

Joseph D. Lakey

# Time–Frequency Methods for Signal Processing

August 21, 2007



---

# Contents

Preface .....	xiii
<b>1 Hilbert space fundamentals</b> .....	1
1.1 Definitions and Examples .....	1
1.1.1 Scaling and limits in $\mathbb{R}^N$ and $\mathbb{C}^N$ .....	3
1.1.2 Unitary equivalence and Gram-Schmidt .....	4
1.2 Angles, Orthogonal matrices and projections .....	5
1.2.1 Orthogonal and unitary matrices .....	6
1.2.2 Rotations .....	6
1.2.3 Variance maximizing rotation .....	8
1.2.4 Projection matrices .....	8
1.3 Infinite scales and infinite dimensions: limits .....	9
1.4 The space $\ell^2(\mathbb{N})$ .....	9
1.4.1 Completeness of $\ell^2(\mathbb{Z}^N)$ .....	9
1.4.2 Orthogonal complements and orthogonal projections in $\ell^2(\mathbb{N})$ .....	10
1.4.3 Complete orthonormal bases .....	11
1.4.4 $\ell^2(\mathbb{Z})$ .....	12
1.4.5 Operators on $\ell^2(\mathbb{Z})$ .....	12
1.4.6 Isomorphisms of Separable Hilbert spaces .....	13
1.4.7 More on Isomorphism .....	14
1.4.8 A nonseparable Hilbert space .....	14
1.5 Limits as $N \rightarrow \infty$ : From Hilbert space to the real world and back .....	14
1.5.1 Linear interpolation from discrete data .....	15
1.5.2 Interpolation errors. ....	15
1.5.3 Filling in values .....	16
1.5.4 Quantization .....	16
1.5.5 Transformation coding .....	17
1.6 Appendix: $\ell^2(\mathbb{Z}_N)$ and finite Fourier series .....	17
<b>2 Other metrics on signals</b> .....	19
2.1 Variation .....	19
2.1.1 Gross change and net change .....	19
2.1.2 Variation and images .....	19
2.1.3 Metric spaces and signal spaces .....	20
2.2 Entropy .....	20
2.2.1 Entropy minimizing transformation .....	21
2.2.2 Direct sum decompositions .....	21
2.2.3 Multiscale transformations .....	21
2.2.4 Additive cost function and tree structures .....	22
2.3 Best $N$ -term approximation and rate of approximation .....	22
2.4 Information .....	23
2.4.1 Rate versus distortion: thresholding and compression .....	23

<b>3</b>	<b>Fourier analysis</b> .....	25
3.1	Discrete Fourier Transform .....	25
3.2	Fast Fourier transform .....	26
3.3	Fourier series .....	27
3.3.1	Centering the approximation .....	27
3.4	Convergence of Fourier series and $L^2(\mathbb{T})$ .....	28
3.5	Convergence of Fourier series .....	29
3.5.1	Trigonometric polynomials .....	29
3.5.2	Convolution and Fourier partial sums .....	30
3.5.3	Convergence of partial sums for nice $f$ .....	30
3.5.4	Density of trigonometric polynomials .....	32
3.5.5	Continuous functions and $L^2(\mathbb{R})$ .....	32
3.5.6	Convergence at jumps and Gibbs phenomenon .....	33
3.6	Separation of variables and the heat equation .....	33
3.7	Fourier transforms .....	33
3.7.1	Properties of Fourier transforms .....	34
3.7.2	Gaussian approximation .....	34
3.7.3	Fourier uniqueness .....	34
3.8	Multipliers .....	34
3.8.1	Deconvolution and applications .....	34
3.9	Plancherel and Parseval .....	34
3.10	Centered DFT and its eigenvectors .....	35
3.11	Bandlimited functions .....	35
3.12	Bandlimited functions on $\mathbb{R}$ .....	35
3.12.1	Why are signals bandlimited .....	35
3.12.2	The Paley-Wiener space .....	35
3.12.3	The classical sampling theorem of Shannon, Whittaker, ... ..	35
3.12.4	Bandlimited wavelets .....	36
3.13	Better Bandlimited wavelets .....	38
3.13.1	Bell functions .....	39
3.14	Time and bandlimited functions .....	39
<b>4</b>	<b>Time-frequency analysis</b> .....	41
4.1	Time-frequency analysis .....	41
4.2	Uncertainty principles .....	41
4.3	Plancherel and Parseval .....	41
4.4	Time-frequency bases and frames .....	41
4.4.1	Gabor bases .....	41
4.4.2	Wilson bases .....	41
4.4.3	Malvar bases .....	41
4.5	Empirical mode decomposition .....	41
4.5.1	Hilbert spectrum .....	41
4.5.2	Implicit modes .....	41
<b>5</b>	<b>Wavelets</b> .....	43
5.1	Wavelets .....	43
5.1.1	Time-scale representation, continuous wavelet transform, inversion .....	43
5.1.2	Discrete wavelet transform and Orthonormal bases .....	43
5.2	Haar Wavelets .....	44
5.2.1	Construction of wavelets .....	44
5.3	Curvelets and other time-frequency tools .....	44

<b>6</b>	<b>Image Segmentation</b> .....	45
6.1	.....	45
6.1.1	Mumford-Shah energy functional .....	46
6.1.2	TV method in wavelet coordinates .....	49
6.1.3	More mathematical properties of ROF and other image decompositions .....	50
6.1.4	TV: general discussion .....	50
6.2	Notes .....	53
<b>7</b>	<b>Extra Chapter</b> .....	55
<b>A</b>	<b>Appendix</b> .....	57
A.1	Notation .....	57
A.2	Miscellany from real and harmonic analysis .....	58
A.3	Miscellany from functional analysis .....	61
A.4	Toolboxes .....	62
A.5	Images Databases .....	62
A.6	Downloadable Speech Databases .....	62
<b>B</b>	<b>Matlab</b> .....	63
B.1	Matlab basics .....	63
B.1.1	Self help .....	63
B.1.2	Vectors and Matrices .....	64
B.1.3	Arithmetic .....	64
B.1.4	Built-in functions .....	65
B.1.5	Comments on graphics .....	65
B.1.6	flow control .....	67
B.1.7	Matlab scripts .....	69
B.1.8	Animation .....	71





## Hilbert space fundamentals

### 1.1 Definitions and Examples

The concept of a Hilbert space was introduced by the mathematician David Hilbert as part of his work on integral equations in the early part of the twentieth century. It has become one of the most fundamental concepts in applied mathematics. Formally, a Hilbert space is a mathematical structure that captures the notions of *length* and *angle* in an abstract way essential to analyzing sequences and functions in a unified fashion. A Hilbert space  $(\mathcal{H}, \langle \cdot, \cdot \rangle)$  consists of a *complete* vector space  $\mathcal{H}$  over  $\mathbb{C}$  (or sometimes over  $\mathbb{R}$ ) together with a mapping  $\langle \cdot, \cdot \rangle : \mathcal{H} \times \mathcal{H} \rightarrow \mathbb{C}$  that assigns to each pair of elements of  $\mathcal{H}$  an element of  $\mathbb{C}$ . The mapping  $\langle \cdot, \cdot \rangle$  is called an *inner product*. It must satisfy four important properties, namely (i) it respects the vector space operations, which is to say, it is bilinear; (ii) it is Hermitian, (iii) it is continuous and (iv) it is nondegenerate.

**Linearity.** Whenever  $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathcal{H}$  and  $\alpha, \beta \in \mathbb{C}$ , one has

$$\langle \alpha \mathbf{x} + \beta \mathbf{y}, \mathbf{z} \rangle = \alpha \langle \mathbf{x}, \mathbf{z} \rangle + \beta \langle \mathbf{y}, \mathbf{z} \rangle.$$

**Hermitian.** This property is named after the mathematician Charles Hermite. The property is expressed as

$$\langle \mathbf{x}, \mathbf{y} \rangle = \overline{\langle \mathbf{y}, \mathbf{x} \rangle}$$

where  $\bar{\alpha}$  is the complex conjugate of the complex number  $\alpha$ . If a mapping taking two inputs is linear in the first input and Hermitian then it is automatically linear in the second input, since then

$$\langle \mathbf{z}, \alpha \mathbf{x} + \beta \mathbf{y} \rangle = \overline{\langle \alpha \mathbf{x} + \beta \mathbf{y}, \mathbf{z} \rangle} = \bar{\alpha} \overline{\langle \mathbf{x}, \mathbf{z} \rangle} + \bar{\beta} \overline{\langle \mathbf{y}, \mathbf{z} \rangle} = \bar{\alpha} \langle \mathbf{z}, \mathbf{x} \rangle + \bar{\beta} \langle \mathbf{z}, \mathbf{y} \rangle$$

which is what it means to be (conjugate) linear in the second argument. *Bilinear* thus means linear in the first argument and conjugate linear in the second.

**Nondegeneracy.** Next we consider what it means to be *nondegenerate* or *definite*. This means that  $\langle \mathbf{x}, \mathbf{x} \rangle \neq 0$  whenever  $\mathbf{x} \neq 0$ . Since  $\langle \mathbf{x}, \mathbf{x} \rangle = \overline{\langle \mathbf{x}, \mathbf{x} \rangle}$  it follows that  $\langle \mathbf{x}, \mathbf{x} \rangle$  is a real number. As such, it is strictly positive or strictly negative when  $x \neq 0$ . In the case of an inner product, we require that  $\langle \mathbf{x}, \mathbf{x} \rangle > 0$  when  $\mathbf{x} \neq 0$ . Then  $\langle \cdot, \cdot \rangle$  is said to be *positive definite*. We can then define  $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$ , called the *norm* of  $f$ .

**Completeness.** This refers to behavior under taking limits. For example, the rational numbers are non complete because, for example,  $\sqrt{2}$  can be expressed as a limit of rational numbers but it is not itself rational. That a Hilbert space is complete means that if a sequence of its elements all bunch up, meaning that, for arbitrarily small  $\epsilon > 0$  there is an  $N$  such that  $\|\mathbf{x}_k - \mathbf{x}_\ell\| < \epsilon$  whenever  $k, \ell > N$ , then this sequence must have a limit point in the space. Completeness is a sometimes subtle, but always important property.

*Example 1.1.1.* The space  $\mathbb{C}^N$  consists of all  $N$ -tuples  $\mathbf{z} = (z_1, \dots, z_N)$  of complex numbers  $z_i$ . We denote by  $\bar{\mathbf{z}} = (\bar{z}_1, \dots, \bar{z}_N)$  the vector whose  $i$ -th coordinate is the complex conjugate of the  $i$ -th coordinate of  $\mathbf{z}$ . The standard inner product on this space is the dot product, namely,

$$\langle \mathbf{z}, \mathbf{w} \rangle = \sum_{i=1}^N z_i \bar{w}_i.$$

In this case,  $\|\mathbf{z}\| = \sqrt{\langle \mathbf{z}, \mathbf{z} \rangle} = \sqrt{\sum_{i=1}^N |z_i|^2}$  where  $|z_i|^2 = z_i \bar{z}_i$ .

The dot product defines an inner product on  $\mathbb{C}^N$  but can there be others? To answer this question, first we rewrite the dot product in the peculiar form

$$(z_1 \ z_2 \ \cdots \ z_N) I_N \begin{pmatrix} \bar{w}_1 \\ \bar{w}_2 \\ \vdots \\ \bar{w}_N \end{pmatrix}$$

where  $I_N$  denotes the  $N \times N$  identity matrix. Suppose that we replace  $I_N$  by any  $N \times N$  matrix and define the form

$$\mathcal{M}(\mathbf{z}, \mathbf{w}) = (z_1 \ z_2 \ \cdots \ z_N) M \begin{pmatrix} \bar{w}_1 \\ \bar{w}_2 \\ \vdots \\ \bar{w}_N \end{pmatrix}$$

**Exercise 1.1.2.** Prove that  $\mathcal{M}(\mathbf{z}, \mathbf{w})$  is bilinear.

**Exercise 1.1.3.** Prove that  $\mathcal{M}(\mathbf{z}, \mathbf{w})$  is Hermitian if and only if  $M_{ij} = \overline{M_{ji}}$ . In other words,  $\mathcal{M}$  is Hermitian if and only if the matrix  $M$  is equal to its conjugate transpose.

The problem of determining when  $\mathcal{M}$  is positive definite is a little more subtle. We want a condition on the matrix  $M$  guaranteeing that whenever  $\mathbf{z} \in \mathbb{C}^N \setminus 0$ , one has  $\mathcal{M}(\mathbf{z}, \mathbf{z}) > 0$ . An obvious thing to try is to express the action of  $M$  in terms of the *standard basis vectors*. Let  $\mathbf{e}_j$  denote the  $j$ -th standard basis vector in  $\mathbb{C}^N$ , that is,  $\mathbf{e}_j$  is the vector whose  $i$ -th coordinate is zero if  $i \neq j$  and whose  $j$ -th coordinate is one. Since  $\mathcal{M}$  is bilinear, assuming it is also Hermitian one has, for any vector  $\mathbf{z} = (z_1, \dots, z_N)$ ,

$$\mathcal{M}(\mathbf{z}, \mathbf{z}) = \sum_{i,j=1}^N z_i \bar{z}_j \mathcal{M}(\mathbf{e}_i, \mathbf{e}_j).$$

This approach leads us nowhere quickly when trying to determine a condition under which  $\mathcal{M}$  is positive definite. Actually, we need a rather powerful result from linear algebra already at this stage, namely the spectral theorem for Hermitian matrices. To make sense of this result, first we have to define what it means to be an orthonormal basis for  $\mathbb{C}^N$ .

**Definition 1.1.4.** A set  $\mathbf{v}_1, \dots, \mathbf{v}_N$  of vectors in  $\mathbb{C}^N$  forms an orthonormal basis for  $\mathbb{C}^N$  if for each  $1 \leq i, j \leq N$ ,  $\langle \mathbf{v}_i, \mathbf{v}_j \rangle = \delta_{ij}$  where  $\delta_{ij} = 1$  if  $i = j$  and  $\delta_{ij} = 0$  if  $i \neq j$ .

The symbol  $\delta_{ij}$  is called the Kronecker  $\delta$  after the mathematician Leopold Kronecker

**Theorem 1.1.5.** Suppose that  $M$  is a Hermitian  $N \times N$  matrix (that is,  $M = \bar{M}^T$ ). Then all of the eigenvalues of  $M$  are nonnegative and there is an orthonormal basis of  $\mathbb{C}^N$  consisting of eigenvectors of  $M$ . In other words, there are vectors  $\mathbf{v}_1, \dots, \mathbf{v}_N$  in  $\mathbb{C}^N$  such that  $\mathbf{v}_i \cdot \mathbf{v}_j = \delta_{ij}$  and  $M\mathbf{v}_i = \lambda_i \mathbf{v}_i$  where  $\lambda_i$  is a positive real number.

In fact,  $M$  has the singular value decomposition  $M = V^T \Sigma V$  in which the columns of  $V$  are the (unit) eigenvectors of  $M$ . Also, in this case  $V^T = V^{-1}$  when  $V$  is invertible, that is, when it has no zero eigenvalues (see the next exercise). Because of the theorem, it follows that if  $M$  is a nonsingular Hermitian matrix then the form  $\mathcal{M}$  is positive definite and therefore defines an inner product on  $\mathbb{C}^N$ .

**Exercise 1.1.6.** Show that if  $M$  is a Hermitian matrix with eigenvalues  $\lambda \neq \mu$  and if  $M\mathbf{v}_\lambda = \lambda\mathbf{v}_\lambda$  and  $M\mathbf{v}_\mu = \mu\mathbf{v}_\mu$  where  $\lambda \neq \mu$  then  $\mathbf{v}_\lambda \cdot \mathbf{v}_\mu = 0$ .

**Exercise 1.1.7.** Show that any inner product defined on  $\mathbb{C}^N$  has to have the form  $\langle \mathbf{z}, \mathbf{w} \rangle = \mathbf{z}^T M \bar{\mathbf{w}}$  for some Hermitian, nonsingular matrix  $M$ .

**Exercise 1.1.8.** (Polarization) Show that if  $\langle \cdot, \cdot \rangle$  is an inner product and  $\|\mathbf{x}\| = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle}$  then

$$\Re \langle \mathbf{x}, \mathbf{y} \rangle = \frac{1}{4} \left[ \|\mathbf{x} + \mathbf{y}\|^2 - \|\mathbf{x} - \mathbf{y}\|^2 \right].$$

Find a corresponding expression for  $\Im \langle \mathbf{x}, \mathbf{y} \rangle$ . Here  $\Re \alpha$  and  $\Im \alpha$  denote the real and imaginary parts of the complex number  $\alpha = \Re \alpha + i \Im \alpha$ .

### 1.1.1 Scaling and limits in $\mathbb{R}^N$ and $\mathbb{C}^N$

We will continue to discuss geometric properties of Hilbert spaces momentarily, but we cannot do so rigorously without some intervening discussion of *completeness*.

#### Completeness of $\mathbb{R}$ and $\mathbb{C}$

The idea of a *limit* is the cornerstone of analysis, and the Archimedean axiom, that *every bounded set of real numbers has a least upper bound* is a precise, if not intuitive, expression of the *completeness* of the real numbers – that a limit of a sequence of real numbers is also a real number.

For us a more concrete expression of this fact will be useful and it is based on the notion of subdivision, namely, that every interval of the real line can be evenly divided into two subintervals of equal length, the so-called left and right subintervals. Thus,  $[0, 1] = [0, 1/2] \cup [1/2, 1]$ . This simple fact leads to the *binary expansion* of any  $x \in [0, 1]$ . First, define a number  $\varepsilon_1$  as follows. Let  $\varepsilon_1 = 0$  if  $0 \leq x < 1/2$  and  $\varepsilon_1 = 1$  if  $1/2 \leq x < 1$ . Then  $x - \varepsilon_1/2 \in [0, 1/2)$  or  $x_1 = 2x - \varepsilon_1 \in [0, 1)$ . Next define  $\varepsilon_2 = 0$  if  $0 \leq x_1 < 1/2$  and  $\varepsilon_2 = 1$  if  $1/2 \leq x_1 < 1$ . Then  $x_1 - \varepsilon_2 \in [0, 1/2)$  or  $x - \varepsilon_1/2 - \varepsilon_2/4 \in [0, 1/4)$ . Define  $x_2 = 4(x - \varepsilon_1/2 - \varepsilon_2/4)$  so  $0 \leq x_2 < 1$ . Define  $\varepsilon_3$  and so on in such a way that  $0 \leq x - \sum_{k=1}^N \frac{\varepsilon_k}{2^k} < 1/2^N$ . In particular, the sequence of numbers  $s_N = \sum_{k=1}^N \frac{\varepsilon_k}{2^k}$  is increasing and bounded above by  $x$  hence converges to its least upper bound. Since  $x - s_N \leq 2^{-N}$  tends to zero, this upper bound has to be  $x$ . This justifies expressing  $x$  as the infinite sum  $x = \sum_{k=1}^{\infty} \frac{\varepsilon_k}{2^k}$ . This sum is called the binary expansion of  $x$ . If the sum terminates, that is,  $\varepsilon_k = 0$  once  $k$  is large enough, then  $x$  is called a *dyadic rational*:  $x = \sum_{k=1}^N \frac{\varepsilon_k}{2^k} = \frac{1}{2^N} \sum_{k=1}^N \varepsilon_k 2^{N-k} = \frac{1}{2^N} \sum_{\nu=0}^{N-1} \varepsilon_{N-\nu} 2^\nu$  expressing  $x$  as the ratio of an integer and a power of 2. For shorthand, one often writes the binary expansion in the form  $x = 0.\varepsilon_1 \varepsilon_2 \varepsilon_3 \dots$ .

If  $x$  is a whole number between  $2^{N-1}$  and  $2^N$  then  $y = x/2^N$  is a dyadic rational number so we can write  $y = 0.\delta_1 \delta_2 \dots \delta_N$  and then  $x = 2^N y = \delta_1 2^{N-1} + \delta_2 2^{N-2} + \dots + \delta_N$  meaning that  $x = \sum_{k=0}^{N-1} \delta_{N-k} 2^k$ . Altogether then, any positive number  $x$  has a dyadic expansion  $\delta_1 2^{N-1} + \delta_2 2^{N-2} + \dots + \delta_N$  plus its *fractional part*  $x - [x] = \sum_{k=1}^{\infty} \frac{\varepsilon_k}{2^k}$ . Completeness of the real numbers is then expressed in terms of the fact that every real number can be expressed in terms of its dyadic infinite series expansion and the partial sums of this series tell us, for each  $k$ , the unique interval of the form  $[\ell/2^k, (\ell+1)/2^k)$ ,  $\ell \in \mathbb{Z}$ , to which  $x$  belongs.

Had we started out by splitting  $[0, 1]$  into ten equal subintervals then we would have obtained the *decimal expansion* of  $x$ . Other expansions are obtained upon subdividing into some other number of equal parts, for example, the *ternary expansion* of  $x$  has the form  $x = \sum \frac{\gamma_k}{3^k}$  where  $\gamma_k \in \{0, 1, 2\}$ .

The completeness of the complex numbers is equivalent to the completeness of  $\mathbb{R}^2$ , the collection of all pairs of real numbers  $(x, y)$  where  $x, y \in \mathbb{R}$ . In this case, when  $z = x + iy$  one expresses each of  $x$  and  $y$  in terms of its dyadic decomposition. For positive  $x$  and  $y$ , the  $N$ -th partial sum of each then gives the vertex  $(x_N, y_N)$  of the unique square in the plane of the form  $[\ell_1/2^N, (\ell_1+1)/2^N) \times [\ell_2/2^N, (\ell_2+1)/2^N)$  in which  $(x, y)$  lies.

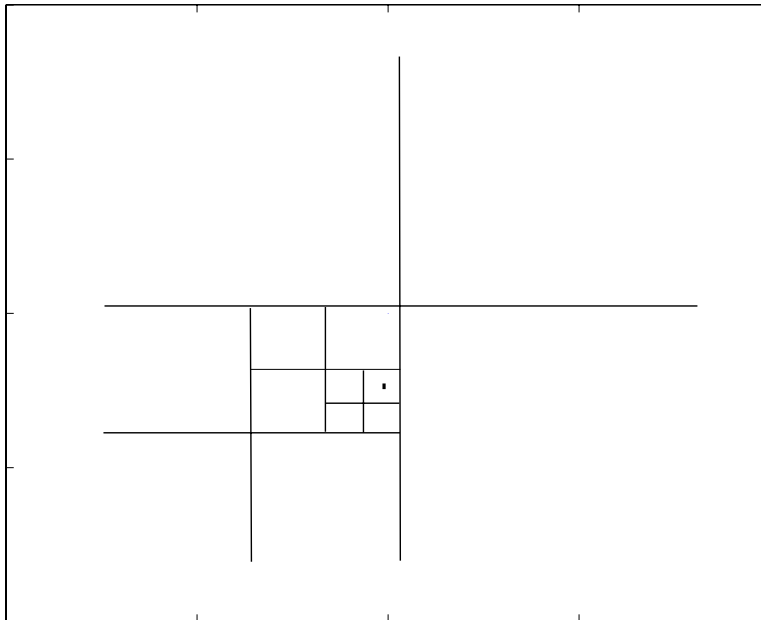
One can carry out similar analysis in higher dimensions. For example, in  $\mathbb{R}^3$  one subdivides a big cube into equal cubes of sidelength  $1/2^N$  and the dyadic decomposition of  $(x_1, x_2, x_3)$  leads to vertices  $(x_{1,N}, x_{2,N}, x_{3,N})$  of cubes of the form  $[\ell_1/2^N, (\ell_1+1)/2^N) \times [\ell_2/2^N, (\ell_2+1)/2^N) \times [\ell_3/2^N, (\ell_3+1)/2^N)$  containing  $(x, y, z)$  and so on in higher dimensions.

A sequence in  $\mathbb{R}^N$  is a collection  $\mathbf{x}^k = (x_1^k, x_2^k, \dots, x_N^k)$  of vectors indexed by  $k = 1, 2, \dots$ . If we fix the  $j$ -th coordinate we get a sequence  $\{x_j^k\}_k$  of real numbers called the  $j$ -th coordinate sequence. This sequence converges to  $x \in \mathbb{R}$  if, for each  $\varepsilon > 0$  there is an  $M$  such that whenever  $k > M$  one has  $|x_j^k - x| < \varepsilon$ . On the other hand, a sequence  $\mathbf{x}^k$  of vectors in  $\mathbb{R}^N$  converges to a vector  $\mathbf{x}$  if for each  $\varepsilon > 0$  there is an  $M$  such that whenever  $k > M$  one has  $\|\mathbf{x}^k - \mathbf{x}\| < \varepsilon$  where  $\|\cdot\|$  denotes the Euclidean distance. Since

$$\|\mathbf{x}\|_\infty \equiv \max_j |x_j| \leq \left( \sum_{j=1}^N |x_j|^2 \right)^{1/2} \equiv \|\mathbf{x}\|_2 \leq N \max_j |x_j| = N \|\mathbf{x}\|_\infty,$$

it follows readily that a sequence  $\mathbf{x}^k$  converges to  $\mathbf{x}$  if and only if  $\{x_j^k\}_k$  converges to  $x_j$  for each  $j = 1, \dots, N$ . Thus convergence in  $\mathbb{R}^N$  is equivalent to coordinatewise convergence. The same holds for  $\mathbb{C}^N$ .

So far we have not addressed the question of when an inner product is continuous. In fact, any inner product on  $\mathbb{C}^N$  is automatically continuous. This means that if we have a sequence  $\mathbf{z}_k$  of vectors in  $\mathbb{C}^N$  that converges to the vector  $\mathbf{z}$  and if  $\mathbf{w}$  is any fixed vector then  $\langle \mathbf{z}_k, \mathbf{w} \rangle \rightarrow \langle \mathbf{z}, \mathbf{w} \rangle$  as  $k \rightarrow \infty$ .



**Fig. 1.1.** Dyadic decomposition in the plane.

**Exercise 1.1.9.** Show that, in  $\mathbb{C}^N$ , any inner product is automatically continuous.

### 1.1.2 Unitary equivalence and Gram-Schmidt

A vector space is finite dimensional if there is an  $N$  such that any collection of  $N + 1$  vectors is linearly dependent. The dimension of a finite dimensional vector space is defined as the maximal number of linearly independent vectors.

**Theorem 1.1.10.** Any finite dimensional inner space over  $\mathbb{R}$  or  $\mathbb{C}$  has an orthonormal basis.

*Proof.* Suppose that  $(V, \langle \cdot, \cdot \rangle)$  is a vector space of dimension  $N$  over  $\mathbb{C}$  and let  $\mathbf{v}_1, \dots, \mathbf{v}_N$  be linearly independent. Define  $\mathbf{w}_1 = \mathbf{v}_1 / \|\mathbf{v}_1\|$  where, as before,  $\|\mathbf{v}\| = \sqrt{\langle \mathbf{v}, \mathbf{v} \rangle}$ . Linear independence implies that we can divide by  $\|\mathbf{v}_i\|$ . Next let  $\mathbf{u}_2 = \mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{w}_1 \rangle \mathbf{w}_1$ . Then  $\mathbf{u}_2 \neq \mathbf{0}$  since  $\mathbf{v}_1$  and  $\mathbf{v}_2$  are linearly independent. Then

$$\langle \mathbf{u}_2, \mathbf{w}_1 \rangle = \langle \mathbf{v}_2 - \langle \mathbf{v}_2, \mathbf{w}_1 \rangle \mathbf{w}_1, \mathbf{w}_1 \rangle = \langle \mathbf{v}_2, \mathbf{w}_1 \rangle - \langle \mathbf{v}_2, \mathbf{w}_1 \rangle \langle \mathbf{w}_1, \mathbf{w}_1 \rangle = 0.$$

Now set  $\mathbf{w}_2 = \mathbf{u}_2 / \|\mathbf{u}_2\|$ . Continuing in this manner, for each  $2 \leq j \leq N$  one sets  $\mathbf{u}_j = \mathbf{v}_j - \sum_{k=1}^{j-1} \langle \mathbf{v}_j, \mathbf{w}_k \rangle \mathbf{w}_k \neq \mathbf{0}$  (by linear independence) and  $\mathbf{w}_j = \mathbf{u}_j / \|\mathbf{u}_j\|$ . One can show, by induction, that  $\mathbf{w}_1, \dots, \mathbf{w}_N$  forms an orthonormal set. Since orthonormal sets are linearly independent, it follows that  $\mathbf{w}_1, \dots, \mathbf{w}_N$  forms an orthonormal basis for  $V$ . This proves the theorem.

The process for constructing the  $\mathbf{w}_i$  is called the *Gram-Schmidt process* after the number theorist Jorgen Pederson Gram and analyst Erhard Schmidt, though use of the technique dates back to Laplace and to Cauchy.

**Exercise 1.1.11.** The polynomials  $\{1, x, \dots, x^N\}$  are linearly independent over  $\mathbb{R}$ . The linear combinations of these polynomials form a vector space  $\mathcal{P}_N$  of dimension  $N + 1$ . The mapping  $\langle p, q \rangle = \int_0^1 p(x)q(x) dx$  defines an inner product on  $\mathcal{P}_N$ . Find an orthonormal basis for this space.

**Definition 1.1.12.** Let  $(V, \langle \cdot, \cdot \rangle_V)$  and  $(W, \langle \cdot, \cdot \rangle_W)$  be a pair of Hilbert spaces over  $\mathbb{C}$ . A linear mapping  $U : V \rightarrow W$  is said to be unitary if, for any pair  $\mathbf{v}_1, \mathbf{v}_2$  in  $V$  one has

$$\langle U\mathbf{v}_1, U\mathbf{v}_2 \rangle_W = \langle \mathbf{v}_1, \mathbf{v}_2 \rangle_V.$$

If  $V$  is  $N$  dimensional then  $W$  necessarily has dimension at least  $N$ . This is by the fact that the dimension of a vector space is the rank plus nullity of any linear mapping defined on the space. If  $\dim W < N$  then the null space of  $U$  would have to have dimension greater than zero so  $U\mathbf{v} = \mathbf{0}_W$  for some  $\mathbf{v} \neq \mathbf{0}_V$ , implying that  $\|U\mathbf{v}\|_W = 0$  but  $\|\mathbf{v}\|_V \neq 0$  contradicting unitarity of  $U$ . Consequently, any unitary mapping is one-to-one onto its range. Suppose now that  $V$  and  $W$  are any pair of  $N$ -dimensional Hilbert spaces. Let  $\mathbf{v}_1, \dots, \mathbf{v}_N$  and  $\mathbf{w}_1, \dots, \mathbf{w}_N$  be orthonormal bases for  $V$  and  $W$  respectively. Let  $U$  be the mapping that maps  $\mathbf{v}_i$  to  $\mathbf{w}_i$  for each  $i = 1, \dots, N$  and extend  $U$  linearly to  $V$ .

**Exercise 1.1.13.** Show that this mapping  $U$  is unitary.

Hilbert spaces  $V$  and  $W$  are said to be *unitarily equivalent* if there is a unitary mapping from  $V$  onto  $W$ .

**Corollary 1.1.14.** Any two  $N$  dimensional Hilbert spaces over  $\mathbb{C}^N$  are unitarily equivalent.

## 1.2 Angles, Orthogonal matrices and projections

Any finite dimensional vector space over  $\mathbb{R}$  (resp.  $\mathbb{C}$ ) is equivalent, as a vector space, to  $\mathbb{R}^N$  (resp.  $\mathbb{C}^N$ ) for appropriate  $N$  – the dimension of the space. A norm on a vector space  $X$  is a mapping  $\|\cdot\|$  that defines to each  $\mathbf{x} \in X$  a nonnegative real number  $\|\mathbf{x}\|$  in such a way that, if  $\lambda \in \mathbb{C}$  then  $\|\lambda\mathbf{x}\| = |\lambda|\|\mathbf{x}\|$  (homogeneity) and if  $\mathbf{x}, \mathbf{y} \in X$  then  $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$  (triangle inequality). Every inner product gives rise to a norm but not every norm gives rise to an inner product. For example, the  $\ell^1$  norm on  $\mathbb{C}^N$ , defined by  $\|\mathbf{z}\|_1 = \sum_{i=1}^N |z_i|$ , is a norm that is not *polarizable*. What is special about an inner product is that it allows one to define an *angle* between two vectors  $\mathbf{z}$  and  $\mathbf{w}$  via

$$\text{angle}(\mathbf{z}, \mathbf{w}) = \arccos\left(\frac{|\langle \mathbf{z}, \mathbf{w} \rangle|}{\|\mathbf{z}\|\|\mathbf{w}\|}\right).$$

**Exercise 1.2.1.** Show that, in an inner product space  $(X, \langle \cdot, \cdot \rangle)$ , if  $\mathbf{w} = \alpha\mathbf{z}$  for some  $\alpha \in X \setminus 0$ , then  $\text{angle}(\mathbf{z}, \mathbf{w}) = 0$ . Also show that if  $\mathbf{z}$  and  $\mathbf{w}$  are *orthogonal*, that is,  $\langle \mathbf{z}, \mathbf{w} \rangle = 0$ , then  $\text{angle}(\mathbf{z}, \mathbf{w}) = \pi/2$ .

**Exercise 1.2.2.** Let  $A \subset \mathbb{R}^N$ . Show that the set  $W_A$  of vectors that are orthogonal to every element of  $A$  forms a subspace of  $\mathbb{R}^N$  (called the orthogonal complement of  $A$ ).

These facts are consistent with the notion of angle in two or three variables. In fact, in the Euclidean space  $\mathbb{R}^N$ , any two vectors determine a two dimensional plane and their angle is simply their angle in that plane. There is one peculiarity with this notion of angle that should be pointed out, namely that we only allow angles between 0 and  $\pi/2$  even though two vectors can, in principle be thought of as having an angle between 0 and  $\pi$ . Instead, we regard the angle in terms of the smallest angle between any scalar multiples ( $\pm$  in  $\mathbb{R}^N$  and  $e^{i\theta}$  in  $\mathbb{C}^N$ ) of the vectors. So in Euclidean space the angle corresponds to the angle between the lines formed by a pair of vectors as opposed to that formed by their positive rays.

In Euclidean space angle really means angle, but in other inner product spaces, especially spaces of *signals*, it is best thought of as a measure of *similarity*. For example, we can define an inner product space on all polynomials of a real variable  $x$  by setting

$$\langle p, q \rangle = \int_0^1 p(x)q(x) dx.$$

**Exercise 1.2.3.** Show that  $\langle p, q \rangle$  defines an inner product on the vector space of all polynomials. The corresponding norm is  $\|p\| = \sqrt{\langle p, p \rangle}$

Polynomials are not vectors in Euclidean space. They are functions. Nevertheless, we can define the angle between  $p$  and  $q$  in terms of inner product as

$$\text{angle}(p, q) = \arccos\left(\frac{|\langle p, q \rangle|}{\|p\|\|q\|}\right).$$

But this angle is no longer an angle in the Euclidean sense. Instead, it is a measure of *similarity* in the sense that the angle is zero if and only if  $q$  is a multiple of  $p$ . If two polynomials have angle zero then we say that they are *orthogonal* by analogy with the Euclidean case.

**Exercise 1.2.4.** Find the angles between 1 and  $x$  and between 1 and  $1 - 2x$ . Explain why two polynomials are orthogonal if and only if the average value of their product on  $[0, 1]$  is zero.

### 1.2.1 Orthogonal and unitary matrices

Recall that, given a pair of bases  $\mathcal{B}_1 = \{\mathbf{x}_1, \dots, \mathbf{x}_N\}$  and  $\mathcal{B}_2 = \{\mathbf{y}_1, \dots, \mathbf{y}_M\}$  for a pair of  $N$  and  $M$  dimensional vector spaces  $V$  and  $W$ , one can define the matrix  $L = L_{\mathcal{B}_1, \mathcal{B}_2}$  of a linear operator  $\mathcal{L} : V \rightarrow W$  with respect to the bases  $\mathcal{B}_1$  and  $\mathcal{B}_2$  to be the matrix whose  $(i, j)$ -entry  $a_{ij}$  is the  $\mathbf{y}_j$  coefficient of  $\mathcal{L}\mathbf{x}_i$ . In other words,  $\mathcal{L}\mathbf{x}_i = \sum_{j=1}^M a_{ij}\mathbf{y}_j$ .

As a particular example one has the so-called *change of basis* matrix when one is representing the identity operator on  $V$  with two different bases  $\mathcal{B}_1$  and  $\mathcal{B}_2$ . This will be an extremely important feature of what we do in this course because the choice of a particular basis brings forth different types of information stored in a signal or image.

We will start by considering a very simple special case. Let  $\mathbb{R}^2$  be the Euclidean plane and consider the *standard basis vectors*  $\mathbf{e}_1 = [1, 0]^T$  and  $\mathbf{e}_2 = [0, 1]^T$ . The matrix of the identity operator in this basis is just the  $2 \times 2$  identity matrix. On the other hand, suppose that we want to represent a vector  $\mathbf{x} = [x_1, x_2]^T = x_1\mathbf{e}_1 + x_2\mathbf{e}_2$  in terms of its average value and the difference from its average. Consider the matrix

$$H = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

We get

$$H\mathbf{x} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} x_1 + x_2 \\ x_1 - x_2 \end{pmatrix}.$$

In other words we have

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \begin{pmatrix} x_1 + x_2 \\ x_1 - x_2 \end{pmatrix} = \frac{x_1 + x_2}{2} \begin{pmatrix} 1 \\ 1 \end{pmatrix} + \frac{x_1 - x_2}{2} \begin{pmatrix} 1 \\ -1 \end{pmatrix} = \frac{x_1 + x_2}{\sqrt{2}} \mathbf{h}_1 + \frac{x_1 - x_2}{\sqrt{2}} \mathbf{h}_2$$

where  $\mathbf{h}_1 = [1, 1]^T/\sqrt{2}$  and  $\mathbf{h}_2 = [1, -1]^T/\sqrt{2}$ . Thus we have expressed  $\mathbf{x}$  in terms of the sum and difference of its standard coordinates.

### 1.2.2 Rotations

One can take a more general geometric view of this. The matrix  $H$  also can be regarded as representing a rotation of a vector  $\mathbf{x}$  through an angle  $-\pi/4$ . In fact, the matrix

$$R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$$

applied to any input vector represents a rotation of the vector through an angle  $\theta$  and  $H = R_{-\pi/4}$ . If any pair of vectors  $\mathbf{x}$  and  $\mathbf{y}$  is input to  $R_\theta$ , they will be rotated by the same amount. For this reason, the matrix  $R_\theta$  is said to *preserve angles* between the input vectors.

**Exercise 1.2.5.** Show that  $\text{angle}(R_\theta\mathbf{x}, R_\theta\mathbf{y}) = \text{angle}(\mathbf{x}, \mathbf{y})$ .

**Exercise 1.2.6.** Shows that if  $\mathcal{B}_1, \mathcal{B}_2, \mathcal{B}_3$  are three bases for a vector space  $X$  and if  $\mathcal{L}$  is a linear operator on  $X$  then  $L_{\mathcal{B}_1, \mathcal{B}_3}$ , the matrix of  $\mathcal{L}$  with respect to  $\mathcal{B}_1$  and  $\mathcal{B}_3$ , satisfies  $L_{\mathcal{B}_1, \mathcal{B}_3} = L_{\mathcal{B}_2, \mathcal{B}_3} I_{\mathcal{B}_1, \mathcal{B}_2}$ , where  $I_{\mathcal{B}_1, \mathcal{B}_2}$  that takes the coordinates of  $\mathbf{x}$  in the basis  $\mathcal{B}_1$  to the coordinates in the basis  $\mathcal{B}_2$ .

On the other hand, it is clear that the columns of  $R_\theta$  are orthogonal to one another. Rotations in higher dimensions can be understood, at least to some extent, by decomposing them *inductively* by dimension. Consider the case of rotations in three real variables. As we learn in calculus or somewhere else perhaps, any rotation has a fixed axis. Let  $\mathbf{v}$  be the fixed direction. Then we can think of our rotation as mapping  $\mathbf{v}$  to  $\mathbf{e}_3 = [0, 0, 1]$  followed by a rotation that fixes  $[0, 0, 1]$  followed by the inverse of the first rotation. Let  $\rho_{\mathbf{v}}$  denote the rotation that maps  $\mathbf{v} = [\cos \vartheta \sin \varphi, \sin \vartheta \sin \varphi, \cos \varphi]$  to  $[0, \sin \varphi, \cos \varphi]$  followed by the one that sends  $[0, \sin \varphi, \cos \varphi]$  to  $[0, 0, 1]$ . Since composition of rotations can be expressed as products of the rotation matrices, this can all be expressed as

$$\rho_{\mathbf{v}} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & \cos \varphi & -\sin \varphi \\ 0 & \sin \varphi & \cos \varphi \end{pmatrix} \begin{pmatrix} \sin \vartheta & -\cos \vartheta & 0 \\ \cos \vartheta & \sin \vartheta & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} \sin \vartheta & -\cos \vartheta & 0 \\ \cos \vartheta \cos \varphi & \sin \vartheta \cos \varphi & -\sin \varphi \\ \cos \vartheta \sin \varphi & \sin \vartheta \sin \varphi & \cos \varphi \end{pmatrix}$$

This rotation is followed by the rotation that fixes  $\mathbf{e}_3$  and rotates through an angle  $\theta$  in the  $(x, y)$  plane, then is followed by the inverse of  $\rho_{\mathbf{v}}$ . Recall that this inverse can be obtained from the general formula for the inverse of a  $3 \times 3$  matrix and this gives

$$\rho_{\mathbf{v}}^{-1} = \begin{pmatrix} \sin \vartheta & \cos \vartheta \cos \varphi & \cos \vartheta \sin \varphi \\ -\cos \vartheta & \sin \vartheta \cos \varphi & \sin \vartheta \sin \varphi \\ 0 & -\sin \varphi & \cos \varphi \end{pmatrix}$$

Remarkably, one sees that  $\rho_{\mathbf{v}}^{-1}$  is nothing but the transpose of  $\rho_{\mathbf{v}}$  in this case. A general rotation of  $\mathbb{R}^3$  can be expressed in the form  $\rho_{\mathbf{v}}^{-1} R_{\theta} \rho_{\mathbf{v}}$  that is,

$$\rho_{\mathbf{v}}^{-1} R_{\theta} \rho_{\mathbf{v}} = \begin{pmatrix} \sin \vartheta & \cos \vartheta \cos \varphi & \cos \vartheta \sin \varphi \\ -\cos \vartheta & \sin \vartheta \cos \varphi & \sin \vartheta \sin \varphi \\ 0 & -\sin \varphi & \cos \varphi \end{pmatrix} \begin{pmatrix} \sin \theta & -\cos \theta & 0 \\ \cos \theta & \sin \theta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} \sin \vartheta & -\cos \vartheta & 0 \\ \cos \vartheta \cos \varphi & \sin \vartheta \cos \varphi & -\sin \varphi \\ \cos \vartheta \sin \varphi & \sin \vartheta \sin \varphi & \cos \varphi \end{pmatrix}$$

In  $N$  dimensions one can proceed in much the same fashion, that is, any rotation of  $\mathbb{R}^N$  has the form  $R = \rho_{\mathbf{v}}^{-1} R_{N-1} \rho_{\mathbf{v}}$  in which  $\rho_{\mathbf{v}}$  maps  $\mathbf{v}$  to the *north pole*  $\mathbf{e}_N = [0, \dots, 0, 1]^T$ , and  $R_{N-1}$  can be regarded as a rotation of  $\mathbb{R}^{N-1}$  — the Euclidean space of one less dimension, regarded here as the span of the standard basis vectors  $\mathbf{e}_1, \dots, \mathbf{e}_{N-1}$  inside of  $\mathbb{R}^N$ . Unfortunately, trying to understand a rotation by such a dimension reduction approach is problematic since writing down the matrices becomes cumbersome for dimensions  $N > 3$ , and even more so in the case of  $\mathbb{C}^N$ . It will help, instead, to take a more abstract approach.

The first thing to notice about the rotation matrices is that they preserve angles between vectors and that they also preserve the lengths of vectors. This is a property of rotations, but let's see how it can be expressed in terms of matrices with respect to the standard basis on  $\mathbb{R}^N$  or  $\mathbb{C}^N$  and the usual dot product for inner product. Suppose then that we have a matrix  $O$  that, regarded as a linear mapping in standard coordinates, preserves lengths of vectors and angles between them. Then, for any vectors  $\mathbf{x}, \mathbf{y}$ ,

$$\text{angle}(O\mathbf{x}, O\mathbf{y}) = \frac{|\langle O\mathbf{x}, O\mathbf{y} \rangle|}{\|O\mathbf{x}\| \|O\mathbf{y}\|} = \frac{|\langle \mathbf{x}, \mathbf{y} \rangle|}{\|\mathbf{x}\| \|\mathbf{y}\|} = \text{angle}(\mathbf{x}, \mathbf{y}).$$

Since lengths are preserved, this also tells us that

$$|\langle O\mathbf{x}, O\mathbf{y} \rangle| = |\langle \mathbf{x}, \mathbf{y} \rangle|.$$

But any matrix  $A$  acting on  $\mathbb{R}^N$  satisfies  $\langle A\mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{x}, A^T \mathbf{y} \rangle$ . Therefore, if  $O$  preserves angles then for any vectors  $\mathbf{x}$  and  $\mathbf{y}$ ,

$$|\langle \mathbf{x}, O^T O \mathbf{y} \rangle| = |\langle \mathbf{x}, \mathbf{y} \rangle|.$$

In particular,  $O^T O \mathbf{y}$  is pointing in the same or the opposite direction as  $\mathbf{y}$ . In particular,  $O^T O \mathbf{e}_i = \pm \mathbf{e}_i$  for each standard basis vector  $\mathbf{e}_i$ . Thus we have shown the following (using the notation  $\text{diag}(\lambda_1, \dots, \lambda_N)$  for the diagonal matrix with  $i$ -th main diagonal entry  $\lambda_i$ ).

**Theorem 1.2.7.** *Suppose that an  $N \times N$  matrix  $O$ , expressed in standard coordinates, preserves angles and lengths between vectors in  $\mathbb{R}^N$ . Then  $O^T O = \text{diag} \pm 1$ .*

A matrix satisfying the condition of the theorem is called an *orthogonal matrix*, although one often reserves the term for the case  $O^T O = I$ , sometimes called a *special orthogonal matrix*. Since the inner product in  $\mathbb{R}^N$  is just the dot product, the condition  $O^T O = I$  just says that the columns of  $O$  are orthonormal, that is, if  $\mathbf{c}_j$  is the  $j$ -th column of  $O$  then  $\mathbf{c}_j \cdot \mathbf{c}_k = \delta_{jk}$ .

**Lemma 1.2.8.** *Any collection of mutually orthogonal vectors is linearly independent. In particular, any collection of  $N$  orthonormal vectors forms an orthonormal basis for  $\mathbb{R}^N$ .*

This tells us that if  $O$  is an orthogonal  $N \times N$  matrix then its columns form an orthonormal basis for  $\mathbb{R}^N$ . In the case of  $\mathbb{C}^N$ , matrices  $U$  that preserve lengths and angles are called *unitary matrices* and are characterized by  $\bar{U}^T U \mathbf{y} = e^{i\theta} \mathbf{y}$  for some  $\theta(\mathbf{y})$  (if  $\theta = 0$  then  $U$  is a *special unitary matrix*).

### 1.2.3 Variance maximizing rotation

In the three-dimensional world, rotations have the effect of changing perspective for a fixed observer. For example, a coin can look like a line segment from one perspective, but like a full disk from another, the latter giving more information such as whose face is replicated on the coin. In linear regression problems in statistics, one seeks a plane or hyperplane that best fits the data in the sense that the sum of the squares of the distances from the data points to the given hyperplane is minimized over all choices of hyperplane. If the data is expressed in terms of  $N$  coordinates and if the hyperplane fit is good, one is able to conclude that one of the coordinates is, up to some uncontrollable factors, a linear function of the other  $N - 1$  (or possibly even fewer) coordinates.

Finding such a hyperplane is often done iteratively. That is, first one seeks a principal axis about which the variances of the data are maximized along this axis. One proceeds to project the data in the hyperplane perpendicular to this axis, and so on. If for example, the data has three dimensions and essentially lies in a plane, then the two principal axis found will determine this plane. The rotation that takes the first two standard basis vectors to the vectors in the direction of these principal axes is then referred to as a variance maximizing rotation. When these axes are determined by experimental data, the terminology suggests that the principal components of the data are essentially uncorrelated. When looking down on the principle axes one sees the face, or back, of the coin.

This sort of analysis only pertains in cases in which the given data lies on a linear hyperplane. In many situations, the variables are actually related in a nonlinear fashion and more sophisticated techniques are required to uncover underlying rules governing given observations.

### 1.2.4 Projection matrices

Orthogonal matrices satisfying  $O^T O = I$  behave the same on all vector pairs. An extreme opposite of this would be a matrix that fixes vectors in certain directions and nulls out vectors in other directions. Such behavior can be captured algebraically with the notion of *idempotency*. A matrix  $E$  that satisfies  $E^2 = E$  is said to be idempotent. If a vector  $\mathbf{x}$  lies in the column space of an idempotent  $E$  then  $E$  acts as the identity on  $\mathbf{x}$ . Because of this property, thought of as an operator,  $E$  is called a *projection* onto its column space. Among idempotent matrices, there are special examples having the property that *the null space of  $E$  is the same as the orthogonal complement of the column space of  $E$* . If  $E$  has this property then, as a linear operator in the standard basis,  $E$  is said to be an *orthogonal projection* operator. Consider here a pair of examples.

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}; \quad E_2 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}$$

For  $\mathbf{x} = [x_1, x_2]^T$ , one has  $E_1 \mathbf{x} = [x_1, 0]^T$  and  $E_2 \mathbf{x} = [x_1 + x_2, 0]^T$ . Both  $E_1$  and  $E_2$  are projections into the subspace of  $\mathbb{R}^2$  of all vectors of the form  $[t, 0]^T$ ; however,  $E_1$  is an orthogonal projection since its null space is precisely all vectors of the form  $[0, t]^T$  which is the same as the orthogonal complement of the column space, whereas the null space of  $E_2$  consists of all vectors of the form  $[t, -t]^T$  which is not the same as the orthogonal complement of the column space. The matrix operator  $E_2$  is called an *oblique* projection.

It would be convenient to have a characterization of projection operators and of orthogonal projection operators in terms of orthonormal bases. For this purpose we need to pull out a second major tool from linear algebra, namely the *singular value decomposition* (SVD) which states that any  $m \times n$  matrix  $M$  can be written

$$M = U \Sigma V^* \tag{1.1}$$

in which  $U$  and  $V$  are unitary  $m \times m$  and  $n \times n$  matrices, respectively, and  $\Sigma$  is an  $m \times n$  matrix that is zero outside the first  $r$  diagonal elements where  $r$  is the rank of  $M$ . Here and below,  $M^*$  will refer to the conjugate transpose  $\overline{M}^T$  of  $M$ . The nonzero values of  $\Sigma$  are referred to as singular values. In the special case when  $M$  is Hermitian, one has  $U = V$  and the singular values are the same as the eigenvalues.

In the case that  $P$  is idempotent with SVD  $P = U \Sigma V^*$ , the condition  $P^2 = P$  translates to

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

Since  $U$  and  $V$  are unitary matrices, one can multiply both sides of this identity on the left by  $U^*$  and on the right by  $V$  to get

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

Since the  $j$ -th column of  $\Sigma$  is  $\lambda_j \mathbf{e}_j$  for some  $\lambda_j \geq 0$ , by applying the equation columnwise, one gets  $\Sigma V^* U \lambda_j \mathbf{e}_j = \lambda_j \mathbf{e}_j$ . If  $\lambda_j = 0$  then both sides equal zero but if  $\lambda_j > 0$  then one has  $\Sigma V^* U \mathbf{e}_j = \mathbf{e}_j$  with  $\Sigma_{jj} = \lambda_j$ . If  $M \leq N$  is the number of nonzero diagonal entries of  $\Sigma$  then it has to be the case that the first  $M$  columns of  $\Sigma V^* U$  are equal to the first  $M$  columns of the identity matrix. In particular, for  $1 \leq j, k \leq M$  the inner product of the  $j$ -th column of  $V$  and  $k$ -th column of  $U$  has to be  $\delta_{jk}$ . Therefore the first  $M$  columns of  $U$  and  $V$  should both be equal to (the same) orthonormal basis for the range of  $P$ . Since  $U$  and  $V$  are unitary, the remaining columns of  $U$  and  $V$  should be equal to an orthonormal basis for the orthogonal complement of the range of  $P$ . The  $N - M$  remaining columns of  $U$  and  $V$  are not required to have the same order, so one need not have  $U = V$ .

**Exercise 1.2.9.** Compute the SVDs of each of the following:

$$E_1 = \begin{pmatrix} 1 & 0 \\ 0 & 0 \end{pmatrix}; \quad E_2 = \begin{pmatrix} 1 & 1 \\ 0 & 0 \end{pmatrix}; \quad E_3 = \begin{pmatrix} 1 & 0 \\ -1 & 0 \end{pmatrix}$$

**Exercise 1.2.10.** Use matlab's `svd` command to compress an image. Start with a  $128 \times 128$  or  $256 \times 256$  image. Look for help under `imread` for reading an image into a file. You should convert the image to *grayscale*: do a matlab documentation search under `grayscale` and/or `colormap(gray)`. Perform the `svd` on the grayimage. Plot the eigenvalues versus index and look for a good cutoff point. Zero out those rows and columns corresponding to small eigenvalues and then reconstruct from the remaining ones by remultiplying the component matrices. Comment on the fidelity, both visually and in terms of  $\ell^2$  error.

### 1.3 Infinite scales and infinite dimensions: limits

#### 1.4 The space $\ell^2(\mathbb{N})$

The space  $\mathbb{C}^N$  is a useful context for many types of signals. For example, in Matlab, any *signal* is really just a finite list of complex numbers. However, there are a lot of contexts in which one does not want to constrain, in advance, the dimensionality of the set of objects with which one is willing to work. A signal could consist of 100 data points, but it could also consist of a million data points or a billion, or more data points than can be stored in a computer's memory at one time. For this reason and others, it is useful not to limit the length of a vector (*e.g.* the number of data points), but instead to say what it means for the magnitude of the vector to be limited.

For this reason we make the following definition.

**Definition 1.4.1.** The space  $\ell^2(\mathbb{N})$  consists of all sequences  $\mathbf{z} = (z_1, z_2, \dots)$  with  $z_k \in \mathbb{C}$  for  $k = 1, 2, \dots$  such that  $\sum_{k=1}^{\infty} |z_k|^2 < \infty$ . One defines the inner product

$$\langle \mathbf{z}, \mathbf{w} \rangle = \sum_{k=1}^{\infty} z_k \bar{w}_k$$

and the norm

$$\|\mathbf{z}\| = \|\mathbf{z}\|_{\ell^2} = \sqrt{\langle \mathbf{z}, \mathbf{z} \rangle}.$$

When  $\mathbf{z}$  and  $\mathbf{w}$  have only finitely many nonzero terms they can be thought of as elements of  $\mathbb{C}^N$  for  $N$  large enough and in that case their inner-product is just the same as their inner product in  $\mathbb{C}^N$ . There are some nontrivial issues, including: (i)  $\ell^2(\mathbb{N})$  is complete and (ii) the inner product on  $\ell^2(\mathbb{N})$  is continuous.

##### 1.4.1 Completeness of $\ell^2(\mathbb{Z}^N)$

The first question we need to answer is: what is a limit point in  $\ell^2(\mathbb{N})$ ? Just as in  $\mathbb{C}^N$ , a limit point is a point in the space such that a sequence of elements of the space converges to it. A sequence of points  $\{\mathbf{z}^k\}$ ,  $k = 1, 2, \dots$  in  $\ell^2(\mathbb{N})$  is a sequence of sequences of numbers. Thus  $\mathbf{z}^k = (z_1^k, z_2^k, z_3^k, \dots)$ . That  $\mathbf{z}^k \rightarrow \mathbf{z}$  in  $\ell^2$  means that for every  $\varepsilon > 0$  there is an  $M$  such that  $k \geq M$  implies that  $\|\mathbf{z}^k - \mathbf{z}\|_{\ell^2} < \varepsilon$ . Since  $\ell^2(\mathbb{N})$  is a vector space,  $\mathbf{z}^k \rightarrow \mathbf{z}$  in  $\ell^2$  is the same as  $\mathbf{z}^k - \mathbf{z} \rightarrow \mathbf{0}$  in  $\ell^2$ . Now set  $\mathbf{w}^k = \mathbf{z}^k - \mathbf{z}$ .

The first observation is that if  $\mathbf{w}^k \rightarrow \mathbf{0}$  in  $\ell^2$  then  $\mathbf{w}^k \rightarrow \mathbf{w}$  componentwise, that is,  $z_j^k \rightarrow z_j$  for each  $j$ . This follows directly from the fact that  $\sup_j |w_j| \leq \|\mathbf{w}\|_{\ell^2}$ . *Warning:* componentwise convergence does not imply convergence in  $\ell^2$ . As an example, let  $\mathbf{e}^k$  be the sequence such that  $e_j^k = \delta_{jk}$ . Then, as soon as  $k > j$  we have  $e_j^k = 0$  and so  $\mathbf{e}^k \rightarrow \mathbf{0}$  componentwise. However,  $\|\mathbf{e}^k - \mathbf{0}\|_{\ell^2} = \|\mathbf{e}^k\|_{\ell^2} = 1$  for all  $k = 1, 2, \dots$  so  $\mathbf{e}^k \not\rightarrow \mathbf{0}$  in  $\ell^2$ .

Nevertheless,  $\ell^2(\mathbb{N})$  is a complete metric space, meaning that any Cauchy sequence in  $\ell^2$  converges to a limit in  $\ell^2$ . By a Cauchy sequence we mean a sequence  $\mathbf{z}^k$  such that, for any  $\varepsilon > 0$ , there is an  $M = M(\varepsilon)$  such that whenever  $k, \ell > M$  one has  $\|\mathbf{z}^k - \mathbf{z}^\ell\|_{\ell^2} < \varepsilon$ . If  $\mathbf{z}^k$  is a Cauchy sequence then each coordinate sequence is a Cauchy sequence of real numbers, hence has a limit, call it  $z_j$  and set  $\mathbf{z} = (z_1, z_2, \dots)$ . One then has to show that  $\mathbf{z}^k \rightarrow \mathbf{z}$ . But this follows since, if  $k > M(\varepsilon)$  then

$$\sum_{j=1}^{\infty} |z_j^k - z_j|^2 \leq \sum_{j=1}^{\infty} \sup_{\ell > k} |z_j^k - z_j^\ell|^2 \leq \varepsilon^2$$

by the definition of a Cauchy sequence. Therefore, if  $k > M(\varepsilon)$  then  $\|\mathbf{z}^k - \mathbf{z}\|_{\ell^2} \leq \varepsilon$ . Since this can be done for any  $\varepsilon > 0$  it follows that  $\mathbf{z}^k \rightarrow \mathbf{z}$  in  $\ell^2$ .

The vectors  $\mathbf{e}^k$  encountered above form an orthonormal set in  $\ell^2(\mathbb{N})$  since

$$\langle \mathbf{e}^k, \mathbf{e}^\ell \rangle = \sum_{j=1}^{\infty} e_j^k e_j^\ell = \sum_{j=1}^{\infty} \delta_{jk} \delta_{j\ell} = \delta_{k\ell}$$

#### 1.4.2 Orthogonal complements and orthogonal projections in $\ell^2(\mathbb{N})$

A convex subset of a Hilbert space  $\mathcal{H}$  is a set  $C$  having the property that if  $\mathbf{x}$  and  $\mathbf{y}$  are in  $C$  and  $0 \leq t \leq 1$  then the *line segment*  $t\mathbf{x} + (1-t)\mathbf{y}$  lies entirely in  $C$ .

**Lemma 1.4.2.** *Suppose that  $C$  is a closed, convex subset of the Hilbert space  $\mathcal{H}$  and suppose that  $\mathbf{y} \notin C$ . Then there is a unique  $\mathbf{x} \in C$  that minimizes  $\|\mathbf{x} - \mathbf{y}\|_{\mathcal{H}}$ .*

*Proof.* Set  $S = \{\|\mathbf{x} - \mathbf{y}\| : \mathbf{x} \in C\}$ . Then  $S$  is a set of strictly positive numbers and, since  $C$  is closed and  $\mathbf{y} \notin C$ , one has  $d = \inf S > 0$ . We can add  $y$  to itself and to all elements of  $C$  so the problem becomes that of finding an element of  $C$  that is closest to  $\mathbf{0}$  when  $\mathbf{0} \notin C$  as we assume now. By definition of  $\inf$  we can find a sequence  $\mathbf{x}_k$  of elements of  $C$  such that  $\|\mathbf{x}_k\| = d_k \downarrow d$  as  $k \rightarrow \infty$ . We claim that  $\mathbf{x}_k$  is a Cauchy sequence. To see this, by the parallelogram law,

$$\|\mathbf{z}_k - \mathbf{w}_k\|^2 = -\|\mathbf{z}_k + \mathbf{w}_k\|^2 + 2(\|\mathbf{z}_k\|^2 + \|\mathbf{w}_k\|^2) \leq -(2d)^2 + 2(\|\mathbf{z}_k\|^2 + \|\mathbf{w}_k\|^2) \rightarrow 0$$

since the convexity of  $C$  implies that

$$\|\mathbf{z}_k + \mathbf{w}_k\|^2 = 2\left\|\frac{1}{2}(\mathbf{z}_k + \mathbf{w}_k)\right\|^2 \geq 2d^2.$$

This proves the lemma.

**Exercise 1.4.3.** (Parallelogram law) Show that in a Hilbert space  $\mathcal{H}$ ,  $\|\mathbf{x} + \mathbf{y}\|^2 + \|\mathbf{x} - \mathbf{y}\|^2 = 2(\|\mathbf{x}\|^2 + \|\mathbf{y}\|^2)$ .

If  $\mathcal{K}$  is a closed subspace of  $\mathcal{H}$  then  $\mathcal{K}$  is convex. Suppose that  $\mathbf{y} \notin \mathcal{K}$ , assuming that  $\mathcal{K}$  is a proper subspace, and let  $\mathbf{x} \in \mathcal{K}$  be the closest element to  $\mathbf{y}$  in  $\mathcal{K}$ . We claim that  $\mathbf{x} - \mathbf{y}$  is orthogonal to  $\mathcal{K}$ . If this were not the case then there would exist some  $\mathbf{k} \in \mathcal{K}$  such that  $\langle \mathbf{x} - \mathbf{y}, \mathbf{k} \rangle \neq 0$ . Set  $\alpha = \langle \mathbf{x} - \mathbf{y}, \mathbf{k} \rangle / \|\mathbf{k}\|^2$ . Then a simple calculation shows that  $\|\mathbf{y} - (\mathbf{x} + \alpha\mathbf{k})\|^2 = \|\mathbf{x} - \mathbf{y}\|^2 - |\langle \mathbf{x} - \mathbf{y}, \mathbf{k} \rangle|^2 / \|\mathbf{k}\|^2 < d^2$  which would contradict minimality of  $\|\mathbf{x} - \mathbf{y}\|$ . This proves the orthogonality of  $\mathbf{x} - \mathbf{y}$  with  $\mathcal{K}$ .

**Corollary 1.4.4.** *Let  $\mathcal{K}$  be a closed proper subspace of the Hilbert space  $\mathcal{H}$ . Then each  $\mathbf{z}$  has a unique decomposition  $\mathbf{z} = \mathbf{k} + (\mathbf{z} - \mathbf{k})$  such that  $\mathbf{k} \in \mathcal{K}$  and  $\mathbf{z} - \mathbf{k} \perp \mathcal{K}$ . The vector  $\mathbf{k}$  is called the orthogonal projection of  $\mathbf{z}$  onto  $\mathcal{K}$ .*

Uniqueness is seen as follows. Suppose that  $\mathbf{z} = \mathbf{k}_1 + (\mathbf{z} - \mathbf{k}_1) = \mathbf{k}_2 + (\mathbf{z} - \mathbf{k}_2)$ . Then  $\mathbf{k}_1 - \mathbf{k}_2 = (\mathbf{z} - \mathbf{k}_2) - (\mathbf{z} - \mathbf{k}_1)$  in which the left hand side lies in  $\mathcal{K}$  and the right hand side lies in the orthogonal complement of  $\mathcal{K}$ . Since  $\mathcal{K} \cap \mathcal{K}^\perp = \{\mathbf{0}\}$  the uniqueness of the decomposition of  $\mathbf{z}$  into an element of  $\mathcal{K}$  plus an element orthogonal to  $\mathcal{K}$  follows. Thus we see that the orthogonal projection of  $\mathbf{z}$  onto  $\mathcal{K}$  is the unique element of  $\mathcal{K}$  closest to  $\mathbf{z}$ .

Just as with  $\mathbb{C}^N$  we can define an idempotent operator on  $\ell^2(\mathbb{N})$  by the condition that  $P^2 = P$  as a linear operator from  $\ell^2(\mathbb{N})$  back into itself. Such an idempotent operator is an *orthogonal projection operator* provided that the null space of  $P$  is the orthogonal complement of the range of  $P$ , that is, if  $\langle x, Py \rangle = 0$  for all  $y \in \mathcal{H}$  if and only if  $Px = 0$ .

### 1.4.3 Complete orthonormal bases

In  $\mathbb{C}^N$  any vector  $\mathbf{z} = (z_1, \dots, z_N) = z_1\mathbf{e}_1 + z_2\mathbf{e}_2 + \dots + z_N\mathbf{e}_N$  where  $\mathbf{e}_k$  is the  $k$ -th standard basis vector. In general one can write  $\mathbf{z} = UU^*\mathbf{z} = \sum_{k=1}^N (U^*\mathbf{z})_k \mathbf{u}_k$  where  $\mathbf{u}_k$  is the  $k$ -th column of the unitary matrix  $U$ . Obviously, but importantly, any unitary matrix is invertible and so any vector in  $\mathbb{C}^N$  can be written as a linear combination of the columns of  $U$ , and any orthonormal basis can be identified, up to an ordering of its elements, with a unitary matrix.

Matters are more complicated in  $\ell^2(\mathbb{N})$  because the dimension is infinite. For example, the collection  $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3, \dots$  forms an orthonormal system and any element  $\mathbf{z} \in \ell^2(\mathbb{N})$  has the form  $\mathbf{z} = \sum_{k=1}^{\infty} z_k \mathbf{e}_k$ . On the other hand, the family  $\mathbf{f}_k = \frac{1}{\sqrt{2}}(\mathbf{e}_{2k} + \mathbf{e}_{2k+1})$  also forms an orthonormal system as one can easily check; however, the family  $\mathbf{f}_k$  is not complete. How can we determine whether an orthonormal family in  $\ell^2(\mathbb{N})$  is complete?

**Definition 1.4.5.** A collection  $\mathbf{g}_k, k = 1, 2, 3, \dots$  is complete in  $\ell^2(\mathbb{N})$  if, for any element  $\mathbf{z} \in \ell^2(\mathbb{N})$  there are numbers  $c_k \in \mathbb{C}$  such that  $\mathbf{z} = \sum_{k=1}^{\infty} c_k \mathbf{g}_k$ . The sequence  $\{\mathbf{g}_k\}$  is a basis for  $\ell^2(\mathbb{N})$  if the coefficients  $c_k$  are unique, that is, if  $\mathbf{z} = \sum_{k=1}^{\infty} c_k \mathbf{g}_k$  and  $\mathbf{z} = \sum_{k=1}^{\infty} d_k \mathbf{g}_k$  then  $c_k = d_k$  for each  $k = 1, 2, \dots$ .

**Exercise 1.4.6.** (Bessel's inequality) Show that if  $\{\mathbf{g}_k\}$  forms an orthonormal family in  $\ell^2(\mathbb{N})$  then, whenever  $\mathbf{z} \in \ell^2(\mathbb{N})$ , the sum  $\sum_{k=1}^{\infty} \langle \mathbf{z}, \mathbf{g}_k \rangle \mathbf{g}_k$  defines an element of  $\ell^2(\mathbb{N})$  and  $\sum_{k=1}^{\infty} |\langle \mathbf{z}, \mathbf{g}_k \rangle|^2 \leq \|\mathbf{z}\|_{\ell^2}^2$ . This estimate is known as *Bessel's inequality*.

**Exercise 1.4.7.** Show that if  $\{\mathbf{g}_k\}$  forms an orthonormal basis for  $\ell^2(\mathbb{N})$  then any  $\mathbf{z} \in \ell^2(\mathbb{N})$  can be written uniquely as  $\mathbf{z} = \sum_{k=1}^{\infty} \langle \mathbf{z}, \mathbf{g}_k \rangle \mathbf{g}_k$ . Show that if  $\{\mathbf{g}_k\}$  is an orthonormal basis then  $\sum_{k=1}^{\infty} |\langle \mathbf{z}, \mathbf{g}_k \rangle|^2 = \|\mathbf{z}\|_{\ell^2}^2$ .

**Exercise 1.4.8.** Show that an orthonormal family  $\{\mathbf{g}_k\}$  in  $\ell^2(\mathbb{N})$  is a basis for  $\ell^2(\mathbb{N})$  if and only if the condition that  $\langle \mathbf{z}, \mathbf{g}_k \rangle = 0$  for all  $k = 1, 2, \dots$  implies that  $\mathbf{z} = \mathbf{0}$ .

**Exercise 1.4.9.** Show that any sequence  $\{\mathbf{g}_k\}$  in  $\ell^2(\mathbb{N})$  is complete in  $\ell^2(\mathbb{N})$  if and only if the condition that  $\langle \mathbf{z}, \mathbf{g}_k \rangle = 0$  for all  $k = 1, 2, \dots$  implies that  $\mathbf{z} = \mathbf{0}$ .

**Exercise 1.4.10.** Let  $\mathbf{f}_k = \frac{1}{\sqrt{2}}(\mathbf{e}_{2k} + \mathbf{e}_{2k+1})$  where  $\mathbf{e}_k$  is the  $k$ -th standard basis vector in  $\ell^2(\mathbb{N})$ . Describe the subspace of vectors that are orthogonal to each of the  $\mathbf{e}_k$ .

### Unitary transformations in $\ell^2(\mathbb{N})$

In  $\mathbb{C}^N$ , a unitary transformation is expressed in terms of the standard basis as any matrix having the property that if its columns are denoted by  $\mathbf{u}_k$  then  $\mathbf{u}_j \cdot \bar{\mathbf{u}}_k = \delta_{jk}$ . The same holds in  $\ell^2(\mathbb{N})$ , if one replaces the dot product by the inner product

$$\langle \mathbf{u}, \mathbf{v} \rangle = \sum_{\nu=1}^{\infty} u_\nu \bar{v}_\nu$$

where  $\mathbf{u} = \sum_{k=1}^{\infty} u_k \mathbf{e}_k$  in standard coordinates on  $\ell^2(\mathbb{N})$ . The added condition is that the *columns* of the matrix  $U$  now have to be elements of  $\ell^2(\mathbb{N})$ . Thus a unitary transformation on  $\ell^2(\mathbb{N})$  is an operator that can be represented in standard coordinates by a matrix  $U$  whose columns  $\mathbf{u}_k$  are elements of  $\ell^2(\mathbb{N})$  and  $\langle \mathbf{u}_j, \mathbf{u}_k \rangle = \delta_{jk}$  for all  $j, k = 1, 2, 3, \dots$ .

Although this seems fairly straightforward there is a major difference from the case of  $\mathbb{C}^N$ . In that case, the columns of  $U$  form an orthonormal basis for  $\mathbb{C}^N$ . In the case of  $\ell^2(\mathbb{N})$  there is no guarantee that the

orthonormal columns of  $U$  will be complete in  $\ell^2(\mathbb{N})$ . In  $N$  dimensions any collection of  $N$  orthonormal vectors will form a basis. The corresponding statement for  $\ell^2(\mathbb{N})$  would seem to be that any infinite collection of orthogonal vectors is complete in  $\ell^2(\mathbb{N})$ . But this statement is false, as exemplified by the family  $\mathbf{f}_k = \frac{1}{\sqrt{2}}(\mathbf{e}_{2k} + \mathbf{e}_{2k+1})$ . The image of  $U$  could be a proper subspace of  $\ell^2(\mathbb{N})$ .

**Exercise 1.4.11.** Suppose that  $\{U_n u\}$ ,  $\nu = 1, 2, \dots$  is a sequence of unitary operator  $N \times N$  matrices. Let  $L$  be the operator that maps the  $k$ th basis vector  $\mathbf{e}_k$  in  $\ell^2(\mathbb{N})$  to the  $j$ -th column of  $U_\nu$  when  $k = \nu N + j$ . Show that this defines an invertible unitary operator on  $\ell^2(\mathbb{N})$  and describe its inverse.

#### 1.4.4 $\ell^2(\mathbb{Z})$

The space  $\ell^2(\mathbb{N})$  is an example of a *separable* Hilbert space. The definition of *separable* is that there exists a countable (infinite in this case) basis for the space. A countable basis is a countable set  $\{\mathbf{b}_1, \mathbf{b}_2, \dots\}$  such that (i) any finite subset is linearly independent and (ii) any element of the space has a unique representation  $\mathbf{z} = \sum_{k=1}^{\infty} c_j \mathbf{b}_j$ . In our case, the vectors  $\mathbf{e}^k$  form a countable orthonormal basis for  $\ell^2(\mathbb{N})$ . The practical meaning of the term separable is that, given two distinct elements of the space, we can find a single basis element whose coefficients are not equal. That is, we can find a single basis element that *separates* the two point in the space. In most of what follows, it will be crucial to think of sequences of numbers that are infinite in both directions. Such a  $\mathbf{z}$  can be expressed as  $(\dots, z_{-1}, z_0, z_1, \dots)$ . In other words, the coordinates of  $\mathbf{z}$  are indexed by the *integers* rather than by the natural numbers. If  $\mathbf{z}$  satisfies the convergence criterion. The standard basis for  $\ell^2(\mathbb{Z})$  consists of the orthogonal sequences  $\mathbf{e}^k$ . The  $j$ -th coordinate of  $\mathbf{e}^k$  is  $\delta_{jk}$ . Here both  $j$  and  $k$  range over all integers, not just over the natural numbers.

Why do we need to define sequence over  $\mathbb{Z}$  and not just  $\mathbb{N}$ ? The main reason is that we need to be able to work with the shift operator  $S$  that sends  $\mathbf{z}$  whose  $j$ -th coordinate is  $z_j$  to the shifted element  $S\mathbf{z}$  whose  $j$ -th element is  $z_{j-1}$ . In particular,  $(S\mathbf{z})_1 = z_0$  and  $(S\mathbf{z})_0 = z_{-1}$ . We would not be able to make sense of this operator if  $\mathbf{z}$  were indexed only by  $\mathbb{N}$  and not  $\mathbb{Z}$ . This operator is of fundamental importance in time-frequency analysis, particularly real-time analysis that involves delays in the propagation of information over communications channels, for example. If  $z_j$  represents a sample of some signal after  $j$  units of time then we might have  $z_j = S^j w_0$  where  $w_0$  was data produced at time zero.

#### 1.4.5 Operators on $\ell^2(\mathbb{Z})$

Just as operators on  $\mathbb{C}^N$  can be represented in the standard basis by matrix multiplication, operators on  $\ell^2(\mathbb{Z})$  can also be represented by matrices with respect to the standard basis  $\{\mathbf{e}^k\}$  for  $\ell^2(\mathbb{Z})$ . Here matrices are infinite dimensional and entries are indexed by pairs of integers. For example, the shift operator acting on the sequence  $\mathbf{z} = (\dots, z_{-1}, z_0, z_1, \dots)$  is  $S\mathbf{z} = (\dots, z_{-2}, z_{-1}, z_0, \dots)$ . In the standard basis, the matrix of  $S$  then has entries  $S_{jk} = \delta_{j,k+1}$  meaning that  $S_{jk} = 1$  if  $j = k + 1$  (the row number is the column number plus one) and  $S_{jk} = 0$  otherwise. The matrix operation of  $S$  is expressed visually as

$$\begin{pmatrix} \vdots \\ z_{-2} \\ z_{-1} \\ z_0 \\ \vdots \end{pmatrix} = \begin{pmatrix} \ddots & \ddots & \ddots & & \\ 0 & 1 & 0 & \cdots & \\ \cdots & 0 & 1 & 0_{(0,0)} & \cdots \\ & \cdots & 0 & 1 & 0 & \cdots \\ & & & \ddots & \ddots & \ddots \end{pmatrix} \begin{pmatrix} \vdots \\ z_{-1} \\ z_0 \\ z_1 \\ \vdots \end{pmatrix}.$$

In general, if  $\mathcal{L}$  is a linear operator from  $\ell^2(\mathbb{Z})$  to itself then the matrix of  $\mathcal{L}$  can be regarded as a function  $L(j, k)$  defined on  $\mathbb{Z} \times \mathbb{Z}$  and its action on  $\mathbf{z}$  is defined as  $\mathcal{L}\mathbf{z} = \mathbf{w}$  where

$$w_j = \sum_{k=-\infty}^{\infty} L(j, k) z_k$$

just as with ordinary matrix multiplication.

### 1.4.6 Isomorphisms of Separable Hilbert spaces

The exercises for this chapter contain several examples of separable Hilbert spaces. At some level, all infinite dimensional, separable Hilbert spaces are the same. What does this mean exactly?

First we consider a very simple example.

*Example 1.4.12.* ( $\ell^2(\mathbb{N})$  is isomorphic to  $\ell^2(\mathbb{Z})$ ). Let  $e_k, k = 1, 2, \dots$  denote the  $k$ -th standard basis element of  $\ell^2(\mathbb{N})$  and let  $e'_k, k = 0, \pm 1, \pm 2, \dots$  be the  $k$ -th standard basis element of  $\ell^2(\mathbb{Z})$ . Define a linear transformation  $T : \ell^2(\mathbb{N}) \rightarrow \ell^2(\mathbb{Z})$  by  $T(e_k) = e'_{k/2}$  if  $k$  is even and  $T(e_k) = e'_{(1-k)/2}$  if  $k$  is odd and extend to all of  $\ell^2(\mathbb{N})$  by making  $T$  linear. Then every  $e_k$  gets mapped to precisely one  $e'_\nu$  and every  $e'_\nu$  is the image of exactly one  $e_k$ . The inverse of  $T$  maps  $e'_\nu$  to  $e_{2\nu}$  if  $\nu = 1, 2, \dots$  and  $e'_\nu$  to  $e_{1-2\nu}$  if  $\nu = 0, -1, -2, \dots$ .

In an abstract sense  $\ell^2(\mathbb{N})$  and  $\ell^2(\mathbb{Z})$  are the same mathematical object. As we just noted, though, it is easier to represent certain concrete operations on  $\ell^2(\mathbb{Z})$ .

*Example 1.4.13.* (A function space isomorphic to  $\ell^2(\mathbb{Z})$ ). Now we consider a less innocent example. Define  $\mathcal{PH}^2(\mathbb{T})$  to consist of all periodic complex-valued functions on the real line having period one (so they can really be thought of as functions on  $[0, 1)$ ) that can be expressed as a finite linear combination of the functions  $E_k(t) = e^{2\pi ikt}$ . By definition,  $e^{i\theta} = \cos \theta + i \sin \theta$  and the law of exponents applies to complex exponents, that is,  $e^z e^w = e^{z+w}$  for any complex numbers  $z, w$ . In addition, the complex conjugate of  $e^{i\theta}$  is  $\cos \theta - i \sin \theta = \cos(-\theta) + i \sin(-\theta) = e^{-i\theta}$ . Therefore, setting

$$\langle f, g \rangle = \langle f, g \rangle_{\mathcal{PH}^2(\mathbb{T})} = \int_0^1 f(t) \bar{g}(t) dt$$

it follows that

$$\langle E_k, E_\ell \rangle = \int_0^1 e^{2\pi ikt} e^{-2\pi i\ell t} dt = \int_0^1 e^{2\pi i(k-\ell)t} dt = \delta_{k\ell}$$

since, if  $k = \ell$  then  $e^{2\pi i(k-\ell)t} \equiv 1$  while if  $k - \ell = n \neq 0$  then

$$\int_0^1 \cos(2\pi int) dt = 0 = \int_0^1 \sin(2\pi int) dt.$$

Now we formally define  $\mathcal{H}^2(\mathbb{T})$  to consist of all infinite sums of the form  $f(t) = \sum_{k=-\infty}^{\infty} c_k E_k(t)$  with the proviso that  $\sum_{k=-\infty}^{\infty} |c_k|^2 < \infty$ . There is a nontrivial question of what it means for the series to converge to  $f$ . For example, if we take  $t = 0$  we get  $f(0) = \sum_{k=-\infty}^{\infty} c_k$  but this sum does not necessarily converge. For example, if  $c_k = 1/k$  when  $k = 1, 2, \dots$  and  $c_k = 0$  otherwise then  $f(0)$  is a divergence harmonic series even though  $\sum_{k=-\infty}^{\infty} |c_k|^2 = \sum_{k=1}^{\infty} 1/k^2$  converges. However, the condition that  $\sum_{k=-\infty}^{\infty} |c_k|^2 < \infty$  implies that the partial sums  $f_M = \sum_{k=-M}^M c_k E_k(t)$  form a Cauchy sequence in the norm on  $\mathcal{H}^2(\mathbb{T})$ . To see this, we have for  $M_1 < M_2$  that

$$\begin{aligned} \|f_{M_1} - f_{M_2}\|^2 &= \langle f_{M_1} - f_{M_2}, f_{M_1} - f_{M_2} \rangle = \sum_{M_1 < |k|, |\ell| \leq M_2} c_k \bar{c}_\ell \int_0^1 E_k \bar{E}_\ell \\ &= \sum_{M_1 < |k|, |\ell| \leq M_2} c_k \bar{c}_\ell \delta_{k\ell} = \sum_{M_1 < |k| \leq M_2} |c_k|^2 < \varepsilon^2 \end{aligned}$$

provided that  $M_1$  is large enough. If we can make sense out of the limit of this Cauchy sequence then  $f$  is well-defined so the elements of  $\mathcal{H}(\mathbb{T})$  make sense.

What do we mean when we say that  $\mathcal{H}(\mathbb{T})$  is isomorphic to  $\ell^2(\mathbb{Z})$ ? Basically we mean that there is a one-to-one and onto mapping from  $\mathcal{H}(\mathbb{T})$  to  $\ell^2(\mathbb{Z})$  that preserves inner products. This is clear since if  $f, g$  are elements of  $\mathcal{H}(\mathbb{T})$  then we can write  $f = \sum_{k=-\infty}^{\infty} c_k E_k(t)$  and  $g = \sum_{k=-\infty}^{\infty} d_k E_k(t)$  where  $\mathbf{c} = \{c_k\}$  and  $\mathbf{d} = \{d_k\}$  define sequences in  $\ell^2(\mathbb{Z})$ .

The fact that this linear mapping preserves inner products is just the statement that

$$\langle f, g \rangle_{\mathcal{H}^2} = \int_0^1 \left( \sum_{k \in \mathbb{Z}} c_k E_k(t) \right) \left( \sum_{\ell \in \mathbb{Z}} \bar{d}_\ell \bar{E}_\ell(t) \right) dt = \sum_{k \in \mathbb{Z}} c_k \bar{d}_k = \langle \mathbf{c}, \mathbf{d} \rangle_{\ell^2(\mathbb{Z})}$$

which follows from the orthogonality of the functions  $E_k$  with respect to the integral over  $[0, 1)$ .

### 1.4.7 More on Isomorphism

Although we have just given one example,  $\mathcal{H}^2$ , of a Hilbert function space that is isomorphic to  $\ell^2(\mathbb{Z})$ , this example is typical in the way that other separable Hilbert spaces are isomorphic to  $\ell^2(\mathbb{Z})$ . But, perhaps more importantly, the space  $\mathcal{H}^2$  is not as innocent as it looks.

This is because  $\ell^2(\mathbb{Z})$  is closed under taking limits and this property is preserved under isomorphism. So it is natural to ask: which functions can be expressed as limits of sequences of trigonometric polynomials (where a trigonometric polynomial is a finite linear combination of the functions  $E_k$ ).

In fact,  $\mathcal{H}^2$  is the same as the space of all one-periodic functions such that

$$\int_0^1 |f(t)|^2 dt$$

exists. Such a function is said to be square integrable over  $[0, 1)$  and the space of all such functions is denoted by  $L^2(\mathbb{T})$ . Here,  $\mathbb{T}$  stands for the unit circle or *torus*. Mathematically, it is the additive group on  $[0, 1)$  with addition taken modulo one. Examples of functions in  $L^2(\mathbb{T})$  include the periodic extensions of functions  $t^{-\alpha} \mathbb{1}_{[0,1)}$  where  $\mathbb{1}_S$  denotes the function equal to one on  $S$  and zero elsewhere. Also included are more complicated functions like  $\sum_{k=1}^{\infty} c_k \mathbb{1}_{A_k}$  such that  $\sum_k |c_k|^2 |A_k|$  converges where  $A_k$  are pairwise disjoint intervals of length  $|A_k|$ . As an example, take  $A_k = [1/(2k), 1/k)$  and  $c_k = 1/\sqrt{k}$ .

A fair amount of technical detail goes into proving that any function in  $L^2(\mathbb{T})$  can be represented by its Fourier series with convergence in  $L^2(\mathbb{T})$ . The major steps of the proof of this fact will be provided in Chapter ??.

### 1.4.8 A nonseparable Hilbert space

Consider the space  $B^2(\mathbb{R})$  consisting of all complex-valued functions defined on the real line and having the property that

$$\lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |f(t)|^2 dt$$

exists and is finite. The functions  $e^{2\pi i \xi}$ , as  $\xi$  ranges over all real numbers, form an orthonormal family with respect to the inner product

$$\langle f, g \rangle = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T f(t) \bar{g}(t) dt.$$

**Exercise** Check this.

However, there is no countable basis for this space since the functions  $e^{2\pi i \xi}$  form an uncountable linearly independent set in  $B^2$ . Therefore  $B^2$  is a nonseparable inner product space. Nonseparability has a concrete meaning here since elements of  $B^2$  can be expressed in the form

$$f(t) = \lim_{S \rightarrow \infty} \frac{1}{2S} \int_{\xi=-S}^S c(\xi) e^{2\pi i t \xi} d\xi$$

However, no single term of this integral expansion can serve to distinguish between two different elements of  $B^2$ . While the space  $B^2$  might appear as an intellectual curiosity, it actually serves as a good model context for *almost periodic functions* such as measured intensities of binary stars.

## 1.5 Limits as $N \rightarrow \infty$ : From Hilbert space to the real world and back

*Sampling and Errors.*

In the *analogue* world, signals or functions vary with respect to *continuous* space or time parameters. In the digital world, quantities depend on discrete parameters. *Sampling* is the mechanism for translating analogue signals to discrete signals. In certain cases it is possible to recover the analog signal entirely from the discrete signal, provided one has some prior information on the analogue signal. Particular circuits called *sample and hold* circuits do the conversion from voltages to numbers. There are errors both in the timing of the circuits,

called jitter errors, and the numerical precision of the samples due to limited precision of sampling devices, to random fluctuations, and to limited precision of circuitry for converting a voltage to a number. Errors due to this last cause are called *quantization errors*. All of these types of errors are often regarded as sources of additive noise in the samples. We will not focus on all of these random aspects at this point but we will keep them in mind as fundamental limitations on precision when converting between analog and digital signals.

*Interpolation.*

Conversely, the method for passing from discrete to analog data is called *interpolation*. There are infinitely many ways to choose a function  $f$  that has prescribed values  $f(s_k)$  at given discrete sample points  $s_k$ . An appropriate choice of interpolant depends on context. For example, how many derivatives one would like the interpolant to have. Mechanically, there are circuits called waveform generators that pass from discrete to continuous signals.

Modern technology allows for very high sampling rates – often higher than simple, inexpensive processors can handle. So issues of *scalability* come into play. How much information is gained when one throws more processing power at a signal?

### 1.5.1 Linear interpolation from discrete data

If one collects infinitely many samples (both forward and backward in time), all separated by the same amount of time  $\Delta t$ , then one obtains a sequence of values  $\{f(t_0 + k\Delta t)\}_{k \in \mathbb{Z}}$ . Really, one cannot collect an infinite number of samples but, in principle, one can collect an arbitrarily large finite number. Therefore it is reasonable to take this data to lie in  $\ell^2(\mathbb{Z})$ . Under what conditions does there exist a unique signal  $f$ , having some prescribed properties, that interpolates these samples? For simplicity suppose that  $t_0 = 0$  and  $\Delta t = 1$ . The function

$$f_{\text{lin}}(t) = \sum_{k \in \mathbb{Z}} ((k+1-t)f(k) + (t-k)f(k+1)) \mathbb{1}_{[k, k+1)}(t)$$

defines a function that is linear on  $[k, k+1)$  for each  $k$  and has the desired sample values. However,  $f_{\text{lin}}$  does not have a continuous derivative since its slope on  $[k, k+1)$  is  $f(k+1) - f(k)$  and this value will change, generally, as we pass from  $k$  to  $k+1$ . On the other hand,  $f_{\text{lin}}$  lies in the space of square integrable functions, that is,

$$\begin{aligned} \int_{-\infty}^{\infty} |f_{\text{lin}}(t)|^2 dt &= \sum_{k=-\infty}^{\infty} \int_k^{k+1} |f_{\text{lin}}(t)|^2 dt \\ &= \sum_{k=-\infty}^{\infty} \int_k^{k+1} |((k+1-t)f(k) + (t-k)f(k+1))|^2 dt \\ &= \sum_{k=-\infty}^{\infty} \int_0^1 |((1-u)f(k) + uf(k+1))|^2 du \\ &= \frac{1}{3} \sum_{k=-\infty}^{\infty} (f(k)^2 + f(k+1)^2 + f(k)f(k+1)) \leq \sum_{k=-\infty}^{\infty} f(k)^2 < \infty \end{aligned}$$

with the line following from the Cauchy-Schwarz inequality and the assumption that  $\{f(k)\} \in \ell^2(\mathbb{Z})$ . Later we will see other possible ways to interpolate a signal  $f$  from the samples.

### 1.5.2 Interpolation errors.

Interpolation produces a *guess*  $f_{\text{int}}$  of the function  $f$  from which the given samples  $f(t_k)$  were produced looks like. There will always be some error between  $f$  and  $f_{\text{int}}$  which can, hopefully, be minimized with some priori knowledge of the nature of  $f$ , as will be discussed presently and in subsequent chapters. If the samples  $f(t_k)$  are known exactly then the interpolation error  $e_{\text{int}}(t) = |f(t) - f_{\text{int}}(t)|$  will vanish at the sample points  $t_k$ .

### 1.5.3 Filling in values

The simplest way to reduce interpolation errors, in principle if not in practice, is to sample at a very high rate. Rather than considering a function defined on all of  $\mathbb{R}$  and sampled at a rate of one sample per unit time, consider the complementary problem of a signal that *lives on*  $[0, 1)$  and for which one has samples at a rate of  $N = 2^L$  samples. In this case, if  $f$  is continuous then the linear interpolants will converge to  $f$  as  $L \rightarrow \infty$ . Suppose now that one only needs to recover  $f$  with finite precision. For example, suppose that one is satisfied to know  $f$  within an error of  $\varepsilon > 0$ . If the signal  $f$  is assumed to be uniformly continuous then there will be some  $\delta > 0$  such that if  $L$  is large enough then, when  $t \in [k/2^L, (k+1)/2^L)$  one will automatically have  $|f(t) - f(k/2^L)| < \varepsilon$ . How large one has to take  $L$  in terms of  $\varepsilon$  depends on something called the *modulus of continuity*, denoted  $\omega(f, \delta; t) = \sup\{|f(t) - f(s)| : |t - s| < \delta\}$ .

This modulus can be pictured in terms of an *envelope* function above and below the graph of  $f$ , see Figure 1.2. The envelope is formed by taking the highest and lowest values of  $f$  in the subinterval  $[t - \delta, t + \delta]$ . If  $f$  is (uniformly) continuous then  $\omega(f, \delta; t) \rightarrow 0$  (uniformly) as  $\delta \rightarrow 0$ . In the figure, the height of the rectangle tends to zero as  $\delta \rightarrow 0$ .

If one assumes that  $f$  is better than merely continuous then one can quantify the scale  $L$  needed to get an error within  $\varepsilon$ . For example, one says that  $f$  is *Hölder continuous of order*  $\alpha > 0$  on  $[0, 1]$  if there is a constant  $C > 0$ , called the Hölder constant, such that whenever  $s, t \in [0, 1)$  one has

$$\frac{|f(s) - f(t)|}{(s - t)^\alpha} \leq C. \quad (1.2)$$

This tells us that if the width of the box is  $\delta$  then the height of the envelope box is at most  $C\delta^\alpha$ . Therefore, in order to achieve an interpolation error of at most  $\varepsilon$  we just need to take  $L = -\log_2(\delta)$  such that  $C\delta^\alpha < \varepsilon$ , or  $L > \log_2(C/\varepsilon)/\alpha$ .

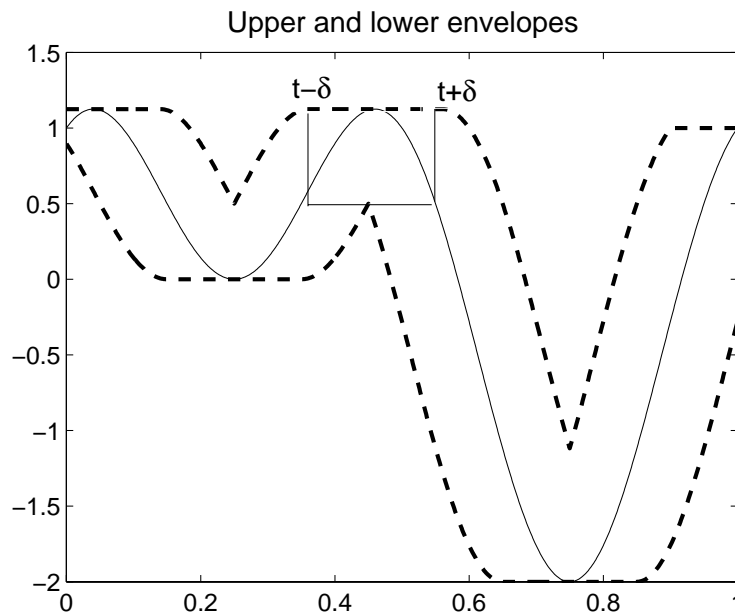


Fig. 1.2. Upper and lower envelopes

### 1.5.4 Quantization

The error estimates just made were based on the premise that the samples  $f(t_k)$  generated by the signal  $f$  are known exactly. Then, by sampling at a high enough rate, one could find an interpolant with an arbitrarily

small error. In reality samples are never known exactly for reasons outlined before and because computer arithmetic has finite precision. Actually these limitations tell us how to fix  $\epsilon$  because we can do no better. Perhaps the most important type of error that we will need to understand is that due to *quantization*.

As we saw in Section 1.1.1, any whole number  $M$  has a *binary* expansion  $M = \sum_{k=0}^{N-1} \delta_k 2^k$  where  $\delta_k \in \{0, 1\}$ . One also writes  $M \sim \delta_1 \delta_2 \dots \delta_{N-1}$  as the *binary representation* of  $M$ .

Also, any  $t \in (0, 1)$  has a possibly nonterminating binary expansion  $t \sim \sum_{k=1}^{\infty} \epsilon_k / 2^k$  or  $t = 0.\epsilon_1 \epsilon_2 \dots$ . By convention,  $0.111\dots = 1$ , expressing the fact that the geometric series  $\sum_{k=1}^{\infty} 2^{-k}$  sums to one. Altogether, any positive number can be expressed as  $x = M + t \sim \delta_1 \delta_2 \dots \delta_{N-1}.\epsilon_1 \epsilon_2 \dots$  and, again, we call this the binary representation of  $x$ . We can add numbers by adding their binary representations, such that if both numbers have a ‘one’ in the  $k$ th slot then we replace that by a zero and add a one in the  $k - 1$ -st slot. The decimal point can be shifted as a result so we call this type of arithmetic *floating point*.

Computers have finite precision, meaning they can only store floating point numbers with a fixed amount of digits (typically 32 or 64) and have to deal with larger numbers in some other clever way. Some large numbers simply cannot be handled. For example, Matlab will return `Inf` when asked to compute 200 factorial. Additionally, several types of data are formatted so only a much smaller degree of precision is used. For example, *grayscale* intensity images – digital counterparts of black and white – have pixel values that are typically stored in `uint8`. This means strings of zeros and ones of length 8 (one byte) corresponding to gray level intensities between zero (black) and 255 (white). All other intensities are rounded to the nearest `uint8`. This rounding is called *quantization* and the rounding errors that result are referred to as *quantization errors*. In many applications these errors are modelled as noise but, in reality, these errors often have a coherent structure and we will discuss the issue of error distributions in more detail later.

Specifying a quantization level at least allows one to specify the amount of storage an image takes in computer memory. For example, if no type of compression is used then a grayscale image of  $256 \times 256$  pixels will require  $2^{16} = 65536$  bytes (about 65 Kb) of storage. A color image uses 1 byte for each of the rgb (red, green, blue) pixel values so it should require about 3 times that. However, if you look at a typical `jpg` format image, you will find that it uses much less – maybe one tenth of the estimated storage requirements. This is because the image is *coded* in such a way that redundant information requires less storage.

**Exercise 1.5.1.** Use Matlab or some other program to compute  $k!$  for  $k = 1, \dots, 40$  or higher, and plot  $\log(k!)$  versus  $k$ . What is the largest integer that your program can handle?

### 1.5.5 Transformation coding

In a sense, transformation coding is what the remainder of this text is all about. There are many ways to encode data such that that data can be recovered from its coded version. By a transformation here we will mean a linear transformation on the vector space in which the given data lies, although nonlinear transformations could also be used.

Any invertible linear transformation, then, is a code. But what comprises a *good* code? As is always the case in these matters, it depends on context. Consider for example the problem of data compression. Suppose that one has a vector  $\mathbf{z} \in \mathbb{C}^N$ . On the one hand the vector contains  $N$  coordinates, so  $N$  pieces of information. When dealing with discrete values we also have to count how many bits, say  $B$ , per coefficient. So totally  $NB$  bits. Suppose now that the data happens to be samples of a pure sinusoid. In this case, the DFT will convert  $\mathbf{z}$  into a new vector  $\mathbf{w}$  that has only one nonzero coefficient, or possibly, due to quantization error, one larger and several very small coefficients. Then, at least up to small error, only  $B$  bits are required to preserve the information since, up to a small error, one can recover  $\mathbf{z}$  by taking the inverse *DFT* of  $\mathbf{w}$ .

This illustrates the fact that transform coding can be very efficient for compressing data, at least in very particular cases. Here is a more general problem. Suppose that we have data that can be assumed to belong to a certain class of data subject to certain rules. For example, the data might be samples of a solution to a differential equation, or might lie along a certain type of curve, etcetera. Then one asks: is there a transformation that compresses or, more generally, efficiently represents the data in that class? What we are looking for are ways to quantify this in mathematical terms that capture *goodness* in measurable ways.

## 1.6 Appendix: $\ell^2(\mathbb{Z}_N)$ and finite Fourier series

We end this chapter with a slight digression on the Hilbert space  $\mathbb{C}^N$  which can also be identified with the collection of all functions defined on the finite group  $\mathbb{Z}_N$ , the *integers modulo N*. If  $0 \leq k, \ell < N$  then the sum

$(k + \ell) \bmod N$  is defined simply as  $k + \ell$  if  $k + \ell < N$  and as  $(k + \ell) - N$  if  $k + \ell \geq N$ . We will just write  $k + \ell$  if the “modulo  $N$ ” is implied. Set  $\omega = \omega_N = e^{2\pi i/N}$ . Then  $\omega$  is called a primitive  $N$ -th root of unity since its  $N$ -th power is one. The group  $\mathbb{Z}_N$  can then be identified with the powers  $\omega^k$  which are  $N$  equally spaced points on the unit circle in the complex plane. Now consider the vector  $\mathbf{w} = (1, \omega, \omega^2, \dots, \omega^{N-1}) \in \mathbb{C}^N$  and denote  $\mathbf{w}^j = (1, \omega^j, \omega^{2j}, \dots, \omega^{(N-1)j}) \in \mathbb{C}^N$ .

**Lemma 1.6.1.** *The vectors  $\mathbf{w}^j$ ,  $j = 0, 1, \dots, N - 1$  form an orthogonal family of vectors, each with norm  $N$  in  $\mathbb{C}^N$ .*

*Proof.* To prove this, first we note that

$$\|\mathbf{w}^j\|^2 = \langle \mathbf{w}^j, \mathbf{w}^j \rangle = \sum_{\nu=0}^{N-1} \omega^{\nu j} \bar{\omega}^{\nu j} = \sum_{\nu=0}^{N-1} \omega^{\nu j} \omega^{-\nu j} = \sum_{\nu=0}^{N-1} \omega^0 = \sum_{\nu=0}^{N-1} 1 = N.$$

If  $j \neq k$  then

$$\langle \mathbf{w}^j, \mathbf{w}^k \rangle = \sum_{\nu=0}^{N-1} \omega^{\nu j} \bar{\omega}^{\nu k} = \sum_{\nu=0}^{N-1} \omega^{\nu(j-k)} = \sum_{\nu=0}^{N-1} (\omega^{j-k})^\nu$$

Multiplying both sides by  $\omega^{j-k}$  one has

$$\omega^{j-k} \langle \mathbf{w}^j, \mathbf{w}^k \rangle = \omega^{j-k} \sum_{\nu=0}^{N-1} (\omega^{j-k})^\nu = \sum_{\nu=0}^{N-1} (\omega^{j-k})^{1+\nu} = \sum_{\nu=0}^{N-1} (\omega^{j-k})^\nu$$

since  $\omega^{(j-k)N} = \omega^0$  so both sides of the last equation contain the sum of all powers of  $\omega^{j-k}$  between 0 and  $N - 1$ . It follows that  $\langle \mathbf{w}^j, \mathbf{w}^k \rangle = \omega^{j-k} \langle \mathbf{w}^j, \mathbf{w}^k \rangle$ . Since  $\omega^{j-k} \notin \{0, 1\}$  it must be that  $\langle \mathbf{w}^j, \mathbf{w}^k \rangle = 0$  and this proves the lemma.

If we divide by  $\sqrt{N}$  it follows that the vectors  $\mathbf{w}^j/\sqrt{N}$  form an orthonormal basis for  $\mathbb{C}^N$  called the *discrete Fourier basis*. Because of its analogy with  $\ell^2(\mathbb{Z})$  we can write  $\mathbb{C}^N \simeq \ell^2(\mathbb{Z}_N)$ . The transformation that takes a vector in  $\ell^2(\mathbb{Z}_N)$  expressed in standard coordinates to its coefficients with respect to the vectors  $\mathbf{w}^j/\sqrt{N}$  is called the *discrete Fourier transform* or DFT.