional space are always linearly dependent. In particular, in this example we have \( g_3 = g_1 - g_2 \). Let us compute the expansion coefficients \( c \) corresponding to \( \{g_1, g_2, g_3\} \):

\[
c = \begin{bmatrix} c_1 \\ c_2 \\ c_3 \end{bmatrix} \triangleq \begin{bmatrix} \langle x, g_1 \rangle \\ \langle x, g_2 \rangle \\ \langle x, g_3 \rangle \end{bmatrix} = \begin{bmatrix} g_1^T \\ g_2^T \\ g_3^T \end{bmatrix} x = \begin{bmatrix} 1 \\ 0 \\ 1 \\ -1 \end{bmatrix} x.
\]

Following Examples 1.1.1 and 1.1.2 we define the analysis matrix

\[
T \triangleq \begin{bmatrix} g_1^T \\ g_2^T \\ g_3^T \end{bmatrix} \triangleq \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 1 & -1 \end{bmatrix}
\]

and rewrite (1.10) as

\[
c = T x.
\]
Note that here, unlike in Examples 1.1.1 and 1.1.2, \(c\) is a redundant representation of \(x\) as we have three expansion coefficients for a two-dimensional signal \(x\).

We next ask if \(x\) can be represented as a linear combination of the form

\[
x = \langle x, g_1 \rangle \tilde{g}_1 + \langle x, g_2 \rangle \tilde{g}_2 + \langle x, g_3 \rangle \tilde{g}_3
\]  

(1.11)

with \(\tilde{g}_1, \tilde{g}_2, \tilde{g}_3 \in \mathbb{R}^2\)? To answer this question (in the affirmative) we first note that the vectors \(g_1, g_2\) form an orthonormal basis (ONB) for \(\mathbb{R}^2\). We therefore know that the following is true:

\[
x = \langle x, g_1 \rangle g_1 + \langle x, g_2 \rangle g_2.
\]  

(1.12)

Setting \(\tilde{g}_1 = g_1, \tilde{g}_2 = g_2, \tilde{g}_3 = 0\) 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):

\[
x = \langle x, g_1 \rangle g_1 + \langle x, g_2 \rangle g_2 + \langle x, g_1 - g_2 \rangle (g_1 - g_1) + 0.
\]

Rearranging terms in this expression, we obtain

\[
x = \langle x, g_1 \rangle 2g_1 + \langle x, g_2 \rangle (g_2 - g_1) - \langle x, g_1 - g_2 \rangle g_1.
\]  

(1.13)

We recognize that \(g_1 - g_2 = g_3\) and set

\[
\tilde{g}_1 = 2g_1, \quad \tilde{g}_2 = g_2 - g_1, \quad \tilde{g}_3 = -g_1.
\]  

(1.14)

This allows us to rewrite (1.13) as

\[
x = \langle x, g_1 \rangle \tilde{g}_1 + \langle x, g_2 \rangle \tilde{g}_2 + \langle x, g_3 \rangle \tilde{g}_3.
\]

The redundant set of vectors \(\{g_1, g_2, g_3\}\) is called a frame. The set \(\{\tilde{g}_1, \tilde{g}_2, \tilde{g}_3\}\) in (1.14) is called a dual frame to the frame \(\{g_1, g_2, g_3\}\). Obviously another dual frame is given by \(\tilde{g}_1 = g_1, \tilde{g}_2 = g_2,\) and \(\tilde{g}_3 = 0\). In fact, there are infinitely many dual frames. To see this, we first define the synthesis matrix corresponding to a dual frame \(\{\tilde{g}_1, \tilde{g}_2, \tilde{g}_3\}\) as

\[
\tilde{T}^T \triangleq \begin{bmatrix} \tilde{g}_1 & \tilde{g}_2 & \tilde{g}_3 \end{bmatrix}.
\]  

(1.15)

It then follows that we can write

\[
x = \langle x, g_1 \rangle \tilde{g}_1 + \langle x, g_2 \rangle \tilde{g}_2 + \langle x, g_3 \rangle \tilde{g}_3
\]

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

Thanks to the redundancy of the frame $\{g_1, g_2, g_3\}$, we obtain design freedom: In order to synthesize the signal $x$ from its expansion coefficients $c_k = \langle x, g_k \rangle$, $k = 1, 2, 3$, in the frame $\{g_1, g_2, g_3\}$, we can choose between infinitely many dual frames $\{\tilde{g}_1, \tilde{g}_2, \tilde{g}_3\}$. 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 1.2.1.** The set of vectors $\{e_k\}_{k=1}^M, e_k \in \mathbb{C}^M$, is called an ONB for $\mathbb{C}^M$ if

1. $\text{span}\{e_k\}_{k=1}^M = \{c_1e_1 + c_2e_2 + \ldots + c_Me_M \ | \ c_1, c_2, \ldots, c_M \in \mathbb{C}\} = \mathbb{C}^M$

2. $\langle e_k, e_j \rangle = \begin{cases} 1, & k = j \\ 0, & k \neq j \end{cases}$ for $k, j = 1, \ldots, M$.

When $\{e_k\}_{k=1}^M$ is an ONB, thanks to the spanning property in Definition 1.2.1, every $x \in \mathbb{C}^M$ can be decomposed as

$$x = \sum_{k=1}^M c_k e_k. \quad (1.16)$$

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

$$\langle x, e_j \rangle = \left\langle \sum_{k=1}^M c_k e_k, e_j \right\rangle = \sum_{k=1}^M c_k \langle e_k, e_j \rangle = c_j.$$

In summary, we have the decomposition

$$x = \sum_{k=1}^M \langle x, e_k \rangle e_k.$$
Just like in Example 1.1.1 in the previous section, we define the analysis matrix

\[ T \triangleq \begin{bmatrix} e_1^H \\ \vdots \\ e_M^H \end{bmatrix}. \]

If we organize the inner products \( \{ \langle x, e_k \rangle \}_{k=1}^M \) into the vector \( c \), we have

\[ c \triangleq \begin{bmatrix} \langle x, e_1 \rangle \\ \vdots \\ \langle x, e_M \rangle \end{bmatrix} = Tx = \begin{bmatrix} e_1^H \\ \vdots \\ e_M^H \end{bmatrix} x. \]

Thanks to the orthonormality of the vectors \( e_1, e_2, \ldots, e_M \) the matrix \( T \) is unitary, i.e., \( T^H = T^{-1} \) and hence

\[ TT^H = \begin{bmatrix} e_1^H \\ \vdots \\ e_M^H \end{bmatrix} \begin{bmatrix} e_1 & \ldots & e_M \end{bmatrix} = \begin{bmatrix} \langle e_1, e_1 \rangle & \cdots & \langle e_M, e_1 \rangle \\ \vdots & \ddots & \vdots \\ \langle e_1, e_M \rangle & \cdots & \langle e_M, e_M \rangle \end{bmatrix} = I_M = T^H T. \]

Thus, if we multiply the vector \( c \) by \( T^H \), we synthesize \( x \) according to

\[ T^H c = T^H Tx = \sum_{k=1}^M \langle x, e_k \rangle e_k = I_M x = x. \]  (1.17)

We shall therefore call the matrix \( T^H \) the synthesis matrix, corresponding to the analysis matrix \( T \). In the \textbf{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 1.2.1 and consider general bases.

**Definition 1.2.2.** The set of vectors \( \{ e_k \}_{k=1}^M, e_k \in \mathbb{C}^M \), is a basis for \( \mathbb{C}^M \) if

1. \( \text{span}\{e_k\}_{k=1}^M = \{ c_1 e_1 + c_2 e_2 + \ldots + c_M e_M \mid c_1, c_2, \ldots, c_M \in \mathbb{C} \} = \mathbb{C}^M \)

2. \( \{ e_k \}_{k=1}^M \) is a linearly independent set, i.e., if \( \sum_{k=1}^M c_k e_k = 0 \) for some scalar coefficients \( \{ c_k \}_{k=1}^M \), then necessarily \( c_k = 0 \) for all \( k = 1, \ldots, M \).

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

\[ c_k \triangleq \langle x, e_k \rangle, \quad k = 1, \ldots, M. \]  (1.18)
Again, it is convenient to introduce the analysis matrix

\[
T \triangleq \begin{bmatrix}
e^H_1 \\
\vdots \\
e^H_M
\end{bmatrix}
\]

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

\[
c = Tx.
\]

Next, let us ask how we can find a set of vectors \(\{\tilde{e}_1, \ldots, \tilde{e}_M\}\), \(\tilde{e}_k \in \mathbb{C}^M\), \(k = 1, \ldots, M\), that is dual to the set \(\{e_1, \ldots, e_M\}\) in the sense that

\[
x = \sum_{k=1}^M c_k \tilde{e}_k = \sum_{k=1}^M \langle x, e_k \rangle \tilde{e}_k
\]

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

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

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

\[
x = \tilde{T}^H c = \tilde{T}^H Tx.
\]

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

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

Next, note that \(TT^H = I_M\) is equivalent to

\[
\begin{bmatrix}
\langle \tilde{e}_1, e_1 \rangle & \cdots & \langle \tilde{e}_1, e_M \rangle \\
\vdots & \ddots & \vdots \\
\langle \tilde{e}_M, e_1 \rangle & \cdots & \langle \tilde{e}_M, e_M \rangle
\end{bmatrix} = I_M
\]

or equivalently

\[
\langle e_k, \tilde{e}_j \rangle = \begin{cases} 
1, & k = j \\
0, & \text{else}
\end{cases}, \quad k, j = 1, \ldots, M.
\]

(1.20)

The sets \(\{e_k\}_{k=1}^M\) and \(\{\tilde{e}_k\}_{k=1}^M\) are biorthonormal bases. ONBs are biorthonormal to themselves in this terminology, as already noted in Example 1.1.2. We emphasize that it is the fact that \(T\) and
\( \tilde{T}^H \) are square and full-rank that allows us to conclude that \( \tilde{T}^H T = I_M \) implies \( T T^H = I_M \) and hence to conclude that (1.20) holds. We shall see below that for redundant expansions \( T \) is a tall matrix and \( \tilde{T}^H T \neq T T^H \) (\( \tilde{T}^H T \) and \( T T^H \) have different dimensions) so that dual frames will not be biorthonormal.

As \( T \) is a square matrix and of full rank, its inverse is unique, which means that for a given analysis set \( \{ e_k \}_{k=1}^M \), the synthesis set \( \{ \tilde{e}_k \}_{k=1}^M \) is unique. Alternatively, for a given synthesis set \( \{ \tilde{e}_k \}_{k=1}^M \), there is a unique analysis set \( \{ e_k \}_{k=1}^M \). This uniqueness property is not always desirable. For example, one may want to impose certain structural properties on the synthesis set \( \{ \tilde{e}_k \}_{k=1}^M \) in which case having freedom in choosing the synthesis set as in Example 1.1.2 is helpful.

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

\[
\| c \|^2 = c^H c = x^H \tilde{T}^H T x = x^H I_M x = \| x \|^2,
\]

where we used (1.17). Bioorthonormal 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

\[
\lambda_{\min}(T^H T) \| x \|^2 \leq \| c \|^2 = x^H \tilde{T}^H T x \leq \lambda_{\max}(T^H T) \| x \|^2.
\]

### 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 1.1.3 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 \( c_3 = \langle x, g_3 \rangle \) in Example 1.1.3 is corrupted or even completely lost. We can still reconstruct \( x \) 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 \( \{ g_1, \ldots, g_N \} \), \( g_k \in \mathbb{C}^M \), \( k = 1, \ldots, N \), with \( N \geq M \). Clearly, when \( N \) is strictly greater than \( M \), the vectors \( g_1, \ldots, g_N \) must be linearly
dependent. Next, consider a signal $x \in \mathbb{C}^M$ and compute the expansion coefficients
\[ c_k = \langle x, g_k \rangle, \quad k = 1, \ldots, N. \] (1.23)

Just as before, it is convenient to introduce the analysis matrix
\[ T \triangleq \begin{bmatrix} g_1^H \\ \vdots \\ g_N^H \end{bmatrix} \] (1.24)
and to stack the coefficients $\{c_k\}_{k=1}^N$ in the vector $c$. Then (1.23) can be written as
\[ c = Tx. \] (1.25)

Note that $c \in \mathbb{C}^N$ and $x \in \mathbb{C}^M$. 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 $x$ and the expansion coefficient vector $c$ will, in general, belong to different Hilbert spaces.

The question now is: How can we find a set of vectors $\{\tilde{g}_1, \ldots, \tilde{g}_N\}$, $\tilde{g}_k \in \mathbb{C}^M$, $k = 1, \ldots, N$, such that
\[ x = \sum_{k=1}^N c_k \tilde{g}_k = \sum_{k=1}^N \langle x, g_k \rangle \tilde{g}_k \] (1.26)
for all $x \in \mathbb{C}^M$? If we introduce the synthesis matrix
\[ \tilde{T}^H \triangleq [\tilde{g}_1 \cdots \tilde{g}_N], \]
we can rewrite (1.26) in vector-matrix notation as follows
\[ x = \tilde{T}^H c = \tilde{T}^H Tx. \] (1.27)

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

First note that $T$ is left-invertible if and only if $\mathbb{C}^M = \text{span}\{g_k\}_{k=1}^N$, i.e., if and only if the set of vectors $\{g_k\}_{k=1}^N$ spans $\mathbb{C}^M$. Next observe that when $N > M$, the $N \times M$ matrix $T$ 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 1.2.3.** Let $A \in \mathbb{C}^{N \times M}$, $N \geq M$. Assume that $\text{rank}(A) = M$. Then $A^\dagger \triangleq (A^H A)^{-1} A^H$ is a left-inverse of $A$, i.e., $A^\dagger A = I_M$. Moreover, the general solution $L \in \mathbb{C}^{M \times N}$ of the equation $LA = I_M$ is given by
\[ L = A^\dagger + M(I_N - AA^\dagger), \] (1.28)
where $M \in \mathbb{C}^{M \times N}$ is an arbitrary matrix.
Proof. Since \( \text{rank}(A) = M \), the matrix \( A^H A \) is invertible and hence \( A^\dagger \) is well defined. Now, let us verify that \( A^\dagger \) is, indeed, a left-inverse of \( A \):

\[
A^\dagger A = (A^H A)^{-1} A^H A = I_M. \quad (1.29)
\]

The matrix \( A^\dagger \) is called the Moore-Penrose inverse of \( A \).

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

\[
LA = (A^\dagger + M(I_N - AA^\dagger)) A
\]

\[
= A^\dagger A + MA - MA A^\dagger I_M
\]

\[
= I_M + MA - MA = I_M;
\]

where we used (1.29) twice.

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

\[
LAA^\dagger = A^\dagger.
\]

Adding \( L \) to both sides of this equation and rearranging terms yields

\[
L = A^\dagger + L - LAA^\dagger = A^\dagger + L(I_N - AA^\dagger),
\]

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

We conclude that for each redundant set of vectors \( \{g_1, \ldots, g_N\} \) that spans \( \mathbb{C}^M \), there are infinitely many dual sets \( \{\tilde{g}_1, \ldots, \tilde{g}_N\} \) such that the decomposition (1.26) holds for all \( x \in \mathbb{C}^M \). These dual sets are obtained by identifying \( \{\tilde{g}_1, \ldots, \tilde{g}_N\} \) with the columns of \( L \) according to

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

where \( L \) can be written as follows

\[
L = T^\dagger + M(I_N - TT^\dagger)
\]

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

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

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

is called the canonical dual of \( \{g_1, \ldots, g_N\} \). Using (1.24), we see that in this case

\[
\tilde{g}_k = (T^HT)^{-1}g_k, \quad k = 1, \ldots, N. \quad (1.30)
\]
Note that unlike in the case of a basis, the equation \( \tilde{T}H\tilde{T} = I_M \) does not imply that the sets \( \{ \tilde{g}_k \}_{k=1}^N \) and \( \{ g_k \}_{k=1}^N \) are biorthonormal. This is because the matrix \( T \) is not a square matrix, and thus, \( \tilde{T}H\tilde{T} \neq TT^H \) (\( T^H \) and \( TT^H \) have different dimensions).

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

\[
\| c \|^2 = x^H\tilde{T}H\tilde{T}x
\]

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

\[
\lambda_{\min}(T^HT) \| x \|^2 \leq \| c \|^2 \leq \lambda_{\max}(T^HT) \| x \|^2
\] (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 \( \{g_k\}_{k \in K} \) (\( K \) 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 \( T \) associated to the (possibly nonorthogonal or redundant) set of vectors \( \{g_1, \ldots, g_N\} \). 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 \( T \) that assigns to each signal \( x \in \mathcal{H} \) the sequence of inner products \( T x = \{ \langle x, g_k \rangle \}_{k \in K} \). Throughout this section, we assume that \( \{g_k\}_{k \in K} \) is a Bessel sequence, i.e.,

\[
\sum_{k \in K} |\langle x, g_k \rangle|^2 < \infty \text{ for all } x \in \mathcal{H}.
\]

**Definition 1.3.1.** The linear operator \( T \) is defined as the operator that maps the Hilbert space \( \mathcal{H} \) into the space \( l^2 \) of square-summable complex sequences\[\] \( T : \mathcal{H} \to l^2 \), by assigning to each signal \( x \in \mathcal{H} \) the sequence of inner products \( \langle x, g_k \rangle \) according to

\[
T : x \to \{ \langle x, g_k \rangle \}_{k \in K}.
\]

\[\]

\[\]\[\]\[The fact that the range space of \( T \) is contained in \( l^2 \) is a consequence of \( \{g_k\}_{k \in K} \) being a Bessel sequence.\]
Note that $\|Tx\|^2 = \sum_{k \in K} |\langle x, g_k \rangle|^2$, i.e., the energy $\|Tx\|^2$ of $Tx$ can be expressed as

$$\|Tx\|^2 = \sum_{k \in K} |\langle x, g_k \rangle|^2.$$  *(1.32)*

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

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

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

$$A \|x - y\|^2 \leq \|Tx - Ty\|^2.$$  *(1.33)*

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

$$A \|z\|^2 \leq \|Tz\|^2$$  *(1.34)*

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

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

$$\|Tx\|^2 = \sum_{k=1}^M |\langle x, e_k \rangle|^2 = \|x\|^2, \text{ for all } x \in \mathbb{C}^M.$$  *(1.35)*

This property is a consequence of the unitarity of $T = T$ and it is clear that it will not hold for general sets $\{g_k\}_{k \in K}$ (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

$$\|Tx\|^2 = \sum_{k \in K} |\langle x, g_k \rangle|^2 \leq B \|x\|^2.$$  *(1.36)*

Together with *(1.33)* this “sandwiches” the quantity $\|Tx\|^2$ according to

$$A \|x\|^2 \leq \|Tx\|^2 \leq B \|x\|^2.$$  *(1.37)*

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

\[\text{Note that if } *(1.35) \text{ is satisfied with } B < \infty, \text{ then } \{g_k\}_{k \in K} \text{ is a Bessel sequence.}\]
CHAPTER 1. A SHORT COURSE ON FRAME THEORY

**Definition 1.3.2.** 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

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

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 1.3.3** ([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 1.3.4** ([21]). Starting from the ONB \( \{e_k\}_{k=1}^{\infty} \), we can construct another redundant set as follows

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

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

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

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

\[
A \|x\|^2 \leq \|Tx\|^2 \leq B \|x\|^2, \quad \text{for all } x \in \mathcal{H}.
\]

This means that the energy in the coefficient sequence \( T x \) 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 \( T \) 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 1.3.5. A set of elements \( \{g_k\}_{k \in K}, g_k \in \mathcal{H}, k \in K, \) is complete for the Hilbert space \( \mathcal{H} \) if \( \langle x, g_k \rangle = 0 \) for all \( k \in K \) 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 \( \{g_k\}_{k \in K} \) is complete for \( \mathcal{H} \), take an arbitrary signal \( x \in \mathcal{H} \) and assume that \( \langle x, g_k \rangle = 0 \) for all \( k \in K \). Due to the existence of a lower frame bound \( A > 0 \) we have
\[
A \|x\|^2 \leq \sum_{k \in K} |\langle x, g_k \rangle|^2 = 0,
\]
which implies \( \|x\|^2 = 0 \) and hence \( x = 0 \).

Finally, note that the existence of an upper frame bound \( B < \infty \) guarantees that \( T \) is a bounded linear operator (see [1, Def. 2.7-1]), and, therefore (see [1, Th. 2.7-9]), continuous (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 \( \{\langle x, g_k \rangle\}_{k \in K} \). In Section 1.2.3 we saw that in the finite-dimensional case, this can be accomplished according to
\[
x = \sum_{k=1}^{N} \langle x, g_k \rangle \tilde{g}_k.
\]
Here \( \{\tilde{g}_1, \ldots, \tilde{g}_N\} \) can be chosen to be the canonical dual to the set \( \{g_1, \ldots, g_N\} \) obtained as follows: \( \tilde{g}_k = (T^H T)^{-1} g_k, k = 1, \ldots, N \). We already know that \( T \) is the generalization of \( T \) to the infinite-dimensional setting. Which operator will then correspond to \( T^H \)? To answer this question we start with a definition.

Definition 1.3.6. The linear operator \( T^\times \) is defined as
\[
T^\times : l^2 \rightarrow \mathcal{H}
\]
\[
T^\times : \{c_k\}_{k \in K} \rightarrow \sum_{k \in K} c_k g_k.
\]

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

Definition 1.3.7. Let \( \mathcal{A} : \mathcal{H} \rightarrow \mathcal{H}' \) be a bounded linear operator between the Hilbert spaces \( \mathcal{H} \) and \( \mathcal{H}' \). The unique bounded linear operator \( \mathcal{A}^* : \mathcal{H}' \rightarrow \mathcal{H} \) that satisfies
\[
\langle \mathcal{A} x, y \rangle = \langle x, \mathcal{A}^* y \rangle \tag{1.37}
\]
for all \( x \in \mathcal{H} \) and all \( y \in \mathcal{H}' \) is called the adjoint of \( \mathcal{A} \).

---

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

4Let \( \mathcal{H} \) and \( \mathcal{H}' \) be Hilbert spaces and \( \mathcal{A} : \mathcal{H} \rightarrow \mathcal{H}' \) a linear operator. The operator \( \mathcal{A} \) 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 \( \|\mathcal{A} x - \mathcal{A} x_0\| < \epsilon \). The operator \( \mathcal{A} \) 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 \( A \in \mathbb{C}^{N \times M}, x \in \mathbb{C}^M, y \in \mathbb{C}^N \), then

\[
\langle Ax, y \rangle = y^H Ax = (A^H y)^H x = \langle x, A^H y \rangle,
\]

which, comparing to (1.37), shows that \( A^H \) corresponds to \( A^* \).

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

\[
\langle T x, \{c_k\}_{k \in K} \rangle = \langle x, T^\times \{c_k\}_{k \in K} \rangle.
\]

This can be established by noting that

\[
\langle T x, \{c_k\}_{k \in K} \rangle = \sum_{k \in K} \langle x, g_k \rangle c_k^* \\
\langle x, T^\times \{c_k\}_{k \in K} \rangle = \left( x, \sum_{k \in K} c_k g_k \right) = \sum_{k \in K} c_k^* \langle x, g_k \rangle.
\]

We therefore showed that the adjoint operator of \( T \) is \( T^\times \), i.e.,

\[
T^\times = T^*.
\]

In what follows, we shall always write \( T^* \) instead of \( T^\times \). 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, \( T^* \) is the generalization of \( T^H \) to the infinite-dimensional setting.

### 1.3.1 The Frame Operator

Let us return to the discussion we had immediately before Definition 1.3.6. We saw that in the finite-dimensional case, the canonical dual set \( \{\tilde{g}_1, \ldots, \tilde{g}_N\} \) to the set \( \{g_1, \ldots, g_N\} \) can be computed as follows:

\[
\tilde{g}_k = (T^H T)^{-1} g_k, \quad k = 1, \ldots, N.
\]

We know that \( T \) is the generalization of \( T \) to the infinite-dimensional case and we have just seen that \( T^* \) is the generalization of \( T^H \). It is now obvious that the operator \( T^* T \) must correspond to \( T^H T \). The operator \( T^* T \) is of central importance in frame theory.

**Definition 1.3.8.** Let \( \{g_k\}_{k \in K} \) be a frame for the Hilbert space \( \mathcal{H} \). The operator \( S : \mathcal{H} \to \mathcal{H} \) defined as

\[
S = T^* T,
\]

\[
S x = \sum_{k \in K} \langle x, g_k \rangle g_k
\]

is called the frame operator.
We note that
\[ \sum_{k \in K} |\langle x, g_k \rangle|^2 = \|Tx\|^2 = \langle T^*Tx, x \rangle = \langle Sx, x \rangle. \] (1.39)

We are now able to formulate the frame condition in terms of the frame operator \(S\) by simply noting that (1.36) can be written as
\[ A\|x\|^2 \leq \langle Sx, x \rangle \leq B\|x\|^2. \] (1.40)

We shall next discuss the properties of \(S\).

**Theorem 1.3.9.** The frame operator \(S\) satisfies the following properties:

1. \(S\) is linear and bounded;
2. \(S\) is self-adjoint, i.e., \(S^* = S\);
3. \(S\) is positive definite, i.e., \(\langle Sx, x \rangle > 0\) for all \(x \in \mathcal{H}\);
4. \(S\) has a unique self-adjoint positive definite square root (denoted as \(S^{1/2}\)).

**Proof.**
1. Linearity and boundedness of \(S\) follow from the fact that \(S\) is obtained by cascading a bounded linear operator and its adjoint (see (1.38)).
2. To see that \(S\) is self-adjoint simply note that
   \[ S^* = (T^*T)^* = T^*T = S. \]
3. To see that \(S\) is positive definite note that, with (1.40)
   \[ \langle Sx, x \rangle \geq A\|x\|^2 > 0 \]
   for all \(x \in \mathcal{H}, x \neq 0\).
4. Recall the following basic fact from functional analysis [1, Th. 9.4-2].

**Lemma 1.3.10.** Every self-adjoint positive definite bounded operator \(A : \mathcal{H} \to \mathcal{H}\) has a unique self-adjoint positive definite square root, i.e., there exists a unique self-adjoint positive-definite operator \(B\) such that \(A = BB\). The operator \(B\) commutes with the operator \(A\), i.e., \(BA = AB\).

Property 4 now follows directly from Property 2, Property 3, and Lemma 1.3.10.

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 \(S\), respectively.
Theorem 1.3.11. Let $A$ and $B$ be the tightest possible frame bounds for a frame with frame operator $S$. Then

$$A = \lambda_{\text{min}} \quad \text{and} \quad B = \lambda_{\text{max}},$$

(1.41)

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

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

(1.42)

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

$$\lambda_{\text{min}} \|x\|^2 \leq \langle Sx, x \rangle \leq \lambda_{\text{max}} \|x\|^2$$

(1.43)

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

It is instructive to compare (1.43) to (1.31). Remember that $S = T^*T$ corresponds to the matrix $T^HT$ in the finite-dimensional case considered in Section 1.2.3. Thus, $\|c\|^2 = x^HT^HTx = \langle Sx, x \rangle$, 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 $\{\tilde{g}_k\}_{k=1}^N$ of the frame $\{g_k\}_{k=1}^N$ can be used to reconstruct the signal $x$ from the expansion coefficients $\{\langle x, g_k \rangle\}_{k=1}^N$ according to

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

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

$$\tilde{g}_k = (T^HT)^{-1}g_k, \quad k = 1, \ldots, N.$$ 

(1.44)

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

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

Theorem 1.3.12. The following properties hold:
1. \( S^{-1} \) is self-adjoint, i.e., \( (S^{-1})^* = S^{-1} \);

2. \( S^{-1} \) satisfies

\[
\frac{1}{B} = \inf_{x \in \mathcal{H}} \frac{\langle S^{-1}x, x \rangle}{\|x\|^2} \quad \text{and} \quad \frac{1}{A} = \sup_{x \in \mathcal{H}} \frac{\langle S^{-1}x, x \rangle}{\|x\|^2},
\]

where \( A \) and \( B \) are the tightest possible frame bounds of \( S \);

3. \( S^{-1} \) is positive definite.

Proof. 1. To prove that \( S^{-1} \) is self-adjoint we write

\[
(SS^{-1})^* = (S^{-1})^*S^* = I_{\mathcal{H}}.
\]

Since \( S \) is self-adjoint, i.e., \( S = S^* \), we conclude that

\[
(S^{-1})^*S = I_{\mathcal{H}}.
\]

Multiplying by \( S^{-1} \) from the right, we finally obtain

\[
(S^{-1})^* = S^{-1}.
\]

2. To prove the first equation in (1.45) we write

\[
B = \sup_{x \in \mathcal{H}} \frac{\langle Sx, x \rangle}{\|x\|^2} = \sup_{y \in \mathcal{H}} \frac{\langle S^{1/2}S^{-1}y, S^{1/2}S^{-1}y \rangle}{\langle S^{1/2}S^{-1}y, S^{1/2}S^{-1}y \rangle} = \sup_{y \in \mathcal{H}} \frac{\langle S^{-1/2}S^*S^{1/2}S^{-1}y, y \rangle}{\langle S^{1/2}S^{-1}y, S^{1/2}S^{-1}y \rangle} = \sup_{y \in \mathcal{H}} \frac{\langle y, y \rangle}{\langle S^{-1}y, y \rangle},
\]

where the first equality follows from (1.41) and (1.42); in the second equality we used the fact that the operator \( S^{1/2}S^{-1} \) is one-to-one on \( \mathcal{H} \) and changed variables according to \( x = S^{1/2}S^{-1}y \); in the third equality we used the fact that \( S^{1/2} \) and \( S^{-1} \) are self-adjoint, and in the fourth equality we used \( S = S^{1/2}S^{1/2} \). The first equation in (1.45) is now obtained by noting that (1.46) implies

\[
\frac{1}{B} = 1 / \left( \sup_{y \in \mathcal{H}} \frac{\langle y, y \rangle}{\langle S^{-1}y, y \rangle} \right) = \inf_{y \in \mathcal{H}} \frac{\langle S^{-1}y, y \rangle}{\langle y, y \rangle}.
\]

The second equation in (1.45) is proved analogously.

3. Positive-definiteness of \( S^{-1} \) 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 1.3.13. 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 \( \mathcal{S} \) be the corresponding frame operator. Then, the set \( \{\tilde{g}_k\}_{k \in \mathcal{K}} \) given by
\[
\tilde{g}_k = \mathcal{S}^{-1}g_k, \quad k \in \mathcal{K},
\]
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
\[
\tilde{T} : \mathcal{H} \to l^2 \quad \tilde{T} : x \to \{\langle x, \tilde{g}_k \rangle\}_{k \in \mathcal{K}}
\]
satisfies
\[
\tilde{T} = \mathcal{S}\mathcal{T}^{-1} = \mathcal{T}(\mathcal{T}^*\mathcal{T})^{-1}.
\]

Proof. Recall that \( \mathcal{S}^{-1} \) is self-adjoint. Hence, we have \( \langle x, \tilde{g}_k \rangle = \langle x, \mathcal{S}^{-1}g_k \rangle = \langle \mathcal{S}^{-1}x, g_k \rangle \) for all \( x \in \mathcal{H} \). Thus, using (1.39), we obtain
\[
\sum_{k \in \mathcal{K}} |\langle x, \tilde{g}_k \rangle|^2 = \sum_{k \in \mathcal{K}} |\langle \mathcal{S}^{-1}x, g_k \rangle|^2 = \langle \mathcal{S}(\mathcal{S}^{-1}x), \mathcal{S}^{-1}x \rangle = \langle x, \mathcal{S}^{-1}x \rangle = \langle \mathcal{S}^{-1}x, x \rangle.
\]
Therefore, we conclude from (1.45) that
\[
\frac{1}{B} \|x\|^2 \leq \sum_{k \in \mathcal{K}} |\langle x, \tilde{g}_k \rangle|^2 \leq \frac{1}{A} \|x\|^2,
\]
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 \( \tilde{T} = \mathcal{S}\mathcal{T}^{-1} \):
\[
\tilde{T}x = \{\langle x, \tilde{g}_k \rangle\}_{k \in \mathcal{K}} = \{\langle x, \mathcal{S}^{-1}g_k \rangle\}_{k \in \mathcal{K}} = \{\langle \mathcal{S}^{-1}x, g_k \rangle\}_{k \in \mathcal{K}} = \mathcal{S}\mathcal{T}^{-1}x.
\]

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 1.3.14. The frame operator associated to the canonical dual frame,
\[
\tilde{\mathcal{S}} = \tilde{T}^*\tilde{T}, \quad \tilde{\mathcal{S}}x = \sum_{k \in \mathcal{K}} \langle x, \tilde{g}_k \rangle \tilde{g}_k
\]
is called the canonical dual frame operator.

Theorem 1.3.15. The canonical dual frame operator \( \tilde{\mathcal{S}} \) satisfies \( \tilde{\mathcal{S}} = \mathcal{S}^{-1} \).
CHAPTER 1. A SHORT COURSE ON FRAME THEORY

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

\[
\tilde{S}x = \sum_{k \in K} \langle x, \tilde{g}_k \rangle \tilde{g}_k = \sum_{k \in K} \langle x, S^{-1}g_k \rangle S^{-1}g_k = \sum_{k \in K} \langle S^{-1}x, g_k \rangle g_k = SS^{-1}x = S^{-1}x,
\]

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 \(S^{-1}\) is self-adjoint, and in the fourth we used the definition of \(S\).

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

\[
\tilde{S}^{-1} \tilde{g}_k = (S^{-1})^{-1} S^{-1} g_k = SS^{-1} g_k = g_k.
\]

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

\[
x = T^* \tilde{T} x = \sum_{k \in K} \langle x, \tilde{g}_k \rangle g_k
\]

(1.50)

Note that, equivalently, we have

\[
T^* \tilde{T} = \tilde{T}^* T = I_{\mathcal{H}}.
\]

**Proof.** We have

\[
T^* \tilde{T} x = \sum_{k \in K} \langle x, \tilde{g}_k \rangle g_k = \sum_{k \in K} \langle x, S^{-1}g_k \rangle g_k = \sum_{k \in K} \langle S^{-1}x, g_k \rangle g_k = SS^{-1}x = x.
\]

This proves that \(T^* \tilde{T} = I_{\mathcal{H}}\). The proof of \(\tilde{T}^* T = I_{\mathcal{H}}\) 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 \(\langle x, g_k \rangle\), \(k \in K\), according to (1.50) is the only way of recovering \(x\) from \(\langle x, g_k \rangle\), \(k \in K\). Recall that we showed in
the finite-dimensional case (see Section 1.2.3) that for each complete and redundant set of vectors \( \{g_1, \ldots, g_N\} \), there are infinitely many dual sets \( \{\tilde{g}_1, \ldots, \tilde{g}_N\} \) that can be used to reconstruct a signal \( x \) from the coefficients \( \langle x, g_k \rangle \), \( k = 1, \ldots, N \), according to (1.26). These dual sets are obtained by identifying \( \{\tilde{g}_1, \ldots, \tilde{g}_N\} \) with the columns of \( L \), where \( L \) is a left-inverse of the analysis matrix \( T \). 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 \( T \), all linear operators \( L \) that satisfy

\[ LTx = x \]

for all \( x \in \mathcal{H} \). In other words, we want to identify all left-inverses \( L \) of the analysis operator \( T \). The answer to this question is the infinite-dimensional version of Theorem 1.2.3 that we state here without proof.

**Theorem 1.3.17.** Let \( A : \mathcal{H} \to l^2 \) be a bounded linear operator. Assume that \( A^*A : \mathcal{H} \to \mathcal{H} \) is invertible on \( \mathcal{H} \). Then, the operator \( A^\dagger : l^2 \to \mathcal{H} \) defined as \( A^\dagger \triangleq (A^*A)^{-1}A^* \) is a left-inverse of \( A \), i.e., \( A^\dagger A = I_{\mathcal{H}} \), where \( I_{\mathcal{H}} \) is the identity operator on \( \mathcal{H} \). Moreover, the general solution \( L \) of the equation \( LA = I_{\mathcal{H}} \) is given by

\[ L = A^\dagger + M(I_{l^2} - AA^\dagger) \]

where \( M : l^2 \to \mathcal{H} \) is an arbitrary bounded linear operator and \( I_{l^2} \) is the identity operator on \( l^2 \).

Applying this theorem to the operator \( T \) we see that all left-inverses of \( T \) can be written as

\[ L = T^\dagger + M(I_{l^2} - TT^\dagger) \]

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

\[ T^\dagger = (T^*T)^{-1}T^* . \]

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

\[ T^\dagger = (T^*T)^{-1}T^* = S^{-1}T^* = \tilde{T}^* . \]

This shows that reconstruction according to (1.50), i.e., by applying the operator \( \tilde{T}^* \) to the coefficient sequence \( Tx = \{\langle x, g_k \rangle\}_{k \in K} \) is nothing but applying the infinite-dimensional analog of the Moore-Penrose inverse \( T^\dagger = (T^*T)^{-1}T^* \). As already noted in the finite-dimensional case the existence of infinitely many left-inverses of the operator \( T \) 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 1.3.18.** The operator

\[ P : l^2 \to \mathcal{R}(T) \subseteq l^2 \]

defined as

\[ P = TS^{-1}T^* \]

satisfies the following properties:
1. \( P \) is the identity operator \( I_{l^2} \) on \( R(T) \).

2. \( P \) is the zero operator on \( R(T)^\perp \), where \( R(T)^\perp \) denotes the orthogonal complement of the space \( R(T) \).

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

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

\[
P \{ c_k \}_{k \in K} = TS^{-1}T^* T x = TS^{-1}S T x = T \Pi_H x = T x = \{ c_k \}_{k \in K}.
\]

This proves that \( P \) is the identity operator on \( R(T) \).

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

\[
T^* \{ c_k \}_{k \in K} = 0
\]

and therefore

\[
P \{ c_k \}_{k \in K} = TS^{-1}T^* \{ c_k \}_{k \in K} = 0.
\]

This proves that \( P \) is the zero operator on \( R(T)^\perp \).

\[ \square \]

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

\[
L = \tilde{T}^* P + M(I_{l^2} - P).
\] (1.52)

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

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

### 1.3.4 Tight Frames

The frames considered in Examples 1.3.3 and 1.3.4 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 1.3.19. A frame $\{g_k\}_{k \in K}$ 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 1.3.20. Let $\{g_k\}_{k \in K}$ be a frame for the Hilbert space $\mathcal{H}$. The frame $\{g_k\}_{k \in K}$ is tight with frame bound $A$ if and only if its corresponding frame operator satisfies $S = A \mathbb{I}_H$, or equivalently, if

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

(1.53)

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

Proof. First observe that $S = A \mathbb{I}_H$ is equivalent to $Sx = A \mathbb{I}_H x = Ax$ 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 $\{g_k\}_{k \in K}$ implies $S = A \mathbb{I}_H$, note that by Definition 1.3.19 using (1.40) we can write

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

Therefore

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

which implies $S = A \mathbb{I}_H$.

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

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

This is equivalent to

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

which shows that $\{g_k\}_{k \in K}$ 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:

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

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

Example 1.3.21 (The Mercedes-Benz frame [20]). The Mercedes-Benz frame (see Figure 1.4) is given by the following three vectors in $\mathbb{R}^2$:

$$g_1 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \quad g_2 = \begin{bmatrix} -\sqrt{3}/2 \\ -1/2 \end{bmatrix}, \quad g_3 = \begin{bmatrix} \sqrt{3}/2 \\ -1/2 \end{bmatrix}.$$ (1.54)
Figure 1.4: The Mercedes-Benz frame.

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

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

The adjoint $T^*$ of the analysis operator is given by the matrix

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

Therefore, the frame operator $S$ is represented by the matrix

$$S = T^*T = \begin{bmatrix} 0 & -\sqrt{3}/2 & \sqrt{3}/2 \\ 1 & -1/2 & -1/2 \end{bmatrix} \begin{bmatrix} 0 & 1 \\ -\sqrt{3}/2 & -1/2 \\ \sqrt{3}/2 & -1/2 \end{bmatrix} = 3/2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} = 3/2 I_2,$$

and hence $S = A I_{R^2}$ with $A = 3/2$, which implies, by Theorem 1.3.20, that $\{g_1, g_2, g_3\}$ is a tight frame (for $R^2$).

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

$$x = \sum_{k \in K} \langle x, S^{-1/2}g_k \rangle S^{-1/2}g_k, \quad \text{for all } x \in H.$$
Proof. Since $S^{-1}$ is self-adjoint and positive definite by Theorem 1.3.12, it has, by Lemma 1.3.10, a unique self-adjoint positive definite square root $S^{-1/2}$ that commutes with $S^{-1}$. Moreover $S^{-1/2}$ also commutes with $S$, which can be seen as follows:

$$S^{-1/2}S^{-1} = S^{-1}S^{-1/2}$$
$$SS^{-1/2}S^{-1} = S^{-1/2}$$
$$SS^{-1/2} = S^{-1/2}S.$$

The proof is then effected by noting the following:

$$x = S^{-1}Sx = S^{-1/2}S^{-1/2}Sx$$
$$= S^{-1/2}SS^{-1/2}x$$
$$= \sum_{k\in\mathcal{K}} \langle S^{-1/2}x, g_k \rangle S^{-1/2}g_k$$
$$= \sum_{k\in\mathcal{K}} \langle x, S^{-1/2}g_k \rangle S^{-1/2}g_k.$$

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 1.3.4. 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 1.3.23.** 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

$$\langle Sg_k, g_k \rangle = A\|g_k\|^2 = \|g_k\|^2$$

with

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

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 1.3.24 (Naimark, [24, Prop. 1.1]). Let $N > M$. Suppose that the set $\{g_1, \ldots, g_N\}$, $g_k \in \mathcal{H}$, $k = 1, \ldots, N$, 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 $\{e_1, \ldots, e_N\}$ for $\mathcal{K}$ such that $P e_k = g_k$, $k = 1, \ldots, N$, where $P : \mathcal{K} \to \mathcal{K}$ is the orthogonal projection onto $\mathcal{H}$.

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

Example 1.3.25. Consider the Hilbert space $\mathcal{K} = \mathbb{R}^3$, and assume that $\mathcal{H} \subset \mathcal{K}$ is the plane spanned by the vectors $[1 0 0]^T$ and $[0 1 0]^T$, i.e.,

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

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 1.3.21. Specifically, consider the vectors $g'_k$, $k = 1, 2, 3$, defined in (1.54) and let $g'_k \triangleq \frac{\sqrt{2}}{3} g_k$, $k = 1, 2, 3$. In the following, we think about the two-dimensional vectors $g'_k$ 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, $\{g'_k\}_{k=1}^3$ is a tight frame for $\mathcal{H}$ with frame bound $A = 1$. Now consider the following three vectors in $\mathcal{K}$:

$$e_1 = \begin{bmatrix} 0 \\ \sqrt{2/3} \\ -1/\sqrt{3} \end{bmatrix}, \quad e_2 = \begin{bmatrix} -1/\sqrt{2} \\ -1/\sqrt{6} \\ -1/\sqrt{3} \end{bmatrix}, \quad e_3 = \begin{bmatrix} 1/\sqrt{2} \\ -1/\sqrt{6} \\ -1/\sqrt{3} \end{bmatrix}.$$ 

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

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

according to $g'_k = P e_k$, $k = 1, 2, 3$. 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 \( \{e_k\}_{k=1}^M \) every signal \( x \in \mathbb{C}^M \) has a *unique* representation according to
\[
x = \sum_{k=1}^M \langle x, e_k \rangle \tilde{e}_k. \tag{1.55}
\]

The basis \( \{e_k\}_{k=1}^M \) and its dual basis \( \{\tilde{e}_k\}_{k=1}^M \) 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 1.3.26.** 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.

**Lemma 1.3.27** ([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
\[
c_k = \langle x, \tilde{g}_k \rangle = \langle x, \mathcal{S}^{-1} g_k \rangle = \langle \mathcal{S}^{-1} x, g_k \rangle = \langle \tilde{x}, g_k \rangle
\]
with \( \tilde{x} = \mathcal{S}^{-1} x \). 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
\[ \langle x, \tilde{x} \rangle = \left\langle \sum_{k \in K} a_k g_k, \tilde{x} \right\rangle = \sum_{k \in K} a_k \langle g_k, \tilde{x} \rangle = \sum_{k \in K} a_k c_k^* . \]

We can therefore conclude that
\[ \sum_{k \in K} |c_k|^2 = \sum_{k \in K} a_k c_k^* = \sum_{k \in K} a_k^* c_k . \quad (1.57) \]

Hence,
\[ \sum_{k \in K} |c_k|^2 + \sum_{k \in K} |c_k - a_k|^2 = \sum_{k \in K} |c_k|^2 + \sum_{k \in K} (c_k - a_k) (c_k^* - a_k^*) \]
\[ = \sum_{k \in K} |c_k|^2 + \sum_{k \in K} |c_k|^2 - \sum_{k \in K} c_k a_k^* - \sum_{k \in K} c_k^* a_k + \sum_{k \in K} |a_k|^2 . \]

Using (1.57), we get
\[ \sum_{k \in K} |c_k|^2 + \sum_{k \in K} |c_k - a_k|^2 = \sum_{k \in K} |a_k|^2 . \]

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

Lemma 1.3.28 ([5]). Let \( \{g_k\}_{k \in K} \) be a frame for the Hilbert space \( \mathcal{H} \) and \( \{\tilde{g}_k\}_{k \in K} \) its canonical dual frame. Then for each \( m \in K \), 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 \( g_m = \sum_{k \in K} a_k g_k \) with \( a_m = 1 \) and \( a_k = 0 \) for \( k \neq m \), so that \( \sum_{k \in K} |a_k|^2 = 1 \). Furthermore, we can write \( g_m = \sum_{k \in K} c_k g_k \) with \( c_k = \langle g_m, \tilde{g}_k \rangle \). From (1.56) it then follows that
\[ 1 = \sum_{k \in K} |a_k|^2 = \sum_{k \in K} |c_k|^2 + \sum_{k \in K} |c_k - a_k|^2 \]
\[ = \sum_{k \in K} |c_k|^2 + |c_m - a_m|^2 + \sum_{k \neq m} |c_k - a_k|^2 \]
\[ = \sum_{k \in K} |\langle g_m, \tilde{g}_k \rangle|^2 + |\langle g_m, \tilde{g}_m \rangle - 1|^2 + \sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 \]
\[ = 2 \sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 + |\langle g_m, \tilde{g}_m \rangle|^2 + 1 - |\langle g_m, \tilde{g}_m \rangle|^2 \]

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 1.3.29 ([5]).** Let \( \{g_k\}_{k \in K} \) be a frame for the Hilbert space \( \mathcal{H} \) and \( \{\tilde{g}_k\}_{k \in K} \) its canonical dual frame. Then,

1. \( \{g_k\}_{k \in K} \) is exact if and only if \( \langle g_m, \tilde{g}_m \rangle = 1 \) for all \( m \in K \);
2. \( \{g_k\}_{k \in K} \) is inexact if and only if there exists at least one \( m \in K \) 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 \( m \in K \), then \( \{g_k\}_{k \in K} \) is incomplete for \( \mathcal{H} \) (for all \( m \in K \)) and hence \( \{g_k\}_{k \in K} \) is an exact frame for \( \mathcal{H} \). Indeed, fix an arbitrary \( m \in K \). From \( \text{Lemma 1.3.28} \) 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 \( \tilde{g}_m \) is orthogonal to all elements of the set \( \{g_k\}_{k \neq m} \).

Next, we show that if there exists at least one \( m \in K \) such that \( \langle g_m, \tilde{g}_m \rangle \neq 1 \), then \( \{g_k\}_{k \in K} \) 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

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

If \( \langle g_m, \tilde{g}_m \rangle \neq 1 \), \( \text{(1.58)} \) 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

\[
|\langle x, g_m \rangle|^2 = \left| \frac{1}{1 - \langle g_m, \tilde{g}_m \rangle} \sum_{k \neq m} \langle g_m, \tilde{g}_k \rangle \langle x, g_k \rangle \right|^2 \leq \frac{1}{|1 - \langle g_m, \tilde{g}_m \rangle|^2} \left[ \sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 \right] \left[ \sum_{k \neq m} |\langle x, g_k \rangle|^2 \right].
\]
With (1.36) it follows that

\[ \sum_{k \in K} |\langle x, g_k \rangle|^2 = |\langle x, g_m \rangle|^2 + \sum_{k \neq m} |\langle x, g_k \rangle|^2 \]

\[ \leq \frac{1}{|1 - \langle g_m, \tilde{g}_m \rangle|^2} \left[ \sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 \right] \left[ \sum_{k \neq m} |\langle x, g_k \rangle|^2 \right] + \sum_{k \neq m} |\langle x, g_k \rangle|^2 \]

\[ = \frac{1}{1 - \langle g_m, \tilde{g}_m \rangle} \sum_{k \neq m} |\langle x, g_k \rangle|^2 \left[ 1 + \frac{1}{|1 - \langle g_m, \tilde{g}_m \rangle|^2} \sum_{k \neq m} |\langle g_m, \tilde{g}_k \rangle|^2 \right] \]

\[ = C \sum_{k \neq m} |\langle x, g_k \rangle|^2 \]

or equivalently

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

With (1.36) it follows that

\[ \frac{A}{C} \|x\|^2 \leq \sum_{k \in K} |\langle x, g_k \rangle|^2 \leq \sum_{k \neq m} |\langle x, g_k \rangle|^2 \leq \sum_{k \in K} |\langle x, g_k \rangle|^2 \leq B \|x\|^2, \quad (1.59) \]

where \( A \) and \( B \) are the frame bounds of the frame \( \{g_k\}_{k \in K} \). 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 \( \{\tilde{g}_k\}_{k \in K} \) 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 \( \{g_k\}_{k \in K} \) implies that \( \langle g_m, \tilde{g}_m \rangle = 1 \) for all \( m \in K \), we suppose that \( \{g_k\}_{k \in K} \) is exact and \( \langle g_m, \tilde{g}_m \rangle \neq 1 \) for at least one \( m \in K \). But the condition \( \langle g_m, \tilde{g}_m \rangle \neq 1 \) for at least one \( m \in K \) implies that \( \{g_k\}_{k \in K} \) is inexact, which results in a contradiction. It remains to show that \( \{g_k\}_{k \in K} \) inexact implies \( \langle g_m, \tilde{g}_m \rangle \neq 1 \) for at least one \( m \in K \). Suppose that \( \{g_k\}_{k \in K} \) is inexact and \( \langle g_m, \tilde{g}_m \rangle = 1 \) for all \( m \in K \). But the condition \( \langle g_m, \tilde{g}_m \rangle = 1 \) for all \( m \in K \) implies that \( \{g_k\}_{k \in K} \) is exact, which again results in a contradiction. \( \square \)

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 1.3.30 (**[5]**). Let \( \{g_k\}_{k \in K} \) be a frame for the Hilbert space \( \mathcal{H} \). If \( \{g_k\}_{k \in K} \) is exact, then \( \{g_k\}_{k \in K} \) and its canonical dual \( \{\tilde{g}_k\}_{k \in K} \) 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 \( \{g_k\}_{k \in K} \) and \( \{\tilde{g}_k\}_{k \in K} \) are biorthonormal, then \( \{g_k\}_{k \in K} \) is exact.
Proof. If \( \{g_k\}_{k \in \mathcal{K}} \) is exact, then biorthonormality follows by noting that Theorem 1.3.29 implies \( \langle g_m, \tilde{g}_m \rangle = 1 \) for all \( m \in \mathcal{K} \), and Lemma 1.3.28 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 1.3.29, 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 1.3.31 ([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. \( \square \)

### 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 \( \mathcal{L}^2 \). 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 \( \text{A/D conversion} \); the process of reconstruction of the signal \( x(t) \) from its samples is called digital-to-analog \( \text{D/A} \) conversion. We shall now formally state and prove the sampling theorem.

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

\[
\hat{x}(f) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi ft} dt.
\]

We say that \( x(t) \) is bandlimited to \( B \) Hz if \( \hat{x}(f) = 0 \) 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 \( L^2 \) functions that are bandlimited to \( B \) Hz is denoted as \( L^2(B) \).

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

\[
\hat{x}_d(f) \triangleq \sum_{k=-\infty}^{\infty} x[k] e^{-i2\pi kf} = \sum_{k=-\infty}^{\infty} x(kT) e^{-i2\pi kf} = \frac{1}{T} \sum_{k=-\infty}^{\infty} \hat{x} \left( \frac{f + k}{T} \right),
\]

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

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

\[
\hat{x}(f/T) = \hat{x}_d(f) T \hat{h}_{\text{LP}}(f)
\]

with

\[
\hat{h}_{\text{LP}}(f) = \begin{cases} 1, & |f| \leq BT \\ 0, & \text{otherwise}. \end{cases}
\]

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

\[
\hat{x}(f) = T \hat{h}_{\text{LP}}(fT) \sum_{k=-\infty}^{\infty} x[k] e^{-i2\pi kfT}.
\]

\[^5\] Strictly speaking \( \text{A/D conversion} \) also involves quantization of the samples.
\[^6\] Let \( x(t) \in L^2 \) with Fourier transform \( \hat{x}(f) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi ft} dt \). The Poisson summation formula states that \( \sum_{k=-\infty}^{\infty} x(k) = \sum_{k=-\infty}^{\infty} \hat{x}(k) \).
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:

$$x(t) = \int_{-\infty}^{\infty} \hat{x}(f)e^{i2\pi tf}df$$

$$= \int_{-\infty}^{\infty} T\hat{h}_{LP}(fT) \sum_{k=-\infty}^{\infty} x[k]e^{-i2\pi kft}e^{i2\pi ft}df$$

$$= \sum_{k=-\infty}^{\infty} x[k] \int_{-\infty}^{\infty} \hat{h}_{LP}(fT)e^{i2\pi f(t/T-k)}d(fT)$$

$$= \sum_{k=-\infty}^{\infty} x[k]h_{LP}\left(\frac{t}{T} - k\right)$$

$$= 2BT \sum_{k=-\infty}^{\infty} x[k] \text{sinc}(2B(t-kT)),$$

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

$$h_{LP}(t) = \int_{-\infty}^{\infty} \hat{h}_{LP}(f)e^{i2\pi ft}df,$$

and

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

Summarizing our findings, we obtain the following theorem.

**Theorem 1.4.1** (Sampling theorem [26, Sec. 7.2]). Let $x(t) \in L^2(B)$. 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

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

**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

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

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

$$x(kT) = \int_{-B}^{B} \hat{x}(f)e^{i2\pi kft}df = \langle \hat{x}, \hat{g_k} \rangle,$$
where
\[
\hat{g}_k(f) = \begin{cases} 
  e^{-12\pi k f T}, & |f| \leq B \\
  0, & \text{otherwise}
\end{cases}
\]
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
\[
  \|x\|^2 = \langle x, x \rangle = \left( T \sum_{k=-\infty}^{\infty} \langle x, g_k \rangle g_k(t), x \right) = T \sum_{k=-\infty}^{\infty} |\langle x, g_k \rangle|^2
\]
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 \( \mathbb{T} : \mathcal{L}^2(B) \rightarrow \ell^2 \) as
\[
  \mathbb{T} : x \rightarrow \{\langle x, g_k \rangle\}_{k \in \mathbb{Z}}, \quad \text{(1.68)}
\]
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 \( \mathbb{T}^* : \ell^2 \rightarrow \mathcal{L}^2(B) \) is to perform interpolation according to
\[
  \mathbb{T}^* : \{c_k\}_{k \in \mathbb{Z}} \rightarrow \sum_{k=-\infty}^{\infty} c_k g_k.
\]
The frame operator \( \mathbb{S} : \mathcal{L}^2(B) \rightarrow \mathcal{L}^2(B) \) is given by \( \mathbb{S} = \mathbb{T}^* \mathbb{T} \) and acts as follows
\[
  \mathbb{S} : x(t) \rightarrow \sum_{k=-\infty}^{\infty} \langle x, g_k \rangle g_k(t).
\]
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 $L^2(B)$ when $1/T < 2B$. This suggests that critical sampling $1/T = 2B$ could implement an exact frame.

Thus, in the case of critical sampling, $\|g_k\|^2 = 1$, for all $k \in \mathbb{Z}$, and Theorem 1.3.29 allows us to conclude that $\{g_k(t)\}_{k \in \mathbb{Z}}$ is an exact frame for $L^2(B)$.

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 $L^2(B)$, 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 $L^2(B)$ 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 1.3.23 allows us to conclude that $\{g_k'(t)\}_{k \in \mathbb{Z}}$ is an ONB for $L^2(B)$.

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 $L^2(B)$ in the oversampled case. Indeed, according to Theorem 1.3.29 the condition

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

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 1.3.29 (1.69) guarantees even more: for every $m \in \mathbb{Z}$, the set $\{g_k(t)\}_{k \neq m}$ is complete for $L^2(B)$. 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.
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 $\hat{x}(f)$ 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

$$x(t) = \sum_{k=-\infty}^{\infty} x[k] h\left(\frac{t}{T} - k\right),$$  \hspace{1cm} (1.70)

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

$$\hat{h}(f) = \begin{cases} 1, & |f| \leq BT \\ \text{arb}(f), & BT < |f| \leq \frac{1}{2} \\ 0, & |f| > \frac{1}{2} \end{cases}.$$  \hspace{1cm} (1.71)

Here and in what follows $\text{arb}(\cdot)$ 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 $\{g_k(t) = 2B \text{sinc}(2B(t - kT))\}_{k \in \mathbb{Z}}$. 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 $L$ of the analysis operator $T$ as discussed in Section 1.3.3. Recall the parametrization (1.52) of all left-inverses of the operator $T$:

$$L = \tilde{T}^*P + M(I_{p2} - P),$$  \hspace{1cm} (1.72)
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 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 (see Exercise 1.61), using Parseval’s theorem, 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 $[-1/2, -BT] \cup [BT, 1/2]$.

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

Let $\{a_k\}_{k \in \mathbb{Z}}, \{b_k\}_{k \in \mathbb{Z}} \in l^2$ with DTFT $\hat{a}(f) = \sum_{k=-\infty}^{\infty} a_k e^{-i2\pi kf}$ and $\hat{b}(f) = \sum_{k=-\infty}^{\infty} b_k e^{-i2\pi kf}$, respectively. Parseval’s theorem states that $\sum_{k=-\infty}^{\infty} a_k b_k = \int_{-1/2}^{1/2} |\hat{a}(f)|^2 |\hat{b}(f)|^2 df$. In particular, $\sum_{k=-\infty}^{\infty} |a_k|^2 = \int_{-1/2}^{1/2} |\hat{a}(f)|^2 df$. 

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

Figure 1.8: The reconstruction filter as a parametrized left-inverse of the analysis operator.
To see the parallels between (1.70) and (1.72), let us decompose $h(t)$ as follows (see Figure 1.8)

$$h(t) = h_{LP}(t) + h_{out}(t),$$

(1.73)

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

$$\hat{h}_{out}(f) = \begin{cases} 
arb(f), & BT \leq |f| \leq \frac{1}{2} \\
0, & \text{otherwise.} 
\end{cases}$$

(1.74)

Now it is clear, and it is left to the reader to verify formally (see Exercise ??), that the operator $A : l^2 \to L^2(B)$ defined as

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

(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 $\tilde{g}_k(t) = h_{LP}(t/T - k)$. In other words $A = \tilde{T}^*P$. Similarly, it is left to the reader to verify formally (see Exercise ??), that the operator $B : l^2 \to L^2(B)$ defined as

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

(1.76)

can be written as $B = M(I_{l^2} - P)$. Here, $(I_{l^2} - P) : l^2 \to l^2$ 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 $M : l^2 \to L^2$ is defined as

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

(1.77)

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

$$\hat{h}_M(f) = \begin{cases} 
arb_2(f), & -\frac{1}{2} \leq |f| \leq \frac{1}{2} \\
0, & \text{otherwise,} 
\end{cases}$$

(1.78)

where $arb_2(f)$ is an arbitrary bounded function that equals $arb(f)$ for $BT \leq |f| \leq \frac{1}{2}$. To summarize, we note that the operator $B$ corresponds to the second term on the right-hand side of (1.72).

We can therefore write the decomposition (1.70) as

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

$$= \underbrace{\sum_{k=-\infty}^{\infty} x[k] h_{LP} \left( \frac{t}{T} - k \right)}_{\tilde{T}^*PT_x(t)} + \underbrace{\sum_{k=-\infty}^{\infty} x[k] h_{out} \left( \frac{t}{T} - k \right)}_{M(I_{l^2} - P)T_x(t)}$$

$$= \mathbb{L}\mathbb{T}x(t).$$
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

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

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 $\hat{h}_{LP}(f)$ of bandwidth $BT$ specified in (1.63), i.e., we reconstruct using the canonical dual frame according to

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

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

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

$$\sigma_{\text{oversampling}}^2 = \mathbb{E}_w |x(t) - x'(t)|^2 = \mathbb{E}_w \left| \sum_{k=-\infty}^{\infty} w[k]h_{LP}\left(\frac{t}{T} - k\right) \right|^2 = \sigma^2 \sum_{k=-\infty}^{\infty} \left| h_{LP}\left(\frac{t}{T} - k\right) \right|^2.$$  \tag{1.80}

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

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

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

$$\sum_{k=-\infty}^{\infty} \hat{h}_{LP}(-f - k) e^{-2\pi i (f+f+k)} = \hat{h}_{LP}(-f) e^{-2\pi i (f/T)f}, \quad \text{for } f \in [-1/2, 1/2].$$  \tag{1.83}
We conclude from (1.82) and (1.83) that the DTFT of the sequence \( \{ a_k \triangleq h_{LP}(t/T - k) \} \) is given (in the fundamental interval \( f \in [-1/2, 1/2) \)) by \( \hat{h}_{LP}(-f) e^{-2\pi i (t/T)f} \) and hence we can apply Parseval’s theorem (as stated in Footnote 8) and rewrite (1.81) according to

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

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 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

\[
\sigma_{\text{oversampling}}^2 = \sigma^2 \int_{-1/2}^{1/2} \left| \hat{h}(f) \right|^2 df \tag{1.85}
\]

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

\[
\sigma_{\text{oversampling}}^2 = \sigma^2 \int_{-BT}^{0} \left| \hat{h}_{LP}(f) \right|^2 df + \sigma^2 \int_{0}^{1/2} \left| \hat{h}_{out}(f) \right|^2 df. \tag{1.86}
\]

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 \( |f| \leq BT \) picked up by the ideal lowpass filter with transfer function \( \hat{h}_{LP}(f) \). 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 \( BT \leq |f| \leq 1/2 \). 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 1.5.1.** The Weyl operator \( \mathbb{W}^{(T,F)}_{m,n} : L^2 \to L^2 \) is defined as

\[
\mathbb{W}^{(T,F)}_{m,n} : x(t) \to e^{i2\pi nFt}x(t - mT),
\]

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 \( g(t) \in L^2 \). 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

\[
g_{m,n}(t) = \mathbb{W}^{(T,F)}_{m,n}g(t) = e^{i2\pi nFt}g(t - mT), \quad m \in \mathbb{Z}, \ n \in \mathbb{Z}.
\]

The set \( \{g_{m,n}(t) = e^{i2\pi nFt}g(t - mT)\}_{m \in \mathbb{Z}, n \in \mathbb{Z}} \) 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 \( L^2 \), it is called a WH frame for \( L^2 \).

Whether or not a WH set \((g, T, F)\) is a frame for \( L^2 \) 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 \( L^2 \). 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 \( L^2 \). Indeed, this intuition is correct and the following fundamental result formalizes it:

**Theorem 1.5.2** ([21] Thm. 8.3.1). Let \( g(t) \in L^2 \) and \( T, F > 0 \) be given. Then the following holds:

- If \( TF > 1 \), then \((g, T, F)\) is not a frame for \( L^2 \).
- If \((g, T, F)\) is a frame for \( L^2 \), 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 \(L^2\) 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 \(L^2\) 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 1.5.3** (Gaussian, [21, Thm. 8.6.1]). Let \(T, F > 0\) and take \(g(t) = e^{-t^2}\). Then the WH set
\[
\{ \mathbb{W}^{(T,F)}_{m,n} g(t) \}_{m \in \mathbb{Z}, n \in \mathbb{Z}}
\]
is a frame for \(L^2\) 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 1.5.4.** The operator \(\mathbb{V}^{(T,S)}_{m,n} : L^2 \to L^2\) is defined as
\[
\mathbb{V}^{(T,S)}_{m,n} : x(t) \mapsto S^{n/2} x(S^n t - mT),
\]
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) \(g(t) \in L^2\). Fix the parameters \(T > 0\) and \(S > 0\) and consider the set of functions
\[
g_{m,n}(t) \triangleq \mathbb{V}^{(T,S)}_{m,n} g(t) = S^{n/2} g(S^n t - mT), \quad m \in \mathbb{Z}, \ n \in \mathbb{Z}.
\]
This set is referred to as a wavelet set. When the wavelet set \(\{g_{m,n}(t) = S^{n/2} g(S^n t - mT)\}_{m \in \mathbb{Z}, n \in \mathbb{Z}}\) with parameters \(T, S > 0\) is a frame for \(L^2\), 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 \(L^2\) 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 1.5.5** (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
\[
\{ \mathbb{V}^{(T,S)}_{m,n} g(t) \}_{m \in \mathbb{Z}, n \in \mathbb{Z}}
\]
is a frame for \(L^2\) [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 \(A \subseteq \{1, \ldots, m\}\), \(D_A\) denotes the \(m \times m\) diagonal matrix with diagonal entries \((D_A)_{i,i} = 1\) for \(i \in A\), and \((D_A)_{i,i} = 0\) else. With \(U \in \mathbb{C}^{m \times m}\) unitary and \(A \subseteq \{1, \ldots, m\}\), we define the orthogonal projection \(P_A(U) = UD_AU^H\) and set \(W_{U^i}^A = R(P_A(U))\). For \(x \in \mathbb{C}^m\) and \(A \subseteq \{1, \ldots, m\}\), we let \(x_A = D_Ax\). With \(A \in \mathbb{C}^{m \times m}\), \(\|A\|_1 = \max_{x: \|x\|_1=1} \|Ax\|_1\) refers to the operator 1-norm, \(\|A\|_2 = \max_{x: \|x\|_2=1} \|Ax\|_2\) designates the operator 2-norm, \(\|A\|_F = \sqrt{\text{tr}(AA^H)}\) is the Frobenius norm, and \(\|A\|_1 = \sum_{i,j=1}^m |A_{i,j}|\). The vector \(x \in \mathbb{C}^m\) 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 \cite{DoSt92} 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” \(D_P\) as an uncertainty relation. More specifically, let \(U \in \mathbb{C}^{m \times m}\) be a unitary matrix, \(P, Q \subseteq \{1, \ldots, m\}\), and consider the orthogonal projection \(P_Q(U)\) onto the subspace \(W_{U^i,Q}\) which is spanned by \(\{u_i : i \in Q\}\). Let\(^1\) \(\Delta_{P,Q}(U) = \|D_PP_Q(U)\|_2\). In the setting of \cite{DoSt92} \(U\) would correspond to the DFT matrix \(F\) and \(\Delta_{P,Q}(F)\) is the operator 2-norm of the band-limitation operator followed by the

---

\(^1\)We note that, for general unitary \(A, B \in \mathbb{C}^{m \times m}\), unitary invariance of \(\| \cdot \|_2\) yields \(\|P_P(A)P_Q(B)\|_2 = \|D_PP_Q(U)\|_2\) with \(U = A^H B\). 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 10 we have
\[
\Delta_{P,Q}(U) = \max_{x \in W^{U,Q} \setminus \{0\}} \frac{\|x_P\|_2}{\|x\|_2}.
\] (2.1)
An uncertainty relation in \((\mathbb{C}^m, \| \cdot \|_2)\) is an upper bound of the form
\[
\Delta_{P,Q}(U) \leq c
\]
with \(c \geq 0\), and states that \(\|x_P\|_2 \leq c\|x\|_2\) for all \(x \in W^{U,Q}\). \(\Delta_{P,Q}(U)\) hence quantifies how well a vector supported on \(Q\) in the basis \(U\) can be concentrated on \(P\). Note that an uncertainty relation in \((\mathbb{C}^m, \| \cdot \|_2)\) is nontrivial only if \(c < 1\). Application of Lemma 11 now yields
\[
\frac{\|DP_PQ(U)\|_2}{\sqrt{\text{rank}(DP_PQ(U))}} \leq \Delta_{P,Q}(U) \leq \|DP_PQ(U)\|_2,
\] (2.2)
where the upper bound constitutes an uncertainty relation and the lower bound will allow us to assess its tightness. Next, note that
\[
\|DP_PQ(U)\|_2 = \sqrt{\text{tr}(DP_PQ(U))}
\] (2.3)
and
\[
\text{rank}(DP_PQ(U)) = \text{rank}(DP_UD_QU^H) 
\leq \min(|P|, |Q|),
\] (2.4)
where (2.5) follows from \(\text{rank}(DP_UD_Q) \leq \min(|P|, |Q|)\) and [48, Property (c), Chapter 0.4.5]. When used in (2.2) this implies
\[
\sqrt{\text{tr}(DP_PQ(U))} \leq \Delta_{P,Q}(U) \leq \sqrt{\text{tr}(DP_PQ(U))}.
\] (2.6)

Particularizing to \(U = F\), we obtain
\[
\sqrt{\text{tr}(DP_PQ(F))} = \sqrt{\text{tr}(DP_FD_QF^H)} = \sqrt{\sum_{i \in P} \sum_{j \in Q} |F_{i,j}|^2} = \sqrt{\frac{|P||Q|}{m}},
\] (2.7)
so that (2.6) reduces to
\[
\sqrt{\frac{\max(|P|, |Q|)}{m}} \leq \Delta_{P,Q}(F) \leq \sqrt{\frac{|P||Q|}{m}}.
\] (2.10)
There exist sets \(P, Q \subseteq \{1, \ldots, m\}\) that saturate both bounds in (2.10), e.g., \(P = \{1\}\) and \(Q = \{1, \ldots, m\}\), which yields \(\sqrt{\frac{\max(|P|, |Q|)}{m}} = \sqrt{\frac{|P||Q|}{m}} = 1\) and therefore \(\Delta_{P,Q}(F) = 1\).
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

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

(2.11)

and

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

(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

$$
\sqrt{|\mathcal{P}| |\mathcal{Q}|} = \frac{n}{\sqrt{m}}.
$$

(2.13)

whereas the lower bound becomes

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

(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 $\Delta_{\mathcal{P}, \mathcal{Q}}(F) \leq \sqrt{|\mathcal{P}| |\mathcal{Q}|} / m$ 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 1.** [32, Theorem 11] Let $n$ divide $m$ and consider

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

(2.15)

and

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

(2.16)

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

**Proof.** We have

$$
\Delta_{\mathcal{P}, \mathcal{Q}}(F) = \|P_{\mathcal{Q}}(F)D_{\mathcal{P}}\|_2
$$

(2.17)

$$
\quad = \|D_{\mathcal{Q}H}D_{\mathcal{P}}\|_2
$$

(2.18)

$$
\quad = \max_{x: \|x\|_2 = 1} \|D_{\mathcal{Q}H}D_{\mathcal{P}}x\|_2
$$

(2.19)

$$
\quad = \max_{x: x \neq 0} \frac{\|D_{\mathcal{Q}H}x_{\mathcal{P}}\|_2}{\|x\|_2}
$$

(2.20)

$$
\quad = \max_{x: x = x_{\mathcal{P}}, x \neq 0} \frac{\|D_{\mathcal{Q}H}x\|_2}{\|x\|_2},
$$

(2.21)
where in (2.17) we applied Lemma 10 and in (2.18) we used unitary invariance of \( \| \cdot \|_2 \). Next, consider an arbitrary but fixed \( x \in \mathbb{C}^m \) with \( x = x_P \) and define \( y \in \mathbb{C}^n \) according to \( y_s = x_{ms/n} \) for \( s = 1, \ldots, n \). It follows that

\[
\| D_Q F^H x \|_2^2 = \frac{1}{m} \sum_{q \in Q} \left| \sum_{p \in P} x_p e^{2\pi i q / m} \right|^2 \quad (2.22)
\]

\[
= \frac{1}{m} \sum_{q \in Q} \left| \sum_{s=1}^n x_{ms/n} e^{2\pi i q / n} \right|^2 \quad (2.23)
\]

\[
= \frac{1}{m} \sum_{q \in Q} \left| \sum_{s=1}^n y_s e^{2\pi i q / n} \right|^2 \quad (2.24)
\]

\[
= \frac{n}{m} \| F^H y \|_2^2 \quad (2.25)
\]

\[
= \frac{n}{m} \| y \|_2^2, \quad (2.26)
\]

where \( F \) 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 \( \| x \|_2 = \| y \|_2 \) in (2.21), we get \( \Delta_{P,Q}(F) = \sqrt{n/m} \). 

### 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 \( U = F \) a more sophisticated upper bound on \( \Delta_{P,Q}(F) \) 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 2.2.1.** Let \( P \subseteq \{1, \ldots, m\} \), \( l, n \in \{1, \ldots, m\} \), and

\[
Q = \{l + 1, \ldots, l + n\} \quad (2.27)
\]

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

\[
\rho(P, \lambda) = \frac{1}{\lambda} \max_{r \in [0, m)} |\bar{P} \cap (r, r + \lambda)|, \quad (2.28)
\]

where \( \bar{P} = P \cup \{m + p : p \in P\} \). Then,

\[
\Delta_{P,Q}(F) \leq \sqrt{\left( \frac{\lambda(n - 1)}{m} + 1 \right) \rho(P, \lambda)} \quad (2.29)
\]

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

$$|x_p|^2 \leq |(Fa)_p|^2$$  \hfill (2.30)

$$= \frac{1}{m} \left| \sum_{q \in \mathcal{Q}} a_q e^{-\frac{2\pi iq}{m}} \right|^2$$  \hfill (2.31)

$$= \frac{1}{m} \left| \sum_{k=1}^{n} a_k e^{-\frac{2\pi ik}{m}} \right|^2$$  \hfill (2.32)

$$= \frac{1}{m} \left| \psi\left(\frac{p}{m}\right) \right|^2 \quad \text{for} \quad p \in \{1, \ldots, m\},$$  \hfill (2.33)

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

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

Next, let $\nu_t$ 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,

$$\|x_p\|_2^2 = \frac{1}{m} \sum_{p \in \mathcal{P}} \left| \psi\left(\frac{p}{m}\right) \right|^2$$  \hfill (2.35)

$$= \frac{1}{m} \int_{[0,1)} |\psi(s)|^2 d\mu(s)$$  \hfill (2.36)

$$\leq \left( \frac{n - 1}{m} + \frac{1}{\lambda} \right) \sup_{r \in [0,1)} \mu\left(\left( r, r + \frac{\lambda}{m} \right) \right)$$  \hfill (2.37)

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 9 with $\delta = \lambda/m$ and $\|a\|_2 = 1$. Now,

$$\sup_{r \in [0,1)} \mu\left(\left( r, r + \frac{\lambda}{m} \right) \right) = \sup_{r \in [0, m)} \sum_{p \in \mathcal{P}} \left( \nu_p((r, r + \lambda)) + \nu_{m+p}((r, r + \lambda)) \right)$$  \hfill (2.38)

$$= \max_{r \in [0, m)} |\tilde{\mathcal{P}} \cap (r, r + \lambda)|$$  \hfill (2.39)

$$= \lambda \rho(\mathcal{P}, \lambda) \quad \text{for all} \quad \lambda \in (0, m],$$  \hfill (2.40)

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

$$\|x_p\|_2^2 \leq \left( \frac{\lambda(n - 1)}{m} + 1 \right) \rho(\mathcal{P}, \lambda) \quad \text{for all} \quad \lambda \in (0, m].$$  \hfill (2.42)
As \( x \in \mathcal{W}^F \) with \( \|x\|_2 = 1 \) was arbitrary, we conclude that

\[
\Delta^2_{P,Q}(F) = \max_{x \in \mathcal{W}^F \setminus \{0\}} \frac{\|x_P\|^2}{\|x\|^2} \leq \left( \frac{\lambda(n-1)}{m} + 1 \right) \rho(P, \lambda) \quad \text{for all } \lambda \in (0, m],
\]

thereby finishing the proof.

Theorem 2.2.1 slightly improves upon [33, Theorem 12] by virtue of applying to more general sets \( Q \) and defining the circular Nyquist density in (2.28) in terms of open intervals \((r, r + \lambda)\).

We next apply Theorem 2.2.1 to specific choices of \( P \) and \( Q \). First, consider \( P = \{1\} \) and \( Q = \{1, \ldots, m\} \), which were shown to saturate the upper and the lower bound in (2.10) leading to \( \Delta_{P,Q}(F) = 1 \). Since \( P \) consists of a single point, \( \rho(P, \lambda) = 1/\lambda \) for all \( \lambda \in (0, m] \). Thus, Theorem 2.2.1 with \( n = m \) yields

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

(2.45)

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

Next, consider \( P \) and \( Q \) 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 \( P \) consists of points spaced \( m/n \) apart, we get \( \rho(P, \lambda) = 1/\lambda \) for all \( \lambda \in (0, m/n] \). The upper bound (2.29) now becomes

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

(2.46)

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

\[
\Delta_{P,Q}(F) \leq \sqrt{(2n-1)/m} \leq \sqrt{2}\sqrt{n/m},
\]

(2.47)

which is tight up to a factor of \( \sqrt{2} \) (cf. Lemma 1). We hasten to add, however, that the large sieve technique applies to \( U = F \) only.

### 2.2.2 Coherence-based Uncertainty Relation

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

**Definition 2.2.2.** For \( A = (a_1 \ldots a_n) \in \mathbb{C}^{m \times n} \) with columns \( \| \cdot \|_2 \)-normalized to 1, the coherence is defined as \( \mu(A) = \max_{i \neq j} |a_i^H a_j| \).

We have the following coherence-based uncertainty relation valid for general unitary \( U \).
Lemma 2. Let $U \in \mathbb{C}^{m \times m}$ be unitary and $\mathcal{P}, \mathcal{Q} \subseteq \{1, \ldots, m\}$. Then,
\[
\Delta_{\mathcal{P}, \mathcal{Q}}(U) \leq \sqrt{\mathcal{P}|\mathcal{Q}| \mu(I \, \!U)}.
\] (2.48)

Proof. The claim follows from
\[
\Delta_{\mathcal{P}, \mathcal{Q}}^2(U) \leq \text{tr}(D_{\mathcal{P}}UD_{\mathcal{Q}}U^H) \quad (2.49)
\]
\[
= \sum_{k \in \mathcal{P}} \sum_{l \in \mathcal{Q}} |U_{k,l}|^2 \quad (2.50)
\]
\[
\leq |\mathcal{P}| |\mathcal{Q}| \max_{k,l} |U_{k,l}|^2 \quad (2.51)
\]
\[
= |\mathcal{P}| |\mathcal{Q}| \mu^2(I \, \!U), \quad (2.52)
\]
where (2.49) is by (2.6) and in (2.52) we used the definition of coherence. □

Since $\mu(I \, \!F) = 1/\sqrt{m}$, Lemma 2 particularized to $U = F$ 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 $p$ and $q$ in $p = Fq$. We will actually study the more general case obtained by replacing $I$ and $F$ by unitary $A \in \mathbb{C}^{m \times m}$ and $B \in \mathbb{C}^{m \times m}$, respectively, and will ask ourselves how well concentrated $p$ and $q$ in $Ap = Bq$ can be. Rewriting $Ap = Bq$ according to $p = Uq$ with $U = A^H B$, we now show how the uncertainty relation in Lemma 2 can be used to answer this question. Let us start by introducing a measure for concentration in $(\mathbb{C}^m, \| \cdot \|_2)$.

Definition 2.2.3. Let $\mathcal{P} \subseteq \{1, \ldots, m\}$ and $\varepsilon_\mathcal{P} \in [0, 1]$. The vector $x \in \mathbb{C}^m$ is said to be $\varepsilon_\mathcal{P}$-concentrated if $\|x - x_\mathcal{P}\|_2 \leq \varepsilon_\mathcal{P}\|x\|_2$.

The fraction of 2-norm of $\varepsilon_\mathcal{P}$-concentrated vector exhibits outside $\mathcal{P}$ is therefore no more than $\varepsilon_\mathcal{P}$. In particular, if $x$ is $\varepsilon_\mathcal{P}$-concentrated with $\varepsilon_\mathcal{P} = 0$, then $x = x_\mathcal{P}$ and $x$ 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 $\Delta_{\mathcal{P}, \mathcal{Q}}(U)$ for unitary matrices $U$ that relate $\varepsilon_\mathcal{P}$-concentrated vectors $p$ to $\varepsilon_\mathcal{Q}$-concentrated vectors $q$ through $p = Uq$. The formal statement is as follows.

Lemma 3. Let $U \in \mathbb{C}^{m \times m}$ be unitary and $\mathcal{P}, \mathcal{Q} \subseteq \{1, \ldots, m\}$. Suppose that there exist a nonzero $\varepsilon_\mathcal{P}$-concentrated $p \in \mathbb{C}^m$ and a nonzero $\varepsilon_\mathcal{Q}$-concentrated $q \in \mathbb{C}^m$ such that $p = Uq$. Then,
\[
\Delta_{\mathcal{P}, \mathcal{Q}}(U) \geq [1 - \varepsilon_\mathcal{P} - \varepsilon_\mathcal{Q}]_+.
\] (2.53)
Proof. We have
\[
\|p - P_Q(U)p_P\|_2 \leq \|p - P_Q(U)p\|_2 + \|P_Q(U)p_P - P_Q(U)p\|_2 \tag{2.54}
\]
\[
\leq \|p - P_Q(U)p\|_2 + \|P_Q(U)\|_2\|p_P - p\|_2 \tag{2.55}
\]
\[
\leq \|p - UD_QU^hp\|_2 + \|p_P - p\|_2 \tag{2.56}
\]
\[
= \|q - q_Q\|_2 + \|p_P - p\|_2, \tag{2.57}
\]
\[
\leq \|q - q_Q\|_2 + \|p_P - p\|_2, \tag{2.58}
\]
\[
= \|q - q_Q\|_2 + \|p_P - p\|_2, \tag{2.59}
\]
where in (2.57) we made use of the unitary invariance of \(\|\cdot\|_2\). It follows that
\[
\|P_Q(U)p_P\|_2 \geq \|p\|_2 - \|p - P_Q(U)p_P\|_2 \tag{2.60}
\]
\[
\geq \|p\|_2[1 - \varepsilon_P - \varepsilon_Q], \tag{2.61}
\]
where (2.60) is by the reverse triangle inequality and in (2.61) we used (2.54)–(2.59). Since \(p \neq 0\) by assumption, (2.60)–(2.61) implies
\[
\|P_Q(U)p_P\|_2 \geq [1 - \varepsilon_P - \varepsilon_Q], \tag{2.62}
\]
which in turn yields \(\|P_Q(U)p_P\|_2 \geq [1 - \varepsilon_P - \varepsilon_Q]\). This concludes the proof as \(\Delta_{P,Q}(U) = \|P_Q(U)p_P\|_2\) by Lemma 10.

Combining Lemma 3 with the uncertainty relation Lemma 2 yields the announced result stating that a nonzero vector can not be arbitrarily well concentrated with respect to two different orthonormal bases.

**Corollary 2.2.4.** Let \(A, B \in \mathbb{C}^{m \times m}\) be unitary and \(P, Q \subseteq \{1, \ldots, m\}\). Suppose that there exist a nonzero \(\varepsilon_P\)-concentrated \(p \in \mathbb{C}^m\) and a nonzero \(\varepsilon_Q\)-concentrated \(q \in \mathbb{C}^m\) such that \(Ap = Bq\). Then,
\[
|P|Q| \geq \frac{[1 - \varepsilon_P - \varepsilon_Q]^2}{\mu^2([A B])}. \tag{2.63}
\]

**Proof.** Let \(U = A^HB\). Then, by Lemmata 2 and 3 we have
\[
[1 - \varepsilon_P - \varepsilon_Q] \leq \Delta_{P,Q}(U) \leq \sqrt{|P|Q|} \mu([I U]). \tag{2.64}
\]
The claim now follows by noting that \(\mu([I U]) = \mu([A B])\).

For \(\varepsilon_P = \varepsilon_Q = 0\), we recover the well-known Elad-Bruckstein result.

**Corollary 2.2.5.** \([34] \text{Theorem 1}\) Let \(A, B \in \mathbb{C}^{m \times m}\) be unitary. If \(Ap = Bq\) for nonzero \(p, q \in \mathbb{C}^m\), then \(\|p\|_0\|q\|_0 \geq 1/\mu^2([A B])\).
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 2.2.4 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 \(p, n \in \mathbb{C}^m\) and \(\mathcal{P} \subseteq \{1, \ldots, m\}\), set \(\mathcal{P}^c = \{1, \ldots, m\} \setminus \mathcal{P}\), and suppose that we observe \(y = p_{\mathcal{P}^c} + n\). Note that the information contained in \(p_{\mathcal{P}}\) is completely lost in the observation. Without structural assumptions on \(p\), it is therefore not possible to recover information on \(p_{\mathcal{P}}\) from \(y\). However, if \(p\) is sufficiently sparse with respect to an orthonormal basis and \(|\mathcal{P}|\) is sufficiently small, it turns out that all entries of \(p\) 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 4.** Let \(U \in \mathbb{C}^{m \times m}\) be unitary, \(\mathcal{Q} \subseteq \{1, \ldots, m\}\), \(p \in \mathcal{W}^{U, \mathcal{Q}}\), and consider

\[
y = p_{\mathcal{P}^c} + n,
\]

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

\[
\|Ly - p\|_2 \leq C\|n_{\mathcal{P}^c}\|_2
\]

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

\[
|\mathcal{P}|\|\mathcal{Q}| < \frac{1}{\mu^2([I, U])}
\]

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

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

\[
\|\|(I - D_{\mathcal{P}}P_{\mathcal{Q}}(U))^{-1}\|\|_2 \leq \frac{1}{1 - \|D_{\mathcal{P}}P_{\mathcal{Q}}(U)\|_2}
\]

\[
= \frac{1}{1 - \Delta_{\mathcal{P}, \mathcal{Q}}(U)}.
\]

We now set \(L = (I - D_{\mathcal{P}}P_{\mathcal{Q}}(U))^{-1}D_{\mathcal{P}^c}\) and note that

\[
Lp_{\mathcal{P}^c} = (I - D_{\mathcal{P}}P_{\mathcal{Q}}(U))^{-1}p_{\mathcal{P}^c}
\]

\[
= (I - D_{\mathcal{P}}P_{\mathcal{Q}}(U))^{-1}(I - D_{\mathcal{P}})p
\]

\[
= (I - D_{\mathcal{P}}P_{\mathcal{Q}}(U))^{-1}(I - D_{\mathcal{P}}P_{\mathcal{Q}}(U))p
\]

\[
= p,
\]
where in (2.72) we used $P_Q(U)p = p$, which is by assumption. Next, we upper-bound $\|L_y - p\|_2$ according to

\[
\|L_y - p\|_2 = \|L_{p^c} + L_n - p\|_2 = \|L_n\|_2 \tag{2.74}
\]

\[
\leq \|(I - D_PP_Q(U))^{-1}\|_2 \|n_{p^c}\|_2 \tag{2.75}
\]

\[
\leq \frac{1}{1 - \Delta_{P,Q}(U)}\|n_{p^c}\|_2, \tag{2.76}
\]

where in (2.75) we used (2.70)–(2.73). Finally, Lemma 2 implies that (2.67) is sufficient for $\Delta_{P,Q}(U) < 1$.

Note that in the noise free case $\Delta_{P,Q}(U) < 1$ is a sufficient condition for perfect recovery of $p$ in Lemma 4. We next particularize Lemma 4 for $U = F$,

\[
\mathcal{P} = \left\{ \frac{m}{n}, \frac{2m}{n}, \ldots, \frac{(n-1)m}{n}, m \right\} \tag{2.78}
\]

with $n$ dividing $m$, and

\[
Q = \{l + 1, \ldots, l + n\} \tag{2.79}
\]

with $l \in \{1, \ldots, m\}$ and $Q$ interpreted circularly in $\{1, \ldots, m\}$. This means that $p$ is $n$-sparse in $F$ and we are missing $n$ entries in the noisy observation $y$. From Lemma 10 we know that $\Delta_{P,Q}(F) = \sqrt{n/m}$. Since $n$ divides $m$ by assumption, stable recovery of $p$ is possible for $n \leq m/2$. In contrast, the coherence-basedcoherence uncertainty relation in Lemma 2 yields $\Delta_{P,Q}(F) \leq \frac{n}{\sqrt{m}}$, 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 $U = (u_1 \ldots u_m) \in \mathbb{C}^{m \times m}$ be a unitary matrix, $\mathcal{P}, Q \subseteq \{1, \ldots, m\}$, and consider the orthogonal projection $P_Q(U)$ onto the subspace $\mathcal{W}^U \subseteq \mathbb{C}^m$, which is spanned by $\{u_i : i \in Q\}$. Let

\[
\Sigma_{\mathcal{P},Q}(U) = \|D_{\mathcal{P}}P_Q(U)\|_1. \tag{2.80}
\]

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

\footnote{In contrast to the operator 2-norm, the operator 1-norm is not invariant under unitary transformations so that we do not have $\|P_{\mathcal{P}}(A)P_Q(B)\|_1 \neq \|D_{\mathcal{P}}P_Q(A^H B)\|_1$ for general unitary $A, B$. This, however, does not constitute a problem as whenever we apply uncertainty relations in $(\mathbb{C}^m, \| \cdot \|_1)$, the case of general unitary $A, B$ can always be reduced directly to $P_{\mathcal{P}}(I) = D_{\mathcal{P}}$ and $P_Q(A^H B)$, simply by rewriting $Ap = Bq$ according to $p = A^H Bq$.}
An uncertainty relation in \((\mathbb{C}^m, \| \cdot \|_1)\) is an upper bound of the form \(\Sigma_{\mathcal{P}, \mathcal{Q}}(U) \leq c\) with \(c \geq 0\) and states that \(\|x_U\|_1 \leq c\|x\|_1\) for all \(x \in \mathcal{W}^{U, \mathcal{Q}}\). \(\Sigma_{\mathcal{P}, \mathcal{Q}}(U)\) hence quantifies how well a vector supported on \(\mathcal{Q}\) in the basis \(U\) 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 12 yields

\[
\frac{1}{m} \|D_{\mathcal{P}} P_{\mathcal{Q}}(U)\|_1 \leq \Sigma_{\mathcal{P}, \mathcal{Q}}(U) \leq \|D_{\mathcal{P}} P_{\mathcal{Q}}(U)\|_1,
\]

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 5.** Let \(U \in \mathbb{C}^{m \times m}\) be a unitary matrix and \(\mathcal{P}, \mathcal{Q} \subseteq \{1, \ldots, m\}\). Then,

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

**Proof.** Let \(\tilde{u}_i\) denote the column vectors of \(U^H\). It follows from Lemma 12 that

\[
\Sigma_{\mathcal{P}, \mathcal{Q}}(U) = \max_{j \in \{1, \ldots, m\}} \|D_{\mathcal{P}} U D_{\mathcal{Q}} \tilde{u}_j\|_1.
\]  

With

\[
\max_{j \in \{1, \ldots, m\}} \|D_{\mathcal{P}} U D_{\mathcal{Q}} \tilde{u}_j\|_1 \leq |\mathcal{P}| \max_{i,j \in \{1, \ldots, m\}} |\tilde{u}_i^H D_{\mathcal{Q}} \tilde{u}_j|
\]

\[
\leq |\mathcal{P}| |\mathcal{Q}| \max_{i,j,k \in \{1, \ldots, m\}} |U_{i,k}| |U_{j,k}|
\]

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

this establishes the proof. \(\square\)

For \(\mathcal{P} = \{1\}, \mathcal{Q} = \{1, \ldots, m\}\), and \(U = F\), the upper bounds on \(\Sigma_{\mathcal{P}, \mathcal{Q}}(F)\) 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, $\mathcal{P} = \{m\}$, $\mathcal{Q} = \{1, \ldots, m/2\}$, and $U = F$. Then, (2.82) becomes $\Sigma_{\mathcal{P}, \mathcal{Q}}(F) \leq 1/2$, whereas

$$||D_\mathcal{P}P_{\mathcal{Q}}(F)||_1 = \frac{1}{m} \sum_{l=1}^{m/2} \left| \sum_{k=1}^{m/2} e^{2\pi i l k/m} \right|$$

(2.87)

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

(2.88)

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

(2.89)

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

(2.90)

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

For $\mathcal{P}$ and $\mathcal{Q}$ as in (2.11) and (2.12), respectively, (2.82) becomes $\Sigma_{\mathcal{P}, \mathcal{Q}}(F) \leq n^2/m$, which for fixed ratio $n/m$ increases linearly in $m$ and becomes trivial for $m \geq (m/n)^2$. A more sophisticated uncertainty relation based on a large sieve inequality exists for strictly band-limited (infinite) $\ell_1$-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 $\ell_1$-norm.

**Definition 2.3.1.** Let $\mathcal{P} \subseteq \{1, \ldots, m\}$ and $\varepsilon_\mathcal{P} \in [0, 1]$. The vector $x \in \mathbb{C}^m$ is said to be $\varepsilon_\mathcal{P}$-concentrated if $||x - x_\mathcal{P}||_1 \leq \varepsilon_\mathcal{P} ||x||_1$.

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 $x$ is $\varepsilon_\mathcal{P}$-concentrated for $\varepsilon_\mathcal{P} = 0$, then $x = x_\mathcal{P}$ and $x$ is $|\mathcal{P}|$-sparse. The zero vector is trivially $\varepsilon_\mathcal{P}$-concentrated for all $\mathcal{P} \subseteq \{1, \ldots, m\}$ and $\varepsilon_\mathcal{P} \in [0, 1]$. In the remainder of Section 2.3 concentration is with respect to the $1$-norm according to Definition 2.3.1.

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

**Lemma 6.** Let $A, B \in \mathbb{C}^{m \times m}$ be unitary and $\mathcal{P}, \mathcal{Q} \subseteq \{1, \ldots, m\}$. Suppose that there exist a nonzero $\varepsilon_\mathcal{P}$-concentrated $p \in \mathbb{C}^m$ and a nonzero $q \in \mathbb{C}^m$ with $q = q_\mathcal{Q}$ such that $Ap = Bq$. Then,

$$||\mathcal{P}||\mathcal{Q} \geq \frac{1 - \varepsilon_\mathcal{P}}{\mu^2(||A||B||)}.$$  

(2.91)
Proof. Rewriting \( A p = B q \) according to \( p = A^H B q \), it follows that \( p \in \mathcal{W}^U, Q \) with \( U = A^H B \). We have

\[
1 - \varepsilon_p \leq \frac{\| p_P \|_1}{\| p \|_1} \leq \Sigma_{P, Q}(U) \leq |P| |Q| \mu^2([I \ U]),
\]

where (2.92) is by \( \varepsilon_p \)-concentration of \( p \), (2.93) follows from (2.80) and \( p \in \mathcal{W}^U, Q \), and in (2.94) we applied Lemma 5. The proof is concluded by noting that \( \mu([I \ U]) = \mu([A \ B]) \).

For \( \varepsilon_p = 0 \), Lemma 6 recovers Corollary 2.2.5.

### 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 \( \varepsilon_p \)-concentrated noise.

**Lemma 7.** Let

\[
y = p + n,
\]

where \( n \in \mathbb{C}^m \) is \( \varepsilon_p \)-concentrated to \( P \subseteq \{1, \ldots, m\} \) and \( p \in \mathcal{W}^U, Q \) for \( U \in \mathbb{C}^{m \times m} \) unitary and \( Q \subseteq \{1, \ldots, m\} \). Denote

\[
z = \operatorname{argmin}_{w \in \mathcal{W}^U, Q} (\| y - w \|_1).
\]

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

\[
|P||Q| < \frac{1}{2 \mu^2([I \ U])},
\]

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

**Proof.** Set \( P^c = \{1, \ldots, m\} \setminus P \) and let \( q = U^H p \). Note that \( q_Q = q \) as a consequence of \( p \in \mathcal{W}^U, Q \), which is by assumption. We have

\[
\| n \|_1 = \| y - p \|_1 \geq \| y - z \|_1 \geq \| n - \tilde{z} \|_1 = \| (n - \tilde{z})_P \|_1 + \| (n - \tilde{z})_{P^c} \|_1 \geq \| n_P \|_1 - \| n_{P^c} \|_1 + \| \tilde{z}_{P^c} \|_1 \end{align}
\]

\[
= \| n \|_1 - 2\| n_{P^c} \|_1 + \| \tilde{z} \|_1 - 2\| \tilde{z}_P \|_1 \geq \| n \|_1 \left(1 - 2\varepsilon_P \right) + \| \tilde{z} \|_1 \left(1 - 2 \Sigma_{P, Q}(U) \right),
\]

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

\[
\|z - p\|_1 = \|\tilde{z}\|_1 \leq \frac{2\varepsilon_p}{1 - 2\Sigma_{P,Q}(U)}\|n\|_1.
\]

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

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

\[
|P||Q| < \frac{m}{2},
\]

and perfect recovery of \( p \) from \( y = p + n \) 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 \( s = Ap = Bq \), where \( A \in \mathbb{C}^{m \times p} \) and \( B \in \mathbb{C}^{m \times q} \) are general matrices, \( p \in \mathbb{C}^p \), and \( q \in \mathbb{C}^q \). We start by introducing the notion of mutual coherence for pairs of matrices.

**Definition 2.3.2.** For \( A = (a_1 \ldots a_p) \in \mathbb{C}^{m \times p} \) and \( B = (b_1 \ldots b_q) \in \mathbb{C}^{m \times q} \), both with columns \( \|\cdot\|_2 \)-normalized to 1, the mutual coherence \( \bar{\mu}(A, B) \) is defined as \( \bar{\mu}(A, B) = \max_{i,j} |a_i^H b_j| \).

The general uncertainty relation we are now ready to state is in terms of a pair of upper bounds on \( \|p_P\|_1 \) and \( \|q_Q\|_1 \) for \( P \subseteq \{1, \ldots, p\} \) and \( Q \subseteq \{1, \ldots, q\} \).

**Theorem 2.3.3.** Let \( A \in \mathbb{C}^{m \times p} \) and \( B \in \mathbb{C}^{m \times q} \), both with column vectors \( \|\cdot\|_2 \)-normalized to 1, and consider \( p \in \mathbb{C}^p \) and \( q \in \mathbb{C}^q \). Suppose that \( Ap = Bq \). Then, we have

\[
\|p_P\|_1 \leq |P| \left( \frac{\mu(A)\|p\|_1 + \bar{\mu}(A, B)\|q\|_1}{1 + \mu(A)} \right)
\]

(2.108)

for all \( P \subseteq \{1, \ldots, p\} \) and, by symmetry,

\[
\|q_Q\|_1 \leq |Q| \left( \frac{\mu(B)\|q\|_1 + \bar{\mu}(A, B)\|p\|_1}{1 + \mu(B)} \right)
\]

(2.109)

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

$$|a_i^H A^* p| = |a_i^H B^* q|. \quad (2.110)$$

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

$$|a_i^H A^* p| = \left| p_i + \sum_{k=1, k \neq i}^p a_i^H a_k p_k \right| \quad (2.111)$$

$$\geq |p_i| - \left| \sum_{k=1, k \neq i}^p a_i^H a_k p_k \right| \quad (2.112)$$

$$\geq |p_i| - \sum_{k=1, k \neq i}^p |a_i^H a_k| |p_k| \quad (2.113)$$

$$\geq |p_i| - \mu(A) \sum_{k=1, k \neq i}^p |p_k| \quad (2.114)$$

$$= (1 + \mu(A))|p_i| - \mu(A)\|p\|_1, \quad (2.115)$$

where (2.112) is by the reverse triangle inequality and in (2.114) we used Definition 2.2.2. Next, we upper-bound the right-hand side of (2.110) according to

$$|a_i^H B^* q| = \left| \sum_{k=1}^q a_i^H b_k q_k \right| \quad (2.116)$$

$$\leq \sum_{k=1}^q |a_i^H b_k| |q_k| \quad (2.117)$$

$$\leq \bar{\mu}(A, B)\|q\|_1, \quad (2.118)$$

where the last step is by Definition 2.3.2. Combining the lower bound (2.111)–(2.115) and the upper bound (2.116)–(2.118) yields

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

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

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

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

$$\|p_P\|_1 \leq |P| \mu([I \ B]) \|q\|_1$$

(2.121)

and

$$\|q_Q\|_1 \leq |Q| \bar{\mu}(I, B) \|p\|_1,$$

(2.122)

respectively. Thus, for arbitrary but fixed $p \in W^B$, $Q \subseteq \{1, \ldots, q\}$, $p \in \mathbb{C}^p$, and $q = B^H p$, we have $q_Q = q$ so that (2.121) and (2.122) taken together yield

$$\|p_P\|_1 \leq |P| \mu^2([I \ B]) \|q\|_1.$$  

(2.123)

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

$$\Sigma_{P, Q}(B) \leq |P| |Q| \mu^2([I \ B]).$$

(2.124)

in Lemma 5.

2.3.5 Concentration Inequalities for Pairs of General Matrices

We next refine the result in Theorem 2.3.3 to vectors that are concentrated in 1-norm according to Definition 2.3.1. The formal statement is as follows.

**Corollary 2.3.4.** Let $A \in \mathbb{C}^{m \times p}$ and $B \in \mathbb{C}^{m \times q}$, both with column vectors $\|\cdot\|_2$-normalized to 1, $P \subseteq \{1, \ldots, p\}$, $Q \subseteq \{1, \ldots, q\}$, $p \in \mathbb{C}^p$, and $q \in \mathbb{C}^q$. Suppose that $Ap = Bq$. Then, the following statements hold.

1. If $q$ is $\varepsilon_Q$-concentrated, then,

$$\|p_P\|_1 \leq \frac{|P|}{1 + \mu(A)} \left( \mu(A) + \frac{\bar{\mu}^2(A, B)|Q|}{(1 + \mu(B))(1 - \varepsilon_Q) - \mu(B)|Q|_+} \right) \|p\|_1.$$  

(2.125)

2. If $p$ is $\varepsilon_P$-concentrated, then,

$$\|q_Q\|_1 \leq \frac{|Q|}{1 + \mu(B)} \left( \mu(B) + \frac{\bar{\mu}^2(A, B)|P|}{(1 + \mu(A))(1 - \varepsilon_P) - \mu(A)|P|_+} \right) \|q\|_1.$$  

(2.126)

3. If $p$ is $\varepsilon_P$-concentrated, $q$ is $\varepsilon_Q$-concentrated, $\bar{\mu}(A, B) > 0$, and $(p^T q^T) \neq 0$, then,

$$|P| |Q| \geq \frac{(1 + \mu(A))(1 - \varepsilon_P) - \mu(A)|P|_+}{\bar{\mu}^2(A, B)} [(1 + \mu(B))(1 - \varepsilon_Q) - \mu(B)|Q|_+] + (1 + \mu(B))(1 - \varepsilon_Q) - \mu(B)|Q|_+.$$  

(2.127)
Proof. By Theorem [2.3.3] we have
\[ \| \mathbf{p}_\mathcal{P} \|_1 \leq |\mathcal{P}| \left( \frac{\mu(A)\|q\|_1 + \mu(A,B)\|p\|_1}{1 + \mu(A)} \right) \] (2.128)
and
\[ \| \mathbf{q}_\mathcal{Q} \|_1 \leq |\mathcal{Q}| \left( \frac{\mu(B)\|q\|_1 + \mu(A,B)\|p\|_1}{1 + \mu(B)} \right). \] (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
\[ \| \mathbf{q} \|_1 \leq \frac{|\mathcal{Q}| \bar{\mu}(A,B) + \mu(A,B)\|q\|_1}{(1 + \mu(B))(1 - \varepsilon_\mathcal{Q}) - \mu(B)\|q\|_1 \} \] (2.130)
Using (2.130) in (2.128) yields (2.125). The relation (2.126) follows from (2.125) by swapping the roles of \( A \) and \( B \), \( \mathbf{p} \) and \( \mathbf{q} \), and \( \mathcal{P} \) and \( \mathcal{Q} \), and upon noting that \( \bar{\mu}(A,B) = \bar{\mu}(B,A) \). 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
\[ \| \mathbf{p} \|_1 [(1 + \mu(A))(1 - \varepsilon_\mathcal{P}) - \mu(A)\|\mathcal{P}\|_1] \leq \bar{\mu}(A,B)\|q\|_1 \|\mathcal{P}\| \] (2.131)
and
\[ \| \mathbf{q} \|_1 [(1 + \mu(B))(1 - \varepsilon_\mathcal{Q}) - \mu(B)\|\mathcal{Q}\|_1] \leq \bar{\mu}(A,B)\|p\|_1 \|\mathcal{Q}\|, \] (2.132)
respectively. Suppose first that \( \mathbf{p} = 0 \). Then, \( \mathbf{q} \neq 0 \) by assumption, and (2.132) becomes
\[ [(1 + \mu(B))(1 - \varepsilon_\mathcal{Q}) - \mu(B)\|\mathcal{Q}\|_1] = 0. \] (2.133)
In this case (2.127) holds trivially. Similarly, if \( \mathbf{q} = 0 \), then \( \mathbf{p} \neq 0 \) again by assumption, and (2.131) becomes
\[ [(1 + \mu(A))(1 - \varepsilon_\mathcal{P}) - \mu(A)\|\mathcal{P}\|_1] = 0. \] (2.134)
As before, (2.127) holds trivially. Finally, if \( \mathbf{p} \neq 0 \) and \( \mathbf{q} \neq 0 \), then we multiply (2.131) by (2.132) and divide the result by \( \bar{\mu}^2(A,B)\|\mathbf{p}\|_1\|\mathbf{q}\|_1 \) 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 cannot be arbitrarily well concentrated with respect to two different general matrices \( A \) and \( B \). For the special case \( \varepsilon_\mathcal{Q} = 0 \) and \( A \) and \( B \) unitary, and hence \( \mu(A) = \mu(B) = 0 \) and \( \bar{\mu}(A,B) = \mu([A \ B]) \), (2.127) recovers Lemma 6.

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

**Corollary 2.3.5.** [36 Lemma 33] Let \( A \in \mathbb{C}^{m \times p} \) and \( B \in \mathbb{C}^{m \times q} \), 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 \( (\mathbf{p}^T \mathbf{q}^T)^T \neq 0 \). Suppose that \( A \mathbf{p} = B \mathbf{q} \). Then, \( \| \mathbf{p} \|_0 \| \mathbf{q} \|_0 \geq f_{A,B}(\| \mathbf{p} \|_0, \| \mathbf{q} \|_0) \), where
\[ f_{A,B}(u, v) = \frac{[1 + \mu(A)(1 - u)][1 + \mu(B)(1 - v)]_+}{\bar{\mu}^2(A,B)}. \] (2.135)
Proof. Let \( P = \{ i \in \{1, \ldots, p \} : p_i \neq 0 \} \) and \( Q = \{ i \in \{1, \ldots, q \} : q_i \neq 0 \} \), so that \( p_P = p \), \( q_Q = q \). \(|P| = \|p\|_0\), and \(|Q| = \|q\|_0\). The claim now follows directly from (2.127) with \( \varepsilon_P = \varepsilon_Q = 0 \).

If \( A \) and \( B \) are both unitary, then \( \mu(A) = \mu(B) = 0 \) and \( \tilde{\mu}(A, B) = \mu([A \ B]) \), and Corollary 2.3.5 recovers the Elad-Bruckstein result in Corollary 2.2.5.

Corollary 2.3.5 admits the following appealing geometric interpretation in terms of a null-space property.

**Lemma 8.** Let \( A \in \mathbb{C}^{m \times p} \) and \( B \in \mathbb{C}^{m \times q} \), both with column vectors \( \| \cdot \|_2 \)-normalized to 1. Then, the set (which actually is a finite union of subspaces)

\[
S = \left\{ \begin{pmatrix} p \\ q \end{pmatrix} : p \in \mathbb{C}^p, q \in \mathbb{C}^q, \|p\|_0\|q\|_0 < f_{A,B}(\|p\|_0, \|q\|_0) \right\}
\]

with \( f_{A,B} \) defined in (2.135) intersects the kernel of \([A \ B]\) trivially, i.e.,

\[
\ker([A \ B]) \cap S = \{0\}.
\]

Proof. The statement of this lemma is equivalent to the statement of Corollary 2.3.5 through a chain of equivalences between the following statements:

1. \( \ker([A \ B]) \cap S = \{0\} \);
2. if \( (p^T - q^T)^T \in \ker([A \ B]) \setminus \{0\} \), then \( \|p\|_0\|q\|_0 \geq f_{A,B}(\|p\|_0, \|q\|_0) \);
3. if \( Ap = Bq \) with \( (p^T q^T)^T \neq 0 \), then \( \|p\|_0\|q\|_0 \geq f_{A,B}(\|p\|_0, \|q\|_0) \).

where \( 1 \Leftrightarrow 2 \) is by definition of \( S \), \( 2 \Rightarrow 3 \) follows from the fact that \( Ap = Bq \) with \( (p^T q^T)^T \neq 0 \) is equivalent to \( (p^T - q^T)^T \in \ker([A \ B]) \setminus \{0\} \), and \( 3 \) is the statement in Corollary 2.3.5.

### 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 9.** Let \( \mu \) be a 1-periodic, \( \sigma \)-finite measure on \( \mathbb{R} \), \( n \in \mathbb{N} \), \( \varphi \in [0, 1) \), \( a \in \mathbb{C}^n \), and consider the 1-periodic trigonometric polynomial

\[
\psi(s) = e^{i2\pi \varphi} \sum_{k=1}^{n} a_k e^{-2\pi i ks}.
\]

Then,

\[
\int_{(0,1)} |\psi(s)|^2 d\mu(s) \leq \left( n - 1 + \frac{1}{\delta} \right) \sup_{r \in [0,1)} \mu((r, r + \delta)) \|a\|^2_2
\]

for all \( \delta \in (0, 1] \).
Proof. Since
\[
|\psi(s)| = \left| \sum_{k=1}^{n} a_k e^{-2\pi i k s} \right| ,
\] (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
\[
G(s) = \int_{-\infty}^{\infty} g(t) e^{-2\pi ist} dt
\] (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
\[
\theta(s) = \sum_{k=1}^{n} \frac{a_k}{g(k)} e^{-2\pi i k s} \] (2.142)
and note that
\[
\int_{-\delta/2}^{\delta/2} G(r) \theta(s-r) dr = \sum_{k=1}^{n} \frac{a_k}{g(k)} e^{-2\pi i k s} \int_{-\infty}^{\infty} G(r) e^{2\pi ikr} dr
\] (2.143)
\[
= \sum_{k=1}^{n} a_k e^{-2\pi i k s}
\] (2.144)
\[
= \psi(s) \quad \text{for all} \ s \in \mathbb{R}.
\] (2.145)
We now have

\[
\int_{[0,1]} |\psi(s)|^2 d\mu(s) = \int_{[0,1]} \left| \int_{-\delta/2}^{\delta/2} G(r) \theta(s - r) dr \right|^2 d\mu(s) 
\]

(2.146)

\[
\leq \|G\|_2^2 \int_{[0,1]} \left( \int_{-\delta/2}^{\delta/2} |\theta(s - r)|^2 dr \right) d\mu(s) 
\]

(2.147)

\[
= \|G\|_2^2 \int_{[0,1]} \left( \int_{s-\delta/2}^{s+\delta/2} |\theta(r)|^2 dr \right) d\mu(s) 
\]

(2.148)

\[
= \|G\|_2^2 \int_{-\delta}^{\delta} \mu((r - \delta/2, r + \delta/2) \cap [0,1]) |\theta(r)|^2 dr 
\]

(2.149)

\[
= \|G\|_2^2 \sum_{i=-1}^{1} \int_{0}^{1+i} \mu((r - \delta/2, r + \delta/2) \cap [0,1]) |\theta(r)|^2 dr 
\]

(2.150)

\[
= \|G\|_2^2 \sum_{i=-1}^{1} \int_{0}^{1} \mu((r - \delta/2, r + \delta/2) \cap [i+1,i]) |\theta(r)|^2 dr 
\]

(2.151)

\[
= \|G\|_2^2 \int_{0}^{1} \mu((r - \delta/2, r + \delta/2) \cap [-1,2]) |\theta(r)|^2 dr 
\]

(2.152)

\[
= \|G\|_2^2 \int_{0}^{1} \mu((r - \delta/2, r + \delta/2)) |\theta(r)|^2 dr 
\]

(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 \(\mu\) is \(\sigma\)-finite by assumption) upon noting that

\[
\{(r,s) : s \in [0,1), r \in (s - \delta/2, s + \delta/2)\} 
\]

(2.154)

\[
= \{(r,s) : r \in [-1,2), s \in (r - \delta/2, r + \delta/2) \cap [0,1)\} 
\]

(2.155)

for all \(\delta \in (0,1]\), in (2.151) we used the 1-periodicity of \(\mu\) and \(\theta\), and (2.152) is by \(\sigma\)-additivity of \(\mu\). Now,

\[
\int_{0}^{1} \mu((r - \delta/2, r + \delta/2)) |\theta(r)|^2 dr \leq \sup_{r \in [0,1)} \mu((r, r + \delta)) \int_{0}^{1} |\theta(r)|^2 dr 
\]

(2.156)

\[
= \sup_{r \in [0,1)} \mu((r, r + \delta)) \sum_{k=1}^{n} \frac{|a_k|^2}{|g(k)|^2} 
\]

(2.157)

\[
\leq \sup_{r \in [0,1)} \mu((r, r + \delta)) ||a||_2^2 
\]

(2.158)

for all \(\delta > 0\), where (2.158) follows from \(|g(t)|^2 \geq 1\) 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 9 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 \([33, \text{Equation (32)}]\) can be strictly larger than that in \((2.139)\) whenever \(\mu\) 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 \(A = I\) and \(B = F\) in our setting.

<table>
<thead>
<tr>
<th></th>
<th>(L_2) analog</th>
<th>(L_1) analog</th>
</tr>
</thead>
<tbody>
<tr>
<td>Upper bound in ((2.10))</td>
<td>([32, \text{Lemma 2}])</td>
<td></td>
</tr>
<tr>
<td>Corollary 2.2.4</td>
<td>([32, \text{Theorem 2}])</td>
<td></td>
</tr>
<tr>
<td>Lemma 4</td>
<td>([32, \text{Theorem 4}])</td>
<td></td>
</tr>
<tr>
<td>Lemma 5</td>
<td>([32, \text{Lemma 3}])</td>
<td></td>
</tr>
<tr>
<td>Lemma 7</td>
<td>([33, \text{Lemma 2}])</td>
<td></td>
</tr>
<tr>
<td>Lemma 9</td>
<td>([33, \text{Theorem 4}])</td>
<td></td>
</tr>
</tbody>
</table>

### 2.6 Results for \(\|\cdot\|_1\) and \(\|\cdot\|_2\)

**Lemma 10.** Let \(U \in \mathbb{C}^{m \times m}\) be unitary, \(P, Q \subseteq \{1, \ldots, m\}\), and consider the orthogonal projection \(P_Q(U) = UD_QU^H\) onto the subspace \(W_{U,Q}\). Then,

\[
\|\|P_Q(U)D_P\|\|_2 = \|\|D_P P_Q(U)\|\|_2. \tag{2.159}
\]

Moreover, we have

\[
\|\|D_P P_Q(U)\|\|_2 = \max_{x \in W_{U,Q} \setminus \{0\}} \frac{\|x_P\|_2}{\|x\|_2} \tag{2.160}
\]

and

\[
\|\|D_P P_Q(U)\|\|_1 = \max_{x \in W_{U,Q} \setminus \{0\}} \frac{\|x_P\|_1}{\|x\|_1}. \tag{2.161}
\]

**Proof.** The identity \((2.159)\) follows from

\[
\|\|D_P P_Q(U)\|\|_2 = \|\|(D_P P_Q(U))^*\|\|_2
\]

\[
= \|\|P_Q(U)D_P^*\|\|_2
\]

\[
= \|\|P_Q(U)D_P\|\|_2, \tag{2.164}
\]

\[
\|\|D_P P_Q(U)\|\|_1 = \max_{x \in W_{U,Q} \setminus \{0\}} \frac{\|x_P\|_1}{\|x\|_1}. \tag{2.161}
\]
where in (2.162) we used that $\| \cdot \|_2$ is self-adjoint \cite[p. 309]{48}, $P_\mathcal{Q}(U)^* = P_\mathcal{Q}(U)$, and $D_P = D_P$.

To establish (2.160), we note that
\[
\|D_P P_\mathcal{Q}(U)\|_2 = \max_{x \in \mathbb{R}^{m \times 1}} \|D_P P_\mathcal{Q}(U)x\|_2 \tag{2.165}
\]
\[
= \max_{x : \|x\|_2 = 1} \|D_P P_\mathcal{Q}(U)x\|_2 \tag{2.166}
\]
\[
\leq \max_{x : \|x\|_2 = 1} \|D_P P_\mathcal{Q}(U)x\|_2 \tag{2.167}
\]
\[
\leq \max_{x : \|x\|_2 = 1} \|D_P P_\mathcal{Q}(U)x\|_2 \tag{2.168}
\]
\[
= \max_{x : \|x\|_2 = 1} \|D_P P_\mathcal{Q}(U)x\|_2 \tag{2.169}
\]
\[
= \|D_P P_\mathcal{Q}(U)\|_2 \tag{2.170}
\]
where in (2.167) we used $\|P_\mathcal{Q}(U)x\|_2 \leq \|x\|_2$, which implies $\|P_\mathcal{Q}(U)x\|_2 \leq 1$ for all $x$ with $\|x\|_2 = 1$. Finally, (2.161) follows by repeating the steps in (2.165)–(2.172) with $\| \cdot \|_2$ replaced by $\| \cdot \|_1$ at all occurrences.

**Lemma 11.** Let $A \in \mathbb{C}^{m \times n}$. Then,
\[
\frac{\|A\|_2}{\sqrt{\text{rank}(A)}} \leq \|A\|_2 \leq \|A\|_2. \tag{2.173}
\]

**Proof.** The proof is trivial for $A = 0$. If $A \neq 0$, set $r = \text{rank}(A)$ and let $\sigma_1, \ldots, \sigma_r$ denote the nonzero singular values of $A$ organized in decreasing order. Unitary invariance of $\|\cdot\|_2$ and $\|\cdot\|_2$ (cf. \cite[Problem 5, p. 311]{48}) yields $\|A\|_2 = \sigma_1$ and $\|A\|_2 = \sqrt{\sum_{i=1}^{r} \sigma_i^2}$. 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 12.** For $A = (a_1 \ldots a_n) \in \mathbb{C}^{m \times n}$, we have
\[
\|A\|_1 = \max_{j \in \{1, \ldots, n\}} \|a_j\|_1 \tag{2.175}
\]
and
\[
\frac{1}{n} \|A\|_1 \leq \|A\|_1 \leq \|A\|_1. \tag{2.176}
\]
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 a 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

\[
\hat{x}[k] := \frac{1}{\sqrt{N}} \hat{x}\left(\frac{k}{N}\right)
\]

\[
= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-2\pi i kn/N}
\]

\[
= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] \omega_N^{kn}, \quad \text{for } k = 0, 1, \ldots, N - 1,
\]

where \(\omega_N = e^{-2\pi i/N}\). It holds that \(\hat{x}[k + N] = \hat{x}[k]\), for all \(k \in \mathbb{Z}\). Now we write this relationship in the vector-matrix form:

\[
\begin{bmatrix}
\hat{x}[0] \\
\hat{x}[1] \\
\vdots \\
\hat{x}[N-1]
\end{bmatrix}
= \frac{1}{\sqrt{N}} \begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & \omega_N & \omega_N^2 & \cdots & \omega_N^{N-1} \\
1 & \omega_N^2 & \omega_N^4 & \cdots & \omega_N^{2(N-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \omega_N^{N-1} & \omega_N^{2(N-1)} & \cdots & \omega_N^{(N-1)^2}
\end{bmatrix}
\begin{bmatrix}
x[0] \\
x[1] \\
\vdots \\
x[N-1]
\end{bmatrix}
= \begin{bmatrix}
\hat{x} \\
x[1] \\
\vdots \\
x[N-1]
\end{bmatrix}
\]

\[
\mathbf{F}_N
\]
The matrix $F_N$ is the $N \times N$ DFT-Matrix and is unitary, i.e., $F_N F_N^H = F_N^H F_N = I_N$, where $I_N$ denotes the $N \times N$ identity matrix. We can thus directly specify how the signal $x$ of length $N$ can be recovered from $\hat{x}$. From $\hat{x} = F_N x$ it follows after multiplication with $F_N^H$ from left that

$$F_N^H \hat{x} = F_N^H F_N x = I_N x = x \quad \Rightarrow \quad x = F_N^H \hat{x}$$

and

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \hat{x}[k] e^{2\pi i kn/N}.$$ 

We thus have the following transformation pair for the discrete Fourier transform:

$$\hat{x}[k] = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} x[n] e^{-2\pi i kn/N}, \quad \text{with } \hat{x}[k + N] = \hat{x}[k]$$

$$x[n] = \frac{1}{\sqrt{N}} \sum_{k=0}^{N-1} \hat{x}[k] e^{2\pi i kn/N}, \quad \text{with } x[n + N] = x[n].$$

In vector matrix form this transformation pair can be written as

$$\hat{x} = F_N x$$

$$x = F_N^H \hat{x}.$$ 

With $F_N^H = \begin{bmatrix} f_0 & f_1 & \cdots & f_{N-1} \end{bmatrix}$ we obtain

$$x = F_N^H F_N x = \begin{bmatrix} f_0 & f_1 & \cdots & f_{N-1} \end{bmatrix} \begin{bmatrix} f_0^H \\ f_1^H \\ \vdots \\ f_{N-1}^H \end{bmatrix} x = \sum_{\ell=0}^{N-1} \langle x, f_\ell \rangle f_\ell.$$ 

Thus we have shown that the DFT is nothing else than the expansion of the vector $x$ into an orthonormal basis (ONB) for $\mathbb{C}^N$.

### 3.1.1 Oversampling

$$\hat{x}[k] = \frac{1}{\sqrt{M}} \hat{x} \left( \frac{k}{M} \right)$$

$$= \frac{1}{\sqrt{M}} \sum_{n=0}^{N-1} x[n] e^{-2\pi i kn/M}, \quad k = 0, 1, \ldots, M - 1, \text{ with } M > N.$$
In vector-matrix form we obtain
\[
\begin{bmatrix}
\hat{x}(0) \\
\hat{x}(1/M) \\
\vdots \\
\hat{x}((M-1)/M)
\end{bmatrix}
= \frac{1}{\sqrt{M}}
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & \omega_M & \omega_M^2 & \cdots & \omega_M^{N-1} \\
1 & \omega_M^2 & \omega_M^4 & \cdots & \omega_M^{2(N-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \omega_M^{M-1} & \omega_M^{2(M-1)} & \cdots & \omega_M^{(N-1)(M-1)}
\end{bmatrix}
\begin{bmatrix}
x[0] \\
x[1] \\
\vdots \\
x[N-1]
\end{bmatrix}
\]

The columns of \(F_o\) are orthonormal since these are the first \(N\) columns of the DFT matrix \(F_M\). This implies that we can recover \(x\) from \(\hat{x}\), but there are infinitely many inverses. A selected inverse is the Moore-Penrose pseudo-inverse given by \((F_o^H F_o)^{-1} F_o^H\). First, we observe that
\[
(F_o^H F_o)^{-1} F_o^H \hat{x} = (F_o^H F_o)^{-1} F_o^H F_o x = x. \quad (3.1)
\]
Because of the orthonormality of the columns of \(F_o\), we have \(F_o^H F_o = I_N\) so that
\[
x = F_o^H \hat{x}. \quad (3.2)
\]
It is noteworthy that, despite oversampling, the inverse transform (corresponding to the pseudo-inverse) is given by
\[
x[n] = \frac{1}{\sqrt{M}} \sum_{k=0}^{M-1} \hat{x}[k] \omega_M^{-kn}.
\]

### 3.1.2 Undersampling

\[
\hat{x}[k] = \frac{1}{\sqrt{M}} \hat{x} \left( \frac{k}{M} \right)
= \frac{1}{\sqrt{M}} \sum_{n=0}^{N-1} x[n] e^{-2\pi i kn/M}, \quad k = 0, 1, \ldots, M - 1, \text{ with } M < N.
\]

In vector-matrix form we obtain
\[
\begin{bmatrix}
\hat{x}(0) \\
\hat{x}(1/M) \\
\vdots \\
\hat{x}((M-1)/M)
\end{bmatrix}
= \frac{1}{\sqrt{M}}
\begin{bmatrix}
1 & 1 & 1 & \cdots & 1 \\
1 & \omega_M & \omega_M^2 & \cdots & \omega_M^{N-1} \\
1 & \omega_M^2 & \omega_M^4 & \cdots & \omega_M^{2(N-1)} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & \omega_M^{M-1} & \omega_M^{2(M-1)} & \cdots & \omega_M^{(N-1)(M-1)}
\end{bmatrix}
\begin{bmatrix}
x[0] \\
x[1] \\
\vdots \\
x[N-1]
\end{bmatrix}
\]

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 $y$ of dimension $\dim y = m \times 1$ obtained when compressing the signal $x$ of dimension $\dim x = n \times 1$ by means of the matrix $D$ of dimension $\dim D = m \times n$ according to

$$y = Dx,$$

we would like to recover the original vector $x$. Clearly, this problem has infinitely many solutions when the set of signals is not constrained. We consider the case where $x$ is $s$-sparse, i.e., $\|x\|_0 \leq s$ or, in words, the signal vector $x$ 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 $D$.

We go back to the case of spectrally sparse signals, this time in the finite-dimensional case.
$m$ measurements in the time-domain

\[ y \ldots m \times 1 \]

\[ F^H \]

\[ D \ldots m \times n \]

\[ x \ldots n \times 1 \]

\[ s \text{ nonzero components} \]

$m = s$: sampling at Landau rate

$m > s$: oversampling

The sampling instants (rows) must be chosen such that $D$ is well-conditioned.

- If the support set of $x$ is known, we can apply the following universal sampling pattern. Choose the first $m = s$ rows of $F$. The resulting $D$-matrix is Vandermonde and hence full-rank, irrespectively of the support set.

- If the support set of $x$ is unknown, we can apply the following universal sampling pattern. Choose the first $m = 2s$ rows of $F$. Any $x_1 - x_2$ with $x_1, x_2$ $s$-sparse and $x_1 \neq x_2$ satisfies $\|D(x_1 - x_2)\|^2 > 0$ since $x_1 - x_2$ is $2s$-sparse and the resulting $D$-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 $y$ depicted in the figure below.

![Diagram](image)

### 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 $x$. In the example depicted below, this is clearly not the case.
Spikes and sinusoids are, e.g., ’incoherent’.

3.2.2 The General Problem

\[ D = \begin{bmatrix} \text{subsampling} \\ \text{sparsity basis} \end{bmatrix} \]
For a given sparsity basis (e.g., wavelets), find a sampling basis such that \( s \)-sparse vectors are distinguishable, i.e., for all \( x_1, x_2 \) that are \( s \)-sparse with \( x_1 \neq x_2 \)

\[
\|D (x_1 - x_2)\|_2^2 > 0.
\]

Hence, all collections of \( 2s \) columns of \( D \) 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 \( D \) is normalized to unit 2-norm.

**Definition 3.2.1.** The spark of a matrix \( A \) denoted by \( \text{spark}(A) \) is defined as the cardinality of the smallest set of linearly dependent columns.

For a given matrix \( D \) of dimension \( m \times n \), uniqueness of recovery of \( s \)-sparse vectors \( x \) from the observation \( y = Dx \) is guaranteed for

\[
s < \frac{\text{spark}(D)}{2}.
\]

### 3.3 The Recovery Problem (P0)

If \( D \) is a (known) ONB, recovering \( x \) from \( y \) is simply

\[
D^H y = D^H Dx = x.
\]

If \( D \) is a (known) basis, we have

\[
D^{-1} y = D^{-1} Dx = x.
\]

If \( D = [A \ d] \) where \( A \) is an ONB and \( d \) is an extra column, then we cannot uniquely determine \( x \) from \( y = Dx \).

However, if \( x \) is \( s \)-sparse and

\[
s < \frac{\text{spark}(D)}{2}
\]

we can recover \( x \) through a combinatorial search:

\[
(P0) \quad \text{find } \arg \min \|\hat{x}\|_0 \text{ subject to } y = D\hat{x}
\]
For any vector \( x \), the quasi-norm \( \| x \|_0 \) denotes the number of nonzero entries.

Suppose that \( \| x \|_0 \leq s \) and \( s < \frac{\text{spark}(D)}{2} \). Let \( \tilde{x} \neq x \) with \( \| \tilde{x} \|_0 \leq s \) and \( y = D\tilde{x} \), then

\[
0 = y - y = Dx - D\tilde{x} = D(x - \tilde{x}) .
\]

Since \( 2s < \text{spark}(D) \), we know, however, that

\[
\| D(x - \tilde{x}) \| > 0, \quad x - \tilde{x} \neq 0
\]

as any set of \( 2s \) columns of \( D \) is linearly independent. Therefore (P0) recovers \( x \) 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 \( \mu(D) \) defined according to (see Definition 2.2.2)

\[
\mu(D) = \max_{r \neq l} |\langle d_l, d_r \rangle| .
\]

**Theorem 3.3.1.** [19]; [53] (P0) applied to \( y = Dx \) recovers \( x \) if

\[
\| x \|_0 \leq s < \frac{1}{2} \left( 1 + \frac{1}{\mu(D)} \right).
\]

**Proof.** We will show that \( \text{spark}(D) \geq 1 + 1/\mu(D) \). Consider \( x \in \mathbb{C}^n \) with \( \| x \|_0 = \text{spark}(D) \) and \( Dx = 0 \). Then, we have

\[
d_l x_l = - \sum_{r \neq l} d_r x_r, \quad \text{for all } l \in \{1, \ldots, n\}.
\]

Left-multiplying both sides by \( d_l^H \) and using \( \| d_l \|_2 = 1 \) yields

\[
x_l = - \sum_{r \neq l} d_l^H d_r x_r ,
\]

which implies

\[
| x_l | = \sum_{r \neq l} d_l^H d_r | x_r | \leq \sum_{r \neq l} | d_l^H d_r | | x_r | \leq \mu(D) \sum_{r \neq l} | x_r | \quad \text{for } l \in \{1, \ldots, n\}.
\]
Adding $\mu(D) |x_l|$ on both sides results in

$$
(1 + \mu(D)) |x_l| \leq \mu(D) \|x\|_1 \quad \text{for } l \in \{1, \ldots, n\}.
$$

Summing over all $l$ for which $x_l \neq 0$ finally leads to

$$
(1 + \mu(D)) \|x\|_1 \leq \mu(D) \|x\|_1 \text{spark}(D)
$$

$$
\Rightarrow \text{spark}(D) \geq 1 + \frac{1}{\mu(D)}.
$$

Notice that determining $\mu(D)$ has the complexity of doing the first step in the computation of $\text{spark}(D)$, i.e., checking whether any two columns are linearly independent.

### 3.4 Basis Pursuit (BP)

In this section, we consider the recovery problem below

(P1) \text{find } \arg \min \|\hat{x}\|_1 \text{ subject to } y = D\hat{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 $\ell_1$-reconstruction:

- Logan, 1965
- Donoho & Logan, 1992

Why does $\ell_1$-reconstruction work?

$$
\arg \min \|\hat{x}\|_1 \text{ subject to } y = D\hat{x}
$$

$\Downarrow$

$$
\arg \min \|\hat{x}\|_1 \text{ subject to } \hat{x} \in (\{x\} + \mathcal{N}(D))
$$
$\ell_1$-ball: $|z_1| + |z_2| = \text{const.}$

Case $z_1, z_2 > 0$: $z_1 + z_2 = \text{const.} \Rightarrow z_2 = \text{const.} - z_1$

By symmetry, the $\ell_1$-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 (P1) finds the correct solution?

We will need a result on the concentration of $\ell_1$ norms.

**Definition 3.4.1.** We denote
\[
P_1(S, D) \triangleq \max_{x \in \mathcal{N}(D), x \neq 0} \frac{\sum_{k \in S} |x_k|}{\sum_{k} |x_k|}.
\]

**Theorem 3.4.2.** Arbitrarily fix $x$ with support set $S$ and let $y = Dx$. If $P_1(S, D) < 1/2$, then $x$ is the unique solution to

(P1) \quad \text{find} \ \arg \min \|\hat{x}\|_1 \text{ subject to } y = D\hat{x}

**Proof.** We need to prove that for all $\alpha \in \mathcal{N}(D)$
\[
\sum_k |x_k + \alpha_k| > \sum_k |x_k|.
\]

Application of the reverse triangle inequality
\[
|a + b| \geq |a| - |b|
\]
to the LHS of the above equation yields
\[
\sum_k |x_k + \alpha_k| = \sum_{k \notin S} |x_k + \alpha_k| + \sum_{k \in S} |x_k + \alpha_k| \\
= \sum_{k \notin S} |\alpha_k| + \sum_{k \in S} |x_k + \alpha_k| \\
\geq \sum_{k \notin S} |\alpha_k| + \sum_{k \in S} |x_k| - \sum_{k \in S} |\alpha_k|.
\]
Therefore, the theorem follows from
\[ \sum_{k \not\in S} |\alpha_k| > \sum_{k \in S} |\alpha_k|. \]

Adding \( \sum_{k \in S} |\alpha_k| \) to both sides of the above equation results in
\[ \sum_k |\alpha_k| > 2 \sum_{k \in S} |\alpha_k| \]
\[ \Rightarrow \frac{\sum_{k \in S} |\alpha_k|}{\sum_k |\alpha_k|} < \frac{1}{2} \]

but this is satisfied for all \( \alpha \in \mathcal{N}(D) \) since \( P_1(S, D) < 1/2 \) by assumption. \( \square \)

Next, we find a sufficient condition for \( P_1(S, D) < \frac{1}{2} \) in terms of the cardinality of the support set \(|S|\) and the dictionary coherence \( \mu \).

Consider \( \alpha \in \mathcal{N}(D) \). Due to the proof of Theorem [3.3.1] we know that
\[ (1 + \mu(D)) |\alpha_l| \leq \mu(D) \| \alpha \|_1, \quad \text{for all} \ l = 1, \ldots, n. \]

Summing over all \( l \in S \), we get
\[ (1 + \mu(D)) \sum_{l \in S} |\alpha_l| \leq \mu(D) \| \alpha \|_1 |S| \]
\[ (1 + \mu(D)) \frac{\sum_{l \in S} |\alpha_l|}{\sum_l |\alpha_l|} \leq \mu(D) |S| \]
\[ (1 + \mu(D)) P_1(S, D) \leq \mu(D) |S| \]
\[ P_1(S, D) \leq \frac{1}{1 + 1/\mu(D)} |S|. \]

Therefore if \( \frac{1}{1 + 1/\mu} |S| < 1/2 \), then \( P_1(S, D) < 1/2 \), i.e., the desired upper bound on the cardinality of the support set is given by
\[ |S| < \frac{1}{2} \left( 1 + \frac{1}{\mu(D)} \right) \]
just as in the case of recovery via (P0).

We therefore proved

**Theorem 3.4.3.** [19]; [53] *(P1) applied to \( y = Dx \) recovers \( x \) if*
\[ \|x\|_0 < \frac{1}{2} \left( 1 + \frac{1}{\mu(D)} \right) . \]
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 $x$ is sparse with unknown support set
- we observe $y = Ax$
- only a subset of the entries of $y = Ax$ 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
  \[ z = \underbrace{Ax + e}_{y} = Ax + Ie \]

and choose $e$ such that the entries of $z = y + e$ corresponding to the missing entries in $y$ 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, $e$ will be sparse.

Hence, we observe $z = Ax + Be$ and know $A, B$ and that $x, e$ are sparse. Based on the observation $z$, we want to recover $x$ and $e$. 
3.5.2 Clipping

Instead of \( y = D x \) we observe \( z = g_a(y) \), where the function \( g_a(y) \) realizes entry-wise signal clipping to the interval \([-a, a]\).

Clipping can equivalently be modeled as

\[
z = y + e
\]

with \( e = g_a(y) - y \). Notice that the error locations can be determined by comparing the entries of \( z \) to the clipping threshold \( a \).

Consequently, we observe

\[
z = Ax + B e
\]

where \( e \) can depend on \( A \) and \( x \).

3.5.3 Signal Separation

We consider the superposition of \( Ax \) and \( Be \), i.e., based on the observation

\[
z = Ax + Be
\]

we want to recover the sparse signals \( x \) and \( e \).

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

\[
z = Ax + Be
\]

where \( A = F \) and \( B = I \).

3.5.5 Recovery of Signals Subject to Narrowband Interference

We observe a sparse signal corrupted by spectrally sparse noise, i.e.,

\[
z = Ax + Be
\]

where \( A = I \) and \( B = F \).
3.5.6 The General Problem

All signal recovery problems and applications discussed above can be embedded into a single model where we observe

\[ z = Ax + Be = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} x \\ e \end{bmatrix} = D\hat{x} \]

and we want to recover \( \hat{x} \) given \( z = D\hat{x} \). To account for all possible scenarios, we cannot assume \( e \) to be independent of \( A \) and \( x \).

In particular, we are interested in dictionaries \( D \) that are the concatenation of two ONBs, e.g., \( D = [I \ F] \). 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 \( D \) is the concatenation of two ONBs \( A \) and \( B \), refined bounds on \( \text{spark}(D) = \text{spark}([A \ B]) \) exist.

For a vector in the null-space of \( D = [A \ B] \), we have

\[ [A \ B] \begin{bmatrix} p \\ q \\ v \end{bmatrix} = 0. \]

Hence,

\[ Ap + Bq = 0 \Rightarrow Ap = B(-q) \triangleq s. \]

The signal \( s \) is represented in two different ways, namely in the dictionary \( A \) and in the dictionary \( B \).

Finding the vector \( v \) with minimum 0-norm among all vectors that satisfy

\[ [A \ B]v = 0 \]

amounts to answering the question: How sparse can \( p \) and \( q \) concurrently be? The uncertainty relation in Corollary 2.2.5 states that \( Ap + Bq = 0 \) is only possible if \( \|p\|_0\|q\|_0 \geq 1/\mu^2([A \ B]) \).

In the case where \( D = [A \ B] \) is a concatenation of two ONBs, the dictionary coherence evaluates to

\[ \mu(D) = \sup_{i \neq j} |\langle d_i, d_j \rangle| = \sup_{i,j} |\langle a_i, b_j \rangle|. \]
How can these uncertainty relations be used to obtain recovery thresholds? By assumption, we have

$$\text{spark}(D) = \|p\|_0 + \|q\|_0 = \left\| \begin{bmatrix} p \\ q \end{bmatrix} \right\|_0.$$ 

The arithmetic-mean – geometric-mean (AM-GM) inequality implies

$$\|p\|_0 + \|q\|_0 \geq 2 \sqrt{\|p\|_0 \|q\|_0} \geq \frac{2}{\mu(D)}.$$ 

This yields a lower bound on the number of nonzero entries a vector in the null-space of $D = [A \ B]$ 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

$$s < \frac{\text{spark}(D)}{2}.$$ 

Together with $\|p\|_0 + \|q\|_0 \geq 2/\mu(D)$, the last inequality results in the following threshold for $(P0)$-uniqueness:

$$s < \frac{1}{\mu(D)}.$$ 

Compare this to the old threshold

$$s < \frac{1}{2} \left(1 + \frac{1}{\mu(D)}\right) \approx \frac{1}{2\mu(D)}$$

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 $x$ and $e$ from the observation

$$z = Ax + Be$$

knowing that $x$ and $e$ are sparse. Let $\mathcal{X} = \text{supp}(x)$ and $\mathcal{E} = \text{supp}(e)$.

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., $A = F$, $B = I$
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., $A = I$, $B = F$

3. knowledge of $\mathcal{X}$ or $\mathcal{E}$

   $$z = Ax + Be$$

   e.g. a clipped spectrally sparse signal with unknown spectrum, i.e., $A = F$ and $B = I$, or inpainting (or super-resolution) where the unknown signal has a sparse representation in $A$ (e.g. 2D-DCT or wavelet transform)

   $$z = A_\mathcal{X}x_\mathcal{X} + Be$$

   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 $\mathcal{E}$

   $$z = A_\mathcal{X}x_\mathcal{X} + Be$$

   e.g. recovery of clipped band-limited signals with known spectral support, i.e., $A = F$, $B = I$.

We start by considering recovery via $(P0)$ in case 3 when $\mathcal{E}$ is known.

**Theorem 3.5.1.** [35] Let $z = Ax + Be$, where $\mathcal{E} = \text{supp}(e)$ is known. Consider the problem

$$(P0, \mathcal{E}) \quad \left\{ \begin{array}{l}
\text{minimize} \|\hat{x}\|_0 \\
\text{subject to} \ A\hat{x} \in ( \{ z \} + R(B\mathcal{E}) ) .
\end{array} \right.$$ 

If $n_x = \|x\|_0$ and $n_e = \|e\|_0$ satisfy

$$2n_xn_e < \frac{[1 - \mu(A)(2n_x - 1)]^+ [1 - \mu(B)(n_e - 1)]^+}{\bar{\mu}^2(A, B)} ,$$

then the unique solution of $(P0, \mathcal{E})$ applied to $z = Ax + Be$ is given by $x$.

**Proof.** Assume that there exists an alternative vector $\tilde{x}$ that satisfies $A\tilde{x} \in ( \{ z \} + R(B\mathcal{E}) )$ with $\|\tilde{x}\|_0 \leq n_x$. Then, there must exist an $\tilde{e}$ with $\text{supp}(\tilde{e}) \subseteq \mathcal{E}$ such that

$$A\tilde{x} + Be = A\tilde{x} + B\tilde{e} ,$$

which implies

$$A(x - \tilde{x}) = B(\tilde{e} - e) .$$
We now apply the uncertainty principle Corollary 2.3.5 to the vectors \((x - \tilde{x})\) and \((e - \tilde{e})\). If both \(x\) and \(\tilde{x}\) have at most \(n_x\) nonzero entries, they can differ in at most \(2n_x\) positions, i.e., \(\|x - \tilde{x}\|_0 \leq 2n_x\). The vectors \(e\) and \(\tilde{e}\) are supported on the same set \(E\) of cardinality \(n_e\). Hence \(e\) and \(\tilde{e}\) can differ in at most \(n_e\) positions, i.e., \(\|e - \tilde{e}\|_0 \leq n_e\).

The uncertainty relation Corollary 2.3.5 sets a limit on how sparse \((x - \tilde{x})\) and \((e - \tilde{e})\) can be. In particular, \((x - \tilde{x})\) and \((\tilde{e} - e)\) cannot both be arbitrarily sparse. Substituting

\[
p = x - \tilde{x}, \quad P = \text{supp}(x - \tilde{x}), \quad |P| \leq 2n_x
\]

\[
q = \tilde{e} - e, \quad Q = \text{supp}(\tilde{e} - e), \quad |Q| \leq n_e
\]

into the uncertainty relation results in

\[
\|p\|_0\|q\|_0 = |P||Q| \geq \frac{[1 - \mu(A) |P| + 1][1 - \mu(B) (|Q| - 1)]}{\mu^2(A, B)}
\]

\[
\geq \frac{[1 - \mu(A) (2n_x - 1)] + [1 - \mu(B) (n_e - 1)]}{\mu^2(A, B)}.
\]

However, since \(\|p\|_0\|q\|_0 \leq 2n_xn_e\), this contradicts the original assumption.

In the theorem below, we state a sufficient condition for recovery via (BP) in case 3 when \(E\) is known.

**Theorem 3.5.2.** [35] Let \(z = Ax + Be\), where \(E = \text{supp}(e)\) is known. Consider the convex problem

\[
(BP, E) \quad \begin{cases}
\text{minimize} & \|\tilde{x}\|_1 \\
\text{subject to} & A\tilde{x} \in (\{z\} + R(B_e)).
\end{cases}
\]

If \(n_x = \|x\|_0\) and \(n_e = \|e\|_0\) satisfy

\[
2n_xn_e < \frac{[1 - \mu(A) (2n_x - 1)] + [1 - \mu(B) (n_e - 1)]}{\mu^2(A, B)},
\]

then the unique solution of \((BP, E)\) applied to \(z = Ax + Be\) is given by \(x\).

**Proof.** Assume that there exists an alternative solution \(\tilde{x}\) with \(\|\tilde{x}\|_1 \leq \|x\|_1\). This would imply

\[
A (x - \tilde{x}) = B (\tilde{e} - e).
\]

Set \(p = x - \tilde{x}\) and let \((\cdot)_A\) denote the projection onto the space of vectors with support \(A\), i.e., \((h_A)_l = h_l I_{l \in A}\). We find the following lower bound on the 1-norm of \(\tilde{x}\):

\[
\|\tilde{x}\|_1 = \|x - p\|_1 = \|(x - p)_x\|_1 + \|p_{x^c}\|_1
\]

\[
\geq \|x_x\|_1 - \|p_x\|_1 + \|p_{x^c}\|_1
\]

\[
= \|x\|_1 - \|p_x\|_1 + \|p_{x^c}\|_1.
\]
Therefore, \( \|\tilde{x}\|_1 \leq \|x\|_1 \) is possible only if
\[
\|P_x\|_1 \geq \|P_{x^c}\|_1,
\]
i.e., if \( p \) is \((1/2)\chi\)-concentrated (see Definition 2.3.1). Let \( q = \tilde{e} - e \), \( Q = \text{supp}(\tilde{e} - e) \subseteq E \Rightarrow |Q| \leq n_e \). Applying the uncertainty principle stated in Part 3 of Corollary 2.3.4 to \( p \) and \( q \) yields
\[
n_x n_e \geq |\mathcal{X}| |Q| \geq \frac{[(1 + \mu(A)) - |\mathcal{X}|\mu(A)]^+ [1 - \mu(B) (|Q| - 1)]^+}{\mu^2} \geq \frac{[(1 + \mu(A))/2 - |\mathcal{X}|\mu(A)]^+ [1 - \mu(B) (|Q| - 1)]^+}{\mu^2(A, B)} \geq \frac{1}{2} \frac{[1 - \mu(A) (2n_x - 1)]^+ [1 - \mu(B) (n_e - 1)]^+}{\mu^2(A, B)}.
\]
Multiplying both sides of the above inequality by a factor of two gives
\[
2n_x n_e \geq \frac{[1 - \mu(A) (2n_x - 1)]^+ [1 - \mu(B) (n_e - 1)]^+}{\mu^2(A, B)},
\]
which contradicts the original assumption
\[
2n_x n_e < \frac{[1 - \mu(A) (2n_x - 1)]^+ [1 - \mu(B) (n_e - 1)]^+}{\mu^2(A, B)}.
\]
\[\square\]

We next present an example saturating the threshold derived in the previous two theorems. For \( A = F_m \) and \( B = I_m \), this threshold becomes
\[
2n_x n_e < \frac{1}{\mu^2(I, F)} = m.
\]
Take
\[
x = \delta_{2\sqrt{m}} - \delta_{\sqrt{m}}
e = \delta_{\sqrt{m}}
\]
where \( \delta_t \) denotes the vector whose \( l \)-th entry satisfies
\[
[\delta_t]_l = \begin{cases} 
1, & \text{if } (l - 1) \mod t = 0 \\
0, & \text{else.}
\end{cases}
\]
Then, we have \( 2n_x n_e = 2\frac{\sqrt{m}}{2} \sqrt{m} = m \). One can verify by straightforward calculation that
\[
F_m \delta_t = \frac{\sqrt{m}}{t} \delta_{m/t}.
\]
We therefore have
\[ z = F_m x + e = F_m \delta_{2\sqrt{m}} + F_m \delta_{\sqrt{m}} - \delta_{\sqrt{m}} e = \frac{1}{2} \delta_{\sqrt{m}/2} - \delta_{\sqrt{m}} + \delta_{\sqrt{m}} = \frac{1}{2} \delta_{\sqrt{m}/2}. \]

Next, consider the vectors \( \tilde{x} = \delta_{2\sqrt{m}} \) and \( \tilde{e} = 0 \) with
\[ F_m \tilde{x} + I_m \tilde{e} = \frac{1}{2} \delta_{\sqrt{m}/2} = F_m x + e. \]

We thus have
\[ (F_m x) \in \{z\} + \mathcal{R}(I_m \tilde{e}) \]
\[ (F_m \tilde{x}) \in \{z\} + \mathcal{R}(I_m \tilde{e}). \]

In addition, we have \( \|x\|_0 = \|\tilde{x}\|_0 \) and \( \|x\|_1 = \|\tilde{x}\|_1 \). Therefore, \((P0, E)\) and \((BP, E)\) cannot distinguish between \(x\) and \(\tilde{x}\).

All sparsity thresholds we obtained so far are proportional to \(1/\mu(D)\). What can we say about the dictionary coherence \(\mu(D)\)?

**Theorem 3.5.3.** Let \(D \in \mathbb{C}^{m \times n}\) be a dictionary with coherence \(\mu(D)\). Then,
\[ \mu(D) \geq \sqrt{\frac{n - m}{m(n - 1)}}, \]
where \(m \leq n\).

**Proof.** Set \(G = D^H D \in \mathbb{C}^{n \times n}\). Then, \(G\) has the following properties:

1. \(G\) has ones along its diagonal (since all dictionary columns have unit \(\ell_2\) norm);

2. \(G\) is positive semi-definite with rank (at most) \(m\).

Let \(\lambda = (\lambda_1, \ldots, \lambda_m)^T\) denote the vector of nonzero eigenvalues \(\lambda_i\) of \(G\). Then, we have
\[ \text{tr } G = \sum_{i=1}^{m} \lambda_i = \|\lambda\|_1 = n \quad (3.3) \]
\[ \|G\|_2^2 = \sum_{i=1}^{m} \lambda_i^2 = \|\lambda\|_2^2. \quad (3.4) \]

Since
\[ \left( \frac{1}{m} \sum_{i=1}^{m} \lambda_i \right)^2 \leq \frac{1}{m} \sum_{i=1}^{m} \lambda_i^2 \quad (3.5) \]
by Jensen’s inequality, it follows that $\|\lambda\|_1^2 \leq m \|\lambda\|_2^2$, which implies in turn that

$$\|G\|_2^2 \geq \frac{n^2}{m}.$$ 

We thus have

$$\|G\|_2^2 = n + \sum_{i=1}^{n} \sum_{j \neq i} |\langle d_i, d_j \rangle|^2 \geq \frac{n^2}{m},$$

which finally yields

$$\mu(D)^2 \geq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} |\langle d_i, d_j \rangle|^2 \geq \frac{1}{n(n-1)} \left( \frac{n^2}{m} - n \right) = \frac{n-m}{m(n-1)}. \quad (3.6)$$

$$\mu(D)^2 \geq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} |\langle d_i, d_j \rangle|^2 \geq \frac{1}{n(n-1)} \left( \frac{n^2}{m} - n \right) = \frac{n-m}{m(n-1)}. \quad (3.7)$$

$$\mu(D)^2 \geq \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} |\langle d_i, d_j \rangle|^2 \geq \frac{1}{n(n-1)} \left( \frac{n^2}{m} - n \right) = \frac{n-m}{m(n-1)}. \quad (3.8)$$

\[ \square \]

For $n \gg m$ the Welch lower bound implies

$$\mu(D) \geq \sqrt{\frac{n-m}{m(n-1)}} \approx \frac{1}{\sqrt{m}}$$

and hence all sparsity thresholds obtained so far obey the fundamental upper bound

$$s \lesssim \sqrt{m} \Rightarrow m \gtrsim s^2.$$ 

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 $\approx 900$ samples to get recovery through (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

$$x(t) = \sum_{k=0}^{K-1} c_k \delta(t - t_k), \quad 0 \leq t_k \leq \tau$$

This signal has bandwidth $\infty$ and according to the classical sampling theorem, we would have to sample it at rate $\infty$ 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-Massys 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 $u_k = e^{-i2\pi \frac{k}{K}}$, that is

$$A(z) = \prod_{k=0}^{K-1} \left( 1 - e^{-i2\pi \frac{k}{K}} z^{-1} \right).$$
Note that this $A(z)$ is the convolution of $K$ elementary filters with impulse responses $(\delta[n] - e^{-i2\pi \frac{t_k}{\tau}} \delta[n-1])_{n \in \mathbb{Z}}$, $k = 0, 1, \ldots, K - 1$. 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:

$$
\begin{align*}
\left( e^{-i2\pi \frac{t_k}{\tau}} \ast \left( \delta[n] - e^{-i2\pi \frac{t_k}{\tau}} \delta[n-1] \right) \right)[n] &= e^{-i2\pi \frac{t_k}{\tau} n} - e^{-i2\pi \frac{t_k}{\tau} e^{-i2\pi \frac{t_k}{\tau} (n-1)}} = e^{-i2\pi \frac{t_k}{\tau} n} - e^{-i2\pi \frac{t_k}{\tau} \sum_{\tau=1}^{n} e^{-i2\pi \frac{t_k}{\tau}}} = 0, \quad \forall n \in \mathbb{Z}
\end{align*}
$$

As the Fourier series coefficients $d_n$ are linear combinations of exponentials $e^{-i2\pi \frac{t_k}{\tau} n}$, it follows that

$$(d_l)_{l \in \mathbb{Z}} \ast (a_l)_{l \in \mathbb{Z}} = 0.$$ 

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

$$
\begin{bmatrix}
\vdots \\
\vdots \\
\vdots \\
\vdots \\
\vdots \\
starrow S
\end{bmatrix}
\begin{bmatrix}
d_0 \\
d_1 \\
\vdots \\
d_K
\end{bmatrix}
= 0.
$$

To solve this linear system of equations, we need at least $2K + 1$ Fourier series coefficients, namely \{d_{-K}, \ldots, d_0, \ldots, d_K\}. In practice, this linear system of equations is solved by identifying the singular vector of $S$ 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 $\alpha_k$ 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

$$\frac{1}{\tau} \begin{bmatrix} 1 & 1 & \ldots & 1 \\ u_0 & u_1 & \ldots & u_{K-1} \\ \vdots & \vdots & \ddots & \vdots \\ u_{K-1}^K & u_{K-1}^{K-1} & \ldots & u_{K-1}^1 \end{bmatrix} \begin{bmatrix} c_0 \\ c_1 \\ \vdots \\ c_{K-1} \end{bmatrix} = \begin{bmatrix} d_0 \\ d_1 \\ \vdots \\ d_{K-1} \end{bmatrix}$$

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:

$$S = \frac{1}{\tau} \sum_{k=0}^{K-1} c_k \begin{bmatrix} 1 & u_{k}^{-1} & \ldots & u_{k}^{-K} \\ u_{k} & 1 & \ldots & u_{k}^{-K+1} \\ u_{k}^2 & u_{k} & \ldots & u_{k}^{-K+2} \\ \vdots & \vdots & \ddots & \vdots \\ u_{k}^K & u_{k}^{K-1} & \ldots & 1 \end{bmatrix}$$

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.
The individual rank-one matrices are linearly independent as the Vandermonde matrices

\[
\begin{bmatrix}
1 & 1 & \ldots & 1 \\
1 & u_1 & \ldots & u_1^K \\
1 & u_2 & \ldots & u_2^K \\
\vdots & \vdots & \ddots & \vdots \\
1 & u^K & \ldots & u^K_K
\end{bmatrix},
\begin{bmatrix}
1 & 1 & \ldots & 1 \\
1 & u^{-1}_1 & \ldots & u^{-1}_K \\
1 & u^{-2}_1 & \ldots & u^{-2}_K \\
\vdots & \vdots & \ddots & \vdots \\
1 & u^{-K}_1 & \ldots & u^{-K}_K
\end{bmatrix},
\]

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

\[ \hat{x}_d(f) \]

- \( f_s > 2f_0: \) oversampling
- \( f_s = 2f_0: \) critical sampling
- \( f_s < 2f_0: \) undersampling
5.2 Sampling Spectrally Sparse Signals

Assume that the spectrum has sparse support in \([-f_0, f_0]\), e.g.

\[
\hat{x}(f) = \begin{cases} 
  0 & \text{if } |f| > f_0 \\
  \frac{1}{4}f_0 & \text{if } f = -\frac{3f_0}{4} \\
  \frac{1}{4}f_0 & \text{if } f = \frac{3f_0}{4} \\
  0 & \text{otherwise}
\end{cases}
\]

Then two-fold undersampling, i.e., \(f_s = f_0\), results in the periodized spectrum below.

\[
\hat{x}_d(f) = \begin{cases} 
  0 & \text{if } |f| > \frac{f_0}{2} \\
  \frac{1}{4}f_0 & \text{if } f = -\frac{f_0}{4} \\
  \frac{1}{4}f_0 & \text{if } f = \frac{f_0}{4} \\
  0 & \text{otherwise}
\end{cases}
\]

Four-fold undersampling, i.e., \(f_s = f_0/2\), yields the following periodized spectrum.

\[
\hat{x}_d(f) = \begin{cases} 
  0 & \text{if } |f| > f_0 \\
  \frac{1}{4}f_0 & \text{if } f = -\frac{f_0}{4} \\
  \frac{1}{4}f_0 & \text{if } f = \frac{f_0}{4} \\
  0 & \text{otherwise}
\end{cases}
\]

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 \(I \subset [-f_0, f_0]\).
Assume the sampling set $\mathcal{P} = \{t_n\}$, i.e., we are given the signal values $\{x(t_n)\}$.

**Theorem 5.2.1** (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

$$\left| N_{t,r} - \frac{r}{T_s} \right| = |N_{t,r} - r f_s| \leq 1.$$
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
\[
\lim_{r \to \infty} \inf_{t \in \mathbb{R}} \frac{|\mathcal{P} \cap [t, t + r]|}{r} = f_s \geq |I|.
\]

### 5.2.2 Stable Sampling

**Definition 5.2.2.** A set of points \( \mathcal{P} = \{t_n\} \) is called a *stable sampling set* if for all \( x_1, x_2 \in \mathcal{H} \)
\[
A\|x_1 - x_2\|_H^2 \leq \|x_1(\mathcal{P}) - x_2(\mathcal{P})\|_2^2 \leq B\|x_1 - x_2\|_H^2
\]
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 \( x \in \mathcal{H} \)
\[
A\|x\|_H^2 \leq \|\mathbb{T}x\|_2^2 \leq B\|x\|_H^2
\]
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 \( x \in \mathcal{H} \)
\[
A\|x\|_H^2 \leq \|\mathbb{T}x\|_2^2 = \langle \mathbb{T}^* \mathbb{T} x, x \rangle = \left\langle \frac{\mathbb{T}^* \mathbb{T} x}{\mathbb{S}} \right\rangle \leq B\|x\|_H^2.
\]

We go back to sampling of multi-band signals and consider the set
\[
\mathcal{B}(\mathcal{I}) \triangleq \{ x(t) \in L^2(\mathbb{R}) : \hat{x}(f) = 0, \forall f \notin \mathcal{I} \}.
\]
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 5.2.1 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.,

$$\mathcal{F}_i = \left[ i\frac{f_0}{L}, (i + 1)\frac{f_0}{L} \right), \quad \text{for } i \in \{0, \ldots, L - 1\}.$$  

\[
\hat{x}(f)
\]

\[
\hat{x}(f)
\]

\[
f
\]

\[
f
\]

\[
f_0 \triangleq \frac{1}{T}
\]

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., $|\mathcal{I}| \approx \frac{s f_0}{L}$.

The signal $x(t)$ is sampled on a periodic nonuniform grid

$$\Psi = \Psi_1 \cup \cdots \cup \Psi_K,$$

which is the union of $K$ subgrids

$$\Psi_k = \{(mL + k)T : m \in \mathbb{Z}\}, \quad \text{for } k = 1, \ldots, K.$$

The samples corresponding to $\Psi_k$ are

$$x_k[m] \triangleq x((mL + k)T), \ m \in \mathbb{Z}.$$  

The overall sampling rate is given by (For every subgrid $\Psi_k$ the sampling rate is $1/(LT) = \frac{f_0}{L}$.)

$$D^-(\mathcal{P}) = \frac{K}{LT} = \frac{K}{L} f_0.$$
For every coset \( \{ x_k[m] \}_{m \in \mathbb{Z}} \), we compute the discrete-time Fourier transform

\[
x_d^{(k)}(f) = \sum_{m \in \mathbb{Z}} x_k[m] e^{-i2\pi fmTL} \\
= \sum_{m \in \mathbb{Z}} x((mL + k) T) e^{-i2\pi fmTL} \\
= e^{i2\pi f kT} \sum_{m \in \mathbb{Z}} x(mTL + \frac{m}{T}) e^{-i2\pi f(mTL + kT)} \\
= e^{i2\pi f kT} \frac{1}{TL} \sum_{m \in \mathbb{Z}} \hat{x}(f + \frac{m}{TL}) e^{i2\pi \frac{mk}{T}}, \quad f \in [0, 1),
\]

where the last equality follows from the Poisson summation formula

\[
\sum_{l \in \mathbb{Z}} s(t + lT) = \frac{1}{T} \sum_{l \in \mathbb{Z}} \hat{s}(\frac{l}{T}) e^{i2\pi \frac{lt}{T}}.
\]

Next, we introduce the functions \( v_k(f) := x_d^{(k)}(f)e^{-i2\pi f kT} TL \)

\[
= \sum_{m \in \mathbb{Z}} \hat{x}(f + \frac{m}{TL}) e^{i2\pi \frac{mk}{T}}, \quad f \in [0, 1/(LT)),
\]

and write

\[
\begin{bmatrix}
  v_1(f) \\
v_2(f) \\
\vdots \\
v_K(f)
\end{bmatrix}
= \begin{bmatrix}
  1 & e^{i2\pi \frac{1}{T}} & \cdots & e^{i2\pi \frac{L-1}{T}} \\
  1 & e^{i2\pi \frac{2}{T}} & \cdots & e^{i2\pi \frac{2(L-1)}{T}} \\
  \vdots & \vdots & \ddots & \vdots \\
  1 & e^{i2\pi \frac{K}{T}} & \cdots & e^{i2\pi \frac{K(L-1)}{T}}
\end{bmatrix}
\begin{bmatrix}
  \hat{x}(f) \\
  \hat{x}(f + \frac{1}{TL}) \\
  \vdots \\
  \hat{x}(f + \frac{L-1}{TL})
\end{bmatrix}, \quad f \in [0, 1/(LT)).
\]

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 \( \hat{x}(f) \) has been recovered for all \( f \in \left[ 0, \frac{1}{TL} \right] \). Noting that \([A]_{k,m} = e^{i2\pi \frac{mk}{T}} \) for \( m \in \{0, \ldots, L - 1\} \) and \( k \in \{1, \ldots, K\} \) and setting \( a_j \triangleq e^{i2\pi \frac{j}{T}} \) we can write

\[
A = \begin{bmatrix}
a_0 & a_1 & a_2 & \cdots & a_{L-1} \\
a_0^2 & a_1^2 & a_2^2 & \cdots & a_{L-1}^2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
a_0^K & a_1^K & a_2^K & \cdots & a_{L-1}^K
\end{bmatrix}.
\]
A Vandermonde matrix \( V(z_0, z_1, \ldots, z_{K-1}) \) has full rank when all \( z_i \) are distinct, where

\[
V(z_0, z_1, \ldots, z_{K-1}) = \begin{bmatrix}
1 & z_0 & z_0^2 & \cdots & z_0^{K-1} \\
1 & z_1 & z_1^2 & \cdots & z_1^{K-1} \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
1 & z_{K-1} & z_{K-1}^2 & \cdots & z_{K-1}^{K-1}
\end{bmatrix}.
\]

For every set \( S = \{s_0, s_1, \ldots, s_{K-1}\} \subset \{1, \ldots, L\} \) of cardinality \(|S| = K\), let \( A_S \) denote the submatrix which contains the columns of \( A \) indexed by \( S \). Since the rank of a Vandermonde matrix is easy to find and for every matrix \( B \) we have \( \text{rank } B = \text{rank } B^T \), it is helpful to express \( A_S^T \) in terms of a Vandermonde matrix. Take \( z_i \equiv a_{s_i} \) for \( i \in \{0, \ldots, K-1\} \) and note that

\[
\begin{bmatrix}
z_0 & z_1 & z_2 & \cdots & z_{K-1} \\
z_0^2 & z_1^2 & z_2^2 & \cdots & z_{K-1}^2 \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
z_0^K & z_1^K & z_2^K & \cdots & z_{K-1}^K
\end{bmatrix}^T = \begin{bmatrix}
z_0 & z_0^2 & z_0^3 & \cdots & z_0^K \\
z_1 & z_1^2 & z_1^3 & \cdots & z_1^K \\
\vdots & \vdots & \vdots & \ddots & \vdots \\
z_{K-1} & z_{K-1}^2 & z_{K-1}^3 & \cdots & z_{K-1}^K
\end{bmatrix}
\]

\( = \text{diag} \left( (z_0, z_1, \ldots, z_{K-1}) \right) \) full-rank

Consequently, every set of \( s \leq K \) columns of the matrix \( A \) is linearly independent as the transpose of the resulting matrix is obtained when multiplying a matrix of full-rank by a Vandermonde matrix \( V(z_0, z_1, \ldots, z_{K-1}) \) with distinct \( z_i \).
Note that the vector $\hat{x}(f)$ is sparse as illustrated by the figure above. Since we know the spectral support $I$ of the original signal $x(t)$ and hence the support set of the vector $\hat{x}(f)$, we can recover the vector $\hat{x}(f)$ (or, more precisely, the entries corresponding to nonzero components) according to

$$v(f) = A\hat{x}(f) = A_\gamma \hat{x}_\gamma(f) \Rightarrow \hat{x}_\gamma(f) = A_\gamma^\dagger v(f),$$

where $\gamma$ denotes the set of indices corresponding to nonzero entries of $\hat{x}(f)$ and $A_\gamma$ denotes the subset of columns of $A$ with indices in $\gamma$.

Given the support set of $\hat{x}(f)$, the minimum $K$ we need is $s$. This corresponds to

$$\mathcal{D}^-(\mathcal{P}) = \frac{K}{LT} \geq \frac{s}{LT} \approx |I|.$$

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 $I$ provided that the number of occupied cells $F_i$ is at most $s$.

### 5.4 Spectrum-Blind Sampling

Let us now consider the case where the support set $\gamma$ is not known a priori, but we know that $|\gamma| \leq s$. This amounts to considering the set

$$\mathcal{X}(C) = \bigcup_{|I| \leq C} B(I).$$
Recall the definition of stable sampling

\[ A\|x_1 - x_2\|^2 \leq \|Tx_1 - Tx_2\|^2 \leq B\|x_1 - x_2\|^2, \quad \forall x_1, x_2 \in \mathcal{X}(C). \]

\[ x_1 - x_2 \notin \mathcal{X}(C) \] in general but \( x_1 - x_2 \in \mathcal{X}(2C) \) so that the stable sampling condition reduces to

\[ A\|x\|^2_{\mathcal{X}(2C)} \leq \|Tx\|^2 \leq B\|x\|^2_{\mathcal{X}(2C)}, \quad \forall x \in \mathcal{X}(2C). \]

To satisfy

\[ \|Tx\|^2 \geq A\|x\|^2_{\mathcal{X}(2C)}, \quad \forall x \in \mathcal{X}(2C) \]

we obtain the necessary condition

\[ D^{-}(\mathcal{P}) \geq 2C \]

from Landau’s Theorem 5.2.1.

How about sufficiency? If \( \|\hat{x}_1(f) - \hat{x}_2(f)\|_0 \leq 2s \leq K, \forall f \in [0, \frac{1}{LT}] \), the Vandermonde structure implies that for all \( x_1 - x_2 \) multicoset sampling is stable. (Recall that spark(\( A \)) = \( K + 1 \), where the spark of a matrix denotes the minimal number of linearly dependent columns. Hence, a vector with positive \( \ell_2 \) norm can only be mapped to zero if it has at least \( \text{spark}(A) \) nonzero entries.)

\[ \|\hat{x}(f)\|_0 = s \leq \frac{K}{2} \]

implies that the cardinality of the spectral occupancy satisfies

\[ |I| = \frac{s}{LT} \leq \frac{1}{LT} \frac{K}{2} = \frac{1}{2} \frac{K}{LT} = \frac{D^{-}(\mathcal{P})}{2}. \]

Since \( |I| \leq C \) holds for all \( x \in \mathcal{X}(C) \) we find that \( D^{-}(\mathcal{P}) \geq 2C \) is also sufficient for stable sampling.
Chapter 6

The restricted isometry property and its implications for compressed sensing

We present the theory developed by E. J. Candès. Assume that we observe $y = \Phi x \in \mathbb{R}^m$, where $x \in \mathbb{R}^n$ is a signal, unknown to us and that we want to reconstruct, and $\Phi \in \mathbb{C}^{m \times n}$ is a known measurement matrix. Here, we consider the underdetermined case with fewer equations than unknowns, i.e., $m < n$.

**Definition 1.** For each integer $s = 1, 2, \ldots$, define the isometry constant $\delta_s$ of a matrix $\Phi$ as the smallest number such that

$$(1 - \delta_s) \|x\|_2^2 \leq \|\Phi x\|_2^2 \leq (1 + \delta_s) \|x\|_2^2$$

holds for all $s$-sparse vectors $x$. A vector is said to be $s$-sparse if it has at most $s$ nonzero entries.

**Theorem 6.0.1** (Noiseless recovery). Assume that $\delta_{2s} < \sqrt{2} - 1$. Then, the solution $x^*$ to

$$\minimize_{\tilde{x} \in \mathbb{R}^n} \|\tilde{x}\|_{\ell_1} \text{ subject to } \Phi \tilde{x} = y$$

obeys

$$\|x^* - x\|_{\ell_1} \leq C_0 \|x - x_s\|_{\ell_1}$$

and

$$\|x^* - x\|_{\ell_2} \leq C_0 s^{-1/2} \|x - x_s\|_{\ell_1}$$

for some constant $C_0$ given explicitly below. In particular, if $x$ is $s$-sparse, recovery is exact.

**Theorem 6.0.2** (Noisy recovery). Assume that $\delta_{2s} < \sqrt{2} - 1$ and $\|n\|_{\ell_2} \leq \varepsilon$. Then, the solution $x^*$ to

$$\minimize_{\tilde{x} \in \mathbb{R}^n} \|\tilde{x}\|_{\ell_1} \text{ subject to } \|y - \Phi \tilde{x}\|_{\ell_2} \leq \varepsilon$$

obeys

$$\|x^* - x\|_{\ell_2} \leq C_0 s^{-1/2} \|x - x_s\|_{\ell_1} + C_1 \varepsilon$$

with the same constant $C_0$ as before and some constant $C_1$ given explicitly below.
Proofs

Lemma 6.0.3. We have
\[ |\langle \Phi v, \Phi v' \rangle| \leq \delta_{s+s'} \|v\|_{\ell^2} \|v'\|_{\ell^2} \]
for all \( v, v' \) supported on disjoint subsets \( Q, Q' \subseteq \{1, \ldots, n\} \) with \( |Q| \leq s \) and \( |Q'| \leq s' \).

Proof. Without loss of generality, let us assume that \( \|v\|_{\ell^2} = \|v'\|_{\ell^2} = 1 \). By definition of the restricted isometry constant \( \delta_{s+s'} \), it holds that
\[ \|v \pm v'\|_{\ell^2}^2 (1 - \delta_{s+s'}) \leq \|\Phi(v \pm v')\|_{\ell^m}^2 \leq \|v \pm v'\|_{\ell^2}^2 (1 + \delta_{s+s'}) \]
Since \( v \) and \( v' \) are disjointly supported and \( \|v\|_{\ell^2} = \|v'\|_{\ell^2} = 1 \) by assumption, we have
\[ \|v \pm v'\|_{\ell^2}^2 = \|v\|_{\ell^2}^2 + \|v'\|_{\ell^2}^2 = 2 \]
It follows that
\[ 2(1 - \delta_{s+s'}) \leq \|\Phi(v \pm v')\|_{\ell^m}^2 \leq 2(1 + \delta_{s+s'}) \]
Applying the polarization identity
\[ \langle u, u' \rangle = \frac{1}{4} \left( \|u + u'\|_{\ell^m}^2 - \|u - u'\|_{\ell^m}^2 \right) \]
to \( u = \Phi v \) and \( u' = \Phi v' \), we obtain
\[ |\langle \Phi v, \Phi v' \rangle| = \frac{1}{4} \left| \|\Phi v + \Phi v'\|_{\ell^m}^2 - \|\Phi v - \Phi v'\|_{\ell^m}^2 \right| \]
Resolving \(|\cdot|\) such that \( \text{arg} > 0 \) yields
\[ |\langle \Phi v, \Phi v' \rangle| = \frac{1}{4} \left( \|\Phi v + \Phi v'\|_{\ell^m}^2 - \|\Phi v - \Phi v'\|_{\ell^m}^2 \right) \leq \frac{1}{4} \cdot 2(1 + \delta_{s+s'}) - \frac{1}{4} \cdot 2(1 - \delta_{s+s'}) = \frac{1}{2} (1 + \delta_{s+s'} - 1 + \delta_{s+s'}) = \delta_{s+s'} \]
Resolving \(|\cdot|\) such that \( \text{arg} < 0 \) yields
\[ |\langle \Phi v, \Phi v' \rangle| = \frac{1}{4} \left( \|\Phi v - \Phi v'\|_{\ell^m}^2 - \|\Phi v + \Phi v'\|_{\ell^m}^2 \right) \leq \frac{1}{4} \cdot 2(1 + \delta_{s+s'}) - \frac{1}{4} \cdot 2(1 - \delta_{s+s'}) = \frac{1}{2} (1 + \delta_{s+s'} - 1 + \delta_{s+s'}) = \delta_{s+s'} \]
Let us denote by $x_Q$ the vector equal to $x$ on the index set $Q$ and zero elsewhere. Let us first prove the noisy case. We start with the basic observation:

$$\|\Phi(x^* - x)\|_{\ell_2^n} \leq \|\Phi x^* - y\|_{\ell_2^n} + \|y - \Phi x\|_{\ell_2^n} \leq 2\varepsilon$$

Write $x^*$ as $x^* = x + h$, and decompose $h$ into a sum of vectors $h_{Q_0}, h_{Q_1}, \ldots$, each of sparsity at most $s$. $Q_0$ corresponds to the locations of the $s$ largest coefficients of $x$, $Q_1$ to the locations of the $s$ largest coefficients of $h_{Q_0}$ and so on. The proof proceeds in two steps:

1. the size of $h$ outside $Q_0 \cup Q_1$ is essentially bounded by that of $h$ on $Q_0 \cup Q_1$.

2. $\|h_{Q_0 \cup Q_1}\|_{\ell_2^2}$ is approximately small.

For the first step, we note that for each $j \geq 2$, we have

$$\|h_{Q_j}\|_{\ell_2^2} \leq s^{1/2} \|h_{Q_j}\|_{\ell_2^2} \leq s^{-1/2} \|h_{Q_{j-1}}\|_{\ell_2^2}$$

because $s \|h_{Q_j}\|_{\ell_2^2} \leq \|h_{Q_{j-1}}\|_{\ell_2^2}$. We therefore get

$$\sum_{j \geq 2} \|h_{Q_j}\|_{\ell_2^2} \leq s^{-1/2} \left( \|h_{Q_1}\|_{\ell_2^1} + \|h_{Q_2}\|_{\ell_2^1} + \ldots \right)$$

$$\leq s^{-1/2} \|h_{Q_0^c}\|_{\ell_2^1}. \quad (6.1)$$

The key point is that $\|h_{Q_0^c}\|_{\ell_2^1}$ cannot be very large as $\|x + h\|_{\ell_1} = \|x^*\|_{\ell_1}$ is minimum. By applying the reverse triangle inequality twice, we obtain

$$\|x\|_{\ell_1} \geq \|x + h\|_{\ell_1} = \sum_{j \in Q_0} |x_j + h_j| + \sum_{j \in Q_0^c} |x_j + h_j|$$

$$\geq \|x_{Q_0}\|_{\ell_1} - \|h_{Q_0}\|_{\ell_1} + \|h_{Q_0}\|_{\ell_2} - \|x_{Q_0^c}\|_{\ell_1}.$$
where the last inequality follows directly from (6.1). Using the fact that \( \| h_{Q_0} \|_{\ell_1^2} \leq s^{1/2} \| h_{Q_0} \|_{\ell_2^2} \), this becomes
\[
2 \| x_{Q_0^c} \|_{\ell_1^1} + s^{1/2} \| h_{Q_0} \|_{\ell_2^2} \geq s^{1/2} \| h_{(Q_0 \cup Q_1)^c} \|_{\ell_2^2}.
\]

By definition, \( x_{Q_0^c} = x - x_s \). Therefore,
\[
2 s^{-1/2} s^{1/2} \| x - x_s \|_{\ell_2^1} + \| h_{Q_0} \|_{\ell_2^2} \geq \| h_{(Q_0 \cup Q_1)^c} \|_{\ell_2^2}.
\]

Next, we bound \( \| h_{(Q_0 \cup Q_1)^c} \|_{\ell_2^2} \). We have the following
\[
\Phi h_{Q_0 \cup Q_1} = \Phi \left( h - \sum_{j \geq 2} h_{Q_j} \right) = \Phi h - \sum_{j \geq 2} \Phi h_{Q_j},
\]
which implies
\[
\| \Phi h_{Q_0 \cup Q_1} \|_{\ell_2^n}^2 = \left\langle \Phi h_{Q_0 \cup Q_1}, \Phi h \right\rangle - \left\langle \Phi h_{Q_0 \cup Q_1}, \sum_{j \geq 2} \Phi h_{Q_j} \right\rangle.
\]

But the Cauchy-Schwarz inequality gives us that
\[
|\left\langle \Phi h_{Q_0 \cup Q_1}, \Phi h \right\rangle| \leq \| \Phi h_{Q_0 \cup Q_1} \|_{\ell_2^n} \| \Phi h \|_{\ell_2^n}.
\]

Moreover, it holds that
\[
\| \Phi \left( x^* - x \right) \|_{\ell_2^n} \leq \| \Phi x^* - y \|_{\ell_2^n} + \| y - \Phi x \|_{\ell_2^n} \leq 2\varepsilon,
\]
which, combined with the definition of the restricted isometry constant, gives
\[
|\left\langle \Phi h_{Q_0 \cup Q_1}, \Phi h \right\rangle| \leq \| \Phi h_{Q_0 \cup Q_1} \|_{\ell_2^n} \cdot 2\varepsilon \leq 2\varepsilon \sqrt{1 + \delta_{2s} \| h_{Q_0 \cup Q_1} \|_{\ell_2^n}}.
\]

It follows from Lemma 6.0.3 that for all \( j \), we have
\[
|\left\langle \Phi h_{Q_0}, \Phi h_{Q_j} \right\rangle| \leq \delta_{2s} \| h_{Q_0} \|_{\ell_2^n} \| h_{Q_j} \|_{\ell_2^n}.
\]

The sets \( Q_0 \) and \( Q_1 \) are disjoint, and therefore,
\[
\| h_{Q_0} \|_{\ell_2^n} + \| h_{Q_1} \|_{\ell_2^n} \leq \sqrt{2} \| h_{Q_0 \cup Q_1} \|_{\ell_2^n}.
\]
This can be seen as follows:

\[ \| h_{Q_0 \cup Q_1} \|_{\ell_2^2}^2 = \| h_{Q_0} \|_{\ell_2^2}^2 + \| h_{Q_1} \|_{\ell_2^2}^2, \]

and we have

\[ \sqrt{2(a^2 + b^2)} \geq a + b \]
\[ 2(a^2 + b^2) \geq a^2 + b^2 + 2ab \]
\[ a^2 + b^2 - 2ab \geq 0 \]
\[ (a - b)^2 \geq 0. \]

Using the triangle inequality, (6.4), and (6.5), we obtain

\[
\left| \left| \Phi h_{Q_0 \cup Q_1} \sum_{j \geq 2} \Phi h_{Q_j} \right| \right| \leq \left| \left| \Phi h_{Q_0} \sum_{j \geq 2} \Phi h_{Q_j} \right| \right| + \left| \left| \Phi h_{Q_1} \sum_{j \geq 2} \Phi h_{Q_j} \right| \right| 
\leq \sum_{j \geq 2} \delta_{2s} \left( \| h_{Q_0} \|_{\ell_2^2} + \| h_{Q_1} \|_{\ell_2^2} \right) \| h_{Q_j} \|_{\ell_2^2} 
\leq \sqrt{2} \delta_{2s} \| h_{Q_0 \cup Q_1} \|_{\ell_2^2} \sum_{j \geq 2} \| h_{Q_j} \|_{\ell_2^2}. \tag{6.6}
\]

Combining (6.2), (6.3), (6.6) and using again the definition of the restricted isometry constant, we get

\[ (1 - \delta_{2s}) \| h_{Q_0 \cup Q_1} \|_{\ell_2^2}^2 \leq \| \Phi h_{Q_0 \cup Q_1} \|_{\ell_2^2}^2 \leq \| h_{Q_0 \cup Q_1} \|_{\ell_2^2} \left( 2\varepsilon \sqrt{1 + \delta_{2s}} + \sqrt{2} \delta_{2s} \sum_{j \geq 2} \| h_{Q_j} \|_{\ell_2^2} \right) \]

Furthermore, we have

\[ \sum_{j \geq 2} \| h_{Q_j} \|_{\ell_2^2} \leq s^{-1/2} \| h_{Q_0} \|_{\ell_1^2}, \]

which implies

\[ \| h_{Q_0 \cup Q_1} \|_{\ell_2^2} \leq \frac{2\sqrt{1 + \delta_{2s}}}{1 - \delta_{2s}} \varepsilon + \frac{\sqrt{2} \delta_{2s}}{1 - \delta_{2s}} s^{-1/2} \| h_{Q_0} \|_{\ell_1^2} = \alpha \varepsilon + \rho s^{-1/2} \| h_{Q_0} \|_{\ell_1^2}. \]

Note that we can divide by \(1 - \delta_{2s}\) because \(\delta_{2s} < 1\).

\[ \| h_{Q_0 \cup Q_1} \|_{\ell_2^2} \leq \alpha \varepsilon + \rho s^{-1/2} \| h_{Q_0} \|_{\ell_1^2} \]

Using

\[ \| h_{Q_0} \|_{\ell_1^2} \leq \| h_{Q_0} \|_{\ell_1^2} + 2 \| x_{Q_0} \|_{\ell_1^2}, \]

\[ = 2\| x - x_s \|_{\ell_1^2}. \]
this becomes
\[ \| h_{Q_0 \cup Q_1} \|_2 \leq \alpha \varepsilon + \rho \| h_{Q_0} \|_2 + \rho s^{-1/2} \| x - x_s \|_1. \]

This gives
\[ \| h_{Q_0 \cup Q_1} \|_2 \leq \alpha \varepsilon + \rho \| h_{Q_0 \cup Q_1} \|_2 + 2 \rho e_0, \]

and therefore, since we assumed that \( \delta_{2s} < \sqrt{2} - 1 \), it holds that \( 1/(1 - \rho) > 0 \), and we have
\[ \| h_{Q_0 \cup Q_1} \|_2 \leq \frac{\alpha \varepsilon + 2 \rho e_0}{1 - \rho}. \]

And finally
\[ \| h \|_2^2 = \| h_{Q_0 \cup Q_1} \|_2^2 + \| h_{(Q_0 \cup Q_1)} \|_2^2 \]
\[ = \| h_{Q_0 \cup Q_1} \|_2^2 + \| h_{(Q_0 \cup Q_1)} \|_2^2 + 2 \rho e_0 \]
\[ = 2 \| h_{Q_0 \cup Q_1} \|_2^2 + 2 \rho e_0 \leq 2 \left( \frac{\alpha \varepsilon + 2 \rho e_0}{1 - \rho} + 1 \right) \]
\[ = 2 \frac{\alpha \varepsilon + e_0 (1 + \rho)}{1 - \rho} = 2 \left( \frac{\alpha \varepsilon}{1 - \rho} + \frac{1 + \rho}{1 - \rho} \right) s^{-1/2} \| x - x_s \|_1. \]

Lemma 6.0.4. Let \( h \) be any vector in the null space of \( \Phi \) and let \( Q_0 \) be any set of cardinality \( s \).

Then,
\[ \| h_{Q_0} \|_1 \leq \rho \| h_{(Q_0 \cup Q_1)} \|_1 \]

with \( \rho = \sqrt{2} \delta_{2s} (1 - \delta_{2s}^{-1}) \).

Proof. We have
\[ \| h_{Q_0} \|_1 \leq s^{-1/2} \| h_{Q_0} \|_2 \leq s^{1/2} \| h_{Q_0 \cup Q_1} \|_2 \]
\[ \leq s^{-1/2} \rho s^{-1/2} \| h_{Q_0} \|_1 \]
\[ = \rho \| h_{Q_0} \|_1. \]

\[ \| h_{Q_0} \|_1 \leq \| h_{Q_0} \|_1 + 2 \| x_{Q_0} \|_1 \leq \frac{2}{1 - \rho} \| x_{Q_0} \|_1. \]
Therefore, in the noiseless case, we have

\[
\|h\|_{\ell^1} = \|h_{Q_0}\|_{\ell^1} + \|h_{Q_c}\|_{\ell^1} \leq (\rho + 1) \|h_{Q_0}\|_{\ell^1} \\
\leq 2 \frac{1 + \rho}{1 - \rho} \|x_{Q_0}\|_{\ell^1} \\
= 2 \frac{1 + \rho}{1 - \rho} \|x - x_s\|_{\ell^1}.
\]

\[
\square
\]
Chapter 7

The Johnson-Lindenstrauss Lemma and concentration of measure

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 $k = O(\log m/\epsilon^2)$ 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 considered in class. As a reference for these notes, see [?].

**Lemma 13** (Johnson-Lindenstrauss Lemma). Choose $\epsilon$ with $0 < \epsilon < 1$ and suppose $k$ satisfies

$$k \geq \frac{8}{\epsilon^2 - \epsilon^3} \log(2m).$$

(7.1)

Then, for every set $\mathcal{U}$ of $m$ points, there is a map $f : \mathbb{R}^n \rightarrow \mathbb{R}^k$ such that for all $u, u' \in \mathcal{U}$,

$$(1 - \epsilon) \|u - u'\|^2 \leq \|f(u) - f(u')\|^2 \leq (1 + \epsilon) \|u - u'\|^2.$$  \hspace{1cm} (7.2)

The JL Lemma is essentially tight according to [?, Thm. 9.3]. The original proof of the JL Lemma, as well as the proof discussed here, is based on random projections. Essentially, it is shown that projecting an arbitrary $m$-point subset into a random subspace only changes the inter-point distances by a factor of $1 \pm \epsilon$ with positive probability.

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 considered in class.

**Lemma 14.** Let $A \in \mathbb{R}^{k \times n}$ be a random matrix with i.i.d. $\mathcal{N}(0, 1/k)$ entries. Then, for $\epsilon$ with $0 < \epsilon < 1$ and a fixed $u \in \mathbb{R}^n$,

$$\mathbb{P} \left( \|Au\|^2 - \mathbb{E} [\|Au\|^2] \geq \epsilon \|u\|^2 \right) < 2e^{-k^2 \epsilon^2/4}.$$  \hspace{1cm} (7.3)
with
\[ E \left[ \| Au \|^2 \right] = \| u \|^2. \tag{7.4} \]

In words, Lemma 14 states that the random variable $\| Au \|^2$ is concentrated around its expectation. An equation of the form (7.3) is called “concentration of measure inequality” or simply “concentration inequality” in the literature. Note that Lemma 14 is not restricted to Gaussian random matrices, but generalizes to other random matrices. E.g., essentially the same inequality holds if each entry $a_{ij}$ of $A$ is i.i.d. sub-Gaussian, i.e., its tail probability satisfies $P(|a_{ij}| > t) \leq c_1 e^{-c_2 t^2}$ for constants $c_1, c_2$.

Before proving Lemma 14 we will show how it implies the JL Lemma.

**Proof of the JL Lemma.** The proof is effected by showing that the (linear) map $f(u) = Au$ with $A \in \mathbb{R}^{k \times n}$ a random matrix with i.i.d. $\mathcal{N}(0, 1/k)$ entries, satisfies (7.2) for all $u, u' \in U$ with non-zero probability.

Applying the union bound over all $m(m - 1)/2 < m^2$ pairs of points in $U$, it follows from Lemma 14 that (7.2) is violated for any pair of points $(u, u')$ with $u, u' \in U$ with probability less than $m^2 e^{-k^2 - \epsilon^4}$. The proof is concluded by showing that $m^2 e^{-k^2 - \epsilon^4} \leq 1/2$ is implied by (7.1), since this ensures that $f(u) = Au$ satisfies (7.2) with probability at least 1/2:

\[ m^2 e^{-k^2 - \epsilon^4} \leq 1/2 \iff -k^2 - \epsilon^4 \leq \log(1/(4m^2)) \iff k \geq \frac{4}{\epsilon^2 - \epsilon^4} \log(2m). \]

**Proof of Lemma 14.** First observe that
\[ E \left[ \| Au \|^2 \right] = E \left[ u^T A^T A u \right] = u^T E \left[ A^T A \right] u = u^T I u = \| u \|^2 \]
which proves (7.4).

Next, let $a_j^T$ be the $j$-th row of $A$, and set $X_j := \frac{\sqrt{k}}{\| u \|} a_j^T u$. Note that $a_j^T u$ is the sum of independent Gaussians and is therefore $\| u \|\mathcal{N}(0, 1/k)$ 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
\[ X = \sum_{j=1}^k X_j^2 = \frac{k}{\| u \|^2} \sum_{j=1}^k |a_j^T u|^2 = \frac{k}{\| u \|^2} \| Au \|^2. \]
Thus, for \( \lambda \geq 0 \),

\[
\mathbb{P}\left( \|A\mathbf{u}\|^2 \geq (1 + \epsilon)\|\mathbf{u}\|^2 \right) = \mathbb{P}\left( X \geq (1 + \epsilon)k \right) \\
= \mathbb{P}\left( e^{\lambda X} \geq e^{\lambda(1+\epsilon)k} \right) \\
\leq \frac{1}{e^{(1+\epsilon)k\lambda}} \mathbb{E}\left[ e^{\lambda X} \right] \\
= \frac{1}{e^{(1+\epsilon)k\lambda}} \prod_{j=1}^{k} \mathbb{E}\left[ e^{\lambda X_j^2} \right] \\
= \frac{1}{e^{(1+\epsilon)k\lambda}} \left( \mathbb{E}\left[ e^{\lambda X_1^2} \right] \right)^k
\]

(7.5)

(7.6)

(7.7)

where we used Markov’s inequality for a nonnegative random variable in (7.5), independence of the \( X_j \) for (7.6) and that all \( X_j \) have the same distribution for (7.7).

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,

\[
\mathbb{E}\left[ e^{\lambda X_1^2} \right] = \int_{-\infty}^{\infty} e^{\lambda x^2} \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}} \, dx \\
= \frac{1}{\sqrt{1 - 2\lambda}} \int_{-\infty}^{\infty} \frac{\sqrt{1 - 2\lambda}}{\sqrt{2\pi}} e^{-\frac{x^2}{2}(1-2\lambda)} \, dx \\
= \frac{1}{\sqrt{1 - 2\lambda}}
\]

(7.8)

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 (7.8) in (7.7) yields

\[
\mathbb{P}\left( \|A\mathbf{u}\|^2 \geq (1 + \epsilon)\|\mathbf{u}\|^2 \right) \leq \left( \frac{e^{-2(1+\epsilon)\lambda}}{1 - 2\lambda} \right)^{\frac{k}{2}}.
\]

(7.9)

We next minimize the right hand side (RHS) of (7.9). To this end, we choose \( \lambda \) 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,

\[
\mathbb{P}\left( \|A\mathbf{u}\|^2 \geq (1 + \epsilon)\|\mathbf{u}\|^2 \right) \leq \left( (1 + \epsilon)e^{-\epsilon} \right)^{\frac{k}{2}} < e^{-(\epsilon^2 - \epsilon^3)^{\frac{k}{4}}}
\]

(7.10)

where for the last inequality we used

\[
1 + \epsilon < e^{\epsilon^2 - \epsilon^3}
\]

which is a consequence of the Taylor expansion of \( \exp(\cdot) \).

Similarly, we obtain

\[
\mathbb{P}\left( \|A\mathbf{u}\|^2 \leq (1 - \epsilon)\|\mathbf{u}\|^2 \right) < e^{-(\epsilon^2 - \epsilon^3)^{\frac{k}{4}}}
\]

(7.11)

Combining (7.10) and (7.11) via the union bound concludes the proof. \( \square \)
Chapter 8

Verifying the RIP from concentration inequalities

We show how to prove the RIP for random matrices.

Given $S$ with $|S| \leq k$, denote by $X_S$ the set of all vectors in $\mathbb{R}^m$ that are zero outside $S$. 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 bounds, through a union bound, and then extend the result from our finite set of points to all possible $k$-dimensional signals.

Lemma 8.0.1. Let $\Phi \in \mathbb{R}^{m \times n}$ be a random matrix with i.i.d. $\mathcal{N}(0, 1/m)$ entries. Then, for any set $S$ with $|S| = k < m$ and any $0 < \delta < 1$, we have

$$(1 - \delta) \|x\|_{\ell^2} \leq \|\Phi x\|_{\ell^2} \leq (1 + \delta) \|x\|_{\ell^2}, \quad \forall x \in X_S$$

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 (8.1) for $\|x\|_{\ell^2} = 1$ as $\Phi x$ is a linear map. Choose a finite set of points $Q_S$ such that $Q_S \subseteq X_S$, $\|q\|_{\ell^2} = 1$ for all $q \in Q_S$, and for all $x \in X_S$ with $\|x\|_{\ell^2} = 1$, we have

$$\min_{q \in Q_S} \|x - q\|_{\ell^2} \leq \delta/4.$$

From the theory of covering numbers, it is known that we can choose such a set $Q_S$ with $|Q_S| \leq (12/\delta)^k$. We use the union bound to apply Lemma 2 of Chapter 7 to this set of points with $\varepsilon = \delta/2$, which yields that

$$(1 - \delta/2) \|q\|_{\ell^2}^2 \leq \|\Phi q\|_{\ell^2}^2 \leq (1 + \delta/2) \|q\|_{\ell^2}^2, \quad \forall q \in Q_S$$
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
\[ \| \Phi x \|_{\ell_2^m} \leq (1 + A) \| x \|_{\ell_2^n}, \quad \forall x \in \mathcal{X}. \]

Our goal is to show that $A \leq \delta$. To this end, recall that for any $x \in \mathcal{X}$ with $\| x \|_{\ell_2^n} = 1$, we can find a $q \in \mathcal{Q}$ such that $\| x - q \|_{\ell_2^n} \leq \delta/4$. Hence, we have
\[ \| \Phi x \|_{\ell_2^m} \leq \| \Phi q \|_{\ell_2^m} + \| \Phi (x - q) \|_{\ell_2^m} \leq 1 + \delta/2 + (1 + A)\delta/4, \]
given that $\| q \|_{\ell_2^n} = 1$. Since by definition, $A$ is the smallest number for which $\| \Phi x \|_{\ell_2^m} \leq (1 + A) \| x \|_{\ell_2^n}$, we have
\[ A \leq \delta/2 + (1 + A)\delta/4 \]
\[ A(1 - \delta/4) \leq \delta/2 + \delta/4 \]
\[ A \leq \frac{\delta/2 + \delta/4}{1 - \delta/4} = \frac{2\delta + \delta}{4 - \delta} \leq \frac{3\delta}{3} = \delta \]
as desired. We therefore proved that
\[ \| \Phi x \|_{\ell_2^m} \leq (1 + \delta) \| x \|_{\ell_2^n} \]
The inequality $\| \Phi x \|_{\ell_2^m} \geq (1 - \delta) \| x \|_{\ell_2^n}$ follows since
\[ \| \Phi x \|_{\ell_2^m} \geq \| \Phi q \|_{\ell_2^m} - \| \Phi (x - q) \|_{\ell_2^m} \geq (1 - \delta/2) - (1 + \delta)\delta/4 \]
\[ = 1 - \delta/2 - \delta/4 - \delta^2/4 \]
\[ \geq 1 - \delta/2 - \delta/4 - \delta/4 = 1 - \delta, \]
which completes the proof.

\[ \square \]

**Theorem 8.0.2.** Suppose that $m$, $n$, and $0 < \delta < 1$ are given. If the pdf generating $\Phi$ satisfies the concentration inequality in Lemma 2 of Chapter 7 then there exist constants $C_1, C_2 > 0$ depending only on $\delta$ such that the RIP holds for $\Phi$ with the prescribed $\delta$ 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 $\mathcal{X}$, the matrix $\Phi$ will fail to satisfy
\[ (1 - \delta) \| x \|_{\ell_2^n} \leq \| \Phi x \|_{\ell_2^m} \leq (1 + \delta) \| x \|_{\ell_2^n}, \quad \forall x \in \mathcal{X} \]

(8.2)
with probability
\[ \leq 2(12/\delta)^k e^{-c_0(\delta/2)m}. \] (8.3)

There are \( \binom{n}{k} \leq (en/k)^k \) such subspaces. Hence, by a union bound argument, (8.2) will fail to hold with probability
\[ \leq 2(e^n/k)^k (12/\delta)^k e^{-c_0(\delta/2)m} = 2e^{-c_0(\delta/2)m + k[\log(en/k) + \log(12/\delta)]}. \]

Thus, for a fixed \( c_1 > 0 \), whenever
\[ k \leq \frac{c_1 m}{\log(n/k)}, \]
we will have that the exponent in (8.3) is smaller than \(-c_2 n\), provided that
\[ c_2 \leq c_0 (\delta/2) - c_1 \left( 1 + \frac{1 + \log(12/\delta)}{\log(n/k)} \right). \]

This is seen as follows:
\[ e^{-c_0(\delta/2)m + k[\log(en/k) + \log(12/\delta)]} \leq e^{-c_2 m}, \]
which gives
\[ k \leq \frac{c_1 m}{\log(n/k)} \]
\[ e^{-m \left( c_0 (\delta/2) - c_1 \frac{\log(en/k) + \log(12/\delta)}{\log(n/k)} \right)} \leq e^{-c_2 m} \]
\[ c_2 \leq c_0 (\delta/2) - c_1 \left( 1 + \frac{1 + \log(12/\delta)}{\log(n/k)} \right). \]

Hence, we can always choose \( c_1 > 0 \) sufficiently small to ensure that \( c_2 > 0 \). This proves that with probability \( 1 - 2e^{-c_2 m} \) the matrix \( \Phi \) will satisfy the RIP. \( \square \)
Bibliography


