Lecture Notes for Math 416

Wojciech Czaja
Department of Mathematics
University of Maryland
College Park, MD 20742 USA

February 12, 2024

Contents

1 Preliminaries **3**
- 1.1 Number systems . . . 3
- 1.2 Complex numbers . . . 4
- 1.3 Vector Spaces . . . 5
- 1.4 Inner product space . . . 7
  - 1.4.1 Finite dimensional inner product spaces . . . 8
  - 1.4.2 The space $\ell^2(\mathbb{Z})$ . . . 10
- 1.5 Metrics and norms . . . 11
  - 1.5.1 Linear transformations and matrices . . . 14

2 Discrete Representations **15**
- 2.1 Frames . . . 15
- 2.2 Principal Component Analysis . . . 31
- 2.3 Laplacian Eigenmaps . . . 37
- 2.4 Problems . . . 42

3 Discrete Time-Frequency Analysis **44**
- 3.1 Motivation . . . 44
- 3.2 Discrete Fourier Transform . . . 47
- 3.3 Fast Fourier Transform . . . 52
- 3.4 Trigonometric Transforms . . . 56
- 3.5 Discrete Hartley Transform . . . 65
- 3.6 Problems . . . 70

4 Discrete Time-Scale Analysis **71**
- 4.1 The Discrete Haar Transform . . . 71
  - 4.1.1 The Haar functions . . . 71
  - 4.1.2 Haar wavelets on $[0, 1]$ . . . 75
  - 4.1.3 Discrete Haar Transform (DHT) . . . 77
- 4.2 Filtering Transforms . . . 82
- 4.3 Multiresolution Analysis and Filters . . . 89
- 4.4 Discrete Wavelet Transforms . . . 95
- 4.5 Problems . . . 98

5 From Discrete to Continuous Representations **99**
5.1 Lagrange Interpolation 99
5.2 Chebyshev Interpolation 101
5.3 Problems 104

Chapter 1

Preliminaries

1.1 Number systems

We begin this venture into the applied harmonic analysis with some basic notations and definitions of objects that we will use throughout this book. Among them the most fundamental object is the concept of numbers. We will use several different sets of numbers:

- $\mathbb{N} = \{0, 1, 2, 3, \ldots\}$ is the set of *natural numbers*;
- $\mathbb{Z} = \{\ldots, -3, -2, -1, 0, 1, 2, 3, \ldots\}$ is the set of *integer numbers*;
- $\mathbb{Q} = \{p/q : p, q \in \mathbb{Z}, q \neq 0\}$ is the set of *rational numbers*;
- $\mathbb{R}$ denotes the set of *real numbers*;
- $\mathbb{C} = \{a + ib : a, b \in \mathbb{R}\}$ is the set of *complex numbers*.

Associated with the concepts of sets are the notions of belonging ($x \in X$) and containment ($X \subset Y$). In this regard we notice that $\mathbb{N} \subset \mathbb{Z} \subset \mathbb{Q} \subset \mathbb{R} \subset \mathbb{C}$.

We note here that all of these sets are *infinite*. However, there are two different kinds of infinity associated with those sets. The sets $\mathbb{N}, \mathbb{Z}, \mathbb{Q}$ are *countably infinite*, whereas $\mathbb{R}$ and $\mathbb{C}$ are *uncountable*. Cantor's diagonal argument provides a beautiful proof that $\mathbb{R}$ is in fact uncountable, and hence so is $\mathbb{C}$.

The essential difference between the properties of countable and uncountable sets leads to numerous questions about our ability to use numerical computations performed on finite-state machines, such as modern computers, to model many continuous real-life phenomena. Some of those fundamental questions lead to problems which we try to address in this book. The most important problem that we shall address here is the problem of efficient data representation.

The concept of data representation has its roots in the seminal work of Joseph Fourier, who conjectured that any function or real argument can be expanded into a series of sine functions. His ideas were soon expanded into series of complex-valued

functions by Laplace and Cauchy, among many others. It is thus natural to explore the complex numbers first, before moving forward.

1.2 Complex numbers

Complex numbers play a very special role in our theory, and so we shall recall some of their fundamental properties.

Definition 1.2.1. a. A *complex number* $z \in \mathbb{C}$ is a number of the form $z = a + ib$ where, $a, b \in \mathbb{R}$ and i is the number such that

$$i^2 = -1.$$

b. We define the *conjugate* of a complex number $z = a + ib \in \mathbb{C}$ to be the complex number denoted by $\bar{z}$ and given by the formula $\bar{z} = a - ib$.

c. *Operations on complex numbers:*

Let $z_k = a_k + ib_k$ be complex numbers for $k = 1, 2$. Then,

(i) $z_1 + z_2 = z_2 + z_1 = (a_1 + ib_1) + (a_2 + ib_2) = (a_1 + a_2) + i(b_1 + b_2)$,

(ii) $z_1 z_2 = z_2 z_1 = (a_1 a_2 - b_1 b_2) + i(a_1 b_2 + a_2 b_1)$.

(iii) In particular, if $z = a + ib$ then $z\bar{z} = a^2 + b^2 \geq 0$.

d. We can use (iii) above to define the *modulus* (*absolute value* for real numbers) of z to be the nonnegative number given by

$$|z| = \sqrt{z\bar{z}} = \sqrt{a^2 + b^2}.$$

Polar form and geometric interpretation of a complex number: Every complex number has a *polar form* given by

$$z = a + ib = re^{i\theta},$$

where $r = |z| = \sqrt{a^2 + b^2}$ and θ is determined by the equations $a = r\cos\theta, b = r\sin\theta$. In particular,

$$z = a + ib = r\cos\theta + ir\sin\theta.$$

It follows that to every complex number $z = a + ib$ one can associate a point P in the $xy-$plane with coordinates $P = (a, b)$. In addition, the polar form of z is equivalent to the fact that $OP = r = |z| = \sqrt{a^2 + b^2}$ and OP makes an angle θ with the positive $x-$axis.

For each natural number n and complex number $z = re^{i\theta}$, where $\theta \in \mathbb{R}$ (although, without loss of generality we can safely assume we will work with $\theta \in [0, 2\pi)$ due to periodicity), *De Moivre's formula* states that

$$z^n = r^n e^{in\theta} = r^n(\cos n\theta + i\sin n\theta).$$

Complex numbers, together with real and rational numbers, form the simplest examples of *number fields*. Formally speaking, a field is a set $\mathbb{F}$ together with two binary operations defined on $\mathbb{F}$, called addition $+$ and multiplication $\cdot$. These operations are required to satisfy the *field axioms*. Let $a, b, c \in \mathbb{F}$. Then:

- Associativity of addition and multiplication: $a + (b + c) = (a + b) + c$, and $a \cdot (b \cdot c) = (a \cdot b) \cdot c$.
- Commutativity of addition and multiplication: $a + b = b + a$, and $a \cdot b = b \cdot a$.
- Additive and multiplicative identity: there exist two distinct elements 0 and 1 in $\mathbb{F}$ such that $a + 0 = a$ and $a \cdot 1 = a$.
- Additive inverses: for every a in $\mathbb{F}$, there exists an element in $\mathbb{F}$, denoted $-a$, called the *additive inverse of* a, such that $a + (-a) = 0$.
- Multiplicative inverses: for every $a \neq 0 \in \mathbb{F}$, there exists an element in $\mathbb{F}$, denoted by a^{-1}, called the *multiplicative inverse of* a, such that $a \cdot a^{-1} = 1$.
- Distributivity of multiplication over addition: $a \cdot (b + c) = (a \cdot b) + (a \cdot c)$.

1.3 Vector Spaces

One of the natural extensions of the idea of systems of numbers is the concept of a *vector space*.

Definition 1.3.1. A *vector space* over a number field $\mathbb{F}$ is a set V together with two operations $+$ and $\cdot$ that satisfy the eight axioms listed below:

- Associativity of addition: $u + (v + w) = (u + v) + w$;
- Commutativity of addition: $v + w = w + v$;
- Identity element of addition: $v + 0 = v$;
- Inverse elements of addition: $v + (-v) = 0$;
- Compatibility of scalar multiplication with field multiplication: $a \cdot (b \cdot v) = (ab) \cdot v$;
- Identity element of scalar multiplication: $1 \cdot v = v$;
- Distributivity of scalar multiplication with respect to vector addition: $a \cdot (u + v) = a \cdot u + a \cdot v$;
- Distributivity of scalar multiplication with respect to field addition: $(a+b) \cdot v = a \cdot v + b \cdot v$.

Elements of the vector space V are called *vectors*. The elements of the number field $\mathbb{F}$ are called *scalars*.

The first operation $+$ is called the *vector addition* and it takes any two vectors $v, w \in V$ and assigns to them a third vector which is commonly written as $v + w$, and is called the sum of these two vectors. Necessarily the sum must be an element of the vector space V. Thus $+ : V \times V \to V$.

The second operation $\cdot$ is called the *scalar multiplication* and it assigns to a pair of a scalar and a vector, another vector. Thus, $\cdot : \mathbb{F} \times V \to V$.

Example 1.3.1. Given an integer $N \geq 1$, we shall denote by $\mathbb{F}^N$, where the field is $\mathbb{F} = \mathbb{R}$ or $\mathbb{C}$, the vector space of all N-tuples of elements of $\mathbb{F}$. That is, $x \in \mathbb{F}^N$ if and only if $x = (x_1, x_2, \ldots, x_N)$ where each $x_k \in \mathbb{F}$. We equip $\mathbb{F}^N$ with two operations: coordinate-wise vector addition:

$$x + y = (x_1, x_2, \ldots, x_N) + (y_1, y_2, \ldots, y_N) = (x_1 + y_1, x_2 + y_2, \ldots, x_N + y_N) \in \mathbb{F}^N,$$

and coordinate-wise scalar multiplication by elements of $\mathbb{F}$:

$$a \cdot x = a \cdot (x_1, x_2, \ldots, x_N) = (a \cdot x_1, a \cdot x_2, \ldots, a \cdot x_N) \in \mathbb{F}^N.$$

The vector additions and scalar multiplications are used to form *linear combinations* of elements of the vector space V: $c \in V$. We say that a set of vectors is *linearly dependent* if one of the vectors in the set can be written as a linear combination of the others:

$$v_i = \sum_{j \neq i} c_j v_j.$$

Otherwise, if no vector in the set can be written in this way, then the vectors are said to be *linearly independent*. That is, a set of vectors $\{v_1, v_2, \ldots, v_n\}$ is said to be linearly independent if it is not linearly dependent, that is, if the equation

$$a_1 \cdot v_1 + \ldots + a_n \cdot v_n = 0,$$

is satisfied only by $a_i = 0$ for $i = 1, \ldots, n$.

If V is a vector space over a field $\mathbb{F}$ and if $W \subset V$ is a subset of V, then W is a *linear subspace of* V if under the addition and constant multiplication operations of V, W is a vector space over $\mathbb{F}$ in its own right.

Given a vector space V over a field $\mathbb{F}$, the *span of a set* $S \subset V$ of vectors (not necessarily a finite set) is defined to be the intersection W of all subspaces of V that contain S. W is the *subspace spanned by* S, or by the vectors in S. Conversely, S is called a *spanning set* of W, and we say that S *spans* W.

Alternatively, the span of S may be defined as the set of all finite linear combinations of elements (vectors) of S:

$$\operatorname{span}(S) = \left\{ \sum_{i=1}^{n} a_i V_i \ : \quad n \in \mathbb{N}, v_i \in S, a_i \in \mathbb{F} \right\}.$$

A set $B \subset V$ of vectors in a vector space V is called a *basis*, if every element of V may be written in a unique way as a finite linear combination of elements of B. The coefficients of this linear combination are referred to as *components (coordinates)* of the vector. The elements of a basis are called *basis vectors*. Equivalently, we can say that $B \subset V$ is a basis if its elements are linearly independent and every element of the vector space V is a linear combination of elements of B. One of our tasks will be to provide competing examples of bases and to introduce concepts which shall generalize bases.

The *dimension of a vector space* V is the cardinality, i.e., the number of elements, of any basis of V. This number is an invariant of the vector space, i.e., different bases of the same vector space must have the same cardinality. There are finite dimensional vector spaces, as well as infinite dimensional ones. This implies that some vector spaces require infinitely many vectors to span the space through finite linear combinations. In the next two sections we shall see some examples of both types, introduced in a more restrictive setting of spaces with an inner product.

1.4 Inner product space

In this section we introduce and study some fundamental properties of inner products, which are a generalization of a concept of a *dot product*.

Definition 1.4.1. We define an *inner product space* to be a vector space V over number field $\mathbb{F}$, together with an *inner product* defined as a map:

$$\langle \cdot, \cdot \rangle : V \times V \to \mathbb{F},$$

which satisfies the following 3 properties: *conjugate symmetry*, *linearity in the first argument* and *positive definiteness*:

- Conjugate symmetry: $\langle x, y \rangle = \overline{\langle y, x \rangle}$. (Note that for $\mathbb{F} = \mathbb{R}$ the conjugation is void.)
- Linearity in the first argument: $\langle ax + by, z \rangle = a\langle x, z \rangle + b\langle y, z \rangle$.
- Positive definiteness: $\langle x, x \rangle \geq 0$ and $\langle x, x \rangle = 0$ if and only if $x = 0$.

As a consequence of the above properties we have, e.g., conjugate linearity in the second argument:

$$\langle x, ay + bz \rangle = \overline{a}\langle x, y \rangle + \overline{b}\langle x, z \rangle.$$

Of particular interest among the properties of the inner product is positive definiteness, as it introduces a new quantity: $\langle x, x \rangle$. This new quantity leads us to the concept of a norm on a vector space.

Definition 1.4.2. Given a vector space V over a number field $\mathbb{F}$, a norm p on V is defined to be a nonnegative-valued function $p : V \to \mathbb{R}$ which, for any $x, y \in V$ and $a \in \mathbb{F}$, satisfies the following properties:

- Subadditivity: $p(x+y) \leq p(x) + p(y)$;
- Absolute homogeneity: $p(ax) = |a|p(x)$;
- Positive definiteness: $p(x) \geq 0$; and if $p(x) = 0$ then $x = 0$.

We then say that V equipped with norm p is a *normed vector space*.

As such, every inner product vector space is a normed vector space, with the norm of a vector, induced by the inner product, defined to be the square root of the inner product of the vector by itself:

$$p(x) = \sqrt{\langle x, x \rangle}.$$

The following inequality is often known as the *reverse triangle inequality*:

$$|p(x) - p(y)| \leq p(x - y).$$

It is of utmost importance to realize that the concept of a norm and a normed vector space is more general than the notion of an inner product and an inner product space and is typically introduced before one introduces an inner product. However, in this text we are primarily concerned with inner product vector spaces. As such it is our primary object of concern.

1.4.1 Finite dimensional inner product spaces

We recall an example of a vector space $\mathbb{F}^N$ introduced in Example 1.3.1. A set of vectors $\{e_k\}_{k=1}^N \subset \mathbb{F}^N$ is a basis for $\mathbb{F}^N$ if and only if it is a linearly independent set and spans $\mathbb{F}^N$. This is equivalent to saying that every $x \in \mathbb{F}^N$ has a unique decomposition

$$x = \sum_{k=1}^{N} c_k e_k,$$

where the coefficients $c_k \in \mathbb{F}$ are unique. We note here that it is a non-trivial result to show that the cardinality of a basis for $\mathbb{F}^N$ is indeed N. We say that $\mathbb{R}^N$ with these operations is a *real coordinate space*, and we call $\mathbb{C}^N$ a *complex coordinate space*. $\mathbb{R}^N$ and $\mathbb{C}^N$ are among the most common examples of vectors spaces. Next, following the definition of an inner product, we shall introduce some additional structure on these spaces.

We define on $\mathbb{F}^N$ the following operation

$$\langle x, y \rangle = \sum_{k=1}^{N} x_k \bar{y}_k,$$

where $x, y \in \mathbb{F}^N$. This is an *inner product*, a *scalar product*, or a *dot product* on $\mathbb{F}^N$. When equipped with this inner product, $\mathbb{F}^N$ becomes an example of an *inner*

product space. In particular, $\mathbb{R}^N$ is called *Euclidean space*. It is left as an exercise for the reader to verify that this inner product we defined, satisfies all properties in Definition 1.4.1.

In the special case of real and complex coordinate spaces defined above, the inner product gives rise to a norm on $\mathbb{F}^N$ in the following way:

$$\forall x = (x_1, x_2, \ldots, x_N) \in \mathbb{F}^N, \quad \|x\|_2 = \sqrt{\sum_{k=1}^{N} |x_k|^2} = \left(\sum_{k=1}^{N} |x_k|^2 \right)^{1/2}.$$

We note that we have used a specific notation reserved for this special norm: $\|\cdot\|_2$. Please note how the subindex 2 corresponds to the exponent in the formula.

We can now use the norm induced by the inner product to define the *Euclidean distance* on $\mathbb{F}^N$: given $x, y \in \mathbb{F}^N$, the distance between x and y is defined as

$$d(x, y) = \|x - y\|_2 = \sqrt{\sum_{k=1}^{N} |x_k - y_k|^2}.$$

Next we state explicitly two important inequalities.

Theorem 1.4.1. *For any $x, y \in \mathbb{F}^N$, we have*

a. **Schwartz inequality**

$$|\langle x, y \rangle| \leq \|x\|_2 \|y\|_2$$

with equality if and only if x or y is a multiple of the other.

b. **Triangle inequality**

$$\|x + y\|_2 \leq \|x\|_2 + \|y\|_2$$

with equality if and only if x or y is a multiple of the other.

The triangle inequality in part b is the subadditivity property for the Euclidean norm.

Orthogonality and orthonormal basis: Vectors $x, y \in V$ are said to be *orthogonal* if and only if

$$\langle x, y \rangle = 0.$$

Given a subspace $E \subset V$, its *orthogonal complement* is the subspace of V denoted by $E^\perp$ and given by

$$E^\perp = \{x \in V : \langle x, y \rangle = 0 \,\text{for all}\, y \in E\}.$$

Recall that $E \subset \mathbb{F}^N$ is a subspace if whenever $x, y \in E$ and $a \in \mathbb{F}$,

- $x + y \in E$
- $ax \in E$.

Thus, $E^\perp$ is a subspace of V whenever E is a subspace of V.

A set of vectors $\{e_k\}_{k=1}^N$ is an *orthonormal basis* (ONB) for an inner product vector space V if and only if $\{e_k\}_{k=1}^N$ is a basis and

$$\langle e_k, e_l \rangle = \delta(k-l),$$

where $\delta(j)$ is the Kronecker delta sequence equal 1 for $j=0$ and 0 otherwise. Alternatively, we can also write $\delta(k-l) = \delta_{k,l}$.

If a set of vectors $\{e_k\}_{k=1}^N \subset \mathbb{F}^N$ is an orthonormal basis, then each $x \in \mathbb{F}^N$ has the (unique) decomposition

$$x = \sum_{k=1}^{N} \langle x, e_k \rangle e_k.$$

Note that if $\{e_k\}_{k=1}^N$ is an ONB for $\mathbb{F}^N$, and if $x \in \mathbb{F}^N$ then

$$\|x\|_2^2 = \sum_{k=1}^{N} |\langle x, e_k \rangle|^2.$$

Moreover, if $x = \sum_{k=1}^N \langle x, e_k \rangle e_k$ and $y = \sum_{k=1}^N \langle y, e_k \rangle e_k$ then

$$\langle x, y \rangle = \sum_{k=1}^{N} \langle x, e_k \rangle \overline{\langle y, e_k \rangle}.$$

Orthogonal projections: Given $1 \leq M \leq N$,

$$P_M x = \sum_{k=1}^{M} \langle x, e_k \rangle e_k,$$

is the orthogonal projection of x onto $E_M = span\{e_k, k = 1, 2, \ldots, M\}$.

Given any basis $\{u_k\}_{k=1}^N$ for $\mathbb{F}^N$, there exists an algorithm, *the Gram-Schmidt* orthogonalization procedure, that transforms this basis to an ONB $\{e_k\}_{k=1}^N$. In particular,

$$e_1 = u_1/\|u_1\|_2, e_2 = \frac{u_2 - \langle u_2, e_1 \rangle e_1}{\|u_2 - \langle u_2, e_1 \rangle e_1\|_2},$$

and having constructed e_l, then,

$$e_{l+1} = \frac{u_{l+1} - \sum_{k=1}^{l} \langle u_{l+1}, e_k \rangle e_k}{\|u_{l+1} - \sum_{k=1}^{l} \langle u_{l+1}, e_k \rangle e_k\|_2}.$$

1.4.2 The space $\ell^2(\mathbb{Z})$

We now provide a classical example of an infinite dimensional inner product vector space that we shall encounter later in this text. This is a space of infinite sequences given by

$$\ell^2(\mathbb{Z}) = \{a = (a_n)_{n=-\infty}^{\infty} : \forall n \in \mathbb{Z}, \;\; a_n \in \mathbb{C}, \;\; \text{and} \;\; \sum_{n=-\infty}^{\infty} |a_n|^2 < \infty\}.$$

An inner product on $\ell^2(\mathbb{Z})$ is defined by: for $a = (a_n)_{n=-\infty}^{\infty}, b = (b_n)_{n=-\infty}^{\infty} \in \ell^2(\mathbb{Z})$ set

$$\langle a, b \rangle = \sum_{n=-\infty}^{\infty} a_n \overline{b_n}.$$

This leads to the following norm:

$$\|a\|_{\ell^2(\mathbb{Z})} = \|a\|_2 = \sqrt{\sum_{n=-\infty}^{\infty} |a_n|^2}.$$

Note that to check if a sequence $a = (a_n)_{n=-\infty}^{\infty}$ belongs to $\ell^2(\mathbb{Z})$ we must check if the series

$$\sum_{n=-\infty}^{\infty} |a_n|^2$$

converges. This is a series whose general term is nonnegative. We can appeal to the convergence theorem for nonnegative series!

1.5 Metrics and norms

A **distance** or a **metric** is a rule that defines how far from each other is any pair of elements of a set X. A set X with a metric d is called a **metric space**.

Formally, a metric is a function, defined as follows:

$$d : X \times X \to \mathbb{R},$$

with the properties:

- Identity of indiscernibles: $d(x, y) = 0 \iff x = y$.
- Symmetry: $d(x, y) = d(y, x)$.
- Subadditivity: $d(x, y) \leq d(x, z) + d(z, y)$.

Typically you will see the assumption that a metric or distance necessarily is assumed to be nonnegative, i.e.,

$$d : X \times X \to \mathbb{R}^+.$$

This is however unnecessary, as it is, in fact, a consequence of our above definition of the metric:

$$\begin{aligned} d(x, x) &\leq d(x, y) + d(y, x); \\ d(x, x) &\leq d(x, y) + d(x, y); \\ 0 &\leq 2d(x, y); \\ 0 &\leq d(x, y). \end{aligned}$$

It is important to realize that the notion of a metric can be defined on arbitrary sets. However, we observe that having defined earlier the concept of a norm on a vector space, $p : V \to \mathbb{F}$, we can easily derive a distance function from the norm, by the following formula:

$$d(x, y) := p(x - y).$$

This special metric is called a metric *induced by the norm* p. As such, normed vector spaces are a natural class of examples of metric spaces and a primary example of interest for us.

A common example of a vector space we will work with in this book is the d-dimensional coordinate vector space. This space can be equipped with many different notions of a distance. Examples of metrics include (but are not limited to) the Euclidean distance, Manhattan distance, or other ℓ_p distance metrics, and we will introduce them next. Recall that

$$d_p(x, y) = \|x - y\|_p := \left(\sum_{i=1}^{d} |x_i - y_i|^p \right)^{1/p}, \quad p \geq 1.$$

Here, we explicitly note that the p-norm of a vector $x \in \mathbb{R}^d$ is defined as:

$$\|x\|_p := \left(\sum_{i=1}^{d} |x_i|^p \right)^{1/p}.$$

When we fix $p = 2$ in the definition of the p-norm, we obtain:

$$\|x\|_2 = \sqrt{|x_1|^2 + \ldots + |x_d|^2},$$

which is the standard Euclidean norm for the d-dimensional vector space, which we introduced earlier. This norm gives rise to the following induced metric:

$$d_2(x, y) = \|x - y\|_2 := \sqrt{\sum_{i=1}^{d} |x_i|^2}.$$

In this regard we note that another interesting example of a metric is given by:

$$d_1(x, y) = \|x - y\|_1 := \sum_{i=1}^{d} |x_i - y_i|.$$

This is the Manhattan distance for the d-dimensional vector space.

Another classical example of a metric is derived from the *supremum norm* $\|x\|_\infty = \inf_{p \geq 1} \|x\|_p$:

$$d_\infty(x, y) = \|x - y\|_\infty := \sup_i |x_i - y_i|.$$

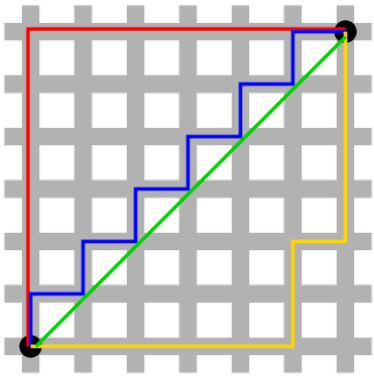

Figure 1.1: L^1 norm, also known as taxicab geometry. Do note that there are many different paths yielding the shortest distance between two points. Source of imagery: Wikipedia

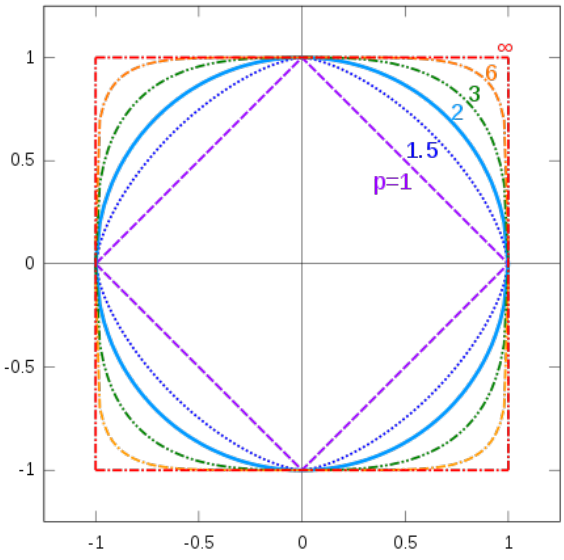

Figure 1.2: Examples of unit circles in different p-norms, $p = 1, 1.5, 2, 3, 6, \infty$. Here, a unit circle is defined as the set of points in $\mathbb{R}^2$ which are equidistant from the origin and the distance is equal to 1. The notion of distance however changes, as we change the value of p. Source of imagery: Wikipedia

1.5.1 Linear transformations and matrices

One of the most important aspects of vector spaces is the existence of a special class of transformations between such spaces, known as *linear transforms* or *linear maps*. Given two possibly different vector spaces V and W over the same field of numbers $\mathbb{F}$, we say that a mapping $f : V \to W$ is a *linear transform* if it satisfies the following two conditions:

- Additivity: for any $x, y \in V$, we have $f(x + y) = f(x) + f(y)$.
- Scalar homogeneity: for any $x \in V$ and $a \in \mathbb{F}$, we have $f(ax) = af(x)$.

In the case when V and W are finite dimensional vector spaces and a basis is defined for each of the two vector spaces, then every linear map f from V to W can be represented by a *matrix* M, where a matrix is a rectangular array or table of coefficients:

$$f(x) = Mx.$$

Some of the typical notations associated with matrices include: A^{-1} for the inverse matrix to A, A^T for the transpose of A, and $A^* = \overline{A^T}$ for the conjugation of a matrix A^T.

The family of m by n matrices will be often denoted by $Mat(m \times n)$. If we want to emphasize the set of matrix coefficients, we will sometimes write $Mat(m \times n, \mathbb{R})$ for real-valued matrices, or $Mat(m \times n, \mathbb{C})$ for complex-valued matrices.

Of particular interest in this text will be a special notion of a matrix transformation, known as the *adjoint transformation*. Consider two inner product vector spaces, $(V, \langle \cdot, \cdot \rangle_V)$ and $(W, \langle \cdot, \cdot \rangle_W)$, and a linear transformation $A : V \to W$. We define the *adjoint map* A^* to be the transformation from W to V which satisfies:

$$\langle A(x), y \rangle_W = \langle x, A^*(y) \rangle_V,$$

for every $x \in V$ and $y \in W$. Such transformation A^* always exists. In particular in the case of finite dimensional vector spaces, the adjoint can be always expressed by the following formula:

$$A^* := \overline{A^T}.$$

In real-valued spaces the adjoint is thus simply the transposition.

An important example of a linear transformation is the one which satisfies the following equality:

$$A = A^*.$$

Such transformations are called *self-adjoint*.

We close this section by introducing a concept of an *orthogonal map*, defined as a linear transformation on an inner product vector space $(V, \langle \cdot, \cdot \rangle)$ and such that

$$AA^* = AA^* = Id.$$

We observe that for finite dimensional vector spaces, $AA^* = Id$ implies $AA^* = Id$, and vice versa. This is generally not true in infinite dimensional spaces.

Chapter 2

Discrete Representations

2.1 Frames

Typical notions of various representation systems that one learns about in linear algebra, such as, e.g., Hamel bases, Schauder bases, Riesz bases, or orthonormal bases, all can be characterized by certain uniqueness of representation and minimality of the system. Although perhaps a little counter-intuitively, it turns out that such features are not always desired in practical applications. In 1952 Duffin and Schaeffer introduced a concept of redundant representations which they called *frames* [DS52]. The concept of frames was reintroduced in the context of applied harmonic analysis by Daubechies et al. [DGM86] and by Benedetto [B87] some 35 years later.

In what follows we shall often rely on the notion of a *Hilbert space*. This is a natural generalization of the concept of Euclidean space and it can be defined as a complete normed vector space with an inner product. The completeness indicates that all Cauchy sequences converge in this space. Moreover, for our purposes, the notion of a *separable* Hilbert space will be of most use. For us this means that separable Hilbert spaces are the Hilbert spaces that possess a countable basis. In view of this observation, we note that in our computations the index set I can be thought of as either finite or countably infinite.

We begin this story about generalizations and extensions of the concept of basis, by recalling one of the more intriguing properties of an orthonormal basis in a Hilbert space H. Indeed, let $\{e_i : i \in I\}$ be such an orthonormal basis. Then, by definition:

$$\langle e_i, e_j \rangle = \begin{cases} 1, & i = j, \\ 0, & i \neq j. \end{cases}$$

In this case the so called *Parseval–Plancherel formula* tells that

$$\forall\, x \in H, \quad x = \sum_{i \in I} \langle x, e_i \rangle e_i$$

and

$$\forall\, x \in H, \quad \|x\|^2 = \sum_{i\in I} |\langle x, e_i\rangle|^2. \tag{2.1}$$

We note that the norm $\|\cdot\|$ is induced by the inner product through $\|x\|^2 = \langle x, x\rangle$. The above is a very useful observation, as it allows us to observe that any separable Hilbert space H is isomorphic to the space $\ell^2(I)$, which significantly simplifies this concept of a very abstract space when presented in terms of basis expansion coefficients. In particular we note that any finite dimensional Hilbert space must be isomorphic to a Euclidean space of the corresponding dimension.

Now, in view of the importance of the Parseval–Plancherel formula, we may ask if it is necessary to have the above equality in (2.1). In other words, what if we introduce there a constant? Or even more, what if we replace the equal components with comparable ones? This is what leads us to the definition of a *frame*.

Definition 2.1.1 (Frames)**.** A *frame* for a separable Hilbert space H is a countable collection of vectors $\{f_i : i \in I\} \subset H$, such that there exist constants A and B, which satisfy $0 < A \leq B < \infty$, and such that for each $x \in H$ we have:

$$A\|x\|^2 \leq \sum_{i\in I} |\langle x, f_i\rangle|^2 \leq B\|x\|^2. \tag{2.2}$$

Constants A and B which satisfy (2.2) are called *frame bounds* of the frame $\{f_i : i \in I\}$.

Intuitively speaking Definition 2.1.1 tells us that the quantities $\sum_{i\in I} |\langle x, f_i\rangle|^2$ and $\|x\|^2$ are comparable with constants A and B, uniformly for all vectors $x \in H$. This is the desired generalization of the concept of an orthonormal basis. Before we proceed to study these systems in some detail, we shall introduce some more definitions and notations.

Given a frame $\{f_i : i \in I\} \subset H$, the optimally chosen values A and B, i.e., such that there is no $A' > A$ and no $B' < B$, for which inequalities in (2.2) also hold, are referred to as the *optimal frame bounds* of the frame $\{f_i : i \in I\} \subset H$.

When only the right inequality in (2.2) is satisfied, i.e., there exists $B > 0$ such that for all $x \in H$, we have:

$$\sum_{i\in I} |\langle x, f_i\rangle|^2 \leq B\|x\|^2, \tag{2.3}$$

then the collection $\{f_i : i \in I\} \subset H$ is referred to as a *Bessel system*. Clearly, every frame is a Bessel system. But the opposite relationship does not hold, as we can construct examples of Bessel systems which are not frames, because they do not satisfy the lower inequality:

$$A\|x\|^2 \leq \sum_{i\in I} |\langle x, f_i\rangle|^2 .$$

For the simplest example of a Bessel system, take a collection consisting of a single vector x in a space of dimension at least 2. Then, given a nonzero vector $y \perp x$, we see that $|\langle x, y\rangle|^2 = 0$, which violates the frame condition.

When $A = B$ in (2.2), the frame $\{f_i : i \in I\}$ is referred to as a *tight frame*. A tight frame with constants $A = B = 1$ is referred to as a *Parseval frame*, as it satisfies the *Parseval formula*:

$$\|x\|^2 = \sum_{i \in I} |\langle x, f_i\rangle|^2 . \tag{2.4}$$

We can also talk of a *normalized frame* $\{f_i : i \in I\} \subset H$, when for each $i \in I$, $\|f_i\| = 1$. We shall see soon in Theorem 2.1.5 that normalized Parseval frames are very closely related to orthonormal bases.

As an example of a frame one may naturally consider any orthonormal basis. Furthermore, any orthonormal basis is in fact a normalized Parseval frame, i.e., a tight frame with constants $A = B = 1$, consisting of vectors of norm 1. Moreover, a union of any two orthonormal bases is a tight frame with constants $A = B = 2$. More generally, a union of N orthonormal bases is a tight frame with constant N. To see this let $\{e_i^j : i = 1, \ldots, d\}, j = 1, \ldots, N$, be the N orthonormal bases. Then, for any $x \in \mathbb{R}^d$, we have:

$$\sum_{j=1}^{N} \sum_{i=1}^{d} |\langle x, e_i^j\rangle|^2 = \sum_{j=1}^{N} \left(\sum_{i=1}^{d} |\langle x, e_i^j\rangle|^2 \right) = \sum_{j=1}^{N} \|x\|^2 = N\|x\|^2.$$

At the same time, note that a union of several orthonormal bases is no longer an orthonormal basis. This is perhaps one of the key reasons why we need a concept like frames. Using several orthonormal bases simultaneously is perhaps not the most frugal thing to do. However, such an approach to data representation has many advantages. On the one hand, it allows us to avoid problems associated with some data loss. If in the transmission we lose one of the coefficients in the expansion in an orthonormal basis, then there is no way for us to recover the original vector from using an expansion in that specific basis. Having more than one basis is a remedy for this issue. This is what we call *redundancy*. Redundancy uses up memory and computational resources, but it has other benefits. In addition to data erasure, one typically such advantage is in noise suppression. Such constructions of redundant representation systems are abundant in engineering applications, where these systems are commonly known as *dictionaries*.

On the other hand, a union of an orthonormal basis with N arbitrarily selected unit norm vectors is a frame with bounds $A = 1$ and $B = N + 1$, but these bounds need not be optimal, depending on the choice of the N vectors. However, it is worth noting that if the Hilbert space is infinite dimensional and N is finite, then this last example is certainly not a tight frame.

Let us now demonstrate how to use the definition of a frame to compute the frame bounds directly from the definition. We will use a vector collection that aligns

with the above structures, and thus we will have an alternate approach to verifying whether a collection of vectors forms a frame or not.

Example 2.1.1. Consider 3 vectors $\{f_1 = (1,0), f_2 = (0,1), f_3 = (1,1)\} \subset \mathbb{R}^2$. For this to be a frame, we need to find lower and upper bounds for the quantity:

$$\sum_{i=1}^{3} |\langle x, f_i \rangle|^2 ,$$

in terms of the norm $\|x\|^2$ for any arbitrary vector $x \in \mathbb{R}^2$. First we rewrite this explicitly as

$$\begin{aligned}\sum_{i=1}^{3} |\langle x, f_i \rangle|^2 &= |\langle x, (1,0) \rangle|^2 + |\langle x, (0,1) \rangle|^2 + |\langle x, (1,1) \rangle|^2 \\ &= |x_1|^2 + |x_2|^2 + |x_1 + x_2|^2 \\ &= x_1^2 + x_2^2 + x_1^2 + 2x_1x_2 + x_2^2 \\ &= 2\|x\|^2 + 2x_1x_2.\end{aligned}$$

Next we estimate $2x_1x_2$ in terms of the norm of vector x. On the one hand, since $(x_1 - x_2)^2 \geq 0$, we have

$$2x_1x_2 \leq x_1^2 + x_2^2 = \|x\|^2.$$

On the other hand, since $(x_1 + x_2)^2 \geq 0$, we have

$$2x_1x_2 \geq -x_1^2 - x_2^2 = -\|x\|^2.$$

Putting all this together, we get

$$\|x\|^2 \leq \sum_{i=1}^{3} |\langle x, f_i \rangle|^2 \leq 3\|x\|^2,$$

and one can show these constants $A = 1$ and $B = 3$ are optimal. This last statement is true because the algebraic inequalities we used are sharp.

Example 2.1.2. It is very easy to modify the previous example to obtain a result matching our observation about frames formed from adding a vector to an ONB. In fact, it suffices to consider $\{f_1 = (1,0), f_2 = (0,1), f_3 = (1/\sqrt{2}, 1/\sqrt{2})\} \subset \mathbb{R}^2$. Then, calculations similar to the ones above yield:

$$\|x\|^2 \leq \sum_{i=1}^{3} |\langle x, f_i \rangle|^2 \leq 2\|x\|^2,$$

which coincides with our previous observation.

Example 2.1.3. Consider 2 vectors in $\mathbb{R}^2$: $f_1 = (1,0)$ and $f_2 = (1,1)$. Repeating the previous computations, we see that

$$\sum_{i=1}^{2} |\langle x, f_i\rangle|^2 = 2x_1^2 + 2x_1x_2 + x_2^2 = \|x\|_2^2 + 2x_1x_2 + x_1^2.$$

From this we can easily obtain an upper estimate of $3\|x\|_2^2$. But what about the lower estimate? A direct attempt to use the previous trick leads us to the following inequality:

$$\|x\|_2^2 + 2x_1x_2 + x_1^2 \geq \|x\|_2^2 - \|x\|_2^2 + x_1^2 = x_1^2.$$

That quantity, however, cannot be bounded from below by a constant multiple of $\|x\|_2^2$. It is simply not true that for an arbitrary vector in $\mathbb{R}^2$, $x_1^2 \geq c\|x\|_2^2$ for any $c > 0$.

What does that mean for us? Does it imply that $\{f_1, f_2\}$ is not a frame? No. It simply means that we need to find a different argument, because this collection happens to be a frame with constants $(3-\sqrt{5})/2$ and $(3+\sqrt{5})/2$. Try to verify that this is the case. In what follows we will seek techniques which will allow us to resolve such questions in a much more straightforward manner.

The above examples and many more similar constructions utilizing orthonormal bases, or bases in an inner product vector space, still do not take full advantage of the concept of a frame. As such, we shall develop further this theory. In this regard we note that in the Parseval-Plancherel formula, we have in fact two separate components: there is the computation of projections of the given vector $x \in H$ onto the basis elements:

$$c_n := \langle x, e_n\rangle,$$

and the linear combination in terms of the basis elements with these coefficients

$$\sum_{i\in I} c_n e_n.$$

These are two different operations, one acting on the vector space H and with values taken in the space $\ell^2(I)$, and the other one acting on the sequences in $\ell^2(I)$, with values in H. When combined, in the case of an ONB, they produce an identity transformation on H. But the situation changes when we introduce frames into this picture. In what follows we shall describe the differences between the ONB and frame representations, and the consequences of these differences. We begin with the following definition.

Definition 2.1.2. Given a frame $\{f_i : i \in I\} \subset H$, the *analysis operator* T, associated with this frame, is defined by

$$T(x) = \big\{\langle x, f_i\rangle\big\}_{i\in I}.$$

Similarly, the *synthesis operator* T^* can be defined by the following formula:

$$T^*(c) = \sum_{i \in I} c_i f_i.$$

Remark 2.1.1. We need to note here that, following the conventions from literature on this subject, we slightly abuse notation at this point in using the symbol T^* for the synthesis operator. This is because in linear algebra $*$ is reserved for the adjoint operation. However, the reason for such notation will become obvious in Theorem 2.1.1.c.

In order to give mathematical meaning to the analysis and synthesis operators, formulas alone are insufficient and we shall provide also the description of the domains of these two operations. Since we no longer have the Parseval-Plancherel formula, as was the case for ONBs, we need to use the definition of the frame in order to more precisely define the domains and ranges of the analysis and synthesis operators.

The analysis operator T clearly acts on vectors from a Hilbert space H. However, the situation is not so clear for the synthesis operator T^*, because the series $\sum_{i \in I} c_i f_i$ may not converge for some choices of coefficients $c = \{c_i\}_{i \in I}$. One simple way to guarantee such convergence would be to require that c is a finitely supported sequence in $\ell^2(I)$. This choice, however, introduces a mismatch between the range of the analysis operator and the domain of the synthesis operator: we simply do not know if finitely supported sequences will suffice to describe the range of T acting on all of H. The following theorem definitively solves our problems in this regard.

Theorem 2.1.1 (Analysis and Synthesis Operators)**.** *Let $\{f_i : i \in I\} \subset H$ be a frame for H. Then the following are satisfied:*
a. *T is a bounded operator from H into $\ell^2(I)$.*
b. *T^* extends to a bounded operator from $\ell^2(I)$ into H.*
c. *T and T^* are adjoint operators of each other.*

Proof. a. We begin by considering an arbitrary $x \in H$. We already know that T is well defined, i.e., the sequence $\{\langle x, f_i \rangle\}_{i \in I}$ consists of well-defined complex numbers. All we need to show is that $T(x) \in \ell^2(I)$. This however follows from the frame property:

$$\|\{\langle x, f_i \rangle\}_{i \in I}\|^2 = \sum_{i \in I} |\langle x, f_i \rangle|^2 \leq B\|x\|^2 < \infty,$$

or, equivalently,

$$\|\{\langle x, f_i \rangle\}_{i \in I}\|_2 \leq \sqrt{B}\|x\|.$$

Hence, we have established that T is a bounded operator from H into $\ell^2(I)$ with norm not exceeding $\sqrt{B}$.

b. Let $c = \{c_i\}_{i \in I} \in \ell^2(I)$. We first need to establish that $T^*(c)$ is well defined, i.e., that $\sum_{i \in I} c_i f_i$ converges in H. For this we shall fix $m, n \in \mathbb{N}$, $m < n$ and use

Schwartz inequality and the definition of a frame to obtain the following.

$$\begin{aligned}
\left\| \sum_{i=1}^{n} c_i f_i - \sum_{i=1}^{m} c_i f_i \right\|^2 &= \left\| \sum_{i=m+1}^{n} c_i f_i \right\|^2 = \left\langle \sum_{i=m+1}^{n} c_i f_i, \sum_{j=m+1}^{n} c_j f_j \right\rangle \\
&= \sum_{i=m+1}^{n} c_i \left\langle f_i, \sum_{j=m+1}^{n} c_j f_j \right\rangle \\
&\leq \sqrt{\sum_{i=m+1}^{n} |c_i|^2} \sqrt{\sum_{i=m+1}^{n} |\langle f_i, \sum_{j=m+1}^{n} c_j f_j \rangle|^2} \\
&\leq \sqrt{\sum_{i=m+1}^{n} |c_i|^2} \sqrt{B \left\| \sum_{i=m+1}^{n} c_i f_i \right\|^2} \\
&= \sqrt{B} \sqrt{\sum_{i=m+1}^{n} |c_i|^2} \left\| \sum_{i=m+1}^{n} c_i f_i \right\|.
\end{aligned}$$

Please note that the use of the Schwarz inequality above does not depend on whether the inner product is real or complex valued. Consequently, we have

$$\left\| \sum_{i=m+1}^{n} c_i f_i \right\| \leq \sqrt{B} \sqrt{\sum_{i=m+1}^{n} |c_i|^2}. \tag{2.5}$$

Since $c \in \ell^2(I)$, we have that $\{\sum_{i=1}^{n} |c_i|^2\}_{n\in\mathbb{N}}$ is a Cauchy sequence in $\mathbb{R}$. This observation, together with the previous estimate, allows us to conclude that we have now shown that $\{\sum_{i=1}^{n} c_i f_i\}_{n\in\mathbb{N}}$ is a Cauchy sequence in H, hence it is convergent. This fact implies that T^* is a well defined operator from $\ell^2(I)$ to H.

Moreover, a calculation similar to the one that yields (2.5), also allows us to conclude that T^* is bounded, and its norm does not exceed $\sqrt{B}$. Indeed, we can show that for all $n \in N$:

$$\left\| \sum_{i=1}^{n} c_i f_i \right\| \leq \sqrt{B} \sqrt{\sum_{i=1}^{n} |c_i|^2} \leq \sqrt{B} \|c\|_2. \tag{2.6}$$

Since the upper bound is independent of n, and the series $\sum_{i=1}^{\infty} c_i f_i$ converges in H, we are done.

c. In the context of bounded operators on a Hilbert space, we can talk about an operator and its adjoint. Recall, that for a linear transformation on an inner product vector space $M : X \to X$, this means that M^* is the conjugate transpose of M, i.e., $M^* = \overline{M^T}$. Alternatively, we can give a more general definition. Let $M : H_1 \to H_2$

be a linear transformation between two Hilbert spaces. Its adjoint operator is the linear operator $M^* : H_2 \to H_1$ which satisfies:

$$\langle Mx_1, x_2\rangle_2 = \langle x_1, M^*x_2\rangle_1.$$

It is not difficult to verify that operators A and S satisfy this adjoint relationship. Indeed,

$$\begin{aligned}\langle T(x), y\rangle_{\ell^2(I)} &= \langle\{\langle x, f_i\rangle_H\}_{i\in I}, y\rangle_{\ell^2(I)} = \sum_{i\in I}\langle x, f_i\rangle_H\overline{y}_i \\ &= \langle x, \sum_{i\in I} y_i f_i\rangle_H = \langle x, T^*(y)\rangle_H.\end{aligned}$$

□

Thus, we have now properly defined the analysis operator $T : H \to \ell^2(I)$ and we shall also identify the synthesis operator with its extension $T^* : \ell^2(I) \to H$. This way we are ready now to define the *frame operator*.

Definition 2.1.3. Given a frame $\{f_i : I \in I\} \subset H$, we define its *frame operator* to be the transformation $S = T^*T : H \to H$.

The reason we introduce the concept of the frame operator is in the definition of the frame itself. It suffices to note for now that for a given frame $\{f_i : i \in I\}$ and an arbitrary vector $x \in H$,

$$S(x) = \sum_{i\in I}\langle x, f_i\rangle f_i,$$

and thus

$$\langle x, S(x)\rangle = \left\langle x, \sum_{i\in I}\langle x, f_i\rangle f_i\right\rangle = \sum_{i\in I}\overline{\langle x, f_i\rangle}\langle x, f_i\rangle = \sum_{i\in I}|\langle x, f_i\rangle|^2.$$

Remark 2.1.2. We note here that since $\langle x, S(x)\rangle$ is nonnegative, we have in particular that $\langle x, S(x)\rangle = \langle S(x), x\rangle$.

The next result describes some of the fundamental properties of the frame operator.

Theorem 2.1.2 (Properties of Frame Operator)**.** *Let $\{f_i : i \in I\} \subset H$ be a frame for H. The frame operator $S = T^*T$ maps H onto H and is a bounded, positive, invertible operator satisfying the following:*

$$A\|x\|^2 \le \langle S(x), x\rangle \le B\|x\|^2.$$

Proof. As a composition of two bounded operators, S is also a bounded operator. Moreover,

$$\|S\| = \|T^*T\| \le \|T^*\|\|T\| \le (\sqrt{B})^2 = B.$$

Furthermore, as note before S can be explicitly written as

$$S(x) = \sum_{i \in I} \langle x, f_i \rangle f_i,$$

for every $x \in H$. Thus the frame condition (2.2), can be restated as

$$A\|x\|^2 \leq \langle S(x), x \rangle \leq B\|x\|^2.$$

This type of inequality we shall abbreviate (with a very good reason) to

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

where, e.g., $A \cdot Id \leq S$ is equivalent to the fact that $S - A \cdot Id \geq 0$, or in other words, $S - A \cdot Id$ is a positive semidefinite operator. Positivity and invertibility of S are, both, consequences of the above inequality. □

Example 2.1.4. For a finite frame $\{f_i : i = 1, \ldots, n\} \subset \mathbb{R}^d$ in a finite-dimensional Euclidean space $\mathbb{R}^d$, it is not difficult to obtain explicit formulas for the analysis, synthesis and frame operators. In fact, the analysis operator T is a $n \times d$ matrix of the form:

$$T = \begin{pmatrix} \ldots f_1 \ldots \\ \ldots f_2 \ldots \\ \ldots f_3 \ldots \\ \ldots \\ \ldots f_n \ldots \end{pmatrix},$$

and the frame operator S is a $d \times d$ matrix, for which:

$$S(i,j) = \sum_{k=1}^{n} \overline{f_k(i)} f_k(j).$$

Example 2.1.5. We recall now the frame from Example 2.1.3: $f_1 = (1, 0)$ and $f_2 = (1, 1)$. This is the simple example in which verification of the frame property was difficult and we needed a hint in the form of frame constants in order to verify that it is indeed a frame for $\mathbb{R}^2$. Now, we shall take advantage of the frame operator in order to find these constants. In this case the analysis operator is

$$T = \begin{pmatrix} 1 & 0 \\ 1 & 1 \end{pmatrix}$$

and the frame operator is equal to

$$S = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}.$$

It is now not difficult to see that the eigenvalues of S are $\lambda_1 = (3 - \sqrt{5})/2$ and $\lambda_2 = (3 + \sqrt{5})/2$. These happen to be our frame bounds. Indeed, it is not difficult

to notice that if e_1, e_2 denote the eigenvectors of S, then for any vector $x \in \mathbb{R}^2$, we have $x = \langle x, e_1\rangle e_1 + \langle x, e_2\rangle e_2$. Thus

$$S(x) = S(\langle x, e_1\rangle e_1 + \langle x, e_2\rangle e_2) = \langle x, e_1\rangle S(e_1) + \langle x, e_2\rangle S(e_2) = \langle x, e_1\rangle \lambda_1 e_1 + \langle x, e_2\rangle \lambda_2 e_2.$$

From this we calculate that

$$\begin{aligned} \langle S(x), x\rangle &= \langle x, e_1\rangle \lambda_1 \langle e_1, x\rangle + \langle x, e_2\rangle \lambda_2 \langle e_2, x\rangle \\ &= \langle x, e_1\rangle \lambda_1 \overline{\langle x, e_1\rangle} + \langle x, e_2\rangle \lambda_2 \overline{\langle x, e_2\rangle} \\ &= |\langle x, e_1\rangle|^2 \lambda_1 + |\langle x, e_2\rangle|^2 \lambda_2. \end{aligned}$$

Now, because $\lambda_1 \leq \lambda_2$, we clearly have

$$\langle S(x), x\rangle \geq |\langle x, e_1\rangle|^2 \lambda_1 + |\langle x, e_2\rangle|^2 \lambda_1 = \lambda_1 \|x\|^2$$

and

$$\langle S(x), x\rangle \leq |\langle x, e_1\rangle|^2 \lambda_2 + |\langle x, e_2\rangle|^2 \lambda_2 = \lambda_2 \|x\|^2.$$

Theorem 2.1.2 provided us with a characterization of frames in terms of the frame operator. This characterization is very useful, as demonstrated in Example 2.1.5, since it allows us to verify whether or not a collection of vectors is a frame and computing its frame bounds by studying its eigenvalue decomposition (EVD). Still, as EVDs are costly to compute, we would like to have a simpler approach. One way to achieve this is by restricting to a special class of frames which may be easier to verify. The special case when $A = B$ leads us to suspect that for tight frames we will obtain an even more explicit characterization. This is indeed the case, as the next result demonstrates.

Theorem 2.1.3 (Characterization of Tight Frames by Frame Operator)**.** *Let $\{f_i : i \in I\} \subset H$ be a frame for H. The frame operator $S = A \cdot Id$ if and only if $\{f_i : i \in I\}$ is a tight frame with constant A.*

Proof. (Sketch) ($\Rightarrow$) If $S = T^*T = AId$, then $\forall x \in H$

$$\begin{aligned} A\|x\|^2 &= A\langle x, x\rangle = \langle Ax, x\rangle = \langle Sx, x\rangle \\ &= \langle T^*Tx, x\rangle = \langle Tx, Tx\rangle \\ &= \|Tx\|^2_{l^2(I)} \\ &= \sum_{i\in I} |\langle x, f_i\rangle|^2. \end{aligned}$$

($\Leftarrow$) If $\{f_i\}_{i\in I}$ is A-tight, then $\forall x \in H$, $A\langle x, x\rangle$ is

$$A\|x\|^2 = \sum_{i\in I} |\langle x, f_i\rangle|^2 = \sum_{i\in I} \langle x, f_i\rangle\langle f_i, x\rangle = \left\langle \sum_{i\in I} \langle x, f_i\rangle f_i, x \right\rangle = \langle Sx, x\rangle.$$

Therefore,

$$\forall x \in H, \quad \langle (S - AId)x, x\rangle = 0.$$

In particular, $S - AId$ is Hermitian and positive semi-definite, so

$$\forall x, y \in H, \quad |\langle (S - AId)x, y\rangle| \leq \sqrt{\langle (S - AId)x, x\rangle \langle (S - AId)y, y\rangle} = 0.$$

Thus, $(S - AId) = 0$, so, $S = AId$. □

We note here that the inverse of the frame operator also satisfies a similar condition, $B^{-1} \cdot Id \leq S^{-1} \leq A^{-1} \cdot Id$. The meaning of this statement is rather profound: the inverse frame operator also generates a frame with constants related to the constants in the original frame. However, the proof for the infinite dimensional case is a bit more involved and we shall leave it for the interested reader to figure it out on their own.

The observation about the inverse frame operator, besides being interesting in its own right, leads us to the following new notion. Given a frame $\{f_i : i \in I\} \subset H$ in a Hilbert space H, a *dual frame* is a collection of vectors $\{f_i^* : i \in I\} \subset H$, such that for all $x \in H$, we have the following reconstruction formula:

$$x = \sum_{i \in I} \langle x, f_i\rangle f_i^*. \tag{2.7}$$

It is perhaps not immediately clear that every frame should possess such a dual frame. In order to obtain a dual frame to a given frame $\{f_i : i \in I\} \subset H$, we will have to take advantage of the *frame operator*. We shall do this in two steps.

The sequence of vectors $\{S^{-1} f_i : i \in I\}$ is called the *canonical dual frame*, and is a dual frame for $\{f_i : i \in I\}$. The reason for this new definition comes from the following simple calculation. Recall that $S(x) = \sum_{i \in I} \langle x, f_i\rangle f_i$. Write $y = S(x)$ and take advantage of invertibility of S to write

$$y = S(x) = \sum_{i \in I} \langle x, f_i\rangle f_i = \sum_{i \in I} \langle S^{-1}(y), f_i\rangle f_i.$$

Since by its definition the frame operator S is self-adjoint ($S^* = S$), the same is true for its inverse. This allows us to write:

$$y = \sum_{i \in I} \langle y, S^{-1}(f_i)\rangle f_i.$$

Alternatively, we may also obtain:

$$y = \sum_{i \in I} \langle y, f_i\rangle S^{-1}(f_i).$$

In the above both sums converge unconditionally in H. We note here that dual frames are, in general, not unique and this underlines the importance of the canonical dual frame.

For any particular given frame, it may not be easy to apply the procedure in the preceding paragraph to obtain a dual frame. As a matter of fact, finding an explicit formula for an inverse for a given operator is typically more difficult than establishing that such an inverse exists. However, one special case in which it is easy to find the dual frame is that of Parseval frames. Recall that a *Parseval frame* is a tight frame consisting with constant 1. For Parseval frames, for every $x \in H$, we have the following perfect reconstruction formula:

$$x = \sum_{i \in I} \langle x, f_i \rangle f_i. \tag{2.8}$$

This is true because according to Theorem 2.1.3, $S(x) = x$. In particular, this observation implies that Parseval frames are dual frames of themselves. For this reason, among others, Parseval frames are the best behaved of frames, and we will present here some of their additional properties.

Most of the basic, general properties of Parseval frames can be derived from the following result, originally (and in a different formulation) due to Mark Naimark.

Theorem 2.1.4. *A collection of vectors $\{f_i : i \in I\} \subset H$ is a Parseval frame for a Hilbert space H if and only if there exists a Hilbert space K containing H as a closed subspace and an orthonormal basis $\{e_i : i \in I\}$ of K such that for all $i \in I$, $Pe_i = f_i$, where P is the orthogonal projection onto H.*

Equation (2.8) follows immediately from Theorem 2.1.4. Indeed, we have for $x \in H$,

$$\begin{aligned} P^2 x &= P(Px) = \sum_{i \in I} \langle Px, e_i \rangle Pe_i \\ &= \sum_{i \in I} \langle x, Pe_i \rangle f_i \\ &= \sum_{i \in I} \langle x, f_i \rangle f_i. \end{aligned}$$

The following result has been discovered by Prof. John J. Benedetto in an unpublished manuscript of Giuseppe Vitali, held in the library of Scuola Normale Superiore di Pisa. Clearly, Vitali did *not* use the word "frame".

Theorem 2.1.5 (Vitali, 1921)**.** *Let H be a Hilbert space, $\{e_n\} \subseteq H$, $\|e_n\| = 1$.*

$$\{e_n\} \text{ is a Parseval frame} \Leftrightarrow \{e_n\} \text{ is an ONB.}$$

Proof. One direction of the proof is immediate.

For the proof of the other implication, we notice that if $\{e_n\}$ is a Parseval frame, then $\forall y \in H$

$$\|y\|^2 = \sum_n |\langle y, e_n \rangle|^2.$$

Since each $\|e_n\| = 1$, we have

$$1 = \|e_n\|^2 = \sum_k |\langle e_n, e_k\rangle|^2 = 1 + \sum_{k,k\neq n} |\langle e_n, e_k\rangle|^2$$

$$\Rightarrow \sum_{k\neq n} |\langle e_n, e_k\rangle|^2 = 0 \Rightarrow \forall n \neq k,\ \langle e_n, e_k\rangle = 0$$

□

In the situation where the explicit computation of the dual frame is impossible, there is the following useful *frame algorithm* - an iterative reconstruction method, thanks to which, we can avoid the direct computation of the dual frame. Instead we can compute an approximation to the value of $S^{-1}(f_i)$, for each element of any given frame $\{f_i : i \in I\}$.

Theorem 2.1.6 (Frame Algorithm)**.** *Let $\{f_i : i \in I\}$ be a given frame for $\mathbb{R}^d$ with known frame constants $0 < A \leq B < \infty$ and let S denote its associated frame operator. Given any $x \in \mathbb{R}^d$, we let $x_0 = \mathbb{0}$, and for $j \geq 1$ we define*

$$x_j := x_{j-1} + \frac{2}{A+B} S(x - x_{j-1}).$$

Then,

$$x = \lim_{j\to\infty} x_j$$

and

$$\|x - x_j\| \leq \left(\frac{B-A}{B+A}\right)^j \|x\|.$$

Proof. Using the frame inequality from the frame definition, we can prove that for every $x \in \mathbb{R}^d$, we have:

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

Indeed, it suffices to recall that for a frame with constants A and B, its frame operator satisfies $AId \leq S \leq BId$, which in turn is equivalent to the fact that $(A-B)Id \leq (A+B)Id - 2S \leq (B-A)Id$. And that yields the above result.

Next, we use the definition of the sequence x_i, to write

$$\begin{aligned}
x - x_j &= x - x_{j-1} - \frac{2}{A+B}S(x - x_{j-1}) \\
&= (Id - \frac{2}{A+B}S)(x - x_{j-1}) = (Id - \frac{2}{A+B}S)(x - x_{j-2} + x_{j-2} - x_{j-1}) \\
&= (Id - \frac{2}{A+B}S)(Id(x - x_{j-2}) - (x_{j-1} - x_{j-2})) \\
&= (Id - \frac{2}{A+B}S)(Id(x - x_{j-2}) - \frac{2}{A+B}S(x - x_{j-2})) \\
&= (Id - \frac{2}{A+B}S)^2(x - x_{j-2}) \\
&= \dots \\
&= (Id - \frac{2}{A+B}S)^j(x - x_0) = (Id - \frac{2}{A+B}S)^j(x).
\end{aligned}$$

Applying the norm to the above vectors, we get

$$\|x - x_j\| \leq \left(\frac{B-A}{B+A}\right)^j \|x\|,$$

as desired. □

Remark 2.1.3. We note that the statement of Theorem 2.1.6 can be generalized as follows.

Let $0 < \lambda < 2/B$ be given. Let $\delta = max(|1 - \lambda A|, |1 - \lambda B|) < 1$. Let $f_0 = 0$. Define

$$f_{n+1} = f_n + \lambda S(f - f_n).$$

Then, $\lim_{n\to\infty} f_n = f$. Moreover, the rate of convergence is geometric:

$$\|f_n - f\| \leq \delta^n \|f\|.$$

Observe that $f_1 = \lambda S(f)$ contains the frame coefficients as input.

In practice it is advisable to combine the frame algorithm with some additional acceleration methods (like, e.g., the conjugate gradient method) in order to speed up the rate of convergence.

We are now ready to use the approximation algorithm, described in Theorem 2.1.6, to compute indirectly the inverse of the frame operator S and the dual frame. As it was mentioned before, it is a difficult task for generic frames. But the frame algorithm helps us avoid these obstacles. We begin by considering $S^{-1}(f)$, where f represents an arbitrary element of the frame. Theorem 2.1.6 allows us to write

$$S^{-1}(f) = \sum_{j=0}^{\infty} \frac{2}{A+B} \sum_{k=0}^{j} (Id - \frac{2}{A+B}S)^k(f) = \lim_{j\to\infty} \frac{2}{A+B} \sum_{k=0}^{j} (Id - \frac{2}{A+B}S)^k(f).$$

Hence, the sequence of approximations of $S^{-1}(f)$ consists of vectors:

$$\begin{aligned}
&\frac{2}{A+B}f,\\
&\frac{2}{A+B}(f)+\frac{2}{A+B}(Id-\frac{2}{A+B}S)(f),\\
&\frac{2}{A+B}(f)+\frac{2}{A+B}(Id-\frac{2}{A+B}S)(f)+\frac{2}{A+B}(Id-\frac{2}{A+B}S)^2(f),\\
&\ldots.
\end{aligned}$$

We shall now simplify the above expressions, by rewriting the jth partial sum of $S^{-1}(f)$, $\frac{2}{A+B}\sum_{k=0}^{j}(Id-\frac{2}{A+B}S)^k(f)$, as

$$\frac{2}{A+B}\sum_{k=0}^{j}(Id-\frac{2}{A+B}S)^k(f)=K_j(f),$$

where K_j is defined inductively by letting $K_0=\frac{2}{A+B}Id$, and

$$K_{j+1}=\frac{2}{A+B}Id+(Id-\frac{2}{A+B}S)K_j.$$

This follows from the fact that

$$\begin{aligned}
K_{j+1}(f)&=\frac{2}{A+B}\sum_{k=0}^{j+1}(Id-\frac{2}{A+B}S)^k(f)\\
&=\frac{2}{A+B}f+\frac{2}{A+B}\sum_{k=1}^{j+1}(Id-\frac{2}{A+B}S)^k(f)\\
&=\frac{2}{A+B}f+(Id-\frac{2}{A+B}S)\frac{2}{A+B}\sum_{k=0}^{j}(Id-\frac{2}{A+B}S)^k(f)\\
&=\frac{2}{A+B}f+(Id-\frac{2}{A+B}S)K_j(f).
\end{aligned}$$

We close this Section with an observation that in finite dimensional Hilbert vector spaces, the notion of a frame becomes intuitively simple: Let $n\geq d$; $\{f_i:i=1,\ldots,n\}$ is a frame for F^d (where F denotes a field of real or complex numbers) if and only if it is a spanning system for F^d. This observation allows us to reduce the burden of studying the eigenvalue decomposition of the frame operator S to something much easier, for the purposes of establishing whether a given collection is a frame or not. Naturally this new observation does not provide us with the values of the optimal frame bounds, which is the price we pay for the simpler method.

Without loss of generality assume that $\{f_i:i=1,\ldots,n\}$ is a frame for $\mathbb{R}^d$. In order to show that this collection spans $\mathbb{R}^d$, assume by contradiction that it is not

true. Then, there exists some $x \in \mathbb{R}^d \setminus \{0\}$, such that $x \perp \text{span } \{f_i : i = 1, \ldots, n\}$. This implies in particular that $\langle x, f_i \rangle = 0$ for each $i = 1, \ldots, n$. As a consequence, $\sum_{i=1}^{n} |\langle x, f_i \rangle|^2 = 0$, which contradicts the assumption that $\{f_i : i = 1, \ldots, n\}$ is a frame.

On the other hand, let $\{f_i : i = 1, \ldots, n\}$ be a spanning set for $\mathbb{R}^d$. Since every finite collection of vectors is Bessel, it only suffices to show that $\{f_i : i = 1, \ldots, n\}$ satisfies the lower frame inequality. Consider the following function defined on the unit sphere in $\mathbb{R}^d$, $f : \mathcal{S}^{d-1} \to \mathbb{R}$:

$$f(x) = \sum_{i=1}^{n} |\langle x, f_i \rangle|^2.$$

This is a continuous function on a closed and bounded subset of $\mathbb{R}^d$, hence it has a global minimum on $\mathcal{S}^{d-1}$. Call this point x_0 and let $f(x_0) = A$, where $A \geq 0$. We claim that $A > 0$. By contradiction assume otherwise, i.e., $A = 0$. Then

$$f(x_0) = \sum_{i=1}^{n} |\langle x_0, f_i \rangle|^2 = 0,$$

which implies that $x_0 \perp f_i$, for $i = 1, \ldots, n$. But this in turn implies that $x_0 \perp \text{span } \{f_i : i = 1, \ldots, n\}$, which means that $x_0 = 0$, which contradicts our assumption that $x_0 \in \mathcal{S}^{d-1}$.

Now, using the fact that $A > 0$, we can write

$$\forall\, x \in \mathbb{R}^d, \quad f(x) = \sum_{i=1}^{n} |\langle x, f_i \rangle|^2 \geq A.$$

This can be rewritten for all $x \neq 0$ as

$$\frac{1}{\|x\|^2} \sum_{i=1}^{n} |\langle x, f_i \rangle|^2 = \sum_{i=1}^{n} \left| \left\langle \frac{x}{\|x\|}, f_i \right\rangle \right|^2 \geq A,$$

or alternatively, as

$$\sum_{i=1}^{n} |\langle x, f_i \rangle|^2 \geq A \|x\|^2,$$

which is what we wanted to obtain.

Notwithstanding the above observation, finite frames still pose many interesting problems. For example the following are true.

- If $\{x_i\}_{i=1}^{N}$ is a *finite unit norm tight frame* (FUNTF) for F^d, with frame constant A, then $A = N/d$.
- If $\{x_i\}_{i=1}^{d}$ is a A-tight frame for F^d, then it is a $\sqrt{A}$-normed orthogonal set.

2.2 Principal Component Analysis

In the standard approach to linear algebra, a basis for the given vector space is usually provided a priori and, often, forever. This approach can be quite impractical in real applications, where computations need to be made efficiently and fast. Therefore, finding an appropriate representation scheme that best describes given data is of utmost importance. There may be many different interpretations of what it means to "best describe" given data. However, most would agree that typically we would like to have as few nonzero coefficients as possible. When the number of nonzero coefficients cannot be made arbitrarily small, we would like such ideal representation to concentrate most of the weight of any given data vector on few coordinates, while distributing the rest somewhat uniformly around the remaining coordinates. Moreover, such a representation system should be easy to recompute provided that the data underwent a simple linear transformation, such as, e.g., dilation, rotation, or translation.

The first example of a method to compute efficient data-dependent representations was proposed by Karl Pearson in 1901, who proposed to use it in "physical, statistical, and biological investigations" [P01]. His idea was based on finding best fitting lines and planes, which he interpreted as those minimizing the variance of the residuals. The idea was independently discovered and developed by Harold Hotelling, who used it in applications to education and psychology. Hotelling called his method *the method of principal components*. The idea was rediscovered several more times, notably by Karhunen and Loève in the context of stochastic processes.

In view of the above, *Principal Component Analysis (PCA)* can be described as a linear transformation or a change of basis, which yields best fit to a given dataset. Once we agree on the measure of the "best" fit, the exsitence of such a basis is not as much of a problem, as is its precise computation. Thus, alternatively, we can think of PCA as the algoritheorem which provides us with the best representation.

We start by assuming that our given data comes in the form of an $m \times n$ matrix X. We can interpret the columns of the matrix X as the independent measurement vectors, and we shall denote them by $x_i \in \mathbb{R}^m$, $i = 1, \ldots, n$. We now make the most crucial assumption for the PCA to work, namely, we assume that the dataset we have at hand is a *linear* transformation of some other, ideal, representation Y. This assumption has already one immediate consequence. As we are going to utilize linear transformations, rather than affine ones, we need to assume that our data is centered at the origin. This is easily done by subtracting the mean of the data X. This is easily done by replacing each x_i with:

$$x_i(j) - \frac{1}{n}\sum_{l=1}^{n} x_l(j), \quad i = 1, \ldots, n, j = 1, \ldots m.$$

It is not difficult to see that, using the matrix notation, we replace X with

$$X - \frac{1}{n} X \mathbb{1}_n \mathbb{1}_n^T,$$

where $\mathbb{1}_n$ denotes the vector of 1s of length n. Here the mean (average) is represented by the matrix

$$\overline{X} = \frac{1}{n}\sum_{l=1}^{n} x_l = \frac{1}{n} X \mathbb{1}_n.$$

Alternatively, we may assume without loss of generality that our data X is already given in the zero mean form.

We next define the *covariance matrix* associated with X to be

$$C_{XX} = \frac{1}{n-1} X X^T.$$

This covariance matrix represents the unbiased estimator of the covariance, and is known in statistics as *sample covariance*. We note here that the above definition takes advantage of the fact that we have subtracted the average of the set of vectors $\{x_i\}$. Otherwise, we would write

$$C_{XX} = \frac{1}{n-1}(X - \overline{X}\mathbb{1}_n^T)(X - \overline{X}\mathbb{1}_n^T)^T.$$

The role of $n-1$ is a little mysterious at this point, but the explanation lies in the estimation process which is required to remove the bias from the dependence of the mean on each component. Such concerns arise in probability and statistics, where typically we assume that the vectors that we deal with are random variables, or in other words, that measurement vectors we have are in fact corrupted with some measurement noise.

We observe that C_{XX} is a square $m \times m$ matrix. This matrix is symmetric by definition. C_{XX} is also positive semidefinite. With these observations in mind, we now return to our assumption that the observed data X is, in fact, a linear transformation of some other data representation Y. We will not use just any linear transformation, but we shall require that the transformation W is orthogonal:

$$X = WY \quad \text{and} \quad W^T W = Id.$$

Thus, we note that

$$\begin{aligned} C_{XX} &= \frac{1}{n-1} X X^T \\ &= \frac{1}{n-1}(WY)(WY)^T \\ &= \frac{1}{n-1} W Y Y^T W^T \\ &= W C_{YY} W^T. \end{aligned}$$

Moreover, by orthogonality, we also have

$$C_{YY} = W^T C_{XX} W.$$

Next, we take advantage of the fact that C_{XX} is symmetric and positive semidefinite. As is well known from linear algebra, such matrices can be factored by means of their eigenvalue decompositions. Let Λ be an $m \times m$ diagonal matrix, containing all eigenvalues of C_{XX} on its diagonal. Without loss of generality we shall assume that the eigenvalues are ordered in a non-increasing way. Let V be the matrix whose columns are formed by the unit normed eigenvectors of the respective eigenvalues. The eigenvectors form an orthonormal basis for $\mathbb{R}^m$ and the eigenvalues are non-negative real numbers.

Recall now the main purpose of PCA: to find the "best" representation of the measured data. We can interpret this, as finding the representation Y (connected by means of an orthogonal transform to the measured data X) which has least correlated vectors. Because we already noticed that correlation matrices are diagonalizable, this means that we are looking for a data representation Y which has diagonal covariance matrix.

Knowing that such a diagonal representation exists is one thing, but how to compute that transformation? The recipe is to choose W defined as:

$$W = V.$$

With this in mind, we compute:

$$\begin{aligned} C_{YY} &= W^T C_{XX} W \\ &= V^T C_{XX} V \\ &= V^T V \Lambda V^T V \\ &= \Lambda. \end{aligned}$$

As such, we have identified PCA with the process of finding an eigenvalue decomposition of the covariance matrix C_{XX}.

An alternative to the above presented approach to computing PCA is to analyze the "best" representation understood as the one that minimizes the reconstruction error. Recall that we require for our consideration that

$$X = WY \quad \text{and} \quad W^T W = Id.$$

Were W a square finite matrix, the condition $W^T W = Id$ would necessarily imply that W is invertible and $W^{-1} = W^T$. However, we may easily imagine a situation where the ideal data representation Y is not of the same dimensionality as collected data X. This is, for example, true in situations, where the data depends on a smaller number of parameters than those that are collected. In such case, W is a rectangular matrix and it does not necessarily have an inverse. Also, although $W^T W = Id$, it is not necessarily true that the other equality, $WW^T = Id$, holds.

Still this simplistic assumption that W be a square invertible matrix gives us a motivation how to proceed. We can use the inverse to rewrite the equation $X = WY$ as

$$Y = W^{-1} X,$$

in order to efficiently compare the original data with its reconstructed form:

$$X - WW^{-1}X.$$

For the case of an invertible matrix this trivially equals to zero. At the same time we are typically dealing with non-square matrices and the above error is going to be our starting point.

Let W be a transformation that converts a vector $y \in \mathbb{R}^p$ into a vector $x \in \mathbb{R}^m$. In other words, W is an $m \times p$ matrix. A natural candidate for the action in the "opposite" direction, i.e., from the space of observed data to the space of its ideal representations is the *left pseudo-inverse* $W^\dagger$ defined for matrices with columns which are linearly independent as

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

Since $W^T W = Id$, we immediately note that $W^\dagger = W^T$. As such, the question of minimizing the reconstruction error that arises when we transform our data and then transform it back, can be described as the problem of minimizing the size of:

$$x - WW^\dagger x = x - WW^T x,$$

where x represents the directly observed quantity. For the measure of size, we shall use the sample mean of the Euclidean norm of the observations of $x - WW^\dagger x$, i,.e.,

$$E = \frac{1}{n} \sum_{i=1}^{n} \|x_i - WW^T x_i\|_2^2.$$

We now follow with a sequence of simple transformations of the error E:

$$\begin{aligned}
E &= \frac{1}{n} \sum_{i=1}^{n} \|x_i - WW^T x_i\|_2^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - WW^T x_i)^T (x_i - WW^T x_i) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(x_i^T x_i - 2 x_i^T WW^T x_i + x_i^T WW^T WW^T x_i \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} \left(x_i^T x_i - x_i^T WW^T x_i \right) \\
&= \frac{1}{n} \sum_{i=1}^{n} x_i^T x_i - \frac{1}{n} \sum_{i=1}^{n} x_i^T WW^T x_i \\
&= \frac{1}{n} tr(X^T X) - \frac{1}{n} tr(X^T WW^T X),
\end{aligned}$$

where $tr(A)$ denotes the trace of the square matrix A.

Since $tr(X^T X)$ is constant for given data X, it is now clear that to minimize the reconstruction error E, we need to maximize

$$\frac{1}{n} tr(X^T WW^T X),$$

over the possible choices of the orthogonal transformation W.

To maximize the last quantity, we will use a trick based on the *singular value decomposition* (SVD) of data matrix X. Recall that SVD of an $m \times n$ matrix X is a factorization of the form:

$$X = V\Sigma U^T,$$

where U and V are unitary matrices of sizes $n \times n$ and $m \times m$ resp., and where Σ is an $m \times n$ matrix with the only non-zero terms appearing on the main diagonal. Without loss of generality we may also assume that the nonzero terms are in the non-increasing order. Thus,

$$\frac{1}{n}(X^T W W^T X) = \frac{1}{n}(U\Sigma^T V^T W W^T V \Sigma U^T).$$

Recall that we are now considering a situation in which $p \leq m$. Thus, we propose to choose the orthogonal matrix W whose columns consist of the p columns of the matrix V associated with the p largest singular values of X. Thus we observe that

$$V^T W W^T V$$

is an $m \times m$ matrix with ones appearing only on the first p positions on the diagonal, and the rest of its entries are zeros. Thus,

$$\Sigma^T V^T W W^T V \Sigma = \Sigma^T \Sigma = \Sigma^2.$$

Therefore, since U is unitary, we have that

$$\frac{1}{n} tr(U\Sigma^T V^T W W^T V \Sigma U^T) = \frac{1}{n} tr(\Sigma^2).$$

Recall that our goal was to minimize E, i.e., to maximize

$$\frac{1}{n} tr(X^T W W^T X).$$

As our choice of W yields

$$\frac{1}{n} tr(X^T W W^T X) = \frac{1}{n} tr(\Sigma^2),$$

we have clearly achieved the maximal possible value by selecting the largest singular values.

We note that the restriction $p \leq m$ could possibly be overcome in the above considerations, by choosing W to be the matrix V augmented by additional columns. However, such a construction will no longer form an orthogonal transformation, which was our fundamental assumption for PCA.

To summarize what we have achieved so far in the construction of various variants of PCA, let us recall that we started by considering a covariance matrix C_{XX} and

first we have identified PCA with the process of finding an eigenvalue decomposition of the matrix C_{XX}.

Next, we have looked at the reconstruction error E, and we chose the singular value decomposition of the data matrix X as our PCA. This creates two competing constructions of PCA. Thus, now we need to show that PCA is well defined, i.e., that the two proposed definitions are equivalent to each other.

Theorem 2.2.1. *Let X be a given $m \times n$ data matrix, which we assume to have zero mean: $\overline{X} = \mathbb{0}$. Let V denote the matrix of normalized eigenvectors for the covariance C_{XX}, ordered by the magnitude of its eigenvalues. Let $\tilde{V}$ denote the matrix of left eigenvectors of a singular value decomposition of the data matrix X, analogically ordered. Then,*

$$\tilde{V} = V.$$

Proof. Let us begin by assuming the existence of a singular value decomposition for X, $X = V\Sigma U^T$. Then,

$$C_{XX} = \frac{1}{n}XX^T = \frac{1}{n}(V\Sigma U^T)(V\Sigma U^T)^T = \frac{1}{n}V\Sigma U^T U\Sigma^T V^T = V(\frac{1}{n}\Sigma\Sigma^T)V^T,$$

and consequently we obtain an eigenvalue decomposition of C_{XX}, with the matrix of eigenvectors given by the left singular vectors of X.

On the other hand, if we assume that the covariance matrix has an eigenvalue decomposition given by $C_{XX} = V\Lambda V^T$, then the left singular vectors of X are the eigenvectors of C_{XX}.

Indeed, we note that by our previous calculation, C_{XX} and $\frac{1}{n}\Sigma\Sigma^T$ are similar, thus they have the same eigenvalues with multiplicities. Therefore, $\frac{1}{n}\Sigma\Sigma^T = \Lambda$, and the left singular vectors are indeed the eigenvectors of their corresponding eigenvalues for C_{XX}, because for any SVD $X = \tilde{V}\Sigma\tilde{U}^T$, $C_{XX} = \tilde{V}\Lambda\tilde{V}^T$, which is equivalent to

$$C_{XX}\tilde{V} = \tilde{V}\Lambda.$$

This computation we did tells us that C_{XX} applied to $\tilde{V}$ produces $\tilde{V}$ times the diagonal matrix of eigenvalues of C_{XX}. But this means that the columns of $\tilde{V}$ are necessarily eigenvectors of C_{XX}. Hence $\tilde{V}$ coincides with V. □

2.3 Laplacian Eigenmaps

Principal Components Analysis algorithm, which we introduced in the previous section, is an example of an orthogonal transformation of the given data. Such transformations faithfully preserve the structure of the dataset. However, sometimes it may be advantageous to apply a transformation which enhances the given relationships. Such transformations may not necessarily be orthogonal. As a matter of fact, such transformations need not even be linear. In this Section we shall take a look at one example of such a nonlinear transformation, called Laplacian Eigenmaps. As the name suggests, this transformation is built on the foundation of the *Laplace operator* or *Laplacian*, which is a differential operator given by the divergence of the gradient field, and which is a natural generalization of the concept of second derivative to higher dimensions.

More specifically, the Laplacian is a second order differential operator, which can be defined on $\mathbb{R}^n$, by the following formula:

$$\Delta(f) = \sum_{k=1}^{n} \frac{\partial^2 f}{\partial x_i^2}.$$

Here, $\partial f/\partial x_i$ denotes the directional derivative in the ith coordinate. Specifically, for $n = 1$, the Laplacian is simply the second derivative:

$$\Delta(f) = f''.$$

Let us recall a well known formula for the computation of the second derivative [Z]:

$$f''(x_0) = \lim_{h \to 0} \frac{f(x_0 + h) + f(x_0 - h) - 2f(x_0)}{h^2}.$$

The limit on the right hand side is often called the second symmetric derivative of f at x_0. It is not difficult to observe that if the second derivative $f''(x_0)$ exists, then so does the second symmetric derivative, and they both have the same values. This is a consequence of Cauchy's Mean Value Theorem, which applied to the ration, with h being the variable, yields

$$\frac{f'(x_0 + k) - f'(x_0 - k)}{2k}$$

for some $0 < k < h$. Now, this type of limit

$$\lim_{h \to 0} \frac{g(x_0 + h) - g(x_0 - h)}{h}$$

which arises naturally here, is called the first symmetric derivative. It is easy to see that it is half of the sum of ratios:

$$\frac{g(x_0 + h) - g(x_0)}{h}$$

and

$$\frac{g(x_0) - g(x_0 - h)}{h}$$

which tend both to $f'(x_0)$, provide it exists. Thus, we complete our observation by noting that f'' is the derivative of f'. We shall now formalize this line of thought as follows.

Proposition 2.3.1. *Let f be a twice differentiable function defined on $(a, b) \subset \mathbb{R}$. Let $x \in (a, b)$. Then*

$$f''(x) = \lim_{h \to 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

The limit on the right hand side is sometimes called the *second symmetric derivative of f*. Please note that the above fact asserts that the second symmetric derivative exists provided the 2nd derivative in the classical sense exists. The opposite statement is false (please find a counterexample).

Proof:

First, consider the limit:

$$\lim_{h \to 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

We now fix x, and treat h as variable. Apply Cauchy's mean value theorem to obtain that

$$\frac{f(x+h) - 2f(x) + f(x-h)}{h^2} = \frac{f'(x+k) - f'(x-k)}{2k},$$

for some $0 < k < h$. Now, however, note that

$$\lim_{k \to 0} \frac{f'(x+k) - f'(x-k)}{2k} = f''(x).$$

To show this last fact, observe that for any differentiable function g, we have:

$$\begin{aligned} \lim_{k \to 0} \frac{g(x+k) - g(x-k)}{2k} &= \frac{1}{2} \left(\frac{g(x+k) - g(x)}{k} + \frac{g(x) - g(x-k)}{k} \right) \\ &= \frac{1}{2}(g'(x) + g'(x)) = g'(x). \end{aligned}$$

Now, complete this argument by taking $g = f'$.

Source of this proof: Antoni Zygmund, Trigonometric Series, Warszawa, 1935

Another Proof:

We can also use de l'Hopital's Rule to verify the existence of the limit:

$$\lim_{h \to 0} \frac{f(x+h) - 2f(x) + f(x-h)}{h^2}.$$

We start by noting that, both, numerator and denominator converge to 0. (Since f is differentiable, it must be in particular continuous.) Moreover, the derivative of the numerator and denominator with respect to h exist. They are $f'(x+h) - f'(x-h)$ and $2h$, respectively. This shows that the derivative of the denominator is different from 0 in the neighborhood of $h = 0$. The last assumption of de l'Hopital's Rule that remains to be checked is the existence of the limit

$$\lim_{h\to 0} \frac{f'(x+h) - f'(x-h)}{2h} = \lim_{h\to 0} \frac{f'(x+h) - f(x) + f(x) - f'(x-h)}{2h} = f''(x),$$

which follows from the existence of $f''(x)$.

Now, having defined this generalization of the 2nd derivative, we can use it to define the analog of the 2nd derivative for a function defined on $\mathbb{Z}$. In such case we let $h = 1$, note that the limit is no longer necessary for a discrete case, and we observe that

$$f''(N) = f(N+1) + f(N-1) - 2f(N), \qquad N \in \mathbb{Z}.$$

We shall now notice that both $N-1$ and $N+1$ can be viewed as neighbors of N. Denote the neighborhood of N by $nbd(N)$. This leads us to the following interpretation of 2nd derivative:

$$f''(N) = \left(\sum_{j\in nbd(N)} f(j) \right) - \left(\sum_{j\in nbd(N)} f(N) \right).$$

Observe that this definition of 2nd derivative makes sense for a function on any graph. As such we shall call it the Graph Laplacian and denote by Δ.

Definition 2.3.1 (Graphs)**.** a) Graph X is an ordered pair (V, E), where $V \subset \mathbb{R}^D$ is a collection of nodes (vertices, points), and the set E is the set of edges connecting these nodes.

b) An undirected graph is a graph in which edges have no orientation.

c) A directed graph is a graph in which edges have orientations, i.e., an edge from A to B is different from edge from B to A.

d) A simple graph is an undirected graph in which there are no multiple edges between the same pairs of nodes, and no loops from a node to itself.

e) A weighted graph is a graph in which a number (the weight) is assigned to each edge.

f) A regular graph is a graph in which each vertex has the same number of neighbors.

g) A complete graph is a graph in which each pair of vertices is joined by an edge.

h) A finite graph is a graph in which both the vertex set and the edge set are finite.

i) A connected graph is an undirected graph in which every unordered pair of vertices in the graph is connected by a continuous path.

Using the following notation,

$$W(i,j) = \begin{cases} 1 & j \in nbd(i) \\ 0 & j \notin nbd(i) \end{cases}$$

and

$$D(i,j) = \begin{cases} \sum j \in nbd(i) & i = j \\ 0 & i \neq j \end{cases}$$

we arrive at the following formula for the Graph Laplacian understood as a matrix acting on vectors which are functions on a (undirected) graph:

$$\Delta = W - D,$$

where W denotes the adjacency matrix and D denotes the degree matrix. (Please note that this is different form typical CS texts, where $\Delta = D - W$, for no good reason :))

We can also generalize now this notation to include weighted (undirected) graphs, i.e., graphs, where each edge (i,j) is assigned a number (weight) $w_{i,j}$:

$$\Delta(f)(n) = \left(\sum_{j \in nbd(n)} w_{j,n} f(j) \right) - \left(\sum_{j \in nbd(n)} w_{j,n} f(N) \right),$$

or, equivalently,

$$\Delta(m,n) = \begin{cases} w_{m,n} & m \neq n, m \in nbd(n) \\ -\sum_{j \in nbd(n)} w_{j,n} & m = n \\ 0 & m \notin nbd(n) \end{cases}.$$

Among some of the basic properties of the matrix Δ we find that it is symmetric, i.e.,

$$\Delta^T = \Delta,$$

and, provided the weights are chosen to be non-negative, the Laplacian is negative semidefinite, i.e.,

$$\forall v \in \mathbb{R}^d, \quad \langle \Delta v, v \rangle \leq 0.$$

These will come in handy when we shall want to compute eigendecompositions of Δ.

Indeed, let $f_0, \ldots, f_{N-1}$ be the eigenvectors of Δ, associated with eigenvalues $\lambda_0, \ldots, \lambda_{N-1}$. Note that N matches the size of the graph G. We may assume that $0 \geq \lambda_0 \geq \ldots \geq \lambda_{N-1}$. Moreover, it is not difficult to see that $\lambda_0 = 0$, necessarily.

Define the following embedding of your data $X = \{x_i; i = 1, \ldots, N\}$:

$$x_i \mapsto (f_1(i), \ldots f_d(i))$$

for some $d < D$. This way we achieve a dimension reduction.

We also note that

$$y^T \Delta y = \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 w_{i,j}.$$

This provides us with an interpretation of the meaning of Laplacian Eigenmaps: an embedding into $\mathbb{R}^d$ which retains the geometric structure of the original data in the sense that representations of points which were close to each other are supposed to remain close, while we are not putting any constraints on the behavior of points which were not connected.

2.4 Problems

1. We say that a collection of vectors $\{e_1, \ldots, e_n\} \subset \mathbb{R}^d$, $n \geq d$ is a *spanning set* for $\mathbb{R}^d$ if every vector in $\mathbb{R}^d$ can be represented as a linear combination of vectors $\{e_1, \ldots, e_n\}$.

We say that a collection of vectors $\{f_1, \ldots, f_n\} \subset \mathbb{R}^d$, $n \geq d$ is a *finite frame* for $\mathbb{R}^d$ if there exist constants $A, B > 0$ $(A < B)$ such that for every vector $x \in \mathbb{R}^d$ the following holds:

$$A\|x\|^2 \leq \sum_{k=1}^{n} |\langle x, f_k \rangle|^2 \leq B\|x\|^2.$$

Show that every finite spanning set for $\mathbb{R}^d$ is a finite frame for $\mathbb{R}^d$.

2. With the definitions of Problem 1, show that every finite frame $\{f_1, \ldots, f_n\} \subset \mathbb{R}^d$, for $\mathbb{R}^d$, $n \geq d$, is a spanning set for $\mathbb{R}^d$.

3. Assume $\{e_1, \ldots, e_d\} \subset \mathbb{R}^d$ is an orthonormal basis for $\mathbb{R}^d$. Is this necessarily a frame for $\mathbb{R}^d$? Prove or give a counter-example.

4. We say that an infinite collection of vectors $\{e_1, \ldots, e_n, \ldots\} \subset \mathbb{R}^d$, $n \geq d$ is a spanning set for $\mathbb{R}^d$ if every vector in $\mathbb{R}^d$ can be represented as a finite linear combination of vectors from the set $\{e_1, \ldots, e_n, \ldots\}$. We say that a collection of vectors $\{f_1, \ldots, f_n, \ldots\} \subset \mathbb{R}^d$, $n \geq d$ is a finite frame for $\mathbb{R}^d$ if there exist constants $A, B > 0 (A < B)$ such that for every vector $x \in \mathbb{R}^d$ the following holds:

$$A\|x\|_2^2 \leq \sum_{n=1}^{\infty} |\langle x, f_k \rangle|^2 \leq B\|x\|_2^2.$$

a) Are infinite spanning sets necessarily frames for $\mathbb{R}^d$? Prove or provide a counterexample.

b) Is every infinite frame necessarily a spanning set for $\mathbb{R}^d$? Prove or provide a counterexample.

5. Show that the collection of vectors $(0, 1), (\sqrt{3}/2, -1/2), (-\sqrt{3}/2, -1/2)$ in $\mathbb{R}^2$ is a tight frame (i.e., a frame with the lower frame bound A equal to the upper frame bound B). Find its frame constant.

6. Provide your own (and interesting) example of a tight frame for $\mathbb{R}^3$.

7. Show that $\{(1, 0, 0), (2, 1, 0), (3, 2, 1), (4, 3, 2)\}$, is a frame for $\mathbb{R}^3$.

8. Let A denote the matrix of an $N \times N$ DCT-II transformation. Let B denote the matrix of an $N \times N$ DST-III transformation. Let M be the matrix of a linear transformation which is defined as follows:

$$M(J, K) = \begin{cases} A(J, k) & if \quad K = 2k, \\ B(J, k) & if \quad K = 2k + 1 \end{cases}.$$

for $J = 0, \ldots, N-1$ and $K = 0, \ldots, 2N-1$.

a) Do columns of M form a frame for $\mathbb{R}^N$? If so, what are the frame constants?

b) Do the rows of M form a frame for $\mathbb{R}^{2N}$? If so, what are the frame constants?

9. We say that a frame $\{f_k : k = 1, \ldots, N\}$ in $\mathbb{R}^d$ is *equiangular* if the angles between all pairs of vectors f_i, f_j, $i \neq j$, are the same. Show an example of an equiangular tight frame in $\mathbb{R}^2$.

10. Show an example of a frame for $\mathbb{R}^d$ which is not a Riesz basis.

11. We say that a frame $\{f_k : k = 1, \ldots, N\}$ in $\mathbb{C}^d$ is equiangular if $|\langle f_i, f_j \rangle|$, $i \neq j$, are all equal to each other. Prove that the Fourier basis $\omega_n \in \mathbb{C}^N$, $n = 0, \ldots, N-1$, where $\omega_n(k) = \frac{1}{\sqrt{N}} e^{2\pi i n k/N}$, $0 \leq k < N$, is an equiangular tight frame for $\mathbb{C}^N$.

Chapter 3

Discrete Time-Frequency Analysis

3.1 Motivation

The theory of the Fourier transformation, which will dominate this section, takes its name from a French mathematician and physicist, Jean-Baptiste Joseph Fourier. Following in the footsteps of some great mathematicians, like Jean le Rond d'Alembert or Carl Friedrich Gauss, he used the concept of trigonometric series to study the heat equation, i.e., the concept of modeling the diffusion of heat through a given region. In this respect, Daniel Bernoulli and Leonhard Euler had already introduced trigonometric representations of functions. Furthermore, Joseph-Louis Lagrange had given the trigonometric series solution to the wave equation. But Fourier's contribution was the claim that an arbitrary function could be represented by a Fourier series, although we do need to note that this result is not correct in generality without some additional assumptions. This claim created a foundation for modern harmonic analysis and representation theory.

To motivate this material we start with the following example of 1D heat flow problem, which you may already know from PDEs.

Example 3.1.1. Let $u(t,x)$ represent the temperature at time $t \geq 0$ and position $x \in [0,1]$ on a piece of wire of length 1 unit. Thus, $u(t,x)$ is a function of two variable: time $t \in [0,\infty)$ and space $x \in [0,1]$. Assume that $u(t,x)$ satisfies the following equation:

$$\begin{cases} u_t(t,x) &= u_{xx}(t,x) \qquad t>0, 0 \leq x \leq 1 \\ u(0,x) &= f(x) \qquad 0 \leq x \leq 1 \\ u(t,0) &= 0 \\ u(t,1) &= 0. \end{cases}$$

Where f is a function defined on $[0,1]$, u_t is the first partial derivative of u with respect to t and $u_x x$ is the second partial derivative of u with respect to x. Find an expression for $u(t,x)$ in terms of f, x and t.

In order to solve this problem we first need to we assume that the solution $u(t,x)$ can be written as $u(t,x) = T(t)X(x)$ where T is only function of time t and X is only function of space x. By substituting this for of $u(t,x)$ in the original equation we obtain:

$$T'(t)X(x) = T(t)X''(x)$$

which is equivalent to

$$\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)} \qquad \forall t > 0, \qquad x \in [0,1].$$

This is only possible if there is a constant c such that

$$\frac{T'(t)}{T(t)} = \frac{X''(x)}{X(x)} = c \qquad \forall t > 0, \qquad x \in [0,1].$$

Now we can solve $\frac{T'(t)}{T(t)} = c$ and get $T(t) = Ce^{ct}$ for all $t > 0$. The constant c must be negative, otherwise the temperature $u(t,x)$ will grow without bound. Thus, $T(t) = Ce^{ct}$ for $t \geq 0$, and $c < 0$.

The second equation now becomes $X''(x) = cX(x)$ where $x \in [0,1]$. This leads to $X(x) = a\cos\sqrt{-c}x + b\sin\sqrt{-c}x$ for some constants a, b. However, the initial conditions now read $T(0)X(x) = f(x)$ and $T(t)X(0) = 0 = T(t)X(1)$ for all $x \in [0,1]$ and $t > 0$. Hence, $X(0) = X(1) = 0$ which implies that $a = 0$, and $b\sin\sqrt{-c} = 0$.

If $b = 0$ we will only have the trivial solution, thus $\sin\sqrt{-c} = 0$ which implies that $\sqrt{-c} = k\pi$, where $k \in \mathbb{N}$. That is, $X_k(x) = b\sin k\pi x$ and so $u(t,x) = Cbe^{-k^2\pi^2 t}\sin k\pi x$.

By the superposition principle, any solution to the above equation is given by

$$u(t,x) = \sum_{k=1}^{\infty} b_k e^{-k^2\pi^2 t}\sin k\pi x.$$

Using the last initial condition we see that

$$u(0,x) = \sum_{k=1}^{\infty} b_k \sin k\pi x = f(x).$$

So the equation will have a solution if the function f can be expressed as an infinite series :

$$f(x) = \sum_{k=1}^{\infty} b_k \sin k\pi x.$$

This is an example of a **trigonometric series**.

Furthermore, what is important is that the coefficients b_k can also be computed by similar means:

$$b_k = 2\int_0^1 f(x)\sin(k\pi x)\,dx.$$

This is an example of a **trigonometric integral transform**. Together, trigonometric series and trigonometric transforms form the foundation of time-frequency analysis.

In Example 3.1.1 we see the trigonometric series and trigonometric transform that utilize the sine function:

$$f(x) = \sum_{k=1}^{\infty} b_k \sin k\pi x \quad \text{and} \quad b_k = 2\int_0^1 f(x)\sin(k\pi x)\,dx.$$

Similarly, it is not difficult to imagine that in the context of even functions (i.e., $f(x) = (f(-x)$, which could happen with a different set of boundary values) we would require a cosine series and a cosine transform:

$$f(x) = \sum_{k=0}^{\infty} a_k \cos k\pi x \quad \text{and} \quad a_0 = \frac{1}{2}\int_0^2 f(x)\,dx, \quad a_k = \int_0^2 f(x)\cos(k\pi x)\,dx,\ k \geq 1.$$

These developments lead us to take advantage of the Euler's formula

$$e^{i\theta} = \cos(\theta) + i\sin(\theta),$$

and to consider the following functions as building blocks of signals:

$$e^{-2\pi ix} = \cos(-2\pi x) + i\sin(-2\pi x).$$

Thus, we shall consider the following **Fourier series** that combine both sine and cosine functions into the complex-valued exponential:

$$\sum_{k=-\infty}^{\infty} c_k e^{-2\pi ikx}.$$

And we shall consider the following **Fourier transform**:

$$\int_{-\infty}^{\infty} f(x)e^{-2\pi ix\xi}\,dx.$$

Unfortunately, both these objects are infinite, which makes them unsuitable for finite numerical computations. This is why we need to replace them with a finite operation. One way to introduce such a simplification would be through considering a finite integral transform $\int_0^1 f(x)e^{-2\pi ix\xi}\,dx$ and replacing it with a finite Riemann sum:

$$\sum_{n=1}^{N} f(n/N)e^{-2\pi in/N\xi_m},$$

where ξ_m is also a finite collection of values uniformly distributed over the interval. This simplification leads to the concept of a **Discrete Fourier Transform**, which we shall discuss next.

3.2 Discrete Fourier Transform

Let N be a given positive integer. We begin by defining the following $N \times N$ complex-valued matrix F_N:

$$F_N = \begin{pmatrix} 1 & 1 & 1 & \dots & 1 \\ 1 & e^{-2\pi i \frac{1}{N}} & e^{-2\pi i \frac{2}{N}} & \dots & e^{-2\pi i \frac{N-1}{N}} \\ 1 & e^{-2\pi i \frac{2}{N}} & e^{-2\pi i \frac{2\cdot 2}{N}} & \dots & e^{-2\pi i \frac{2(N-1)}{N}} \\ \vdots & \vdots & \vdots & & \vdots \\ 1 & e^{-2\pi i \frac{N-1}{N}} & e^{-2\pi i \frac{2(N-1)}{N}} & \dots & e^{-2\pi i \frac{(N-1)(N-1)}{N}} \end{pmatrix}.$$

In short, we can write the following formula for the coefficients of F_N as follows:

$$F_N(m,n) = e^{-2\pi i \frac{mn}{N}}, \quad \text{for} \quad m,n = 0,\dots,N-1.$$

We can now identify the matrix F_N with a linear transformation $F_N : \mathbb{C}^N \to \mathbb{C}^N$, which acts on a vector $v = (v(0),\dots,v(N-1)) \in \mathbb{C}^N$ as:

$$F_N(v)(k) = \sum_{j=0}^{N-1} v(j) e^{-2\pi i \frac{jk}{N}}, \quad \text{for} \quad k = 0,\dots,N-1.$$

One of the natural questions regarding any matrix is finding its determinant and verifying whether or not the transformation is invertible.

Proposition 3.2.1. *The transformation $F_N : \mathbb{C}^N \to \mathbb{C}^N$ is an invertible linear transformation.*

Proof. Let us first introduce the following auxiliary notation:

$$e_m = e^{-2\pi i \frac{m}{N}}, \quad \text{for} \quad m = 0,\dots,N-1,$$

noting that it implies that the nth power of e_m is equal to:

$$e_m^n = (e_m)^n = e^{-2\pi i \frac{mn}{N}}.$$

Thus, $F_N(m,n) = e_m^n$, for $m,n = 0,\dots,N-1$. This is the structure of the so called **Vandermonde matrix**. From linear algebra we know that the determinant of the Vandermonde matrix can be expressed as

$$\det(F_N) = \prod_{N>m>n\geq 0} (e_m - e_n).$$

But as for every $m > n$, $e_m \neq e_n$, the determinant is clearly different from zero and the matrix is invertible. □

Proposition 3.2.1 provides us only with the information that F_N is invertible. But the exact estimation of either the determinant of the matrix F_N, or finding its inverse is impossible from just knowing that the determinant of F_N is non-zero. Thus, we require an entirely different approach to finding the inverse of F_N. For this reason we start by introducing the following definition and formalizing the concept of the **Discrete Fourier Transform**.

Definition 3.2.1 (Discrete Fourier Transform)**.** Let $\mathcal{F}_N$ be an $N \times N$ matrix defined by the formula:

$$\mathcal{F}_N(m,n) = \frac{1}{\sqrt{N}} F_N(m,n) = \frac{1}{\sqrt{N}} e^{-2\pi i \frac{mn}{N}},$$

for $m, n = 0, \ldots, N-1$. We say that $\mathcal{F}_N$ is the **Discrete Fourier Transform** and we often use the acronym **DFT**.

We begin by making a few simple observations about the Discrete Fourier Transform.

Proposition 3.2.2. *Let $\mathcal{F}_N$ be the $N \times N$ matrix of the Discrete Fourier Transform. Then, we have:*

a)

$$\mathcal{F}_N(m,n) = \frac{1}{\sqrt{N}} (\cos(-2\pi mn/N) + i \sin(-2\pi mn/N)).$$

b)

$$\overline{\mathcal{F}_N(m,n)} = \frac{1}{\sqrt{N}} e^{2\pi i \frac{mn}{N}} = \frac{1}{\sqrt{N}} (\cos(2\pi mn/N) + i \sin(2\pi mn/N)).$$

c)

$$\mathcal{F}_N = \mathcal{F}_N^T.$$

We are now ready to present one of the most fundamental properties of the Discrete Fourier Transform.

Theorem 3.2.1 (Unitarity of DFT)**.** *Let $\mathcal{F}_N$ be the $N \times N$ matrix of the Discrete Fourier Transform. Then, $\mathcal{F}_N : \mathbb{C}^N \to \mathbb{C}^N$ is a unitary transformation.*

Proof. We start by noting that, by the symmetry of the DFT established in Proposition 3.2.2.b, it suffices to show the orthogonality of columns of the matrix of DFT, and to show that all rows have the same norm equal to 1.

Let F_N^k denote the kth column vector of F_N, and let $\mathcal{F}_N^k$ denote the kth column of the Discrete Fourier Transform, $k = 0, \ldots, N-1$. To show that each column vector (and consequently each row vector) of $\mathcal{F}_N$ is a vector of length 1, we note that,

$$\sum_{n=0}^{N-1} \left| \frac{1}{\sqrt{N}} e^{-2\pi i \frac{mn}{N}} \right|^2 = \sum_{n=0}^{N-1} \frac{1}{N} = 1.$$

Next, let $k \neq l$.

$$\begin{aligned}
\langle F_N^k, F_N^l \rangle &= \sum_{n=0}^{N-1} F_N(n,k)\overline{F_N(n,l)} \\
&= \sum_{n=0}^{N-1} e^{-2\pi i \frac{nk}{N}} e^{2\pi i \frac{nl}{N}} \\
&= \sum_{n=0}^{N-1} e^{-2\pi i \frac{n(k-l)}{N}} \\
&= \sum_{n=0}^{N-1} e_{k-l}^n \\
&= \frac{1 - e_{k-l}^N}{1 - e_{k-l}} = \frac{1 - e^{-2\pi i (k-l)\frac{N}{N}}}{1 - e_{k-l}} = \frac{1-1}{1-e_{k-l}} = 0.
\end{aligned}$$

The above calculation also implies that column (or row) vectors of $\mathcal{F}_N$ are mutually orthogonal to each other. As such we have shown that the rows (or columns) of the matrix $\mathcal{F}_N$ form an ONB for $\mathbb{C}^N$. This is equivalent to the unitarity of $\mathcal{F}_N$. □

Unitarity of a transformation is helpful in establishing a direct relationship between the transformation and its inverse, due to the following property of unitary matrices:

$$UU^* = U^*U = \mathrm{Id},$$

where

$$U^* := \overline{U^T}.$$

Equivalently, we can conclude that for a unitary matrix U:

$$U^{-1} = U^*,$$

which allows us to establish the following useful fact.

Corollary 3.2.1. *Let $\mathcal{F}_N$ be the $N \times N$ matrix of the Discrete Fourier Transform. Then,*

$$\mathcal{F}_N^{-1} = \overline{\mathcal{F}}_N^T = \overline{\mathcal{F}}_N.$$

Corollary 3.2.1 implies in particular that

$$\mathcal{F}_N^{-1}(m,n) = \overline{\mathcal{F}}_N(m,n) = e^{2\pi i \frac{mn}{N}}, \quad \text{for} \quad m,n = 0,\ldots,N-1.$$

However, the unitarity of $\mathcal{F}_N$ has much further reaching consequences. It allows us to establish a relationship between inner products of vectors and the inner products of their Fourier transforms. This is known as the **Parseval theorem**.

Theorem 3.2.2 (Parseval Theorem)**.** *Let $\mathcal{F}_N$ be the $N \times N$ matrix of the Discrete Fourier Transform, and let $v, w \in \mathbb{C}^N$. Then,*

$$\langle v, w \rangle = \langle \mathcal{F}_N(v), \mathcal{F}_N(w) \rangle.$$

Proof. From linear algebra we recall that for any unitary matrix U, we have

$$\langle v, w \rangle = \langle U(v), U(w) \rangle,$$

and we complete the proof by noting that $\mathcal{F}_N$ is in fact unitary. □

Theorem 3.2.3 (Plancherel Theorem)**.** *Let $\mathcal{F}_N$ be the $N \times N$ matrix of the Discrete Fourier Transform, and let $v \in \mathbb{C}^N$. Then,*

$$\sum_{n=0}^{N-1} |v(n)|^2 = \sum_{n=0}^{N-1} |\mathcal{F}_N(v)(n)|^2 = \frac{1}{N} \sum_{n=0}^{N-1} |F_N(v)(n)|^2.$$

Proof. From the Parseval theorem we can conclude that

$$\|v\|_2^2 = \langle v, v \rangle = \langle \mathcal{F}_N(v), \mathcal{F}_N(v) \rangle = \|\mathcal{F}_N(v)\|_2^2,$$

which implies that

$$\sum_{n=0}^{N-1} |v(n)|^2 = \sum_{n=0}^{N-1} |\mathcal{F}_N(v)(n)|^2,$$

and the last equality in the hypothesis follows from the definition of $\mathcal{F}_N$. □

Proposition 3.2.3 (Periodicity)**.**

$$\begin{aligned} \mathcal{F}_N(v)(m+N) &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} v(n) e^{-2\pi i \frac{(m+N)n}{N}} \\ &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} v(n) e^{-2\pi i \frac{mn}{N}} e^{-2\pi i \frac{Nn}{N}} \\ &= \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} v(n) e^{-2\pi i \frac{mn}{N}} = \mathcal{F}_N(v)(m). \end{aligned}$$

Definition 3.2.2 (Translation)**.** We define a shift by n modulo N to be the following operation:

$$T_n(v)(m) = v(m - n \mod N).$$

As such a vector of coefficients $(v(0), v(1), v(2), \ldots, v(N-1))$, becomes $(v(N-n), v(N-n+1), v(N-n+2), \ldots, v(N-n-1))$.

Definition 3.2.3 (Modulation)**.** We define a modulation by m to be the following operation:

$$M_m(v)(n) = v(n) e^{2\pi i m \frac{n}{N}}.$$

Theorem 3.2.4.

$$\mathcal{F}_N(M_m(v)) = T_m(\mathcal{F}_N(v)).$$
$$\mathcal{F}_N(T_n(v)) = M_{-n}(\mathcal{F}_N(v)).$$

Definition 3.2.4 (Circular Convolution)**.**

$$v * w(m) := \sum_{n=0}^{N-1} v(m)w(m - n \ \text{ modulo } \ N).$$

Theorem 3.2.5.

$$v * w = {F_N}^{-1}(F_N(v) \cdot F_N(w)),$$

where $\cdot$ *denotes pointwise multiplication of vector coordinates. Equivalently,*

$$F_N(v * w) = F_N(v) \cdot F_N(w).$$

3.3 Fast Fourier Transform

We start by observing that for every m, n, $\mathcal{F}(m,n) = \frac{1}{\sqrt{N}}$. This implies that each coefficient of $\mathcal{F}$ is in particular different from zero. As such, the cost of applying the DFT to a vector of length N is high:

$$\mathcal{F}(v)(m) = \frac{1}{\sqrt{N}} \sum_{n=0}^{N-1} v(n) e^{-2\pi i \frac{mn}{N}}$$

has $N+1$ multiplications and $N-1$ additions, for each coefficient $m = 0, \ldots, N-1$. As such the cost of applying DFT to a vector of length N is equal to $2N^2$. Even if we discount the obviously redundant operations, like, e.g., by $1/\sqrt{N}$, by factoring it into precomputed terms of the form $1/\sqrt{N} e^{-2\pi i \frac{mn}{N}}$, or by using some similar tricks, we are not going to change the fact that the computation of the DFT for any given vector v requires on the magnitude of N^2 elementary arithmetic operations. This means that for a Fourier transform of a vector with thousands of coefficients, we need millions of elementary operations. For a CD-like quality of sound, we require 44,100 samples per second. This means a vector in 44100 dimensions. So now we are talking about billions of operations to compute the Fourier transform of 1 second of music.

For a transformation to be useful we need something faster. The following result is a key step in this regard. Whenever in doubt, we are going to specify the dimension of the DFT by a subscript: let F_N denotes the $N \times N$ DFT.

Theorem 3.3.1 (Danielson–Lanczos Lemma)**.** *Let N be an even positive integer. Let, for all $0 \leq n \leq N/2$,*

$$v^e(n) = v(2n)$$

and

$$v^o(n) = v(2n+1).$$

Then, for all $m = 0, \ldots, N-1$,

$$F_N(v)(m) = F_{\frac{N}{2}}(v^e)(m) + e^{-2\pi i m/N} F_{\frac{N}{2}}(v^o)(m).$$

Proof.

$$\begin{aligned}
F_N(v)(m) &= \sum_{n=0}^{N-1} v(n)e^{-2\pi i\frac{mn}{N}} \\
&= \sum_{n=0}^{N/2-1} v(2n)e^{-2\pi i\frac{m2n}{N}} + \sum_{n=0}^{N/2-1} v(2n+1)e^{-2\pi i\frac{m(2n+1)}{N}} \\
&= \sum_{n=0}^{N/2-1} v(2n)e^{-2\pi i\frac{m2n}{N}} + e^{-2\pi i\frac{m}{N}} \sum_{n=0}^{N/2-1} v(2n+1)e^{-2\pi i\frac{m2n}{N}} \\
&= \sum_{n=0}^{N/2-1} v(2n)e^{-2\pi i\frac{mn}{N/2}} + e^{-2\pi i\frac{m}{N}} \sum_{n=0}^{N/2-1} v(2n+1)e^{-2\pi i\frac{mn}{N/2}} \\
&= \sum_{n=0}^{N/2-1} v^e(n)e^{-2\pi i\frac{mn}{N/2}} + e^{-2\pi i\frac{m}{N}} \sum_{n=0}^{N/2-1} v^o(n)e^{-2\pi i\frac{mn}{N/2}} \\
&= F_{\frac{N}{2}}(v^e)(m) + e^{-2\pi im/N} F_{\frac{N}{2}}(v^o)(m).
\end{aligned}$$

□

This seemingly "insignificant" and straightforward result is crucial for our ability to find a better algorithm for implementation of DFT. Indeed, the cost of application of $F_{\frac{N}{2}}$ is equal to $2(N/2)^2 = N^2/2$. As such the above result tells us that instead of approx. $2N^2$ elementary operations we only need approximately N^2. If we iterate this formula we can further lower the computational complexity of implementing DFT significantly. And this is the key idea behind the fast implementation of the Fourier transform.

How to do this? Assume that $N = 2^q$. This assumption allows us to repeat halving the integer until we reach 1. Therefore, we can iterate the application of the Danielson-Laczos Lemma exactly q times. Each time we perform the following operations:

- Partition the input vector into two vectors of half length each. Call this operation P_n, for each dimension $n = q, q-1, q-2, \ldots, 1$.
- Apply the Fourier matrix of half the size to the even and odd vectors separately. Let F^n denote the block diagonal matrix with the half size Fourier matrices $F_{2^{n-1}}$, where $n = q, q-1, q-2, \ldots, 1$.
- Recombine the two pieces into one by summing the components. Denote this operation by S_n, for each dimension $n = q, q-1, q-2, \ldots, 1..$

So after the first implementation of the Danielson-Lanczos lemma, we obtain:

$$F_N = S_q F^q P_q,$$

where F_N, S_q, F^q, P_q are all $N \times N$ matrices of the following structure:

- S_q consists of 4 blocks each of size $N/2 \times N/2$
- F^q is a block diagonal matrix, which has two blocks of $F_{N/2}$ on the diagonal
- P_q is a block matrix consisting of two rectangular $N/2 \times N$ blocks with 1s on the diagonals.

After two iterations, we have

$$F_N = S_q S_{q-1} F^{q-1} P_{q-1} P_q,$$

where, in addition to S_q and P_q, we also have

- S_{q-1} consists of 2 diagonal blocks, each consisting of 4 blocks each of size $N/4 \times N/4$
- F^{q-1} is a block diagonal matrix, which has four blocks of $F_{N/4}$ on the diagonal
- P_{q-1} is a block matrix consisting 2 diagonal blocks, each consisting of of two rectangular $N/4 \times N/2$ blocks with 1s on the diagonals.

Now, after all the iterations, we have:

$$F = S_q S_{q-1} \dots S_2 S_1 F^1 P_1 P_2 \dots P_{q-1} P_q.$$

We note that each of the above listed matrices is very sparse: they all have at most 2 nonzero terms in every row and every column.

Indeed, matrix P_n takes the form of the following $2^q \times 2^q$ matrix with $2^n \times 2^n$ diagonal blocks of the form:

$$\begin{pmatrix}
1 & 0 & 0 & 0 & 0 & 0 & 0 & ... & 0 & 0 \\
0 & 0 & 1 & 0 & 0 & 0 & 0 & ... & 0 & 0 \\
0 & 0 & 0 & 0 & 1 & 0 & 0 & ... & 0 & 0 \\
... & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & ... & 1 & 0 \\
0 & 1 & 0 & 0 & 0 & 0 & 0 & ... & 0 & 0 \\
0 & 0 & 0 & 1 & 0 & 0 & 0 & ... & 0 & 0 \\
0 & 0 & 0 & 0 & 0 & 1 & 0 & ... & 0 & 0 \\
... & & & & & & & & & \\
0 & 0 & 0 & 0 & 0 & 0 & 0 & ... & 0 & 1
\end{pmatrix}$$

On the other hand S_n takes the form of the following $2^q \times 2^q$ matrix with $2^n \times 2^n$

diagonal blocks of the form:

$$\begin{pmatrix}
1 & 0 & 0 & 0 & ... & 1 & 0 & ... & 0 & 0 \\
0 & 1 & 0 & 0 & ... & 0 & e^{-2\pi i\frac{1}{2^n}} & ... & 0 & 0 \\
0 & 0 & 1 & 0 & ... & 0 & 0 & e^{-2\pi i\frac{2}{2^n}} & ... & 0 \\
... & & & & & & & & & \\
0 & 0 & 0 & ... & 1 & 0 & 0 & 0 & ... & e^{-2\pi i\frac{2^{n-1}-1}{2^n}} \\
1 & 0 & 0 & 0 & ... & e^{-2\pi i\frac{2^{n-1}}{2^n}} & 0 & 0 & 0 & 0 \\
0 & 1 & 0 & 0 & ... & 0 & e^{-2\pi i\frac{2^{n-1}+1}{2^n}} & 0 & ... & 0 \\
0 & 0 & 1 & 0 & ... & 0 & 0 & e^{-2\pi i\frac{2^{n-1}+2}{2^n}} & ... & 0 \\
... & & & & & & & & & \\
0 & 0 & 0 & ... & 1 & 0 & 0 & 0 & ... & e^{-2\pi i\frac{2^n-1}{2^n}}
\end{pmatrix}$$

As such application of each of these matrices costs at most $3N$ elementary operations. Since there is $2q+1$ such transformations and since $q = \log_2(N)$, we can estimate the computational complexity of this algorithm for computation of the DFT at most as

$$3N(2\log_2(N)+1).$$

Function such as the above is often denoted by

$$O(N\log(N)).$$

This is much faster than the **direct** matrix multiplication of DFT.

Definition 3.3.1 (FFT)**.** The algorithmic implementation of the DFT as described above is called the **Fast Fourier Transform** (FFT).

We do emphasize that this is the same transformation as the original DFT, just a different implementation.

3.4 Trigonometric Transforms

Although the definition of the Fourier Transform takes advantage of complex frequencies, most of the time the types of signals that we analyze have real values. This is due to the fact that most real-life measurement systems are based on counting principles. For example, in image analysis we typically count photons. After some pre- and post-processing, these integers may be averaged or the resulting values can be normalized. However, these procedures will not lead to complex values. Therefore one may ask the following question: Why to analyze real signals using complex-valued transforms? This leads to interpretational difficulties and unnecessarily increases the complexity.

The above considerations led researchers to introduce real-valued analogs of the Fourier Transform. They are known as **discrete sine and cosine transforms**. We begin by introducing auxiliary normalizing constants, which are needed to assure orthogonality of resulting matrices:

$$\beta(k) = \begin{cases} 0, & \text{if } k < 0 \text{ or } k > N, \\ 1/\sqrt{2} & \text{if } k = 0 \text{ or } k = N, \\ 1 & \text{if } 0 < k < N. \end{cases}$$

With this let us introduce the following 4 discrete sine and 4 discrete cosine transforms, which we define after M. V. Wickerhauser.

Definition 3.4.1 (DCT-I)**.** For any $N > 0$ define $C_I^{N+1} : \mathbb{R}^{N+1} \to \mathbb{R}^{N+1}$ by letting:

$$C_I^{N+1}(m,n) = \beta(m)\beta(n)\cos\left(\frac{\pi mn}{N}\right),$$

for $m, n = 0, \ldots, N$.

Definition 3.4.2 (DCT-II)**.** For any $N > 1$ define $C_{II}^N : \mathbb{R}^N \to \mathbb{R}^N$ by letting:

$$C_{II}^N(m,n) = \beta(m)\cos\left(\frac{\pi m(n+\frac{1}{2})}{N}\right),$$

for $m, n = 0, \ldots, N-1$.

Definition 3.4.3 (DCT-III)**.** For any $N > 1$ define $C_{III}^N : \mathbb{R}^N \to \mathbb{R}^N$ by letting:

$$C_{III}^N(m,n) = \beta(n)\cos\left(\frac{\pi(m+\frac{1}{2})n}{N}\right),$$

for $m, n = 0, \ldots, N-1$.

Definition 3.4.4 (DCT-IV)**.** For any $N > 1$ define $C_{IV}^N : \mathbb{R}^N \to \mathbb{R}^N$ by letting:

$$C_{IV}^N(m,n) = \cos\left(\frac{\pi(m+\frac{1}{2})(n+\frac{1}{2})}{N}\right),$$

for $m, n = 0, \ldots, N-1$.

Definition 3.4.5 (DST-I)**.** For any $N > 2$ define $S_I^{N-1} : \mathbb{R}^{N-1} \to \mathbb{R}^{N-1}$ by letting:

$$S_I^{N-1}(m,n) = \sin\left(\frac{\pi mn}{N}\right),$$

for $m, n = 1, \ldots, N-1$.

Definition 3.4.6 (DST-II)**.** For any $N > 1$ define $S_{II}^N : \mathbb{R}^N \to \mathbb{R}^N$ by letting:

$$S_{II}^N(m,n) = \beta(m+1)\sin\left(\frac{\pi(m+1)(n+\frac{1}{2})}{N}\right),$$

for $m, n = 0, \ldots, N-1$.

Definition 3.4.7 (DST-III)**.** For any $N > 1$ define $S_{III}^N : \mathbb{R}^N \to \mathbb{R}^N$ by letting:

$$S_{III}^N(m,n) = \beta(n+1)\sin\left(\frac{\pi(m+\frac{1}{2})(n+1)}{N}\right),$$

for $m, n = 0, \ldots, N-1$.

Definition 3.4.8 (DST-IV)**.** For any $N > 1$ define $S_{IV}^N : \mathbb{R}^N \to \mathbb{R}^N$ by letting:

$$S_{IV}^N(m,n) = \sin\left(\frac{\pi(m+\frac{1}{2})(n+\frac{1}{2})}{N}\right),$$

for $m, n = 0, \ldots, N-1$.

It is not difficult to verify that all 8 transforms which we introduced satisfy various interesting symmetry relationships, are real-valued, and - with appropriate normalization - are orthogonal transforms. As such, when analyzing real-valued signals, they will yield real-valued coefficients.

Example 3.4.1 (Discrete Sine Transform of Type I)**.** As an illustrative example, we can check the orthogonality of Discrete Sine Transform of the 1st type: DST-I. We recall that $S_I^{N-1} : \mathbb{R}^{N-1} \to \mathbb{R}^{N-1}$ is defined as:

$$S_I^{N-1}(m,n) = \sin\left(\frac{\pi mn}{N}\right) = \sin\left(\frac{\pi nm}{N}\right) = S_I^{N-1}(n,m) = \left(S_I^{N-1}\right)^T(m,n).$$

As a first step, we compute the squared norm of the mth column of S_I^{N-1}:

$$\begin{aligned}
\langle S_I^{N-1}(\cdot,m), S_I^{N-1}(\cdot,m)\rangle &= \sum_{n=1}^{N-1} \left(S_I^{N-1}(n,m)\right)^2 \\
&= \sum_{n=1}^{N-1} \sin^2\left(\frac{\pi nm}{N}\right) \\
&= \sum_{n=1}^{N-1} \frac{1}{2}\left(1-\cos\left(2\pi mn/N\right)\right) = \frac{N-1}{2} - \frac{1}{2}\sum_{n=1}^{N-1}\cos\left(-2\pi mn/N\right) \\
&= \frac{N-1}{2} - \frac{1}{2}\sum_{n=1}^{N-1} Re(F_N(m,n)) = \frac{N-1}{2} - \frac{1}{2}Re\left(\sum_{n=1}^{N-1} F_N(m,n)\right) \\
&= \frac{N-1}{2} - \frac{1}{2}Re\left(\sum_{n=1}^{N-1} e^{-2\pi imn/N}\right) = \frac{N-1}{2} + \frac{1}{2} = \frac{N}{2}
\end{aligned}$$

Therefore, we can now ask if the normalized Discrete Sine Transform of type 1:

$$\sqrt{\frac{2}{N}} S_I^{N-1}$$

has mutually orthogonal columns (rows)? Due to symmetry, it suffices to show that the column vectors of S_I^{N-1} are mutually orthogonal to each other.

$$\begin{aligned}
\langle S_I^{N-1}(\cdot,m), S_I^{N-1}(\cdot,n)\rangle &= \sum_{k=1}^{N-1} S_I^{N-1}(m,k) S_I^{N-1}(n,k) \\
&= \sum_{k=1}^{N-1} \sin\left(\frac{\pi mk}{N}\right)\sin\left(\frac{\pi nk}{N}\right) \\
&= \sum_{k=1}^{N-1} \frac{1}{2}\left(\cos(\pi(m-n)k/N) - \cos(\pi(m+n)k/N)\right) \\
&= \frac{1}{2}\sum_{k=1}^{N-1}\left(Re(e^{\pi i(m-n)k/N}) - Re(e^{\pi i(m+n)k/N})\right) \\
&= \frac{1}{2}\sum_{k=0}^{N-1}\left(Re(e^{\pi i(m-n)k/N}) - Re(e^{\pi i(m+n)k/N})\right) \\
&= \frac{1}{2}Re\left(\sum_{k=0}^{N-1} e^{\pi i(m-n)k/N}\right) - \frac{1}{2}Re\left(\sum_{k=0}^{N-1} e^{\pi i(m+n)k/N}\right)
\end{aligned}$$

To establish this, we compute separately:

$$\begin{aligned}\sum_{k=0}^{N-1} e^{\pi i(m-n)k/N} &= \frac{1-e^{\pi i(m-n)(N)/N}}{1-e^{\pi i(m-n)/N}} \\ &= \frac{1-e^{\pi i(m-n)}}{1-e^{\pi i(m-n)/N}} \\ &= \frac{1-e^{\pi i(m-n)}}{1-e^{\pi i(m-n)/N}} \cdot \frac{1-e^{-\pi i(m-n)/N}}{1-e^{-\pi i(m-n)/N}} \\ &= \frac{1}{2(1-\cos(\pi(m-n)/N))}(1-(-1)^{m-n})(1-e^{-\pi i(m-n)/N}).\end{aligned}$$

Thus,

$$\begin{aligned}Re\left(\sum_{k=0}^{N-1} e^{\pi i(m-n)k/N}\right) &= \frac{1}{2(1-\cos(\pi(m-n)/N))}(1-(-1)^{m-n})(1-\cos(\pi(m-n)/N)) \\ &= \frac{1}{2}(1-(-1)^{m-n}).\end{aligned}$$

Similarly, we compute that

$$\begin{aligned}Re\left(\sum_{k=0}^{N-1} e^{\pi i(m+n)k/N}\right) &= \frac{1}{2(1-\cos(\pi(m+n)/N))}(1-(-1)^{m+n})(1-\cos(\pi(m+n)/N)) \\ &= \frac{1}{2}(1-(-1)^{m+n}).\end{aligned}$$

Because the difference between $m+n$ and $m-n$ is an even integer, both real parts are equal to each other, and we thus conclude that

$$\langle S_I^{N-1}(\cdot,m), S_I^{N-1}(\cdot,n)\rangle = 0,$$

for $m \neq n$.

Example 3.4.2 (Discrete Sine Transforms of Type II and III)**.** It is straightforward to notice that Discrete Sine Transforms of Type II and III are not symmetric matrices, but instead they are transpositions of each other:

$$S_{II}^N = (S_{III}^N)^T.$$

Next, we shall compute the norms of rows and columns for these matrices, while noting that for non-symmetric matrices we need to do 2 separate computations. This is because having all rows (resp. columns) of a square matrix of the same constant norm, does not imply that columns (resp. rows) will have the same property. Recall the following example from linear algebra:

$$\begin{pmatrix}1 & 2 & 3\\ 1 & 2 & 3\\ 1 & 2 & 3\end{pmatrix}.$$

Fix $0 \leq n \leq N-1$. Our first step is to compute the squared norm of the nth column of S_{II}^N. We proceed taking advantage of the following property of trigonometric functions: $\sin(\theta) = (-1)^n \cos(\pi(n+\frac{1}{2}) - \theta)$.

$$\begin{aligned}
\|S_{II}^N(\cdot,n)\|^2 &= \sum_{k=0}^{N-1} \beta^2(k+1) \left(S_{II}^N(k,n)\right)^2 \\
&= \sum_{k=0}^{N-1} \beta^2(k+1) \sin^2\left(\frac{\pi(k+1)(n+\frac{1}{2})}{N}\right) \\
&= \sum_{k=0}^{N-1} \beta^2(k+1) \cos^2\left(\frac{\pi(n+\frac{1}{2})(N-k-1)}{N}\right),
\end{aligned}$$

where we recall that $\beta^2(k+1) = 1$ for $0 \leq k < N-1$, and it equals to $1/2$ for $k = N-1$. Next, we make a discrete substitution in the summation, replacing $N-k-1$ with $k'+1$, to get

$$\begin{aligned}
&\sum_{k=0}^{N-1} \beta^2(k+1) \cos^2\left(\frac{\pi(n+\frac{1}{2})(N-k-1)}{N}\right) \\
&= \sum_{k'=-1}^{N-2} \beta^2(N-k'-1) \cos^2\left(\frac{\pi(n+\frac{1}{2})(k'+1)}{N}\right).
\end{aligned}$$

This way we have that

$$\|S_{II}^N(\cdot,n)\|^2 = \sum_{k=0}^{N-1} \beta^2(k+1) \sin^2\left(\frac{\pi(k+1)(n+\frac{1}{2})}{N}\right) = \sum_{k'=-1}^{N-2} \beta^2(N-k'-1) \cos^2\left(\frac{\pi(n+\frac{1}{2})(k'+1)}{N}\right),$$

where

$$\sum_{k=0}^{N-1} \beta^2(k+1) \sin^2\left(\frac{\pi(k+1)(n+\frac{1}{2})}{N}\right) = \sum_{k=0}^{N-2} \sin^2\left(\frac{\pi(k+1)(n+\frac{1}{2})}{N}\right) + \frac{1}{2},$$

and

$$\sum_{k'=-1}^{N-2} \beta^2(N-k'-1) \cos^2\left(\frac{\pi(n+\frac{1}{2})(k'+1)}{N}\right) = \sum_{k'=0}^{N-2} \cos^2\left(\frac{\pi(n+\frac{1}{2})(k'+1)}{N}\right) + \frac{1}{2}.$$

Hence, we have that

$$\sum_{k=0}^{N-2} \sin^2\left(\frac{\pi(k+1)(n+\frac{1}{2})}{N}\right) = \sum_{k'=0}^{N-2} \cos^2\left(\frac{\pi(n+\frac{1}{2})(k'+1)}{N}\right).$$

Now, using this fact and the fundamental trigonometric identity $\sin^2(\theta)+\cos^2(\theta) = 1$, we obtain that

$$\sum_{k=0}^{N-2} \sin^2\left(\frac{\pi(k+1)(n+\frac{1}{2})}{N}\right) = \frac{1}{2}\sum_{n=0}^{N-2} 1 = \frac{N-1}{2}.$$

So, adding all together, we see that for any $0 \leq n \leq N-1$,

$$\|S_{II}^N(\cdot, n)\| = \sqrt{\frac{N}{2}}.$$

Next, let us look at the norms of rows of S_{II}^N. We have to split this computation into two separate cases. Namely, we consider $0 \leq m < N-1$ and the special case of $m = N-1$. Thus, the corresponding rows have the following norms. For the case $m = N-1$, we have

$$\|S_{II}^N(N-1,\cdot)\|^2 = \sum_{k=0}^{N-1} \beta^2(N)\sin^2\left(\frac{\pi N(k+\frac{1}{2})}{N}\right) = \frac{1}{2}\sum_{k=0}^{N-1}\sin^2\left(\pi(k+\frac{1}{2})\right) = \frac{N}{2}.$$

For the general case $0 \leq m < N-1$, we get:

$$\|S_{II}^N(m,\cdot)\|^2 = \sum_{k=0}^{N-1} \beta^2(m+1)\sin^2\left(\frac{\pi(m+1)(k+\frac{1}{2})}{N}\right) = \sum_{k=0}^{N-1}\sin^2\left(\frac{\pi(m+1)(k+\frac{1}{2})}{N}\right).$$

We now take advantage of the following trigonometric identity $\sin^2(x) = \frac{1-\cos(2x)}{2}$, which allows us to write:

$$\sum_{k=0}^{N-1}\sin^2\left(\frac{\pi(m+1)(k+\frac{1}{2})}{N}\right) = \sum_{k=0}^{N-1}\frac{1-\cos\left(2\frac{\pi(m+1)(k+\frac{1}{2})}{N}\right)}{2} = \frac{N}{2} - \frac{N}{2}\sum_{k=0}^{N-1}\cos\left(\frac{2\pi(m+1)(k+\frac{1}{2})}{N}\right).$$

To compute this last sum we observe that

$$\begin{aligned}
\sum_{k=0}^{N-1}\cos\left(\frac{2\pi(m+1)(k+\frac{1}{2})}{N}\right) &= Re\left(\sum_{k=0}^{N-1}\exp\left(\frac{2\pi(m+1)(k+\frac{1}{2})}{N}\right)\right) \\
&= Re\left(e^{\frac{\pi i(m+1)}{N}}\sum_{k=0}^{N-1}\exp^k\left(\frac{2\pi(m+1)}{N}\right)\right) \\
&= Re\left(e^{\frac{\pi i(m+1)}{N}}\frac{1-e^{2\pi i(m+1)}}{1-e^{\frac{2\pi i(m+1)}{N}}}\right) \\
&= Re(0) = 0.
\end{aligned}$$

Hence,

$$\|S_{II}^N(m,\cdot)\|^2 = \frac{N}{2}.$$

As such, we can claim that all rows and columns of transforms S_{II}^N and S_{III}^N have norm equal to $\sqrt{\frac{N}{2}}$.

Example 3.4.3 (Discrete Sine Transform of Type IV)**.** We begin by recalling that the Discrete Sine Transform of Type IV $S_{IV}^N : \mathbb{R}^N \to \mathbb{R}^N$ is defined by letting:

$$S_{IV}^N(m,n) = \sin\left(\frac{\pi(m+\frac{1}{2})(n+\frac{1}{2})}{N}\right),$$

for $m, n = 0, \ldots, N-1$. As such the symmetry of this transform holds by definition: $S_{IV}^N(m,n) = S_{IV}^N(n,m)$.

Therefore, as before, our first step is to compute the squared norm of the mth column of S_{IV}^N. We proceed as before, taking advantage of the following property of trigonometric functions: $\sin(\theta) = (-1)^n \cos(\pi(n+\frac{1}{2}) - \theta)$.

$$\begin{aligned}
\langle S_{IV}^N(\cdot,m), S_{IV}^N(\cdot,m)\rangle &= \sum_{n=0}^{N-1} \left(S_{IV}^N(n,m)\right)^2 \\
&= \sum_{n=0}^{N-1} \sin^2\left(\frac{\pi(m+\frac{1}{2})(n+\frac{1}{2})}{N}\right) \\
&= \sum_{n=0}^{N-1} \cos^2\left(\frac{\pi(m+\frac{1}{2})(N-1-n+\frac{1}{2})}{N}\right).
\end{aligned}$$

Next, we make a discrete substitution in the summation, replacing $N-1-n$ with n', to get

$$\begin{aligned}
&\sum_{n=0}^{N-1} \cos^2\left(\frac{\pi(m+\frac{1}{2})(N-1-n+\frac{1}{2})}{N}\right) \\
&= \sum_{n'=0}^{N-1} \cos^2\left(\frac{\pi(m+\frac{1}{2})(n'+\frac{1}{2})}{N}\right).
\end{aligned}$$

This way we have that

$$\sum_{n=0}^{N-1} \sin^2\left(\frac{\pi(m+\frac{1}{2})(n+\frac{1}{2})}{N}\right) = \sum_{n=0}^{N-1} \cos^2\left(\frac{\pi(m+\frac{1}{2})(n'+\frac{1}{2})}{N}\right).$$

Now, using the fundamental trigonometric identity $\sin^2(\theta) + \cos^2(\theta) = 1$, we obtain that

$$\sum_{n=0}^{N-1} \sin^2\left(\frac{\pi(m+\frac{1}{2})(n+\frac{1}{2})}{N}\right) = \frac{1}{2}\sum_{n=0}^{N-1} 1 = \frac{N}{2}.$$

So,

$$\langle S_{IV}^N(\cdot,m), S_{IV}^N(\cdot,m)\rangle = \frac{N}{2}.$$

Therefore, what remains to show is that for $m \neq n$:

$$\langle S_{IV}^N(\cdot,m), S_{IV}^N(\cdot,n)\rangle = 0.$$

To verify this formula, we begin by making the following observation:

$$\begin{aligned}\langle S_{IV}^N(\cdot,m), S_{IV}^N(\cdot,n)\rangle &= \sum_{k=0}^{N-1} \sin\left(\frac{\pi(m+\frac{1}{2})(k+\frac{1}{2})}{N}\right)\sin\left(\frac{\pi(n+\frac{1}{2})(k+\frac{1}{2})}{N}\right)\\ &= \frac{1}{2}\sum_{k=0}^{N-1} 2\sin\left(\frac{\pi(m+\frac{1}{2})(k+\frac{1}{2})}{N}\right)\sin\left(\frac{\pi(n+\frac{1}{2})(k+\frac{1}{2})}{N}\right)\\ &= \frac{1}{2}\sum_{k=0}^{N-1}\left(\cos\left(\frac{\pi(m+\frac{1}{2})(k+\frac{1}{2})}{N}-\frac{\pi(n+\frac{1}{2})(k+\frac{1}{2})}{N}\right)\right.\\ &\quad\left.-\cos\left(\frac{\pi(m+\frac{1}{2})(k+\frac{1}{2})}{N}+\frac{\pi(n+\frac{1}{2})(k+\frac{1}{2})}{N}\right)\right),\end{aligned}$$

where, in the last equality, we used the fact that $2\sin(\alpha)\sin(\beta) = \cos(\alpha-\beta)-\cos(\alpha+\beta)$. We can now rearrange the above expression as follows:

$$\begin{aligned}&\frac{1}{2}\sum_{k=0}^{N-1}\left(\cos\left(\frac{\pi(m+\frac{1}{2})(k+\frac{1}{2})}{N}-\frac{\pi(n+\frac{1}{2})(k+\frac{1}{2})}{N}\right)-\cos\left(\frac{\pi(m+\frac{1}{2})(k+\frac{1}{2})}{N}+\frac{\pi(n+\frac{1}{2})(k+\frac{1}{2})}{N}\right)\right)\\ &= \frac{1}{2}\sum_{k=0}^{N-1}\left(\cos\left(\frac{\pi(m-n)(k+\frac{1}{2})}{N}\right)-\cos\left(\frac{\pi(m+n+1)(k+\frac{1}{2})}{N}\right)\right)\\ &= \frac{1}{2}\sum_{k=0}^{N-1}\cos\left(\frac{\pi(m-n)(k+\frac{1}{2})}{N}\right)-\frac{1}{2}\sum_{k=0}^{N-1}\cos\left(\frac{\pi(m+n+1)(k+\frac{1}{2})}{N}\right).\end{aligned}$$

We now deal with each of the two above terms separately. First, we see that because $m \neq n$, we can use the geometric summation formula to obtain:

$$\begin{aligned}\sum_{k=0}^{N-1}\cos\left(\frac{\pi(m-n)(k+\frac{1}{2})}{N}\right) &= Re\left(\sum_{k=0}^{N-1} e^{\frac{\pi i(m-n)(k+\frac{1}{2})}{N}}\right)\\ &= Re\left(e^{\frac{\pi i(m-n)}{2N}}\frac{1-e^{\pi i(m-n)}}{1-e^{\pi i(m-n)/N}}\right)\\ &= Re\left(\frac{1-e^{\pi i(m-n)}}{e^{-\frac{\pi i(m-n)}{2N}}-e^{\frac{\pi i(m-n)}{2N}}}\right)\\ &= Re\left(\frac{1-e^{\pi i(m-n)}}{-2i\sin(\pi i(m-n)/2N)}\right)\\ &= Re\left(\frac{1-(-1)^{m-n}}{-2i\sin(\pi i(m-n)/2N)}\right) = 0.\end{aligned}$$

We note that since $1 \leq m+n+1 \leq 2N-1$, $(m+n+1)/N$ cannot be an even integer. Hence $e^{\frac{\pi i(m+n+1)}{N}} \neq 1$. Thus, we proceed with the second term similarly as before to

get:

$$\begin{aligned}
\sum_{k=0}^{N-1} \cos\left(\frac{\pi(m+n+1)(k+\frac{1}{2})}{N}\right) &= Re\left(e^{\frac{\pi i(m+n+1)(k+\frac{1}{2})}{N}}\right) \\
&= Re\left(e^{\frac{\pi i(m+n+1)}{2N}} \frac{1-e^{\pi i(m+n+1)}}{1-e^{\pi i(m+n+1)/N}}\right) \\
&= Re\left(\frac{1-e^{\pi i(m+n+1)}}{e^{-\frac{\pi i(m+n+1)}{2N}} - e^{\frac{\pi i(m+n+1)}{2N}}}\right) \\
&= Re\left(\frac{1-e^{\pi i(m+n+1)}}{-2i\sin(\pi i(m+n+1)/2N)}\right) \\
&= Re\left(\frac{1-(-1)^{m+n+1}}{-2i\sin(\pi i(m+n+1)/2N)}\right) = 0.
\end{aligned}$$

These last two observations now complete the proof of the fact that

$$\langle S_{IV}^N(\cdot, m), S_{IV}^N(\cdot, n)\rangle = 0.$$

3.5 Discrete Hartley Transform

We introduce the following real-valued transformation which is an analogue of the Discrete Fourier Transform, and which is suitable for processing real-valued input vectors. We begin by defining a new function - the so-called *cosine and sine* function:

$$\text{cas}(x) = \cos(x) + \sin(x).$$

Sometimes in engineering literature, the function **cas** is also known as the **Hartley kernel**.

Next, we let H_N to be an $N \times N$ matrix with real-valued coefficients defined by the formula:

$$H_N(m,n) = \ \text{cas}(2\pi mn/N) = \cos(2\pi mn/N) + \sin(2\pi mn/N),$$

where $m, n = 0, \ldots, N-1$. Finally, we define the **Discrete Hartley Transform (DHT)** to be an $N \times N$ matrix with real-valued coefficients defined by the formula:

$$\mathcal{H}_N(m,n) = \frac{1}{\sqrt{N}} H_N(m,n) = \frac{1}{\sqrt{N}}\ \text{cas}(2\pi mn/N),$$

where $m, n = 0, \ldots, N-1$. We immediately note that this is a symmetric matrix:

$$\mathcal{H}_N(m,n) = \mathcal{H}_N(n,m),$$

i.e., $\mathcal{H}_N = \mathcal{H}_N^T$.

We can also introduce an auxiliary new function, *cosine minus sine*, using the obvious formula of:

$$\text{cms}(x) = \text{cas}(-x) = \cos(x) - \sin(x).$$

(We can in fact check that cms is the derivative of cas!) The functions cas and cms satisfy some interesting relationships (which can be verified using basic trigonometric identities):

1. $\text{cas}(x+y) = \cos(x)\text{cas}(y) + \sin(x)\text{cms}(y)$;
2. $\text{cas}(x-y) = \cos(x)\text{cms}(y) + \sin(x)\text{cas}(y)$;
3. $\text{cas}(x)\text{cas}(y) = \cos(x-y) + \sin(x+y)$;
4. $\text{cas}(x) + \text{cas}(y) = 2\text{cas}((x+y)/2)\cos((x-y)/2)$;
5. $\text{cas}(x) - \text{cas}(y) = 2\text{cms}((x+y)/2)\sin((x-y)/2)$.

These relationships provide us with some additional insight into the cas function, showing for example that it satisfies many properties analogous to standard trigonometric functions. This is not surprising, as we can easily establish that in fact cas function is a form of scaled sinusoid:

$$\text{cas}(x) = \sqrt{2}\sin(x + \pi/4) = \sqrt{2}\cos(x - \pi/4).$$

What we are really interested in showing, however, is the fact that this new DHT transformation is very closely related to the Discrete Fourier Transform. For this purpose let us observe first the following relationship:

$$\begin{aligned} H_N(m,n) &= \mathrm{cas}(2\pi mn/N) = \cos(2\pi mn/N) + \sin(2\pi mn/N) \\ &= Re(e^{2\pi imn/N}) + Im(e^{2\pi imn/N}) \\ &= Re(\overline{F_N(m,n)}) + Im(\overline{F_N(m,n)}), \end{aligned}$$

where Re and Im denote the real and imaginary parts of a complex number, resp., and where $F_N(m,n) = e^{-2\pi imn/N}$ is the (m,n)th coefficient of the matrix F_N, which we used to define the DFT. This last identity indicates that we can use DFT to compute DHT: All we need to do is to first compute the DFT, then to switch the signs of all imaginary parts of every coefficient to produce the inverse of the DFT, and finally to add its real and imaginary parts. This way we can obtain a fast implementation for the DHT, without relying on a version of Danielson-Lanczos lemma for this new transformation.

Next, we will use some additional information about parity of sines and cosines, to modify this last relationship in the following manner:

$$\begin{aligned} H_N(m,n) &= \mathrm{cas}(2\pi mn/N) = \cos(2\pi mn/N) + \sin(2\pi mn/N) \\ &= \cos(-2\pi mn/N) - \sin(-2\pi mn/N) \\ &= Re(e^{-2\pi imn/N}) - Im(e^{-2\pi imn/N}) \\ &= Re(F_N(m,n)) - Im(F_N(m,n)). \end{aligned}$$

This formula serves the same purpose, as the previous one, but - by using some properties of sine and cosine - we no longer need to compute the inverse DFT. Just the DFT itself.

But what to do if we do not want to mix real and imaginary parts of DFT variants? Well, there is another formula for that. We start by expanding the coefficients of DFT:

$$\begin{aligned} F_N(m,n) &= \cos(-2\pi mn/N) + i\sin(-2\pi mn/N), \\ \overline{F_N}(m,n) &= \cos(2\pi mn/N) + i\sin(2\pi mn/N). \end{aligned}$$

Furthermore, by multiplying the first equation above by i, we obtain:

$$iF_N(m,n) = i\cos(-2\pi mn/N) - \sin(-2\pi mn/N) = \sin(2\pi mn/N) + i\cos(2\pi mn/N).$$

Now, we add $\overline{F_N}(m,n)$ and $iF_N(m,n)$ to get:

$$\begin{aligned} \overline{F_N}(m,n) + iF_N(m,n) &= \cos(2\pi mn/N) + i\sin(2\pi mn/N) + \sin(2\pi mn/N) + i\cos(2\pi mn/N) \\ &= (\cos(2\pi mn/N) + \sin(2\pi mn/N)) + i\,(\cos(2\pi mn/N) + \sin(2\pi mn/N)) \\ &= \mathrm{cas}(2\pi mn/N) + i\mathrm{cas}(2\pi mn/N) \\ &= H_N(m,n) + iH_N(m,n). \end{aligned}$$

This leads us to the following conclusion:

$$H_N(m,n) = Re\left(\overline{F_N}(m,n) + iF_N(m,n)\right)$$
$$H_N(m,n) = Im\left(\overline{F_N}(m,n) + iF_N(m,n)\right).$$

In other words, H_N is simultaneously the real and the imaginary part of the complex-valued matrix $\overline{F_N}+iF_N$. Now this relationship will be beneficial for us in the following way: we will use it to show the **orthogonality of rows** (or columns - because of symmetry) of **DHT**. Let us denote the mth row of matrix H_N by H_N^m, and similarly for the mth row of F_N, we shall denote it by F_N^m, $m = 0, \ldots, N-1$.

We start by doing the following calculation for $m \neq n$, $m, n = 0, \ldots, N-1$:

$$\begin{aligned}\langle \overline{F_N^m} + iF_N^m; \overline{F_N^n} + iF_N^n\rangle &= \langle \overline{F_N^m}; \overline{F_N^n}\rangle + \langle \overline{F_N^m}; iF_N^n\rangle + \langle iF_N^m; \overline{F_N^n}\rangle + \langle iF_N^m; iF_N^n\rangle \\ &= \overline{\langle F_N^m; F_N^n\rangle} - i\langle \overline{F_N^m}; F_N^n\rangle + i\langle F_N^m; \overline{F_N^n}\rangle + \langle F_N^m; F_N^n\rangle,\end{aligned}$$

where the last equality follows from the properties of the complex inner product.

We can now use the orthogonality of the DFT to observe that

$$\overline{\langle F_N^m; F_N^n\rangle} = \langle F_N^m; F_N^n\rangle = 0.$$

Furthermore, we can easily see that

$$\langle \overline{F_N^m}; F_N^n\rangle = \overline{\langle F_N^m; \overline{F_N^n}\rangle},$$

which, in turn, implies that

$$-i\langle \overline{F_N^m}; F_N^n\rangle + i\langle F_N^m; \overline{F_N^n}\rangle = i\left(\langle F_N^m; \overline{F_N^n}\rangle - \overline{\langle F_N^m; \overline{F_N^n}\rangle}\right)$$

is a real number.

We now summarize our findings so far:

1. coefficients $H_N(m,n)$ are real-valued and we want to compute the real-valued inner product between H_N^m and H_N^n for $m \neq n$;
2. for real-valued vectors, the real- and complex-valued inner products coincide;
3. the complex-valued inner product of $\overline{F_N^m} + iF_N^m$ with $\overline{F_N^n} + iF_N^n$ is real;
4. since $\overline{F_N}(m,n) + iF_N(m,n) = H_N(m,n) + iH_N(m,n)$, we have

   $$\langle \overline{F_N^m} + iF_N^m; \overline{F_N^n} + iF_N^n\rangle = 2\langle H_N^m; H_N^n\rangle,$$

   where the inner product on the right hand side can be treated as a real-valued inner product, because all coefficients of H_N are real; this however, can be

rewritten as:

$$\begin{aligned}
\langle H_N^m; H_N^n\rangle &= \frac{1}{2}\langle \overline{F_N^m} + iF_N^m; \overline{F_N^n} + iF_N^n\rangle \\
&= \frac{i}{2}\left(\langle F_N^m; \overline{F_N^n}\rangle - \langle \overline{F_N^m}; F_N^n\rangle\right) \\
&= \frac{i}{2}\sum_{k=0}^{N-1}\left(F_N(m,k)F_N(n,k) - \overline{F_N(m,k)F_N(n,k)}\right) \\
&= \frac{i}{2}\left(\sum_{k=0}^{N-1} F_N(m,k)F_N(n,k) - \overline{\sum_{k=0}^{N-1} F_N(m,k)F_N(n,k)}\right) \\
&= \frac{i}{2}2iIm\left(\sum_{k=0}^{N-1} F_N(m,k)F_N(n,k)\right) = -Im\left(\sum_{k=0}^{N-1} F_N(m,k)F_N(n,k)\right) \\
&= -Im\left(\sum_{k=0}^{N-1} F_N(m,k)F_N(k,n)\right) = -Im\left(F_N^2(m,n)\right).
\end{aligned}$$

The second to last equality follows from the symmetry of F_N.

5. Finally, we recall from one of the HWs that $F_N^2(m,n)$ is always a real number, equal to either 0 or N, depending on the sum of $m+n$. Thus,

$$Im\left(F_N^2(m,n)\right) = 0.$$

We are now ready to complete the argument that rows (or, equivalently, columns) of DHT are mutually orthogonal. We have derived in (5) the following equality:

$$\langle H_N^m; H_N^n\rangle = Im\left(F_N^2(m,n)\right) = 0,$$

which completes our proof.

Additionally, we note here that the above calculation can be used to show that the norm of the mth column of matrix H_N is equal to N, hence, indeed

$$\mathcal{H}_N(m,n) = \frac{1}{\sqrt{N}}H_N(m,n) = \frac{1}{\sqrt{N}}\,\mathrm{cas}(2\pi mn/N)$$

is a real-valued, symmetric, orthogonal transform.

Remark: *Please do note that we are not using here the fact that most of the coefficients of F_N^2 are zeros. We are only using the fact that F_N^2 is a real-valued matrix, which is an interesting corollary of our HW problem in its own right.*

We shall close this section by reversing our direction and showing that not only we can use the DFT to compute the DHT, but *vice versa*, the DFT can be computed from the DHT. Or, to be more precise, from the trigonometric functions **cas** and **cms**. In fact, we have that

$$\cos(x) = \frac{1}{2}\left(\mathrm{cas}(x) + \mathrm{cms}(x)\right).$$

Similarly,

$$\sin(x) = \frac{1}{2}\left(\mathrm{cas}(x) - \mathrm{cms}(x)\right).$$

Therefore, we can write:

$$\begin{aligned} F_N(m,n) &= \frac{1}{2}\left((\mathrm{cas}(2\pi mn/N) + \mathrm{cms}(2\pi mn/N)) + i\left(\mathrm{cas}(2\pi mn/N) - \mathrm{cms}(2\pi mn/N)\right)\right) \\ &= \frac{1}{2}((H_N(m,n) + H_N(-m,n)) + i(H_N(m,n) - H_N(-m,n))), \end{aligned}$$

where $-m$ is understood in the sense of periodic extension of DHT, similarly to how we treated DFT.

Thus, DFT can be computed from DHT, which is advantageous, as all computations are done in real numbers only, and then some of the results are multiplied at the end by i.

3.6 Problems

1. Write explicitly the matrix of 6×6 DFT. Apply it to a vector $(-1, 0, 1, 0, -1, 0)$.

2. What is the matrix of the 4th power of the $N \times N$ DFT matrix? Give a formula for a generic N and prove it.

3. Find the eigenvalues of 64×64 DFT matrix.

4. Let A denote the matrix of an $N \times N$ unitary DCT-IV transformation (i.e., $A(m,n) = \sqrt{\frac{2}{N}} \cos(\pi(m+0.5)(n+0.5)/N)$). Let B denote the matrix of an $N \times N$ unitary DST-IV transformation (i.e., $B(m,n) = \sqrt{\frac{2}{N}} \sin(\pi(m+0.5)(n+0.5)/N)$). Let M be the matrix of a linear transformation which is defined as follows:

$$M(j,k) = \begin{cases} A(j,m) & if \;\; k = 2m, \\ B(j,m) & if \;\; k = 2m+1 \end{cases}.$$

for $j = 0, \ldots, N-1$ and $k = 0, \ldots, 2N-1$.

a) Do columns of M form a frame for $\mathbb{R}^N$? If so, what are the frame constants?

b) Do the rows of M form a frame for $\mathbb{R}^{2N}$? If so, what are the frame constants?

5. Prove that the $N \times N$ discrete Hartley transform matrix (recall $H(k,n) = \cos(2\pi nk/N) + \sin(2\pi nk/N)$) has mutually orthogonal columns. (You can use unitarity of DFT, i.e., you can assume that columns of DFT are mutually orthogonal.)

6. Compute the 8×8 DFT applied as a matrix multiplication to the following vectors: $v_1(k) = \sin(2\pi k/8), k = 0, \ldots, 7$, $v_2(k) = \sin(4\pi k/8), k = 0, \ldots, 7$, $v_3(k) = \cos(2\pi k/8), k = 0, \ldots, 7$. Plot the results in form of a function graph. Draw conclusions. (You may use matlab, but you are not allowed to use any built-in DFT command.)

7. Show that $\mathcal{F}^2(m,n) = 1$ when $m = N - n$ and 0 otherwise.

8. Let $v(k), k = 0, \ldots, N-1$ be a given vector of length N. Let $w(k) = v(k)e^{2\pi ikm/N}$. Express the DFT of w in terms of the DFT of v. Show all your work.

9. Compute the 8×8 FFT of the following vectors: $v_1(k) = \sin(2\pi k/8), k = 0, \ldots, 7$, $v_2(k) = \sin(4\pi k/8), k = 0, \ldots, 7$, $v_3(k) = \cos(2\pi k/8), k = 0, \ldots, 7$. Plot the results in form of a function graph and compare with your results of HW 5 (You may use Matlab's fft function for the FFT.)

10. Compute the 100×100 FFT of the following vector: $v_1(k) = \sin(2\pi k/100), k = 0, \ldots, 99$, and compare with the 101×101 FFT of the following vector: $v_1(k) = \sin(2\pi k/100), k = 0, \ldots, 100$.

11. Compare experimentally the computational cost of your 1-dimensional DFT implementation from HW 5, with that of the native Matlab FFT implementation (fft). For this problem, depending on what computer you use, you must carefully choose the size of data to be processed, so that your time measurements are, both, non-negligible and finite.

Chapter 4

Discrete Time-Scale Analysis

4.1 The Discrete Haar Transform

4.1.1 The Haar functions

We begin this story with the concept of the **dyadic intervals**, i.e., intervals of which endpoints are formed by dyadic rational numbers.

Definition 4.1.1 (Dyadic Interval)**.** Let $j, k \in \mathbb{Z}$. Define the interval $I_{j,k}$ to be:

$$I_{j,k} = [2^{-j}k, 2^{-j}(k+1)).$$

We note that these intervals are formed by two simple operations or shifting and dilating of the unit interval $[0, 1)$.

Let $\mathbf{D}_j$, $j \in \mathbb{Z}$, be the operation of dyadic dilation of a set $A \subset \mathbb{R}$:

$$\mathbf{D}_j(A) = 2^{-j}A = \{x \in \mathbb{R} : 2^j x \in A\}.$$

Let $\mathbf{T}_k$ be the operation of an integer shift of a set $A \subset \mathbb{R}$:

$$\mathbf{T}_k(A) = A + k = \{x \in \mathbb{R} : x - k \in A\}.$$

Then,

$$I_{j,k} = \mathbf{D}_j\mathbf{T}_k([0, 1)) = \mathbf{D}_j([k, k+1)) = [2^{-j}k, 2^{-j}(k+1)).$$

This specific structure comes handy in proving some of the properties of the dyadic intervals.

First, we say that $(j, k) = (j', k')$ if $j = j'$ and $k = k'$. This implies that $(j, k) \neq (j', k')$ if $j \neq j'$ or $k \neq k'$. In particular, this means it it suffices for one of the coefficients to differ, in order to produce a different pair.

Clearly, if $(j, k) = (j', k')$ then $I_{j,k} = I_{j',k'}$, but much more can be deduced about the intervals from their indices, as the next lemma indicates.

Lemma 4.1.1. *Let $j, j', k, k' \in \mathbb{Z}$ be such that $(j,k) \neq (j',k')$. Then, one of the following is true:*
(i) $I_{j,k} \cap I_{j',k'} = \emptyset$, or
(ii) $I_{j,k} \subset I_{j',k'}$, or
(iii) $I_{j',k'} \subset I_{j,k}$.

Given the integers $j, k \in \mathbb{Z}$ and the interval $I_{j,k}$, we define $I^l_{j,k}$ (resp., $I^r_{j,k}$) to be the left half of $I_{j,k}$ (resp., the right half of $I_{j,k}$). Thus,

$$I^l_{j,k} = [2^{-j}k, 2^{-j}k + 2^{-j-1}),$$
$$I^r_{j,k} = [2^{-j}k + 2^{-j-1}, 2^{-j}(k+1)).$$

It is not difficult to see that:

$$I^l_{j,k} = I_{j+1,2k} = [2^{-(j+1)}2k, 2^{-(j+1)}(2k+1)) = [2^{-j}k, 2^{-j}k + 2^{-(j+1)})$$

and

$$I^r_{j,k} = I_{j+1,2k+1} = [2^{-(j+1)}(2k+1), 2^{-(j+1)}((2k+1)+1)) = [2^{-j}k + 2^{-(j+1)}, 2^{-j}(k+1)).$$

With this observation in mind, we have the following lemma.

Lemma 4.1.2. *Let $j, j', k, k' \in \mathbb{Z}$ be such that $(j,k) \neq (j',k')$. If $I_{j,k} \subset I_{j',k'}$ then*
(i) either $I_{j,k} \subset I^l_{j',k'}$
(ii) or $I_{j,k} \subset I^r_{j',k'}$.

We shall now transform our observations about subsets of $\mathbb{R}$, into results about functions on $\mathbb{R}$.

Definition 4.1.2 (Characteristic Function)**.** Given set $A \subset \mathbb{R}$, we say that $\mathbb{1}_A$ is the *characteristic function* of the set A, if

$$\mathbb{1}_A(x) = \begin{cases} 1, & x \in A, \\ 0, & \text{otherwise.} \end{cases}$$

Let $p(x) = \mathbb{1}_{[0,1)}(x)$ be the **characteristic function** of the the interval $[0,1)$, i.e.,

$$p(x) = \mathbb{1}_{[0,1)}(x) = \begin{cases} 1, & x \in [0,1), \\ 0, & \text{otherwise.} \end{cases}$$

Furthermore, let $h(x) = \mathbb{1}_{[0,1/2)}(x) - \mathbb{1}_{[1/2,1)}(x)$, i.e.,

$$h(x) = \mathbb{1}_{[0,1/2)}(x) - \mathbb{1}_{[1/2,1)}(x) = \begin{cases} 1, & x \in [0,1/2), \\ -1, & x \in [1/2,1), \\ 0, & \text{otherwise.} \end{cases}$$

Some of the results about dyadic intervals can be now restated in terms of functions. To do this we will introduce first the analogues for characteristic functions of dilations and translations of intervals.

Definition 4.1.3 (Dilations and Translations)**.** Let for $a > 0$, $D_a f(x) = \sqrt{a} f(ax)$ be the dilation operator which is an isometry on the space of square-integrable functions $L^2(\mathbb{R})$.

Let for $b \in \mathbb{R}$, $T_b f(x) = f(x - b)$ be the translation operator which also is an isometry on $L^2(\mathbb{R})$.

Definition 4.1.4 (Haar scaling functions)**.** For each $j, k \in \mathbb{Z}$, let

$$p_{j,k} = D_{2^j} T_k p(x) = 2^{j/2} p(2^j x - k).$$

The collection $\{p_{j,k}(x)\}_{j,k \in \mathbb{Z}}$ is called the system of *Haar scaling functions*. For each $j \in \mathbb{Z}$, the collection $\{p_{j,k}(x)\}_{k \in \mathbb{Z}}$ is called the system of scale j *Haar scaling functions*.

We now note that

$$p_{j,k}(x) = 2^{j/2} \mathbb{1}_{I_{j,k}}(x),$$

where we recall that $I_{j,k} = [2^{-j}k, 2^{-j}(k+1))$. Furthermore, we observe that

$$p_{j,k}(x) = D_{2^j} T_k p(x) = D_{2^j} T_k \mathbb{1}_{[0,1)}(x) = 2^{j/2} \mathbb{1}_{\mathbf{D}_j \mathbf{T}_k([0,1))},$$

or, in particular, that

$$D_{2^j} T_k p(x) = 2^{j/2} \mathbb{1}_{\mathbf{D}_j \mathbf{T}_k([0,1))},$$

which is an interesting connection between dilations and translations of functions and of sets.

It is also not difficult to observe now that for all $j, k \in \mathbb{Z}$:

$$\int_{\mathbb{R}} p_{j,k}(x)\, dx = 2^{-j/2}$$

and

$$\int_{\mathbb{R}} |p_{j,k}(x)|^2\, dx = 1.$$

This last equation is the real reason for the introduction of the normalizing factor of $\sqrt{a}$ in the definition of dilation of a function by a. This is not necessary in the case of translations, which do not alter the integrals.

Definition 4.1.5 (Haar Wavelet Functions)**.** For each $j, k \in \mathbb{Z}$, let

$$h_{j,k}(x) = 2^{j/2} h(2^j x - k) = D_{2^j} T_k h(x).$$

The collection $\{h_{j,k}(x)\}_{j,k \in \mathbb{Z}}$ is called the *Haar wavelet system* on $\mathbb{R}$. For each $j \in \mathbb{Z}$, the collection $\{h_{j,k}(x)\}_{k \in \mathbb{Z}}$ is called the collection of scale j *Haar wavelet functions*.

Note that

$$h_{j,k}(x) = 2^{j/2}(\mathbb{1}_{I_{j+1,2k}}(x) - \mathbb{1}_{I_{j+1,2k+1}}(x)),$$

where again $I_{j,k} = [2^{-j}k, 2^{-j}(k+1))$. It is also clear that

$$\int_{\mathbb{R}} h_{j,k}(x)dx = 0$$

and

$$\int_{\mathbb{R}} |h_{j,k}(x)|^2 dx = 1.$$

Alternatively, we can write:

$$h_{j,k} = 2^{j/2}\left(\mathbb{1}_{I^l_{j,k}} - \mathbb{1}^r_{j,k}\right), \quad j,k \in \mathbb{Z}.$$

Also, because of our earlier observations about $I^l_{j,k}$ and $I^r_{j,k}$, the Haar wavelets and scaling functions satisfy the following relations:

$$p_{l,k} = \frac{1}{\sqrt{2}}(p_{l+1,2k} + p_{l+1,2k+1}) \tag{4.1}$$

and

$$h_{l,k} = \frac{1}{\sqrt{2}}(p_{l+1,2k} - p_{l+1,2k+1}). \tag{4.2}$$

The above equations, (4.1) and (4.2), are perhaps our most important observation about the Haar systems. We shall further exploit it later.

Lemma 4.1.3. *Let $j \in \mathbb{Z}$ and let $g_j(x)$ be a step function, constant on all dyadic intervals $I_{j,k}$ $k \in \mathbb{Z}$. Then $g_j(x) = r_{j-1}(x) + g_{j-1}(x)$ where*

$$r_{j-1}(x) = \sum_k a_{j-1}(k)h_{j-1,k}(x),$$

for some coefficients $\{a_{j-1}(k)\}_{k\in\mathbb{Z}}$, where $g_{j-1}(x)$ is a scale $j-1$ dyadic step function.

Proof. Let $g_j(x) = c_j(k)$ on $I_{j,k}$. Define $g_{j-1}(x)$ on $I_{j-1,k}$ by the following formula:

$$g_{j-1}(x) = \begin{cases} 2^{j-1}\int_{I_{j-1,k}} g_j(t)dt = \frac{1}{2}(c_j(2k) + c_j(2k+1)), & \text{for } x \in I_{j-1,k}, \\ 0, & \text{otherwise.} \end{cases}$$

Let $r_{j-1}(x) = g_j(x) - g_{j-1}(x)$, and observe that $\int_{I_{j-1,k}} r_{j-1}(x)dx = 0$ and so on $I_{j-1,k}$, r_{j-1} is a constant multiple of the Haar function $h_{j-1,k}(x)$. □

Theorem 4.1.1. *The Haar wavelet system $\{h_{j,k} : j,k \in \mathbb{Z}\}$ on $\mathbb{R}$ is an orthonormal basis for $L^2(\mathbb{R})$.*

We shall only look at the sketch of the proof of orthonormality of the Haar system on $\mathbb{R}$, leaving the completeness part for Advanced Calculus classes.

First, fix $j \in \mathbb{Z}$ and suppose that $k \neq k'$. We have that $h_{j,k}(x)h_{j,k'}(x) = 0$ for all $x \in R$, since the function $h_{j,k}$ assumes non-zero values only in the interval $I_{j,k}$, and for $k \neq k'$,

$$I_{j,k} \cap I_{j,k'} = \emptyset.$$

If $k = k'$, then

$$< h_{j,k}, h_{j,k} >= \int_{\mathbb{R}} |h_{j,k}(x)|^2 \, dx = 1.$$

Assume now that $j \neq j'$. Without loss of generality it is sufficient to consider the case $j > j'$. Then, it follows from the first two Lemmas that there are 3 distinct possibilities: $I_{j,k} \cap I_{j',k'} = \emptyset$, or $I_{j,k} \subset I^l_{j',k'}$, or $I_{j,k} \subset I^r_{j',k'}$.

Case $I_{j,k} \cap I_{j',k'} = \emptyset$ is elementary, as in this case $h_{j,k}(x)h_{j',k'}(x) = 0$ and so

$$< h_{j,k}, h_{j',k'} >= 0.$$

Case $I_{j,k} \subset I^l_{j',k'}$ implies that whenever $h_{j,k}$ is non-zero, $h_{j',k'}$ is constantly equal to 1. Thus,

$$< h_{j,k}, h_{j',k'} >= \int_{I_{j,k}} h_{j,k}(x)h_{j',k'}(x) \, dx = \int_{I_{j,k}} h_{j,k}(x) \, dx = 0.$$

Similarly, the case $I_{j,k} \subset I^r_{j',k'}$ implies that whenever $h_{j,k}$ is non-zero, $h_{j',k'}$ is constantly equal to -1, and so,

$$< h_{j,k}, h_{j',k'} >= \int_{I_{j,k}} h_{j,k}(x)h_{j',k'}(x) \, dx = -\int_{I_{j,k}} h_{j,k}(x) \, dx = 0.$$

This completes the proof of orthonormality of the Haar wavelet system on $\mathbb{R}$.

For the Haar wavelet systems of scale J, we have analogous result.

Theorem 4.1.2. *The Haar wavelet systems of scale J, $\{p_{J,k}, h_{j,k} : k \in \mathbb{Z}, j \geq J\}$, on $\mathbb{R}$ is an orthonormal basis for $L^2(\mathbb{R})$.*

(The proof of this result is similar to the previous one.)

4.1.2 Haar wavelets on $[0, 1]$

Fix an integer $J \geq 0$. The scale J Haar system on $[0, 1]$ the collection

$$\{p_{J,k}(x) : 0 \leq k \leq 2^J - 1\} \cup \{h_{j,k}(x) : j \geq J; 0 \leq k \leq 2^j - 1\}.$$

When $J = 0$ this collection will be referred to as the Haar system on $[0, 1]$.

Here, the choice of k's and the assumption about $J \geq 0$ are necessary so that the system we have created is a collection of functions which are non-zero only in the interval $[0, 1]$.

Remark. Note that each and every Haar system on $[0,1]$ consists of **both** Haar wavelet functions and Haar scaling functions. This is to compensate the fact that we have restricted the set of possible parameters j, k.

Theorem 4.1.3. *For each $J \geq 0$, the scale J Haar system on $[0,1]$ is a complete orthonormal system on $[0,1]$.*

Proof. First approximate a function $f \in L^2[0,1]$ by a scale j dyadic step function g_j with $\|f - g_j\|_\infty$ small for some $j > 0$. Assume without loss of generality that $j \geq J$. Write $g_j = \sum_{\ell=J}^{j-1} r_\ell(x) + g_\ell(x)$. □

For $j \in \mathbb{Z}$ define the approximation operator P_j on $L^2(\mathbb{R})$ by

$$P_j(x) = \sum_{k \in \mathbb{Z}} \langle f, p_{j,k} \rangle p_{j,k}(x).$$

The approximation space

$$V_j = \overline{span}\{p_{j,k}\}_{k \in \mathbb{Z}}.$$

$P_j f$ is the best approximation of $f \in L^2(\mathbb{R})$ by an element of V_j.

Lemma 4.1.4. *(a) P_j is an orthogonal projection.*

(b) For $j \leq j'$ and $g \in V_j$, $P_{j'} g(x) = g(x)$.

(c) $\|P_j f\|_2 \leq \|f\|_2$

(d) If $f \in \mathcal{C}_c$, $\lim_{j \to \infty} \|P_j f - f\|_2 = 0$ and $\lim_{j \to \infty} \|P_{-j} f\|_2 = 0$.

or each $j \in \mathbb{Z}$ define the detail operator Q_j by

$$Q_j f(x) = P_{j+1} f - P_j f(x)$$

for $f \in L^2(\mathbb{R})$. And let the wavelet space W_j be given by

$$W_j = \overline{span}\{h_{j,k} k \in \mathbb{Z}\}$$

$Q_j f$ is the best approximation in W_j of $f \in L^2(\mathbb{R})$.

Lemma 4.1.5. *(a) Q_j is an orthogonal projection.*

(b) For $j \neq j'$ and $g \in W_j$, $Q_{j'} g(x) = 0$.

(c) $\|Q_j f\|_2 \leq \|f\|_2$.

(d) If $f \in \mathcal{C}_c$,

$$Q_j f(x) = \sum_{k \in \mathbb{Z}} \langle f, h_{j,k} \rangle h_{j,k}(x).$$

4.1.3 Discrete Haar Transform (DHT)

Recall that for a fixed integer $J \geq 0$, we can define the scale J Haar system on $[0,1]$ to be the collection

$$\{p_{J,k}(x) : 0 \leq k \leq 2^J - 1\} \cup \{h_{j,k}(x) : j \geq J; 0 \leq k \leq 2^j - 1\}.$$

We typically assume $J = 0$ for convenience.

Let f be a signal of finite energy, which we want to numerically analyze. To motivate the Discrete Haar Transform (DHT) assume that f is supported on $[0,1]$ and that it has the following Haar decomposition:

$$f(x) = \sum_{j=J}^{\infty} \sum_{k=0}^{2^j-1} \langle f, h_{j,k} \rangle h_{j,k}(x) + \sum_{k=0}^{2^J-1} \langle f p_{J,k} \rangle p_{J,k}(x)$$

in $L^2[0,1]$. Suppose that for a fixed $N > J$,

$$f(x) \simeq P_N f(x) = \sum_{k=0}^{2^N-1} \langle f, p_{N,k} \rangle p_{N,k}(x).$$

The idea is to approximate the Haar coefficients of f by those of $P_N f$. This is not an unreasonable idea in the situations, when our data has some degree for finest resolution, here represented by integer N. On the other hand, the scale J represents the coarsest possible resolution.

Consider $\{c_0(k)\}_{k=0}^{2^N-1}$ where $c_0(k) = \langle f, p_{N,k} \rangle$. This is our starting finite sequence of length 2^N. One may think of this sequence as of a finite approximation on dyadic intervals of scale N, to a given signal f of finite energy. Alternatively, we can identify the coefficients $c_0(k)$ with the values of the finest piece-wise constant approximation to signal f.

Suppose now that for the previously fixed scale $J \in \mathbb{N}, J < N$, we pick $K \in \mathbb{N}$ such that $N - K = J$. Next, for each $1 \leq j \leq K$ we define the following coefficients:

$$c_j(k) = \langle f, p_{N-j,k} \rangle \quad \text{and} \quad d_j(k) = \langle f, h_{N-j,k} \rangle,$$

where $k = 0, \ldots, 2^{N-j} - 1$. This definition is consistent with the definition of c_0:

$$c_0(k) = \langle f, p_{N,k} \rangle = \langle f, p_{N-0,k} \rangle, \quad k = 0, \ldots, 2^N - 1.$$

We use now the earlier observation, see equations (4.1) and (4.2), that

$$p_{\ell,k} = \tfrac{1}{\sqrt{2}}(p_{\ell+1,2k} + p_{\ell+1,2k+1})$$

and

$$h_{\ell,k} = \tfrac{1}{\sqrt{2}}(p_{\ell+1,2k} - p_{\ell+1,2k+1}),$$

to obtain the following relationship for the coefficients c_j and d_j:

$$c_j(k) = \frac{1}{\sqrt{2}}(c_{j-1}(2k) + c_{j-1}(2k+1))$$

and

$$d_j(k) = \frac{1}{\sqrt{2}}(c_{j-1}(2k) - c_{j-1}(2k+1)).$$

These equations can be rewritten equivalently in the matrix-vector form:

$$\begin{bmatrix} c_j(k) \\ d_j(k) \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} c_{j-1}(2k) \\ c_{j-1}(2k+1) \end{bmatrix}.$$

Conversely, by taking the inverse of the above matrix (which has a non-zero determinant, and so, is invertible), we can re-write the matrix-vector equation above as follows:

$$\begin{bmatrix} c_{j-1}(2k) \\ c_{j-1}(2k-1) \end{bmatrix} = \frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} c_j(k) \\ d_j(k) \end{bmatrix}.$$

Here we want to emphasize the role of the matrix:

$$\frac{1}{\sqrt{2}} \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}.$$

This is an orthogonal matrix. Furthermore, the whole transformation from vector c_{j-1} to vector (c_j, d_j), can be expressed as an action of a block-diagonal matrix with these 2×2 blocks:

$$\begin{pmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} & 0 & 0 & \dots \\ 0 & \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} & 0 & \dots \\ 0 & 0 & \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix} & \dots \\ \dots & & & \end{pmatrix}.$$

As a block diagonal matrix consisting of orthogonal blocks, this is also a $(2^{N-j+1} \times 2^{N-j+1})$ orthogonal matrix. The results of its action on vector c_{j-1} is the vector:

$$(c_j(0), d_j(0), c_j(1), d_j(1), \dots, c_j(2^{N-j}-1), d_j(2^{N-j}-1))^T.$$

Because the mixing of c_j and d_j coefficients can be sometimes viewed as suboptimal, we would like to permute the rows of the matrix to obtain the following transformation:

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 1 & 0 & \dots \\ \dots & & & & & \\ 0 & \dots & \dots & 0 & 1 & 1 \\ 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots \\ \dots & & & & & \\ 0 & \dots & \dots & 0 & 1 & -1 \end{pmatrix}.$$

The permutation of rows preserves the orthogonality of the matrix. At the same time the advantage is that the output of th eabove transformation is now exactly:

$$(c_j, d_j)^T.$$

In order to simplify the notation, we introduce the following auxiliary $L/2 \times L$ matrices, where L represents any integer of the form 2^{N-j+1}, where $j = 1, \dots, K$:

$$H_j = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & 1 & 0 & \dots \\ \dots & & & & & \\ 0 & \dots & \dots & 0 & 1 & 1 \end{pmatrix}$$

and

$$G_j = \frac{1}{\sqrt{2}}\begin{pmatrix} 1 & -1 & 0 & 0 & \dots & 0 \\ 0 & 0 & 1 & -1 & 0 & \dots \\ \dots & & & & & \\ 0 & \dots & \dots & 0 & 1 & -1 \end{pmatrix}.$$

Definition 4.1.6 (Single Level of Discrete Haar Transform)**.** Given integers $N > J \geq 0$, we define the following $2^{N-j+1} \times 2^{N-j+1}$ matrix, where $j = 1, \dots, N - J$:

$$\begin{pmatrix} H_j \\ G_j \end{pmatrix},$$

to be the matrix representing a *single level of the Discrete Haar Transform.* In particular,

$$\begin{pmatrix} H_j \\ G_j \end{pmatrix} [c_{j-1}] = \begin{bmatrix} c_j \\ d_j \end{bmatrix},$$

for $j = 1, \dots, N - J$.

In the future our convention will be to drop the index j in H_j and G_j, whenever the dimensions of matrices are clear. Then we will simply use just H and G.

Example 4.1.1. In the case when $J = 0$, matrices H and G have the dimensions ranging from 1×2 to $2^{N-1} \times 2^N$.

Definition 4.1.7 (Discrete Haar Transform)**.** Given integers $N > J \geq 0$ and a finite sequence $c_0 = \{c_0(k)\}_{k=0}^{2^N-1}$. We define the output of the *Discrete Haar Transform (DHT)* of c_0 to be the collection of vectors:

$$\{c_{N-J}(k) : 0 \leq k \leq 2^{N-J} - 1\} \cup \{d_j(k) : 1 \leq j \leq N - J; 0 \leq k \leq 2^{N-j} - 1\},$$

where

$$c_j(k) = \tfrac{1}{\sqrt{2}}(c_{j-1}(2k) + c_{j-1}(2k+1)),$$
$$d_j(k) = \tfrac{1}{\sqrt{2}}(c_{j-1}(2k) - c_{j-1}(2k+1)).$$

In the vector notation, the **Discrete Haar Transform (DHT)** produces the vector

$$(c_{N-J},\ d_{N-J},\ d_{N-J+1},\ \ldots\ d_1)^T,$$

where, as before, each d_j is a vector of length 2^{N-j}, $1 \leq j \leq N - J$, and c_{N-J} is a vector of length 2^{N-J}.

In the extreme case when $J = 0$, $d_{N-J} = d_N$ and $c_{N-J} = c_N$ are both just numbers (vectors of length 1).

The inverse of DHT is computed by iteratively applying the formula:

$$c_{j-1} = H^*(c_j) + G^*(d_j),$$

up to $j = 1$. This is related to the fact that the matrix

$$\frac{1}{\sqrt{2}}\begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}$$

is its own inverse; hence, the inverse of

$$\begin{pmatrix} H \\ G \end{pmatrix}$$

is (H^*, G^*).

Example 4.1.2. In many situations labeling the Haar transform by means of scales is inconvenient. This is in particular true, when we are given a vector of length which is a power of 2 and we just want to compute the complete **Full Discrete Haar Transform**, for all possible coefficients, assuming that $J = 0$. In such cases we can re-label the previous notation as follows.

Given any integer L of the form $L = 2^N$, and $x = (x(0), \ldots, x(L-1)) \in \mathbb{R}^L$, let the *Full DHT* be defined as:

$$H^N(x)(n) = c^N(k) = \frac{1}{\sqrt{2}}(x(2k) + x(2k+1)),$$

for $k \in \{0, 1, \ldots, L/2 - 1\}$, and

$$G^N(x)(n) = d^N(k - N/2) = \frac{1}{\sqrt{2}}(x(2(k - \tfrac{L}{2})) - x(2(k - \tfrac{L}{2}) + 1)),$$

for $k \in \{L/2, L/2 + 1, \ldots, L - 1\}$. Next, for any vector of length 2^q, apply H_q, G_q, $q = N, N - 1, \ldots, 1$, consecutively N times to the resulting sequence of coefficients c^q. Define the output of the full Haar transform as a vector of length L:

$$(c^1, d^1, d^2, \ldots, d^N).$$

Note that except for c^1 and d^1, all other entries in the above are vectors.

Example 4.1.3. Use $x = (1, 2, 3, 4)$ and observe that the difference between the Haar transform and standard sum or difference transform is the specific structure of consecutive transformations acting on previous outputs.

4.2 Filtering Transforms

In signal processing, a *filter* is often a device or a process that modifies and transforms the given signal. Such modification can for example take the form of removal of certain undesired components or features from a signal, like noise. In other cases filtering is removing some specified frequencies, also known as frequency bands. However, filters can do much more, especially in the field of image processing. Filters are a popular tool in a range of applications, from electronics and telecommunication, to radio, television, audio recording, radar, control systems, music synthesis, image processing, and to computer graphics.

Some examples of commonly used filters include:

- *Low-pass filter*: retains low frequencies while high frequencies are suppressed.
- *High-pass filter*: retains high frequencies while low frequencies are suppressed.
- *Band-pass filter*: retains frequencies in a specified frequency band.
- *Band-stop filter*: frequencies in a specified frequency band are suppressed.
- *Notch filter*: suppresses a single frequency.

Given a sequence $f = \{f(k)\}_{k\in\mathbb{Z}}$, we say that its support is defined as the smallest interval which contains all integers $k \in \mathbb{Z}$ such that $f(k) \neq 0$. If the support of a sequence is a finite interval $[a, b]$, where $a, b \in Z$, we say that the sequence f is *finitely supported*. Otherwise, the support of f is an infinite set.

Please note that this allows for f to take some zero values inside the supporting interval (a, b). For example, the sequence f defined as follows: $f(k) = 0$ for $k < 0$, $f(0) = 1$, $f(k) = 0$ for $0 < k < 10$, $f(10) = 1$, $f(k) = 0$ for $k > 10$, has the interval $[0, 10]$ as its support, even though it takes zero values inside this interval.

Definition 4.2.1. Let $f = \{f(k)\}_{k\in\mathbb{Z}}$ be a finitely supported sequence. We define the *filtering transform* F, defined on sequences $u = \{u(k)\}_{k\in\mathbb{Z}}$ to be

$$\forall n \in \mathbb{Z}, \quad F(u)(n) = \sum_{k\in\mathbb{Z}} f(k - 2n)u(k).$$

Equivalently, by a discrete change of variables, we have

$$\forall n \in \mathbb{Z}, \quad F(u)(n) = \sum_{k\in\mathbb{Z}} f(k)u(k + 2n).$$

Indeed, consider the change $k - 2n \to k'$. Then $k \to k' + 2n$ and

$$F(u)(n) = \sum_{k\in\mathbb{Z}} f(k - 2n)u(k) = \sum_{k'\in\mathbb{Z}} f(k')u(k' + 2n) = \sum_{k\in\mathbb{Z}} f(k)u(k + 2n),$$

where the last equality follows from the fact that k and k' both represent the index of summation in the series and can be freely exchanged by any other symbol representing the index of summation.

We say that f is the *filter* associated with the transform F.

Example 4.2.1. Consider the filter h defined using the Haar sequence:

$$h(k) = \begin{cases} 1/\sqrt{2} & k = 0, 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then, given a sequence $u = \{u(k) : k \in \mathbb{Z}\}$, using the latter formula for F, we see that the associated filtering transform H takes form of:

$$H(u)(n) = \sum_{k\in\mathbb{Z}} h(k)u(k+2n) = h(0)u(2n) + h(1)u(2n+1) = \frac{1}{\sqrt{2}}(u(2n) + u(2n+1)).$$

This is an example where using the other formulation for the filtering transform may be advantageous. Using the original definition would require a change of summation.

Similarly, for the Haar filter g defined using the other Haar sequence:

$$g(k) = \begin{cases} 1/\sqrt{2} & k = 0, \\ -1/\sqrt{2} & k = 1, \\ 0 & \text{otherwise.} \end{cases}$$

the associated filtering transform G takes form of:

$$G(u)(n) = \sum_{k\in\mathbb{Z}} g(k)u(k+2n) = g(0)u(2n) + g(1)u(2n+1) = \frac{1}{\sqrt{2}}(u(2n) - u(2n+1)).$$

The above example demonstrate that even though the filter can be finitely supported, the output of the associated filtering transformation does not have to be.

Remark. *Please note that in some cases it is advantageous to use symbols, rather than exact formulas. In those cases, we shall use the following convention: Lower case letters, like f, g, h, denote filters (filter sequences) and upper case letters denote their associated filtering transformations, like F, G, H.*

Because f has a finite support, F is a linear transformation which is well defined on every sequence space. Because of the finite support of f, the second formula above shows that the sum is always finite. Therefore, without loss of generality we may assume that the space of interest for us is the space $\ell^2(\mathbb{Z})$ - the space of square summable sequences on $\mathbb{Z}$:

$$\ell^2(\mathbb{Z}) = \left\{ u : \mathbb{Z} \to \mathbb{C}, \quad s.t. \quad \sum_{k\in\mathbb{Z}} |u(k)|^2 < \infty \right\}.$$

This space has a well defined inner product:

$$\langle u, v\rangle = \sum_{k\in\mathbb{Z}} u(k)\overline{v(k)}.$$

In this space, F has an adjoint operator F^*, which satisfies the relationship:

$$\langle F(u), v\rangle = \langle u, F^*(v)\rangle. \tag{4.3}$$

Using this definition we can derive the explicit formula for F^*:

$$F^*(v)(k) = \sum_{n\in\mathbb{Z}} \overline{f(k-2n)}v(n). \tag{4.4}$$

Guessing the formula for F^* is not difficult, but having the formula (4.4) given, it is very straightforward to verify that it is correct. Indeed, we need to consider the left and right sides of the equation (4.3). First, note that for the left-had side in (4.3) we have:

$$\begin{aligned}\langle F(u), v\rangle &= \sum_{n\in\mathbb{Z}} F(u)(n)\overline{v(n)} = \sum_{n\in\mathbb{Z}} \left(\sum_{k\in\mathbb{Z}} f(k-2n)u(k)\right)\overline{v(n)} \\ &= \sum_{n\in\mathbb{Z}}\sum_{k\in\mathbb{Z}} f(k-2n)u(k)\overline{v(n)}.\end{aligned}$$

Next, for the right-had side in (4.3) we have:

$$\begin{aligned}\langle u, F^*(v)\rangle &= \sum_{k\in\mathbb{Z}} u(k)\overline{F^*(v)(k)} = \sum_{k\in\mathbb{Z}} u(k)\overline{\left(\sum_{n\in\mathbb{Z}} \overline{f(k-2n)}v(n)\right)} \\ &= \sum_{k\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} u(k)f(k-2n)\overline{v(n)},\end{aligned}$$

where we used the fact that for a complex number z: $\overline{\overline{z}} = z$. Thus, we now observe that both sides in (4.3) are equal, after a change of summation. Note that even though we use the series notation, all involved sums are finite, so we can always change the order of summations.

Using a discrete change of variables we can rewrite the above formula for the adjoint of the filter transformation as:

$$F^*(v)(k) = \begin{cases} \sum_{n\in\mathbb{Z}} \overline{f(2n)}v\left(\frac{k}{2} - n\right), & \text{if } \ k \ \text{ even}, \\ \sum_{n\in\mathbb{Z}} \overline{f(2n+1)}v\left(\frac{k-1}{2} - n\right), & \text{if } \ k \ \text{ odd}. \end{cases}$$

It is worth noting that in this example the discrete substitution results in a more complicated formula, since we need to remember that the output of the substitution formula must be an integer.

To verify that the alternative definition of the adjoint formula is correct, it now suffices to only consider the right-hand side in (4.3), as we have already simplified its

left-hand side. In the first step, we split the sum into even and odd coefficients and apply separate formulas for F^* in each case:

$$\begin{aligned}
\langle u, F^*(v)\rangle &= \sum_{k\in\mathbb{Z}} u(k)\overline{F^*(v)(k)} = \sum_{k\in\mathbb{Z}} u(2k)\overline{F^*(v)(2k)} + \sum_{k\in\mathbb{Z}} u(2k+1)\overline{F^*(v)(2k+1)} \\
&= \sum_{k\in\mathbb{Z}} u(2k)\overline{\sum_{n\in\mathbb{Z}} \overline{f(2n)} v\left(\frac{2k}{2} - n\right)} \\
&\quad + \sum_{k\in\mathbb{Z}} u(2k+1)\overline{\sum_{n\in\mathbb{Z}} \overline{f(2n+1)} v\left(\frac{(2k+1)-1)}{2} - n\right)} \\
&= \sum_{k\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} u(2k) f(2n)\overline{v\left(\frac{2k}{2} - n\right)} \\
&\quad + \sum_{k\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} u(2k+1) f(2n+1)\overline{v\left(\frac{(2k+1)-1)}{2} - n\right)} \\
&= \sum_{k\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} u(2k) f(2n)\overline{v\,(k-n)} + \sum_{k\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} u(2k+1) f(2n+1)\overline{v\,(k-n)}.
\end{aligned}$$

In the last two steps we have simplified the formula. Now, we use the following discrete substitution in both sums over n: $k - n \to n'$. Then, $n \to k - n'$ and the right-hand side becomes:

$$\begin{aligned}
&\sum_{k\in\mathbb{Z}}\sum_{n'\in\mathbb{Z}} u(2k) f(2(k-n'))\overline{v\,(n')} \\
&\quad + \sum_{k\in\mathbb{Z}}\sum_{n'\in\mathbb{Z}} u(2k+1) f(2(k-n')+1)\overline{v\,(n')} \\
&= \sum_{k\in\mathbb{Z}}\sum_{n'\in\mathbb{Z}} u(2k) f(2k-2n'))\overline{v\,(n')} \\
&\quad + \sum_{k\in\mathbb{Z}}\sum_{n'\in\mathbb{Z}} u(2k+1) f((2k+1)-2n')\overline{v\,(n')} \\
&= \sum_{k\in\mathbb{Z}}\sum_{n\in\mathbb{Z}} u(k) f(k-2n))\overline{v\,(n')},
\end{aligned}$$

where in the last step we have combined two sums over even and odd integers $2k$ and $2k+1$, into one sum over $k \in \mathbb{Z}$. We also relabeled n' as n to match the original adjoint formula and we conclude that they are both equal.

We have noted after Example 4.2.1 that the output of a finitely supported filtering transform needs not be finitely supported. However, it is not difficult to see that if both the filter and the input sequence are finitely supported, then so it the output of the filtering transform. In fact, knowing the supports of the filter and the input, we can estimate the support of the output sequence, as is demonstrated by the following result. In the process we shall use the following notation:

$$\lfloor x \rfloor = \max\{m \in \mathbb{Z} : m \le x\}$$

and

$$\lceil y \rceil = \min\{n \in \mathbb{Z} : n \geq y\}.$$

These are known as the so-called *floor and ceiling functions*.

Lemma 4.2.1. *Let $a, b, x, y \in \mathbb{Z}$. Suppose that the filter f is supported on the interval $[a, b]$ and that the sequence u is supported on the interval $[x, y]$. Then:*

- *$F(u)$ is supported on the interval $[\lceil \frac{x-b}{2} \rceil, \lfloor \frac{y-a}{2} \rfloor]$.*
- *$F^*(u)$ is supported on the interval $[a + 2x, b + 2y]$.*

Proof. We are only going to verify one of these statements, leaving the other to the interested reader, as the technicalities involved in both arguments are very similar.

Consider

$$F(u)(n) = \sum_{k \in \mathbb{Z}} f(k)u(k + 2n).$$

Since u is supported on the interval $[x, y]$, then the shifted sequence $u(\cdot + 2n)$ is supported on the interval $[x - 2n, y - 2n]$. Thus, since f is supported on $[a, b]$, $F(u)(n) = 0$ whenever the two supports are disjoint, i.e., when $b < x - 2n$ or $y - 2n < a$. This can be rewritten as $n < (x - b)/2$ and $n > (y - a)/2$, respectively. Because n must be an integer, we obtain the inequalities:

$$n < \left\lceil \frac{x-b}{2} \right\rceil \quad \text{and} \quad n > \left\lfloor \frac{y-a}{2} \right\rfloor.$$

It remains to show that $F(u)(\lceil \frac{x-b}{2} \rceil) \neq 0$ and $F(u)(\lfloor \frac{y-a}{2} \rfloor) \neq 0$, which can be done by a simple analysis considering the parity of $x - b$ and $y - a$. □

Example 4.2.2. Consider now the case of one of Haar-type filters of length 2, h or g froom Example 4.2.1. Without loss of generality we may assume that both h and g are supported on $[0, 1]$. Suppose that the input signal u has length equal to 4 and is supported on $[0, 3]$. According to our formula in Lemma 4.2.1, the support of $H(u)$ is $[\lceil \frac{0-1}{2} \rceil, \lfloor \frac{3-0}{2} \rfloor] = [\lceil \frac{-1}{2} \rceil, \lfloor \frac{3}{2} \rfloor] = [0, 1]$. This implies that in this case both H and G are transformations from $\mathbb{R}^4$ into $\mathbb{R}^2$.

Furthermore, it is not difficult to see that the matrices of these filtering transforms take the form of:

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}$$

and

$$G = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & -1 & 0 & 0 \\ 0 & 0 & 1 & -1 \end{pmatrix}.$$

Compare this with the definitions of the Haar transform.

Example 4.2.3. Suppose now that we modify the previous example and consider a filter f of length 4, supported on $[0, 3]$. Then, for the input signal u which also has length equal to 4 and is supported on $[0, 3]$, the filtering transform $F(u)$ is supported on the interval $[-1, 1]$, i.e., it maps a vector from $\mathbb{R}^4$ into $\mathbb{R}^3$.

Furthermore, it is not difficult to see that the matrix equivalent of such a filtering transform takes the form of:

$$F = \begin{pmatrix} f(2) & f(3) & 0 & 0 \\ f(0) & f(1) & f(2) & f(3) \\ 0 & 0 & f(0) & f(1) \end{pmatrix}.$$

So, we have a lot of zeros that we would like to avoid. In what follows, we will present a way to do that.

Because the filter f is finite, and because its output is also finite, we may define a related filtering transform that operates on finite dimensional vectors rather than on infinite sequences. One way to do so was just demonstrated in the examples above.

We shall now present an alternative approach. The first step is to assume that the filter is given as a periodic sequence. Not to create any confusion with the previous case, we shall use the notation f_P for a periodic filter with period P. The process of creating a periodic sequence is known as periodization and it can be described using the following formula:

$$u_P(m) = \sum_{k \in \mathbb{Z}} f(m + kP),$$

where $\{u(k) : k \in \mathbb{Z}\}$ is a given summable sequence. We note that if the sequence u is supported on $[0, P - 1]$, then the periodization formula above coincides with the shifts of the nonzero values of u by integer multiples of P.

We can now define a periodic filtering transformation acting on P-periodic sequences u_P via:

$$F_P(u_P)(n) = \sum_{k=0}^{P-1} f_P(k - 2n)u_P(k).$$

We note that if P is even, then $F_P(u_P)$ is $P/2$-periodic, and so we let $P = 2p$ and note that

$$F_{2p}(u_{2p})(n) = \sum_{k=0}^{2p-1} f_{2p}(k - 2n)u_{2p}(k) = \sum_{k=0}^{2p-1} f_{2p}(k)u_{2p}(k + 2n).$$

Indeed,

$$F_{2p}(u_{2p})(n+p) = \sum_{k=0}^{2p-1} f_{2p}(k-2n-2p)u_{2p}(k) = \sum_{k=0}^{2p-1} f_{2p}(k-2(n))u_{2p}(k) = F_{2p}(u_{2p})(n).$$

Then, similarly as before, we compute the adjoint of F_{2p} to be:

$$F_{2p}^*(v_{2p})(k) = \sum_{n=0}^{p-1} f_{2p}(k-2n)v(n)$$

With a change of variables we get that

$$F_{2q}^*(v)(k) = \begin{cases} \sum_{n=0}^{q-1} f_{2q}(2n)\overline{v(\frac{k}{2}-n)}, & if \quad k \in [0, 2q-2] \quad even, \\ \sum_{n=0}^{q-1} f_{2q}(2n+1)\overline{v(\frac{k-1}{2}-n)}, & if \quad k \in [1, 2q-1] \quad odd. \end{cases}$$

Finally we observe that periodization commutes with filtering transforms:

$$(F(u))_p = F_{2p}(u_{2p})$$

and

$$(F^*(v))_{2p} = F_{2p}^*(v_{2p}).$$

Example 4.2.4. Suppose again that we use the previous example and consider a filter f of length 4, supported on $[0, 3]$, and the input signal u which also has length 4 and is supported on $[0, 3]$. Next, we periodize both sequences with period $P = 4$. According to our observation above, the resulting periodic filtering transform of the signal u is 2-periodic and the associated matrix takes the form of the following 2×4 matrix:

$$F = \begin{pmatrix} f(0) & f(1) & f(2) & f(3) \\ f(2) & f(3) & f(0) & f(1) \end{pmatrix}.$$

We note that this matrix essentially contains the same information as the 3×4 matrix for the original filtering transform, but it has fewer zero coefficients.

4.3 Multiresolution Analysis and Filters

Motivated by the example of Haar system, we now introduce the following definition.

Definition 4.3.1. Let $\{V_j : j \in \mathbb{Z}\}$ be a sequence of subspaces of $L^2(X)$. The collection $\{V_j : j \in \mathbb{Z}\}$ is called a *Multiresolution Analysis* (MRA) with scaling function φ if the following conditions hold:

(1) $V_j \subset V_{j+1}$ for each $j \in \mathbb{Z}$

(2) $\overline{\cup_{j=-\infty}^{\infty} V_j} = L^2(\mathbb{R})$

(3) $\cap_{j=-\infty}^{\infty} V_j = \{0\}$

(4) The function $f(x) \in V_j$ if and only if the function $f(2^{-j}x) \in V_0$

(5) The function $\varphi \in V_0$ and the set $\{\varphi(x-k), k \in \mathbb{Z}\}$ is an orthonormal basis for V_0.

It is not difficult to see that the function $p = p_{0,0}$ and the spaces V_j defined using Haar functions, form an MRA.

Our goal is not to introduce a transformation that will be more general than the concept of the Discrete Haar Transform. We will achieve this by means of generalizing the Haar ideas by replacing the sums and differences with more general filtering transforms.

In this regard, we begin our construction by noting that, because $V_0 \subset V_1$ and because $\{D_2 T_k \varphi : k \in \mathbb{Z}\}$ forms an ONB per MRA conditions, we have that $\varphi \in V_1$ and

$$\varphi(x) = \sum_k h(k) D_2 T_k \varphi(x) = \sum_k h(k)\sqrt{2}\varphi(2x-k). \tag{4.5}$$

If we assume that φ has a compact support, then necessarily for all but finitely many k, $h(k) = 0$. This sequence is called the *low pass filter*. Thus we can define the following filtering transform:

$$H(u)(n) = \sum_{k\in\mathbb{Z}} h(k-2n)u(k) = \sum_{k\in\mathbb{Z}} h(k)u(k+2n).$$

We begin by observing that this filter satisfies a range of properties, all due to the fact that it is derived form an MRA.

Theorem 4.3.1 (Finiteness)**.** *Assume that φ has a compact support. Then, the sequence $\{h(k) : k \in Z\}$ is different form zero for all but finitely many k's.*

Theorem 4.3.2 (Self-Orthonormality)**.**

$$\forall\ m, n \in \mathbb{Z}, \quad \sum_k h(k+2n)\overline{h(k+2m)} = \delta_{m,n}.$$

Proof. We begin by noting that due to orthogonality of the set the set $\{\varphi(x-k), k \in \mathbb{Z}\}$, we have:

$$\begin{aligned} 0 = \langle \varphi(x), \varphi(x-k) \rangle &= \sum_k \sum_l h(k)\overline{h(l)} \langle \sqrt{2}\varphi(2x-k), \sqrt{2}\varphi(2x-2n-l) \rangle \\ &= \sum_k \sum_l h(k)\overline{h(l)}\delta_{k,2n-l} = \sum_k h(k)\overline{h(2n+k)}. \end{aligned}$$

Similarly

$$\begin{aligned} 1 = \langle \varphi(x), \varphi(x) \rangle &= \sum_k \sum_l h(k)\overline{h(l)} \langle \sqrt{2}\varphi(2x-k), \sqrt{2}\varphi(2x-l) \rangle \\ &= \sum_k \sum_l h(k)\overline{h(l)}\delta_{k,l} = \sum_k h(k)\overline{h(k)} = \sum_k |h(k)|^2. \end{aligned}$$

This essentially finishes the proof. □

Theorem 4.3.3 (Normalization)**.** *Assume that φ is integrable. Then, the sequence $\{h(k) : k \in Z\}$ satisfies the following:*

$$\sum_k h(2k) = \sum_k h(2k+1) = \frac{1}{\sqrt{2}}.$$

In particular,

$$\sum_k h(k) = \sqrt{2}.$$

Proof. Integrating both sides of the (4.5), we obtain:

$$\int_X \varphi(x)\,dx = \sum_k h(k)\sqrt{2} \int_X \varphi(2x-k)\,dx = \sum_k h(k)\frac{1}{\sqrt{2}} \int_X \varphi(x)\,dx.$$

Assuming the integral is different form zero, we obtain that

$$1 = \sum_k h(k)\frac{1}{\sqrt{2}} \equiv \sum_k h(k) = \sqrt{2}.$$

In order to show that even and odd sums are equal, we begin by noting that the above condition is equivalent to

$$\begin{aligned} 2 = \left| \sum_k h(k) \right|^2 &= \sum_k \sum_n h(k)\overline{h(n)} = \sum_k \sum_n h(k)\overline{h(k+n)} \\ &= \sum_k \sum_n h(k)\overline{h(k+2n)} + \sum_k \sum_n h(k)\overline{h(k+2n+1)}. \end{aligned}$$

Because of Self-Orthonormality, we have that

$$\sum_k \sum_n h(k)\overline{h(k+2n)} = 1.$$

Hence,

$$\begin{aligned}
1 &= \sum_k \sum_n h(k)\overline{h(k+2n+1)} = \sum_k h(k) \sum_n \overline{h(k+2n+1)} \\
&= \sum_k h(2k) \sum_n \overline{h(2k+2n+1)} + \sum_k h(2k+1) \sum_n \overline{h(2k+2n+2)} \\
&= \left(\sum_k h(2k)\right)\left(\sum_n \overline{h(2n+1)}\right) + \left(\sum_k h(2k+1)\right)\left(\sum_n \overline{h(2n)}\right).
\end{aligned}$$

If we let $A = \sum_k h(2k)$ and $B = \sum_k h(2k+1)$, then the above equation can be restated as:

$$A\overline{B} + \overline{A}B = 1.$$

On the other hand we already know that

$$A + B = \sqrt{2}.$$

Thus,

$$|A - B|^2 = |A + B|^2 - 2(A\overline{B} + \overline{A}B) = 2 - 2 = 0.$$

So,

$$A = B \quad \Longrightarrow \quad A = B = \frac{1}{\sqrt{2}}.$$

□

We now define another filter, related to the *low pass filter* h.

Definition 4.3.2. Let M be any fixed integer. Let $\{h(k) : k \in Z\}$ be the low pas filter. Define

$$g(k) = (-1)^k \overline{h(2M - 1 - k)}.$$

This sequence is called the *high pass filter*.

For the purposes of remaining arguments it suffices to assume the special case that $M = 1$, i.e., that

$$g(k) = (-1)^k \overline{h(1-k)}.$$

Theorem 4.3.4 (Normalization of High Pass Filter)**.** *The high pass filter sequence* $\{g(k) : k \in Z\}$ *satisfies the following:*

$$\sum_k g(2k) = -\sum_k g(2k+1) = \frac{1}{\sqrt{2}}.$$

In particular,

$$\sum_k g(k) = 0.$$

Proof.

$$\sum_k g(2k) = \sum_k (-1)^{2k}\overline{h(1-2k)} = \overline{\sum_k h(1-2k)} = \overline{\sum_k h(2k+1)} = \frac{1}{\sqrt{2}}.$$

Similarly,

$$\sum_k g(2k+1) = \sum_k (-1)^{2k+1}\overline{h(1-2k-1)} = -\overline{\sum_k h(-2k)} = -\overline{\sum_k h(2k)} = -\frac{1}{\sqrt{2}}.$$

□

Theorem 4.3.5 (Self-Orthonormality of High Pass Filter)**.**

$$\forall\ m,n \in \mathbb{Z}, \quad \sum_k g(k+2n)\overline{g(k+2m)} = \delta_{m,n}.$$

Proof.

$$\begin{aligned}
\sum_k g(k+2n)\overline{g(k+2m)} &= \sum_k (-1)^{k+2n}\overline{h(1-k-2n)}(-1)^{k+2m}h(1-k-2m) \\
&= \sum_k (-1)^k\overline{h(1-k-2n)}(-1)^k h(1-k-2m) \\
&= \sum_k \overline{h(1-k-2n)}h(1-k-2m) \\
&= \sum_k \overline{h(k-2n)}h(k-2m) \\
&= \sum_k h(k-2m)\overline{h(k-2n)} = \delta_{m,n}.
\end{aligned}$$

□

Theorem 4.3.6 (Independence)**.** *Let $\{h(k) : k \in Z\}$ be the low pas filter and let $\{g(k) : k \in Z\}$ be the high pass filter sequence. Then*

$$\forall\ m,n \in \mathbb{Z}, \quad \sum_k g(k+2n)\overline{h(k+2m)} = 0.$$

Proof.

$$\sum_k g(k+2n)\overline{h(k+2m)} = \sum_k (-1)^{k+2n}\overline{h(1-k-2n)h(k+2m)}$$

$$= \overline{\sum_k (-1)^k h(1-k-2n)h(k+2m)}$$

$$= \overline{\sum_k (-1)^{2k}h(1-2k-2n)h(2k+2m) + \sum_k (-1)^{2k+1}h(1-2k-1-2n)h(2k+1+2m)}$$

$$= \overline{\sum_k h(1-2k-2n)h(2k+2m) - \sum_k h(-2k-2n)h(2k+1+2m)}$$

$$= \overline{\sum_k h(1-2k-2n+2m)h(2k) - \sum_k h(2k)h(-2k-2n+1+2m)} = 0.$$

□

Theorem 4.3.7 (Completeness)**.** *Let $\{h(k) : k \in Z\}$ be the low pas filter and let $\{g(k) : k \in Z\}$ be the high pass filter sequence. Then*

$$\forall\, m, n \in \mathbb{Z}, \quad \sum_k h(2k+n)\overline{h(2k+m)} + \sum_k g(2k+n)\overline{g(2k+m)} = \delta_{m,n}.$$

Proof. We first use the definition of g and we let $p = m - n$ so that $m + n = p + 2n$ to obtain:

$$\begin{aligned}
&\sum_k h(2k+n)\overline{h(2k+m)} + \sum_k g(2k+n)\overline{g(2k+m)} \\
&= \sum_k h(2k+n)\overline{h(2k+m)} + \sum_k (-1)^{m+n}\overline{h(1-(2k+n))}h(1-(2k+m)) \\
&= \sum_k h(2k+n)\overline{h(2k+n+p)} + (-1)^p \sum_k \overline{h(1-2k-n))}h(1-2k-n-p)).
\end{aligned}$$

We now consider several cases.

1. If $n = 2q$ is even we perform two different substitutions in the sums to obtain
$$\sum_k h(2k)\overline{h(2k+p)} + (-1)^p \sum_k \overline{h(1+2k))}h(1+2k-p)).$$
   - If $p = 2r$ is even then a substitution $k := k - r$ in the second sum leads to:
   $$\begin{aligned}
   &\sum_k h(2k)\overline{h(2k+p)} + \sum_k h(2k+1))\overline{h(2k+1+p))} \\
   &= \sum_k h(k)\overline{h(k+p)} = \sum_k h(k)\overline{h(k+2r)} = \delta_r = \delta_{m,n}.
   \end{aligned}$$
   - If $p = 2r + 1$ is odd then a similar substitution leads to
   $$\sum_k h(2k)\overline{h(2k+p)} - \sum_k h(2k))\overline{h(2k+p))} = 0 = \delta_{m,n},$$
   because for odd $p = m - n$, necessarily $m \neq n$.

2. If $n = 2q + 1$ is odd then again performing two different substitutions on both sums leads to
$$\sum_k h(2k+1)\overline{h(2k+1+p)} + (-1)^p \sum_k \overline{h(2k))}h(2k-p)).$$
   - If $p = 2r$ is even then a substitution $k := k - r$ in the second sum leads to:
   $$\begin{aligned}
   &\sum_k h(2k+1)\overline{h(2k+1+p)} + \sum_k h(2k))\overline{h(2k+p))} \\
   &= \sum_k h(k)\overline{h(k+p)} = \sum_k h(k)\overline{h(k+2r)} = \delta_r = \delta_{m,n}.
   \end{aligned}$$

- If $p = 2r - 1$ is odd then a similar substitution leads to

$$\sum_k h(2k+1)\overline{h(2k+1+p)} - \sum_k h(2k+1))\overline{h(2k+1+p))} = 0 = \delta_{m,n},$$

because for odd $p = m - n$, necessarily $m \neq n$.

□

Definition 4.3.3 (Orthogonal Conjugate Quadrature Filters)**.** The pair of filter sequences $\{h(k) : k \in Z\}$ and $\{g(k) : k \in Z\}$, which are finitely supported, which satisfy the relationship from Definition 4.3.2 and which satisfy the conditions from Theorems 4.3.2, 4.3.3, 4.3.4, 4.3.5, 4.3.6, 4.3.7, are called *Orthogonal Conjugate Quadrature Filters* (OCQF).

Definition 4.3.4. Let $\{f(k), k \in \mathbb{Z}\}$ be a finitely supported filter sequence. We say that the support of the filter sequence f is the smallest integer interval which contains all the non-zero values of f. We say that the length of the support of the filter sequence f is the number of integers in this interval.

For example, let $f(k) = 0$ for all $k < 0$, let $f(0) \neq 0$, let $f(k) = 0$ for all $k > L$ and let $f(L) \neq 0$. Then the support of this filter sequence is $[0, L]$ and the length is $L + 1$.

Theorem 4.3.8. *Let $\{h(k), k \in \mathbb{Z}\}$ and $\{g(k), k \in \mathbb{Z}\}$ be a pair of finitely supported Orthogonal Conjugate Quadrature Filters. Then the length of each filter must always be an even number.*

4.4 Discrete Wavelet Transforms

Our goal is to construct a generalization of the Haar transform by replacing the averaging and difference filters, with more general transformations. The concept of a generalization has been presented in the previous section, where we introduced the Orthogonal Conjugate Quadrature Filters (OCQFs). It is not difficult to see that the filter sequences

$$h(k) = \frac{1}{\sqrt{2}},\ k = 0, 1, \quad \textit{and} \quad h(k) = 0,\ k \neq 0, 1,$$

and

$$g(0) = \frac{1}{\sqrt{2}},\ g(1) = -\frac{1}{\sqrt{2}}, \quad \textit{and} \quad g(k) = 0,\ k \neq 0, 1,$$

are the simplest example of an OCQF pair.

An equally simple calculation shows actually something stronger.

Theorem 4.4.1. *The Haar filter pair is the only pair of OCQFs of length 2, up to a translation.*

Proof. Without loss of generality assume that the nonzero filter coefficients are $h(0)$ and $h(1)$, resp. $g(0)$ and $g(1)$. Then the normalization conditions tell us that $h(0) = h(1)$ and $g(0) = -g(1)$. Furthermore the 0th coefficients are the only nonzero even coefficients and the 1st coefficients are the only nonzero odd coefficients. As such, $h(0) = h(1) = \frac{1}{\sqrt{2}}$ and $g(0) = -g(1) = \frac{1}{\sqrt{2}}$. □

Our next task is to find the OCQ filters of length 4. Since the high pass filter is always determined by the low pass filter, it suffices to find all low pass filters of length 4. Without loss of generality we may assume that the sequence is supported on the interval $[0, 3]$. Hence we are looking for $h(0), h(1), h(2), h(3)$ which satisfy conditions of Theorems 4.3.2 and 4.3.3. This implies the following constraints:

$$\begin{aligned} |h(0)|^2 + |h(1)|^2 + |h(2)|^2 + |h(3)|^2 &= 1, \\ h(0)h(2) + h(1)h(3) &= 0, \\ h(0) + h(2) &= \frac{1}{\sqrt{2}}, \\ h(1) + h(3) &= \frac{1}{\sqrt{2}}. \end{aligned}$$

It is not difficult to see that the solution of such a quadratic system of equations is

the following collection of filter values:

$$\begin{aligned} h(0) &= \frac{1-c}{\sqrt{2}(c^2+1)}, \\ h(1) &= \frac{1+c}{\sqrt{2}(c^2+1)}, \\ h(2) &= \frac{c(c+1)}{\sqrt{2}(c^2+1)}, \\ h(3) &= \frac{c(c-1)}{\sqrt{2}(c^2+1)}, \end{aligned}$$

for any parameter $c \in \mathbb{R}$. Indeed, let $h(2) = \frac{1}{\sqrt{2}} - h(0)$ and $h(3) = \frac{1}{\sqrt{2}} - h(1)$. Then,

$$\begin{aligned} |h(0)|^2 + |h(1)|^2 + |\frac{1}{\sqrt{2}} - h(0)|^2 + |\frac{1}{\sqrt{2}} - h(1)|^2 = 1, \\ h(0)(\frac{1}{\sqrt{2}} - h(0)) + h(1)(\frac{1}{\sqrt{2}} - h(1)) = 0. \end{aligned}$$

We see now that both equations are equal to each other. Hence, we only need to solve

$$h(0)(\frac{1}{\sqrt{2}} - h(0)) + h(1)(\frac{1}{\sqrt{2}} - h(1)) = 0,$$

which we treat as a quadratic equation with one unknown $h(0)$ and one parameter $h(1)$.

It is worth noting that for specific choices of the parameter c we can obtain very special filter sequences:

$$\begin{aligned} c = -1 \quad &\Longrightarrow \quad h = (\frac{1}{\sqrt{2}}, 0, 0\frac{1}{\sqrt{2}}) \\ c = 0 \quad &\Longrightarrow \quad h = (\frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0, 0) \\ c = 1 \quad &\Longrightarrow \quad h = (0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}, 0) \\ c = \pm\infty \quad &\Longrightarrow \quad h = (0, 0, \frac{1}{\sqrt{2}}, \frac{1}{\sqrt{2}}). \end{aligned}$$

Definition 4.4.1 (Daubechies 4 Filter)**.** A specific and important choice for a low pass filter is obtained by choosing the value for the parameter $c = 2 - \sqrt{3}$. This

yields:

$$h(0) = \frac{1-\sqrt{3}}{4\sqrt{2}},$$
$$h(1) = \frac{3+\sqrt{3}}{4\sqrt{2}},$$
$$h(2) = \frac{3-\sqrt{3}}{4\sqrt{2}},$$
$$h(3) = \frac{1-\sqrt{3}}{4\sqrt{2}},$$

Now, in order to define the discrete wavelet transform, we recall the periodic filtering. E.g., for a filter of length 4 the associated transform will act on a space of dimension 2 as

$$F = \begin{pmatrix} f(0) & f(1) & f(2) & f(3) \\ f(2) & f(3) & f(0) & f(1) \end{pmatrix}.$$

Thus, given low pass filter h and high pass filter g we construct a 4×4 matrix

$$W = \begin{pmatrix} H \\ G \end{pmatrix} = \begin{pmatrix} h(0) & h(1) & h(2) & h(3) \\ h(2) & h(3) & h(0) & h(1) \\ g(0) & g(1) & g(2) & g(3) \\ g(2) & g(3) & g(0) & g(1) \end{pmatrix}.$$

This matrix W forms one step of the wavelet transform. If the dimension of data is higher, we repeat the filtering of the output of the H transform until no longer possible.

4.5 Problems

1. Let h be a low pass OCQF filter of length 4, with nonzero values $h(0) = a, h(1) = b, h(2) = c, h(3) = d$. Write explicitly the 8×8 matrix of the associated **full** Discrete Wavelet Transform.

2. Find the inverse of the 4×4 full Discrete Wavelet Transform generated by the Haar filter.

3. Find an example of a low-pass OCQF filter of length 6, which has exactly 2 non-zero coefficients. Prove that it satisfies the normalization and self-orthonormality conditions.

4. Find all real-valued filters of length 2, which satisfy *only* the Self-Orthonormality condition ($\sum_k f(k+2n)\overline{f(k+2m)} = \delta_{m,n}$).

5. Are there any real-valued low-pass OCQF filters of length 4 satisfying the antisymmetry condition $h(0) = -h(3)$ and $h(1) = -h(2)$? Provide an example or disprove.

6. Let f be a finite filter such that $f(0) = \frac{1+\sqrt{3}}{4\sqrt{2}}, f(1) = \frac{3+\sqrt{3}}{4\sqrt{2}}, f(2) = \frac{3-\sqrt{3}}{4\sqrt{2}}, f(3) = \frac{1-\sqrt{3}}{4\sqrt{2}}$, and all other values equal to 0. Compute $F(u)$ for a sequence u such that $u(0) = 1, u(1) = 2, u(2) = 1$, and all other values of u are equal to 0.

7. Let f be a finite filter such that $f(0) = \frac{1+\sqrt{3}}{4\sqrt{2}}, f(1) = \frac{3+\sqrt{3}}{4\sqrt{2}}, f(2) = \frac{3-\sqrt{3}}{4\sqrt{2}}, f(3) = \frac{1-\sqrt{3}}{4\sqrt{2}}$, and all other values equal to 0. Compute $F^*(v)$ for a sequence v such that $v(0) = 1, v(1) = 2, v(2) = 1$, and all other values of v are equal to 0.

8. Show explicitly that

$$F^*(v)(k) = \begin{cases} \sum_n f(2n)\overline{v(\frac{k}{2} - n)}, & if \quad k \quad even, \\ \sum_n f(2n+1)\overline{v(\frac{k-1}{2} - n)}, & if \quad k \quad odd. \end{cases}$$

9. Let f be a filter with support of length 7 and let u be a vector with support of length 6. Find the length of the support of $F(u)$.

10. Show explicitly that

$$F^*_{2q}(v)(k) = \begin{cases} \sum_{n=0}^{q-1} f_{2q}(2n)\overline{v(\frac{k}{2} - n)}, & if \quad k \in [0, 2q-2] \quad even, \\ \sum_{n=0}^{q-1} f_{2q}(2n+1)\overline{v(\frac{k-1}{2} - n)}, & if \quad k \in [1, 2q-1] \quad odd. \end{cases}$$

Chapter 5

From Discrete to Continuous Representations

5.1 Lagrange Interpolation

An important problem in the analysis of discrete and finite signals is how to naturally connect to higher dimensional and even continuous functions. A readily available solution comes to us from the field of **algebra** in the form of **polynomial interpolation**. The basic idea behind this concept is represented by the following simple algebraic fact.

Theorem 5.1.1. *Let P be a polynomial of degree not exceeding $d > 0$. If P has d distinct roots (i.e., there exist d numbers $x_1, \ldots, x_d$, such that $x_i \neq x_j$ for $i \neq j$, and such that $P(x_i) = 0$ for $i = 1, \ldots, d$), then all coefficients of P must be equal to 0 and hence $P(x) = 0$ for all x.*

Let

$$P(x) = \sum_{k=0}^{d} a_k x^k.$$

Based on our previous observation, the coefficients a_k are uniquely determined by the values of P at $d+1$ distinct locations. The process of determining these coefficients from a given set of values provides us with a polynomial formula to fill in the values of all points between the finite collection that we started with. This is called the **polynomial interpolation**.

The following algorithm for finding the coefficients of the polynomial based on knowing its values is called the **Lagrange interpolation**.

For $k = 0, \ldots, d$, we let:

$$y_k = P(x_k), \quad \text{for the given fixed numbers } x_0, \ldots, x_d,$$

and

$$\lambda_k(x) = \prod_{j \neq k} \frac{x - x_j}{x_k - x_j}.$$

It is not difficult to see that every λ_k is a polynomial of degree d, and so is their linear combination:

$$\Lambda_d(x) = \sum_{k=0}^{d} y_k \lambda_k(x). \tag{5.1}$$

Theorem 5.1.2 (Lagrange Interpolation)**.** *If P is a polynomial of degree not exceeding d and if Λ_d is as defined in (5.1), then for every $x \in \mathbb{R}$,*

$$P(x) = \Lambda_d(x).$$

The algorithmic implementation of the Lagrange Interpolation requires $O(d^2)$ arithmetic operations in order to evaluate the polynomial at any given input x. This is because we have to compute $d+1$ individual Lagrange polynomials λ_k, and each one of them is a product of d monomials, so the computation of the value of each polynomial requires no more than $2d$ subtractions, d divisions, and $2d-1$ multiplications.

5.2 Chebyshev Interpolation

Because of the last remark in the previous section, about the computational cost of Lagrange Interpolation being on level of $O(d^2)$, we are interested in finding a more efficient way to implement Lagrange interpolation.

Definition 5.2.1 (Chebyshev Polynomials)**.** We define the following family of functions on $[-1, 1]$:

$$T_n(x) = \cos(n \arccos(x)), \quad \text{for} \quad n = 1, \ldots.$$

It is not difficult to see that:

$$T_0(x) = \cos(0) = 1,$$

and

$$T_1(x) = \cos(\arccos(x)) = x.$$

It begs the question of whether it may be true that every function defined in Definition 5.2.1 is a polynomial?

Proposition 5.2.1. *For every integer $n \geq 2$, we have:*

$$T_n(x) = 2xT_{n-1}(x) - T_{n-2}(x).$$

Proof. Given $x \in [-1, 1]$, we let $\xi = \arccos(x)$. Thus, $\cos(\xi) = x$, and we can write:

$$\begin{aligned} 2xT_{n-1}(x) - T_{n-2}(x) &= 2\cos(\xi)\cos((n-1)\xi) - \cos((n-2)\xi) \\ &= 2\cos(\xi)\cos((n-1)\xi) - \cos((n-1)\xi - \xi) \\ &= 2\cos(\xi)\cos((n-1)\xi) - \cos((n-1)\xi)\cos(\xi) - \sin((n-1)\xi)\sin(\xi) \\ &= \cos((n-1)\xi)\cos(\xi) - \sin((n-1)\xi)\sin(\xi) \\ &= \cos((n-1)\xi + \xi) = \cos(n\xi) = \cos(n\arccos(x)) = T_n(x). \end{aligned}$$

□

From Proposition 5.2.1 we can conclude that every T_n is in fact a polynomial.

Proposition 5.2.2. *For every integer $n \geq 0$, we have that T_n is a polynomial of degree n, with the leading coefficient equal to 2^{n-1}.*

Corollary 5.2.1. *Every polynomial $P(x)$, $x \in [-1, 1]$, can be written as a linear combination of Chebyshev polynomials T_n.*

Proof. Indeed it suffices to show that every monomial x^n can be written using Chebyshev polynomials of degree no more than n. It is true for every polynomial of degree 1, as $P(x) = ax + b = aT_1(x) + bT_0(x)$. For the inductive step, we assume that every polynomial of degree less or equal to $n - 1$ can be expressed as a linear combination of Chebyshev polynomials, and we use Proposition 5.2.2, to note that

$$x^n = T_n(x) - Q(x),$$

where Q is a polynomial of degree at most $n - 1$, and so the induction follows. □

Corollary 5.2.1 implies that, in particular, the Lagrange interpolation polynomial for any given data points $\{(x_k, y_k), k = 1, \ldots, N\}$, can also be expressed in terms of Chebyshev polynomials:

$$\Lambda_d(x) = \sum_{k=0}^{d} c(k) T_k(x), \quad \text{for some coefficients} \quad c(k) \in \mathbb{R}.$$

We shall take advantage of this observation, by constructing a specific notion of an inner product for polynomials, which, in turn, will allow us to simplify the computation of expansions in terms of Chebyshev polynomials. We begin by making the following observation.

Proposition 5.2.3. *For every integer $n \geq 0$, we have that the roots of the nth Chebyshev polynomial T_n take the form of:*

$$x_k = \cos\left(\frac{\pi(k+\frac{1}{2})}{n}\right), \quad \textit{for} \quad k = 0, \ldots, n-1.$$

Definition 5.2.2 (Inner product for polynomials)**.** Given any two polynomials P, Q of degree no more than $d > 0$, we let

$$\langle P, Q \rangle = \sum_{k=0}^{d-1} P(x_k) Q(x_k),$$

where $\{x_k\}$ denotes the set of roots of T_d.

For $0 \leq m, n \leq d$, we now compute

$$\begin{aligned} \langle T_m, T_n \rangle &= \sum_{k=0}^{d-1} \cos\left(\frac{\pi(k+\frac{1}{2})m}{d}\right) \cos\left(\frac{\pi(k+\frac{1}{2})n}{d}\right) \\ &= \begin{cases} 0 & \text{if } m \neq n \\ d & \text{if } m = n = 0 \\ \frac{d}{2} & \text{if } m = n \neq 0, \end{cases} \end{aligned}$$

because we realize that the expression on the right hand side is the usual inner product of rows of the matrix of the Discrete Cosine Transform of type II, (DCT-II, which we also denoted by C_d^{II}).

This simple observation implies that Chebyshev polynomials of degree at most d, form an ONB with our new inner product for the space of polynomials of degree no more than d. We shall use this fact in the proof of our main result for this section.

Theorem 5.2.1 (Chebyshev Interpolation)**.** *Let $x_k = \cos\left(\frac{\pi(k+\frac{1}{2})}{d+1}\right)$, for $k = 0, \ldots, d$, and let $y_k, k = 0, \ldots, d$ be an arbitrary collection of given real values. Then,*

$$\Lambda_d(x) = \sum_{k=0}^{d} c(k) T_k(x),$$

where,

$$c(0) = \frac{1}{d+1} \sum_{k=0}^{d} y_k,$$

and

$$c(m) = \frac{2}{d+1} \sum_{k=0}^{d} \cos\left(\frac{\pi(k+\frac{1}{2})m}{d+1}\right) y_k, \quad \textit{for} \quad m = 1, \ldots, d.$$

Proof. On. the one hand, we have that

$$\langle T_m, \Lambda_d \rangle = c(m) \langle T_m, T_m \rangle.$$

On the other hand,

$$\begin{aligned} \langle T_m, \Lambda_d \rangle &= \sum_{k=0}^{d} T_m(x_k) \Lambda_d(x_k) \\ &= \sum_{k=0}^{d} \cos\left(\frac{\pi(k+\frac{1}{2})m}{d+1}\right) \Lambda_d(x_k) \\ &= \sum_{k=0}^{d} \cos\left(\frac{\pi(k+\frac{1}{2})m}{d+1}\right) y_k. \end{aligned}$$

Combining the two observations together completes the proof of our theorem. □

We now observe that the coefficients computed in Theorem 5.2.1 are in fact the coefficients computed by means of the Discrete Cosine Transform of type II, applied to the given data vector $y = \{y_k : k = 0, \ldots, d\}$:

$$c = C^{II}_{d+1}(y).$$

This simple fact produces significant savings, as the computational cost of DCT-II is $O(d \log(d))$ for when the dimension is a power of 2. Once we have precomputed the coefficients c of the Chebyshev polynomials, then the actual evaluation of all Chebyshev polynomials at each given point x requires only $O(d)$ arithmetic operations, due to the recurrent structure of Chebyshev polynomials, see Proposition 5.2.1. This is significantly less than the cost of computing the Lagrange Interpolation, which was $O(d^2)$. This efficient process is known as the **Chebyshev interpolation**.

These savings come at the cost of loosing freedom of choosing the distribution of our data y along the real line. Instead we are forced to use the roots of Chebyshev polynomials as the arguments.

5.3 Problems

1. Find the 2nd degree Lagrange polynomial for the following data points:

$$(-1, 0), (0, 0), (1, 0).$$

2. Find a polynomial of exactly degree 3 which interpolates the following data points:

$$(-1, 0), (0, 0), (1, 0)..$$

3. Find the 3rd degree Lagrange polynomial for the following data points:

$$(0, -1), (1, 0), (2, 5), (3, 20).$$

4. Find the 6th degree Lagrange polynomial for the following data points:

$$(0, 0), (1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6).$$

5. Find the expansion of the polynomial $4x^3 - 2x^2 - x$ in terms of Chebyshev polynomials.

6. Find the explicit form of the 4th Chebyshev polynomial T_4.

7. Let $P(x) = x^3 - x$. Find all the zeros of the 3rd degree Lagrange polynomial for data points $(1, P(1)), (2, P(2)), (3, P(3)), (4, P(4))$.

[B87] J. J. Benedetto, "Gabor Representations and Wavelets", Technical Report 87-36, University of Maryland College Park, MD, 1987.

[B92] J. J. Benedetto, "Irregular sampling and frames", in Wavelets: A Tutorial in Theory and Applications, C. K. Chui, Ed. pp. 445–507, CRC Press, Boca Raton, FL, 1992.

[B94] J. J. Benedetto, "Frame decompositions, sampling and uncertainty principle inequalities", in: J. Benedetto, M. Frazier (Eds.), Wavelets: Mathematics and Applications, CRC Press, Boca Raton, FL, 1994, pp. 247–304.

[C00] P. Casazza, "The Art of Frame Theory", 2000.

[C03] O. Christensen, "An Introduction to Frames and Riesz Bases", Birkhauser, 2003.

[DGM86] I. Daubechies, A. Grossman, and Y. Meyer, "Painless Nonorthogonal Expansions", J. Math. Physics, vol. 27, 1271–1283, 1986.

[DS52] R. J. Duffin and A. C. Schaeffer, "A class of nonharmonic Fourier series", Trans. Amer. Math. Soc., vol. 72, 341–366, 1952.

[Z] A. Zygmunt, "Trigonometric Series", Cambridge University Press, 2002