

The projection theorem in control and signal processing

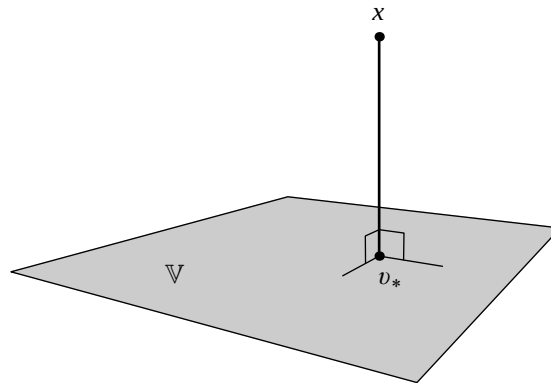
Gjerrit Meinsma

University of Twente

<http://www.math.utwente.nl/>

Current version: July 2, 2007

These notes were written while on sabbatical leave at the Faculty of Mechanical Engineering, Technion-IIT, Haifa. This sabbatical leave is supported by the Lady Davis Fellowship Trust.



Preface

The *projection theorem* in $\mathbb{R}^3$ is intuitive: the point v_* in a plane $\mathbb{V}$ that is closest to some exterior point x is precisely the one for which $x - v_*$ is perpendicular to this plane, see the above figure. The purpose of this course is to formalize and generalize the notions of “plane” and “perpendicular” and “closest” etc. in such a way that these notions become available in many other situations as well. The second part of this course is a collection of worked examples to demonstrate how powerful the general projection theorem is. Most examples are drawn from Systems and Control Theory and Signal Processing.

I assume the reader has some familiarity with MATLAB and some basic knowledge of Calculus and Linear Algebra and here and there (but not too often I hope) I will need some Statistical notions like expectation and variance.

I wrote a large part of these notes especially for this course (in Haifa) and though I try my best (well) it will certainly not be ideal. One part that worries me a bit is the lack of Topology (open and closed sets, Cauchy-sequence, complete spaces, etc.) in this course. I do this on purpose because it would require too much math, but it comes at some cost. If you want to know more about these matters and its beautiful applications I suggest these books:

Young (1988) A beautifully written introduction to Hilbert space. It is a math book.

Luenberger (1969) Old but great and well written book. Also covers optimization.

Saxe (2002) A first course on functional analysis.

Mallat (1998) A very good and complete book on Wavelets and other Signal processing tools. A tough read though.

Hayes (1996) An excellent first course on statistical signal processing.

Zhou et al. (1996) A standard book on optimal control of finite dimensional systems. It takes a linear algebra approach to control.

And of course there is wikipedia.

Contents

1	Intro: elementary geometry	1
1.1	Perpendicular vectors in $\mathbb{R}^2$	1
1.2	Perpendicular vectors in $\mathbb{R}^3$	1
1.3	Length in $\mathbb{R}^2$ and $\mathbb{R}^3$	2
1.4	Angle	3
1.5	Problems	4
2	The dot product on $\mathbb{R}^n$	5
2.1	The dot product and orthogonality	5
2.2	Euclidean norm	5
2.3	Pythagoras	5
2.4	Orthogonal complement in $\mathbb{R}^n$	6
2.5	Subspaces of $\mathbb{R}^n$ and im and ker	6
2.6	Projection	8
2.7	Transpose	9
2.8	Explicit projection: matrices and the normal equations	10
2.9	Under-determined equations	11
2.10	Orthogonal matrices and QR-factorization	11
2.11	If you want to know more...	12
2.12	Problems	13
3	Vector space and subspace	15
3.1	Real vector space	15
3.2	Complex vector space	17
3.3	Subspace	18
3.4	Linear combination, span, basis and dimension	19
3.5	If you want to know more...	21
3.6	Problems	21
4	Linear transformation – operator	23
4.1	Linear transformation	23
4.2	Familiar linear transformations	23
4.3	Linear transformation on $\mathbb{R}^n$	25
4.4	Matrices as representations of operators	26
4.5	If you want to know more	29
4.6	Problems	29
5	Inner product, norm and orthogonal projection	33
5.1	Real inner product	33
5.2	Complex inner products	33
5.3	Norm	34
5.4	Orthogonal complement	34
5.5	Pythagoras	35
5.6	Cauchy-Schwarz	35
5.7	Examples of inner product	36
5.8	Inner product on function space: $\mathbb{L}^2$	37
5.9	Vector valued functions	39
5.10	Orthogonal projection	39

5.11	Projection onto finite dimensional space – Gram matrix	40
5.12	Orthonormal sequences and Parseval	41
5.13	Adjoint operator	42
5.14	Normal equations revisited	44
5.15	Under determined equations	45
5.16	Gram-Schmidt and the matrix of inner products	46
5.17	Hilbert space	46
5.18	If you want to know more...	48
5.19	Problems	48
6	Fourier	51
6.1	DFT — discrete to discrete Fourier transform	51
6.2	Harmonic decomposition — the DFT as a change of basis	52
6.3	Extra: the FFT	53
6.4	Fourier — continuous to discrete	55
6.5	Fourier transform — discrete to continuous (DTFT)	57
6.6	Fourier transform — continuous to continuous	58
6.7	Bilateral Laplace transform of $\mathbb{L}^2(\mathbb{R})$ -signals	60
6.8	Stable systems, causal or noncausal	62
6.9	LTI transformations, adjoint and conjugate	62
6.10	Extra: some technical details...	64
6.11	Problems	65
7	An Introduction to stochastic processes	67
7.1	Expectation, variance and covariance	67
7.2	Inner product and the covariance matrix	68
7.3	Orthogonal — uncorrelated	69
7.4	Norm — standard deviation	69
7.5	Cauchy-Schwarz — correlation	70
7.6	Pythagoras — variance of uncorrelated variables	70
7.7	Projection — optimum linear estimator	71
7.8	Normal equations	71
7.9	Stochastic processes	72
7.10	Wide-sense stationary stochastic processes	73
7.11	White noise	75
7.12	Power spectral density, Parseval	76
7.13	LTI systems, power and the ℓ^2 -norm	78
7.14	Models for spectral densities	80
7.15	WSS processes are responses of LTI systems driven by white noise	81
7.16	Extra: Continuous-time stochastic processes — simplistic overview	82
7.17	IYWTKM	84
7.18	Problems	85
8	Least squares	87
8.1	Approximating data pairs by straight lines	87
8.2	Approximating data pairs by a linear combination of functions	89
8.3	Approximate functions by other functions	90
8.4	Stochastic – linear estimators	91
8.5	If you want to know more	92
8.6	Problems	93

9	AR-models, Linear predictive coding	95
9.1	Stable AR systems	96
9.2	Least squares for AR systems, part 1	96
9.3	Least squares for AR systems, part 2: Toeplitz	96
9.4	Problems	99
10	Finite horizon LQ control	101
10.1	Control the state	101
10.2	A first solution — controllability Gramian	101
10.3	More explicit solution — matrix exponential	102
10.4	Minimal control input – embellishments	103
10.5	Minimal control input – state representation of adjoint	104
10.6	Finite horizon LQ	105
10.7	Solution of finite horizon LQ	105
10.8	IYWTKM...	108
10.9	Problems	108
11	$\mathbb{H}^2$ optimization in frequency domain	111
11.1	LQ in frequency domain	111
11.2	SISO LQ and return difference inequality	111
11.3	MIMO systems – spectral factorization	114
11.4	MIMO LQ and the return difference inequality	119
11.5	SISO Filtering	121
11.6	MIMO filtering	124
11.7	Extra: Standard $\mathbb{H}^2$ problem	125
11.8	Extra: Riccati equation	126
11.9	IYWTKM.. . . .	127
11.10	Problems	128
12	Wiener filtering and the matched filter	129
12.1	Constant filter	129
12.2	FIR filter — matrix form of Wiener-Hopf equations	130
12.3	Classic Wiener Filtering	131
12.4	Noise cancellation	132
12.5	Linear prediction	133
12.6	IIR Wiener filter — causal linear prediction	135
12.7	Standard biased estimate of covariances — least squares	137
12.8	Matched filter and the signal-to-noise ratio	138
12.9	Matched filters for noisy signals	139
12.10	Problems	140
13	JPEG and 2D-Fourier	143
13.1	Images as matrices	143
13.2	Introduction: bases for 2×2 images	143
13.3	Bases functions for images (2D-signals)	144
13.4	Two-dimensional Fourier series	145
13.5	Cosine transform	145
13.6	Discrete two-dimensional cosine transform	146
13.7	JPEG	147

14	A very brief introduction to wavelets	151
14.1	Multiresolution approximation and detail—Haar	151
14.2	Orthogonal Haar basis	152
14.3	Down and up-sampling	154
14.4	Haar basis (continuous time)	154
14.5	Generalizations	155
14.6	Some more snippets	156
A	Matrices and Matlab	159
A.1	Matrices and vectors	159
A.2	Positive definite matrices	160
A.3	Trace	161
A.4	Matlab	161
B	Bibliography	163
	Index	165

1 Intro: elementary geometry

Some intuitive ideas from elementary geometry are useful for what eventually will become a full fledged projection theorem.

1.1 Perpendicular vectors in $\mathbb{R}^2$

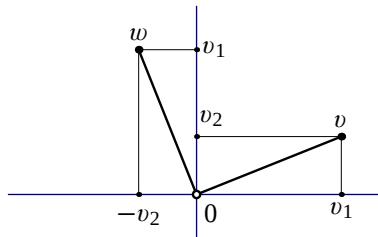


Figure 1.1: Perpendicular vectors

Consider the vectors v and w shown in Fig. 1.1, that is (with respect to the standard basis) that

$$v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix} \quad \text{and} \quad w = \begin{bmatrix} w_1 \\ w_2 \end{bmatrix} = \begin{bmatrix} -v_2 \\ v_1 \end{bmatrix}.$$

Wouldn't you say that these two vectors are perpendicular? You probably would. Convince yourself of the fact that two vectors $v, w \in \mathbb{R}^2$ are perpendicular iff

$$v_1 w_1 + v_2 w_2 = 0.$$

Incidentally, we identify elements of $\mathbb{R}^2$ with the *column* vector of its entries, such as $v = \begin{bmatrix} v_1 \\ v_2 \end{bmatrix}$ but once in while, for notational convenience, we write $v = (v_1, v_2)$. Similar conventions apply to $\mathbb{R}^3$ and higher dimensions. To avoid clutter in figures, vectors on axes are labeled by the appropriate coordinate only (i.e. we label the vector $(v_1, 0)$ simply by the number v_1 , see Fig. 1.1).

1.2 Perpendicular vectors in $\mathbb{R}^3$

A way to intuitively construct vectors that are perpendicular to a given vector $v = (a, b, c) \in \mathbb{R}^3$ is shown in Fig. 1.2. The construction goes as follows: the vector v is an element of a plane that contains the z -axis and the vector $s := (a, b, 0)$, i.e., the grey-colored plane in Fig. 1.2. Now $w^z := (b, -a, 0)$ is perpendicular to the z -axis – since w^z is in the (x, y) -plane – and it is perpendicular to $s = (a, b, 0)$ according to the previous section. Hence w^z is perpendicular to the entire grey-plane of Fig. 1.2 and therefore in particular is perpendicular to v . In summary:

$$w^z := \begin{bmatrix} b \\ -a \\ 0 \end{bmatrix} \text{ is perpendicular to } v = \begin{bmatrix} a \\ b \\ c \end{bmatrix}.$$

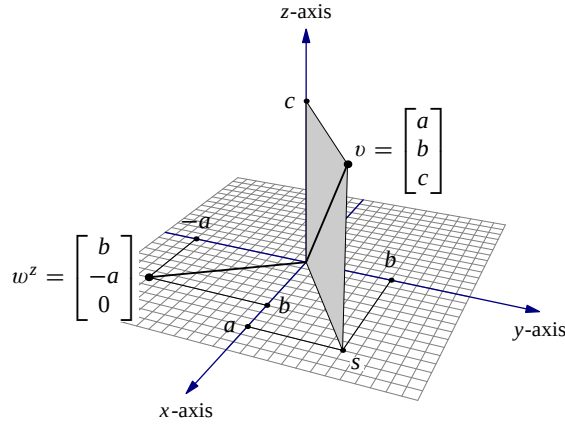


Figure 1.2: Since w is perpendicular to the z -axis and vector s it is perpendicular to the entire gray plane, including v

In terms of the coordinates: take $v = (a, b, c)$, set its last coordinate to zero, $(a, b, 0)$, then interchange the first two and take one of them with a minus sign, $(b, -a, 0)$. Then you end up with a vector that is perpendicular to the v you started with. Obviously we can interchange the order of the coordinates so all of these three vectors are perpendicular to (a, b, c) :

$$w^z = \begin{bmatrix} b \\ -a \\ 0 \end{bmatrix}, \quad w^y = \begin{bmatrix} -c \\ 0 \\ a \end{bmatrix}, \quad w^x = \begin{bmatrix} 0 \\ c \\ -b \end{bmatrix}. \quad (1.1)$$

Combined the three vectors span the plane of vectors

$$\{ w \in \mathbb{R}^3 \mid w = \lambda w^x + \mu w^y + \eta w^z, \quad (\mu, \lambda, \eta \in \mathbb{R}) \}$$

that are perpendicular to v . (In fact two of them are enough to span this plane.) Now each of the three vectors w^z, w^y, w^x have the property that

$$v_1 w_1 + v_2 w_2 + v_3 w_3 = 0. \quad (1.2)$$

and it is now easy verify that a vector w is perpendicular to v if and only if (1.2) holds.

1.3 Length in $\mathbb{R}^2$ and $\mathbb{R}^3$

The Pythagorean theorem says that in a right-angled triangle with sides of length a, b and c the longest side c satisfies $c^2 = a^2 + b^2$. Figure 1.3 is a *proof by picture* of the Pythagorean theorem.

Therefore in $\mathbb{R}^2$ the length $|v|$ of a vector $v = (v_1, v_2)$ equals

$$|v| := \sqrt{v_1^2 + v_2^2}.$$

Likewise the length of a vector $v = (v_1, v_2, v_3)$ in $\mathbb{R}^3$ equals

$$|v| = \sqrt{v_1^2 + v_2^2 + v_3^2}.$$

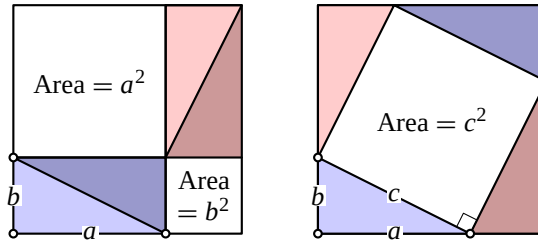


Figure 1.3: Pythagoras: $a^2 + b^2 = c^2$

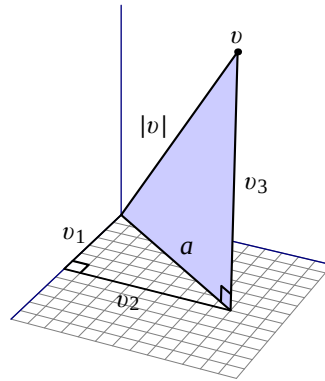


Figure 1.4: Pythagoras in $\mathbb{R}^3$

This follows from a nested application of Pythagoras: first $|v|^2 = a^2 + v_3^2$ where a is the length of the vector $(v_1, v_2, 0)$ – see Fig. 1.4 – and the length a in turn, again by ordinary Pythagoras, satisfies $a^2 = v_1^2 + v_2^2$.

1.4 Angle

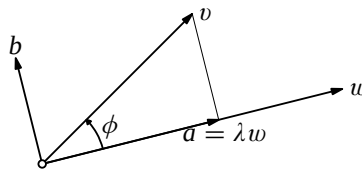


Figure 1.5: Angles

The expression $v_1w_1 + v_2w_2$ not only captures perpendicular vectors, but also angles:

Lemma 1.4.1 (Angle). For any $v, w \in \mathbb{R}^2$ we have that

$$v_1w_1 + v_2w_2 = |v||w| \cos(\phi) \quad (1.3)$$

where ϕ is the angle between the vectors v and w , see Fig. 1.5. □

Proof. It trivially holds if v or w are zero. Now suppose $v \neq 0, w \neq 0$. Write v as $v = a + b$ with $a = \lambda w$ a multiple of w and b perpendicular to w , see Fig. 1.5. Suppose the angle is acute or straight: $0 \leq \phi \leq \pi/2$. Then $\lambda \geq 0$ and

$$\begin{aligned} v_1 w_1 + v_2 w_2 &= (a_1 + b_1)w_1 + (a_2 + b_2)w_2 \\ &= a_1 w_1 + a_2 w_2 + \underbrace{b_1 w_1 + b_2 w_2}_{= 0 \text{ because perp.}} \\ &= \lambda(w_1^2 + w_2^2) = \lambda|w|^2 = |a||w| \\ &= |v| \cos(\phi) |w|. \end{aligned}$$

Here we used that $|a| = |v| \cos(\phi)$ which is immediate from Fig. 1.5. For obtuse angles, $\phi > \pi/2$, the factor λ is negative and so is $\cos(\phi)$ so that equality is maintained. ■

This cumbersome proof begs for more efficient notation.

1.5 Problems

- 1.1 *Representation free dot product.* We define the *dot product* $x \cdot y$ for $x, y \in \mathbb{R}^2$ as $x \cdot y = |x||y| \cos(\phi)$ where ϕ is the angle between the two vectors and where it is assumed that we know what length is. Show from scratch (that is, using nothing of the above) that then $x \cdot y = x_1 y_1 + x_2 y_2$.
- 1.2 *Another intuitive derivation of perpendicular vectors.* Fig. 1.6 suggests that v and w are perpendicular iff the two vectors $v + w$ and $v - w$ have equal norm. So we could have defined v to be *perpendicular* to w if $|v - w| = |v + w|$. Show that with this definition, v and w are perpendicular iff $v_1 w_1 + v_2 w_2 = 0$.

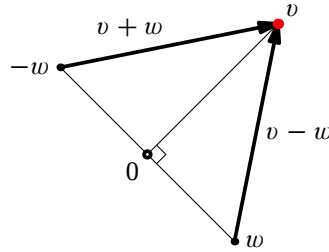


Figure 1.6: $|v + w| = |v - w|$

2 The dot product on $\mathbb{R}^n$

Now it is time for a more axiomatic approach to “perpendicular” (to be called *orthogonal*) and “length” (to be called *norm*).

The set $\mathbb{R}^n$ is the set of ordered n -tuples $(x_1, x_2, \dots, x_n)$ with $x_k \in \mathbb{R}$, $k \in \{1, 2, \dots, n\}$. Gloss through Appendix A for some standard matrix notation and results because we are going to need it.

2.1 The dot product and orthogonality

Given the previous chapter it is natural define

Definition 2.1.1 (Dot product and orthogonality in $\mathbb{R}^n$). The *dot product* $x \cdot y$ of two vectors $x, y \in \mathbb{R}^n$ is defined as

$$x \cdot y = x_1y_1 + x_2y_2 + \dots + x_ny_n$$

and we say that two vectors $x, y \in \mathbb{R}^n$ are *orthogonal* or *perpendicular* – notation $x \perp y$ – if $x \cdot y = 0$. $\square$

For $\mathbb{R}^2$ and $\mathbb{R}^3$ the dot product being zero agrees with our intuition of orthogonality but from now on we simply take $x \cdot y = 0$ to be the *definition* of orthogonality for any $\mathbb{R}^n$.

2.2 Euclidean norm

In $\mathbb{R}^3$ Pythagoras suggests that the *length* of a vector (x_1, x_2, x_3) is $\sqrt{x_1^2 + x_2^2 + x_3^2}$. This is precisely $\sqrt{x \cdot x}$ so our notion of dot product appears to encompass length already. It is customary to refer to $\sqrt{x \cdot x}$ not as a length of x but as the (*Euclidean*) *norm* of x .

Definition 2.2.1 (Euclidean norm). The *Euclidean norm* $|x|$ of $x \in \mathbb{R}^n$ is defined as

$$|x| = \sqrt{x \cdot x} = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}.$$

$\square$

2.3 Pythagoras

So forget about geometric intuition (well) and realize that orthogonality is now *defined* as having zero dot product. With that out of the way the Pythagorean theorem is trivial:

Theorem 2.3.1 (Pythagorean theorem). $x \perp y$ iff $|x + y|^2 = |x|^2 + |y|^2$.

Proof.

$$|x + y|^2 = (x + y) \cdot (x + y) \quad (2.1)$$

$$= x \cdot (x + y) + y \cdot (x + y) \quad (2.2)$$

$$= x \cdot x + 2(x \cdot y) + y \cdot y \quad (2.3)$$

$$= |x|^2 + 2(x \cdot y) + |y|^2. \quad (2.4)$$

(Please realize that the dot product has these properties.) ■

2.4 Orthogonal complement in $\mathbb{R}^n$

The *orthogonal complement* of some set $S \subset \mathbb{R}^n$ is the set of *all* vectors that are orthogonal to *all* elements of S . The orthogonal complement is denoted $S^\perp$.

Example 2.4.1 (Orthogonal complement). Consider

$$x := \begin{bmatrix} 1 \\ 3 \\ 10 \end{bmatrix} \in \mathbb{R}^3.$$

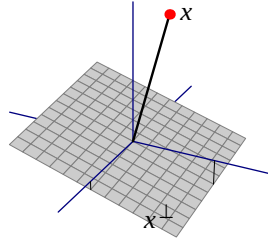
Its orthogonal complement is

$$x^\perp := \{y \in \mathbb{R}^3 \mid x \cdot y = 0\}$$

$$= \{y \in \mathbb{R}^3 \mid y_1 + 3y_2 + 10y_3 = 0\}$$

$$= \{y \in \mathbb{R}^3 \mid y_3 = -\frac{1}{10}y_1 - \frac{3}{10}y_2\}$$

$$= \left\{ \begin{bmatrix} a \\ b \\ -\frac{1}{10}a - \frac{3}{10}b \end{bmatrix} \mid a, b \in \mathbb{R} \right\} =$$



The orthogonal complement here is a two-dimensional plane. □

We write $\mathbb{V} \perp \mathbb{W}$ whenever all elements of $\mathbb{V}$ are perpendicular to all elements of $\mathbb{W}$.

2.5 Subspaces of $\mathbb{R}^n$ and im and ker

Loosely speaking a subset of $\mathbb{R}^n$ is a *subspace* if it is “flat” if it extends in all directions and if it contains the origin, see Fig. 2.1. It is not too hard to formalize subspace:

Definition 2.5.1 (Subspace). A subset $\mathbb{S}$ of $\mathbb{R}^n$ is a *subspace* if

1. $0 \in \mathbb{S}$,

2. $u, v \in \mathbb{S}$ implies $u + v \in \mathbb{S}$,

closed under addition

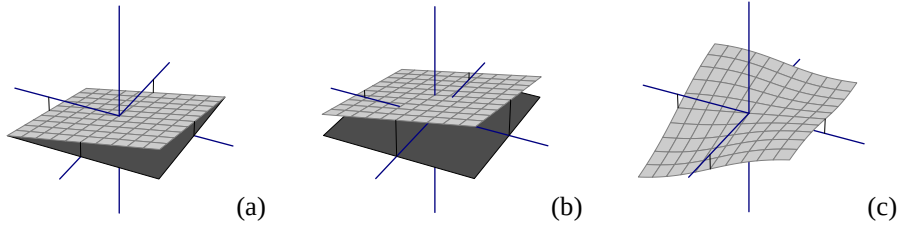


Figure 2.1: (a) subspace; (b) affine variety; (c) not a subspace

3. $u \in \mathbb{S}, \alpha \in \mathbb{R}$ implies $\alpha u \in \mathbb{S}$.

closed under scaling

□

It is customary use 0 for both the *origin* (i.e. the zero vector) and the zero *number*. The first of the three conditions seems superfluous but it is needed to guarantee that the subspace is not empty. The first part states that the origin is part of the subset, the second roughly that it is flat and the third that it extends to infinity and zero. If $\mathbb{S}$ is a subspace then $x + \mathbb{S}$ referred to as an *affine variety*, see Fig. 2.1(b).

Example 2.5.2 (Image and kernel). The set

$$\mathbb{S} := \{v \in \mathbb{R}^3 \mid v = (v_1, v_2, 0), v_1, v_2 \in \mathbb{R}\}$$

is a subspace of $\mathbb{R}^3$. It is the “(x, y)-plane”. Let us verify the three defining properties of subspace:

1. Clearly $0 = (0, 0, 0) \in \mathbb{S}$
2. If $v, w \in \mathbb{S}$ then $v = (v_1, v_2, 0)$ and $w = (w_1, w_2, 0)$. Hence $v + w = (v_1, v_2, 0) + (w_1, w_2, 0) = (v_1 + w_1, v_2 + w_2, 0)$ and since its last entry is zero also this vector is in $\mathbb{S}$.
3. If $v \in \mathbb{S}$ then $v = (v_1, v_2, 0)$ so that $\lambda v = \lambda(v_1, v_2, 0) = (\lambda v_1, \lambda v_2, 0)$ and this clearly is again an element of $\mathbb{S}$.

This subspace can be represented in many different ways:

- $\mathbb{S} = \text{im}(A) := A\mathbb{R}^2 := \{x \mid x = Ay, y \in \mathbb{R}^2\}$ for $A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \in \mathbb{R}^{3 \times 2}$.
- $\mathbb{S} = \mathbb{W}^\perp$ for the subspace $\mathbb{W} = \text{im}\left(\begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 0 \\ 1 \end{bmatrix} \mathbb{R}$.
- $\mathbb{S} = \ker(W) := \{x \in \mathbb{R}^3 \mid Wx = 0\}$ for $W = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix} \in \mathbb{R}^{1 \times 3}$.

□

We hope you appreciate the compactness of the notation $\text{im}(A) = A\mathbb{R}^k$. It is the *image* of A , i.e. the set of all possible linear combinations of the columns of the matrix A . The *kernel* $\ker(W)$ is the set (subspace) of vectors x that annihilate Wx . The following lemma states that any subspace of $\mathbb{R}^n$ can be represented by matrices, and the use of such representations are that it allows us to translate properties of subspaces etcetera into properties of matrices, making it explicit, i.e. “Matlabable”.

Lemma 2.5.3 (Matrix representation of subspace). Let $\mathbb{S}$ be a subset of $\mathbb{R}^n$. The following statements are equivalent:

- $\mathbb{S}$ is a subspace
- $\mathbb{S} = \text{im}(A)$ for some matrix $A \in \mathbb{R}^{n \times k}$ and some $k \in \mathbb{N}$
- $\mathbb{S} = \mathbb{W}^\perp$ for some set $\mathbb{W} \in \mathbb{R}^n$
- $\mathbb{S} = \ker(W)$ for some $W \in \mathbb{R}^{m \times n}$ and some $m \in \mathbb{N}$

Given any subspace $\mathbb{S}$ there are many matrices A and W for which $\mathbb{S} = \text{im}(A) = \ker(W)$. Among these there are A, W that are minimal in the following sense: if $\mathbb{S} = \text{im}(A) = \ker(W)$ then the following are equivalent:

1. If $\mathbb{S} = \text{im}(\hat{A}) = \ker(\hat{W})$ then $\hat{A}$ has at least as many columns as A and $\hat{W}$ has at least as many rows as W ,
2. $\begin{bmatrix} A & W' \end{bmatrix}$ is square,
3. $\begin{bmatrix} A & W' \end{bmatrix}$ is square and invertible,
4. $A'A$ and WW' are invertible.

□

Example 2.5.4. In Example 2.5.2 we have

$$\mathbb{S} = \text{im}(A) = \ker(W) \text{ with } A = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \text{ and } W = \begin{bmatrix} 0 & 0 & 1 \end{bmatrix}.$$

Clearly $\begin{bmatrix} A & W' \end{bmatrix}$ is the 3×3 identity matrix I_3 hence by Lemma 2.5.3 A and W both are minimal in that no “smaller” $\hat{A}$ and $\hat{W}$ exist that represent $\mathbb{S}$. Minimality can also be inferred from invertibility of $A'A = I_2$ and $WW' = 1$. □

2.6 Projection

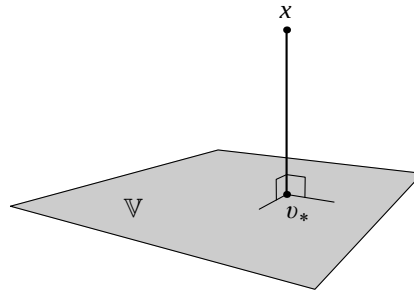


Figure 2.2: Projection in $\mathbb{R}^3$

With orthogonality, norm and subspace defined it should now be possible to mathematically formulate our intuition that connects minimal distance (norm) with orthogonality. Here is our first version. Have a look at the proof because it is the basis for later generalizations.

Definition 2.6.1 (Best approximation). An element $v_* \in \mathbb{V} \subseteq \mathbb{R}^n$ is a *best approximation* in $\mathbb{V}$ of $x \in \mathbb{R}^n$ if

$$|x - v_*| \leq |x - v| \quad \forall v \in \mathbb{V}.$$

□

Theorem 2.6.2 (A projection theorem). Let $x \in \mathbb{R}^n$ and let $\mathbb{V}$ be a subspace of $\mathbb{R}^n$.

1. A vector v_* is a best approximation in $\mathbb{V}$ of x iff $(x - v_*) \perp \mathbb{V}$.
2. The best approximation is unique and satisfies $|x - v_*|^2 = |x|^2 - |v_*|^2$.

Proof. If $(x - v_*) \perp \mathbb{V}$ then for any $v \in \mathbb{V}$,

$$|x - v|^2 = |(x - v_*) - \underbrace{(v - v_*)}_{\in \mathbb{V}}|^2 = |x - v_*|^2 + |v - v_*|^2 \quad \text{because } (x - v_*) \perp (v - v_*).$$

Hence if $v \neq v_*$ the norm $x - v$ exceeds that of $x - v_*$ making v_* the unique best approximation.

Conversely, suppose $(x - v_*) \not\perp \mathbb{V}$. Then there is a $v \in \mathbb{V}$ such that $(x - v_*) \not\perp v$ i.e. such that $(x - v_*) \cdot v \neq 0$. In particular such $v \neq 0$. Then

$$\begin{aligned} |x - (v_* + \alpha v)|^2 &= |(x - v_*) - \alpha v|^2 \\ &= |x - v_*|^2 - 2(x - v_*) \cdot (\alpha v) + |\alpha v|^2 \\ &= |x - v_*|^2 - 2\alpha[(x - v_*) \cdot v] + \alpha^2|v|^2. \end{aligned}$$

This quadratic expression in α is minimized for $\alpha = \frac{(x - v_*) \cdot v}{|v|^2}$ rendering it equal to

$$|x - (v_* + \alpha v)|^2 = |x - v_*|^2 - \frac{[(x - v_*) \cdot v]^2}{|v|^2} < |x - v_*|^2$$

so that v_* is not the best approximation. The equality $|x - v_*|^2 = |x|^2 - |v_*|^2$ is a restatement of Pythagoras. ■

The lemma avoids the issue of *existence* of the best approximation v_* because we prefer not to worry about it now. Here it does exist though. Existence can for instance be inferred by construction one, which we will do soon, but first a restatement of the projection theorem, a version that appears often in applications, see Fig. 2.3.

Theorem 2.6.3 (Projection theorem – minimum norm element of affine variety). Let $x \in \mathbb{R}^n$ and let $\mathbb{V}$ be a subspace of $\mathbb{R}^n$. A vector v_* is the minimal norm element of the affine variety $x + \mathbb{V}$ iff $v_* \perp \mathbb{V}$.

Proof. Do it yourself: Problem 2.1. ■

2.7 Transpose

For explicit representations of the best approximation we remind you of a trivial fact: the dot product can be also compactly be written using *transpose*:

$$x \cdot y = y'x = \begin{bmatrix} y_1 & y_2 & \cdots & y_n \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix}.$$

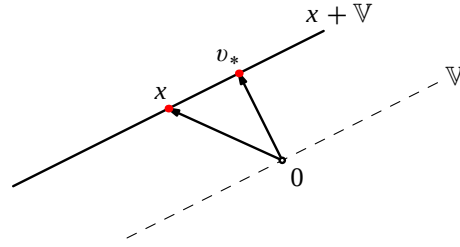


Figure 2.3: Minimum norm element v_* of affine variety $x + \mathbb{V}$

Then we get the handy rule that for any $n \times k$ -matrix M and any $z \in \mathbb{R}^k$ that

$$x \cdot (Mz) = (Mz)'x = z'(M'x) = (M'x) \cdot z.$$

2.8 Explicit projection: matrices and the normal equations

If we have the subspace given in the explicit form

$$\mathbb{V} = A\mathbb{R}^k.$$

then the best approximation of $v_* \in \mathbb{V}$ of x can be obtained rather explicitly:

Lemma 2.8.1 (Explicit projection – normal equations). Let $x \in \mathbb{R}^n$ and $A \in \mathbb{R}^{n \times k}$. Then

$$y_* = \operatorname{argmin}_{y \in \mathbb{R}^k} |x - Ay|$$

iff $y_* \in \mathbb{R}^k$ satisfies the *normal equations*

$$A'Ay_* = A'x. \tag{2.5}$$

The best approximation $v_* \in A\mathbb{R}^k$ of x then is $v_* = Ay_*$.

Proof. This is the projection theorem for $\mathbb{V} = A\mathbb{R}^k$ and $v_* = Ay_*$. By the projection theorem we need only establish that $(x - Ay_*) \perp \mathbb{V}$.

$$\begin{aligned} (x - Ay_*) \perp \mathbb{V} &\iff (x - Ay_*) \cdot (Ay) = 0 \quad \forall y \in \mathbb{R}^k \\ &\iff y'A'(x - Ay_*) = 0 \quad \forall y \in \mathbb{R}^k \\ &\iff A'(x - Ay_*) = 0, \quad \text{see Problem 2.2.} \end{aligned}$$

■

This result clearly shows that the transpose is a convenient notion. With it, projections can be formulated “explicitly”, something we’ll come back to later (at which point we generalize transpose to something called *adjoint operator*). The equations (2.5) are known as the *normal equations* (plural). The lemma does not require that $A'A$ is invertible and indeed the solution y_* of the normal equation need not be unique but if $A'A$ is invertible then (2.5) yields

$$y_* = (A'A)^{-1}A'x$$

and hence the best approximation $v_* = Ay_*$ equals

$$v_* = A(A'A)^{-1}A'x. \tag{2.6}$$

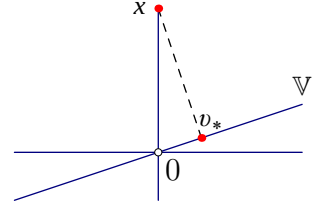
Realize that the best approximation $v_* = Ay_*$ according to Thm. 2.6.2 is *always* unique, even if y_* is not unique.

Example 2.8.2 (Projection in $\mathbb{R}^2$). Let

$$\mathbb{V} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \mathbb{R} \quad \text{and} \quad x = \begin{bmatrix} 0 \\ 1 \end{bmatrix}.$$

Then

$$v_* = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \left(\begin{bmatrix} 3 & 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} \right)^{-1} \begin{bmatrix} 3 & 1 \end{bmatrix} \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \frac{1}{10} = \begin{bmatrix} 0.3 \\ 0.1 \end{bmatrix}$$



□

2.9 Under-determined equations

Lemma 2.5.3 assures that *any* subspace $\mathbb{V}$ has a representation $\mathbb{V} = A\mathbb{R}^k$ with $A'A$ invertible. In applications we often encounter representations $\mathbb{V} = A\mathbb{R}^k$ with A a strictly *wide* matrix which rules out invertibility of $A'A$ but which may be such that AA' is invertible (note the difference).

Example 2.9.1. Let

$$A = \begin{bmatrix} 1 & 3 \end{bmatrix} \mathbb{R}^2, \quad x = 1.$$

Now there are many $y \in \mathbb{R}^2$ that satisfy $x = Ay$, for instance $y = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$. The solution set $\{y \in \mathbb{R}^2 \mid x = Ay\}$ for given x is an affine variety

$$\{y \in \mathbb{R}^2 \mid x = Ay\} = \left\{ \begin{bmatrix} 1 - 3\alpha \\ \alpha \end{bmatrix} \mid \alpha \in \mathbb{R} \right\} = \begin{bmatrix} 1 \\ 0 \end{bmatrix} + \ker(A).$$

□

In such cases one may opt for the smallest (in norm) y that solves $x = Ay$:

Lemma 2.9.2 (Minimal norm solution). If AA' is invertible then for every $x \in \mathbb{R}^n$ the equation $x - Ay = 0$ has at least one solution y and the unique minimal norm solution y_* is

$$y_* = A'(AA')^{-1}x. \quad (2.7)$$

This y_* is also the unique solution y_* in $\text{im}(A')$.

Proof. This is Theorem 2.6.3 for $\mathbb{V} = \text{im}(A)$ and $v_* = y_*$. Clearly the y_* of (2.7) satisfies $Ay_* = AA'(AA')^{-1}x = x$ and $y_* \perp \ker(A)$. ■

2.10 Orthogonal matrices and QR-factorization

The normal equations are particularly easy if $A'A = I$. The ramifications of this basic idea is what is known as *QR-factorization*. This factorization at present is the numerically most reliable method for solving least squares problems.

Definition 2.10.1 (orthogonal matrix). A matrix Q is *orthogonal* if it is square and $Q'Q = I$. □

Lemma 2.10.2. The following statements are equivalent.

- Q is orthogonal matrix
- Q is invertible and $Q'Q = I$
- Q is invertible and $QQ' = I$
- $Q'Q = QQ' = I$
- Q is square and all its columns have unit norm and are mutually orthogonal: $|Q_{*k}| = 1 \forall k$ and $Q_{*k} \perp Q_{*j} = 0 \forall j \neq k$.

Proof. Problem 2.5. ■

The term *orthogonal matrix* is unfortunate. Better would have been *orthonormal* matrix, since orthonormal means orthogonal *and* having norm 1.

Lemma 2.10.3 (QR-factorization). Every $A \in \mathbb{R}^{n \times k}$ has a factorization

$$A = QR$$

with $Q \in \mathbb{R}^{n \times n}$ orthogonal and $R \in \mathbb{R}^{n \times k}$ upper triangular. It is known as the *QR-factorization*.

Since R is upper triangular it naturally splits into a non-zero matrix R_1 and a zero matrix R_2 ,

$$R = \begin{bmatrix} R_1 \\ R_2 = 0 \end{bmatrix} \quad \text{with } R_2 \text{ of maximal dimension (possibly empty).}$$

Partitioning Q compatibly, the QR -factorization becomes

$$A = QR = \begin{bmatrix} Q_1 & Q_2 \end{bmatrix} \begin{bmatrix} R_1 \\ R_2 = 0 \end{bmatrix} = Q_1 R_1.$$

Proof. Consult a linear algebra book. ■

If R_2 is of maximal dimension then R_1 is of full row rank (equivalently: $R_1' R_1$ invertible). If $A'A$ is invertible then R_1 is square invertible, in which case the least squares solution (2.5) is unique and can efficiently be solved via

$$\begin{aligned} A'Ay_* = A'x &\iff R_1' Q_1' Q_1 R_1 y_* = R_1' Q_1' x \\ &\iff R_1' R_1 y_* = R_1' Q_1' x \\ &\iff R_1 y_* = Q_1' x. \end{aligned}$$

This last equation is easily solved for y_* because R_1 is upper triangular: solve y_* element-wise beginning with the final element of y_* and then moving up.

Another implication of the existence of QR -factorization is that

$$\text{im}(A) = \text{im}(Q_1 R_1) = \text{im}(Q_1) \quad \text{since } R_1 \text{ is full row rank.}$$

That is, subspaces $\text{im}(A)$ of $\mathbb{R}^n$ can always be seen as generated by orthogonal unit norm vectors (the columns of Q_1). Later we will generalize this to what is known as the Gram-Schmidt procedure.

2.11 If you want to know more...

- The mapping $P := A(A'A)^{-1}A'$ from x to its best approximation $v_* = Px$ is a projection. A *projection* (matrix) on $\mathbb{R}^n$ is a matrix P with the property that $P^2 = P$. It is an *orthogonal projection* if in addition $P = P'$, which is the case in (2.6).
- More variations of QR-factorization: google for Moore-Penrose inverse and SVD (singular value decomposition).

2.12 Problems

2.1 Prove Theorem 2.6.3.

2.2 Show that $x \cdot y = 0$ for all $y \in \mathbb{R}^n$ implies $x = 0$.

2.3 Let $W \in \mathbb{R}^{m \times n}$. Is $\ker(W)$ a subspace?

2.4 Show that a set $\mathbb{S} \subseteq \mathbb{R}^n$ is a subspace iff $(\mathbb{S}^\perp)^\perp = \mathbb{S}$.

2.5 Prove Lemma 2.10.2.

2.6 Consider

$$A = \begin{bmatrix} 1 & 0 \\ 1 & 4 \\ 0 & 1 \end{bmatrix}$$

- MATLAB: Compute the QR-factorization of A
 - Verify that the first and second column of Q are orthogonal
 - Sketch the columns of A and the first two columns of Q as vectors of $\mathbb{R}^3$
 - Do these two columns of A and two first columns of Q span the same plane?
 - MATLAB: Compute the best approximation in $\mathbb{V} = \text{im}(A)$ of $x = (0, 0, 1)$
- 2.7 Sketch the affine variety $\begin{bmatrix} 0 \\ 1 \end{bmatrix} + \text{im}\begin{bmatrix} 1 \\ -2 \end{bmatrix}$ and determine the minimal norm element of this set.
- 2.8 Let $A \in \mathbb{R}^{n \times 3}$ and let (as always) A_{*k} denote its k column. Show that

$$A'A = \begin{bmatrix} |A_{*1}|^2 & A_{*2} \cdot A_{*1} & A_{*3} \cdot A_{*1} \\ A_{*1} \cdot A_{*2} & |A_{*2}|^2 & A_{*3} \cdot A_{*2} \\ A_{*1} \cdot A_{*3} & A_{*2} \cdot A_{*3} & |A_{*3}|^2 \end{bmatrix}.$$

2.9 Let $x = (0, 0, 1) \in \mathbb{R}^3$ and

$$\mathbb{V} = \ker\begin{bmatrix} 1 & 1 & 1 \end{bmatrix}.$$

- Express $\mathbb{V}$ as $\mathbb{V} = \text{im}(A)$ for some matrix A
 - MATLAB: Determine with MATLAB the best approximation in $\mathbb{V}$ of $x = (0, 0, 1)$.
 - Sketch $\mathbb{V}$ and both $x = (0, 0, 1)$ and its best approximation. (It need not be artwork.)
- 2.10 Show that $Q \in \mathbb{R}^{2 \times 2}$ is orthogonal iff it is of the form

$$Q = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ \sin(\alpha) & -\cos(\alpha) \end{bmatrix}.$$

3 Vector space and subspace

If we are to generalize the projection theorem we should generalize the four players in the projection theorem:

- space $\mathbb{R}^n$,
- subspace $\mathbb{V}$ of $\mathbb{R}^n$,
- dot product,
- Euclidean norm.

In this chapter we generalize space $\mathbb{R}^n$ (to be called vector space) and subspace $\mathbb{V}$ (still to be called subspace). As always $\mathbb{R}$ denotes the set of real numbers and the set of complex numbers is denoted by $\mathbb{C}$. The set of positive integers $\{1, 2, 3, \dots\}$ is denoted by $\mathbb{N}$ (by $\mathbb{N}_0$ if zero is included) and all integers $\{\dots, -1, 0, 1, \dots\}$ by $\mathbb{Z}$.

3.1 Real vector space

The question thus is: what properties of $\mathbb{R}^n$ did we (implicitly) use in the projection theorem and its proof? Have a look at Thm. 2.6.2 and its proof and you'll probably agree that the following eight properties of $\mathbb{R}^n$ are sufficient:

Definition 3.1.1 (Real vector space). A *real vector space* $\mathbb{X}$ is a set of elements, called *vectors*, on which vector *addition* $\mathbb{X} \times \mathbb{X} \rightarrow \mathbb{X}$ and real scalar *multiplication* $\mathbb{R} \times \mathbb{X} \rightarrow \mathbb{X}$ is defined with the following eight properties for all $v, w \in \mathbb{X}$ and all scalars $\alpha, \beta \in \mathbb{R}$:

1. $u + v = v + u$ *commutative*
2. $(u + v) + w = u + (v + w)$. *associative*
3. There is a unique *zero vector* or *origin* $0 \in \mathbb{X}$ such that $u + 0 = u \ \forall u \in \mathbb{X}$
4. For each $v \in \mathbb{X}$ there is a unique *additive inverse* $-v \in \mathbb{X}$ such that $v + (-v) = 0$
5. $0v = 0$ and $1v = v$.
6. $\alpha(\beta v) = (\alpha\beta)v$. *associative*
7. $(\alpha + \beta)v = \alpha v + \beta v$ *distributive*
8. $\alpha(u + v) = \alpha u + \alpha v$ *distributive*

□

Actually that $0v = 0$ for any vector v is implied by the other properties (see Problem 3.18). If this is your first contact with such a formal definition then please realize this: we have the freedom to define or own “addition” and “multiplication” and we may dream up really weird sets $\mathbb{X}$; but the moment that $\mathbb{X}$ with that addition and multiplication satisfies the eight axioms of vector space then automatically all results we will derive for general vector spaces hold for our weird $\mathbb{X}$ as well. That's the beauty of generality and abstraction.

Example 3.1.2 ($\mathbb{R}^n$). The space $\mathbb{R}^n$ is the set of ordered sequences of length n with entries in $\mathbb{R}$,

$$\mathbb{R}^n = \{ u \mid u = (u_1, u_2, \dots, u_n), u_k \in \mathbb{R} \}.$$

It is a vector space under the addition and scalar multiplication defined elementwise as

$$\begin{aligned} u + v &:= (u_1 + v_1, u_2 + v_2, \dots, u_n + v_n), \\ \alpha u &:= (\alpha u_1, \alpha u_2, \dots, \alpha u_n). \end{aligned}$$

The subtlety is that the plus-sign in $u + v$ represents addition of two vectors whereas the plus-sign in $u_1 + v_1$ represents ordinary addition of two real numbers. Likewise αu is a product of scalar α and vector u while αu_1 simply means product of real numbers. $\square$

Example 3.1.3 ($\ell_{\mathbb{R}}(\mathbb{N})$). The space $\ell_{\mathbb{R}}(\mathbb{N})$ is the set of one-sided infinite sequences

$$\ell_{\mathbb{R}}(\mathbb{N}) = \{ u \mid u = (u_1, u_2, u_3, \dots), u_k \in \mathbb{R}, k \in \mathbb{N} \}.$$

As in $\mathbb{R}^n$ it is a vector space under the addition and scalar multiplication defined elementwise as

$$\begin{aligned} u + v &:= (u_1 + v_1, u_2 + v_2, u_3 + v_3, \dots), \\ \alpha v &:= (\alpha u_1, \alpha u_2, \alpha u_3, \dots). \end{aligned}$$

$\square$

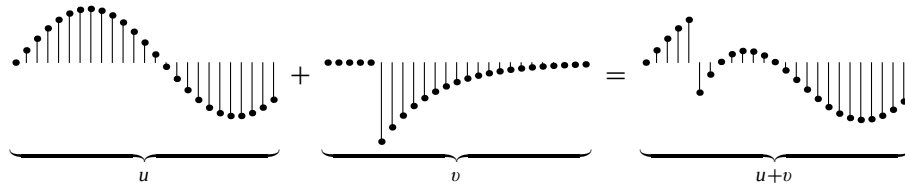


Figure 3.1: Two vectors $u, v \in \mathbb{R}^{25}$ and sum $u + v$

Figure 3.1 depicts vector addition in $\mathbb{R}^{25}$. (In fact the reason for us to include this figure is to convince you of the fact that also function spaces can be seen as vector spaces, see the next example, and that conceptually the step from $\mathbb{R}^n$ to function space is marginal.)

Example 3.1.4 (function space). Also sets of real-valued functions defined on, say, $[0, 1]$

$$\mathfrak{F}([0, 1], \mathbb{R}) := \{ f : [0, 1] \rightarrow \mathbb{R} \}$$

can be seen as a vector space: under addition and scalar multiplication defined pointwise,

$$\begin{aligned} f + g &:= h && \text{in which } h(t) = f(t) + g(t) \text{ for each } t \in [0, 1], \\ \alpha f &:= h && \text{in which } h(t) = \alpha f(t) \text{ for each } t \in [0, 1] \end{aligned}$$

the space $\mathfrak{F}([0, 1], \mathbb{R})$ is a vector space, see Fig. 3.2.

It is a bit of a bore to verify the eight defining rules of vector space, but once we have to do it: for our function space rules 1 and 2 are trivial; there is only one function “0” with the property that $v + 0 = v \forall v$: it is the zero function $0(t) = 0 \forall t$; given any function v only one function g exists such that $v + g$ is the zero function: define g pointwise as $g(t) = -v(t)$, we denote this g as $-v$; rule 5 is trivial; in fact rules 6, 7 and 8 are trivial as well. Done. $\square$

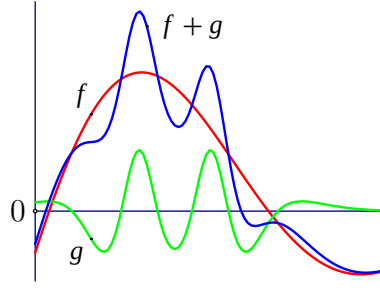


Figure 3.2: Two functions f and g and their sum $f + g$

The notation $\mathfrak{F}(\mathbb{T}, \mathbb{V})$ will be used to denote the set of mappings from $\mathbb{T}$ to $\mathbb{V}$. For the particular case that $\mathbb{T}$ is countable, $\mathbb{T} \subseteq \mathbb{Z}$, we use instead the notation $\ell_{\mathbb{V}}(\mathbb{T})$.

3.2 Complex vector space

A complex vector space differs from a real vector space only in that the scalars — the α s and β s — in complex vector space are taken from $\mathbb{C}$ instead of $\mathbb{R}$. For completeness: a *complex vector space* $\mathbb{X}$ is a set of elements, called *vectors*, on which vector *addition* $\mathbb{X} \times \mathbb{X} \rightarrow \mathbb{X}$ and complex scalar *multiplication* $\mathbb{C} \times \mathbb{X} \rightarrow \mathbb{X}$ is defined that satisfy the 8 properties of Definition 3.1.1 for all $v, w \in \mathbb{X}$ and all $\alpha, \beta \in \mathbb{C}$. From the context it will be clear whether we deal with real or complex vector spaces and we refer to the “ α s and β s” simply as *scalars*.

Example 3.2.1 ($\mathbb{C}^n$). The space $\mathbb{C}^n$ is the set of ordered sequences of complex numbers of length n ,

$$\mathbb{C}^n = \{u \mid u = (u_1, u_2, \dots, u_n), u_1, \dots, u_n \in \mathbb{C}\}.$$

It is a vector space under the addition and scalar multiplication defined elementwise as

$$\begin{aligned} u + v &= (u_1 + v_1, u_2 + v_2, \dots, u_n + v_n), \\ \alpha u &= (\alpha u_1, \alpha u_2, \dots, \alpha u_n). \end{aligned}$$

□

Example 3.2.2 ($\ell_{\mathbb{C}}(\mathbb{Z})$). The space $\ell_{\mathbb{C}}(\mathbb{Z})$ is the set of doubly infinite ordered sequences

$$\ell_{\mathbb{C}}(\mathbb{Z}) = \{u \mid u = (\dots, u_{-1}, u_0, u_1, u_2, \dots), u_k \in \mathbb{C}, k \in \mathbb{Z}\}.$$

It is a vector space under the addition and scalar multiplication defined elementwise as

$$\begin{aligned} u + v &= (\dots, u_{-1} + v_{-1}, u_0 + v_0, u_1 + v_1, u_2 + v_2, \dots), \\ \alpha u &= (\dots, \alpha u_{-1}, \alpha u_0, \alpha u_1, \alpha u_2, \dots). \end{aligned}$$

□

Example 3.2.3 (function space). Complex-valued functions defined on, say, $[0, 1]$

$$\mathfrak{F}([0, 1], \mathbb{C}) = \{f : [0, 1] \rightarrow \mathbb{C}\}$$

can be seen as a vector space with addition and scalar multiplication defined pointwise as

$$\begin{aligned} f + g = h & \quad \text{in which } h(t) = f(t) + g(t) \text{ for each } t \in [0, 1], \\ \alpha f = h & \quad \text{in which } h(t) = \alpha f(t) \text{ for each } t \in [0, 1]. \end{aligned}$$

□

3.3 Subspace

Actually subspace hardly needs any modification.

Definition 3.3.1 (Subspace). A subset $\mathbb{V}$ of a vector space $\mathbb{X}$ is a *subspace* if for all $u, v \in \mathbb{V}$ and scalar α :

1. $0 \in \mathbb{V}$,
2. $u + v \in \mathbb{V}$, *closed under addition*
3. $\alpha v \in \mathbb{V}$. *closed under scaling*

□

In a non-empty set $\mathbb{V}$ the third condition implies the first (take $\alpha = 0$). Therefore the first condition in effect only says that subspaces are not allowed to be empty.

Example 3.3.2 (function space). The set

$$\mathbb{S} = \{ f : \mathbb{R} \rightarrow \mathbb{R} \mid \exists a, b \in \mathbb{R} : \text{such that } f(t) = a \cos(t) + b \sin(t) \forall t \in \mathbb{R} \}$$

is a subspace of $\mathfrak{F}(\mathbb{R}, \mathbb{R})$. Indeed (1) the zero function is an element of $\mathbb{S}$ (take $a = b = 0$), (2) it is closed under addition for if $f_k(t) = a_k \cos(t) + b_k \sin(t) \in \mathbb{S}$ then so is their sum $(f_1 + f_2)(t) = (a_1 + a_2) \cos(t) + (b_1 + b_2) \sin(t) \in \mathbb{S}$. Finally it is closed under scalar multiplication for if $f(t) := a \cos(t) + b \sin(t) \in \mathbb{S}$ then so is $\alpha f(t) = (\alpha a) \cos(t) + (\alpha b) \sin(t) \in \mathbb{S}$.

Our intuition from $\mathbb{R}^3$ that says that a subspace is “something flat” fails for function spaces. It is a subspace nonetheless. □

Example 3.3.3 ($\ell_{\mathbb{R}}^{\text{finite}}(\mathbb{N})$). The set

$$\ell_{\mathbb{R}}^{\text{finite}}(\mathbb{N}) = \{ u : \mathbb{N} \rightarrow \mathbb{R} \mid \text{only finitely many } u_k \text{ are nonzero} \}$$

is a subspace of $\ell_{\mathbb{R}}(\mathbb{N})$. See problem 3.10. □

Example 3.3.4 (Subspaces of matrices). The set of real square $n \times n$ matrices is denoted $\mathbb{R}^{n \times n}$ and it is a vector space under the usual matrix addition and scalar multiplication. The subset $\mathbb{S} \subset \mathbb{R}^{n \times n}$ of upper-triangular matrices

$$\mathbb{S} = \{ A \in \mathbb{R}^{n \times n} \mid A_{i,j} = 0 \text{ for any } 0 < j < i \leq n \}$$

is a subspace of $\mathbb{R}^{n \times n}$. See problem 3.11. □

Example 3.3.5 (Differential equation). The space

$$\mathbb{S} := \{ f \in \mathfrak{F}(\mathbb{R}, \mathbb{R}) \mid f \text{ is differentiable and } f'(t) + f(t) = 0 \forall t \}$$

is a subspace of $\mathfrak{F}(\mathbb{R}, \mathbb{R})$. That is, the set of solutions of the homogeneous linear differential equation $f' + f = 0$ is a subspace.

Proof: clearly the zero function $0(t) \equiv 0$ satisfies the differential equation. Suppose $f, g \in \mathbb{S}$ so they both are differentiable and satisfy the differential equation, then $f + g$ is differentiable as well and we have

$$\begin{aligned}(f + g)'(t) + (f + g)(t) &= f'(t) + g'(t) + f(t) + g(t) \\ &= [f'(t) + f(t)] + [g'(t) + g(t)] = 0 + 0 = 0.\end{aligned}$$

Similarly if $f \in \mathbb{S}$ then so is $g := \alpha f$ for any scalar $\alpha \in \mathbb{R}$ because g is then differentiable and

$$g + g' = \alpha f + [\alpha f]' = \alpha(f + f') = 0.$$

This completes the proof. □

Subspaces inherit the defining properties (Dn. 3.1.1) of the vector space they are part of and are themselves vector spaces. For instance the (x, y) -plane of Example 2.5.2 is a subspace of $\mathbb{R}^3$ but is also a vector space in its own right and it can be identified with vector space $\mathbb{R}^2$.

The real vector space $\mathbb{R}^n$ is not a subspace of the complex vector space $\mathbb{C}^n$. It is a subset, $\mathbb{R}^n \subseteq \mathbb{C}^n$, and it is closed under addition, but $\mathbb{R}^n$ is not closed under scaling since the scalars are allowed to be complex.

3.4 Linear combination, span, basis and dimension

Lemma 2.5.3 says that in $\mathbb{R}^n$ subspaces can be represented as $A\mathbb{R}^k$ that is as the set of *linear combinations* of a *finite* number of columns $\{A_{*1}, A_{*2}, \dots, A_{*k}\}$. With $\mathbb{R}^n$ in mind we similarly define linear combination, span, finite basis and dimension on arbitrary vector spaces as follows.

Definition 3.4.1 (linear combination, linear dependence, span, basis and dimension). Let $\mathbb{X}$ be a vector space and let $v_1, v_2, \dots, v_n$ be a finite set of elements in $\mathbb{X}$.

- A *linear combination* of $v_1, v_2, \dots, v_n$ is a sum of the form

$$\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_n v_n$$

with α_k scalars.

- The set $\{v_1, v_2, \dots, v_n\}$ is *linearly independent* if none of its elements v_k equals a linear combination of the others. (The set $\{0\}$ is defined to be dependent.)
- The *span* of $v_1, v_2, \dots, v_n$ is the set of all possible linear combinations

$$\text{span}\{v_1, v_2, \dots, v_n\} := \{\alpha_1 v_1 + \alpha_2 v_2 + \dots + \alpha_n v_n \mid \alpha_k \text{ scalar}\}.$$

- The set $\{v_1, v_2, \dots, v_n\}$ is a *basis* of $\mathbb{X}$ if
 - $\text{span}\{v_1, v_2, \dots, v_n\} = \mathbb{X}$, and
 - the set $\{v_1, v_2, \dots, v_n\}$ is linear independent.
- If $\mathbb{X}$ has a finite basis then the *dimension* $\dim(\mathbb{X})$ of $\mathbb{X}$ is defined as the number of basis vectors. A space $\mathbb{X}$ is *finite dimensional* if it has a finite basis, otherwise $\mathbb{X}$ is said to be *infinite dimensional*.

□

Of course this assumes that the so defined dimension of $\mathbb{X}$ does not depend on the choice of basis. Indeed, that is the case. Linear independence can equivalently be characterized as follows.

Lemma 3.4.2 (Linear independence). A set $\{v_1, v_2, \dots, v_n\}$ is linearly independent iff $\sum_{i=1}^n \alpha_i v_i = 0$ implies all scalars are zero: $\alpha_i = 0 \forall i = 1, 2, \dots, n$. □

You may wonder why in this section we do not consider *infinite* sets of vectors, $v_1, v_2, \dots$. The reason is that we do not yet know how, on general vector spaces, infinite sums are defined! This requires some form of convergence. It will be resolved when we consider inner products — more particularly norms — in Chapter 5.

Example 3.4.3 ($\mathbb{R}^3$). Let $v = (1, 2, 3)$ and $w = (0, 1, -1)$ both in $\mathbb{R}^3$. Then $(1, 5, 1)$ is a linear combination of v and w while $(0, 1, 5)$ is not.

It is also interesting to see what the set of *all* linear combinations of v and w are:

$$\text{span}\{v, w\} = \{\alpha_1 v + \alpha_2 w = \alpha_1 \begin{bmatrix} 1 \\ 2 \\ 3 \end{bmatrix} + \alpha_2 \begin{bmatrix} 0 \\ 1 \\ -1 \end{bmatrix} = \begin{bmatrix} \alpha_1 \\ 2\alpha_1 + \alpha_2 \\ 3\alpha_1 - \alpha_2 \end{bmatrix} \mid \alpha_1, \alpha_2 \in \mathbb{R}\}.$$

Clearly a linear combination $\alpha_1 v + \alpha_2 w$ is zero only if its first entry (which is α_1) is zero. Hence $\alpha_1 v + \alpha_2 w = 0$ implies $\alpha_1 = 0$ but then $\alpha_2 w = 0$ implying $\alpha_2 = 0$. The two vectors v, w therefore are linearly independent and hence the v, w form a basis for the subspace $\text{span}\{v, w\}$ and $\dim \text{span}\{v, w\} = 2$.

(We could also have invoked Lemma 2.5.3 and deduce that $A'A$ is invertible for

$$A := [v \quad w] = \begin{bmatrix} 1 & 0 \\ 2 & 1 \\ 3 & -1 \end{bmatrix}.)$$

□

Example 3.4.4 (Span of monomials). Consider the polynomials $p_k : \mathbb{R} \rightarrow \mathbb{R}$,

$$p_k(t) = t^k, \quad k = 0, 1, 2, \dots$$

For ease of reference we will refer to these functions p_k simply as t^k . Then

$$\text{span}\{1, t, t^2, \dots, t^{10}\}$$

is the set of all polynomials of degree 10 or less. □

Any finite span is necessarily a subspace.

Example 3.4.5. $\mathbb{R}^n$ has dimension n . □

Example 3.4.6 (Polynomials). The vector space of polynomials of degree n or less,

$$\mathbb{P}_n := \{p : \mathbb{R} \rightarrow \mathbb{R} \mid p(t) = p_n t^n + p_{n-1} t^{n-1} + \dots + p_1 t + p_0\}$$

has dimension $n + 1$. Indeed, the dimension is at most $n + 1$ because it is spanned by $n + 1$ elements¹,

$$\mathbb{P}_n = \text{span}\{1, t, t^2, \dots, t^n\}.$$

Convince yourself of the fact that the dimension is not less than $n + 1$, see Problem 3.12. □

¹Why is the dimension $\leq m$ if the space is spanned by m elements? Problem 3.20

Example 3.4.7 (Infinite dimensional). Consider the vector space of polynomials of *any* degree

$$\mathbb{P} := \{p : \mathbb{R} \rightarrow \mathbb{R} \mid \exists n \in \mathbb{N}, p_k \in \mathbb{R} : p(t) = p_n t^n + p_{n-1} t^{n-1} + \cdots + p_1 t + p_0 \forall t \in \mathbb{R}\}.$$

Clearly for any $n \in \mathbb{N}$ we have $\mathbb{P}_n \subset \mathbb{P}$ so that the dimension is at least $n + 1$. Since n is arbitrary $\mathbb{P}$ is infinite dimensional. $\square$

Example 3.4.8. What is the dimension of $\mathbb{S} := \text{span}\{1, e^t, e^{2t}\} \subset \mathfrak{F}(\mathbb{R}, \mathbb{R})$? Surely it is at most 3. Let us verify linear independence. So consider

$$\alpha_1 + \alpha_2 e^t + \alpha_3 e^{2t} = 0 \quad \forall t \in \mathbb{R}.$$

In particular this should hold for $t = 0, t = 1, t = 2$:

$$\begin{cases} \alpha_1 + \alpha_2 + \alpha_3 = 0 \\ \alpha_1 + \alpha_2 e + \alpha_3 e^2 = 0 \\ \alpha_1 + \alpha_2 e^2 + \alpha_3 e^4 = 0 \end{cases} \quad (3.1)$$

These are three equations in three unknowns and as is well known its solution is unique iff its determinant is nonzero:

$$\det \begin{bmatrix} 1 & 1 & 1 \\ 1 & e & e^2 \\ 1 & e^2 & e^4 \end{bmatrix} = (e - 1)(e^2 - 1)(e^2 - e) \neq 0.$$

The determinant is nonzero, hence (3.1) has unique solution $\alpha_1 = \alpha_2 = \alpha_3 = 0$. The set $\{1, e^t, e^{2t}\}$ therefore is a linear independent, i.e., it is a basis of $\mathbb{S} := \text{span}\{1, e^t, e^{2t}\}$ and it has dimension 3. $\square$

3.5 If you want to know more...

- http://en.wikipedia.org/wiki/Vector_space

3.6 Problems

- 3.1 Is $\mathbb{Z}$ is real vector space?
- 3.2 Are the three vectors w^z, w^y, w^x of (1.1) linearly independent?
- 3.3 Consider the set $\mathbb{V} = \{f : \mathbb{R} \rightarrow \mathbb{C} \mid f(t) = e^{ct} \text{ for some } c \in \mathbb{C}\}$. On this set we define the weird vector addition and scalar multiplication

$$e^{c_1 t} + e^{c_2 t} := e^{(c_1 + c_2)t}, \quad \alpha e^{ct} := e^{\alpha ct}.$$

With this addition and multiplication, is $\mathbb{V}$ a complex vector space?

- 3.4 Is $\{f : \mathbb{R} \rightarrow \mathbb{R} \mid \exists M, |f(t)| < M \forall t\}$ a subspace of $\mathfrak{F}(\mathbb{R}, \mathbb{R})$?
- 3.5 Show that $\{f : \mathbb{R} \rightarrow \mathbb{R} \mid \int_0^1 f(t) dt = 0\}$ is a subspace of $\mathfrak{F}(\mathbb{R}, \mathbb{R})$.

- 3.6 Show that $\{f : \mathbb{R} \rightarrow \mathbb{R} \mid \int_{-\infty}^{\infty} |f(t)| dt < \infty\}$ is a vector space (under standard addition and multiplication).
- 3.7 Consider the subset of $\mathbb{P}_{10}$
- $$\mathbb{S} = \{f \in \mathbb{P}_{10} \mid f(2) = 0\}.$$
- Show that $\mathbb{S}$ is a subspace of $\mathbb{P}_{10}$.
 - What is the dimension of $\mathbb{S}$?
 - Determine a basis for $\mathbb{S}$.
- 3.8 Consider the three functions $f_1(t) = 1$, $f_2(t) = e^t$, $f_3(t) = e^{-t}$.
- Check whether or not the three functions are independent (as elements of $\mathfrak{F}(\mathbb{R}, \mathbb{R})$).
 - Determine a basis for $\text{span}\{f_1, f_2, f_3\} \subseteq \mathfrak{F}(\mathbb{R}, \mathbb{R})$.
- 3.9 What is the dimension of the subspace of Example 3.3.2.
- 3.10 Show that the $\ell_{\mathbb{R}}^{\text{finite}}(\mathbb{N})$ of Example 3.3.3 is a subspace of $\ell_{\mathbb{R}}(\mathbb{N})$.
- 3.11 Show that the $\mathbb{S}$ of Example 3.3.4 is a subspace.
- 3.12 Consider Example 3.4.6. Show that $\text{span}\{1, t, \dots, t^n\}$ has dimension $n + 1$.
- 3.13 Let $\mathbb{S}_1$ and $\mathbb{S}_2$ be subspaces of some vector space $\mathbb{X}$. Is $\mathbb{S}_1 \cap \mathbb{S}_2$ a subspace of $\mathbb{X}$?
- 3.14 Let $\mathbb{S}_1$ and $\mathbb{S}_2$ be subspaces of some vector space $\mathbb{X}$. Is $\mathbb{S}_1 + \mathbb{S}_2 := \{s_1 + s_2 \mid s_1 \in \mathbb{S}_1, s_2 \in \mathbb{S}_2\}$ a subspace of $\mathbb{X}$?
- 3.15 what is the smallest subspace of $\mathbb{R}^3$ that contains the unit circle $\{(x, y, z) \mid x^2 + y^2 = 1, z = 0\}$?
- 3.16 Is the set of diagonal matrices in $\mathbb{R}^{n \times n}$ a subspace of $\mathbb{R}^{n \times n}$?
- 3.17 Is the set of symmetric matrices in $\mathbb{R}^{n \times n}$ a subspace of $\mathbb{R}^{n \times n}$?
- 3.18 Show that the property $0v = 0$ for any vector v in a vector space is implied by the other defining properties of vector space.
- 3.19 Is $\ell_{\mathbb{C}}(\mathbb{N})$ finite dimensional?
- 3.20 Show that $\dim(\mathbb{V}) \leq m$ if $\mathbb{V}$ is spanned by $m \in \mathbb{N}$ elements.

4 Linear transformation – operator

Linear transformations (aka linear operators) are everywhere. For instance the projection theorem (page 9) states that the best approximation v_* of any x is unique so we can consider the mapping $\mathcal{F}$ that sends x to its best approximation $v_* = \mathcal{F}(x)$. This is just one of the many mappings $\mathcal{F}$ that turn out to be linear. The main reason we are reviewing properties of linear transformations is that we want to generalize the projection theorem (next chapter) to cases where the subspace $\mathbb{V}$ is given as the image of a linear map (remember Section 2.8 for the $\mathbb{R}^n$ -version.)

4.1 Linear transformation

Definition 4.1.1 (Operator). Let $\mathbb{V}$ and $\mathbb{W}$ be two vector spaces (both real or both complex). A mapping $\mathcal{F}$ from $\mathbb{V}$ to $\mathbb{W}$ is *linear* if for every $v_1, v_2, v \in \mathbb{V}$ and scalar α :

1. $\mathcal{F}(v_1 + v_2) = \mathcal{F}(v_1) + \mathcal{F}(v_2)$, *additive*
2. $\mathcal{F}(\alpha v) = \alpha \mathcal{F}(v)$. *homogeneous*

□

If $\mathcal{F}$ maps from $\mathbb{V}$ to $\mathbb{W}$ we write $\mathcal{F} : \mathbb{V} \rightarrow \mathbb{W}$. We can apply $\mathcal{F}$ to elements (vectors) $v \in \mathbb{V}$ but also to sets $\mathbb{S} \subseteq \mathbb{V}$, and we write

$$\mathcal{F}\mathbb{S} := \{\mathcal{F}(v) \mid v \in \mathbb{S}\}.$$

The *range* of $\mathcal{F} : \mathbb{V} \rightarrow \mathbb{W}$ is defined as $\mathcal{F}\mathbb{V}$, i.e. it is the set of all possible outcomes of the mapping. The range is also known as the *image* (of its domain) and is denoted as $\text{im}(\mathcal{F})$. The set $\mathbb{W}$ to which $\mathcal{F}$ maps is sometimes referred to as the *codomain*.

This is an attempt to explain what linearity means on function space: suppose we know two outcomes of $\mathcal{F}$:

$$\mathcal{F}(\text{step}) = \text{slope} \quad \text{and} \quad \mathcal{F}(\text{bump}) = \text{spike}$$

then additivity implies that

$$\mathcal{F}(\text{step} + \text{bump}) = \text{slope} + \text{spike}$$

and homogeneity that

$$\mathcal{F}(2 \cdot \text{bump}) = 2 \cdot \text{spike}.$$

4.2 Familiar linear transformations

Well, the most familiar linear transformations are the ones that map from $\mathbb{R}^n$ to $\mathbb{R}^k$ —see the next section—but here are some other standard ones. It is easy to verify that they are indeed linear. Following the colloquial definition some trickier issues regarding domain and codomain are added.

Example 4.2.1 (Fourier transform). The *Fourier transformation* is a linear transformation that sends continuous time functions $f : \mathbb{R} \rightarrow \mathbb{R}$ to continuous frequency functions $\hat{f} : \mathbb{R} \rightarrow \mathbb{R}$, defined as

$$\hat{f} = \mathcal{F}(f) : \quad \hat{f}(\omega) = \int_{-\infty}^{\infty} f(t) e^{-i\omega t} dt.$$

As domain $\mathbb{V}$ we could take the set of absolutely integrable functions $\{f : \mathbb{R} \rightarrow \mathbb{R} \mid \int_{\mathbb{R}} |f(t)| dt < \infty\}$ (with standard addition and multiplication) because then $\mathcal{F}(f)$ is well defined for every $\omega \in \mathbb{R}$. See Problem 3.6. $\square$

Example 4.2.2 (Fourier series). The *Fourier series* can be seen as a linear mapping that sends continuous time functions on finite interval, $f : [0, T] \rightarrow \mathbb{R}$ to countably many frequency coefficients $\hat{f} : \mathbb{Z} \rightarrow \mathbb{C}$,

$$\hat{f} = \mathcal{F}(f) : \quad \hat{f}_k = \frac{1}{T} \int_0^T f(t) e^{-ik \frac{2\pi}{T} t} dt, \quad k \in \mathbb{Z}.$$

As domain $\mathbb{V}$ we could take the set of continuous functions on $[0, T]$ (but other sensible domains can be dreamed up). Codomain $\ell_{\mathbb{C}}(\mathbb{Z})$ is natural. $\square$

Example 4.2.3 (Laplace transform). Likewise the unilateral Laplace transform $\mathcal{L}$ is linear as well,

$$F = \mathcal{L}(f) : \quad F(s) = \int_0^{\infty} f(t) e^{-st} dt.$$

Domain and codomain are tricky here: if as domain we take $\mathbb{V}_{s_0} = \{f \mid \exists c > 0 : |f(t)| < ce^{s_0 t}\}$ then as codomain $\mathbb{W}$ it suffices to take $\mathbb{W}_{s_0} = \{f : ((s_0, \infty) + i\mathbb{R}) \rightarrow \mathbb{C}\}$. $\square$

Example 4.2.4 (Convolution and Fredholm). Here is another familiar linear mapping: the *convolution* $\mathcal{C}_h$,

$$y = \mathcal{C}_h(u) : \quad y(t) = (h * u)(t) := \int_{-\infty}^{\infty} h(\tau) u(t - \tau) d\tau$$

The convolution is in fact a special case of the “general” linear mapping from $\mathfrak{F}(\mathbb{R}, \mathbb{R})$ to $\mathfrak{F}(\mathbb{R}, \mathbb{R})$,

$$y = \mathcal{F}_{\text{fredholm}}(u) : \quad y(t) = \int_a^b K(t, s) u(s) ds.$$

If a and b are finite and $K(t, s)$ is continuous and u is continuous as well, then the operator is well defined and its outcome is continuous. The equation relating u and y is often called *Fredholm equation* (and the game then is to find u for given K and y). $\square$

Example 4.2.5 (Differentiator). Also linear is the differentiator $\mathcal{D}$,

$$f = \mathcal{D}(g) : \quad f(t) = \frac{\partial}{\partial t} g(t).$$

As domain $\mathbb{V}$ we should take a set whose elements are differentiable. Let us verify linearity. For one it is additive,

$$\mathcal{D}(g + h)(t) = \frac{\partial}{\partial t} [g(t) + h(t)] = \frac{\partial}{\partial t} g(t) + \frac{\partial}{\partial t} h(t) = (\mathcal{D}g)(t) + (\mathcal{D}h)(t)$$

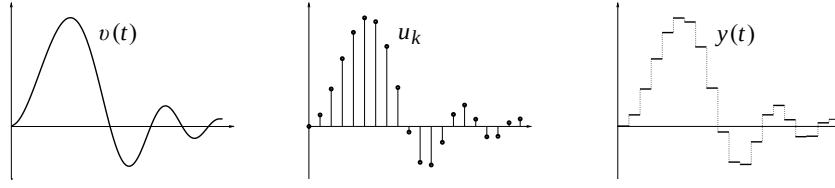


Figure 4.1: Original signal, sampled signal, held signal (zero order hold)

and it is homogeneous as well,

$$\mathcal{D}(\alpha g)(t) = \frac{\partial}{\partial t}[\alpha g(t)] = \alpha \frac{\partial}{\partial t}g(t) = \alpha(\mathcal{D}g)(t).$$

□

Example 4.2.6 (Sampler). The ideal sampler $\mathcal{S}_h$ with sampling period $h > 0$ is defined as

$$u = \mathcal{S}_h(v) : \quad u_k = v(kh), \quad k \in \mathbb{Z}.$$

It maps functions to sequences. See Fig. 4.1. It is a linear transformation $\mathcal{S}_h : \mathbb{V} \rightarrow \mathbb{U}$ for, for instance, the vector spaces $\mathbb{V} = \{v : \mathbb{R} \rightarrow \mathbb{R} \mid v \text{ is continuous}\}$ and $\mathbb{U} = \ell_{\mathbb{R}}(\mathbb{Z})$ with their standard addition and multiplication. Linearity in words means “the samples of the sum equals the sum of the samples”. Indeed,

$$\mathcal{S}_h(f + g)_k = (f + g)(kh) = f(kh) + g(kh) = \mathcal{S}_h(f)_k + \mathcal{S}_h(g)_k.$$

It is also homogeneous: “the samples of a scaled signal is are scaled samples of the signal” (or “scaling commutes with sampling”):

$$\mathcal{S}_h(\alpha f)_k = (\alpha f)(kh) = \alpha(f(kh)) = \alpha \mathcal{S}_h(f)_k.$$

(In digital communication the samples are digitized, i.e. mapped to some finite set, which makes it a nonlinear transformation.) □

Example 4.2.7 (Hold). The zero-order hold $\mathcal{H}_h : \mathbb{U} \rightarrow \mathbb{Y}$ is best understood from Fig. 4.1 and is formally defined as

$$y = \mathcal{H}_h(u) : \quad y(kh + \tau) = u_k \quad \forall k \in \mathbb{Z}, \tau \in [0, h).$$

It sends sequences to continuous-time functions. It is a linear transformation from domain

$$\mathbb{U} = \ell_{\mathbb{R}^n}(\mathbb{Z})$$

to codomain, say,

$$\mathbb{Y} := \{y : \mathbb{R} \rightarrow \mathbb{R}^n \mid y(kh + \tau) = y(kh) \quad \forall k \in \mathbb{Z}, \tau \in [0, h)\}.$$

The elements in this codomain are piecewise constant functions (but $\mathbb{Y} = \mathfrak{F}(\mathbb{R}, \mathbb{R}^n)$ suffices as codomain as well). □

4.3 Linear transformation on $\mathbb{R}^n$

On $\mathbb{R}^n$ linear mappings are often identified with matrices.

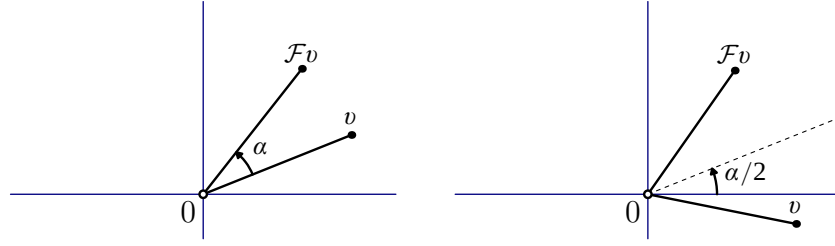


Figure 4.2: Rotation and reflection

Example 4.3.1 (Rotation in $\mathbb{R}^2$). Figure 4.2(a) illustrates the *rotation* $\mathcal{F} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ operator. It rotates its argument over an angle of α (counter clockwise). It is a linear mapping (verify this). In particular it maps the unit vector $e_1 := \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ to $y_1 := \begin{bmatrix} \cos(\alpha) \\ \sin(\alpha) \end{bmatrix}$ and the unit vector $e_2 := \begin{bmatrix} 0 \\ 1 \end{bmatrix}$ to $y_2 := \begin{bmatrix} -\sin(\alpha) \\ \cos(\alpha) \end{bmatrix}$. Combining the two outcomes in a matrix

$$F_{\text{rotation}} := [y_1 \ y_2] = \begin{bmatrix} \cos(\alpha) & -\sin(\alpha) \\ \sin(\alpha) & \cos(\alpha) \end{bmatrix}$$

is the standard way of representing linear mappings from $\mathbb{R}^k$ to $\mathbb{R}^n$. Identifying linear mappings with their matrix has to do with the fact that the linear mapping is completely specified by its matrix. The drawback of such a matrix approach is that it assumes we all agree on what the “standard basis” is and while this may be so (well) in $\mathbb{R}^n$, for other vector spaces this may not be so obvious. $\square$

Example 4.3.2 (Reflection in $\mathbb{R}^2$). Figure 4.2(b) depicts the reflection transformation $\mathcal{F} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$. It reflects its argument with respect to the line with angle $\alpha/2$. The matrix F now becomes

$$F_{\text{reflection}} = \begin{bmatrix} \cos(\alpha) & \sin(\alpha) \\ \sin(\alpha) & -\cos(\alpha) \end{bmatrix}.$$

Note that this is an orthogonal matrix (see Lemma 2.10.1). $\square$

Example 4.3.3 (transformation on $\mathbb{R}^3$). Suppose we have a mapping $\mathcal{T}$ that we know to be linear and that sends the unit cube to a stretched version, see Fig. 4.3. The figure indicates that

$$\mathcal{T}e_1 = e_1, \quad \mathcal{T}e_2 = 2e_2, \quad \mathcal{T}e_3 = e_3.$$

This way the mapping (since assumed linear) is fully specified. The matrix T associated with this mapping (with respect to the standard basis) is

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 2 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Note that the volume of the transformed unit cube equals $|\det T|$ (in our case 2). This holds for any transformation $\mathcal{T}$. $\square$

4.4 Matrices as representations of operators

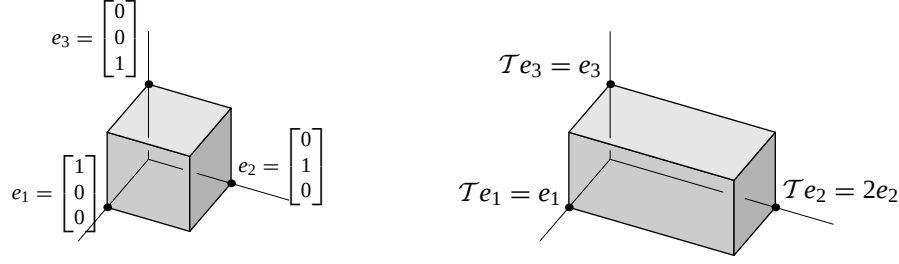


Figure 4.3: Unit cube linearly transformed

A message of the previous section is: once we settle on a “standard basis” then the linear mapping may be identified with “its matrix”. As mentioned earlier, a drawback of such an approach is that it assumes we all agree on “the basis.” On the other hand, an advantage is that it translates the linear mapping into a matrix of numbers which makes it explicit (matlabable). Since the linear mapping is fully specified by its matrix, any property that the linear mapping may have can in principle be verified from its matrix. For computation in MATLAB it is often needed to represent the linear mapping by matrices. Consistent with the previous section we define:

Definition 4.4.1 (Coordinates and matrices). Let $\mathbb{V}$ be a vector space with finite ordered basis $\{v_1, v_2, \dots, v_n\}$. For any $x \in \mathbb{V}$ let $x_{\#} \in \mathbb{C}^n$ denote the column vector of coordinates of x with respect to the basis, that is,

$$x = [v_1 \ v_2 \ \cdots \ v_n] x_{\#}.$$

For any linear transformation $\mathcal{F} : \mathbb{V} \rightarrow \mathbb{V}$ define the *matrix* $F_{\#\#} \in \mathbb{C}^{n \times n}$ of $\mathcal{F}$ with respect to the basis by the equality

$$[\mathcal{F}(v_1) \ \mathcal{F}(v_2) \ \cdots \ \mathcal{F}(v_n)] = [v_1 \ v_2 \ \cdots \ v_n] F_{\#\#}.$$

□

In other words, the ij th element of the matrix $F_{\#\#}$ of $\mathcal{F}$ is the contribution of basis element v_i in $\mathcal{F}(v_j)$. The matrix clearly depends on the choice of basis. This definition is consistent with that of the matrices of Examples 4.3.1 and 4.3.2, because of the following lemma. Roughly speaking the lemma says that linear transformations on finite dimensional vector space can be identified with matrices, much like what we commonly do in $\mathbb{R}^n$:

Lemma 4.4.2 (Matrix representation of linear transformation). Let $\mathbb{V}$ be vector space with finite ordered basis $\{v_1, \dots, v_n\}$. For any linear transformation $\mathcal{F} : \mathbb{V} \rightarrow \mathbb{V}$ we have

$$y = \mathcal{F}(x) \iff y_{\#} = F_{\#\#} x_{\#}.$$

Proof. By definition of $x_{\#}$ we have that $x = [v_1 \ \cdots \ v_n] x_{\#}$. Using linearity we get $\mathcal{F}(x) = \mathcal{F}([v_1 \ \cdots \ v_n] x_{\#}) = [\mathcal{F}(v_1) \ \cdots \ \mathcal{F}(v_n)] x_{\#} = [v_1 \ \cdots \ v_n] F_{\#\#} x_{\#}$. So $y = \mathcal{F}(x)$ iff $[v_1 \ \cdots \ v_n] y_{\#} = [v_1 \ \cdots \ v_n] F_{\#\#} x_{\#}$. As the $\{v_1, \dots, v_n\}$ are linearly independent this last equality holds iff $y_{\#} = F_{\#\#} x_{\#}$. ■

Lemma 4.4.3 (Eigenvalue and eigenvector). Let $\lambda \in \mathbb{C}$. Consider a linear mapping $\mathcal{F} : \mathbb{V} \rightarrow \mathbb{V}$ and let $F_{\#\#}$ be the matrix associated with this mapping (given some basis of $\mathbb{V}$). The following statements are equivalent.

1. There is a $v \in \mathbb{V}$, $v \neq 0$ such that $\mathcal{F}(v) = \lambda v$.

2. λ is an eigenvalue of the matrix $F_{\# \#}$.

Proof. Apply Lemma 4.4.2 for $y = \lambda x$. ■

The lemma implies that the eigenvalues of $F_{\# \#}$ do not depend on the choice of basis!

Example 4.4.4 (Differentiator). Consider the differentiator $\mathcal{D} : \mathbb{P}_n \rightarrow \mathbb{P}_n$ that sends polynomials $p \in \mathbb{P}_n$ of degree at most n to their derivative $\mathcal{D}(p)$. A basis for $\mathbb{P}_n$ clearly is

$$\{1, t, t^2, \dots, t^n\}$$

and they map to

$$\{0, 1, 2t, \dots, nt^{n-1}\}.$$

With respect to this basis the matrix that represents the differentiator on $\mathbb{P}_n$ can be derived from

$$\begin{aligned} [\mathcal{D}(1) \quad \mathcal{D}(t) \quad \mathcal{D}(t^2) \quad \dots \quad \mathcal{D}(t^n)] &= [0 \quad 1 \quad 2t \quad \dots \quad nt^{n-1}] \\ &= \begin{bmatrix} 1 & t & t^2 & \dots & t^n \end{bmatrix} \begin{bmatrix} 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 2 & \ddots & \vdots \\ 0 & 0 & \ddots & \ddots & 0 \\ \vdots & \ddots & \ddots & \ddots & n \\ 0 & \dots & \dots & \dots & 0 \end{bmatrix}. \end{aligned}$$

The matrix (the one on the far right in the above equation) is not invertible hence neither is the differentiator. Indeed the differentiator is not invertible since $\mathcal{D}(1) = 0$. The only eigenvalue that the matrix has is $\lambda = 0$ hence the differentiator has no eigenvalues other than $\lambda = 0$. Indeed, the derivative of any polynomial is of lower degree so nontrivial eigenfunctions do not exist. The eigenfunctions with eigenvalue 0 are the constant functions.

If we would have chosen as domain $\mathbb{V} = \text{span}\{e^t, e^{2t}\}$ with obvious basis $\{e^t, e^{2t}\}$ then the matrix of the differentiator would have been $\begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}$ because

$$[\mathcal{D}(e^t) \quad \mathcal{D}(e^{2t})] = [e^t \quad 2e^{2t}] = \begin{bmatrix} e^t & e^{2t} \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 2 \end{bmatrix}.$$

Now the matrix is invertible, hence the differentiator is invertible on $\text{span}\{e^t, e^{2t}\}$, indeed it is. Also, its eigenvalues are 1 and 2 hence $f \in \text{span}\{e^t, e^{2t}\}$ exist with $\mathcal{D}(f) = f$ and $\mathcal{D}(f) = 2f$, indeed such f exist. □

The connection between mappings and their matrix representation can be quite useful.

Example 4.4.5 (Eigenfunctions of polynomials). Consider the mapping $\mathcal{F} : \mathbb{P}_2 \rightarrow \mathbb{P}_2$ defined as

$$g = \mathcal{F}(p) : \quad g(t) = p(1 - t).$$

The graph $(t, g(t))$ hence is the graph $(t, p(t))$ reflected in the vertical $t = 1/2$ -axis. With respect to the standard basis $\{1, t, t^2\}$ the matrix becomes

$$[\mathcal{F}(1) \quad \mathcal{F}(t) \quad \mathcal{F}(t^2)] = [1 \quad 1 - t \quad (1 - t)^2] = \begin{bmatrix} 1 & t & t^2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & -1 & -2 \\ 0 & 0 & 1 \end{bmatrix}.$$

Because of the upper-triangular structure the eigenvalues are the diagonal elements of the matrix: $\lambda = 1, -1, 1$. It is readily verified that the eigenvectors (modulo scaling etc.) are

$$\lambda = 1 : v_1 = \begin{bmatrix} 1 \\ 0 \\ 0 \end{bmatrix}, v_{1'} = \begin{bmatrix} 0 \\ -1 \\ 1 \end{bmatrix} \quad \text{and} \quad \lambda = -1 : v_{-1} = \begin{bmatrix} -1 \\ 2 \\ 0 \end{bmatrix}.$$

This corresponds to the eigenfunctions

$$p_1 = 1, \quad p_{1'}(t) = t^2 - t, \quad \text{and} \quad p_{-1}(t) = 2t - 1.$$

See Fig. 4.4. Since the eigenvector v_{-1} for $\lambda = -1$ is unique (up to scaling) the eigenfunction p_{-1} with eigenvalue -1 is unique as well (up to scaling). The eigenfunctions with eigenvalue 1 are the linear combinations of p_1 and $p_{1'}$. $\square$

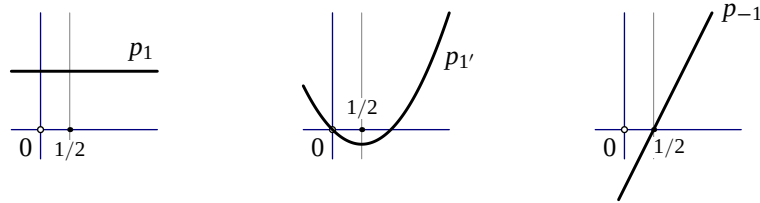


Figure 4.4: Three eigenfunctions (Example 4.4.5)

4.5 If you want to know more

- Let $\mathcal{F} : \mathbb{V} \rightarrow \mathbb{W}$ be a linear transformation on vector space, then $\text{im}(\mathcal{F})$ and $\text{ker}(\mathcal{F})$ are subspaces and we have

$$\dim(\text{ker } \mathcal{F}) + \dim(\mathcal{F}\mathbb{V}) = \dim(\mathbb{V}).$$

- If $\mathbb{V} = \mathbb{W}$ and if they have finite dimension, then the above says that $\mathcal{F}$ is surjective iff it is injective.
- This implies that there is a unique polynomial $p(t)$ of degree n that interpolates $n + 1$ pairs (t_i, p_i) , $t_i \neq t_j$.
- A sets of eigenvectors all with different eigenvalues is linearly independent. In particular on finite dimensional $\mathbb{V}$ a mapping $\mathcal{F}$ has at most finitely many different eigenvalues.
- http://en.wikipedia.org/wiki/Linear_transformation

4.6 Problems

4.1 Determine which of the following mappings are linear:

- a) $\mathcal{F} : \mathbb{R} \rightarrow \mathbb{R}; \mathcal{F}(t) = 3t + 1$.

- b) $\mathcal{A} : \mathbb{P}_n \rightarrow \mathbb{R}; \mathcal{F}(p) = p^{(1)}(3)$.
- c) $\mathcal{B} : \mathbb{P} \rightarrow \mathbb{P}; \mathcal{B}(p) = p^{(1)}$.
- d) $\mathcal{G} : \mathbb{C}^n \rightarrow \mathbb{C} : \mathcal{G}(x) = a'x$ (for some given $a \in \mathbb{C}^n$).

4.2 Let $B \in \mathbb{C}^{n \times n}$. Show that the mapping $\mathcal{L} : \mathbb{C}^{n \times n} \rightarrow \mathbb{C}^{n \times n}$ defined as $\mathcal{L}(A) = AB - BA$ is linear.

4.3 Prove Lemma 4.4.3.

4.4 Consider the linear transformation $\mathcal{F} : \mathbb{P}_1 \rightarrow \mathbb{P}_1$,

$$\mathcal{F}(\alpha_0 + \alpha_1 t) = \alpha_0 + (8\alpha_0 - \alpha_1)t.$$

- a) Determine the matrix of $\mathcal{F}$ with respect to the standard basis of $\mathbb{P}_1$.
 - b) Determine the matrix of $\mathcal{F}$ with respect to basis $\{t + 1, t - 1\}$.
 - c) Determine the eigenvalues of the above two matrices.
 - d) Determine the eigenvalues of $\mathcal{F}$ without using the matrices.
- 4.5 Let $C([a, b], \mathbb{R})$ denote the set of continuous functions in $\mathfrak{F}([a, b], \mathbb{R})$. Is the integral operator $\mathcal{I} : C[a, b] \rightarrow C[a, b]$ defined as

$$f = \mathcal{I}(g) : \quad f(t) = \int_a^t g(\tau) d\tau$$

linear?

4.6 Let $\delta \in \ell_{\mathbb{C}}(\mathbb{N})$. Show that $\Delta : \ell_{\mathbb{C}}(\mathbb{N}) \rightarrow \ell_{\mathbb{C}}(\mathbb{N})$ defined below is linear:

$$f = \Delta(g) : \quad f_k = \delta_k g_k.$$

4.7 Let $\mathbb{V}$ be a subspace of $\mathbb{R}^n$. Show that the orthogonal projection from x to its best approximation v_* (Thm. 2.6.2) is linear and do this without using normal equations (because without it the proof is more elegant).

4.8 Assume $\mathcal{F}$ is linear. Show that for any $m \in \mathbb{N}$:

$$\mathcal{F}(a_1 v_1 + a_2 v_2 + \cdots + a_m v_m) = a_1 \mathcal{F}(v_1) + a_2 \mathcal{F}(v_2) + \cdots + a_m \mathcal{F}(v_m).$$

4.9 Consider the mapping $\mathcal{F} : \mathbb{P}_2 \rightarrow \mathbb{P}_2$ defined as¹ $\mathcal{F}(p(t)) = p(-t)$.

- a) Find a basis of $\mathbb{P}_2$
 - b) Determine the matrix $F_{\#\#}$ of the mapping with respect to this basis
 - c) Find the eigenvalues and eigenvectors of F
 - d) Find the eigenvalues λ and eigenfunctions $p \in \mathbb{P}_2$ of $\mathcal{F}$.
- 4.10 Repeat the previous question but now for mapping $\mathcal{F}(p(t)) = p(t + 1)$.
- 4.11 Figure 4.5 shows the effect of four linear mappings $\mathcal{F}_1, \mathcal{F}_2, \mathcal{F}_3, \mathcal{F}_4$. Approximately determine the four 2×2 matrices F_1, F_2, F_3, F_4 .

¹Since $\mathcal{F}$ acts on functions the notation $\mathcal{F}(p(t))$ may seem awkward, but it is clear anyway and avoids more cumbersome forms like $\mathcal{F}(p)(t)$

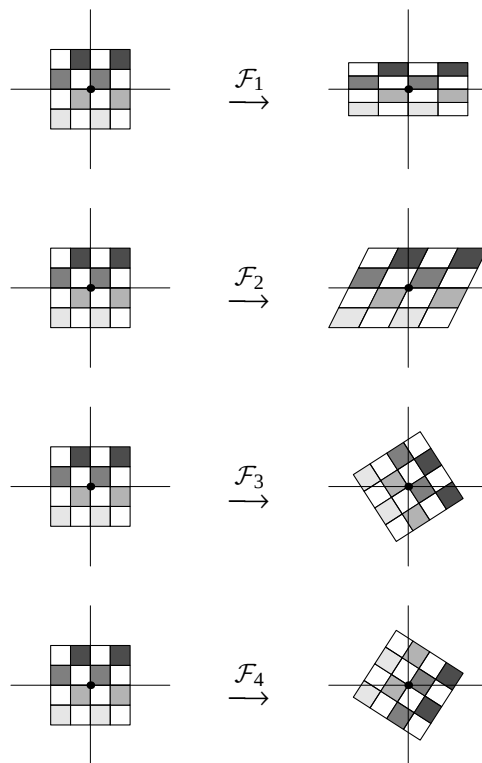


Figure 4.5: Four transformations on $\mathbb{R}^2$

5 Inner product, norm and orthogonal projection

Next we generalize the remaining two players in the projection theorem:

- the dot product (to be called *inner product*)
- length or Euclidean norm (to be called *norm*)

Then, finally, we can formulate a full fledged version of the projection theorem.

5.1 Real inner product

Let us see what we used in the proof of Theorem 2.6.2: (a) $\mathbb{V}$ was assumed to be a *subspace* of the *vector space* $\mathbb{R}^n$; (b) of the dot product we used that $(\alpha x + z) \cdot y = \alpha(x \cdot y) + (z \cdot y)$ and that $x \cdot y = y \cdot x$ and of course that $x \cdot x \geq 0$ with $x \cdot x = 0$ iff $x = 0$. *These we now simply take as the definition of inner product on real vector space:*

Definition 5.1.1 (Real inner product). Let $\mathbb{V}$ be a real vector space. A functional $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{R}$ is a (real) *inner product* if

1. $\langle x, y \rangle = \langle y, x \rangle$, *reflexive*
2. $\langle \alpha x + z, y \rangle = \alpha \langle x, y \rangle + \langle z, y \rangle$ for every $\alpha \in \mathbb{R}$, *linear in first argument*
3. $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0 \iff x = 0$. *positive definite*

□

That's all. Once we settle on this definition we are forced to conclude that there are many inner products on $\mathbb{R}^n$ other than the dot product. Indeed we could equip $\mathbb{R}^2$ for instance with:

Example 5.1.2. Consider as inner product on $\mathbb{R}^2$

$$\langle x, y \rangle := x_1 y_1 + 3x_2 y_2.$$

It is an inner product because

1. $\langle x, y \rangle = x_1 y_1 + 3x_2 y_2 = y_1 x_1 + 3y_2 x_2 = \langle y, x \rangle$,
2. $\langle x + \alpha z, y \rangle = (x_1 + \alpha z_1)y_1 + 3(x_2 + \alpha z_2)y_2 = (x_1 y_1 + 3y_2 x_2) + \alpha(z_1 y_1 + 3z_2 y_2) = \langle x, y \rangle + \alpha \langle z, y \rangle$,
3. $\langle x, x \rangle = x_1^2 + 3x_2^2 \geq 0$ and it equals zero iff $x_1 = x_2 = 0$ i.e. iff $x = 0$.

□

5.2 Complex inner products

There is one little further generalization that is useful: extension to complex numbers (and functions and vectors). An inner product is now a functional that maps to the complex numbers.

Since we want to maintain $\langle x, x \rangle \geq 0$ we should guarantee that $\langle x, x \rangle$ is a real number. This is automatic if we impose that $\langle x, y \rangle = \overline{\langle y, x \rangle}$.

Definition 5.2.1 (Inner product on a complex vector space). Let $\mathbb{V}$ be a complex vector space. A functional $\langle \cdot, \cdot \rangle : \mathbb{V} \times \mathbb{V} \rightarrow \mathbb{C}$ is a (complex) inner product if

1. $\langle x, y \rangle = \overline{\langle y, x \rangle}$, *conjugate reflexive*
2. $\langle \alpha x + z, y \rangle = \alpha \langle x, y \rangle + \langle z, y \rangle$ for every $\alpha \in \mathbb{C}$, *linear in first argument*
3. $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0 \iff x = 0$. *positive definite*

□

5.3 Norm

Definition 5.3.1 (norm associated with inner product). The norm $\|x\|$ on a vector space with inner product is defined as $\|x\| = \sqrt{\langle x, x \rangle}$. □

So the norm depends on the choice of inner product: first we settle on an inner product and then the norm follows.

Lemma 5.3.2 (Three properties of norm). Let $\mathbb{V}$ be a real or complex vector space with inner product. The norm $\|\cdot\|$ associated with the inner product has the following properties

1. $\|\alpha x\| = |\alpha| \|x\|$ for every $x \in \mathbb{V}$ and scalar α *homogeneous*
2. $\|x + y\| \leq \|x\| + \|y\|$ *triangle inequality*
3. $\|x\| \geq 0$, and $\|x\| = 0 \iff x = 0$. *positive definite*

□

Proof. The first and third are obvious. The second (triangle inequality) follows from Cauchy-Schwarz' inequality, see Section 5.6. ■

The triangle inequality loosely speaking says that traveling from 0 to $x + y$ via x or y can only mean a detour.

5.4 Orthogonal complement

Obviously $S^\perp$ is used to denote the set of all vectors perpendicular to all elements in S . The orthogonal complement is thus dependent on the choice of inner product.

Example 5.4.1. Consider again the inner product of Example 5.1.2 on $\mathbb{R}^2$:

$$\langle x, y \rangle := x_1 y_1 + 3x_2 y_2.$$

The boundary of the unit ball $\{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}$ is not a circle but an ellipse

$$\{x \in \mathbb{R}^2 \mid \|x\| \leq 1\} = \{x \in \mathbb{R}^2 \mid x_1^2 + 3x_2^2 \leq 1\}$$

see Fig. 5.1(a), and the orthogonal complement of $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$ is now

$$\begin{bmatrix} 1 \\ -1 \end{bmatrix}^\perp = \{x \mid \langle \begin{bmatrix} 1 \\ -1 \end{bmatrix}, x \rangle = x_1 - 3x_2 = 0\} = \text{span} \begin{bmatrix} 3 \\ 1 \end{bmatrix},$$

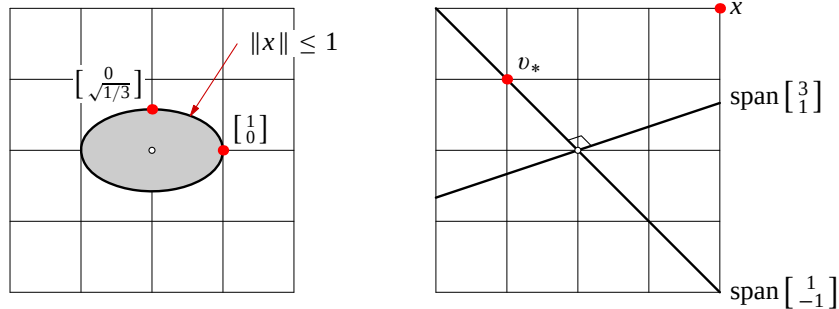


Figure 5.1: (a) Unit ball for inner product $\langle x, y \rangle = x_1 y_1 + 3x_2 y_2$; (b) two orthogonal vectors in this inner product

see Fig. 5.1(b). Orthogonality with respect to our inner product here does not mean “having an angle of 90 degree”.

It is instructive at this point to compute the best approximation v_* in $\mathbb{V} = \text{span}[\frac{1}{-1}]$ of $x = [\frac{2}{2}]$, see Fig. 5.1(b). Express candidate v_* as $\alpha[\frac{1}{-1}]$. Then

$$\frac{\partial}{\partial \alpha} \|x - v_*\|^2 = \frac{\partial}{\partial \alpha} \left\| \begin{bmatrix} 2-\alpha \\ 2+\alpha \end{bmatrix} \right\|^2 = \frac{\partial}{\partial \alpha} [(2-\alpha)^2 + 3(2+\alpha)^2] = 2(4\alpha + 4).$$

This is zero iff $\alpha = -1$ making $v_* = -1[\frac{1}{-1}] = [\frac{-1}{1}]$ the best approximation, see Fig. 5.1(b). Of course we should be able to compute v_* more efficiently and in greater generality with the projection theorem. Soon. $\square$

5.5 Pythagoras

Pythagoras’ theorem remains valid on arbitrary inner product space:

Lemma 5.5.1 (Pythagoras). $x \perp y$ implies $\|x + y\|^2 = \|x\|^2 + \|y\|^2$. $\square$

Proof. If $\langle x, y \rangle = 0$ then

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x + y \rangle + \langle y, x + y \rangle \\ &= \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\ &= \|x\|^2 + 2\text{Re}(\langle x, y \rangle) + \|y\|^2 = \|x\|^2 + \|y\|^2. \end{aligned}$$

■

In fact on *real* vector spaces $x \perp y$ is *equivalent* to $\|x + y\|^2 = \|x\|^2 + \|y\|^2$, but on complex vector space it is not equivalent, see Problem 5.10.

5.6 Cauchy-Schwarz

In $\mathbb{R}^2$ the dot product $x \cdot y$ was shown to satisfy $x \cdot y = |x||y| \cos(\phi)$, see Lemma 1.4.1. Now we have the following.

Theorem 5.6.1 (Cauchy-Schwarz inequality). Any inner product satisfies

$$1. |\langle x, y \rangle| \leq \|x\| \|y\|, \quad \text{Cauchy-Schwarz inequality}$$

$$2. |\langle x, y \rangle| = \|x\| \|y\| \text{ iff } y = \lambda x \text{ or } x = \lambda y \text{ for some scalar } \lambda, \quad \text{aligned}$$

Here (as always) $\|\cdot\|$ is the norm associated with the inner product.

Proof. The result is trivial if $y = 0$. Suppose $y \neq 0$. For any scalar λ ,

$$0 \leq \|x - \lambda y\|^2 = \langle x - \lambda y, x - \lambda y \rangle = \|x\|^2 - \lambda \langle y, x \rangle - \overline{\lambda} \langle y, x \rangle + |\lambda|^2 \|y\|^2.$$

In particular the right-hand side is ≥ 0 for $\lambda = \overline{\langle y, x \rangle} / \|y\|^2$:

$$0 \leq \|x\|^2 - \frac{|\langle x, y \rangle|^2}{\|y\|^2}$$

which proves the Cauchy-Schwarz inequality. Equality occurs if and only if $\|x - \lambda y\| = 0$ i.e. iff $x = \lambda y$. ■

The Cauchy-Schwarz inequality is very important and shows up in many applications. A consequence of Cauchy-Schwarz is for instance the triangle inequality.

Lemma 5.6.2 (Triangle inequality). For any inner product and associated norm we have

$$\|x + y\| \leq \|x\| + \|y\|.$$

Proof.

$$\begin{aligned} \|x + y\|^2 &= \langle x + y, x + y \rangle \\ &= \langle x, x \rangle + \langle x, y \rangle + \langle y, x \rangle + \langle y, y \rangle \\ &\leq \langle x, x \rangle + |\langle x, y \rangle| + |\langle y, x \rangle| + \langle y, y \rangle \\ &\leq \|x\|^2 + 2\|x\| \|y\| + \|y\|^2 \\ &= (\|x\| + \|y\|)^2. \end{aligned}$$

■

It is intriguing that in this abstract and general setting the triangle inequality remains.

5.7 Examples of inner product

Example 5.7.1 (scaled $\mathbb{R}^n$). Let $T \in \mathbb{R}^{n \times n}$ be an invertible matrix. Then

$$\langle x, y \rangle_{T'T} := (Tx) \cdot (Ty) = y' T' T x \quad (5.1)$$

is an inner product on $\mathbb{R}^n$. In fact on $\mathbb{R}^n$ a functional is an inner product *iff* it is of the form (5.1) for some nonsingular T , see Problem 5.5. □

There is more to this example than meets the eye: a serious drawback of the dot product $x \cdot y$ is that it considers all components $x_1, x_2, \dots$ of a vector on equal footing. But what if x_1 is in meters and x_2 is in millimeters? We would like to have the notions of norm and orthogonality at our disposal irrespective of the scaling (choice of units) or more generally, irrespective of linear transformation of the components x_k , and our inner product allows exactly that (as opposed to the, in hind side, rather limited dot product).

Example 5.7.2 (Not $\ell_{\mathbb{R}}(\mathbb{N})$ but $\ell_{\mathbb{R}}^2(\mathbb{N})$). There is a little subtlety if we want to extend the standard inner product of $\mathbb{R}^n$ (that is, the dot product) to $\ell_{\mathbb{R}}(\mathbb{N})$. The natural candidate is

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

but the infinite sum is not always defined, take for instance $x = y = (1, 1, 1, \dots)$. The standard remedy is to allow this inner product only on the smaller set

$$\ell_{\mathbb{R}}^2(\mathbb{N}) := \{u \mid u = (u_1, u_2, \dots), \|u\|^2 = \sum_{k=1}^{\infty} |u_k|^2 < \infty\}.$$

This set is a vector space which can be proved using Cauchy-Schwarz: the set $\ell_{\mathbb{R}}^2(\mathbb{N})$ is closed under addition for if $u, v \in \ell_{\mathbb{R}}^2(\mathbb{N})$ then

$$\|u + v\| \leq \|u\| + \|v\| < \infty.$$

(That $\ell_{\mathbb{R}}^2(\mathbb{N})$ is closed under scaling and contains the zero element is easy.) It is therefore a subspace of vector space $\ell_{\mathbb{R}}(\mathbb{N})$ and hence is itself a vector space. $\square$

5.8 Inner product on function space: $\mathbb{L}^2$

What is the function space equivalent of $\ell_{\mathbb{R}}^2(\mathbb{N})$? It is the space $\mathbb{L}_{\mathbb{R}}^2([a, b])$:

Lemma 5.8.1 ($\mathbb{L}^2$). Let $a, b \in \mathbb{R}$, $a < b$ (including $a = -\infty$ and $b = \infty$). The set

$$\mathbb{L}_{\mathbb{R}}^2([a, b]) := \{f : [a, b] \rightarrow \mathbb{R} \mid \int_a^b f^2(t) dt < \infty\}$$

is a vector space and the functional

$$\langle f, g \rangle = \int_a^b g(t) f(t) dt \tag{5.2}$$

is an inner product on this vector space. $\square$

We should first fix an unsettling problem: part of the definition of inner product (and its norm) is that $\|f\| = 0 \iff f = 0$ but here that is not the case! For instance functions that are zero *almost* everywhere such as

$$f(t) = 0 \text{ if } t \neq 0, \text{ and } f(0) = 3$$

are not zero but do have zero norm. The easy way out of this problem is to simply *identify* every function f with zero norm with the zero function. That is not far fetched because if $\|f\| = 0$ then

$$\|f\|^2 = \int_a^b |f(t)|^2 dt = 0$$

implying that “ $f(t)$ is zero almost everywhere”. From now on we do not distinguish between functions f and g when their difference has norm zero:

$$f = g \text{ for } f, g \in \mathbb{L}_{\mathbb{R}}^2([a, b]) \text{ is considered to mean that } \|f - g\| = 0.$$

We leave it to the reader (Problem 5.17) that (5.2) is indeed an inner product (given our identification of zero-equivalent functions).

Once we have an inner product we have the notion of orthogonality and we can talk about projections: we are now able to project functions onto subspaces of functions and apparently there is such a thing as functions being orthogonal to one-another!

Example 5.8.2 (Orthogonal sinusoids). Consider $\mathbb{L}_{\mathbb{R}}^2([0, 2\pi])$. On this space sine and cosine are orthogonal,

$$\sin(t) \perp \cos(t).$$

Why? Because

$$\langle \sin(t), \cos(t) \rangle = \int_0^{2\pi} \cos(t) \sin(t) dt = \int_0^{2\pi} \frac{1}{2} \sin(2t) dt = \left. \frac{-\cos(2t)}{4} \right|_0^{2\pi} = 0.$$

See Fig. 5.2 for a “proof by picture” of orthogonality of these two functions. □



Figure 5.2: $\sin(t) \perp \cos(t)$ on $\mathbb{L}_{\mathbb{R}}^2([0, 2\pi])$

Example 5.8.3. On $\mathbb{L}_{\mathbb{R}}^2([0, 1])$ the norm of e^t is

$$\|e^t\| = \sqrt{\int_0^1 (e^t)^2 dt} = \sqrt{\int_0^1 e^{2t} dt} = \sqrt{\left. \frac{e^{2t}}{2} \right|_0^1} = \sqrt{\frac{e^2 - 1}{2}} \approx 1.787.$$

□

Example 5.8.4. Consider $\mathbb{L}_{\mathbb{R}}^2([0, 1])$ with two of its elements

$$p_0(t) = 1, \quad p_1(t) = t.$$

They are not orthogonal because

$$\langle p_0, p_1 \rangle = \int_0^1 p_1(t) p_0(t) dt = \int_0^1 t dt = \frac{1}{2} \neq 0.$$

If we subtract the mean of p_1 from p_1 ,

$$p_{1'}(t) := p_1(t) - \frac{1}{2} = t - \frac{1}{2}$$

then we end up with two functions $p_0, p_{1'}$ that are orthogonal

$$\langle p_0, p_{1'} \rangle = \int_0^1 t - \frac{1}{2} dt = 0.$$

The norms of p_0 and $p_{1'}$ are

$$\|p_0\| = 1, \quad \|p_{1'}\| = 2\sqrt{3}.$$

□

5.9 Vector valued functions

To vectorize an inner product is often straightforward. For instance for the functions mapping from $[a, b]$ to $\mathbb{R}^n$ (mind the n in $\mathbb{R}^n$) we can take as inner product

$$\langle f, g \rangle = \int_a^b g'(t) f(t) dt. \quad (5.3)$$

Here g' denotes the complex conjugate transpose so $g'(t) f(t)$ for each t is a number, hence $\langle f, g \rangle \in \mathbb{R}$ as it should (on real vector space). We denote the space in which the norm associated with this inner product is finite also with $\mathbb{L}^2$:

$$\mathbb{L}_{\mathbb{R}^n}^2([a, b]) = \{ f : [a, b] \rightarrow \mathbb{R}^n \mid \langle f, f \rangle < \infty \text{ defined in (5.3)} \}.$$

Of course this is just one of the many inner products for vector valued functions.

5.10 Orthogonal projection

As before an element $v \in \mathbb{V}$ is a *best approximation* in $\mathbb{V}$ of x if

$$\|x - v_*\| \leq \|x - v\| \quad \forall v \in \mathbb{V}.$$

Theorem 5.10.1 (A projection theorem). Let $\mathbb{X}$ be a vector space with inner product, $\mathbb{V}$ a subspace of $\mathbb{X}$. Then for any $x \in \mathbb{X}$

1. If a best approximation v_* in $\mathbb{V}$ of x exists then it is unique.
2. v_* is a best approximation in $\mathbb{V}$ of x iff $(x - v_*) \perp \mathbb{V}$.
3. The norm of the best approximation error $\|x - v_*\|$ equals $\sqrt{\|x\|^2 - \|v_*\|^2}$.

Proof. An easy way out would be to say that the proof is *exactly* like that of Thm. 2.6.2 because that is how we designed our notion of vector space, subspace and inner product, but since the projection theorem is so important – it arguably is the most important theorem in this course – we should re-do the proof and verify that our generalization indeed was successful.

If $(x - v_*) \perp \mathbb{V}$ then because of subspace properties and Pythagoras we have for any $v \in \mathbb{V}$,

$$\|x - v\|^2 = \|(x - v_*) - \underbrace{(v - v_*)}_{\in \mathbb{V}}\|^2 = \|x - v_*\|^2 + \|v - v_*\|^2.$$

Hence if $v \neq v_*$ the norm $x - v$ exceeds that of $x - v_*$ making v_* the unique best approximation.

Conversely, suppose $(x - v_*) \not\perp \mathbb{V}$. Then there is a $v \in \mathbb{V}$ such that $(x - v_*) \not\perp v$ i.e. such that $\langle x - v_*, v \rangle \neq 0$. In particular such $v \neq 0$. Then for any α ,

$$\begin{aligned} \|x - (v_* + \alpha v)\|^2 &= \|(x - v_*) - \alpha v\|^2 \\ &= \|x - v_*\|^2 - \langle x - v_*, \alpha v \rangle - \langle \alpha v, x - v_* \rangle + \|\alpha v\|^2 \\ &= \|x - v_*\|^2 - 2\operatorname{Re}[\bar{\alpha} \langle x - v_*, v \rangle] + |\alpha|^2 \|v\|^2. \end{aligned}$$

For $\alpha = \frac{\langle x - v_*, v \rangle}{\|v\|^2}$ this becomes

$$\|x - (v_* + \alpha v)\|^2 = \|x - v_*\|^2 - \frac{|\langle x - v_*, v \rangle|^2}{\|v\|^2} < \|x - v_*\|^2$$

so that v_* is not the best approximation. The equality $\|x - v_*\| = \sqrt{\|x\|^2 - \|v_*\|^2}$ is a restatement of Pythagoras: $\|x - v_*\|^2 + \|v_*\|^2 = \|x\|^2$. ■

Example 5.10.2. Consider Example 5.4.1 with $x = (2, 2)$ and $v_* = (-1, 1)$ as indicated in the figure. Clearly $x - v_* \perp \text{span} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ hence by the projection theorem v_* is the unique best approximation in $\text{span} \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ of x . Done. $\square$

For infinite dimensional subspace $\mathbb{V}$ it may happen — but not necessarily — that a best approximation does not exist:

Example 5.10.3 (No solution exists). Take $\mathbb{X} = \ell_{\mathbb{R}}^2(\mathbb{N})$ and the subset $\mathbb{V} = \ell_{\mathbb{R}}^{\text{finite}}(\mathbb{N})$ — see Example 3.3.3. There is no best approximation in $\ell_{\mathbb{R}}^{\text{finite}}(\mathbb{N})$ of

$$x = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \frac{1}{5}, \dots)$$

because it can be approximated arbitrarily well by $v_N := (1, \frac{1}{2}, \frac{1}{3}, \dots, \frac{1}{N}, 0, 0, \dots)$ but the approximation error cannot be zero because $x \notin \ell_{\mathbb{R}}^{\text{finite}}(\mathbb{N})$. $\square$

As we shall see next, for *finite* dimensional subspace $\mathbb{V}$ a best approximation always does exist and for many infinite dimensional subspaces as well.

5.11 Projection onto finite dimensional space – Gram matrix

Consider projecting x onto a finite dimensional space $\mathbb{V} = \text{span} \{v_1, \dots, v_n\}$. The candidate best approximation v_* in $\mathbb{V}$ we write as $v_* = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} y_*$ with $y_* \in \mathbb{C}^n$. By the projection theorem this problem is easily solved:

Lemma 5.11.1 (Projection onto finite dimensional subspace). On any vector space $\mathbb{X}$ with inner product, a finite set $\{v_1, v_2, \dots, v_n\}$ is linearly independent iff the *Gram-matrix* defined as

$$\begin{bmatrix} \langle v_1, v_1 \rangle & \langle v_1, v_2 \rangle & \cdots & \langle v_1, v_n \rangle \\ \langle v_2, v_1 \rangle & \langle v_2, v_2 \rangle & \cdots & \langle v_2, v_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_n, v_1 \rangle & \langle v_n, v_2 \rangle & \cdots & \langle v_n, v_n \rangle \end{bmatrix} \in \mathbb{C}^{n \times n} \quad (5.4)$$

is invertible (in fact positive definite). In that case the best approximation v_* in $\text{span} \{v_1, \dots, v_n\}$ of $x \in \mathbb{X}$ is

$$v_* = \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} y_*, \quad y_* = \begin{bmatrix} \langle v_1, v_1 \rangle & \langle v_1, v_2 \rangle & \cdots & \langle v_1, v_n \rangle \\ \langle v_2, v_1 \rangle & \langle v_2, v_2 \rangle & \cdots & \langle v_2, v_n \rangle \\ \vdots & \vdots & \ddots & \vdots \\ \langle v_n, v_1 \rangle & \langle v_n, v_2 \rangle & \cdots & \langle v_n, v_n \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle x, v_1 \rangle \\ \langle x, v_2 \rangle \\ \vdots \\ \langle x, v_n \rangle \end{bmatrix}.$$

Proof. By the projection theorem v_* is a best approximation of x iff:

$$\begin{aligned} x - v_* \perp \mathbb{V} &\iff x - \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} y_* \perp \text{span} \{v_1, \dots, v_n\} \\ &\iff x - \begin{bmatrix} v_1 & \cdots & v_n \end{bmatrix} y_* \perp v_k \quad \forall k = 1, 2, \dots, n \\ &\iff [\langle v_1, v_k \rangle \quad \cdots \quad \langle v_n, v_k \rangle] y_* = \langle x, v_k \rangle \quad \forall k = 1, 2, \dots, n \\ &\iff \begin{bmatrix} \langle v_1, v_1 \rangle & \langle v_1, v_2 \rangle & \cdots \\ \langle v_2, v_1 \rangle & \langle v_2, v_2 \rangle & \cdots \\ \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ \vdots \end{bmatrix} = \begin{bmatrix} \langle x, v_1 \rangle \\ \langle x, v_2 \rangle \\ \vdots \end{bmatrix}. \end{aligned} \quad (5.5)$$

Remains to show is that the Gram-matrix is positive definite iff $\{v_1, \dots, v_n\}$ is a linearly independent set. This follows from the fact that $\| [v_1 \ \dots \ v_n] y \|^2 = y' R y$ where R is the Gram-matrix. Hence $[v_1 \ \dots \ v_n] y \neq 0$ for any nonzero y iff $R > 0$. ■

Equation (5.5) is again called *normal equations* (for reasons that will become clear later). For $n = 1$ the normal equations simply state that

$$v_* = v_1 \frac{\langle x, v_1 \rangle}{\langle v_1, v_1 \rangle}.$$

Example 5.11.2. Let us redo the projection of Example 5.4.1, that is, we want find the best approximation in $\text{span}\{v_1\}$ for $v_1 = \begin{bmatrix} 1 \\ -1 \end{bmatrix}$ of $x = \begin{bmatrix} 2 \\ 2 \end{bmatrix}$ with the weird inner product. According to the normal equations we have as best approximation

$$v_* = v_1 \frac{\langle x, v_1 \rangle}{\langle v_1, v_1 \rangle} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \frac{2 \times 1 + 3 \times 2 \times (-1)}{1 + 3} = \begin{bmatrix} 1 \\ -1 \end{bmatrix} \frac{-4}{4} = \begin{bmatrix} -1 \\ 1 \end{bmatrix}.$$

□

Example 5.11.3. Consider $\mathbb{L}_{\mathbb{R}}^2([-1, 1])$ and let us project e^t onto the subspace of polynomials of degree at most 1:

$$\min_{y_1, y_2} \|e^t - (y_1 + y_2 t)\|_{\mathbb{L}^2}.$$

With respect to the obvious basis $\{1, t\}$ for $\mathbb{P}_1$ the normal equations become

$$\begin{bmatrix} \langle 1, 1 \rangle & \langle 1, t \rangle \\ \langle t, 1 \rangle & \langle t, t \rangle \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \langle e^t, 1 \rangle \\ \langle e^t, t \rangle \end{bmatrix}.$$

Its solution is (verify this):

$$y_1 = \frac{1}{2}(e - \frac{1}{e}), \quad y_2 = \frac{3}{e}.$$

The best linear approximation on $[-1, 1]$ of e^t by 1-degree polynomial hence is $p_*(t) = \frac{1}{2}(e - \frac{1}{e}) + \frac{3}{e}t$, see Fig. 5.3. □

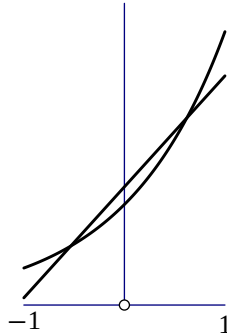


Figure 5.3: Approximation of e^t by degree-1 polynomial

5.12 Orthonormal sequences and Parseval

A convenient special case is if the Gram matrix (5.4) of inner products is the identity matrix.

Definition 5.12.1 (Orthonormal sequence). A sequence $\{e_1, e_2, \dots, e_n\}$ is an *orthonormal sequence* if all elements have unit norm and are mutually orthogonal,

$$\|e_k\| = 1 \quad \text{and} \quad e_k \perp e_j \quad \forall k \neq j.$$

□

A sequence is orthonormal iff the Gram matrix (5.4) is the identity matrix. Lemma 5.11.1 then specializes to the following elegant result:

Lemma 5.12.2. If $\{e_1, e_2, \dots, e_n\}$ is an orthonormal sequence then the best approximation in $\text{span}\{e_1, \dots, e_n\}$ of x is

$$v_* = \langle x, e_1 \rangle e_1 + \langle x, e_2 \rangle e_2 + \dots + \langle x, e_n \rangle e_n. \quad (5.6)$$

Moreover then $\|x - v_*\|^2 = \|x\|^2 - \sum_{k=1}^n |\langle x, e_k \rangle|^2$.

Proof. Follows from Lemma 5.11.1 and Pythagoras, $\|v_*\|^2 = \sum_{k=1}^n |\langle x, e_k \rangle|^2$. ■

The numbers $\langle x, e_k \rangle$ in (5.6) are the coordinates of v_* with respect to the basis $\{e_1, e_2, \dots\}$. If x happens to be an element of $\text{span}\{e_1, \dots, e_n\}$ then the minimal approximation error is obviously zero so then

$$\|x\|^2 = \sum_k |\langle x, e_k \rangle|^2. \quad (5.7)$$

This is known as *Parseval's identity* (which also holds more general, see Section 5.17) and it should remind you of the fact that in $\mathbb{R}^n$ the length of the vector can be expressed by its coordinates with respect to the standard (orthonormal) basis as $|x| = (x_1^2 + x_2^2 + \dots)^{\frac{1}{2}}$.

Example 5.12.3 (Legendre polynomials). Without proof we claim that the polynomials

$$p_0(t) = 1, \quad p_1(t) = t, \quad p_{k+1}(t) = \frac{(2k+1)tp_k(t) - kp_{k-1}(t)}{k+1}, \quad k > 1$$

defines an *orthogonal* sequence of polynomials on $\mathbb{L}_{\mathbb{R}}^2([-1, 1])$ with norm

$$\|p_k\| = \sqrt{\frac{2}{2k+1}}.$$

See Fig. 5.4. So then $e_k := \sqrt{\frac{2k+1}{2}}p_k$ is an orthonormal sequence. These are known as *Legendre polynomials*.

5.13 Adjoint operator

In $\mathbb{R}^n$ it turned out to be useful to be able to pull matrices M in the dot product from one side of the dot product to the other:

$$(Mx) \cdot z = x \cdot M'z.$$

(See Section 2.7 and the ensuing results). Given its use it makes sense to generalize this notion.

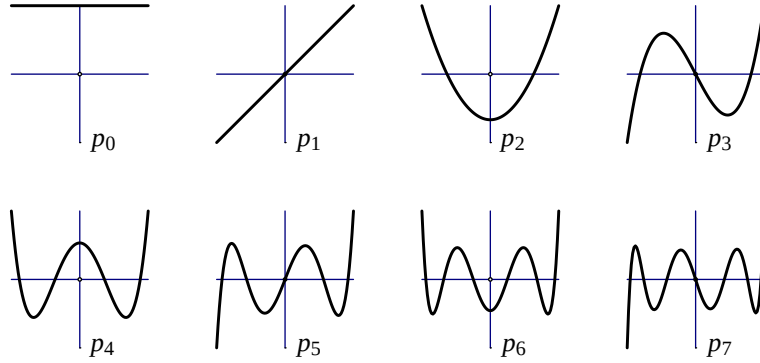


Figure 5.4: The first eight Legendre polynomials $p_k : [-1, 1] \rightarrow \mathbb{R}$. (not normalized: $\|p_k\|^2 = \frac{2}{2k+1}$).

Definition 5.13.1 (Adjoint). Let $\mathbb{V}$ and $\mathbb{W}$ be vector spaces (both real or both complex) and both with inner product. Let $\mathcal{A} : \mathbb{W} \rightarrow \mathbb{V}$ be a linear mapping. Then a linear mapping $\mathcal{A}^* : \mathbb{V} \rightarrow \mathbb{W}$ is said to be the *adjoint* of $\mathcal{A}$ if

$$\langle \mathcal{A}v, w \rangle_{\mathbb{W}} = \langle v, \mathcal{A}^*w \rangle_{\mathbb{V}}, \quad \forall v \in \mathbb{V}, w \in \mathbb{W}.$$

□

Given $\mathcal{A}$ the adjoint (if it exists) is unique so we can talk about *the* adjoint, see Problem 5.9.

Example 5.13.2 (Shift operator). Consider the linear shift operator $\mathcal{S} : \ell_{\mathbb{R}}^2(\mathbb{N}) \rightarrow \ell_{\mathbb{R}}^2(\mathbb{N})$,

$$y = \mathcal{S}(u) : \quad (y_1, y_2, y_3, \dots) = (0, u_1, u_2, \dots).$$

Given the standard inner product the adjoint of this mapping is

$$\tilde{u} = \mathcal{S}^*(\tilde{y}) : \quad (\tilde{u}_1, \tilde{u}_2, \tilde{u}_3, \dots) = (\tilde{y}_2, \tilde{y}_3, \dots).$$

Indeed:

$$\langle \mathcal{S}u, \tilde{y} \rangle = \sum_{k=1}^{\infty} u_k \tilde{y}_{k+1} = \langle u, \mathcal{S}^* \tilde{y} \rangle.$$

That is, the adjoint of the forward shift is the backward shift.

□

Example 5.13.3 (Integral operator). We determine the adjoint of $\mathcal{A} : \mathbb{L}_{\mathbb{R}n_u}^2(\mathbb{R}) \rightarrow \mathbb{L}_{\mathbb{R}n_y}^2(\mathbb{R})$,

$$y = \mathcal{A}(u) : \quad y(t) = \int_{-\infty}^{\infty} K(t, s)u(s) \, ds.$$

Note that for each t, s the matrix $K(t, s)$ is a $n_y \times n_u$ matrix. (Implicit here also is that the K is

such that it maps $\mathbb{L}^2$ signals to $\mathbb{L}^2$ signals¹.) Then

$$\begin{aligned}
\langle \mathcal{A}u, y \rangle &= \int_{t \in \mathbb{R}} y'(t) \left[\int_{s \in \mathbb{R}} K(t, s) u(s) ds \right] dt \\
&= \int_{t \in \mathbb{R}} \int_{s \in \mathbb{R}} [K'(t, s) y(t)]' u(s) ds dt \\
&= \int_{t \in \mathbb{R}} \int_{s \in \mathbb{R}} [K'(s, t) y(s)]' u(t) ds dt \\
&= \int_{t \in \mathbb{R}} \underbrace{\left[\int_{s \in \mathbb{R}} K'(s, t) y(s) ds \right]'}_{\mathcal{A}^* y :=} u(t) dt = \langle u, \mathcal{A}^* y \rangle.
\end{aligned}$$

That is, taking the adjoint corresponds to transforming the kernel as

$$K(t, s) \longrightarrow K'(s, t).$$

This is similar to the adjoint M' of matrices M , with the added fact that the two time indices need to be interchanged as well (which can also be understood as transpose). $\square$

Example 5.13.4 (From finite dimensional to infinite dimensional). Consider the linear mapping

$$\mathcal{P} : \mathbb{R}^2 \rightarrow \mathbb{L}_{\mathbb{R}}^2([0, 1]); \quad \mathcal{P}\left(\begin{bmatrix} a_1 \\ a_2 \end{bmatrix}\right) = a_1 + a_2 t$$

(with $\mathbb{R}^2$ and $\mathbb{L}^2$ equipped with standard inner product). Then

$$\langle \mathcal{P}a, f \rangle_{\mathbb{L}^2} = \int_0^1 (a_1 + a_2 t) f(t) dt = a_1 \langle f, 1 \rangle_{\mathbb{L}^2} + a_2 \langle f, t \rangle_{\mathbb{L}^2} =: \langle a, \mathcal{P}^* f \rangle_{\mathbb{R}^2}$$

The adjoint of $\mathcal{P}^* : \mathbb{L}_{\mathbb{R}}^2([0, 1]) \rightarrow \mathbb{R}^2$ hence maps functions to vectors as

$$b = \mathcal{P}^*(f) : \quad b = \begin{bmatrix} \langle f, 1 \rangle_{\mathbb{L}^2} \\ \langle f, t \rangle_{\mathbb{L}^2} \end{bmatrix}.$$

$\square$

5.14 Normal equations revisited

If we have the subspace $\mathbb{V} \subseteq \mathbb{X}$ given in the form

$$\mathbb{V} = \mathcal{A}\mathbb{Y}$$

for some vector space $\mathbb{Y}$ and linear operator $\mathcal{A} : \mathbb{Y} \rightarrow \mathbb{X}$ then the projection of $x \in \mathbb{X}$ onto this subspace $\mathbb{V}$ can be characterized again by the normal equations:

Lemma 5.14.1 (Normal equations). Let $\mathbb{X}$ and $\mathbb{Y}$ be a vector spaces with inner product. Let $x \in \mathbb{X}$ and assume $\mathcal{A} : \mathbb{Y} \rightarrow \mathbb{X}$ is linear and that the adjoint $\mathcal{A}^*$ exists. Then

$$y_* = \operatorname{argmin}_{y \in \mathbb{Y}} \|x - \mathcal{A}y\|$$

iff $y_* \in \mathbb{Y}$ satisfies

$$\mathcal{A}^* \mathcal{A} y_* = \mathcal{A}^* x. \tag{5.8}$$

¹sufficient is that $\int_{s, t \in \mathbb{R}} \operatorname{tr} K'(t, s) K(t, s) dt ds < \infty$

Proof. This is the projection theorem for $\mathbb{V} = \mathcal{A}\mathbb{Y}$ and $v_* = \mathcal{A}y_*$. By the projection theorem we need only establish that $(x - \mathcal{A}y_*) \perp \mathbb{V}$.

$$\begin{aligned}(x - \mathcal{A}y_*) \perp \mathbb{V} &= \langle x - \mathcal{A}y_*, \mathcal{A}y \rangle = 0 \quad \forall y \in \mathbb{Y} \\ &= \langle \mathcal{A}^*(x - \mathcal{A}y_*), y \rangle = 0 \quad \forall y \in \mathbb{Y} \\ &= \mathcal{A}^*(x - \mathcal{A}y_*) = 0.\end{aligned}$$

■

Since best approximations do not necessarily exist, the normal equations need not have a solution. Indeed:

Example 5.14.2 (No solution of normal equations exists). Consider again $\mathbb{X} = \ell_{\mathbb{R}}^2(\mathbb{N})$ and the subset $\mathbb{Y} = \ell_{\mathbb{R}}^{\text{finite}}(\mathbb{N})$ and let $\mathcal{A}$ be the identity mapping from $\mathbb{Y}$ to $\mathbb{X}$. Clearly $\mathcal{A}^*$ is identity mapping as well and the normal equations become

$$y_* = x.$$

However $\mathbb{Y} \neq \mathbb{X}$ so for some $x \in \mathbb{X}$ the normal equation has no solution $y_* \in \mathbb{Y}$. Indeed it does not for $x = (1, \frac{1}{2}, \frac{1}{3}, \frac{1}{4}, \dots)$, see Example 5.10.3. □

A pragmatic approach to approximation is that *if* the normal equations have a solution then the problem is solved. The Gram-matrix is a special case of the general normal equations:

Example 5.14.3 (Gram-matrix). Consider $\mathbb{L}_{\mathbb{R}}^2([-1, 1])$ and let us project e^t onto the subspace of polynomials of degree at most 1,

$$\min_{y_1, y_2} \|e^t - (y_1 + y_2 t)\|_{\mathbb{L}^2}.$$

This problem can be solved using Gram-matrices but for the sake of generality we will set up the problem using adjoints. Borrowing the operator $\mathcal{P}$ (and its adjoint) from Example 5.13.4, the subspace of polynomials $\mathbb{P}_1$ of degree at most 1 can be seen as the image,

$$\mathbb{P}_1 = \mathcal{P}(\mathbb{R}^2)$$

so the normal equations are

$$\mathcal{P}^* \mathcal{P} y_* = \mathcal{P}^* e^t.$$

These are equations in $\mathbb{R}^2$ (the domain of $\mathcal{P}$) and it readily verified that these are exactly the normal equations derived earlier in Example 5.11.3:

$$\begin{bmatrix} \langle 1, 1 \rangle & \langle 1, t \rangle \\ \langle t, 1 \rangle & \langle t, t \rangle \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \begin{bmatrix} \langle e^t, 1 \rangle \\ \langle e^t, t \rangle \end{bmatrix}.$$

□

The normal equations discussed in this section may also apply to cases where $\mathbb{Y}$ is infinite dimensional. In that sense the normal equations here generalize the Gram-matrix theorem. Applications of the normal equations are discussed in following chapters.

5.15 Under determined equations

Theorem 2.6.3 from Chapter 2 and the related Lemma 2.9.2 carry over verbatim:

Theorem 5.15.1 (Projection theorem – minimum norm element of affine variety). Consider a vector space $\mathbb{X}$ with inner product and a subspace $\mathbb{V} \subseteq \mathbb{X}$ and $x \in \mathbb{X}$. A vector v_* is the minimal norm element of the affine variety $x + \mathbb{V}$ iff $v_* \perp \mathbb{V}$.

Proof. Do it yourself. ■

Lemma 5.15.2 (Minimal norm solution). Let $\mathbb{X}$ and $\mathbb{Y}$ be vector spaces, both with inner product. If linear $\mathcal{A} : \mathbb{Y} \rightarrow \mathbb{X}$ has adjoint, and if $\mathcal{A}\mathcal{A}^* : \mathbb{X} \rightarrow \mathbb{X}$ is invertible, then for every $x \in \mathbb{X}$ the unique minimal norm solution y_* of the equation $x = \mathcal{A}y$ exists and equals

$$y_* = \mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}x. \quad (5.9)$$

Moreover the minimal norm squared can be expressed in terms of x as

$$\|y_*\|_{\mathbb{Y}}^2 = \langle (\mathcal{A}\mathcal{A}^*)^{-1}x, x \rangle_{\mathbb{X}}. \quad (5.10)$$

Proof. This is Theorem 5.15.1 for $\mathbb{V} = \ker(\mathcal{A})$ and $v_* = y_*$. Clearly the y_* of (5.9) satisfies $\mathcal{A}y_* = \mathcal{A}\mathcal{A}^*(\mathcal{A}\mathcal{A}^*)^{-1}x = x$ and $y_* \perp \ker(\mathcal{A})$. Equation (5.10) follows easily. ■

The right-hand side of the expression (5.10) is a useful one because typically $\mathbb{X}$ is of lower dimension than $\mathbb{Y}$, see Chapter 10 for an extensive example.

5.16 Gram-Schmidt and the matrix of inner products

The generalization of QR -factorization is the Gram-Schmidt procedure, see Problem 5.3.

5.17 Hilbert space

In Chapter 2 we managed to find best approximations by solving normal equations and the normal equations could in principle always be formulated such that the inverses involved exist. Hence these projection problems are essentially solvable. On general subspace best approximations need not exist (Example 5.10.3). Here is an other example.

Example 5.17.1 (No best approximation exists). Consider $\mathbb{X}$ the set of bounded functions on $[-1, 1]$ and consider its subspace of continuous functions $\mathbb{V} = C([-1, 1], \mathbb{R})$. What is the best approximation of the unit step function

$$x(t) = \mathbb{1}(t) := \begin{cases} 1 & t \geq 0 \\ 0 & t < 0 \end{cases}.$$

No best approximation exist because the approximation error $\|x - v_*\|$ can not be zero (for then $v_* = x$ but $x \notin \mathbb{V}$) but it can be made arbitrarily close to zero. For instance these functions

$$v_n(t) = \frac{1}{\pi} \arctan(nt) + \frac{1}{2}$$

are continuous and satisfy $\lim_{n \rightarrow \infty} \|x - v_n\|_{\mathbb{L}^2([-1, 1])} = 0$. □

The problem of nonexistence can be traced back to the fact that Cauchy sequences in $\mathbb{V}$ such as the v_n in the above example need not have a limit in $\mathbb{V}$. What to do? How to rule out such undesirable problems? Here is the by now standard route. We mention this route here because it is really standard and powerful, but for this course you do not have to master the details:

- Assume we have a vector space $\mathbb{X}$ with inner product (hence norm). Then an infinite series $\sum_{k=1}^{\infty} \alpha_k v_k$ is said to *converge* to some $x \in \mathbb{X}$ if $\lim_{M \rightarrow \infty} \|x - \sum_{k=1}^M \alpha_k v_k\| = 0$.
- Check whether or not every Cauchy sequence in $\mathbb{X}$ has a limit in $\mathbb{X}$, that is,

$$\lim_{N \rightarrow \infty} \sup_{n, m > N} \|x_n - x_m\| = 0 \implies \exists x \in \mathbb{H} : \lim_{n \rightarrow \infty} \|x - x_n\| = 0.$$

Inner product spaces with this property are known as *Hilbert spaces*.

- Check whether or not the subspace $\mathbb{V}$ is a Hilbert space. (If $\mathbb{X}$ is a Hilbert space then $\mathbb{V}$ is Hilbert iff it is a *closed* subspace, meaning that the complement $\{x \in \mathbb{X} | x \notin \mathbb{V}\}$ is *open* meaning that for any x in this complement a small enough neighborhood does not contain any element of $\mathbb{V}$. A sufficient test for closedness of a subspace is that it have finite dimension.)
- Theorem: If $\mathbb{V}$ and $\mathbb{X}$ are Hilbert then for any $x \in \mathbb{X}$ there is a unique best approximation $v_* \in \mathbb{V}$. The theorem does *not* provide a means to construct best approximations (via normal equations or whatever) it only guarantees its existence.

This is all rather involved but for each inner product space the Hilbert space property needs only to be verified once (and we will not do it). The following inner product spaces are Hilbert spaces:

- $\mathbb{R}^n, \mathbb{C}^n$ in fact *any* finite dimensional inner product space is a Hilbert space;
- $\ell_{\mathbb{C}^n}^2(\mathbb{Z})$ and $\mathbb{L}_{\mathbb{C}^n}^2([a, b])$ including if $a = -\infty, b = \infty$

The Hilbert-space property of $\mathbb{L}_{\mathbb{C}^n}^2([a, b])$ did not come about easily. In fact given the standard Riemann-sum definition of integral, the space of square integrable functions is *not* a Hilbert space! This prompted mathematicians to actually wanting to redefine the notion of integral (different from Riemann sums). In the beginning of the 20th century issue was settled by the mathematician Henri Lebesgue. He came up with complicated notion of integral that did not suffer the annoying convergence problems that Riemann's definition suffered from. The *Lebesgue measure* was born and with respect to this measure the space $\mathbb{L}^2$ was shown to be a Hilbert space. The "L" in the notation $\mathbb{L}^2$ is in honor of its inventor Lebesgue. We, however, do not need to bother about the difference between Riemann and Lebesgue and we certainly are not going to treat measure theory here.

Another advantage of having a Hilbert space is that some more familiar results from $\mathbb{R}^n$ carry over. For instance this central result:

Theorem 5.17.2 (Complete orthonormal basis). Let $\mathbb{H}$ be a Hilbert space and let $\{e_1, e_2, \dots\}$ be an orthonormal sequence (possibly infinite) in $\mathbb{H}$. The following conditions are equivalent:

1. $x \perp e_k$ for all e_k iff $x = 0$ sequence is *complete*
2. $x = \sum_k \langle x, e_k \rangle e_k$ for every $x \in \mathbb{H}$ function equals its *Fourier series*
3. $\|x\|^2 = \sum_k |\langle x, e_k \rangle|^2$ for every $x \in \mathbb{H}$. *Parseval's identity*

□

Proof. (1) $\implies$ (2): any finite sum $x_N = \sum_{k=1}^N \langle x, e_k \rangle e_k$ is the best approximation of x in $\text{span}\{e_1, \dots, e_N\}$ hence by Thm. 5.10.1 the norm of the approximation x_N does not exceed that of x , so $\|\sum_{k=1}^N \langle x, e_k \rangle e_k\| \leq \|x\| < \infty$. By Pythagoras this implies that $\sum_{k=1}^N |\langle x, e_k \rangle|^2 \leq \|x\|^2 < \infty$ (for $N = \infty$ this inequality is known as *Bessel's inequality*). Consequently x_N is a Cauchy sequence because $\|x_N - x_M\|^2 = \sum_{N < k \leq M} |\langle x, e_k \rangle|^2 \rightarrow 0$ as $\min\{N, M\} \rightarrow \infty$.

By the Hilbert property then $x_\infty := \sum_{k=1}^{\infty} \langle x, e_k \rangle e_k$ exists in $\mathbb{H}$. This x_∞ equals x because we assume $\{e_1, \dots\}$ to be complete and

$$\langle x - x_\infty, e_n \rangle = \langle x, e_n \rangle - \sum_{k=1}^{\infty} \langle x, e_k \rangle \langle e_k, e_n \rangle = \langle x, e_n \rangle - \langle x, e_n \rangle = 0 \quad \forall n.$$

That (2) $\implies$ (3) is immediate from Pythagoras and orthonormality of $\{e_1, e_2, \dots\}$. We prove (3) $\implies$ (1) by contradiction: suppose $\{e_1, e_2, \dots\}$ is not complete, i.e. a nonzero x exists with $x \perp e_k \forall k$. Then $\|x\| > 0$ while $\sum_k |\langle x, e_k \rangle|^2 = 0$. This contradicts (3). ■

For finite dimensional space such as $\mathbb{R}^n$ the theorem is intuitive and says that the $\{e_1, e_2, \dots\}$ span the entire space iff the orthogonal complement of $\text{span}\{e_1, e_2, \dots\}$ is zero.

5.18 If you want to know more...

Some other snippets of results:

- $\ker \mathcal{A} = [\text{im } \mathcal{A}^*]^\perp$ and $\ker \mathcal{A}^* = [\text{im } \mathcal{A}]^\perp$.
- If v_* is best approximation in $\mathbb{V}$ then $x - v_*$ is a best approximation in $\mathbb{V}^\perp$. This can be useful if $\mathbb{V}^\perp$ is easier to work with (for instance if $\mathbb{V}^\perp$ is finite dimensional).
- For linear operators continuity is equivalent to being bounded (in operator norm).
- If e_i is an orthonormal set then $\sum a_i e_i$ is well defined (if Hilbert) iff $\sum |a_i|^2 < \infty$
- Riesz-Fréchet: If $\mathbb{X}$ is Hilbert then a functional $\mathcal{F} : \mathbb{X} \rightarrow \mathbb{R}$ is bounded (continuous) iff it is of the form $\mathcal{F}(x) = \langle x, y \rangle$ for some $y \in \mathbb{X}$.
- Any bounded linear operator from Hilbert space to Hilbert space has a unique adjoint and the adjoint is bounded as well.
- The notion of *norm* also exists outside the context of inner product: any real functional on some vector space $\mathbb{X}$ is defined to be a *norm* if it satisfies the three enumerated properties of Lemma 5.3.2.

And finally we should warn the reader about the unfortunate habit in the systems and control community to define inner products as the complex conjugate of the inner products used elsewhere. So in systems and control literature one may read $\langle f, g \rangle = \int \overline{f(t)} g(t) dt$ which on complex vector space is *not* linear in its first argument f , but linear in its second argument g .

5.19 Problems

5.1 Prove the *parallelogram law*: $\|u + v\|^2 + \|u - v\|^2 = 2\|u\|^2 + 2\|v\|^2$.

5.2 Verify that on $\mathbb{L}_{\mathbb{R}}^2([0, 1])$ with its standard inner product we have that $\langle t^k, e^t \rangle = e - k \langle t^{k-1}, e^t \rangle$ for any $k \in \mathbb{N}, k > 0$. Use this and MATLAB to find the optimal degree 10 polynomial $p(t)$ approximation of e^t and plot the error $e^t - p(t)$.

5.3 Let $\{v_1, v_2, \dots\}$ be a linearly independent set. Prove that the *Gram-Schmidt* procedure:

$$e_1 = \frac{v_1}{\|v_1\|}, \quad g_n = v_n - \sum_{k=1}^{n-1} \langle v_n, e_k \rangle e_k \quad e_n = \frac{g_n}{\|g_n\|},$$

produces an orthonormal set $\{e_1, e_2, \dots\}$ and that $\text{span}\{v_1, \dots, v_k\} = \text{span}\{e_1, \dots, e_k\}$ for each k .

5.4 Apply the Gram-Schmidt procedure (Problem 5.3) on $\{1, t, t^2\} \in \mathbb{L}_{\mathbb{R}}^2([0, 1])$ and sketch the graphs of the so obtained e_1, e_2, e_3 .

5.5 Show that on $\mathbb{R}^n$:

$$\langle x, y \rangle \text{ is an inner product } \iff \langle x, y \rangle = y' P x \text{ for some } P = T' T \text{ with } T \text{ nonsingular.}$$

5.6 Suppose $\langle x, y \rangle = 0$ for all $x \in \mathbb{X}$. Show that $y = 0$.

5.7 Is $\langle A, B \rangle := \text{tr } B' A$ an inner product on $\mathbb{R}^{n \times k}$?

5.8 Is it $\mathbb{S}^{\perp\perp} \subseteq \mathbb{S}$ or $\mathbb{S} \subseteq \mathbb{S}^{\perp\perp}$?

5.9 Show that the adjoint $\mathcal{A}^*$ (if it exists) of a mapping $\mathcal{A}$ is unique (Lemma 5.9).

5.10 Show that on real vector space $\|x + y\|^2 = \|x\|^2 + \|y\|^2$ iff $x \perp y$ and find a counter example to this equality for the complex vector space $\mathbb{C}$.

5.11 Show that a $c > 0$ exists such that

$$\left| \int_0^\pi f(t) \cos(t) dt \right| \leq c \sqrt{\int_0^\pi |f(t)|^2 dt} \quad \forall f$$

and determine the smallest c_* possible and a function $f(t)$ that achieves equality for this smallest c_* .

5.12 Consider the vector space of continuous functions defined on the unit circle,

$$f, g : \{z \in \mathbb{C} \mid |z| = 1\} \rightarrow \mathbb{C}.$$

Show that

$$\langle f, g \rangle = \frac{1}{2\pi i} \int_{|z|=1} \overline{g(z)} f(z) \frac{1}{z} dz \quad (\text{where } z \text{ encircles counter clockwise})$$

is an inner product.

5.13 Consider a set of continuously differentiable functions $[0, 1] \rightarrow \mathbb{R}$. Is

$$\langle f, g \rangle := \int_0^1 g(t) f(t) + g^{(1)}(t) f^{(1)}(t) dt$$

an inner product?

5.14 Prove Lemma 5.11.1.

5.15 Consider $\mathfrak{F}([0, 1], \mathbb{R})$. Does a best approximation of e^t in the space of polynomials of any degree $\mathbb{P}$ exist?

5.16 Does $2|\langle x, y \rangle| \leq \|x\|^2 + \|y\|^2$ hold for any inner product?

5.17 Prove that (5.2) is indeed an inner product on $\mathbb{L}_{\mathbb{R}}^2([a, b])$.

5.18 Consider the convolution mapping $\mathcal{C}_h : \mathbb{L}_{\mathbb{R}}^2(\mathbb{R}) \rightarrow \mathbb{L}_{\mathbb{R}}^2(\mathbb{R})$,

$$y = \mathcal{C}_h(u) : \quad y(t) = \int_{-\infty}^{\infty} h(\tau)u(t - \tau) \, d\tau$$

Determine the adjoint in integral form, that is, determine k :

$$u = \mathcal{C}_h^*(y) : \quad u(t) = \int_{-\infty}^{\infty} k(t)y(t - \tau) \, d\tau.$$

5.19 Write a MATLAB function that returns the solution $y \in \mathbb{R}^n$ of the optimization problem

$$\min_y y'Qy, \quad \text{subject to the } Ay = x$$

for given positive definite Q , matrix A and vector x . (You may assume that AA' is invertible.)

5.20 An extension of Parseval's equality is that for a complete orthonormal sequence on Hilbert space $\mathbb{H}$ we have for any $x, y \in \mathbb{H}$ that

$$\langle x, y \rangle = \sum_k \langle x, e_k \rangle \langle e_k, y \rangle.$$

Prove it.

6 Fourier

There are four classic fourier transformations (three, if you count differently) and they all have intriguing orthogonality properties. They can be categorized by their domain and range:

- From discrete to discrete (DFT, FFT),
- From continuous to discrete (Fourier series),
- From continuous to continuous (Fourier transform),
- From discrete to continuous (DTFT).

In this chapter we cover the first two in detail and the remaining two only briefly.

6.1 DFT — discrete to discrete Fourier transform

The n -point *discrete Fourier transform (DFT)* is the linear transformation $\mathcal{F}_{\text{dd}} : \mathbb{C}^n \rightarrow \mathbb{C}^n$ defined as

$$\hat{x} = Fx : \underbrace{\begin{bmatrix} \hat{x}_0 \\ \hat{x}_1 \\ \hat{x}_2 \\ \vdots \\ \hat{x}_{n-1} \end{bmatrix}}_{\hat{x}} = \underbrace{\begin{bmatrix} 1 & 1 & 1 & \cdots & 1 \\ 1 & b^1 & b^2 & \cdots & b^{n-1} \\ 1 & b^2 & b^4 & \cdots & b^{2(n-1)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & b^{n-1} & b^{2n-2} & \cdots & b^{(n-1)^2} \end{bmatrix}}_F \underbrace{\begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \\ x_{n-1} \end{bmatrix}}_x \quad (6.1)$$

Here F is the $n \times n$ matrix whose (k, n) th entry is equal to $b^{(k-1)(n-1)}$ with b the complex number

$$b = e^{-i2\pi/n}. \quad (6.2)$$

Note that the first index of x and $\hat{x}$ is chosen to be 0 instead of 1, and also note that for each $n \in \mathbb{N}$ we have a different DFT. The $n \times n$ matrix F is completely determined by b which in turn is completely determined by n . If we distribute n complex numbers uniformly on the unit circle with 1 being one of the numbers, then b is the first number below 1 and b^2 is the next in line and then b^3 , see Fig. 6.1. Repeating this process we eventually end up in $a := b^{n-1}$ and then, finally, back in $1 = b^n$. The number $a := b^{n-1}$ is

$$a = e^{+i2\pi/n}. \quad (6.3)$$

The numbers a and b are each-others inverse, $ab = 1$, and each-others complex conjugate: $\bar{a} = b$ and $\bar{b} = a$.

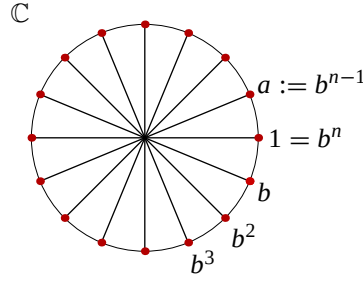


Figure 6.1: The unit circle equidistantly divided into $n = 16$ parts

The DFT has the property that it is invertible for any n and that its inverse is similar to the DFT. This follows from the fact that F is orthogonal (modulo a factor n):

$$F^*F = \begin{bmatrix} 1 & 1 & 1 & \cdots \\ 1 & a^1 & a^2 & \cdots \\ 1 & a^2 & a^4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & \cdots \\ 1 & b^1 & b^2 & \cdots \\ 1 & b^2 & b^4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} = \begin{bmatrix} n & 0 & 0 & \cdots \\ 0 & n & 0 & \cdots \\ 0 & 0 & n & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix}.$$

This is readily verified: the $(1, 1)$ element of F^*F is the product of the vectors $(1, 1, \dots, 1)$ and $(1, 1, \dots, 1)$ hence equals n . Likewise the $(2, 2)$ -entry is the product of $(1, a, a^2, \dots)$ and $(1, b, b^2, \dots)$ and hence becomes $1 + ab + a^2b^2 + \dots = 1 + 1 + 1 + \dots = n$. All diagonal entries similarly turn out to equal n . All non-diagonal entries of F^*F are zero. For instance the $(1, 2)$ -element is the product of the first row of F^* and the second column of F . That gives $1 + b + b^2 + \dots + b^{n-1}$. These are the n numbers on the unit circle and since they circle around zero they add up to zero, see Fig 6.1. Likewise the $(1, 3)$ -entry of the right-hand side is $1 + b^2 + b^4 + \dots + b^{2n-2}$. All these n numbers circle around zero as well, be it now twice as fast, nevertheless due to symmetry they again add up to zero. All non-diagonal entries are zero. The conclusion is

Lemma 6.1.1 (Inverse DFT). DFT (6.1) for any $n \in \mathbb{N}$ is invertible with inverse

$$x = F^{-1}\hat{x} : \begin{bmatrix} x_0 \\ x_1 \\ x_2 \\ \vdots \end{bmatrix} = \frac{1}{n} \begin{bmatrix} 1 & 1 & 1 & \cdots \\ 1 & a^1 & a^2 & \cdots \\ 1 & a^2 & a^4 & \cdots \\ \vdots & \vdots & \vdots & \ddots \end{bmatrix} \begin{bmatrix} \hat{x}_0 \\ \hat{x}_1 \\ \hat{x}_2 \\ \vdots \end{bmatrix} \quad (6.4)$$

where $a = 1/b = e^{i2\pi/n}$. □

Since $F^*F = nI$ we have for any $x \in \mathbb{C}^n$ that

$$|\hat{x}|^2 = \hat{x} \cdot \hat{x} = (Fx) \cdot (Fx) = (F^*Fx) \cdot x = n|x|^2.$$

This equality is known as Parseval's identity (though in preceding chapters we preserved Parseval's identity for orthonormal bases).

6.2 Harmonic decomposition — the DFT as a change of basis

Instead of interpreting the DFT as a linear transformation we can also interpret it as moving from one basis to the other. With the DFT we determine the coordinates $\hat{x}_k$ of the vector x , with

respect to the basis formed by the columns of F^{-1} ,

$$x = \frac{\hat{x}_0}{n} \begin{bmatrix} 1 \\ 1 \\ 1 \\ \vdots \end{bmatrix} + \frac{\hat{x}_1}{n} \begin{bmatrix} 1 \\ a \\ a^2 \\ \vdots \end{bmatrix} + \frac{\hat{x}_2}{n} \begin{bmatrix} 1 \\ a^2 \\ a^4 \\ \vdots \end{bmatrix} + \cdots + \frac{\hat{x}_{n-1}}{n} \begin{bmatrix} 1 \\ a^{n-1} \\ a^{2(n-1)} \\ \vdots \end{bmatrix}$$

and since F^{-1} is invertible (obviously) we know that the columns of F^{-1} are linearly independent and that they span the entire space $\mathbb{C}^n$, that is, any $x \in \mathbb{C}^n$ has such an expansion. In fact, since F^*F is diagonal we have that the columns of F^{-1} form an orthogonal basis of $\mathbb{C}^n$.

Figure 6.2 shows the DFT of two n -vectors for $n = 16$. The DFT spreads any x_m over all $\hat{x}_k$, $k = 1, \dots, n$. Inversely, as can be seen from Fig. 6.3, each $\hat{x}_k$ is distributed over all entries of $x = F^{-1}\hat{x}$. As the bottom two plots of Fig. 6.3 suggests, the DFT for real valued x_k obeys the symmetry property

$$\operatorname{Re} \hat{x}_k = \operatorname{Re} \hat{x}_{n-k}, \quad \operatorname{Im} \hat{x}_k = -\operatorname{Im} \hat{x}_{n-k} \quad \forall 0 < k < n$$

and that $\hat{x}_0$ is real.

6.3 Extra: the FFT

For general $n \times n$ matrices A and vectors $x \in \mathbb{R}^n$ it requires at least n^2 operations to form the product Ax because every of the n^2 elements A_{ij} affect the product. For the DFT the story is quite different. The point is that the matrix in an n -point DFT only has n different elements. Indeed, since $b^n = 1$ any term b^k is one of $\{1, b, \dots, b^{n-1}\}$. This gives the DFT-product Fx a lot of structure, a structure that can be exploited to significantly reduce the numbers of operations required. The resulting algorithm is known as the *fast fourier transform* or simply as *FFT*; a truly famous algorithm. It was invented in 1965 during the cold war by John Tukey, and was published together with James Cooley. It was immediately recognized as a breakthrough and soon computers appeared whose sole purpose was to compute FFTs. Among them the impressive, *Fourier Analyzer 5154A* of Hewlett-Packard.

It is interesting fact that the derivation of the FFT is best seen not from a matrix representation but from a summation representation,

$$\hat{x}_k = \sum_{m=0}^{n-1} e^{-i\frac{2\pi}{n}mk} x_m \quad k = 1, 2, \dots, n-1,$$

and its inverse

$$x_m = \frac{1}{n} \sum_{k=0}^{n-1} e^{i\frac{2\pi}{n}mk} \hat{x}_k \quad m = 1, 2, \dots, n-1.$$



Hewlett-Packard 5451A Fourier Analyzer, December 1972.

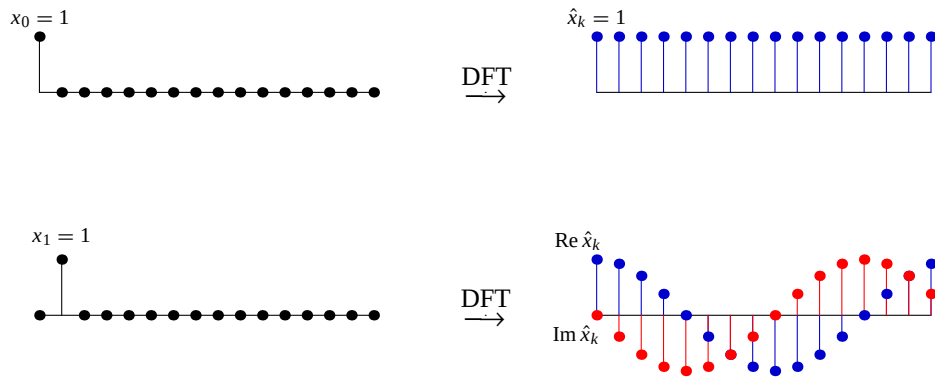


Figure 6.2: The data x_0 or x_1 is spread over all $\hat{x}_k$,
(blue = real part; red = imaginary part)

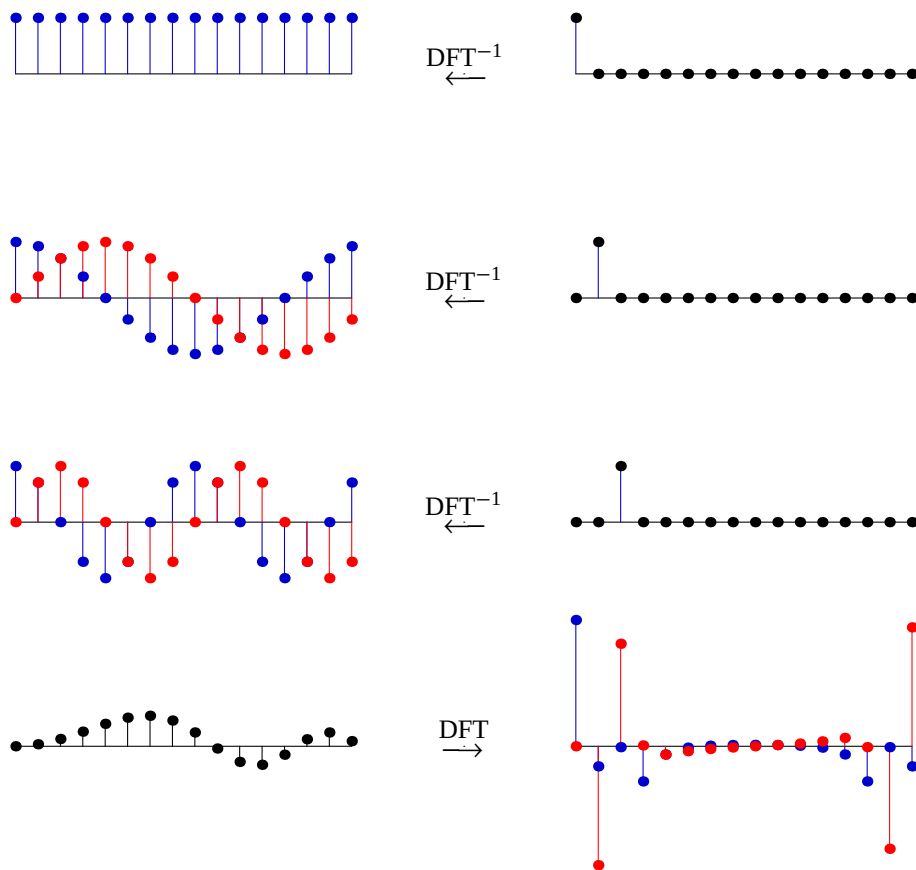


Figure 6.3: $\hat{x}_k$ is the coordinate of the $(k + 1)$ st column of F^{-1} ,
(blue = real part; red = imaginary part)

Lemma 6.3.1 (FFT). Suppose n is an integral power of 2: $n = 2^M$. Then the DFT can be computed in $\mathcal{O}(n \log(n))$ number of operations.

Proof. The crux of the FFT is that for all even $k = 2j$ we may rewrite (6.5) as

$$\hat{x}_{2j} = \sum_{m=0}^{n/2-1} (x_m + x_{m+n/2}) e^{-i \frac{2\pi}{n} m j}, \quad j = 0, 1, \dots, n/2 - 1$$

and for all odd indices $m = 2j + 1$,

$$\hat{x}_{2j+1} = \sum_{m=0}^{n/2-1} e^{-i \frac{2\pi}{n} m} (x_m - x_{m+n/2}) e^{-i \frac{2\pi}{n} m j}, \quad j = 0, 1, \dots, n/2 - 1.$$

Thus the n -point DFT (6.5) may be reduced to computation of two $n/2$ -point DFTs, one of

$$z_{\text{even},m} := x_m + x_{m+n/2}, \quad m = 0, 1, \dots, n/2 - 1$$

and one of

$$z_{\text{odd},m} := e^{-i \frac{2\pi}{n} m} (x_m - x_{m+n/2}), \quad m = 0, 1, \dots, n/2 - 1.$$

Forming the $z_{\text{even},m}$ from x_m takes $n/2$ (complex) additions. Forming $z_{\text{odd},m}$ takes besides additions also multiplications. Nowadays multiplication requires about as much computing time as addition, in any event, forming $z_{\text{even},m}$ and $z_{\text{odd},m}$ together takes dn units of operation for some constant $d > 0$. Now let $C(n)$ denote the number of operations needed to compute the n -point DFT. Using the above procedure we have that

$$C(n) = dn + 2C(n/2).$$

As $C(1) = 0$ we get that

$$C(n) = dn \log_2(n).$$

■

6.4 Fourier — continuous to discrete

It is time for some elegant math. We will not rely on the results of the previous section but we prove from scratch (well) an $\mathbb{L}^2$ version of the famous Fourier series. The story begins with the following easy result:

Lemma 6.4.1. On $\mathbb{L}^2_{\mathbb{C}}([0, 1])$ the doubly infinite sequence $\{\dots, e_{-1}, e_0, e_1, e_2, \dots\}$ of harmonic signals

$$e_k(t) := e^{i2\pi kt}$$

is an orthonormal set.

Proof. $\langle e_k, e_n \rangle = \int_0^1 e^{i2\pi(k-n)t} dt$ which is 0 if $k \neq n$ and is 1 if $k = n$. ■

So immediately we get that the best approximation f_* in $\text{span}\{\dots, e_{-1}, e_0, e_1, \dots\}$ of a function $f \in \mathbb{L}^2([0, 1])$ is

$$f_* = \sum_{k=-\infty}^{\infty} \langle f, e_k \rangle e_k,$$

that is,

$$f_*(t) = \sum_{k=-\infty}^{\infty} \hat{f}_k e^{i2\pi kt}$$

where we used the short hand

$$\hat{f}_k = \langle f, e_k \rangle = \int_0^1 e^{-i2\pi kt} f(t) dt.$$

The obvious question now is: is the Fourier series f_* of f equal to f ? Or, to put it differently, is the sequence $\{\dots, e_{-1}, e_0, e_1, \dots\}$ a *complete* orthonormal sequence in $\mathbb{L}_{\mathbb{C}}^2([0, 1])$? It is:

Lemma 6.4.2 (Classical Fourier series). Let $T > 0$. Every $f \in L_{\mathbb{R}}^2([0, T])$ equals its Fourier series,

$$f(t) = \sum_{k \in \mathbb{Z}} \hat{f}_k e^{ik \frac{2\pi}{T} t} \quad (6.6)$$

where $\hat{f}_k$ are the *Fourier coefficients* defined as

$$\hat{f}_k = \frac{1}{T} \langle f, e^{ik \frac{2\pi}{T} t} \rangle = \frac{1}{T} \int_0^T e^{-ik \frac{2\pi}{T} t} f(t) dt.$$

Proof. We prove the $T = 1$ case (other T follow from scaling). According to Theorem 5.17.2 we need only show that the only function orthogonal to all e_k is the zero function. Assume for the moment that f is continuous. If f is nonzero then f is > 0 (or < 0) throughout some small enough interval $[c - \Delta, c + \Delta]$, see Fig. 6.4. Take the harmonic function $h(t)$ of the form

$$h(t) := b + \frac{1}{2} \cos(2\pi t + a) = b + \frac{1}{4} (e^{i(2\pi t + a)} + e^{-i(2\pi t + a)})$$

where a, b are such that $h(c - \Delta) = h(c + \Delta) = 1$, and $h(t) > 1$ inside this interval, and < 1 elsewhere in $[0, 1]$, see Fig. 6.4. Clearly for large enough n we have

$$\langle f, h^n \rangle \neq 0.$$

The h^n is a linear combination of harmonic signals: $h^n \in \text{span}\{e_{-n}, \dots, e_n\}$ hence for at least one k we have $\langle f, e_k \rangle \neq 0$. Therefore the only continuous f that is orthogonal to all e_k is the zero function, $f = 0$.

Now let $f \in \mathbb{L}^2$, not necessarily continuous. If $f \perp e_k$ for all k then the integrated $F(t) := \int_0^t f(\tau) d\tau$ — which is continuous — is orthogonal to all e_k as well. Indeed, since $F(1) = F(0) + \int_0^1 f(t) dt = F(0) + \langle f, e_0 \rangle = F(0) = 0$ we get using integration by parts

$$\langle F, e_k \rangle = \int_0^1 e^{-i2\pi kt} F(t) dt = \left[\frac{e^{-i2\pi kt}}{-i2\pi k} F(t) \right]_0^1 - \frac{1}{-i2\pi k} \langle f, e_k \rangle = 0 \quad \forall k \neq 0.$$

So $F(t) - \langle F, e_0 \rangle e_0$ is orthogonal to all e_k , including $k = 0$. By continuity this implies that $F - \langle F, e_0 \rangle$ is the zero function, i.e. $F(t)$ is constant, $F(t) = F(0) = 0$. Its derivative f hence is the zero function¹, which is what we set out to prove. ■

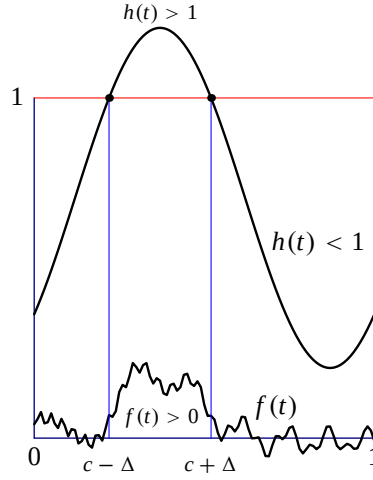


Figure 6.4: A particular cosine $h(t) = 1 + \frac{\cos(2\pi(t-c)) - \cos(2\pi\Delta)}{2}$

Be aware that equality in (6.6) is to be understood in $\mathbb{L}^2$ -sense. I.e. $f(t)$ and its Fourier series could differ at some values of t , as long as they are the same for “almost all t ”.

Parseval’s identity for our case becomes:

Lemma 6.4.3 (Parseval and power).

$$\frac{1}{T} \|f\|_{\mathbb{L}^2([0, T])}^2 = \sum_{k \in \mathbb{Z}} |\hat{f}_k|^2$$

for any $f \in \mathbb{L}^2([0, T])$ and corresponding Fourier coefficients $\hat{f}_k$. □

Proof. For $T = 1$ it follows immediately from Theorem 5.17.2. Other T follow from scaling. ■

For $f : [0, T] \rightarrow \mathbb{C}^n$, the quantity $\frac{1}{T} \|f\|_{\mathbb{L}^2[0, T]}^2$ is often referred to as the *power*. Since the chopped Fourier series

$$f_N(t) = \sum_{k=-N}^N \hat{f}_k e^{i \frac{2\pi}{T} kt}$$

is a projection of f — onto $\text{span}\{e_{-N}, \dots, e_N\}$ — its power never exceeds that of f and by Parseval we have an explicit description for the power loss

$$\frac{1}{T} \|f - f_N\|^2 = \sum_{|k| > N} |\hat{f}_k|^2.$$

We think of $|\hat{f}_k|^2 + |\hat{f}_{-k}|^2$ — for $k \neq 0$ — as the power of the signal at angular frequency $\frac{2\pi}{T}k$.

The Fourier series can also be interpreted as a linear transformation from functions to Fourier coefficients, $\mathcal{F}_{\text{cd}} : \mathbb{L}^2([0, T]) \rightarrow \ell^2(\mathbb{Z})$, $\hat{f} = \mathcal{F}_{\text{cd}}(f)$.

6.5 Fourier transform — discrete to continuous (DTFT)

¹It is a *theorem* that the derivative of the integral F of a Lebesgue integrable function f equals f almost everywhere!

Very briefly: fixing $T = 2\pi$ and interchanging the role of time and frequency readily gives us that the discrete-to-continuous Fourier transform $\hat{u} = \mathcal{F}_{\text{dc}}(u)$ is

$$\hat{u}(\omega) = \sum_{m=-\infty}^{\infty} u_m e^{-i\omega m}, \quad \omega \in [-\pi, \pi].$$

It is better known as the *discrete-time Fourier transform (DTFT)*. It transforms infinite (discrete-time) sequences to continuous-frequency functions and it is invertible for any $u \in \ell^2$ with inverse

$$u_m = \frac{1}{2\pi} \int_{-\pi}^{\pi} \hat{u}(\omega) e^{i\omega m} d\omega, \quad m \in \mathbb{Z}.$$

Parseval's identity becomes

$$\|u\|_{\ell^2(\mathbb{Z})} = \frac{1}{\sqrt{2\pi}} \|\hat{u}\|_{\mathbb{L}^2[-\pi, \pi]}$$

or, in terms of inner products,

$$\langle u, y \rangle_{\ell^2(\mathbb{Z})} = \frac{1}{2\pi} \langle \hat{u}, \hat{y} \rangle_{\mathbb{L}^2[-\pi, \pi]}.$$

Mapping $\frac{1}{\sqrt{2\pi}} \mathcal{F}_{\text{dc}} : \ell^2(\mathbb{Z}) \rightarrow \mathbb{L}^2[-\pi, \pi]$ is *unitary* (inner product preserving, linear, invertible).

6.6 Fourier transform — continuous to continuous

For $T \rightarrow \infty$ the Fourier series thickens to an integral.

Theorem 6.6.1. For any $u \in \mathbb{L}_{\mathbb{C}^n}^2(\mathbb{R})$ the Fourier transform $\hat{u} = \mathcal{F}_{\text{cc}}(u)$ defined as

$$\hat{u}(i\omega) = \int_{-\infty}^{\infty} u(t) e^{-i\omega t} dt, \quad \omega \in \mathbb{R} \quad (6.7)$$

is invertible, with

$$u(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{u}(i\omega) e^{i\omega t} d\omega, \quad t \in \mathbb{R}. \quad (6.8)$$

where equality is understood in $\mathbb{L}^2$ sense. The Fourier transform maps $\mathbb{L}_{\mathbb{C}^n}^2(\mathbb{R})$ to $\mathbb{L}_{\mathbb{C}^n}^2(i\mathbb{R})$.

Proof (sketch). Consider the restriction u_T of u to the time interval $[-\frac{T}{2}, \frac{T}{2}]$. The restriction u_T is in $\mathbb{L}^2([-\frac{T}{2}, \frac{T}{2}])$ and so equals its Fourier series,

$$u(t) = u_T(t) = \sum_{k \in \mathbb{Z}} \hat{u}_k e^{i\frac{2\pi}{T}kt} \quad \forall t \in [-\frac{T}{2}, \frac{T}{2}]$$

where $\hat{u}_k$ are the Fourier coefficients of u_T : $\hat{u}_k = \frac{1}{T} \int_{-T/2}^{T/2} u(t) e^{-i\frac{2\pi}{T}kt} dt$. Given that the Fourier transform of u_T is $\hat{u}_T(i\omega) = \int_{-T/2}^{T/2} u(t) e^{-i\omega t} dt$ we get $\hat{u}_k = \frac{1}{T} \hat{u}_T(i\frac{2\pi}{T}k)$, so

$$u(t) = \frac{1}{2\pi} \sum_{k \in \mathbb{Z}} \frac{2\pi}{T} \hat{u}_T(i\frac{2\pi}{T}k) e^{i\frac{2\pi}{T}kt} \quad \forall t \in [-\frac{T}{2}, \frac{T}{2}].$$

With the substitution $\Delta := \frac{2\pi}{T}$ we recognize this as a Riemann-sum approximation of integral,

$$u(t) = \frac{1}{2\pi} \sum_{k \in \mathbb{Z}} \Delta \hat{u}_T(ik\Delta) e^{i\Delta k t} \quad \forall t \in [-\frac{T}{2}, \frac{T}{2}]$$

Let $T \rightarrow \infty$ (hence $\Delta \rightarrow 0$) and the above sum thickens to the integral (6.8). That it maps $u \in \mathbb{L}^2_{\mathbb{C}^n}(\mathbb{R})$ to $\hat{u} \in \mathbb{L}^2_{\mathbb{C}^n}(i\mathbb{R})$ follows from Plancherel's theorem (done next). ■

In this context Parseval's identity is often called *Plancherel's theorem* after its inventor Michel Plancherel.

Lemma 6.6.2 (Plancherel, $\mathbb{L}^2$ -inner product in frequency domain). For any $u, y \in \mathbb{L}^2(\mathbb{R})$ and their Fourier transforms $\hat{u}, \hat{y}$ we have

$$\langle u, y \rangle_{\mathbb{L}^2(\mathbb{R})} = \frac{1}{2\pi} \langle \hat{u}, \hat{y} \rangle_{\mathbb{L}^2(i\mathbb{R})}$$

and hence for any $u \in \mathbb{L}^2(\mathbb{R})$,

$$\|u\|_{\mathbb{L}^2(\mathbb{R})} = \frac{1}{\sqrt{2\pi}} \|\hat{u}\|_{\mathbb{L}^2(i\mathbb{R})}. \quad (6.9)$$

Proof (sketch). The inverse Fourier transform says that $u(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \hat{u}(i\omega) e^{i\omega t} d\omega$. Therefore

$$\begin{aligned} 2\pi \langle u, y \rangle &= \int y^*(t) \int \hat{u}(i\omega) e^{i\omega t} d\omega dt = \iint (e^{-i\omega t} y(t))^* \hat{u}(i\omega) dt d\omega = \int \hat{y}(i\omega)^* \hat{u}(i\omega) d\omega \\ &= \langle \hat{u}, \hat{y} \rangle. \end{aligned}$$

(See the comment in Section 6.10 , or, on second thought, don't see this comment!) ■

So, up to a scaling factor, the Fourier transform from $\mathbb{L}^2(\mathbb{R})$ to $\mathbb{L}^2(i\mathbb{R})$ is a *unitary* transformation (inner-product preserving invertible linear transformation) and because of this strong connection we usually are ambiguous about which domain we are considering, time domain or frequency domain.

Example 6.6.3. The Fourier transform of $u(t) = e^{-t} \mathbb{1}_{[0, \infty)}(t)$ is

$$\hat{u}(i\omega) = \frac{1}{1 + i\omega}.$$

In time domain the norm (squared) computes as

$$\|u\|^2 = \int_{-\infty}^{\infty} u^2(t) dt = \int_0^{\infty} e^{-2t} dt = \frac{1}{2}$$

while in frequency domain we get

$$\frac{1}{2\pi} \|\hat{u}\|^2 = \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{1}{1 + \omega^2} d\omega = \frac{1}{2\pi} \arctan(\omega) \Big|_{-\infty}^{\infty} = \frac{1}{2}.$$

The two results are the same, as they should. □

6.7 Bilateral Laplace transform of $\mathbb{L}^2(\mathbb{R})$ -signals

The Fourier transform naturally extends to the bilateral Laplace transform. The *bilateral Laplace transform* also known as the *two-sided Laplace transform* is the linear mapping $f : \mathbb{R} \rightarrow \mathbb{V}$

$$F = \mathfrak{L}(f) : \quad F(s) = \int_{-\infty}^{\infty} f(t) e^{-st} dt$$

for those $s \in \mathbb{C}$ for which the integral is well defined. It is an extension of Fourier in that it may not only be defined on the imaginary axis, $s = i\omega$, but possibly for other $s \in \mathbb{C}$ as well. The (largest open?) set of $\mathbb{C}$ for which $F(s)$ is defined is called the *region of convergence (ROC)*.

Example 6.7.1 (Region of convergence).

1. The bilateral Laplace transform of the causal signal $f(t) = e^{-t} \mathbb{1}(t) = \text{⌋}$ is well defined for $\operatorname{Re} s > -1$,

$F(s) = \frac{1}{s+1},$


2. The bilateral Laplace transform of the anticausal signal $g(t) = e^t \mathbb{1}(-t) = \text{⏏}$ is well defined for $\operatorname{Re} s < 1$,

$$G(s) = \frac{1}{-s + 1}, \quad \text{ROC} = \text{Re}(s) < 1$$

3. The Laplace transform of $h(t) = e^{t^2}$ not defined for any $s \in \mathbb{C}$. The ROC is empty.
4. The Laplace transform of $p(t) = e^{-t^2}$ is well defined for every $s \in \mathbb{C}$. The ROC is $\mathbb{C}$.

The Fourier transform $\hat{f}$ will from now on usually be seen as a special case of the bilateral Laplace transform and we use $F(i\omega)$ to denote the Fourier transform of a signal f . Another example of Plancherel's theorem reveals something special:

Example 6.7.2 (Orthogonal rational Fourier transform). For any $a, b > 0$ these two functions

$$f(t) = e^{-at} \mathbb{1}_{[0,\infty)}(t) = \begin{array}{c} \text{graph of } f(t) \\ 0 \end{array} \quad \text{and} \quad g(t) = e^{bt} \mathbb{1}_{(-\infty,0]}(t) = \begin{array}{c} \text{graph of } g(t) \\ 0 \end{array}$$

are orthogonal, and are in $\mathbb{L}^2(\mathbb{R})$. Then Plancherel states that their Fourier transforms are orthogonal in $\mathbb{L}^2(i\mathbb{R})$,

$$F(i\omega) = \frac{1}{a + i\omega} \perp \frac{1}{b - i\omega} = G(i\omega).$$

(With Complex Functions Theory this can also be verified entirely in frequency domain.) $\square$

The findings of the above example generalizes as follows. In what follows *stability* of a rational complex function $F(s)$ is understood to mean that $F(s)$ has all its poles in the open left-half plane. *Antistability* means its poles all have strictly *positive* real part.

Lemma 6.7.3 (Orthogonal rational Fourier transform). Let $F(s)$ and $G(s)$ be two strictly proper rational functions. If $F(s)$ is stable and $G(s)$ anti-stable, then they are orthogonal on $\mathbb{L}^2(i\mathbb{R})$:

$$F(i\omega) \perp G(i\omega).$$

If $H(s)$ is a stable rational function with relative degree at least 2, then

$$\int_{\omega \in \mathbb{R}} H(i\omega) d\omega = 0.$$

Proof. Associate with F and G a region of convergence that contains the imaginary axis (i.e. a right-half plane for F and a left-half plane for G). With partial fraction expansion the inverse Laplace transform of F and G may be shown to be causal and anticausal respectively (zero for negative time and zero for positive time, respectively) hence F and G are orthogonal as elements of $\mathbb{L}^2(i\mathbb{R})$. Split $H(s)$ as $H(s) = F(s)G'(-s)$ with F, G strictly proper and F stable and G antistable. Then $\int_{\omega \in \mathbb{R}} H(i\omega) d\omega = \int_{\omega \in \mathbb{R}} F(i\omega)G'(-i\omega) d\omega = \langle F, G \rangle_{\mathbb{L}^2(i\mathbb{R})} = 0$. ■

The set of all stable strictly proper rational functions

$$\mathbb{H}_{\text{rat}}^2 = \{F : i\mathbb{R} \rightarrow \mathbb{C} \mid F(s) \text{ is rational, stable, strictly proper}\}$$

can on the one hand be characterized as being the rational functions that are Fourier transforms of causal² signals and on the other hand can obviously also be characterized as the strictly proper rational functions that have their poles in the open left-half complex plane. The generalization of this equivalence is due to Paley and Wiener (stated below without proof). In what follows $\mathbb{C}_+$ is the open right-half complex plane and $\mathbb{C}_-$ the open left-half complex plane,

$$\mathbb{C}_+ = \{s \in \mathbb{C} \mid \text{Re}(s) > 0\}, \quad \mathbb{C}_- = \{s \in \mathbb{C} \mid \text{Re}(s) < 0\}.$$

Theorem 6.7.4 (Extra: Paley-Wiener). The set of Laplace transforms of causal signals

$$\mathfrak{L}\left(\mathbb{L}_{\mathbb{C}^n}^2[0, \infty)\right)$$

is well defined in $\mathbb{C}_+$ and equals the *Hardy space* $\mathbb{H}^2$,

$$\mathbb{H}^2 = \{F : \mathbb{C}_+ \rightarrow \mathbb{C}^n \mid F \text{ is analytic in } \mathbb{C}_+ \text{ and } \sup_{\sigma > 0} \|F\|_{\mathbb{L}^2(\sigma + i\mathbb{R})} < \infty\}.$$

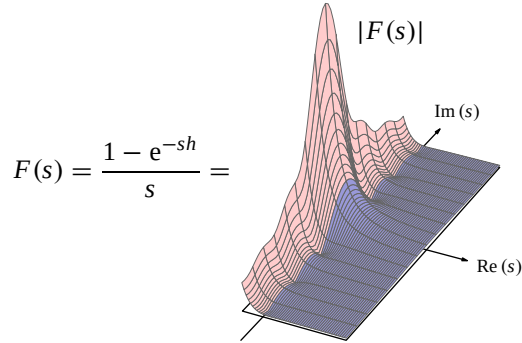
Moreover then $\sup_{\sigma > 0} \|F\|_{\mathbb{L}^2(\sigma + i\mathbb{R})} = \|F(i\omega)\|_{\mathbb{L}^2(i\mathbb{R})}$. Similarly

$$\mathfrak{L}\left(\mathbb{L}_{\mathbb{C}^n}^2(-\infty, 0]\right) = \mathbb{H}^{2\perp} := \{F : \mathbb{C}_- \rightarrow \mathbb{C}^n \mid F \text{ analytic in } \mathbb{C}_- \text{ and } \sup_{\sigma < 0} \|F\|_{\mathbb{L}^2(\sigma + i\mathbb{R})} < \infty\}.$$

□

²A signal $f : \mathbb{R} \rightarrow \mathbb{V}$ is *causal* if $f(t) = 0$ for all $t < 0$.

Example 6.7.5. The bilateral Laplace transform of $f(t) = \mathbb{1}_{[0,h]}(t) = \begin{cases} 1 & 0 \leq t \leq h \\ 0 & \text{otherwise} \end{cases}$ is



The candidate-pole $s = 0$ of $F(s)$ cancels against the zero in the numerator. This $F(s)$ is defined everywhere in the complex plane $\mathbb{C}$. Over the *left*-half complex plane it is unbounded (red part of above figure). This function F is in $\mathbb{H}^2$. $\square$

6.8 Stable systems, causal or noncausal

A linear mapping (system) $y = \mathcal{G}u$ from $\{\mathbb{R} \rightarrow \mathbb{C}^{n_u}\}$ to $\{\mathbb{R} \rightarrow \mathbb{C}^{n_y}\}$ is said to be $\mathbb{L}^2$ -*stable* — or simply, *stable* — if the $\mathbb{L}^2$ -gain of the mapping is finite, that is, if an $M > 0$ exists such that

$$\|\mathcal{G}u\|_{\mathbb{L}^2(\mathbb{R})} \leq M \|u\|_{\mathbb{L}^2(\mathbb{R})} \quad \forall u \in \mathbb{L}^2(\mathbb{R}).$$

The infimal such M is known as the *operator norm* of $\mathcal{G}$.

The set of $\mathbb{L}^2(\mathbb{R})$ -stable LTI systems is denoted

$$\mathbb{L}^\infty.$$

This set contains both causal and noncausal systems. The set of $\mathbb{L}^2$ -stable LTI systems that in addition are causal is denoted

$$\mathbb{H}^\infty.$$

Both $\mathbb{L}^\infty$ and $\mathbb{H}^\infty$ have clear frequency-domain characterizations but that does not concern us now.

For systems with rational transfer matrices $G(s) = \mathcal{L}(g)$, to be in $\mathbb{L}^\infty$ it is necessary and sufficient that the region of convergence of $G(s)$ contains the imaginary axis and that $G(s)$ is proper. In other words, with every proper rational matrix that does not have poles on $i\mathbb{R}$ one can associate a (possibly noncausal) system $y = g * u$ that is $\mathbb{L}^2(\mathbb{R})$ -stable. For systems with rational transfer matrices $G(s) = \mathcal{L}(g)$, to be in $\mathbb{H}^\infty$ it is necessary and sufficient that the region of convergence of $G(s)$ contains the entire closed right-half plane and that it is proper. Conversely to any proper $G(s)$ with its poles in the open LHP one can associate a causal system $y = g * u$ that is $\mathbb{L}^2(\mathbb{R})$ -stable.

6.9 LTI transformations, adjoint and conjugate

In what follows we take the adjoint with respect to $\mathbb{L}^2$.

Lemma 6.9.1 (Adjoint in frequency domain). Let $\mathcal{G}$ be a stable LTI mapping with frequency response $G(i\omega)$. Then the adjoint $\mathcal{G}^*$ has frequency response $[G(i\omega)]'$.

Proof. $\langle y, \mathcal{G}u \rangle = \int_{-\infty}^{\infty} [G(i\omega)u(i\omega)]' y(i\omega) d\omega = \int_{-\infty}^{\infty} u(i\omega)' [G(i\omega)' y(i\omega)] d\omega = \langle \mathcal{G}^* y, u \rangle$. ■

The adjoint frequency response $[G(i\omega)]'$ by itself is defined only on the imaginary axis. It will be useful to extend the adjoint frequency response to transfer matrices (i.e. defined on a larger set than just the imaginary axis). If the LTI mapping is given in time domain (i.e. as a convolution)

$$y = \mathcal{G}(u) : \quad y = h * u$$

then according to Example 5.13.3 the adjoint is a convolution as well with the only difference that the impulse response is time-reversed and that its complex conjugate transpose is taken,

$$u = \mathcal{G}^*(y) : \quad u = g * y, \quad g(t) = h'(-t).$$

The bilateral Laplace transform of $h'(-t)$ is denoted $H^\sim(s)$,

$$H^\sim(s) = \int_{-\infty}^{\infty} h'(-t) e^{-st} dt = \int_{-\infty}^{\infty} h'(t) e^{+st} dt = [H(-\bar{s})]'$$

The commutative diagram below expresses that we can determine the, say, frequency response of the adjoint in different ways.

$$\begin{array}{ccccc} h(t) & \xrightarrow{\text{Laplace}} & H(s) & \xrightarrow{s=i\omega} & H(i\omega) \\ \text{adjoint} \downarrow & & \downarrow \text{conjugate} & & \downarrow \text{adjoint} \\ h'(-t) & \xrightarrow{\text{Laplace}} & H^\sim(s) & \xrightarrow{s=i\omega} & H^\sim(i\omega) \end{array}$$

We can bypass any time-domain interpretations and directly move from transfer matrix $H(s)$ to $H^\sim(s)$ according to the rule

$$H^\sim(s) = [H(-\bar{s})]'$$

The so defined $H^\sim(s)$ in terms of $H(s)$ is known as the *conjugate* of $H(s)$. We see that it has the property

$$H^\sim(i\omega) = [H(i\omega)]'$$

which is consistent with Lemma 6.9.1. If $H(s)$ is defined on the closed right-half plane (closed RHP) the conjugate is defined on the closed left-half plane (closed LHP) and vice versa. If the domain of $H(s)$ includes the imaginary axis then so does that of $H^\sim(s)$. The conjugate of a rational function is rational.

Example 6.9.2 (Conjugate of real-rational $G(s)$). The conjugate of the stable system with transfer function

$$G(s) = \frac{1}{s+1}, \quad (\text{Re}(s) > -1)$$

is

$$G^\sim(s) = \left[\frac{1}{-\bar{s}+1} \right]^* = \frac{1}{-s+1} = G(-s), \quad (\text{Re}(s) < 1).$$

Note that $G(s) = \mathcal{L}(e^{-t} \mathbb{1}(t))$ and $G^\sim(s) = \mathcal{L}(e^t \mathbb{1}(-t))$. □

Likewise one readily gets that for real LTI systems (systems with real-valued impulse response) the conjugate $G^\sim$ may be determined as

$$G^\sim(s) = [G(-s)]^T.$$

In particular, if $G(s)$ is real-rational then so is $G^\sim(s)$ and the poles of $G^\sim(s)$ are minus those of $G(s)$.

Example 6.9.3 (Three norm preserving systems).

1. The stable LTI system $y = \mathcal{G}(u)$ with transfer function

$$G(s) = \frac{s-a}{s+1}$$

is norm preserving — that is $\|y\|_{\mathbb{L}^2} = \|u\|_{\mathbb{L}^2}$ for all u — iff $a = \pm 1$. Indeed,

$$\langle \mathcal{G}u, \mathcal{G}u \rangle_{\mathbb{L}^2(\mathbb{R})} = \langle G\hat{u}, G\hat{u} \rangle_{\mathbb{L}^2(i\mathbb{R})} = \langle G^\sim G\hat{u}, \hat{u} \rangle_{\mathbb{L}^2(i\mathbb{R})}$$

and this equals $\langle \hat{u}, \hat{u} \rangle_{\mathbb{L}^2(i\mathbb{R})}$ for every $\hat{u}$ iff $G^\sim G = I$ on the imaginary axis,

$$G^\sim(s)G(s) = \frac{-s-a}{-s+1} \frac{s-a}{s+1} = \frac{a^2 - s^2}{1 - s^2}.$$

2. The stable LTI system with transfer matrix

$$G(s) = \begin{bmatrix} \frac{1}{s+1} \\ \frac{s}{s+1} \end{bmatrix}$$

is a norm preserving mapping from $\mathbb{L}_{\mathbb{C}}^2(i\mathbb{R})$ to $\mathbb{L}_{\mathbb{C}^2}^2(i\mathbb{R})$ because

$$G^\sim(s)G(s) = \begin{bmatrix} \frac{1}{-s+1} & \frac{-s}{-s+1} \end{bmatrix} \begin{bmatrix} \frac{1}{s+1} \\ \frac{s}{s+1} \end{bmatrix} = \frac{1-s^2}{-s^2+1} = 1.$$

3. The delay operator $y(t-h) = u(t)$ is obviously norm preserving on $\mathbb{L}^2(\mathbb{R})$. Its transfer function is $G(s) = e^{-sh}$ and so

$$G^\sim(s)G(s) = e^{sh}e^{-sh} = 1$$

which once again proves the norm preserving property.

□

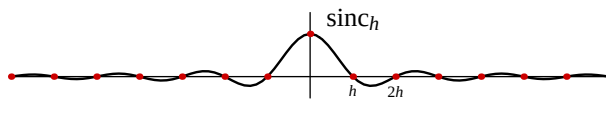
6.10 Extra: some technical details...

- A proper proof of Theorem 6.6.1 and Plancherel's theorem is surprisingly involved. The problem with the sketch of proofs given above is that they rely on Riemann-sum approximations (while we are supposed to work with Lebesgue integrable functions) and in the proof of Plancherel we secretly assumed that the order of integration may be interchanged, which requires Fubini's theorem. A proper proof follows these lines: first show that the theorems are valid for $u = \mathbb{L}^1 \cap \mathbb{L}^2$ (the set $\mathbb{L}^1$ is the set of absolutely integrable functions). For $u \in \mathbb{L}^2$ use that $\mathbb{L}^2 \cap \mathbb{L}^1$ is dense in $\mathbb{L}^2$ to construct a sequence of $u_k \in \mathbb{L}^2 \cap \mathbb{L}^1$ such that $u_k \rightarrow u$ in $\mathbb{L}^2$ -sense. Because of Plancherel $\hat{u}_k \in \mathbb{L}^2(i\mathbb{R})$ and since $\mathbb{L}^2$ is a Hilbert space it has a limit $\hat{u} \in \mathbb{L}^2$. This is the extension of the Fourier transform of u .

- Note that the domain of $\mathbb{H}^2$ does not contain the imaginary axis so we can not consider $F(i\omega)$ for $F \in \mathbb{H}^2$ (which we secretly did in Theorem 6.7.4). We should have said: if $F \in \mathbb{H}^2$ then for almost every $\omega \in \mathbb{R}$ the nontangential limit $\hat{F}(i\omega) := \lim_{\sigma \downarrow 0} F(\sigma + i\omega)$ exists and the so defined function $\hat{F}$ is an element of $\mathbb{L}^2(i\mathbb{R})$.

6.11 Problems

- 6.1 Write a MATLAB script that approximates the Fourier transform $\mathcal{F}_{cc}$ using the FFT.
- 6.2 Consider the $\mathcal{F}_{cc}$ and subspace $\mathbb{V}_A = \{f \in \mathbb{L}^2_{\mathbb{R}}(\mathbb{R}) \mid \hat{f}(i\omega) = 0 \forall |\omega| > A\}$. Find the best approximation in $\mathbb{V}_A$ (in both time domain and frequency domain) of

$$\text{sinc}_h(t) = \frac{\sin(\frac{\pi}{h}t)}{\frac{\pi}{h}t} =$$


The answer depends on h . [Hint: what is the Fourier transform of $\text{sinc}_h(t)$?]

- 6.3 Show that $\text{sinc}_h(t - kh)$, $k \in \mathbb{Z}$ is an orthogonal sequence in $\mathbb{L}^2(\mathbb{R})$.
- 6.4 Construct a function $f \in \mathbb{L}^2_{\mathbb{R}}(\mathbb{R})$ whose Fourier transform $\hat{f}(i\omega)$ is not defined at $\omega = 1$ (but $\hat{f}$ is nevertheless in $\mathbb{L}^2_{\mathbb{C}}(i\mathbb{R})$ as Plancherel dictates). [You do not have to determine $\hat{f}$ explicitly.]
- 6.5 Consider $\mathbb{L}^2_{\mathbb{R}}([-1, 1])$ and suppose we apply the Gram-Schmidt procedure to the sequence $\text{span}\{1, t, t^2, t^3, \dots\}$ to obtain an orthonormal set $\{e_0, e_1, \dots\}$ of polynomials. Prove that this sequence is complete (i.e. every $f \in \mathbb{L}^2_{\mathbb{R}}([-1, 1])$ equals its series expansion $f = \sum_{k=0}^{\infty} \langle f, e_k \rangle e_k$). [Hint: do not apply Gram-Schmidt and have a look at the proof of Thm. 6.4.2.]
- 6.6 Show that the sequence $\{\hat{e}_1, \hat{e}_2, \dots\}$,

$$\hat{e}_k(i\omega) = \frac{1}{i\omega + 1} \left(\frac{i\omega - 1}{i\omega + 1} \right)^k$$

is an orthogonal sequence on $\mathbb{L}^2(i\mathbb{R})$, and determine $\|\hat{e}_k\|$. Is the sequence complete?

- 6.7 Show that the stable causal delay-differential system

$$\dot{y}(t) + \frac{1}{2}y(t - \pi/2) = u(t) - \frac{1}{2}\dot{u}(t - \pi/2)$$

is $\mathbb{L}^2$ -norm preserving.

- 6.8 Let

$$H(s) = c \frac{\prod_{i=1}^n (s - a_i)}{\prod_{j=1}^m (s - b_j)}$$

For which $n, m \in \mathbb{N}$, $c, a_1, \dots, a_n, b_1, \dots, b_m \in \mathbb{R}$ is this the transfer function of a norm preserving LTI system?

- 6.9 $(H^\sim)^\sim = H$?

7 An Introduction to stochastic processes

Also spaces of stochastic variables can be equipped with an inner product. We will denote stochastic variables (random variables) by lower case symbols.

7.1 Expectation, variance and covariance

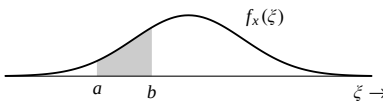
For a real valued stochastic variable x the *probability distribution* $F_x(\xi)$ is defined as the probability that $x \leq \xi$:

$$F_x(\xi) = \Pr(x \leq \xi).$$

Its derivative

$$f_x(\xi) = \frac{\partial}{\partial \xi} F_x(\xi)$$

(which we assume exists) is known as the *probability density function*. It is nonnegative and we get the familiar picture (with $a < b \in \mathbb{R}$):

$$\Pr(a < x < b) = \int_a^b f_x(\xi) d\xi$$


The *expected value* $E(x)$, also known as the *mean*, of a stochastic variable x is defined as

$$E(x) = \int_{-\infty}^{\infty} \xi f_x(\xi) d\xi$$

and one can prove that for any¹ function $g : \mathbb{R} \rightarrow \mathbb{R}$

$$E g(x) = \int_{-\infty}^{\infty} g(\xi) f_x(\xi) d\xi.$$

The three for us important expectations are

$$\begin{aligned} \text{mean:} & E(x) \\ \text{second moment:} & E(x^2) \\ \text{variance:} & \text{var}(x) := E(x - E x)^2 \end{aligned}$$

The mean is also known as the first moment.

For two real valued stochastic variables x_1, x_2 the (*joint*) *probability distribution* $F_{x_1, x_2}(\xi_1, \xi_2)$ is defined as

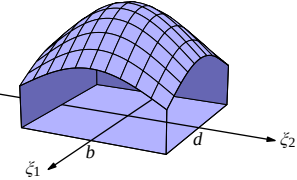
$$F_{x_1, x_2}(\xi_1, \xi_2) = \Pr(x_1 \leq \xi_1; x_2 \leq \xi_2).$$

¹well, sufficient is that g have bounded variation?

Its derivative

$$f_{x_1, x_2}(\xi_1, \xi_2) := \frac{\partial^2}{\partial \xi_1 \partial \xi_2} F_{x_1, x_2}(\xi_1, \xi_2)$$

(we assume it exists and that it is continuous) is known as the *(joint) probability density function*. It is nonnegative and we get the familiar picture (with $a < b \in \mathbb{R}, c < d \in \mathbb{R}$):

$$\Pr(a < x_1 < b; c < x_2 < d) = \int_{\xi_1=a}^{\xi_1=b} \int_{\xi_2=c}^{\xi_2=d} f_{x_1, x_2}(\xi_1, \xi_2) d\xi_1 d\xi_2$$


As before one can prove that for functions $g : \mathbb{R}^2 \rightarrow \mathbb{R}$ the stochastic variable $g(x_1, x_2)$ has expectation

$$\mathbb{E} g(x_1, x_2) = \iint_{\mathbb{R}^2} g(\xi_1, \xi_2) f_{x_1, x_2}(\xi_1, \xi_2) d\xi_1 d\xi_2.$$

The three for use important expectations are

$$\begin{aligned} \text{correlation:} & \quad \mathbb{E}(xy) \\ \text{covariance:} & \quad \text{cov}(x, y) := \mathbb{E}(x - \mathbb{E} x)(y - \mathbb{E} y) \\ \text{correlation coefficient:} & \quad \rho(x, y) := \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}} \end{aligned}$$

We mostly work directly with mean and (co)variance and we hardly ever have to worry about underlying probability density functions.

7.2 Inner product and the covariance matrix

Given a finite number m of $\mathbb{R}^n$ -valued stochastic variables $x_1, \dots, x_m$ all with finite first and second moment

$$\mathbb{E} x_{kj} \in \mathbb{R}, \quad \mathbb{E} x_{kj}^2 < \infty$$

we form the vector space of linear combinations of these m stochastic vectors x_k 's:

$$\mathbb{X}^n := \text{span}\{x_1, \dots, x_m\} := \{x \mid x = \sum_{k=1}^m a_k x_k, a_1, \dots, a_m \in \mathbb{R}\}. \quad (7.1)$$

Any sample (observation, realization) from $\mathbb{X}^n$ is an element of $\mathbb{R}^n$. The space $\mathbb{X}^n$ satisfies the usual vector space properties (we do not want to dwell on it).

Definition 7.2.1 (inner product). Let $\mathbb{X}^n$ be the above vector space (7.1). A functional $\langle \cdot, \cdot \rangle : \mathbb{X}^n \times \mathbb{X}^n \rightarrow \mathbb{R}$ is an *inner product* if for every $x, y, z \in \mathbb{X}^n$

1. $\langle x, y \rangle = \langle y, x \rangle$,
2. $\langle \alpha x + z, y \rangle = \alpha \langle x, y \rangle + \langle z, y \rangle$ for every $\alpha \in \mathbb{R}$,
3. $\langle x, x \rangle \geq 0$, and $\langle x, x \rangle = 0 \implies \Pr(x = 0) = 1$.

□

We make no distinction between zero and stochastic variables that are zero with probability 1. The only inner product we are going to consider is

$$\langle x, y \rangle := E y'x$$

Lemma 7.2.2. $\langle x, y \rangle := E y'x$ is an inner product on $\mathbb{X}^n$:

1. $E y'x = E x'y$
2. $E y'(x + az) = E y'x + a E y'z$.
3. $E x'x \geq 0$
4. $E x'x = 0$ iff $\sum_{k=1}^n E x_k^2 = 0$ iff $\Pr(x = 0) = 1$ zero with probability 1

□

Elements $x, y \in \mathbb{X}^n$ are considered *column* vectors, so $y'x$ is a 1×1 stochastic variable and xy' is a $n \times n$ matrix of stochastic variables.

Definition 7.2.3 (covariance matrix). The *covariance matrix* of $x, y \in \mathbb{X}^n$ is defined as

$$\text{cov}(x, y) = E(x - E x)(y - E y)' = E(\tilde{x}\tilde{y}') = \begin{bmatrix} E(\tilde{x}_1\tilde{y}_1) & E(\tilde{x}_1\tilde{y}_2) & \cdots & E(\tilde{x}_1\tilde{y}_n) \\ E(\tilde{x}_2\tilde{y}_1) & E(\tilde{x}_2\tilde{y}_2) & \cdots & E(\tilde{x}_2\tilde{y}_n) \\ \vdots & \vdots & \ddots & \vdots \\ E(\tilde{x}_n\tilde{y}_1) & E(\tilde{x}_n\tilde{y}_2) & \cdots & E(\tilde{x}_n\tilde{y}_n) \end{bmatrix} \in \mathbb{R}^{n \times n}.$$

Here we use the short hands $\tilde{x} = x - E x$, $\tilde{y} = y - E y$. □

For zero mean x, y the covariance matrix is a matrix of inner products and we have the useful identity

$$\text{tr } E xy' = \sum_{k=1}^n E(x_k y_k) = E y'x.$$

7.3 Orthogonal — uncorrelated

Elements $x, y \in \mathbb{X}^n$ are said to be *orthogonal* if the inner product $E y'x$ is zero. Two stochastic variables $x, y \in \mathbb{X}^1$ are said to be *uncorrelated* if

$$E(xy) = (E x)(E y).$$

Lemma 7.3.1. $x, y \in \mathbb{X}^1$ are uncorrelated iff $(x - E x) \perp (y - E y)$.

Proof. $\langle x - E x, y - E y \rangle = E(x - E x)(y - E y) = E(xy) - (E x)(E y)$. ■

7.4 Norm — standard deviation

The norm is defined as

$$\|x\| = \sqrt{\langle x, x \rangle}$$

Hence for our inner product

$$\|x\| = \sqrt{E x' x} = \sqrt{E(x_1^2) + E(x_2^2) + \cdots + E(x_n^2)} = \sqrt{E |x|^2}.$$

For zero mean $x \in \mathbb{X}^1$ this equals the *standard deviation*

$$\text{std}(x) := \sigma_x := \sqrt{E(x - E x)^2}.$$

7.5 Cauchy-Schwarz — correlation

Lemma 7.5.1 (Cauchy-Schwarz). For stochastic variables $x, y \in \mathbb{X}^1$ there holds that

1. $|E(xy)|^2 \leq E(x^2) E(y^2)$
2. $|E(xy)|^2 = E(x^2) E(y^2)$ if and only if $y = 0$ or $x = ky$ for some $k \in \mathbb{R}$.

Proof. Let $\sigma_x^2 := E(x^2)$ and $\sigma_y^2 := E(y^2)$. The result is trivial if $x = 0$ or $y = 0$. Consider $x \neq 0$ and $y \neq 0$. Then for any $\gamma \in \mathbb{R}$,

$$0 \leq E\left(\frac{x}{\sigma_x} - \gamma \frac{y}{\sigma_y}\right)^2 = E\left(\frac{x^2}{\sigma_x^2} - 2\gamma \frac{xy}{\sigma_x \sigma_y} + \gamma^2 \frac{y^2}{\sigma_y^2}\right) = 1 - 2\gamma \frac{E(xy)}{\sigma_x \sigma_y} + \gamma^2. \quad (7.2)$$

As the above is nonnegative it follows that

$$2\gamma \frac{E(xy)}{\sigma_x \sigma_y} \leq 1 + \gamma^2. \quad (7.3)$$

For $\gamma = \pm 1$ this reads $\frac{|E(xy)|}{\sigma_x \sigma_y} \leq 1$. This proves Condition 1.

If equality $|E(xy)|^2 = E(x^2) E(y^2)$ holds then for $\gamma = \text{sgn}(E(xy))$ the expression (7.2) is zero, so necessarily $\frac{x}{\sigma_x} - \gamma \frac{y}{\sigma_y} = 0$. Take $k = \gamma \sigma_x / \sigma_y$. ■

An immediate consequence is:

Lemma 7.5.2 (Correlation coefficient). The correlation coefficient $\rho(x, y) := \frac{\text{cov}(x, y)}{\sqrt{\text{var}(x)\text{var}(y)}}$ satisfies the bounds

$$-1 \leq \rho(x, y) \leq 1.$$

□

7.6 Pythagoras — variance of uncorrelated variables

Pythagoras become the familiar rule:

Lemma 7.6.1. Let $x, y \in \mathbb{X}^1$. If $E xy = 0$ then $E(x + y)^2 = E(x^2) + E(y^2)$. When applied to mean zero x and y it says: zero mean, uncorrelated implies

$$\text{var}(x + y) = \text{var}(x) + \text{var}(y).$$

Proof. Easy. ■

7.7 Projection — optimum linear estimator

Consider the finitely generated $\mathbb{X}^n$ of (7.1) and suppose that $x \in \mathbb{X}^n$ is not observed but that we want to estimate x on the basis of other (related) stochastic variables in some set $\mathbb{V} \subset \mathbb{X}^n$.

Definition 7.7.1 (Optimum linear estimator). Let $\mathbb{V}$ be a subspace of $\mathbb{X}^n$ of (7.1) with norm $\|x\| = \sqrt{E|x|^2}$. A stochastic variable $v_* \in \mathbb{V} \subseteq \mathbb{X}^n$ is an *optimum linear estimator*² in $\mathbb{V}$ of $x \in \mathbb{X}^n$ if

$$\|x - v_*\| \leq E\|x - v\| \quad \forall v \in \mathbb{V}.$$

□

Theorem 7.7.2 (A projection theorem). Let $x \in \mathbb{X}^n$ and let $\mathbb{V}$ be a subspace of the $\mathbb{X}^n$ of (7.1).

1. A stochastic variable v_* is an optimal linear estimator in $\mathbb{V}$ of x iff $(x - v_*) \perp \mathbb{V}$.
2. The optimum linear estimator is unique and the *minimum mean square error (MMSE)* $E(|x - v_*|^2)$ equals $E(|x - v_*|^2) = E(|x|^2) - E(|v_*|^2)$.

Proof. Essentially the same as that of Theorem 2.6.2. ■

State differently, v_* is an optimum linear estimator iff $E(x - v_*)v = 0$ for every $v \in \mathbb{V}$. For our finitely generated space (7.1) the optimum linear estimator always exists, which follows from the normal equations derived next.

7.8 Normal equations

For a brief moment let $x, z_1, \dots, z_n$ denote real numbers. Then the problem

$$\min_{y \in \mathbb{R}^n} \|x - [z_1 \ \cdots \ z_n] y\|$$

is an underdetermined problem and typically many $y \in \mathbb{R}^n$ exist that achieve zero norm. With stochastic variables the story is quite different.

Lemma 7.8.1 (Normal equations). Let $x, z_1, \dots, z_n$ be 1×1 stochastic variables. Then

$$y_* = \operatorname{argmin}_{y \in \mathbb{R}^n} \|x - [z_1 \ \cdots \ z_n] y\|$$

iff y_* satisfies the *normal equations*

$$\begin{bmatrix} E z_1 x \\ E z_2 x \\ \vdots \\ E z_n x \end{bmatrix} - \begin{bmatrix} E z_1 z_1 & E z_1 z_2 & \cdots & E z_1 z_n \\ E z_2 z_1 & E z_2 z_2 & \cdots & E z_2 z_n \\ \vdots & \vdots & \ddots & \vdots \\ E z_n z_1 & E z_n z_2 & \cdots & E z_n^2 \end{bmatrix} \begin{bmatrix} y_{*1} \\ y_{*2} \\ \vdots \\ y_{*n} \end{bmatrix} = 0$$

²The term “linear” emphasizes that we seek a linear combination of elements of $\mathbb{V}$; in general if the entire joint probability distribution is known, then $E|x - v_*|^2$ is minimized for the conditional expectation $E(x|\mathbb{V})$ and this need not be a linear combination of elements of $\mathbb{V}$!

or, more succinctly (with z the column vector of z_k 's),

$$E(zx') - E(zz')y_* = 0.$$

The MMSE $E(x - z'y_*)^2$ equals

$$E(x^2) - E(xz')[E(zz')]^{-1}E(zz'x).$$

Proof. According to the projection theorem we have

$$\begin{aligned} z'y_* \text{ optimal} &\iff x - [z_1 \ \cdots \ z_n]y_* \perp z_k \quad \forall k \\ &\iff E(z_k x) - [E(z_k z_1) \ \cdots \ E(z_k z_n)]y_* = 0 \quad \forall k. \end{aligned}$$

The MMSE is $E(x - z'y_*)^2 = E(x^2) - E((z'y_*)^2) = E(x^2) - y_*' E(zz')y_*$. ■

Example 7.8.2 (All that hair). Let us say there is a correlation between a persons IQ, EQ and amount of hair x_h . We assume that z_{iq} , z_{eq} and x_h have been centered (normalized) to have mean zero³. Of course the correlation between x_h and z_{iq} is negative.

We aim to estimate x_h on the basis of z_{iq} and z_{eq} . The normal equations for our problem are

$$y_* = \begin{bmatrix} E(z_{iq}^2) & E(z_{iq}z_{eq}) \\ E(z_{eq}z_{iq}) & E(z_{eq}^2) \end{bmatrix}^{-1} \begin{bmatrix} E(z_{iq}x_h) \\ E(z_{eq}x_h) \end{bmatrix}.$$

Case 1: when z_{iq} and z_{eq} are uncorrelated. Then the covariance matrix is diagonal and the optimum linear estimator of x_h is

$$x_{*h} = \frac{E(z_{iq}x_h)}{E(z_{iq}^2)}z_{iq} + \frac{E(z_{eq}x_h)}{E(z_{eq}^2)}z_{eq}.$$

Case 2: when $z_{iq} = z_{eq}$. Then the covariance matrix that we want to invert is singular and indeed in this case any $\lambda z_{iq} + (1 - \lambda)z_{eq}$ is the same, i.e. the optimum linear estimator is not unique. The optimum linear estimator is as in the scalar case,

$$x_{*h} = \frac{E(z_{iq}x_h)}{E(z_{iq}^2)}z_{iq}.$$

Case 3: when $E(z_{iq}^2) = E(z_{eq}^2) = E(z_{iq}x_h) = E(z_{eq}x_h) = 1$. Then

$$x_{*h} = \frac{z_{iq} + z_{eq}}{1 + E(z_{iq}z_{eq})}.$$

Cauchy-Schwarz in this case says that $-1 \leq E(z_{iq}z_{eq}) \leq 1$. □

7.9 Stochastic processes

A discrete *stochastic process* is a sequence of stochastic variables $x[t]$ with $t \in \mathbb{Z}$ (or some other countable set). At each time $t \in \mathbb{Z}$ the $x[t]$ is a stochastic variable. Its stochastic properties may change with time as indicated in Fig. 7.1 where we see that the probability density function $f(x, t)$ changes with time.

³so for most of us here at the Technion the amount of hair is negative and probably one out of (z_{iq}, z_{eq}) is positive.

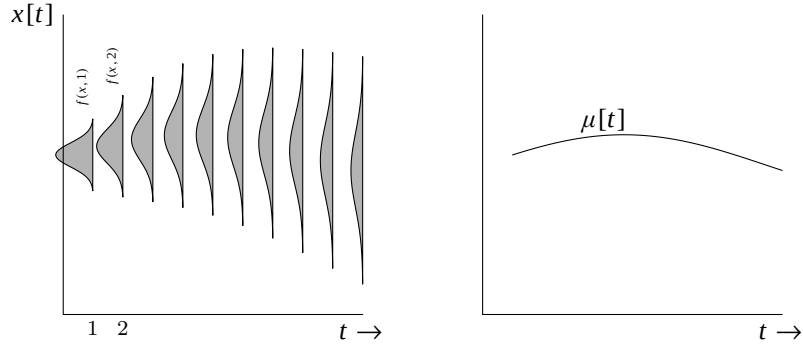


Figure 7.1: The probability density $f(x; t)$ and mean $\mu[t]$ etc. depend on time

The mean

$$\mu_{x[t]}[t] := E x[t]$$

and variance and standard deviation

$$\sigma_{x[t]}^2[t] = \text{var}(x[t]), \quad \sigma_{x[t]}[t] = \text{std}(x[t])$$

may depend on time. We, though, will work exclusively with stochastic processes that are wide-sense stationary, considered next. By the way, Figure 7.1 does *not* capture temporal relatedness of a stochastic process, i.e. that samples at subsequent times may be related.

7.10 Wide-sense stationary stochastic processes

Definition 7.10.1 (WSS). A stochastic process $x[t]$, $t \in \mathbb{Z}$ is *wide sense stationary (WSS)* if

1. mean $\mu_x[t] := E x[t]$ is constant, $\mu_x[t] = \mu_x$,
2. $\text{cov}(x[t], x[s])$ depends on lag $t - s$ only: $\text{cov}(x[t + \tau], x[t]) = r_x[\tau]$.

In this case we refer to $r_x[\tau]$ as the *covariance function* of the process x . It is also known as the *autocorrelation* (function). $\square$

The covariance function $r[\tau]$ of a WSS process describes the temporal relatedness of the process. For example, if $r_x[24] > 0$ then $x[t + 24]$ and $x[t]$ are positively correlated. The covariance function is even, $r[\tau] = r[-\tau]$, and by Cauchy-Schwarz it is maximal at $\tau = 0$,

$$r[0] \geq |r[\tau]| \quad \forall \tau \in \mathbb{Z}.$$

This expresses that $x[t]$ correlates best with itself. An example of a covariance function is

$$r[\tau] = \sigma^2 a^{|\tau|} \quad \text{for } |a| < 1. \quad (7.4)$$

for some $|a| < 1$. The number $\sigma = \sqrt{r[0]} = \sqrt{\text{var}(x[t])}$ is the *standard deviation* of the process. The number a , with $|a| < 1$, determines how fast the function decreases to zero, and, hence, how fast the temporal relatedness is lost. Figure 7.2(a) shows the plot of r for $\sigma = 1$ and $a = 0.9$.

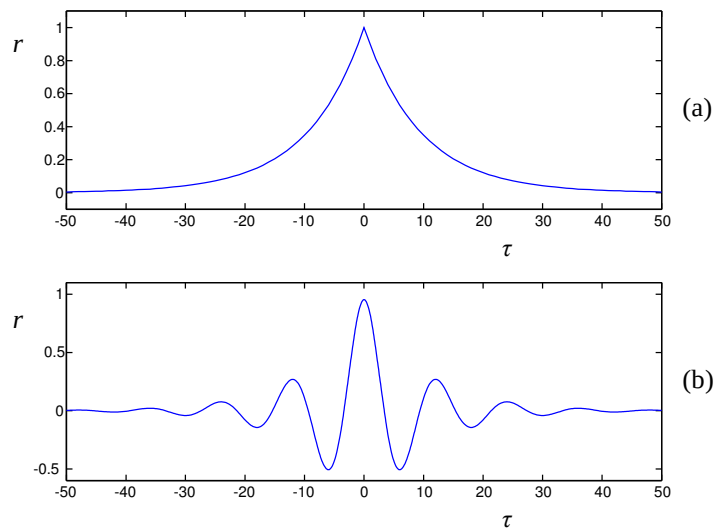


Figure 7.2: (a) Exponential covariance function
(b) Damped harmonic covariance function

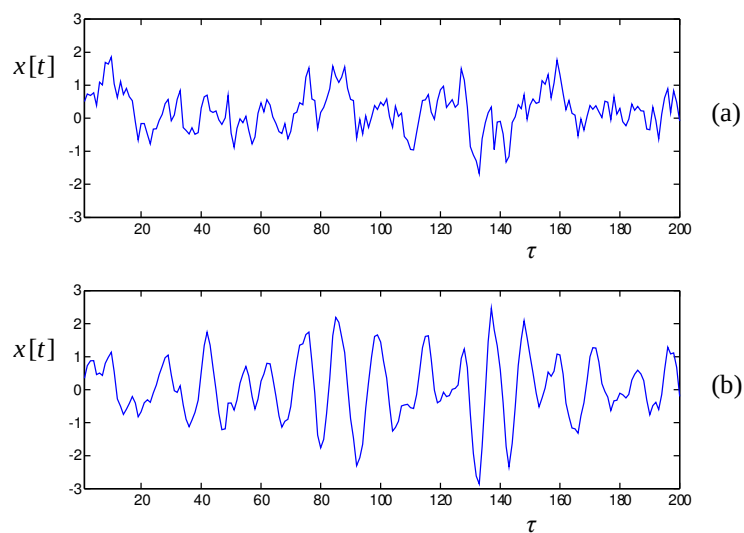


Figure 7.3: (a) Realization of a process with exponential covariance function
(b) Realization of a process with damped harmonic covariance function

Figure 7.3(a) shows an example of a realization of the corresponding WSS process with mean value $\mu = 0$. Later we will see how we can generate such realizations.

A different example of a covariance function is (without proof):

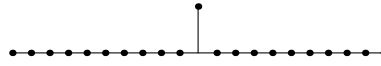
$$r[\tau] = \sigma^2 a^{|\tau|} [A \cos(\frac{2\pi \tau}{T}) + B \sin(\frac{2\pi |\tau|}{T})], \quad \tau \in \mathbb{Z}, \quad (7.5)$$

with σ , a , T , A , and B constants. Again $\sigma \geq 0$ is the standard deviation of the process. The number a , with $|a| < 1$, again determines how fast the function decreases to zero. The number $T > 0$ represents a periodic characteristic of the signal. Figure 7.2(b) displays the behavior of the plot of r for $\sigma = 1$, $a = 0.9$, $T = 12$, $A = 1$ and $B = 0.18182$. Figure 7.3(b) shows a possible realization of the process with mean value $m = 0$. The irregularly periodic behavior is unmistakable.

7.11 White noise

A WSS process has minimal temporal relatedness if

$$r[\tau] = \begin{cases} \sigma^2 & \text{for } \tau = 0, \\ 0 & \text{for } \tau \neq 0, \end{cases} \quad (7.6)$$



This process consists of a sequence of uncorrelated stochastic variables. Such a process is sometimes said to be “purely random.” In a physical or engineering context it is often called *white noise*. This name is explained later in Example 7.12.2. Usually the mean value of white noise is assumed to be 0 and the standard deviation σ taken to be 1. Sometimes it is assumed that the stochastic variables are not only uncorrelated but even mutually independent. In this case the process is completely characterized by its distribution $F_{x[t]}(\xi) = \Pr(x[t] \leq \xi)$. Often it is assumed that this distribution is normal.

The white noise process plays an important role in the theory of stochastic processes and signal processing and time series. In a sense it is the most elementary stochastic process. We shall soon see that other stochastic processes often may be thought to originate from white noise, that is, they are generated by a “driving” white noise process.

Figure 7.4 shows an example of a realization of white noise with mean zero and standard deviation $\sigma = 1$. Realizations of white noise may easily be generated with a random number generator.

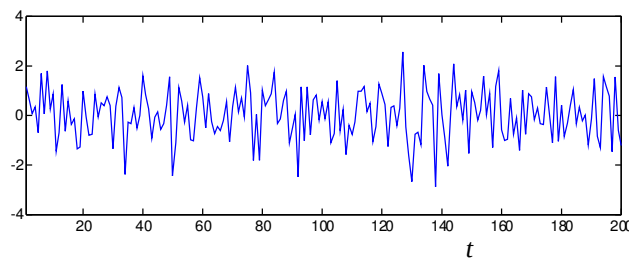


Figure 7.4: Realization of white noise with mean 0 and standard deviation 1

```
x= rand(1,200); % 200 samples white, uniformly distributed over [0,1]
x=randn(1,200); % 200 samples white, normally distributed, mean 0, var 1
plot(x);        % plot it
```

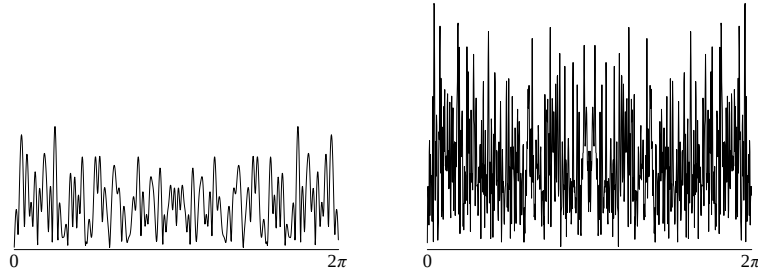


Figure 7.5: Magnitude of DTFT of a realization of white noise (a) $N = 100$ samples, (b) $N = 400$ samples

7.12 Power spectral density, Parseval

Realizations of WSS processes $x[t]$ do not have a *Discrete-time-Fourier-Transform (DTFT)*

$$\sum_{t=-\infty}^{\infty} x[t]e^{-i\omega t}$$

because the infinite sum is not well defined. Direct frequency analysis of stochastic processes therefore is not immediately feasible. At first sight it may seem a good idea to take a *finite* realization $x[0], \dots, x[N-1]$ and compute the DTFT of this finite realization. Figure 7.5 shows the magnitude of the DTFT of a realization of unit variance zero mean white noise for two different N . What we notice is that the DTFT is very irregular and that the magnitude appears to increase with N but not linearly. The following shows that it grows linearly in $\sqrt{N}$ and that we should have taken the DTFT of the covariance function:

Lemma 7.12.1 (Spectral density function). If $x[t]$ is a zero mean WSS process with limited memory in the sense that $\sum_{\tau \in \mathbb{Z}} |\tau r[\tau]| < \infty$ then

$$\lim_{N \rightarrow \infty} \frac{1}{N} \mathbb{E} \left| \sum_{t=0}^{N-1} x[t]e^{i\omega t} \right|^2 = \phi(\omega) \quad (7.7)$$

where $\phi(\omega)$ is the (power) *spectral density* of $x[t]$ defined as the DTFT of the covariance function $r[\tau]$ of the process,

$$\phi(\omega) = \sum_{\tau=-\infty}^{\infty} r[\tau]e^{-i\omega\tau}, \quad \omega \in [-\pi, \pi). \quad (7.8)$$

Proof.

$$\begin{aligned} \frac{1}{N} \mathbb{E} \left| \sum_{t=0}^{N-1} x[t]e^{-it\omega} \right|^2 &= \frac{1}{N} \mathbb{E} \sum_{t=0}^{N-1} x[t]e^{-it\omega} \sum_{s=0}^{N-1} x[s]e^{is\omega} \\ &= \frac{1}{N} \mathbb{E} \sum_{t=0}^{N-1} \sum_{s=0}^{N-1} x[t]x[s]e^{-i(t-s)\omega} = \frac{1}{N} \sum_{t=0}^{N-1} \sum_{s=0}^{N-1} r[t-s]e^{-i(t-s)\omega}. \end{aligned}$$

The above double sum equals the sum of all entries in the $N \times N$ matrix

$$\frac{1}{N} \begin{bmatrix} r[0]e^0 & r[1]e^{-i\omega} & \dots & r[N-1]e^{-i(N-1)\omega} \\ r[1]e^{-i\omega} & r[2]e^{-i2\omega} & \dots & r[N-2]e^{-i(N-2)\omega} \\ \dots & \dots & \dots & \dots \\ r[N-1]e^{-i(N-1)\omega} & r[N-2]e^{-i(N-2)\omega} & \dots & r[0]e^0 \end{bmatrix}. \quad (7.9)$$

This matrix is constant along diagonals. If we sum the entries of the matrix per diagonal we get

$$\sum_{k=-N+1}^{N-1} (1 - \frac{|k|}{N}) r[k] e^{-i\omega k} = \sum_{k=-N+1}^{N-1} r[k] e^{-i\omega k} - \frac{1}{N} \sum_{k=-N+1}^{N-1} |k| r[k] e^{-i\omega k}. \quad (7.10)$$

The right-hand side converges to $\phi(\omega)$ as $N \rightarrow \infty$. ■

This does not explain why the DTFTs of Fig. 7.5 are so irregular. We will not go into that here. The name “power spectral density” will soon become clear. The spectral density takes over the role of the Fourier transform. A noticeable difference between DTFT’s of signals and spectral density functions of processes is that spectral densities are real valued and in fact nonnegative as we see from (7.7). A further mathematically trivial property is that

$$\text{var}(x[t]) = r[0] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi(\omega) d\omega. \quad (7.11)$$

This follows immediately from the inverse DTFT applied to $r[\tau]$ for $\tau = 0$ — See Section 6.5. For non-stochastic signals $x[t]$, the *power* is commonly defined as the average of $x^2[t]$. In a stochastic setting the power of a zero mean process becomes $\text{var}(x[t]) = \mathbb{E} x^2[t] = r[0]$. Equation (7.11) shows that the power can be seen as distributed over frequency, hence the name power spectral density and we think of the number $\frac{1}{2\pi} \phi(\omega_0)$ as the power at frequency ω_0 . Equation (7.11) is a type of Parseval identity.

Example 7.12.2 (Why white noise is white). Because the covariance function of white noise $x[t]$ with standard deviation σ is given by

$$r[\tau] = \text{cov}(x[t + \tau], x[t]) = \begin{cases} \sigma^2 & \text{for } \tau = 0, \\ 0 & \text{for } \tau \neq 0, \end{cases} \quad \tau \in \mathbb{Z}, \quad (7.12)$$

the spectral density function of white noise equals

$$\phi(\omega) = \sigma^2, \quad \omega \in [-\pi, \pi]. \quad (7.13)$$

Therefore, all frequencies are equally represented in white noise. I.e. its power is equally distributed over all frequencies. The fact that white light supposedly also has this property explains the name “white noise.” □

Not every function $r[\tau]$ is a covariance function of some WSS $x[t]$. Clearly it needs to be symmetric $r[\tau] = r[-\tau]$ and $r[0]$ needs to be nonnegative because $r[0] = \text{var}(x[t])$ and by Cauchy-Schwarz’s inequality we also need that $|r[\tau]| \leq r[0]$. These conditions are not sufficient. The following theorem summarizes some elegant necessary and sufficient characterizations of covariance functions.

Theorem 7.12.3 (Covariance functions). Suppose $r[\tau]$ is an exponentially decaying. The following four statements are equivalent.

1. $r[\tau]$ is a covariance function of some WSS process.

- $$R_n := \begin{bmatrix} r[0] & r[1] & \cdots & r[n-1] \\ r[1] & r[0] & \cdots & r[n-2] \\ \vdots & \vdots & \ddots & \vdots \\ r[n-1] & r[n-2] & \cdots & r[0] \end{bmatrix} \quad (7.14)$$

Proof. Equivalence of 2) and 3) follows from inverse DTFT and DTFT:

$$R_n = \frac{1}{2\pi} \int_{-\pi}^{\pi} \begin{bmatrix} e^{0i\omega} \\ e^{1i\omega} \\ \vdots \\ e^{(n-1)i\omega} \end{bmatrix} \phi(\omega) \begin{bmatrix} e^{-0i\omega} & e^{-1i\omega} & \dots & e^{-(n-1)i\omega} \end{bmatrix} d\omega$$

7.13 LTI systems, power and the ℓ^2 -norm

The diagram illustrates the convolution operation. On the left, the input signal u is shown as a sequence of impulses $u[n]$ and $u[m]$. These impulses pass through a system block labeled $\mathcal{H}$. On the right, the output signal y is shown as a sequence of impulses. Each impulse in y is a weighted sum of the input impulses, where the weights are the impulse response h evaluated at different time shifts. Specifically, the impulse at time t in y is the sum of $u[n]h[t-n]$ for all n such that $u[n]$ is non-zero. The diagram shows several such terms, with the most prominent one being $u[m]h[t-m]$.

78

here is this: since the white noise samples are uncorrelated (orthogonal), their contribution to $y[t]$:

$$h[t-m]u[m] \quad \text{and} \quad h[t-n]u[n]$$

are uncorrelated (orthogonal) as well and so each term attributes to the power (variance) of $y[t]$ irrespective of the other,

$$\text{var}(h[t-m]u[m] + h[t-n]u[n]) = h^2[t-m] + h^2[t-n].$$

This is for two samples. When we take all white noise samples into account we get the total power of $y[t]$:

$$\text{var}(y^2[t]) = \sum_n h^2[t-n] = \|h\|_{\ell^2}^2.$$

The power does not depend on time t . Slightly more general:

Lemma 7.13.1 (ℓ^2 -norm). Let u be a white noise process with zero mean and unit variance. Suppose the impulse response $h \in \ell^2(\mathbb{Z})$. Then the response $y = \mathcal{H}(u)$ to u is a well defined zero-mean WSS process with covariance function

$$r_y[\tau] = \sum_{k \in \mathbb{Z}} h[k]h[k-\tau].$$

In particular the power of the output equals

$$\text{var}(y[t]) = r_y[0] = \|h\|_{\ell^2(\mathbb{Z})}^2 := \sum_{k \in \mathbb{Z}} |h[k]|^2.$$

Proof. Just plug it in (boring):

$$\begin{aligned} \mathbb{E} y[t]y[s] &= \mathbb{E} \left(\sum_m h[m]u[t-m] \sum_n h[n]u[s-n] \right) \\ &= \mathbb{E} \left(\sum_m \sum_n h[m]h[n]u[t-m]u[s-n] \right) \\ &= \sum_m \sum_n h[m]h[n] \mathbb{E} u[t-m]u[s-n] = \sum_m \sum_n h[m]h[n]r_u[t-s+n-m]. \end{aligned}$$

All summations are from $-\infty$ to ∞ . Inspection shows that the right-hand side only depends on the difference $t-s$. It shows that the process $y[t]$ is WSS with covariance function r_y given by

$$r_y[\tau] = \sum_m \sum_n h[m]h[n]r_u[\tau+n-m], \quad \tau \in \mathbb{Z}. \quad (7.15)$$

For white noise, $r_u[\tau+n-m]$ is nonzero only if $m = \tau+n$, rendering $r_y[\tau] = \sum_n h[\tau+n]h[n]$. ■

Example 7.13.2. Consider LTI system $y = \mathcal{H}(u)$ with u a zero-mean unit variance white noise process. For $h[n] = a^n \mathbb{1}[n]$ with $-1 < a < 1$ we get (after some manipulation)

$$r_y[\tau] = \frac{1}{1-a^2} a^{-|\tau|}.$$

The variance (power) of the output y equals $r[0] = 1/(1-a^2)$. Clearly for $a \approx 1$ the system has a long memory and the power of the output becomes significant. □

7.14 Models for spectral densities

Since ℓ^2 -stable LTI systems transform any WSS process into a WSS process — this is shown in the proof of Lemma 7.13.1 — there should be a connection between input and output spectral densities.

Lemma 7.14.1 (Output spectral density). The output $y = \mathcal{H}(u)$ of a stable LTI system driven by a WSS stationary u with spectral density ϕ_u is again WSS and its spectral density is

$$\phi_y(\omega) = |H(e^{i\omega})|^2 \phi_u(\omega) \quad (7.16)$$

with $H(e^{i\omega})$ the frequency response of the system.

Proof.

$$r_u[\tau] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_u(\omega) e^{i\omega\tau} d\omega, \quad \tau \in \mathbb{Z}. \quad (7.17)$$

Substitution of (7.17) into the right-hand side of (7.15) yields

$$\begin{aligned} r_y[\tau] &= \sum_m \sum_n h[m]h[n] \left(\frac{1}{2\pi} \int_{-\pi}^{\pi} \phi_u(\omega) e^{i\omega(\tau+n-m)} d\omega \right) \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \underbrace{\left(\sum_m h[m]e^{-i\omega m} \right)}_{H(e^{i\omega})} \underbrace{\left(\sum_n h[n]e^{i\omega n} \right)}_{H(e^{-i\omega})} \phi_u(\omega) e^{i\omega\tau} d\omega \\ &= \frac{1}{2\pi} \int_{-\pi}^{\pi} |H(e^{i\omega})|^2 \phi_u(\omega) e^{i\omega\tau} d\omega, \quad \tau \in \mathbb{Z}. \end{aligned} \quad (7.18)$$

Here we use the conjugate symmetry of the frequency response function H . Closer inspection shows that the right-hand side of (7.18) is the inverse DTFT of r_y . ■

Example 7.14.2 (Second order system). Let u be a zero mean, unit variance white noise process. Consider the LTI system described by difference equation

$$y[t] - a_1 y[t-1] - a_2 y[t-2] = u[t]. \quad (7.19)$$

As we all know it has frequency response

$$H(e^{i\omega}) = \frac{1}{1 - a_1 e^{-i\omega} - a_2 e^{-i2\omega}}.$$

For $a_1 = 2a \cos(2\pi/T)$ and $a_2 = -a^2$ the transfer function $H(z)$ has poles at $z_{1,2} = ae^{\pm i2\pi/T}$. We take $T = 12$. The power spectral density becomes

$$\phi_z(\omega) = \frac{1}{|1 - a_1 e^{-i\omega} - a_2 e^{-i2\omega}|^2}.$$

It has a peak near frequency $2\pi/T = 0.5236$ confirming that $T = 12$. With the MATLAB command `filter` one can generate realizations of this process (see below). Figures 7.2(b) and 7.3(b) respectively show $r[\tau]$ and a possible realization for this process. □

An example of Octave/Matlab code to generate WSS processes is:

```

N=1024;
u=randn(1,N);           % N samples of zero-mean, unit-var white noise
A=[1 -.75 0.5];         % system y[n]-0.75y[n-1]+0.5y[n-2]=u[n]
y=filter(1,A,u);         % find y (by stability it converges to WSS)
w=2*pi*(0:(N-1))/N;
plot(w,abs(fft(y)).^2/N) % estimated spectral density (known as periodogram)

```

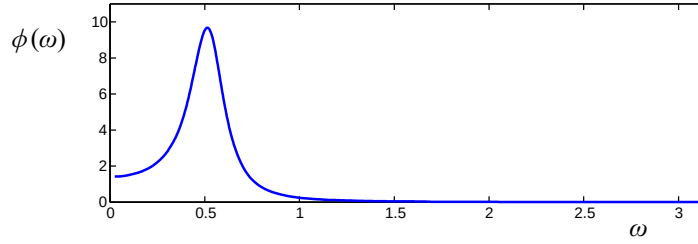


Figure 7.7: Spectral density function of process (7.19), Example 7.14.2

7.15 WSS processes are responses of LTI systems driven by white noise

The following lemma is useful. It says that effectively every WSS process can be seen as the output of a stable LTI system driven by white noise.

Lemma 7.15.1. For any exponentially decaying covariance function r there is zero-mean, unit-variance white noise process u and impulse response $h \in \ell^2(\mathbb{Z})$ such that $y = h * u$ has covariance function r .

Proof. If $r[\tau]$ is exponentially decaying then its spectral density $\phi(\omega)$ exists (and is continuous) and $r[0] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \phi(\omega) d\omega < \infty$. Therefore $\psi(\omega) := \sqrt{\phi(\omega)}$ is square integrable and hence its inverse Fourier transform—call it $h[t]$ —is in ℓ^2 . The spectral density of $y := h * u$ according to Lemma 7.14.1 equals $|H(e^{i\omega})|^2 = |\psi(\omega)|^2 = \phi(\omega)$. ■

If in addition the spectral density function is *strictly* positive then the system may in fact be taken causal and causally invertible⁴ (no proof; we do an example soon). The important practical implication of this lemma is this:

find a discrete-stable rational function H such that $|H(e^{i\omega})|^2$ resembles the (estimated) spectral density of a stationary y then we can think of y as the output of a system $y = \mathcal{H}(u)$ driven by a process u with unit spectral density, $\phi_u(\omega) = 1$, i.e. white noise.

The inverse filter $u = \mathcal{H}^{-1}(y)$ is called a *whitening filter*.

Example 7.15.2 (Spectral factorization.). Suppose $\phi_y(\omega)$ is a continuous smooth spectral density function that is > 0 on $[-\pi, \pi]$. Take a Fourier series approximation of high enough order N such that it is > 0 as well,

$$\phi_{*y}(\omega) = \sum_{k=-N}^N c_k e^{ik\omega}, \quad c_N \neq 0.$$

⁴Another sufficient condition of existence of such a bicausal, bistable $y = \mathcal{H}(u)$ for given process y is that it satisfies the Paley-Wiener condition that $\int_{-\pi}^{\pi} |\log(\phi_y(\omega))| d\omega < \infty$.

As $\phi_y(\omega)$ is real and even, so is its Fourier series approximation and this means that the c_k are real and even as well. We want a real bistable bicausal system $y = \mathcal{H}(u)$ such that

$$H(e^{i\omega})H(e^{-i\omega}) = \phi_{*y}(\omega).$$

That is,

$$H(z)H\left(\frac{1}{z}\right) = \Phi(z) := \sum_{k=-N}^N c_k z^k.$$

The right-hand side is rational and $\Phi(z) = \Phi(\frac{1}{z})$ by symmetry of ϕ_{*y} . Therefore its zeros come in conjugate pairs $(z, 1/z)$ and since $\phi_{*y} > 0$ it has no zeros on the unit circle,

$$\Phi(z) = c \prod_{i=1}^N (z - z_i) \left(\frac{1}{z} - \bar{z}_i\right), \quad 0 < |z_i| < 1.$$

Take

$$H(z) = \pm \sqrt{c} \prod_{i=1}^N \left(\frac{1}{z} - \bar{z}_i\right), \quad 0 < |z_i| < 1.$$

It is a FIR system of order N and its inverse is stable and causal. □

The conditions in this section can probably be relaxed. There is a lot more to say about the problems considered in this section.

7.16 Extra: Continuous-time stochastic processes — simplistic overview

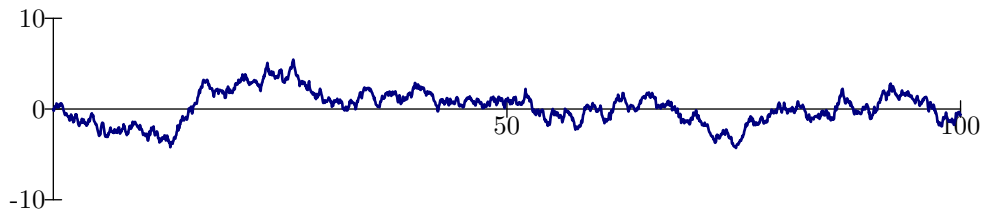


Figure 7.8: A realization of continuous-time stationary noise

If we restrict the noisy signal $z(t)$ of Fig. 7.8 to

$$z_T(t) = \begin{cases} z(t) & t \in [0, T] \\ 0 & \text{otherwise} \end{cases} \quad (7.20)$$

for some choice of T then the Fourier transform of z_T depends critically on this choice of T no matter how large we take it.

If the signal $z(t)$ is stationary—which we colloquially define to mean that its stochastic properties do not vary with time—then the power of the signal,

$$\mathbb{E} z^2(t)$$

does not depend on time and we can estimate this power by averaging over time

$$\mathbb{E} z^2(t) \approx \frac{1}{T} \int_0^T z^2(v) dv.$$

Stationarity also implies that the covariance function $\mathbb{E} z(t)z^*(t + \tau)$ does not depend on time t . The covariance function is a measure for the temporal relatedness in the signal and typically it decays to zero as $\tau \rightarrow \infty$. If this decay is fast enough—a technical condition—then the Fourier transform of z_T converges in the following sense.

Theorem 7.16.1 (SISO spectral density). For any continuous-time zero-mean stationary signal $z(t)$ with finite power and whose covariance function decays exponentially fast⁵ the (power) spectral density function⁶ equals

$$\phi_z(\omega) = \lim_{T \rightarrow \infty} \mathbb{E} \frac{1}{T} |z_T(i\omega)|^2 \quad (7.21)$$

where $z_T(t)$ is defined as in (7.20) and $z_T(i\omega)$ is its Fourier transform. The power $\mathbb{E} z^2(t)$ can be expressed as an integral over frequency,

$$\mathbb{E} z^2(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_z(\omega) d\omega. \quad (7.22)$$

□

Because of (7.22) and the fact that $\phi(\omega) \geq 0$ we can think of the function $\phi_z(\omega)$ as the way that the power of z is distributed over frequency. The theorem shows us how to properly normalize the Fourier transform of $z_T(t)$. The function $\frac{1}{T} |z_T(i\omega)|^2$ is known as the *periodogram* and it serves as an estimate of $\phi_z(\omega)$. Note that any information about the phase of $z_T(i\omega)$ is lost in the spectral density. This is in line with the assumed stationarity of the signal, implying that time shifts do not affect its stochastic properties.

It is a classic result that the output $z = \mathcal{H}(w)$ of a stable LTI system driven by a stationary input w is again stationary and that its spectral density equals

$$\phi_z(\omega) = |H(i\omega)|^2 \phi_w(\omega).$$

This allows us to shape the spectral density function. The practical implication again is:

find a stable rational function H such that $|H(i\omega)|^2$ resembles the (estimated) spectral density of a stationary z then we can think of z as the output of a system $z = Hw$ driven by a process w with unit spectral density, $\phi_w(\omega) = 1$.

A stationary process whose spectral density is 1 is known as unit intensity *white noise*. White noise in continuous time is a somewhat illusive process because of its infinite power, but for the importance is foremost that it allows to consider essentially any stationary signal as the output of a system driven by a normalized (namely white) noise process, and that we can compute the power of z via the $\mathbb{L}^2$ -norm:

Theorem 7.16.2 (power and the $\mathbb{L}^2$ -norm). The power $\mathbb{E} z^2(t)$ of the output of a stable system $z = Hw$ driven by unit intensity white noise input w equals

$$\mathbb{E} z^2(t) = \|H\|_{\mathbb{L}^2}^2.$$

□

⁵ $|\mathbb{E} z(t)z(t + \tau)| < M e^{-\sigma|\tau|}$ for some $M, \sigma > 0$. This is a weak assumption.

⁶ ϕ_z is defined as the Fourier transform of the covariance function, $\phi_z(\omega) = \int_{-\infty}^{\infty} \mathbb{E} z(t)z^*(t + \tau) e^{-i\omega\tau} d\tau$, but for our purposes we can take (7.21) to be its definition.

Example 7.16.3. Generating continuous-time white noise w is problematic because of its infinite power. An approximation can be obtained by taking w piecewise constant on each $[k\varepsilon, (k+1)\varepsilon]$ and independent on disjoint intervals, all with variance $1/\varepsilon$. If ε times the bandwidth of the system H is much smaller than 1 then the output $z = Hw$ is close to what we want. The data of Fig. 7.8 was generated with this MATLAB code:

```
ep=0.01;
t=0:ep:100;
w=randn(size(t))/sqrt(ep);
A=-.1; B=1; C=1; D=0;           % H(s)=1/(s+0.1), bandwidth=0.1
x=0*w;                           % initialization: x(1)=0
for j=1:(length(t)-1)
    x(j+1)=x(j)+ep*(A*x(j)+B*w(j)); % ensure no smoothing of input
end
z=x;                               % z=x in our case
plot(t,z)
%z=lsim(A,B,C,D,w,t);             % should give similar results
```

The transfer function here is $H(s) = 1/(s + 0.1)$ so that

$$\|H\|_{\mathbb{L}^2} = \sqrt{5}.$$

The power of z hence is 5 which is consistent with Fig. 7.8. □

MIMO systems and signals. In case the signals z and w have more than one component the above theorem should be modified to:

Theorem 7.16.4 (MIMO spectral density). For any zero-mean stationary signal $z(t)$ with finite power and whose covariance function $E z(t)z'(t + \tau) \in \mathbb{R}^{n_z \times n_z}$ decays exponentially fast, the (power) spectral density

$$\phi_z(\omega) = \lim_{T \rightarrow \infty} E \frac{1}{T} z_T(i\omega) z_T'(-i\omega)$$

is a well defined matrix and

$$E z(t)z'(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \phi_z(\omega) d\omega.$$

In particular the sum of powers $E z_i^2(t)$ of all its components equals

$$\sum_i E z_i^2(t) = \frac{1}{2\pi} \int_{-\infty}^{\infty} \text{tr} \phi_z(\omega) d\omega.$$

□

Here tr denotes the *trace* of the matrix, i.e. the sum of the diagonal entries of the matrix.

7.17 IYWTKM

- There is a stochastic equivalent of Hilbert.
- It is tempting to introduce $\mathbb{W}(0, 1)$ as a short-hand for: white noise process of zero mean and standard deviation 1.

- cross covariance, cross spectral density.
- A pleasant read is *Levinson's Algorithm, Wold's Decomposition, and Spectral Estimation* by Papoulis from 1985, available on the internet.
- Missing in this chapter is connections between finite dimensional representations and covariance functions

7.18 Problems

7.1 Consider the system

$$y[n] - 1.97y[n-1] + 0.98y[n-2] = u[n]$$

- Matlab: Generate 1000 samples of zero-mean unit-variance u and use it to find 1000 samples y . Plot y as a function of time and relate its behavior to the poles of the system
- Plot the spectral density function of the process y for $\omega \in [0, \pi)$
- Numerically compute the power of y (based on the samples) and compare it with the plot of the spectral density function.
- Find in MATLAB a command to compute the ℓ^2 -norm of the system and relate the outcome to the above power.
- Estimate the power spectral density using the periodogram (see code at the end of Section 7.14). What do you notice?

7.2 a) For which $\eta \in \mathbb{R}$ is

$$\phi(\omega) := 1 + \eta \cos(\omega)$$

a possible spectral density function? (I.e. that a WSS process exists having this spectral density, see also the next part).

- Find a stable system $y = \mathcal{H}(u)$ (driven by zero-mean, unit-variance white noise) that has spectral density

$$\phi_y(\omega) = 1 + \eta \cos(\omega).$$

(assuming η satisfies the conditions of part 1 of this exercise).

7.3 For which $a \in \mathbb{R}$ is

$$r[\tau] = \begin{cases} 1 & \text{if } \tau = 0 \\ a & \text{if } \tau = 1 \\ 0 & \text{elsewhere} \end{cases}$$

a covariance function?

7.4 Fill in the missing details of Example 7.13.2.

7.5 Show that $\phi(\omega) = \phi(-\omega)$.

8 Least squares

The remaining chapters sketches a series of applications of the theory presented in the preceding chapters. The first application is approximation of data by simpler forms.

8.1 Approximating data pairs by straight lines

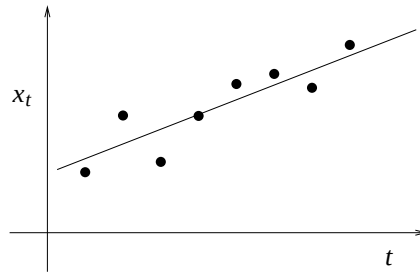


Figure 8.1: Approximating data by a single straight line

Figure 8.1 depicts one of the classical approximation problems:

Application–1: Approximate many pairs (t_k, x_k) , $k = 1, 2, \dots, N$ by a single straight line.

This can be seen as an application of the projection theorem. Write the candidate straight line (affine variety) as

$$x(t) = y_1 + y_2 t \quad \text{with } y_1, y_2 \in \mathbb{R}.$$

Then the problem is to find y_1, y_2 such that ideally $x_k = y_1 + y_2 t_k$ for all k or, practically, such that the approximation errors $\varepsilon_k := x_k - (y_1 + y_2 t_k)$ are small in some sense. In the context of projection this is achieved by minimizing for instance the Euclidean norm of the vector ε of approximation errors,

$$y_* = \operatorname{argmin}_y \left\| \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}}_x - \underbrace{\begin{bmatrix} 1 & t_1 \\ 1 & t_2 \\ \vdots & \vdots \\ 1 & t_N \end{bmatrix}}_A \underbrace{\begin{bmatrix} y_1 \\ y_2 \end{bmatrix}}_y \right\|.$$

Lemma 2.8.1 says that the solution y is the solution of the normal equations

$$\begin{aligned}
 A' Ay = A' x &\iff \begin{bmatrix} N & \sum t_k \\ \sum t_k & \sum t_k^2 \end{bmatrix} y = \begin{bmatrix} \sum x_k \\ \sum t_k x_k \end{bmatrix} \\
 &\iff y = \begin{bmatrix} N & \sum t_k \\ \sum t_k & \sum t_k^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum x_k \\ \sum t_k x_k \end{bmatrix} \\
 &\iff y = \frac{1}{N \sum t_k^2 - (\sum t_k)^2} \begin{bmatrix} \sum t_k^2 & -\sum t_k \\ -\sum t_k & N \end{bmatrix} \begin{bmatrix} \sum x_k \\ \sum t_k x_k \end{bmatrix} \quad (8.1)
 \end{aligned}$$

Or, in MATLAB:

```

N=20; % some integer>0
t=((1:N)'+rand(N,1))/N; % some vector t
x=-1+2*t+rand(N,1)/5; % some vector x
A=[ones(N,1) t];
y=A\x

```

Sensitivity. For the sake of analysis let us fix N , t_k and x_k :

$$N = 10, \quad t_k = \frac{k}{10}, \quad x_k = e^{t_k}$$

Straightforward application of the projection theorem gives us that

$$0.82886 + 1.77594t$$

is the best fit, see Fig. 8.2(a). Part (b) of this figure shows the (y_1, y_2) plane with the optimal y_* indicated, but, more importantly, it shows level sets of approximation error $\{y \mid |x - Ay| = \text{constant}\}$, so it shows how changes in y_1, y_2 affect the error. Clearly a change in y_2 has a more severe effect than a change in y_1 . The level sets are ellipses and they can be computed as follows: let Δ denote the vector of change $\Delta = y - y_*$, then

$$|\varepsilon_y|^2 = |x - A(y_* + \Delta)|^2 = |(x - Ay_*) + A\Delta|^2 = |x - Ay_*|^2 + \Delta A' A \Delta.$$

Thus the level sets (of constant error norm $|\varepsilon_y|$) are

$$\{\Delta \in \mathbb{R}^2 \mid \Delta' A' A \Delta = \text{constant} > 0\} \quad (8.2)$$

and these are ellipses, see Problem 8.7.

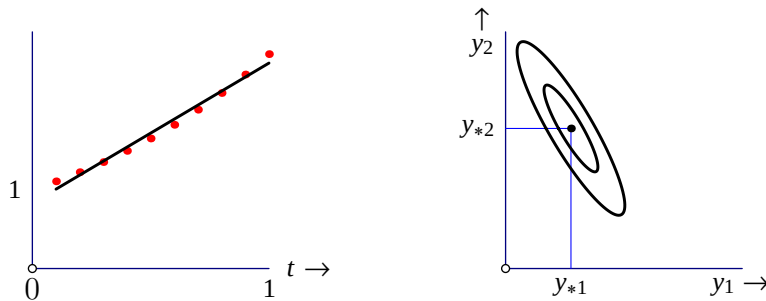


Figure 8.2: Least squares fit and sensitivity

Normalized error. For this particular problem the error norm $|\varepsilon|$ is not an appropriate indicator. Better is to compute the *root mean square (RMS)*

$$\sqrt{\frac{1}{N}|\varepsilon|^2} = \sqrt{\frac{1}{N} \sum_k (x_k - x(t_k))^2}.$$

It has the same dimension as x_k (meters, seconds or whatever) and it typically converges as $N \rightarrow \infty$ (whereas typically $|\varepsilon| \rightarrow \infty$ if $N \rightarrow \infty$).

8.2 Approximating data pairs by a linear combination of functions

Similar is the following problem:

Application–2: Approximate many pairs (t_k, x_k) , $k = 1, 2, \dots, N$ by a single polynomial of certain degree or by a single linear combination of known functions.

This is best explained on an example. Suppose we want to approximate the (t_k, x_k) by a polynomial $p(t)$ of degree at most n ,

$$p(t) = y_1 + y_2 t + \dots + y_{n+1} t^n, \quad y_1, y_2, \dots, y_{n+1} \in \mathbb{R}.$$

To find y we again minimize the Euclidean norm $|\varepsilon|$ of the vector of equation errors $\varepsilon_k := x_k - p(t_k)$:

$$y_* = \operatorname{argmin}_y \left\| \underbrace{\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_N \end{bmatrix}}_{x:=} - \underbrace{\begin{bmatrix} 1 & t_1 & t_1^2 & t_1^3 & \dots & t_1^n \\ 1 & t_2 & t_2^2 & t_2^3 & \dots & t_2^n \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & t_N & t_N^2 & t_N^3 & \dots & t_N^n \end{bmatrix}}_{A:=} \underbrace{\begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ \vdots \\ y_{n+1} \end{bmatrix}}_{y:=} \right\|.$$

(This matrix A with its special structure is known as a *Vandermonde matrix*.) Thus finding a suitable polynomial approximation of data can be seen as projecting the data vector x onto the subspace spanned by the columns of A .

Example 8.2.1 (Matlab). Let us approximate e^t on the interval $[0, 1]$ by a polynomial of degree $n = 4$. For the moment we restrict ourselves to approximating (t_k, e^{t_k}) for some uniformly gridded $t_1 = 0, \dots, t_{20} = 1$. The MATLAB code is:

```
N=21; % lets take 21 pairs
t=(0:(N-1))/(N-1); % t = 0, 1/20 ... , 1
x=exp(t); % x_k=exp(t_k)
n=4; % degree of polynomial
A=ones(N,n+1); % first column of A
for k=1:n
    A(:,k+1)=A(:,k).*t; % other columns of A
end
ys=(A'*A)\A'*x; % y_star (we may also do ys=A\x)
xs=A*ys; % polynomial p(t) evaluated at t_k
plot(t,x-xs,'*-'); % plot approximation error
```

The outcome of the MATLAB code is $y_* = (1.000031, 0.998639, 0.510167, 0.139870, 0.069542)$ so the optimal approximating polynomial is

$$p_*(t) = 1.000031 + 0.998639t + 0.510167t^2 + 0.139870t^3 + 0.069542t^4.$$

Notice the similarity with the Taylor series expansion of $e^t = 1 + t + \frac{1}{2!}t^2 + \dots$ around zero, but whereas this Taylor series is designed to approximate well around zero, the least squares fit $p_*(t)$ is designed to be “equally good” over the entire range $t \in [0, 1]$. According to the projection theorem the vector of approximation errors $\varepsilon = x - Ay_*$ satisfies

$$|\varepsilon|^2 = |x|^2 - |Ay_*|^2 = 8.322 \times 10^{-9}$$

The *root mean square* of the error is

$$\sqrt{\frac{1}{N}|\varepsilon|^2} = 1.99 \times 10^{-5}.$$

The error ε is shown in Fig. 8.3. □

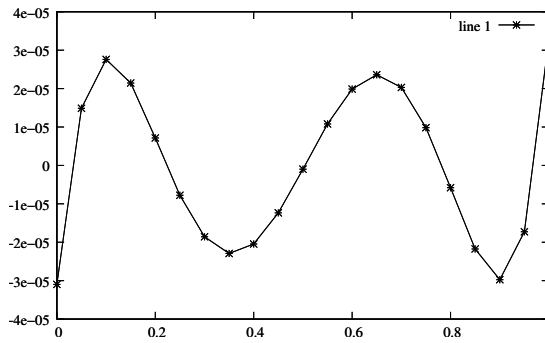


Figure 8.3: Matlab (octave) plot of error $\varepsilon := x - Ay_*$, as a function of t_k ;
Example 8.2.1

8.3 Approximate functions by other functions

The difference between a finite number of pairs (t_k, x_k) and a graph $(t, f(t))$ is conceptually small.

Application–3: Approximate a function $f(t)$, $t \in [a, b]$ by a single polynomial of certain degree or by a single linear combination of known functions.

Example 8.3.1 (continuous function). Let us approximate e^t on $[0, 1]$ by a polynomial $p(t)$ of degree $n = 3$,

$$p(t) = y_1 + y_2t + y_3t^2 + y_4t^3.$$

That is, we want to project the exponential function onto the subspace $\mathbb{P}_3$ of polynomials of degree at most 3. According to Lemma 5.11.1 given any inner product $\langle \cdot, \cdot \rangle$ on $L^2_{\mathbb{R}}([0, 1])$ the

optimal y_* satisfies

$$y_* = \begin{bmatrix} \langle 1, 1 \rangle & \langle 1, t \rangle & \langle 1, t^2 \rangle & \langle 1, t^3 \rangle \\ \langle t, 1 \rangle & \langle t, t \rangle & \langle t, t^2 \rangle & \langle t, t^3 \rangle \\ \langle t^2, 1 \rangle & \langle t^2, t \rangle & \langle t^2, t^2 \rangle & \langle t^2, t^3 \rangle \\ \langle t^3, 1 \rangle & \langle t^3, t \rangle & \langle t^3, t^2 \rangle & \langle t^3, t^3 \rangle \end{bmatrix}^{-1} \begin{bmatrix} \langle e^t, 1 \rangle \\ \langle e^t, t \rangle \\ \langle e^t, t^2 \rangle \\ \langle e^t, t^3 \rangle \end{bmatrix}$$

which if we take the standard inner product $\langle f, g \rangle = \int_0^1 g(t) f(t) dt$ becomes

$$y_* = \begin{bmatrix} \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{bmatrix}^{-1} \begin{bmatrix} e - 1 \\ 1 \\ e - 2 \\ 6 - 3e \end{bmatrix}$$

The matrix here is known as the *Hilbert matrix* (of order 4). We also used here that $\langle e^t, 1 \rangle = e - 1$ and the recursion $\langle e^t, t^k \rangle = e - k \langle e^t, t^{k-1} \rangle$ for $k > 0$ (see Problem 8.1).

$$p_*(t) = 0.99906 + 1.01830 t + 0.42125 t^2 + 0.27863 t^3.$$

and the norm of the error $e^t - p_*(t)$ can be expressed in many different ways,

$$\begin{aligned} \|e^t - p_*\|^2 &= \|e^t\|^2 - \|p_*\|^2 = \|e^t\|^2 - \langle p_*, p_* \rangle = \|e^t\|^2 - \langle p_*, e^t \rangle = \langle e^t - p_*, e^t \rangle \\ &= \|e^t\|^2 - \langle \mathcal{P}y_*, e^t \rangle = \|e^t\|^2 - \langle y_*, \mathcal{P}^* e^t \rangle \\ &= \|e^t\|^2 - y_*' \begin{bmatrix} e - 1 \\ 1 \\ e - 2 \\ 6 - 3e \end{bmatrix} \\ &= \frac{e^2 - 1}{2} - \begin{bmatrix} e - 1 \\ 1 \\ e - 2 \\ 6 - 3e \end{bmatrix}' \begin{bmatrix} \frac{1}{1} & \frac{1}{2} & \frac{1}{3} & \frac{1}{4} \\ \frac{1}{2} & \frac{1}{3} & \frac{1}{4} & \frac{1}{5} \\ \frac{1}{3} & \frac{1}{4} & \frac{1}{5} & \frac{1}{6} \\ \frac{1}{4} & \frac{1}{5} & \frac{1}{6} & \frac{1}{7} \end{bmatrix}^{-1} \begin{bmatrix} e - 1 \\ 1 \\ e - 2 \\ 6 - 3e \end{bmatrix} \\ &= 1.1 \times 10^{-7}. \end{aligned}$$

Up to the last step the computation is exact. The “root mean square” is now simply the norm (since the interval has length 1),

$$\|e^t - p_*\| = 3.3 \times 10^{-4}.$$

□

We could also choose to weight the errors, for instance perhaps the *relative* error needs to be small. To minimize the relative error of $x(t) - p(t)$ we could choose to minimize

$$\min_y \left\| 1 - \frac{p(t)}{x(t)} \right\|.$$

This is again a projection problem. Of course it assumes that the function $x(t)$ we want to approximate is nonzero on the interval, or does it?

8.4 Stochastic – linear estimators

Later.....

Suppose now that we have some knowledge of stochastic properties of ε . To begin with, assume that the entries ε_t of ε come from a zero mean white noise process. Then

$$\text{var}(\varepsilon) = E(\varepsilon\varepsilon^T) = \sigma^2 I_N. \quad (8.3)$$

Because of the relation $X = W\theta + \varepsilon$ it follows that also X is random, and therefore any sensible estimator $\hat{\theta} = S(X)$ as well. In the least squares problem the estimator $\hat{\theta}$ turns out to be linear in the observations X . To keep the problems tractable we will now limit the class of estimators to *linear* estimators, that is, we want the estimator $\hat{\theta}$ to be of the form

$$\hat{\theta} = KX, \quad \text{for some matrix } K. \quad (8.4)$$

Since we know that $\text{var}(\varepsilon) = \sigma^2 I_N$, the least squares solution that aims to minimize $\|\varepsilon\|^2$ is no longer well motivated. What is well motivated is to find an estimate $\hat{\theta}$ of θ that minimizes

$$E \|\theta - \hat{\theta}\|^2 = \sum_j E(\theta_j - \hat{\theta}_j)^2. \quad (8.5)$$

Lemma 8.4.1 (Linear unbiased minimum variance estimator). Consider $X = W\theta + \varepsilon$ and assume that $E\varepsilon = 0$ and $\text{var}(\varepsilon) = \sigma^2 I_N$. The following holds.

1. A linear estimator $\hat{\theta} = KX$ of θ is unbiased if and only if $KW = I$. In particular this shows that linear unbiased estimators exist if and only if W has full column rank;
2. The linear unbiased estimator $\hat{\theta} = KX$ that minimizes $E \|\theta - \hat{\theta}\|^2$ is

$$\hat{\theta} = (W^T W)^{-1} W^T X \quad (8.6)$$

in which case

$$\text{var}(\hat{\theta}) = \sigma^2 (W^T W)^{-1} \quad \text{and} \quad E \|\theta - \hat{\theta}\|^2 = \sigma^2 \text{tr}(W^T W)^{-1}; \quad (8.7)$$

3. For any linear unbiased estimator p of θ there holds that

$$\text{var}(p) \geq \text{var}(\hat{\theta}) \quad (8.8)$$

where $\hat{\theta}$ is the estimator defined in (8.6).

□

Proof. Follows later... ■

8.5 If you want to know more

- In statistics, this form least squares is closely related to *Linear regression*.
- See http://en.wikipedia.org/wiki/Least_squares
- See http://en.wikipedia.org/wiki/Curve_fitting

- See <http://www.dartmouth.edu/~chemlab/info/resources/linear/linear.html> for a nice and simple linear least squares java applet.
- There are million of application of least squares fit. For instance, google for image "least squares" to learn about least squares in image processing
- See http://en.wikipedia.org/wiki/Linear_least_squares
- We do not bother about numerical problems (the Hilbert matrix is badly conditioned meaning that numerical inversion is error prone). Having a orthogonal basis reduces numerical problems.

8.6 Problems

- 8.1 MATLAB: approximate $f(t) = \sin(t)$ on $[-1, 1]$ with a polynomial $p(t)$ of degree $n = 7$ using a "fine enough" grid. Plot the error $f(t) - p(t)$. Prove that the optimal polynomial $p_*(t)$ is odd [that is, $p_*(t) = p_1t + p_3t^3 + p_5t^5 + p_7t^7$]
- 8.2 Generate 100 pairs (x, y, z) in $\mathbb{R}^3$ with

```
N=100;
x=10*randn(1,N);
y=10*randn(1,N);
z=x+y+randn(1,N);
```

and then find a least squares approximation "plane" $z = a_1x + a_2y$.

- 8.3 Use Cauchy-Schwarz to show that the determinant $N \sum t_k - (\sum t_k)^2 \geq 0$ (used in the least squares problem (8.1)) and under what conditions on vector t is the determinant zero? And what does this condition mean for the least squares problem of Section 8.1.
- 8.4 MATLAB approximate on $[-1, 1]$ the function

$$f(t) = \begin{cases} 1 & \text{if } |t| \leq 0.5 \\ 0 & \text{elsewhere} \end{cases}$$

by a linear combination of cosines

$$c(t) = \sum_{k=0}^{10} y_k \cos(\pi t)$$

Plot both $f(t)$ and $c(t)$. (It is okay to grid the time axis.)

- 8.5 MATLAB: approximate on $[0, 1]$ the function t^4 by a polynomial of degree 3. Use as inner product $\langle f, g \rangle = \int_0^1 g(t) f(t) dt$.
- 8.6 Repeat the previous problem but now solve it analytically. [Hint: use Legendre polynomials up to and including degree 4.]
- 8.7 Show that (8.2) is an ellipse in $\mathbb{R}^2$ iff $A'A \in \mathbb{R}^2$ is invertible.

9 AR-models, Linear predictive coding

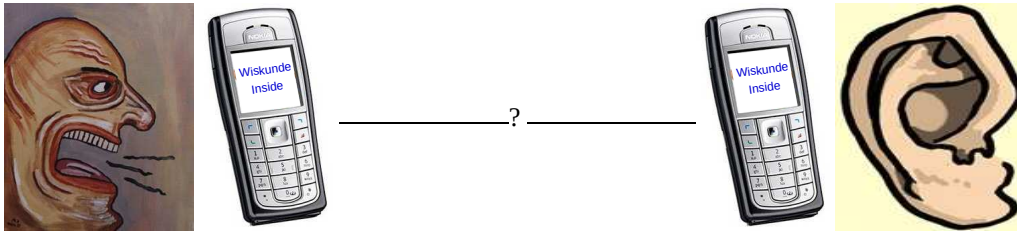


Figure 9.1: What is being transmitted?

Application–4: Encode speech so that it is (a) cheap to code at one end, (b) cheap to transmit and (c) cheap and robust to decode at the other end.

We will explain the gsm-standard, at least the main part (the work horse) of the gsm-standard known as *linear predictive coding (LPC)*.

A mobile telephone samples speech 8000 times per second and per set of $N = 160$ samples $x_1, x_2, \dots, x_N$ it tries to squeeze out structure that this sequence may have before transmitting the data. Quite ideal would be if the data satisfies a difference equation for certain coefficients $a_{n1}, \dots, a_{nn}$:

$$x_k = -a_{n1}x_{k-1} - a_{n2}x_{k-2} - \dots - a_{nn}x_{k-n}, \quad k = n+1, \dots, N.$$

That way the entire sequence of N samples is determined by the coefficients $a_{n1}, \dots, a_{nn}$ and initial conditions $x_1, \dots, x_n$. In reality some residuals c_k (small correction terms) remain:

$$x_k = -a_{n1}x_{k-1} - a_{n2}x_{k-2} - \dots - a_{nn}x_{k-n} \boxed{+c_k} \quad k = n+1, \dots, N. \quad (9.1)$$

Mobile telephones take $N = 160$ and $n = 8$ and they determine eight numbers $a_{81}, \dots, a_{88}$ so that $c_9^2 + c_{10}^2 + \dots + c_{160}^2$ is minimal. This is a least-squares problem but one with a special structure, a structure that we shall exploit (and so does your mobile telephone) because it will allow us to ensure that speech can be safely decoded on the other end of the line. In summary so far: your mobile telephone transmits¹

$$x_1, \dots, x_n, \quad a_{n1}, \dots, a_{nn}, \quad c_{n+1}, \dots, c_N. \quad (9.2)$$

The bulky part are the residuals c_k but since these c_k are small we can hope to that these can be rounded off without much quality degradation of the speech. These transmitted data are enough to recover at the other end of the line the speech by passing it through (9.1).

Due to rounding of numbers (intrinsic in digital communication) the data (9.2) is never going to be received error free at the other end of the line. In particular it could happen that the a_{nk} perceived at the other end of the line makes system (9.1) an unstable system! Below we develop a method (part of the gsm-standard) such that

¹save some details that don't concern us here

- The c_k are “easy” to compute;
- The data that is being transmitted is such that by construction the system (9.1) received at the other end is guaranteed to be stable.

9.1 Stable AR systems

Lemma 9.1.1 (stability). The system (9.1) with input c and output x is (*forward*) *stable* (in any sensible i/o-sense) iff the polynomial in z^{-1} ,

$$P(z) := 1 + a_{n1}z^{-1} + \cdots + a_{nn}z^{-n}$$

has all its zeros $z \in \mathbb{C}$ inside the unit circle, $|z| < 1$.

Proof. Since the transfer function is rational we need not be too precise about stability here (we all know what we mean). With *forward* stability we mean stability of the scheme when the x_k are computed forward in time (i.e. for increasing k). The transfer function from c to x is $1/P(z)$. ■

By the fundamental theorem of algebra $P(z) = 1 + a_{n1}z^{-1} + \cdots + a_{nn}z^{-n}$ has n zeros $\lambda_i \in \mathbb{C}$ and we have

$$P(z) = \prod_{i=1}^n (1 - \lambda_i z^{-1}).$$

In particular this shows that $|a_{nn}| < 1$ for stable systems because

$$a_{nn} = \prod_{i=1}^n -\lambda_i.$$

9.2 Least squares for AR systems, part 1

Since we want c_k to be small we want to express the c ’s as a function of the rest:

$$c_k = x_k + a_{n1}x_{k-1} + a_{n2}x_{k-2} + \cdots + a_{nn}x_{k-n}, \quad k = n+1, \dots, N$$

so the relevant equations become

$$a_* = \operatorname{argmin}_a \left\| \begin{bmatrix} x_{n+1} & x_n & \cdots & x_1 \\ x_{n+2} & x_{n+1} & \cdots & x_2 \\ \vdots & \vdots & \ddots & \vdots \\ x_N & x_{N-1} & \cdots & x_{N-n} \end{bmatrix} \begin{bmatrix} 1 \\ a_{n1} \\ a_{n2} \\ \vdots \\ a_{nn} \end{bmatrix} \right\|$$

This is not yet in standard form (but who cares). The dimensions of the matrices are dictated by the order of the scheme n and the available data x_k : stacking a row on top are at the bottom would require x_0 or x_{N+1} respectively, both of which are unavailable. In some (though unlikely) cases the so determined a_* may render (9.1) an unstable system.

9.3 Least squares for AR systems, part 2: Toeplitz

An elegant method that will guarantee stability of the system for optimal a_* is to enforce more symmetry. For that we're going to expand the LS equations somewhat: let us assume for the sake of argument that the unavailable x_k (so for $k \leq 0$ and $k > N$) are zero. Then the LS equations can be expanded to

$$a_* = \operatorname{argmin}_{a \in \mathbb{R}^n} \left| \begin{array}{c} \begin{bmatrix} x_1 & 0 & \dots & \dots & 0 \\ x_2 & x_1 & \ddots & & \vdots \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ x_n & x_{n-1} & \dots & x_1 & 0 \\ \hline x_{n+1} & x_n & \dots & \dots & x_1 \\ x_{n+2} & x_{n+1} & \dots & \dots & x_2 \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ x_N & x_{N-1} & \dots & \dots & x_{N-n} \\ \hline 0 & x_N & \ddots & & \vdots \\ \vdots & 0 & x_N & \ddots & \vdots \\ 0 & \dots & \dots & 0 & x_N \end{bmatrix} \\ \underbrace{\hspace{10em}}_W \end{array} \right| \begin{bmatrix} 1 \\ a_{n1} \\ a_{n2} \\ \vdots \\ a_{nn} \end{bmatrix} \quad (9.3)$$

Realize that the middle part is the bulk of the equation so the solution a_* may be expected to be close to that of the previous section. The beauty of the current expansion however is that the columns of the big matrix W are identical up to a shift in position. Each column contains the entire data vector $(x_1, x_2, \dots, x_N)$ appended and/or prepended by zeros. The implication (two lemmas) are interesting, to say the least:

Lemma 9.3.1. Let W be as defined in (9.3). Then a_* solves (9.3) iff it satisfies the linear equation in $a_* \in \mathbb{R}^n$ and $\gamma_n \in \mathbb{R}$,

$$W'W \begin{bmatrix} 1 \\ a_* \end{bmatrix} = \begin{bmatrix} \gamma_n \\ 0 \\ \vdots \\ 0 \end{bmatrix}. \quad (9.4)$$

Moreover in this case γ_n equals the minimal norm squared, $\gamma_n = |W \begin{bmatrix} 1 \\ a_* \end{bmatrix}|^2$.

Proof. Problem 9.3. ■

The fact is that $W'W$ is both *Toeplitz* (i.e. constant along diagonals) and symmetric,

$$W'W = \begin{bmatrix} r_{11} & r_{12} & r_{13} & \ddots \\ r_{12} & r_{11} & r_{12} & \ddots \\ r_{13} & r_{12} & r_{11} & \ddots \\ \ddots & \ddots & \ddots & \ddots \end{bmatrix} \in \mathbb{R}^{(n+1) \times (n+1)}$$

For general Toeplitz, symmetric matrices there holds the following:

Lemma 9.3.2 (Toeplitz-symmetric matrices). Suppose $R \in \mathbb{R}^{(n+1) \times (n+1)}$ symmetric Toeplitz matrix. The following statements are equivalent.

1. $R = W'W$ for some W , and R is invertible.
2. All eigenvalues of R are real and strictly positive.
3. There is a unique solution $a_* \in \mathbb{R}^n$, $\gamma_n \in \mathbb{R}$ of $R \begin{bmatrix} 1 \\ a \end{bmatrix} = \begin{bmatrix} \gamma_n \\ 0 \end{bmatrix}$, with $\gamma_n > 0$ and $P(z) := 1 + a_{*1}z^{-1} + \dots + a_{*n}z^{-n}$ has all its zeros inside the unit circle ($|z| < 1$).
4. R has a decomposition of the form

$$\begin{bmatrix} 1 & a_{n1} & a_{n2} & \dots & a_{nn} \\ 0 & 1 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & a_{21} & a_{22} \\ \vdots & & \ddots & 1 & a_{11} \\ 0 & \dots & \dots & 0 & 1 \end{bmatrix} R \begin{bmatrix} 1 & 0 & \dots & \dots & 0 \\ a_{n1} & 1 & \ddots & & \vdots \\ a_{n2} & \ddots & \ddots & \ddots & \vdots \\ \vdots & & a_{21} & 1 & 0 \\ a_{nn} & \dots & a_{22} & a_{11} & 1 \end{bmatrix} = \begin{bmatrix} \gamma_n & 0 & \dots & \dots & 0 \\ 0 & \gamma_{n-1} & \ddots & & \vdots \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \gamma_1 & 0 \\ 0 & \dots & \dots & 0 & \gamma_0 \end{bmatrix} \quad (9.5)$$

with $\gamma_0 > 0$ and $|a_{kk}| < 1$ for all $k = 1, \dots, n$.

Proof. We prove the for us important implication: that $P(z)$ has its zeros inside the unit circle if the other conditions hold. The lemma is correct for $n = 1$ (see Problem 9.4). Now let $\lambda \in \mathbb{C}$ be a zero of $P(z)$, that is

$$P(z) = (1 - \lambda z^{-1})(1 + c_1 z^{-1} + \dots + c_{n-1} z^{-1}).$$

This means in terms of the coefficients

$$\begin{bmatrix} 1 \\ a_{n1} \\ a_{n2} \\ \vdots \\ a_{nn} \end{bmatrix} = \underbrace{\begin{bmatrix} 1 & 0 \\ c_1 & 1 \\ c_2 & c_1 \\ \vdots & \vdots \\ 0 & c_n \end{bmatrix}}_C \begin{bmatrix} 1 \\ -\lambda \end{bmatrix}.$$

Then $C'RC \begin{bmatrix} 1 \\ -\lambda \end{bmatrix}$ is of the form $\begin{bmatrix} * \\ 0 \end{bmatrix} \in \mathbb{R}^2$ and since $C'RC$ is symmetric and Toeplitz it follows from the $n = 1$ case (Problem 9.4) that $|\lambda| < 1$. Since λ was an arbitrary zero of $P(z)$, all zeros of $P(z)$ are inside the unit circle. ■

An implication of Lemma 9.3.1 is that the γ_m in (9.5) equals the square of the norm of the vector of residuals if we take an m th order model (9.1). Obviously with more degrees of freedom we can achieve more, so the γ_k are non-increasing,

$$\gamma_0 \geq \gamma_1 \geq \gamma_2 \geq \dots \geq \gamma_n \geq 0$$

if $R = W'W$. Also we have $\det R = \prod_{m=0}^n \gamma_m$. The nesting of (9.4) of Lemma 9.3.1 for different orders m into a single matrix factorization (9.5) suggests that the models for different orders can be solved recursively. Indeed, that is what the fast *Levinson-Durbin* algorithm (in its many different forms) is about. Along the way the algorithm computes the n numbers

$$a_{11}, a_{22}, \dots, a_{nn}$$

and the lemma guarantees that all these are in $(-1, 1)$. It is these (so called *reflection coefficients*) that your mobile telephone transmits, and not the n coefficients a_{nk} , $k = 1, 2, \dots, n$. The

reflection coefficients have the advantage of being scaled and that digitizing (rounding in $(-1, 1)$) guarantees a stable system. The reflection coefficients (together with initial conditions $x_1, \dots, x_n$ and residuals c_k) are enough to decode the sound at the other end of the line.

9.4 Problems

9.1 MATLAB: Generate artificial speech with

```
N=160;  
k=1:N;  
y=sin(k/10)+randn(size(k))/100;  
plot(k,y);
```

- a) Determine the models 9.1 for $n = 1, 2, 3, 4, 5, 6, 7, 8$ using the method of Section 9.2 and verify stability of all models.
- b) Determine the models 9.1 for $n = 1, 2, 3, 4, 5, 6, 7, 8$ using the method of Section 9.3 and verify stability of all models.

9.2 MATLAB: Write a MATLAB function that generates the matrix factorization (9.5) for a given symmetric Toeplitz R .

9.3 Prove Lemma 9.3.1.

9.4 Prove Lemma 9.3.2 for $n = 1$.

10 Finite horizon LQ control

This chapter emphasizes the application of projection theorems to finite horizon LQ control. There is a lot more to say about LQ than covered here.

10.1 Control the state

One of the fundamental problems in control is to control the state x of a system in state space form

$$\dot{x}(t) = Ax(t) + Bu(t), \quad x(t) \in \mathbb{R}^n, \quad u(t) \in \mathbb{R}^{n_u}, \quad t \in \mathbb{R}.$$

We consider the following problem.

Application–5: Given matrices A, B and zero initial state $x(0) = 0$ and desired final state x_1 at desired time $T > 0$, find an input $u : [0, T] \rightarrow \mathbb{R}^{n_u}$ that achieves $x(T) = x_1$.

How to tackle this problem? The functional analysis routine would be:

- $x(T)$ is linear in $u : [0, T] \rightarrow \mathbb{R}^{n_u}$. Hence we can write $x(T) = \mathcal{G}(u)$ for some linear mapping $\mathcal{G}$.
- The set of inputs u that achieve $x(T) = x_1$ is an affine variety (if a solution u exists): $u \in u_{\text{particular}} + \ker(\mathcal{G})$.
- Equip the set of inputs with an inner product, say consider $u \in L^2([0, T])$ with standard inner product.
- Equip the state space with inner product, say the dot product $\langle x, y \rangle_{\mathbb{R}^n} := y'x$.
- Then if the adjoint $\mathcal{G}^*$ is well defined, i.e. if for some $\mathcal{G}^*$ we have $\langle x, \mathcal{G}u \rangle_{\mathbb{R}^n} = \langle \mathcal{G}^*x, u \rangle_{L^2([0, T])}$ and if the normal equations have a solution: $u_* = \mathcal{G}^*(\mathcal{G}\mathcal{G}^*)^{-1}x_1$ then it is the minimal norm solution of Problem 5 [Theorem 5.15.2].
- Then the minimal norm solution u_* satisfies $\|u\|_{L^2[0, T]}^2 = \langle (\mathcal{G}^*\mathcal{G})^{-1}x_1, x_1 \rangle_{\mathbb{R}^n}$ [Theorem 5.15.2] i.e. the minimal input energy is quadratic in the desired final state.

What remains is to make $\mathcal{G}$ more explicit and hopefully in such a way that the solution u_* is well defined.

10.2 A first solution — controllability Gramian

Let $\Phi(t)$ be the *state transition matrix*, that is, the $n \times n$ solution of the differential equation $\dot{\Phi}(t) = A\Phi(t)$, $\Phi(0) = I_n$. Then

$$x(T) = \int_0^T \Phi(T - \tau)Bu(\tau) d\tau$$

is the solution of $\dot{x} = Ax + Bu$, $x(0) = 0$ (which follows from differentiation of the above $x(t)$). The adjoint of this mapping $\mathcal{G}$ is

$$u = \mathcal{G}^* \tilde{x} : \quad u(t) = B' \Phi'(T - t) \tilde{x}$$

(verify this). Note that the adjoint maps from $\mathbb{R}^n$ to functions on $[0, T]$. Now

$$\mathcal{G} \mathcal{G}^* \tilde{x} = \left(\int_0^T \Phi(T - \tau) B B' \Phi'(T - \tau) d\tau \right) \tilde{x}$$

which can also be expressed as

$$\mathcal{G} \mathcal{G}^* \tilde{x} = \left(\int_0^T \Phi(\tau) B B' \Phi'(\tau) d\tau \right) \tilde{x}.$$

Lemma 10.2.1. Let matrices A , B and final time T be given. The following statements are equivalent.

1. For any $x_1 \in \mathbb{R}^n$ Problem 5 has a solution u .
2. $\text{im}(\mathcal{G}) = \mathbb{R}^n$.
3. $\mathcal{G} \mathcal{G}^*$ is invertible on $\mathbb{R}^n$, that is, the $n \times n$ *controllability Gramian* $P_T \in \mathbb{R}^{n \times n}$ defined as

$$P_T := \int_0^T \Phi(\tau) B B' \Phi'(\tau) d\tau$$

is invertible.

4. For any $x_1 \in \mathbb{R}^n$ the minimal norm solution exists and equals

$$u_*(t) = \mathcal{G}^* (\mathcal{G} \mathcal{G}^*)^{-1} x_1 = B' \Phi'(T - t) P_T^{-1} x_1.$$

Moreover in that case $\|u_*\|^2 = x_1' P_T^{-1} x_1$.

Proof. (1) $\implies$ (2) is immediate. (2) $\implies$ (3) because if P_T would have been singular then $v' P_T = 0$ for some nonzero $v \in \mathbb{R}^n$ implying $0 = v' P_T v = \int_0^T \|v' \Phi(T - t) B\|^2 dt$. By continuity of Φ this means $v' \Phi(T - t) B = 0$ for all t so $x_1 := v \notin \text{im}(\mathcal{G})$. The remaining implications (3) $\implies$ (4) $\implies$ (1) and that $\|u_*\|^2 = x_1' P_T^{-1} x_1$ is an application of Theorem 5.15.2. ■

10.3 More explicit solution — matrix exponential

As we all know the transition matrix is the matrix exponential,

$$\Phi(t) = e^{At}.$$

So the controllability Gramian is

$$P_T = \left(\int_0^T e^{A\tau} B B' e^{A'\tau} d\tau \right).$$

and so Problem 5 is solvable iff P_T is invertible in which case the minimal norm solution is

$$u_*(t) = B' e^{A'(T-t)} \left(\int_0^T e^{A\tau} B B' e^{A'\tau} d\tau \right)^{-1} x_1$$

with $\|u_*\|^2 = x_1' P_T^{-1} x_1$. Since u_* is the minimal norm solution on $[0, T]$ we immediately get that the longer we allow the control to act the smaller the norm of u_* is:

Lemma 10.3.1. Assume A is Hurwitz and that Problem 5 has a solution for some $T > 0$ (i.e. $P_T > 0$ for some $T > 0$). The minimal input energy needed to steer $x(0) = 0$ to x_1 at some time $T > 0$ is achieved at $T = \infty$ and the minimal energy required is

$$\|u_*\|_{L^2(\mathbb{R}_+)}^2 = x_1 P^{-1} x_1 \quad \text{where} \quad P := \int_0^\infty e^{At} B B' e^{A't} dt.$$

Proof. Stability of A is sufficient to guarantee existence of P . The rest is immediate from the preceding discussion. ■

For this reason we can interpret the eigenvalues of P as a “measure” for controllability. The “closer” an eigenvalue of P is to zero the more difficult it is to activate the corresponding state eigenvector through u . This idea lies at the basis of Balanced Truncation and state space solution of Hankel norm approximations.

10.4 Minimal control input – embellishments

Lemma 10.4.1. Assume A is Hurwitz and let P be as defined in Lemma 10.3.1. The following statements are equivalent

1. $\dot{x} = Ax + Bu$ is controllable
2. $P = P' > 0$

Proof. Do it yourself. ■

Theorem 10.4.2 (Computation of Controllability Gramian). Assume A is Hurwitz. Then P of Lemma 10.3.1 is the unique solution of the linear Lyapunov equation

$$AP + PA' + BB' = 0.$$

Proof.

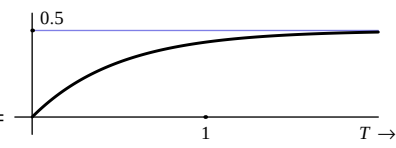
$$AP + PA' = \int_0^\infty A e^{At} B B' e^{A't} + e^{At} B B' e^{A't} A' dt = e^{At} B B' e^{A't} \Big|_0^\infty = -BB'.$$

(That the solution of the Lyapunov equation is unique for stable A can for instance be seen by this tricked argument: if $AP + PA' + BB' = 0$ and $AQ + QA' + BB' = 0$ then by linearity $\Delta = P - Q$ satisfies $A\Delta + \Delta A' = 0$. Now $\frac{d}{dt} e^{At} \Delta e^{A't} = 0$ so $e^{At} \Delta e^{A't}$ is constant. It is zero at $t = \infty$ and Δ at $t = 0$. Hence $\Delta = 0$.) ■

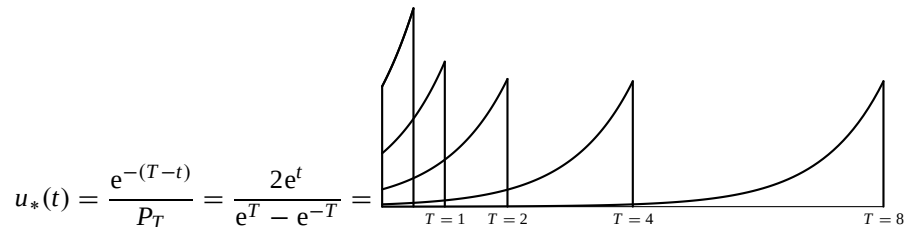
Example 10.4.3. Consider the scalar differential equation

$$\dot{x} = -x + u, \quad x(0) = 0, \quad x(T) = 1.$$

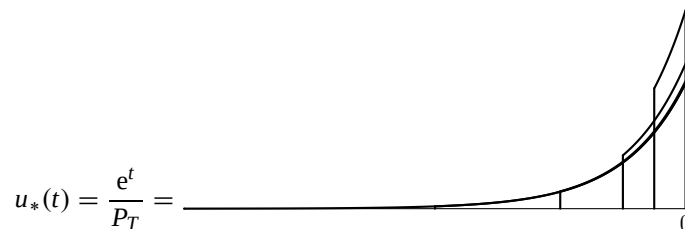
The finite-horizon controllability Gramian is

$$P_T = \int_0^T e^{-2t} dt = \frac{e^{-2T} - 1}{-2} = \frac{1 - e^{-2T}}{2} =$$


In the limit $T \rightarrow \infty$ the Gramian approaches $P = \frac{1}{2} > 0$ which indeed is the solution of the Lyapunov equation $AP + PA + BB' = -P - P + 1 = 0$. The optimal $u_* : [0, T] \rightarrow \mathbb{R}$ for various T is



The limit $T \rightarrow \infty$ of u_* is not well defined. For better comparison of the different u_* shift the problem from $[0, T]$ to $[-T, 0]$. Conclusion: the minimal norm input $u : [-T, 0] \rightarrow \mathbb{R}$ that steers $x(-T) = 0$ to $x(0) = x_1$ is



and as $T \rightarrow \infty$ the optimal solution converges to $u_*(t) = \frac{e^t}{P_\infty} = 2e^t$ for $t \in (-\infty, 0]$. $\square$

Example 10.4.4 (Controllability Gramian in Matlab/Octave). Assume A is stable. To compute the observability Gramian $P := P_\infty$ in OCTAVE you can do:

```
A=[ -1  1;  0 -1];
B=[ 1;  2];
P=lyap(A',B*B')
```

and in MATLAB:

```
A=[ -1  1;  0 -1];
B=[ 1;  2];
P=lyap(A,B*B')
```

(Spot the difference). The outcome is $P = \begin{bmatrix} 2.5 & 2 \\ 2 & 2 \end{bmatrix}$. $\square$

10.5 Minimal control input – state representation of adjoint

The optimal input

$$u_*(t) = B' \Phi'(T-t) P_T^{-1} x_1$$

can be seen as the output of a dynamical system. Indeed,

$$\frac{d}{dt} \Phi'(T-t) = -\dot{\Phi}'(T-t) = -A' \Phi'(T-t)$$

so that:

Lemma 10.5.1 (Adjoint dynamical system). The optimal $u_*(t)$ of Problem 5 is the output $u_*(t) = B'p(t)$ of the dynamical system

$$\dot{p}(t) = -A'p(t), \quad t \in [0, T]$$

with final condition $p(T) = P_T^{-1}x_1$. □

This system operates backwards in time.

10.6 Finite horizon LQ

Now it is time for the classical finite-horizon LQ problem. Consider the system

$$\begin{cases} \dot{x} = Ax + Bu, & x(0) = x_0 \\ z = Cx + Du \end{cases} \quad (10.1)$$

Application–6: Given real matrices A, B, C, D and horizon $T > 0$ and initial state $x(0) = x_0 \in \mathbb{R}^n$ find an input $u : [0, T] \rightarrow \mathbb{R}^{n_u}$ that minimizes

$$\|z\|_{L^2[0, T]}^2 = \int_0^T z'(t)z(t) dt.$$

This covers for instance the classic LQ problem if we choose matrices C and D such that

$$z = \begin{bmatrix} u \\ y = C_y x \end{bmatrix}.$$

What to do? Here are some ideas:

- Given $x(0)$ the output z is affine $u : [0, T] \rightarrow \mathbb{R}^{n_u}$:

$$z = Ce^{At}x_0 + \mathcal{G}u$$

- Equip the input and output space with inner product.
- Hence the projection theorem states that u_* is optimal iff

$$\langle Ce^{At}x_0 + \mathcal{G}u_*, \mathcal{G}(L^2[0, T]) \rangle = \langle \mathcal{G}^*(Ce^{At}x_0 + \mathcal{G}u_*), L^2[0, T] \rangle = 0.$$

- That is the case iff the normal equations are satisfied $\mathcal{G}^*(Ce^{At}x_0 + \mathcal{G}u_*) \equiv 0$.
- $\mathcal{G}^*\mathcal{G}$ maps functions to functions so the optimal $u_* = -(\mathcal{G}^*\mathcal{G})^{-1}\mathcal{G}^*Ce^{At}x_0$ may be difficult?
- Express the adjoint $\mathcal{G}^*$ in state space. That may help?

10.7 Solution of finite horizon LQ

We now will “explicitly” solve Problem 6. The formulae may appear messy but we believe that they arrive naturally given our understanding of the projection theorem and adjoints, etc. It makes sense to first try to come up with a state representation of the adjoint system $\mathcal{G}^*$.

Lemma 10.7.1 (Adjoint). The adjoint $\mathcal{G}^*$ of the mapping $\mathcal{G} : L^2[0, T] \rightarrow L^2[0, T]$ defined as

$$z = \mathcal{G}(u) : \begin{cases} \dot{x} = Ax + Bu \\ z = Cx + Du \end{cases}, \quad x(0) = 0$$

satisfies a state representation with final condition:

$$v = \mathcal{G}^*(z), \quad \begin{cases} \dot{p} = -A'p - C'z \\ v = B'p + D'z \end{cases}, \quad p(T) = 0. \quad (10.2)$$

Proof. Assume $D = 0$ (the $D \neq 0$ case you can do yourself, see Problem 10.3). Let $h(t)$ be the impulse response of $\mathcal{G}$: $h(t) = Ce^{At}B\mathbb{1}(t)$. Then

$$\begin{aligned} \langle \mathcal{G}u, z \rangle &= \int_0^T z'(t) \left[\int_0^T h(t-\tau)u(\tau) d\tau \right] dt \\ &= \int_0^T \int_0^T z'(t)h(t-\tau)u(\tau) dt d\tau \\ &= \int_0^T u'(\tau) \left[\int_0^T h'(t-\tau)z(t) dt \right] d\tau \\ &= \int_0^T u'(t) \left[\int_0^T h'(\tau-t)z(\tau) d\tau \right] dt \\ &= \int_0^T u'(t) \left(\int_t^T B'e^{-A'(t-\tau)}C'z(\tau) d\tau \right) dt =: \langle u, \mathcal{G}^*z \rangle. \end{aligned}$$

Therefore $\mathcal{G}^*z = \int_t^T B'e^{-A'(t-\tau)}C'z(\tau) d\tau$. Differentiating $p(t) := \int_t^T e^{-A'(t-\tau)}C'z(\tau) d\tau$ shows that $\dot{p} = -A'p - C'z$ with $p(T) = 0$. ■

The system interconnection $v = \mathcal{G}^*(Ce^{At}x_0 + \mathcal{G}u)$ (needed in the normal equation) is simply the series interconnection of $Ce^{At}x_0 + \mathcal{G}u$ and $\mathcal{G}^*$.

Lemma 10.7.2. u_* satisfies the normal equations

$$\mathcal{G}^*(Ce^{At}x_0 + \mathcal{G}u_*) = 0$$

if and only if there exist (x, p) such that

$$\begin{cases} \begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C'C & -A' \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix} + \begin{bmatrix} B \\ -C'D \end{bmatrix} u \\ 0 = \begin{bmatrix} D'C & B' \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix} + D'Du, \quad \forall t \in [0, T], \text{ with } x(0) = x_0, p(T) = 0. \end{cases} \quad (10.3)$$

Proof. Feed the output $z := (Ce^{At}x_0 + \mathcal{G}u_*)$ to the adjoint $v = \mathcal{G}^*(z)$ in (10.2) and set v to zero. ■

If $D'D > 0$ then u_* in the normal equations (10.3) necessarily satisfies

$$u_*(t) = -(D'D)^{-1} \begin{bmatrix} D'C & B' \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix}, \quad \forall t \in [0, T] \quad (10.4)$$

which when inserted into the dynamic part of (10.3) results in the *Hamiltonian* system

$$\begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \underbrace{\left(\begin{bmatrix} A & 0 \\ -C'C & -A' \end{bmatrix} - \begin{bmatrix} B \\ -C'D \end{bmatrix} (D'D)^{-1} \begin{bmatrix} D'C & B' \end{bmatrix} \right)}_H \begin{bmatrix} x \\ p \end{bmatrix}, \quad x(0) = x_0, p(T) = 0. \quad (10.5)$$

Lemma 10.7.3 (Solution finite horizon LQ). Assume $D'D$ is invertible. Then the optimal $u_* \in L^2[0, T]$ of Problem 6 exists and equals (10.4) where (x, p) is the solution of the state equation with two-point boundary condition (10.5). The solution (x, p) of (10.5) exists and is unique:

$$\begin{bmatrix} x(t) \\ p(t) \end{bmatrix} = \Sigma(t) \begin{bmatrix} I \\ P_T(T) \end{bmatrix} x_0, \quad P_T(T) = -\Sigma_{22}(T)^{-1} \Sigma_{21}(T).$$

Here $\Sigma(t) := \begin{bmatrix} \Sigma_{11}(t) & \Sigma_{12}(t) \\ \Sigma_{21}(t) & \Sigma_{22}(t) \end{bmatrix} := e^{Ht}$, and the square $\Sigma_{22}(T)$ is guaranteed to be nonsingular.

Proof. There is only the problem of existence and uniqueness of the solution (x, p) to this two-point boundary condition problem. This is not entirely trivial (but not too complicated either).

Consider for the moment the case $x_0 = 0$, so the normal equations reduce to

$$\mathcal{G}^* \mathcal{G} u_* = 0 \quad \text{and} \quad x(0) = p(T) = 0.$$

Clearly $\mathcal{G}^* \mathcal{G} u_* = 0$ implies $z := \mathcal{G} u_* = 0$ which in turn implies $u_* = 0$ because $\mathcal{G}$ is left invertible if $D'D > 0$. Now $u_* = 0$ implies $\begin{bmatrix} \dot{x} \\ \dot{p} \end{bmatrix} = \begin{bmatrix} A & 0 \\ -C'C & -A' \end{bmatrix} \begin{bmatrix} x \\ p \end{bmatrix}$ which, given that $x(0) = p(T) = 0$ in turn implies both x and p are trivial. Therefore for $x_0 = 0$ the solutions (x, p, u_*) of the normal equations are unique and are all equal to zero. For $\Sigma(t)$ defined as $\Sigma(t) = e^{Ht}$ the solution of (10.5) in terms of *initial* conditions is

$$\begin{bmatrix} x(t) \\ p(t) \end{bmatrix} = \Sigma(t) \begin{bmatrix} x(0) \\ p(0) \end{bmatrix}.$$

Since no non-trivial solution (x, p) exists for which $x(0) = p(T) = 0$ we must have that $\Sigma_{22}(T)$ is invertible. Now we can proceed.

Back to the general case ($x_0 \neq 0$). Since $\Sigma_{22}(T)$ is invertible we can solve $p(0)$ from

$$0 = p(T) = \Sigma_{21}(T)x_0 + \Sigma_{22}(T)p(0).$$

Its solution is $p(0) = -\Sigma_{22}(T)^{-1} \Sigma_{21}(T)x_0$. ■

In contrast to Problem 5 here the dynamic equations that determine the optimal solution $u_*(t)$ (backward in time) is coupled with dynamic equations for x (forward in time). It is possible decouple the equations, but we will not do it.

Example 10.7.4 (First order system). For the system

$$\dot{x} = u, \quad z = \begin{bmatrix} u \\ x \end{bmatrix}$$

the Hamiltonian H of (10.3) becomes

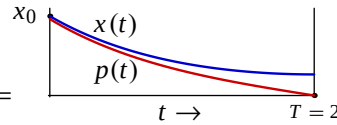
$$H = \begin{bmatrix} 0 & -1 \\ -1 & 0 \end{bmatrix}$$

which is simple enough to allow explicit solution of the (symplectic) matrix exponential,

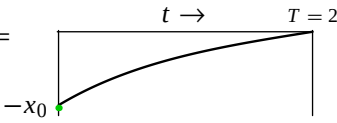
$$\Sigma(t) = e^{Ht} = \frac{1}{2} \begin{bmatrix} e^t + e^{-t} & -e^t + e^{-t} \\ -e^t + e^{-t} & e^t + e^{-t} \end{bmatrix}.$$

The state-costate solution (10.3) become

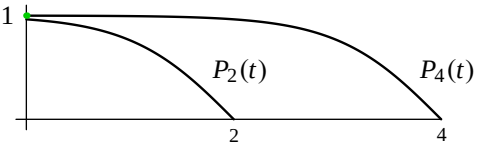
$$\begin{bmatrix} x(t) \\ p(t) \end{bmatrix} = \frac{1}{2} \begin{bmatrix} e^t + e^{-t} & -e^t + e^{-t} \\ -e^t + e^{-t} & e^t + e^{-t} \end{bmatrix} \begin{bmatrix} 1 \\ \frac{e^T - e^{-T}}{e^T + e^{-T}} \end{bmatrix} x_0 =$$



The optimal input

$$u_*(t) = -B'p(t) = -p(t) =$$


which we can also see as a function of state $x(t)$ as $u(t) = -B'P_T(t)x(t)$ with

$$P_T(t) = \frac{p(t)}{x(t)} =$$


The larger T is, the closer $P_T(t)$ is to 1 initially. It is now tempting to think that for the infinite horizon case the optimal $u(t)$ is a static state feedback $u_*(t) = -B'x(t) = -x(t)$. It is. $\square$

10.8 IYWTKM...

- The finite horizon solution can be extended to the infinite horizon case but the derivation is lengthy. We will solve the infinite horizon case for LTI systems from scratch (but many arguments will look similar to those presented in this chapter)
- The solution (x, p) provided in Lemma 10.7.3 in terms of matrix exponentials is not numerically reliable. Better (numerically) is to work with Riccati differential equations. The idea then being to express $p(t)$ as a static linear function of $x(t)$, that is, $p(t) = P(t)x(t)$ and then to derive a differential equation for $P(t)$. See for instance the classic books by Kwakernaak & Sivan and by Anderson & Moore (both available on line)

10.9 Problems

10.1 Consider the system

$$\dot{x} = x + 2u, \quad y = 2x$$

Given x_0 and horizon $T > 0$ find the input u that minimizes $\int_0^T y^2(t) + u^2(t) dt$. (That is, derive equation similar to that of Example 10.7.4). What would the infinite horizon feedback solution $u_*(t) = -Fx(t)$ be, i.e. what is F ?

10.2 Fill in the details of Example 10.7.4.

10.3 Lemma 10.7.1 is only proven for the case that $D \neq 0$. Extend the proof to the $D = 0$ case (and the extension of the proof must be short!).

10.4 Consider Problem 6. Show that $\frac{d}{dt} p'(t)x(t) = -z'(t)z(t)$ and use it to prove that $p'(t)x(t) = \|z\|_{L^2[t,T]}^2$.

10.5 Write a MATLAB script that computes the minimal norm $\|z_*\|_{L^2[0,T]}$ of Problem 6 given matrices A, B, C, D and time $T > 0$.

10.6 Make precise what is used in the proof of Lemma 10.7.3: $\mathcal{G}$ is left invertible over $[0, T]$ is $D'D > 0$.

10.7 Redo (as best as possible) the results obtained in Section 10.3 for Problem 5 slightly altered:

Application–7: Given matrices A, B and nonzero initial state $x(0) = x_0$ and desired final state x_1 at desired time $T > 0$, find an input $u : [0, T] \rightarrow \mathbb{R}^{n_u}$ that achieves $x(T) = x_1$.

11 $\mathcal{H}^2$ optimization in frequency domain

This chapter needs a re-write.

11.1 LQ in frequency domain

The approach taken in this and the following section is not standard, but elegant nonetheless (we think). It is an input-output approach to LQ, and relies heavily on frequency domain arguments. In doing so we naturally arrive at the familiar guaranteed gain and phase margins and, for SISO systems, a simple rule to determine the closed loop poles based on the open loop plant.

As a motivating example consider an LTI system with input u and output y ,

$$y = \mathcal{P}(u) \quad (11.1)$$

and suppose the plant was operational over the past $(-\infty, 0]$. The LQ-problem colloquially is to minimize $\int_0^\infty |u(t)|^2 + |y(t)|^2 dt$ with respect to all future-time inputs $u : [0, \infty) \rightarrow \mathbb{R}^{n_u}$. As in the finite-horizon case (Chapter 10) it is convenient to work with system representations in which both y and u are seen as outputs, i.e.,

$$\begin{bmatrix} u \\ y \end{bmatrix} = \begin{bmatrix} I \\ \mathcal{P} \end{bmatrix} u.$$

This way the input (here u) is to be optimized over and the output is to be minimized. The LQ-problem can be formulated as follows:

Application–8: Consider a causal system $z = \mathcal{G}(v)$ with given past $v_- : (-\infty, 0] \rightarrow \mathbb{C}^{n_v}$. Determine the $v_+ : [0, \infty) \rightarrow \mathbb{C}^{n_v}$ that minimizes $\|z\|_{\mathbb{L}^2[0, \infty)}$.

This formulation of the LQ-problem

- requires no notion of state,
- has a solution that depends on the given past v_- ,
- is *not* a feedback control problem, that is, we are not demanding that the minimizers of $\|z\|_{\mathbb{L}^2[0, \infty)}$ are to be sought within some class of feedback controllers. (It will be a *result* that for the usual state-space systems the optimal solutions can be implemented as a state-feedback controller.)

11.2 SISO LQ and return difference inequality

Definition 11.2.1 (Inner transfer matrix). A rational matrix G is said to be *inner* if it has its poles in the open left-half plane and $G^\sim G = I$. □

Example 11.2.2 (Inner matrix). Suppose $z = \mathcal{G}(v)$ is the causal stable system with transfer matrix

$$G(s) = \begin{bmatrix} \frac{s}{s+1} \\ \frac{1}{s+1} \end{bmatrix}.$$

The mapping is norm preserving because G is inner (see Example 6.9.3):

$$G^\sim(s)G(s) = \begin{bmatrix} -s & 1 \\ -s+1 & -s+1 \end{bmatrix} \begin{bmatrix} \frac{s}{s+1} \\ \frac{1}{s+1} \end{bmatrix} = \frac{-s^2}{-s^2+1} + \frac{1}{-s^2+1} = 1.$$

That is the input and output energy are the same

$$\|z\|_{\mathbb{L}^2(\mathbb{R})} = \|v\|_{\mathbb{L}^2(\mathbb{R})}.$$

Given the past v_- , the past of z_- is determined (because it is causal) hence the minimal future energy $\|z\|_{\mathbb{L}^2(\mathbb{R}_+)}^2$ is achieved by taking $v_+ = 0$.

For completeness (but you may skip this part if you don't care about it, this part is not essential for what is to come), the signal z for positive time can be computed using residues (assuming $v_- \in \mathbb{L}^2(-\infty, 0]$),

$$z(s) = \begin{bmatrix} \frac{s}{s+1} \\ \frac{1}{s+1} \end{bmatrix} v_-(s) = \underbrace{\begin{bmatrix} 1 \\ 0 \end{bmatrix} v_-(s) + \begin{bmatrix} \frac{-1}{s+1} \\ \frac{1}{s+1} \end{bmatrix} (v_-(s) - v_-(-1))}_{\in \mathbb{H}^{2\perp}} + \underbrace{\begin{bmatrix} \frac{-1}{s+1} \\ \frac{1}{s+1} \end{bmatrix} v_-(-1)}_{\mathbb{H}^2}.$$

The first two terms are in $\mathbb{H}^{2\perp}$ because the stable pole $s = -1$ cancels against the zero at $s = -1$ in $v_-(s) - v_-(-1)$. The future time solution is the inverse Laplace transform of the last (third) term

$$z(t) = \begin{bmatrix} -e^{-t} \\ e^{-t} \end{bmatrix} v_-(-1).$$

□

Effectively what this example demonstrates is that with inner transfer matrices the orthogonality criterion (that we know shows up in the projection theorem) reduced to that of orthogonality of $\mathbb{L}^2[0, \infty)$ and $\mathbb{L}^2(-\infty, 0]$, which is trivial. This simple example in fact solves essentially any classical SISO LQ-problem, see the following example and ensuing theorem.

Example 11.2.3 (Integrator). Consider the SISO integrator $y = \frac{1}{s}u$. To minimize $\int_0^\infty y^2(t) + u^2(t) dt$ given the past we rewrite the system as

$$z = \begin{bmatrix} u \\ y \end{bmatrix} = \begin{bmatrix} s \\ 1 \end{bmatrix} y.$$

The transfer matrix $G(s) = \begin{bmatrix} s \\ 1 \end{bmatrix}$ here is not inner, it satisfies

$$G^\sim(s)G(s) = -s^2 + 1.$$

This polynomial has zeros at $s = 1$ and $s = -1$ so can be spectrally factored as

$$G^\sim(s)G(s) = (-s+1)(s+1).$$

Aha, if we now divide $G(s)$ with the stable factor $s + 1$ then we do get, by construction, an inner transfer matrix,

$$z = \begin{bmatrix} u \\ y \end{bmatrix} = \begin{bmatrix} s/(s+1) \\ 1/(s+1) \end{bmatrix} v, \quad v = (s+1)y.$$

Here we recognize again the transfer matrix $G(s)$ of the previous example, so optimal is $v_+ = 0$. Done. One may object about “dividing by polynomials”, but it is all right which we will see by redoing the example in state space, done next. $\square$

Lemma 11.2.4 (LQ control for SISO systems, pole placement). Consider a SISO system $y = \mathcal{P}(u)$ with strictly proper minimal transfer function

$$P(s) = \frac{N(s)}{M(s)}$$

and minimal realization

$$\begin{bmatrix} A & B \\ C & 0 \end{bmatrix}.$$

The solution to the LQ-problem $\min_{u \in \mathbb{L}^2[0, \infty)} \int_0^\infty |u(t)|^2 + |y(t)|^2 dt$ can be implemented as a static state feedback $u = -Fx$ where $F \in \mathbb{R}^{1 \times n}$ is the unique matrix for which $A - BF$ has its eigenvalues at the stable zeros of

$$M(-s)M(s) + N(-s)N(s). \quad (11.2)$$

Proof. In state space the system with input u and output $\begin{bmatrix} u \\ y \end{bmatrix}$ become

$$\begin{bmatrix} u \\ y \end{bmatrix} = \left[\begin{array}{c|c} A & B \\ \hline 0 & 1 \\ C & 0 \end{array} \right] u$$

Let $u = -Fx + v$ where F is as specified above (such F exists by minimality of the realization, and is unique because the system is SISO). Then

$$\begin{bmatrix} u \\ y \end{bmatrix} = \left[\begin{array}{c|c} A - BF & B \\ \hline -F & 1 \\ C & 0 \end{array} \right] v.$$

Let $\chi(s) := \det(sI - (A - BF))$. Assume without loss of generality that $M(s)$ is monic (this does not change F). Clearly the transfer function from u to y has not been modified by the change of input $u = -Fx + v$, so the transfer function from v to $\begin{bmatrix} u \\ y \end{bmatrix}$ equals

$$\begin{bmatrix} u \\ y \end{bmatrix} = \underbrace{\begin{bmatrix} M(s)/\chi(s) \\ N(s)/\chi(s) \end{bmatrix}}_{G(s)} v.$$

This renders G inner,

$$G^\sim(s)G(s) = \frac{M(-s)M(s) + N(-s)N(s)}{\chi(-s)\chi(s)} = 1$$

hence rendering $v_+ = 0$ optimal. The condition $v_+ = 0$ is equivalent to $u = -Fx + v$ being $-Fx$ for positive time. $\blacksquare$

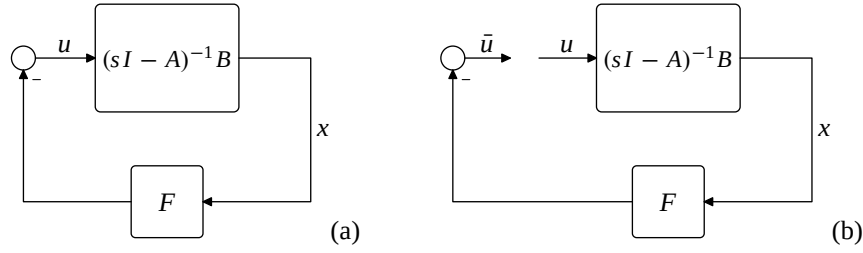


Figure 11.1: State feedback and loop gain $\bar{u} = -Lu$

Example 11.2.5 (Integrator). The integrator $y = \frac{1}{s}u$ has minimal state-space realization $\dot{x} = u$, $y = x$ and transfer function $\frac{N(s)}{M(s)} = \frac{1}{s}$. We have $M(-s)M(s) + N(-s)N(s) = -s^2 + 1 = (-s + 1)(s + 1)$ so we need to place the pole $A - BF$ at -1 . This requires $F = 1$ so $u_*(t) = -Fx(t) = -x(t) = -y(t)$ is the solution of the LQ-problem. $\square$

The core of the proof is that the transfer function of the mapping from v to $\begin{bmatrix} u \\ y \end{bmatrix}$,

$$\begin{bmatrix} G_u \\ G_y \end{bmatrix} := \left[\begin{array}{c|c} A - BF & B \\ \hline -F & 1 \\ C & 0 \end{array} \right]$$

is inner, that is (besides stability) that

$$G_u^* G_u + G_y^* G_y = 1 \quad \text{on the imaginary axis.}$$

An immediate consequence of this equality is that

$$G_u^* G_u = 1 - G_y^* G_y \leq 1 \quad \text{on the imaginary axis.}$$

or, equivalently,

$$|G_u^{-1}(i\omega)| \geq 1 \quad \forall \omega \in \mathbb{R}.$$

This is known as the *return difference inequality* and in terms of the realization says

$$|1 + F(i\omega I - A)^{-1}B| \geq 1 \quad \forall \omega \in \mathbb{R}.$$

The classic interpretation is that LQ is robust in the sense that the loop gain $L(i\omega) = F(sI - A)^{-1}B$ in the feedback loop of Fig. 11.1 stays outside the complex unit circle with center at the critical point -1 :

$$|1 + L(i\omega)| \geq 1 \quad \forall \omega \in \mathbb{R}.$$

11.3 MIMO systems – spectral factorization

Algorithmically, the bottle-neck in the LQ-problem is *spectral factorization*, that is, splitting the polynomial $M(-s)M(s) + N(-s)N(s)$ into a product $\chi(-s)\chi(s)$ with $\chi(s)$ Hurwitz. Nowadays the problem of computation of zeros of polynomials—i.e. computation of eigenvalues of matrices—is considered solved for modest sized problems. Thus the SISO LQ-problem is

essentially a solved problem. The beauty of LQ is that it works for MIMO systems as well and still the bottle-neck is a form of eigenvalue computation, so is essentially a solved problem as well. Theoretically the bottle-neck is to show that spectral factorization works (save some minor details) for MIMO systems. That is the topic of this section. It involves quite a bit of algebra. We will do spectral factorization not for polynomials but for proper rational functions because then the powerful machinery of state-space is at our disposal.

Application–9: Given a stable rational G find necessary and sufficient conditions for the existence of a bistable G_0 such that

$$G^\sim G = G_0^\sim G_0 \quad (11.3)$$

and determine the inner GG_0^{-1} .

A rational matrix G_0 is said to be *bistable* if it is invertible and both G_0 and G_0^{-1} are stable. A bistable solution G_0 of (11.3) known as a *spectral factor* of $G^\sim G$. Our quest begins with a realization of G

$$G = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right] \quad (11.4)$$

with A Hurwitz. Since both G and G_0 are stable it is to be expected that G_0 can be taken to have the same ‘ A -matrix’ as G . We propose a realization of the form

$$G_0 = \left[\begin{array}{c|c} A & B \\ \hline L & E \end{array} \right]$$

with L and E yet to be determined. It is readily verified that

$$G^\sim G = \left[\begin{array}{cc|c} A & 0 & B \\ -C'C & -A' & -C'D \\ \hline D'C & B' & D'D \end{array} \right] \quad (11.5)$$

and similarly

$$G_0^\sim G_0 = \left[\begin{array}{cc|c} A & 0 & B \\ -L'L & -A' & -L'E \\ \hline E'L & B' & E'E \end{array} \right].$$

Obviously equating $G^\sim G$ with $G_0^\sim G_0$ does not imply equating the realizations. There is, for one, the non-uniqueness due to state transformations. In order to preserve the structure of the realization we propose a state transformation

$$\begin{bmatrix} I_n & 0 \\ X & I_n \end{bmatrix}, \quad X \in \mathbb{R}^{n \times n}.$$

Under this transformation the realization of $G_0^\sim G_0$ transforms into

$$G_0^\sim G_0 = \left[\begin{array}{cc|c} \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} A & 0 \\ -L'L & -A' \end{bmatrix} \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} & \begin{bmatrix} I & 0 \\ X & I \end{bmatrix} \begin{bmatrix} B \\ -L'E \end{bmatrix} \\ \hline \begin{bmatrix} E'L & B' \end{bmatrix} \begin{bmatrix} I & 0 \\ -X & I \end{bmatrix} & E'E \end{array} \right] \quad (11.6)$$

$$= \left[\begin{array}{cc|c} A & 0 & B \\ A'X + XA - L'L & -A' & -L'E + XB \\ \hline E'L - B'X & B' & E'E \end{array} \right] \quad (11.7)$$

Whenever L, E, X are such that (11.7) equals (11.5) entry by entry, that is, if

$$E'E = D'D \quad (11.8)$$

$$L'E - XB = C'D \quad (11.9)$$

$$E'L - B'X = D'C \quad (11.10)$$

$$A'X + XA - L'L = -C'C \quad (11.11)$$

then $G \sim G = G_0 \sim G_0$. Remember that G_0 is supposed to be bistable. In particular this means that $G_0(\infty) = E$ is supposed to be invertible. Now let E be any square nonsingular solution of $E'E = D'D$ (such E exists iff D has full column rank). Then from (11.9) we can express L' as $L' = (C'D + XB)E^{-1}$, and then plug this expression into (11.11) to obtain

$$A'X + XA - (C'D + XB)E^{-1}E^{-'}(D'C + B'X) + C'C = 0. \quad (11.12)$$

Equation 11.10 follows from (11.9) if $X = X'$.

Lemma 11.3.1 (Sufficient conditions – Riccati). Consider G with realization (11.4) and assume $D'D > 0$. Let E be any nonsingular solution $E'E = D'D$. If $X = X' \in \mathbb{R}^{n \times n}$ solves the Algebraic Riccati Equation (ARE)

$$A'X + XA - (C'D + XB)(D'D)^{-1}(D'C + B'X) + C'C = 0 \quad (11.13)$$

then $G \sim G = G_0 \sim G_0$ for

$$G_0 = \left[\begin{array}{c|c} A & B \\ \hline L & E \end{array} \right] \quad \text{with } L := E^{-'}(D'C + B'X).$$

The matrix X is said to be the *stabilizing solution* of the Riccati equation if

$$A - BE^{-1}L := A - B(D'D)^{-1}(D'C + B'X)$$

is Hurwitz. In that case G_0^{-1} is stable. If in addition A is Hurwitz then G_0 is a spectral factor of $G \sim G$.

Proof. Follows essentially from the previous analysis. Note that $E^{-1}E^{-'}$ equals $(D'D)^{-1}$ so (11.13) is (11.12). The realization of G_0^{-1} is

$$G_0^{-1} = \left[\begin{array}{c|c} A - BE^{-1}L & * \\ \hline * & * \end{array} \right].$$

hence G_0^{-1} is stable if $A - BE^{-1}L$ is Hurwitz, and bistable if in addition A is Hurwitz. ■

Realize that we have not yet proved existence of any solution X of the ARE, let alone existence of a stabilizing solution. This we will do at the end of this section.

This is nice: G and G_0 have a joint realization

$$\begin{bmatrix} G \\ G_0 \end{bmatrix} = \left[\begin{array}{c|c} A & B \\ \hline C & D \\ L & E \end{array} \right].$$

Therefore

$$\begin{bmatrix} z \\ v \end{bmatrix} = \begin{bmatrix} G \\ G_0 \end{bmatrix} w \implies \begin{bmatrix} \dot{x} \\ z \\ v \end{bmatrix} = \begin{bmatrix} A & B \\ C & D \\ L & E \end{bmatrix} \begin{bmatrix} x \\ w \end{bmatrix}.$$

Now what is the transfer matrix from v to z ? On the one hand this can be computed using $z = Gw$ and $w = G_0 v$:

$$z = GG_0^{-1}v$$

and this transfer matrix is precisely the inner matrix GG_0^{-1} that we are after. On the other hand in terms of realizations we have

$$v = Lx + Ew \implies w = E^{-1}(v - Lx)$$

so that

$$z = \left[\begin{array}{c|c} A - BE^{-1}L & BE^{-1} \\ \hline C - DE^{-1}L & DE^{-1} \end{array} \right] v. \quad (11.14)$$

Therefore this must be a realization of the inner GG_0^{-1} . The ‘A-matrix’ here is the same as the ‘A-matrix’ of the G_0^{-1} . Surprising? Well at least it is good to see that the state dimension of GG_0^{-1} does not increase beyond that of G (which would normally happen if G and G_0 are not ‘related’). Now we can state and prove the famous theorem:

Lemma 11.3.2 (Spectral factorization). Let $G(s) = [A, B, C, D]$ be a realization of $G(s)$ with A Hurwitz and suppose the realization has no imaginary zeros, including at infinity:

$$\begin{bmatrix} A - i\omega I & B \\ C & D \end{bmatrix} \quad \text{and} \quad D$$

have full column rank for all $\omega \in \mathbb{R}$. Then the following holds.

1. The Hamiltonian matrix

$$H := \begin{bmatrix} A & 0 \\ -C'C & -A' \end{bmatrix} - \begin{bmatrix} B \\ -C'D \end{bmatrix} (D'D)^{-1} [D'C \quad B']$$

has no imaginary eigenvalues.

2. The Riccati equation

$$XA + A'X - (XB + C'D)(D'D)^{-1}(D'C + B'X) + C'C = 0$$

has a unique solution X for which $A - B(D'D)^{-1}(D'C + B'X)$ is Hurwitz. This X is referred to as the *stabilizing* solution and it is symmetric $X = X'$.

3. G_0 is bistable and such that $G \sim G = G_0 \sim G_0$ iff

$$G_0 = \left[\begin{array}{c|c} A & B \\ \hline L & E \end{array} \right], \quad L = E^{-'}(D'C + B'X), \quad E'E = D'D \quad (11.15)$$

with E square and X the stabilizing solution of the above ARE.

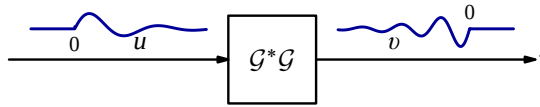
Proof (extensive sketch). Prior to this proof we already established that if X is a stabilizing solution of the ARE then any G_0 in the last item is a solution. Remains to show existence and uniqueness of stabilizing X and that (11.15) represents *all* spectral factors. (The following proof of existence of X is essentially the same as that $\Sigma_{22}(T)$ is nonsingular in the proof of Lemma 10.7.3.) Let us collect realizations for $G \sim G$ and its inverse,

$$G \sim G = \left[\begin{array}{c|c} A & 0 \\ \hline -C'C & -A' \end{array} \right] \left[\begin{array}{c} B \\ -C'D \end{array} \right], \quad (G \sim G)^{-1} = \left[\begin{array}{c|c} H & \left[\begin{array}{c} B \\ -C'D \end{array} \right] (D'D)^{-1} \\ \hline -(D'D)^{-1} [D'C \quad B'] & (D'D)^{-1} \end{array} \right].$$

(The A -matrix of $(G^\sim G)^{-1}$ is the Hamiltonian H defined in Item 1.) By stability of A the realizations have no unobservable or uncontrollable modes on the imaginary axis. By assumption G has no imaginary zeros, hence $(G^\sim G)^{-1}$ has no imaginary poles and thus H has no imaginary eigenvalues. Being the A -matrix of a self-conjugate $(G^\sim G)^{-1}$ the Hamiltonian matrix H exhibits some symmetry as well. Indeed $\hat{H}(s) := (sI - H) \begin{bmatrix} 0 & I \\ -I & 0 \end{bmatrix}$ is readily seen to be self-conjugate: $\hat{H}^\sim(s) = \hat{H}(s)$ so H has as many stable eigenvalues as anti-stable eigenvalues. Since H is $2n \times 2n$ and has no imaginary eigenvalues, the stable eigenspace of H is n -dimensional. Write the stable eigenspace as

$$\text{im} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix}, \quad X_1, X_2 \in \mathbb{C}^{n \times n}.$$

If X_1 is nonsingular it is easy to verify that $X := X_2 X_1^{-1}$ is a stabilizing solution to the ARE and that it is unique and symmetric (Thm. 11.8.2). So why is X_1 nonsingular? Suppose, to the contrary, that X_1 is singular. We will show that then state $x \in \mathbb{L}_{\mathbb{R}^{2n}}^2(\mathbb{R})$ exists and orthogonal external $\mathbb{L}^2$ -signals u and $v = \mathcal{G}^* \mathcal{G} u$ exist of the form



That is, with $u(t)$ zero for negative time and v zero for positive time. Such u, v can not exist because $\|\mathcal{G}u\|^2 = \langle G^\sim \mathcal{G}u, u \rangle = \langle v, u \rangle = 0$ which contradicts $\mathcal{G}u \neq 0$. If X_1 is singular there exist nonzero x_0 of the form $\begin{bmatrix} 0 \\ \hat{x}_0 \end{bmatrix}$ in the stable eigenspace of H . Given this “initial” state setting $u(t) = 0$ for negative time gives us $v := \mathcal{G}^* \mathcal{G} u$ for negative time:

$$t < 0, u(t) = 0 \implies v(t) = B' \hat{x}(t), \quad \hat{x}^{(1)} = -A' \hat{x}, \quad \hat{x}(0) = \hat{x}_0.$$

By antistability of $-A'$ this v is in $\mathbb{L}^2(-\infty, 0]$. For $t > 0$ we set $v(t) = 0$ and equate $u = (\mathcal{G}^* \mathcal{G})^{-1} v$:

$$t > 0, v(t) = 0 \implies u(t) = -(D' D)^{-1} [D' C \quad B'] x(t), \quad \dot{x} = H x, \quad x(0) = x_0.$$

Since $x(0)$ is in the stable eigenspace of H this u is in $\mathbb{L}^2[0, \infty)$. So both u and v are in $\mathbb{L}^2(\mathbb{R})$, and v (hence u) is not identical to zero because if it would have been zero then for $t < 0$ we would have had $\dot{x} = H x$, $x(0) = x_0$ which implies $\lim_{t \rightarrow -\infty} |x(t)| = \infty$ while we just showed that $x(t) \rightarrow 0$ as $t \rightarrow -\infty$.

If G_0 and G_1 are two bistable solutions then $G_0^\sim G_0 = G_1^\sim G_1$ so that $G_0 G_1^{-1} = G_0^{-\sim} G_0^\sim$. The left-hand side of this equation is bistable, the right-hand side is bi-antistable. Hence both sides are constant, i.e. $G_1 = W G_0$ for some constant W . The constant W is orthogonal, $W' W = I$. This constant W can be absorbed into E . In other words, letting E range over all solutions of $E' E = D' D$ gives us all solutions G_0 . ■

Dually one gets:

Lemma 11.3.3 (Spectral co-factorization). Let $G(s) = [A, B, C, D]$ be a realization of $G(s)$ with A Hurwitz and suppose the realization has full row rank on the imaginary axis, including at infinity:

$$\begin{bmatrix} A - i\omega I & B \\ C & D \end{bmatrix} \quad \text{and} \quad D$$

have full row rank for all $\omega \in \mathbb{R}$. Then the following holds.

1. There exist bistable G_0 such that $GG^\sim = G_0G_0^\sim$

2. The Riccati equation

$$AY + YA' - (YC' + B'D)(D'D)^{-1}(CY + DB') + B'B = 0$$

has a unique solution Y for which $A - (YC' + BD')(DD')^{-1}C$ is Hurwitz, called the stabilizing solution Y , and it satisfies $Y = Y'$.

3. G_0 is a spectral cofactor iff

$$G_0 = \left[\begin{array}{c|c} A & L \\ \hline C & E \end{array} \right], \quad L = (BD' + YC)E^{-'}, \quad EE' = DD'$$

with Y the unique stabilizing solution of the above ARE and E the non-unique square solution of $EE' = DD'$.

4. G_0 and G have a joint realization

$$\begin{bmatrix} G & G_0 \end{bmatrix} = \left[\begin{array}{c|c|c} A & B & L \\ \hline C & D & E \end{array} \right]$$

and $M := G_0^{-1}G$ is *co-inner* (meaning stable and $MM^\sim = I$) and has realization

$$\left[\begin{array}{c|c|c} A - LE^{-1}C & B - LE^{-1}D & 0 \\ \hline E^{-1}C & E^{-1}D & I \end{array} \right]$$

□

At the end of this chapter we comment upon the computation of Riccati equations.

Example 11.3.4 (Matlab – mu-tools).

```
H=[-1 -1; -1 1];      % some Hamiltonian
[X1,X2]=ric_schr(H);   % ric_schr is part of mu-tools
X=X2/X1;               % stabilizing solution
```

□

11.4 MIMO LQ and the return difference inequality

The solution for the LQ-problem (Application 8) for MIMO systems can now be stated.

Theorem 11.4.1. Consider a system $y = \mathcal{P}(u)$ with stabilizable realization

$$P(s) = \left[\begin{array}{c|c} A & B \\ \hline C & D \end{array} \right]$$

and assume (A, C) has no unobservable modes on the imaginary axis. Then the *Control Algebraic Riccati Equation (CARE)*

$$A'X + XA + C'C - (C'D + XB)(I + D'D)^{-1}(D'C + B'X) = 0 \quad (11.16)$$

has a unique stabilizing solution X and the optimal input u is a state feedback

$$u = -Fx, \quad F = (I + D'D)^{-1}(D'C + B'X).$$

Moreover the minimal $\int_0^\infty |y(t)|^2 + |u(t)|^2 dt$ equals $x(0)'Xx(0)$.

Proof. The realization of the transfer matrix from u to $\begin{bmatrix} u \\ y \end{bmatrix}$ is

$$\left[\begin{array}{c|c} A & B \\ \hline 0 & I \\ C & D \end{array} \right].$$

Since (A, C) has no unobservable modes the above realization has no imaginary zeros. Also the direct feedthrough term $\begin{bmatrix} I \\ D \end{bmatrix}$ has full column rank. Hence if A were stable Lemma 11.3.2 applies and says that the CARE has a stabilizing solution and that state feedback $u = -Fx + E^{-1}v$ with $F = (I + D'D)^{-1}(D'C + B'X)$ renders the system from v to $\begin{bmatrix} u \\ y \end{bmatrix}$ inner,

$$\begin{bmatrix} u \\ y \end{bmatrix} = \left[\begin{array}{c|c} A - BF & BE^{-1} \\ \hline -F & E^{-1} \\ C - DF & DE^{-1} \end{array} \right] v. \quad (11.17)$$

The matrix E does not concern us. Since the system is inner the optimal $v_+ = 0$ so that $u = -Fx$ for positive time.

If A is not stable then first apply a preliminary stabilizing state feedback $u = -F_{\text{stab}}x + \tilde{u}$. It can be verified (boring) that this preliminary state feedback cancels in the inner system (11.17). That $\|z\|^2 = x_0'Xx_0$ is easily verified by showing that $\frac{d}{dt}x'Xx = -z'(t)z(t)$ for the optimal u . This result appears not to come naturally in our “projection approach” and I do not know how to fix it. ■

Example 11.4.2 (Integrator). Consider once again the integrator $y = \frac{1}{s}u$. A state realization is

$$\dot{x} = u, \quad y = x.$$

The Riccati equation (11.16) becomes

$$1 - X^2 = 0.$$

The two solutions are $X = \pm 1$. Only $X = +1$ is stabilizing, giving us $u_*(t) = -B'Xx(t) = -x(t)$ as the unique optimal solution. □

As in the SISO case, also in the MIMO case the crux is that

$$\begin{bmatrix} G_u \\ G_y \end{bmatrix} := \begin{bmatrix} I \\ G \end{bmatrix} G_0^{-1} = \left[\begin{array}{c|c} A - BE^{-1}L & BE^{-1} \\ \hline -F & E^{-1} \\ C - DE^{-1}L & DE^{-1} \end{array} \right]$$

is inner. If the plant $P(s)$ is strictly proper then we have $D = 0$ and $E = I$. In this case the inner matrix is

$$\begin{bmatrix} G_u \\ G_y \end{bmatrix} := \left[\begin{array}{c|c} A - BL & B \\ \hline -F & I \\ C & 0 \end{array} \right].$$

Being inner we have that $G_u' G_u = I - G_y' G_y \leq I$ on the imaginary axis. This brings us once again at the return difference inequality, but now for MIMO systems:

Lemma 11.4.3 (Return difference inequality). For strictly proper plants we have

$$[I + L^\sim(i\omega)]'[I + L(i\omega)] \geq I \quad \forall \omega \in \mathbb{R}$$

where now L is the loop gain $L(s) = F(sI - A)^{-1}B$.

Proof. $G_u = (1 + L)^{-1}$. ■

11.5 SISO Filtering

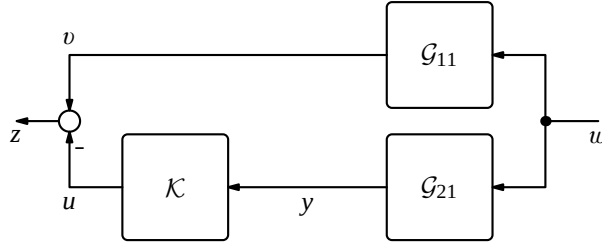


Figure 11.2: Filtering configuration

Filtering is the problem of learning about a signal v on the basis of related signal y . A systems approach to filtering is depicted in Fig. 11.2 where both v and y are seen as the output a system driven by a common source w . For the sake of argument, say that v is the noise a passenger hears sitting in an airplane and that y is the noise picked up by a microphone at some other point in the airplane. Clearly v and y are related meaning that G_{11} and G_{21} share certain dynamic features. Continuing with our little airplane-example, the aim of active noise canceling is to find a system K (called *filter*) such that $u := Ky$ ideally imitates v perfectly so that if its output u is transmitted “in antiphase” in the passengers headphone we ideally hear nothing $z = v - u = 0$ (or at least nothing in some desired frequency band). The game, in short, with filtering is to choose the filter K so that the error signal $z = v - u$ as small as possible. Small in what sense? An *extremely* popular approach — the reason is explained in Chapter 7 — is to minimize the $\mathbb{L}^2$ -norm of the impulse response of the error system,

$$\min_k \|g_{11} - k * g_{21}\|_{\mathbb{L}^2(\mathbb{R})}.$$

This assumes all systems G_{11}, G_{21}, K to be LTI, with respective impulse responses g_{11}, g_{21}, k . Because of Plancherel we can for $\mathbb{L}^2$ -stable systems also express this in frequency domain as

$$\min_K \|G_{11} - KG_{21}\|_{\mathbb{L}^2(\mathbb{i}\mathbb{R})}.$$

We pose and solve this problem for various assumptions imposed on the filter. We mostly work with transfer function representations G_{11}, G_{21}, K of the systems.

Application–10: Consider rational SISO systems $G_{11}, G_{21} \in \mathbb{L}^\infty$. Find $K \in \mathbb{L}^\infty$ that minimizes $\|G_{11} - KG_{21}\|_{\mathbb{L}^2(\mathbb{i}\mathbb{R})}$.

This particular application is so easy that projections are not needed. Simply taking

$$K_* = \frac{G_{11}}{G_{21}}$$

will do the job and render the error system identical to zero,

$$G_{11} - K_* G_{21} = 0,$$

or, at least, in many cases:

Example 11.5.1 (Four examples).

- If $G_{11}, G_{21} \in \mathbb{H}^\infty$ with G_{21} bistable, say,

$$G_{11}(s) = 1/(s + 1) \quad \text{and} \quad G_{21} = (s + 2)/(s + 3)$$

we have the happy situation that $K_* = \frac{G_{11}}{G_{21}} = \frac{s+3}{(s+1)(s+2)}$ is stable and causal. Nothing scary here.

- If $G_{11} = 1/(s + 1)$ and $G_{21} = (s - 2)/(s + 3)$ — both in $\mathbb{H}^\infty$ — we have that $K_* = \frac{s+3}{(s+1)(s-2)}$. This K_* has a pole at $s = 2$ and is thus not in $\mathbb{H}^\infty$ i.e. it is not both stable and causal. Pity. We insist that K is stable (i.e. in $\mathbb{L}^\infty$) but then our K_* is not causal. Mathematically there is nothing wrong with this choice. Being stable we have to assume that the region of convergence contains the imaginary axis. From partial fraction expansion

$$K_*(s) = \frac{s+3}{(s+1)(s-2)} = \frac{5/3}{s-2} - \frac{2/3}{s+1} \quad -1 < \operatorname{Re}(s) < 2$$

we can derive the impulse response

$$k_*(t) = \frac{5}{3}e^{2t}\mathbb{1}(-t) - \frac{2}{3}e^{-t}\mathbb{1}(t) = \text{graph of } k_*(t) \text{ on } \mathbb{R}.$$

Indeed it is not causal, but it is $\mathbb{L}^2(\mathbb{R})$ -stable.

- If $G_{11} = 1$ and $G_{21} = 1/(s+3)$ — both in $\mathbb{H}^\infty$ — we have the problem that $K_* = s+3$ which is non-proper implying that K is not stable (i.e. not in $\mathbb{L}^\infty$, it has differentiating action). This type of problems can occur since we divide by *strictly* proper G_{21} .

In this example in fact no $K \in \mathbb{L}^\infty$ exists that renders $\|G_{11} - KG_{21}\|$ finite. This is due to the fact that G_{21} is strictly proper and G_{11} is not strictly proper.

- If G_{21} has imaginary zeros then essentially the same problem occurs. For instance if $G_{11} = 1/(s+1)$ and $G_{21} = \frac{s^2+1}{(s+1)^2}$ we have $K_* = \frac{s+1}{s^2+1}$ which has poles at $\pm i$. Causal or not, it is not a stable system. It can be approximated arbitrarily well by a stable system though, in fact stable and causal,

$$K_\varepsilon(s) = \frac{s+1}{s^2 + \varepsilon s + 1}, \quad \varepsilon > 0$$

rendering the error system $G_{11} - K_\varepsilon G_{21}$ arbitrarily close to zero (in norm) for $\varepsilon > 0$ but small.

The last two examples are more-or-less the reason that in applications we prefer to rule out the situation that G_{21} has imaginary zeros or zeros at infinity. After all, what is an optimization problem whose “optimal” solution K is (close to) unstable? For other problems we can easily summarize:

Lemma 11.5.2. If $K_* := G_{11}/G_{21}$ is in $\mathbb{L}^\infty$ then it is a solution of Application 10. If this K has its poles in the open LHP it is also causal. In all other cases it is not causal. $\square$

☐

Next bring in causality constraints.

Application–11: Consider rational, SISO systems $G_{11} \in \mathbb{H}^2$ and $G_{21} \in \mathbb{H}^\infty$. Minimize $\|G_{11} - K G_{21}\|_{\mathbb{L}^2}$ with respect to $K \in \mathbb{H}^\infty$.

Its solution is an example of projection, be it with one little subtlety. The set of filters $\mathbb{H}^\infty$ is a vector space so minimizing the $\mathbb{L}^2$ -norm of $G_{11} - K G_{21}$ is that of finding the minimal norm solution an affine variety, and it looks like that Theorem 5.15.1 applies. Not so, because $\mathbb{H}^\infty$ is not an inner product space. The common way out of this paradox is to assume that

G_{21} has no imaginary zeros and no zeros at ∞ .

Then $\|G_{11} - K G_{21}\|_{\mathbb{L}^2} < \infty$ implies $K \in L^2(i\mathbb{R})$. So we may without loss of generality restrict our search over the smaller set

$$K \in \mathbb{H}^\infty \cap \mathbb{L}^2 \subseteq \mathbb{H}^2.$$

Now we can proceed with the $\mathbb{L}^2$ -inner product:

$$\begin{aligned} K_* \in \mathbb{H}^\infty \cap \mathbb{L}^2 \text{ is optimal} &\iff G_{11} - K_* G_{21} \perp \mathbb{H}^2 G_{21} \\ &\iff \langle G_{11} - K_* G_{21}, K G_{21} \rangle = 0, \quad \forall K \in \mathbb{H}^2 \\ &\iff \langle G_{11} G_{21}^\sim - K_* G_{21} G_{21}^\sim, \mathbb{H}^2 \rangle = 0 \\ &\iff G_{11} G_{21}^\sim - K_* G_{21} G_{21}^\sim \in \mathbb{H}^{2\perp}. \end{aligned}$$

If we can solve the last inclusion for some $K_* \in \mathbb{H}^\infty \cap \mathbb{L}^2$ then K_* is optimal.

Example 11.5.3. Suppose

$$G_{11}(s) = \frac{1}{s+3}, \quad G_{21}(s) = \frac{s-1}{s+1}.$$

(We are lucky because G_{21} is inner.) The inclusion to solve for is

$$\frac{-s-1}{(s+3)(-s+1)} - K_* \in \mathbb{H}^{2\perp}.$$

Partial fraction expansion (PFE) of the first term gives the equivalent inclusion

$$\frac{1/2}{s+3} + \frac{1/2}{s-1} - K_* \in \mathbb{H}^{2\perp}.$$

The PFE splits the first term into a stable part (in $\mathbb{H}^2$ in fact) and an anti-stable part (in $\mathbb{H}^{2\perp}$ in fact). Simply setting K_* equal to the stable term solves the problem

$$K_*(s) = \frac{1/2}{s+3}$$

rendering the the error system $G_{11} - K_* G_{21}$ equal to the unstable part divided by $G_{21}^\sim$,

$$G_{11}(s) - K_*(s) G_{21}(s) = \frac{1/2}{s-1} \times \frac{s-1}{s+1} = \frac{1/2}{s+1}.$$

Our K_* is in $\mathbb{H}^\infty$ as required, and it is in $\mathbb{L}^2$ as guaranteed. □

Lemma 11.5.4 (SISO Inner-outer cofactorization). Consider rational, SISO systems $G_{11} \in \mathbb{H}^2$ and $G_{21} \in \mathbb{H}^\infty$ and assume G_{21} has no zeros on $i\mathbb{R} \cup \{\infty\}$. Then G_{21} has a (co)inner-outer factorization

$$G_{21} = G_0 G_i, \quad G_i, G_0, G_0^{-1} \in \mathbb{H}^\infty \text{ and } G_i G_i^\sim = 1$$

and the solution K_* of Application 11 exists and is unique,

$$K_* = \pi_+(G_{11} G_i^\sim) G_0^{-1}.$$

□

Here π_+ is the orthogonal projection operator from $\mathbb{L}^2$ to $\mathbb{H}^2$, or more concretely $\pi_+(G)$ is the sum of strictly proper stable terms in a PFE of G .

11.6 MIMO filtering

Now we have enough to solve the (almost) most general Filtering problem:

Application–12: Consider rational, $G_{11} \in \mathbb{H}^2$ and $G_{21} \in \mathbb{H}^\infty$ (both possibly MIMO). Minimize $\|G_{11} - K G_{21}\|_{\mathbb{L}^2}$ over $K \in \mathbb{H}^\infty$.

Theorem 11.6.1 (Filtering). Consider rational systems G_{11} and G_{21} with G_{21} strictly proper and having joint realization

$$\begin{bmatrix} G_{11} \\ G_{21} \end{bmatrix} = \left[\begin{array}{c|c} A & B \\ \hline C_z & 0 \\ C & D \end{array} \right]$$

with (A, C) detectable and assume the realization of G_{21} has full row rank on the imaginary axis, including at infinity:

$$\begin{bmatrix} A - i\omega I & B \\ C & D \end{bmatrix}, \quad D.$$

Then the *Filtering ARE (FARE)*

$$AY + YA' - (YC' + B'D)(D'D)^{-1}(CY + DB') + B'B = 0 \quad (11.18)$$

has a unique stabilizing solution Y and the optimal filter K_* solving Application 12 exists and has realization

$$K_* = \left[\begin{array}{c|c} A - LC & B - LD \\ \hline C_z & 0 \end{array} \right], \quad L = (B'D + YC)E^{-'}, \quad EE' = DD'$$

and the optimal error systems is

$$G_{11} - K_* G_{21} = \left[\begin{array}{c|c} A - LC_y & B - LD_y \\ \hline C_z & 0 \end{array} \right]$$

Proof. Later. ■

This theorem actually does not require G_{11} and G_{21} to be stable. All it does is make $G_{11} - K_* G_{21} \in \mathbb{H} \cap \mathbb{H}^2$ and minimize its norm over $K_* \in \mathbb{H}^\infty \cap \mathbb{H}^2$.

11.7 Extra: Standard $\mathbb{H}^2$ problem

A combination of LQ and Filtering. No time to talk about it in detail only brief summary:

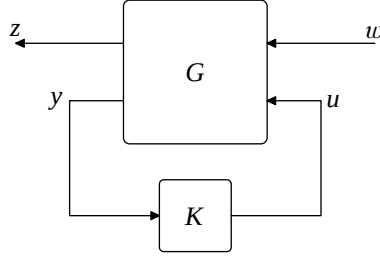


Figure 11.3: The standard $\mathbb{H}^2$ problem

Application–13: The *standard* $\mathbb{H}^2$ optimization problem is the problem of choosing the controller K in the block diagram of Fig. 11.3 such that it

1. stabilizes the closed-loop system, and
2. minimizes the $\mathbb{H}^2$ -norm of the closed-loop system (with w as input and z as output).

We represent the generalized plant G of Fig. 11.3 in state space form as

$$\dot{x}(t) = Ax(t) + B_1 w(t) + B_2 u(t), \quad (11.19)$$

$$z(t) = C_1 x(t) + D_{11} w(t) + D_{12} u(t), \quad (11.20)$$

$$y(t) = C_2 x(t) + D_{21} w(t) + D_{22} u(t). \quad (11.21)$$

The $\mathbb{H}^2$ problem may be solved by seeing at as nested LQ and Filtering problem.

Lemma 11.7.1 (Solution of the $\mathbb{H}^2$ problem). Consider the standard $\mathbb{H}^2$ optimization problem for the generalized plant

$$\dot{x}(t) = Ax(t) + B_1 w(t) + B_2 u(t), \quad (11.22)$$

$$z(t) = C_1 x(t) + D_{12} u(t), \quad (11.23)$$

$$y(t) = C_2 x(t) + D_{21} w(t) + D_{22} u(t). \quad (11.24)$$

Assumptions:

- The system $\dot{x}(t) = Ax(t) + B_2 u(t)$, $y(t) = C_2 x(t)$ is stabilizable and detectable.
- The matrix $\begin{bmatrix} A-sI & B_1 \\ C_2 & D_{21} \end{bmatrix}$ has full row rank for every $s = i\omega$, and D_{21} has full row rank.
- The matrix $\begin{bmatrix} A-sI & B_2 \\ C_1 & D_{12} \end{bmatrix}$ has full column rank for every $s = i\omega$, and D_{12} has full column rank.

Under these assumptions the optimal output feedback controller $u = Ky$ is

$$\dot{\hat{x}}(t) = A\hat{x}(t) + B_2u(t) + L[y(t) - C_2\hat{x}(t) - D_{22}u(t)] \quad (11.25)$$

$$u(t) = -F\hat{x}(t). \quad (11.26)$$

The observer and state feedback gain matrices are

$$F = (D'_{12}D_{12})^{-1}(B'_2X + D'_{12}C_1), \quad L = (YC'_2 + B_1D'_{21})(D_{21}D'_{21})^{-1}. \quad (11.27)$$

The symmetric matrices X and Y are the unique positive-definite solutions of the algebraic Riccati equations

$$\begin{aligned} A'X + XA + C'_1C_1 - (XB_2 + C'_1D_{12})(D'_{12}D_{12})^{-1}(B'_2X + D'_{12}C_1) &= 0, \\ AY + AY' + B_1B'_1 - (YC'_2 + B_1D'_{21})(D_{21}D'_{21})^{-1}(C_2Y + D_{21}B'_1) &= 0. \end{aligned} \quad (11.28)$$

□

The condition that D_{12} has full column rank means that there is “direct feedthrough” from the input u to the error signal z . Dually, the condition that D_{21} has full row rank means that the some noise w is directly fed through to the observed output y .

11.8 Extra: Riccati equation

An *algebraic Riccati equation* (ARE for short) is a quadratic equation of the form

$$PA + A'P + PRP + Q = 0. \quad (11.29)$$

Here the matrices $A, Q, R \in \mathbb{R}^{n \times n}$ are assumed given, with R and Q symmetric, and the equation has to be solved for $P \in \mathbb{R}^{n \times n}$. Riccati equations, being quadratic, usually have more than one solution P . There is however at most one *stabilizing* solution.

Definition 11.8.1 (stabilizing solution of an ARE). Let $A, R, Q \in \mathbb{R}^{n \times n}$ with $R = R', Q = Q'$. A matrix $P \in \mathbb{R}^{n \times n}$ is said to be a *stabilizing* solution of (11.29) if $A + RP$ is Hurwitz. □

Theorem 11.8.2 (Existence of the stabilizing solution of ARE). If the Riccati equation (11.29) has a stabilizing solution, then the $2n \times 2n$ Hamiltonian matrix

$$\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix}$$

has no imaginary eigenvalues, and there is a full rank matrix $V \in \mathbb{R}^{2n \times n}$ and a Hurwitz matrix $\Lambda \in \mathbb{R}^{n \times n}$ such that

$$\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix} V = V\Lambda.$$

Further, given any such V decompose V as $V = \begin{bmatrix} V_1 \\ V_2 \end{bmatrix}$ with $V_1, V_2 \in \mathbb{R}^{n \times n}$. Then a stabilizing solution P of (11.29) exists iff V_1 is nonsingular. In this case P is determined uniquely as $P = V_2V_1^{-1}$ and it is symmetric. Moreover the eigenvalues of $A + RP$ coincide with those of Λ . □

Proof. If P is a stabilizing solution of the ARE, then

$$\begin{bmatrix} I & 0 \\ -P & I \end{bmatrix} \begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix} \begin{bmatrix} I & 0 \\ P & I \end{bmatrix} = \begin{bmatrix} A + RP & R \\ 0 & -(A + RP)' \end{bmatrix}. \quad (11.30)$$

Interestingly $\begin{bmatrix} I & 0 \\ -P & I \end{bmatrix} \begin{bmatrix} I & 0 \\ P & I \end{bmatrix} = I_{2n}$ so the above equation is a similarity transformation and this tells us that the eigenvalues of the Hamiltonian $\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix}$ are those of $A + RP$ and $-(A + RP)'$. If P is stabilizing then none of these eigenvalues lie on the imaginary axis. Furthermore it shows that there are n stable eigenvalues (those of $A + RP$) and n antistable eigenvalues (those of $-(A + RP)'$). In addition it shows that $\text{im} \begin{bmatrix} I \\ P \end{bmatrix}$ is the stable eigenspace of the Hamiltonian and hence that stabilizing P is unique.

Now for simplicity assume that all eigenvalues are distinct, and let $\lambda_1, \dots, \lambda_n$ denote the n stable eigenvalues. To each eigenvalue λ_i there corresponds an eigenvector v_i . That is, $\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix} v_i = \lambda_i v_i$, or in matrix form

$$\begin{bmatrix} A & R \\ -Q & -A' \end{bmatrix} \underbrace{\begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}}_V = \underbrace{\begin{bmatrix} v_1 & v_2 & \cdots & v_n \end{bmatrix}}_V \underbrace{\begin{bmatrix} \lambda_1 & 0 & 0 & 0 \\ 0 & \lambda_2 & 0 & 0 \\ 0 & 0 & \ddots & 0 \\ 0 & 0 & 0 & \lambda_n \end{bmatrix}}_\Lambda.$$

As (11.30) is a similarity transformation, the eigenvectors v_i may also be written as $v_i = \begin{bmatrix} I & 0 \\ P & I \end{bmatrix} w_i$ where w_i is an eigenvector of $\begin{bmatrix} A+RP & R \\ 0 & -(A+RP)' \end{bmatrix}$ with eigenvalue λ_i . Now all λ_i are stable, and they are the eigenvalues of $A + RP$ and *not* of $-(A + RP)'$, so w_i must be of the form $w_i = \begin{bmatrix} * \\ 0 \end{bmatrix}$. Hence

$$V = \begin{bmatrix} I & 0 \\ P & I \end{bmatrix} \begin{bmatrix} W_1 \\ 0 \end{bmatrix}, \quad \text{where} \quad \begin{bmatrix} W_1 \\ 0 \end{bmatrix} = \begin{bmatrix} w_1 & w_2 & \cdots & w_n \end{bmatrix}.$$

Written out this becomes $V = \begin{bmatrix} W_1 \\ P W_1 \end{bmatrix}$. Now W_1 is nonsingular because the matrix of eigenvectors $\begin{bmatrix} W_1 \\ 0 \end{bmatrix}$ has full column rank. Therefore $V_1 = W_1$ is nonsingular and P follows uniquely as $V_2 V_1^{-1} = (P W_1) W_1^{-1} = P$. Conversely if V_1 is nonsingular then it is easy to verify that $P := V_2 V_1^{-1}$ satisfies the ARE and that is stabilizing with $A + RP = V_1 \Lambda V_1^{-1}$.

To verify symmetry of P , define $T = P' - P$. We have

$$T(A + RP) = (P' - P)(A + RP) = \begin{bmatrix} I & P' \end{bmatrix} \begin{bmatrix} 0 & -I \\ I & 0 \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix} (A + RP) = \begin{bmatrix} I & P' \end{bmatrix} \begin{bmatrix} Q & A' \\ A & R \end{bmatrix} \begin{bmatrix} I \\ P \end{bmatrix}$$

The last term is symmetric hence so is $T(A + RP)$ which combined with skew symmetry $T' = -T$ of T gives $T(A + RP) + (A + RP)'T = 0$. This is a Lyapunov equation and its solution is unique because $A + RP$ is Hurwitz (see the proof Theorem 10.4.2.) That is, $T = 0$, i.e., $P = P'$. ■

This theorem is the vantage point for numerical algorithms for finding P . Indeed, the above proof of the theorem is practically an algorithm. Solving AREs is essentially an eigenvalue problem.

11.9 IYWTKM..

- Another approach to infinite horizon LQ is to consider $u, z \in \mathbb{L}^2[0, \infty)$ and then to solve the normal equations $\mathcal{G}^*(Ce^{At}x_0 + \mathcal{G}(u_*)) = 0$. Here we need the adjoint with respect to $\mathbb{L}^2[0, \infty)$ and this is not so easy to work with.
- In LQ we optimize over *signals* whereas in Filtering we optimize over *systems*.
- Many formulas about these problems can be found in Zhou et al. (1996).

11.10 Problems

11.1 Consider the double integrator $y = \frac{1}{s^2}u, \dots$

11.2 Prove ... gain-phase margin

11.3 integral control.

11.4 Extend Lemma 11.2.4 to the case that $P(s)$ is proper but not necessarily strictly proper.

11.5 Solve the LQ problem for the first-order system

$$\dot{x} = ax + u, \quad y = x.$$

Comment on the dependence of u_* and minimal norm on a .

12 Wiener filtering and the matched filter

At the end of this chapter we review the matched filter, but the main topic of this chapter is discrete filtering in the ℓ^2 -norm from a stochastic perspective, a problem known as *Wiener filtering*. The emphasis is more on signals and less on systems and the signals are assumed to be stochastic processes. A motivating noise-canceling example was briefly mention in Section 11.5 and another example more in the Wiener spirit is:

Example 12.0.1 (Constant desired signal). If a desired signal v is known to be constant and η is known to be zero mean white noise then averaging $y := v + \eta$ should recover v well, see Fig. 12.1. In the class of causal FIR systems $\hat{v} = \mathcal{K}(y)$ one could for instance use a moving average $\hat{v}[n] := \frac{1}{M} \sum_{i=0}^{M-1} y[n-i]$ with M large. $\square$

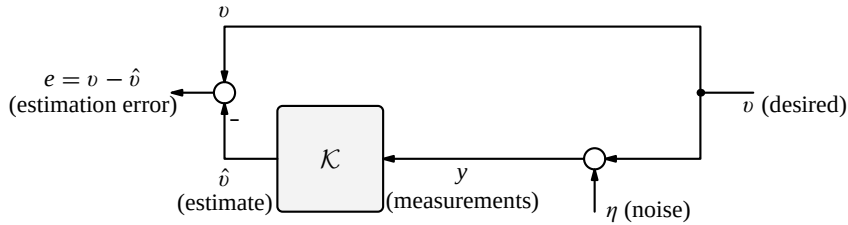


Figure 12.1: Recover desired v from noise corrupted $y = v + \eta$

The message of this little example is: exploit known or estimated structure (memory) in v , η (and hence $y = v + \eta$) to find $\hat{v} := \mathcal{K}(y)$ so as to make $e := v - \hat{v}$ small in some sense. In *Wiener filtering* the assumptions are that

1. Signal v needs to be estimated;
2. Signal v and (additive) noise η are WSS stochastic processes with known spectral characteristics or known covariances and cross-covariances.
3. The performance criterium is the minimum mean-square estimation error: $E(v[n] - \hat{v}[n])^2$ which by assumed wide-sense stationarity does not depend on n .

In this chapter it will be convenient to use the following two short hands for covariance matrices and vectors of given WSS processes y and v :

$$R_y^{p \times p} := \begin{bmatrix} r_y[0] & r_y[1] & \cdots & r_y[p-1] \\ r_y[1] & r_y[0] & \ddots & r_y[p-2] \\ \vdots & \ddots & \ddots & \vdots \\ r_y[p-1] & r_y[p-2] & \cdots & r_y[0] \end{bmatrix}, \quad R_{yv}^{p \times 1} = \begin{bmatrix} r_{yv}[0] \\ r_{yv}[1] \\ \vdots \\ r_{yv}[p-1] \end{bmatrix}$$

where, as always, $r_y[\tau] = \text{cov}(y[t + \tau], y[t])$ and $r_{yv}[\tau] = \text{cov}(y[t + \tau], v[t])$.

12.1 Constant filter

Application–14: Given zero mean WSS processes v and y find linear static filter $u = \mathcal{K}(y)$ that minimizes $E(v[n] - u[n])^2$. One may assume to know all possible second order moments of the processes.

Lemma 12.1.1 (Constant filter). The solution to Application 14 is

$$u[n] = k_* y[n], \quad k_* = \frac{E(v[n]y[n])}{E(y^2[n])}.$$

The k_* by WSS is independent of $n \in \mathbb{Z}$. The estimation error equals

$$E(v[n] - u[n])^2 = E(v^2[n] - u^2[n]) = E(v^2[n]) - k'_* E(y[n]v[n]) = E(v^2[n]) - \frac{(E(v[n]y[n]))^2}{E(y^2[n])}.$$

Proof. Express the filter as $u[n] = k y[n]$ with $k \in \mathbb{R}$ and let $y = (\dots, y[-1], y[0], y[1], \dots)$ considered as a infinite column vector.

$$\begin{aligned} k_* \text{ is optimal} &\iff (v - k_* y) \perp k_0 y \quad \forall k_0 \in \mathbb{R} \\ &\iff E y'(v - k_* y) = 0 \\ &\iff \sum_{n \in \mathbb{Z}} E y[n](v[n] - k_* y[n]) = 0. \end{aligned}$$

The expression $E y'(v - k_* y)$ is potentially scary because $E y' y = \infty$. However for $k_* = E(y[n]v[n]) / E(y^2[n])$ the terms $E y[n](v[n] - k_* y[n])$ are zero for every n . By orthogonality of v and u we have $E(v[n] - u[n])^2 = E(v^2[n] - u^2[n])$ which can be expressed in many different ways. ■

The lemma says that constant filtering achieves *something*—that is, achieves $E(v - u)^2 < E(v^2)$ —iff $v[n]$ and $y[n]$ are correlated.

12.2 FIR filter — matrix form of Wiener-Hopf equations

Application–15: Given zero mean WSS processes v and y find a FIR filter $u = \mathcal{K}(y)$ of given order p ,

$$u[n] = \underbrace{[k_0 \quad \dots \quad k_{p-1}]}_{k'} \begin{bmatrix} y[n] \\ \vdots \\ y[n - p + 1] \end{bmatrix}, \quad n \in \mathbb{Z}$$

that minimizes $E(v[n] - u[n])^2$. One may assume to know all possible second order moments of the processes.

Again this problem is well defined and $E(v[n] - u[n])^2$ is independent of n by the fact that all signals are WSS. The normal equations in this respect is known as the matrix form of the *Wiener-Hopf* equations (there is another form as well).

Lemma 12.2.1 (FIR filter). A $k_* \in \mathbb{R}^p$ is a solution to Application 15 iff

$$R_y^{p \times p} k_* = R_{yv}^{p \times 1}$$

or, more explicitly (with $r_y[m] = \mathbb{E} y[n+m]y[n]$ and $r_{yv}[m] = \mathbb{E} y[n+m]v[n]$)

$$\begin{bmatrix} r_y[0] & r_y[1] & \cdots & r_y[p-1] \\ r_y[1] & r_y[0] & \ddots & r_y[p-2] \\ \vdots & \ddots & \ddots & \vdots \\ r_y[p-1] & r_y[p-1] & \cdots & r_y[0] \end{bmatrix} \underbrace{\begin{bmatrix} k_0 \\ k_1 \\ \vdots \\ k_{p-1} \end{bmatrix}}_{k_*} = \begin{bmatrix} r_{yv}[0] \\ r_{yv}[1] \\ \vdots \\ r_{yv}[p-1] \end{bmatrix} \quad (12.1)$$

and the MMSE equals

$$\mathbb{E}(v[n] - u[n])^2 = \mathbb{E}(v^2[n] - u^2[n]) = \mathbb{E}(v^2[n]) - k_*' R_y^{p \times p} k_* = \mathbb{E}(v^2[n]) - [R_{yv}^{p \times 1}]' [R_y^{p \times p}]^{-1} R_{yv}^{p \times 1}$$

Proof. See Lemma 7.8.1. ■

12.3 Classic Wiener Filtering

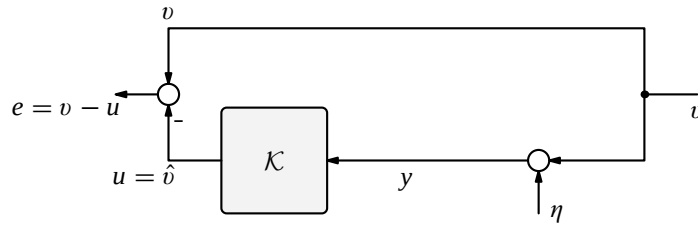


Figure 12.2: Recover desired v from noise corrupted $y = v + \eta$

A special case of Application 15 and probably the best known Wiener filtering problem is what we considered at the beginning of this chapter,

$$y = v + \eta \quad \text{for some zero-mean “noise” } \eta \text{ uncorrelated with } v.$$

That is, the available measurements y is the desired v additively corrupted by zero-mean η that is uncorrelated with v , see Fig. 12.2. The Wiener-Hopf equations

$$R_y^{p \times p} k_* = R_{yv}^{p \times 1}$$

for this case specialize to

$$(R_v^{p \times p} + R_\eta^{p \times p}) k_* = R_{yv}^{p \times 1} = R_{vv}^{p \times 1}. \quad (12.2)$$

The covariance matrix $R_y^{p \times p}$ can be estimated since y is assumed available (see Section 12.7). The signal v obviously is not measured — otherwise there would be no point in trying to estimate it — but the solution (12.2) does assume we have some knowledge (namely $R_{vv}^{p \times 1}$) of v . A practical situation where we do have access to (estimates of) $R_{vv}^{p \times 1}$ is noise cancellation covered in the next section.

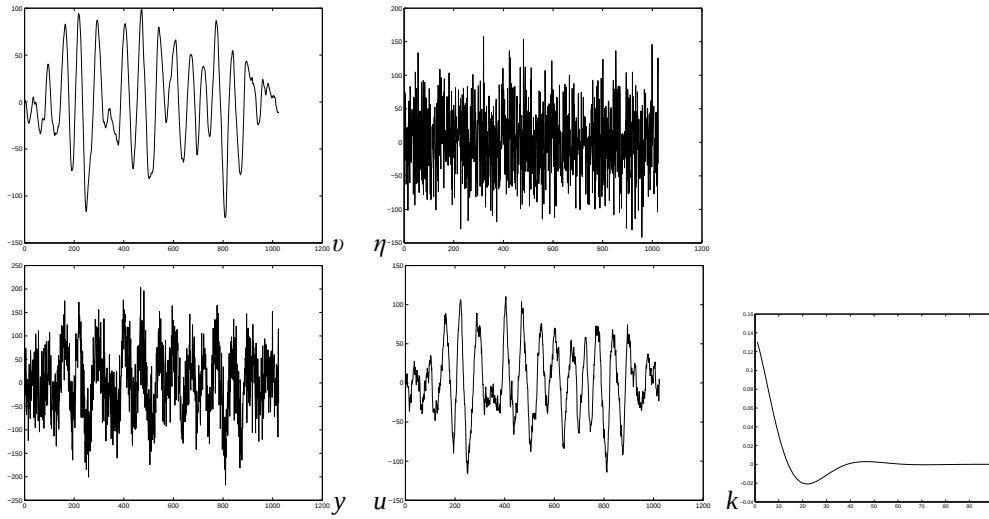


Figure 12.3: Realizations of v , η , y , u where $u = \mathcal{K}(y)$ a FIR filter of order $p = 100$. Bottom right: the optimal FIR

Example 12.3.1. Figure 12.3 shows realizations of WSS processes v and η . The fluctuations in η are much stronger than that of v so we expect that the filter $\hat{v} := u := \mathcal{K}(y)$ based on $y = v + \eta$ (shown as well in Fig. 12.3) is some form of moving averaging. Using known covariances (not spelled out here) of v and η and assumed orthogonality of v and η the optimal FIR system of order $p = 100$ determined from (12.2) results in a impulse response shown in the bottom-right part of Fig. 12.3 and the estimate $\hat{v} := u$ is (visually at least) close to v . $\square$

12.4 Noise cancellation

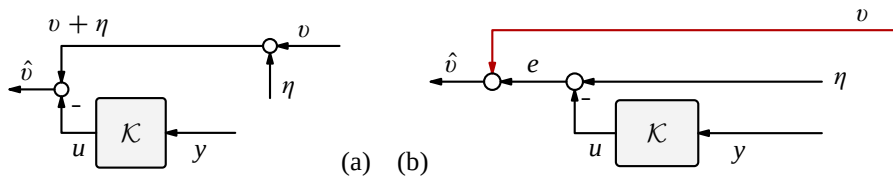


Figure 12.4: Try to remove noise η from $v + \eta$ using related $u = \mathcal{K}(y)$

Application–16: See Fig. 12.4(a). Let v , η , y be zero mean WSS processes and assume that (desired) v is uncorrelated (orthogonal) to (noise) η and (measurements) y . Recover v from $v + \eta$ by cancelling η optimally through minimization with respect FIR filters of order p ,

$$\min_{u=\mathcal{K}(y)} E(\eta[n] - u[n])^2.$$

One may assume to know all possible second order moments of processes $v + \eta$ and y .

The idea here is that we measure our desired v in noise, $v + \eta$ but that the noise is related to a secondary signal y that we measure as well. Using the relation between η and y we can hope to remove the noise η to some extent from our measurements $v + \eta$ rendering the final output $\hat{v}$ close to our desired v .

This problem is a standard Wiener filtering problem in disguise: reorder the configuration of part (a) of Fig. 12.4 to that of part (b) so that

$$\|\hat{v} - v\|^2 = \|e\|^2, \quad e = \eta - u.$$

So we once again apply can Lemma 12.2.1 which says that the problem has an explicit FIR solution once we know the covariances

$$R_y^{p \times p} \quad \text{and} \quad R_{y\eta}^{p \times 1}.$$

Here there is a catch in that we do not measure (y, η) so we can not readily estimate the covariances that we need. We measure only $(y, v + \eta)$. However by the assumption that v as a process is uncorrelated (orthogonal) to η and y we have

$$R_{y\eta}^{p \times 1} = R_{y(v+\eta)}^{p \times 1}.$$

Hence:

Lemma 12.4.1 (FIR filter for noise cancellation). A $k_* \in \mathbb{R}^p$ is a solution to Application 16 iff

$$R_y^{p \times p} k_* = R_{y(v+\eta)}^{p \times 1}.$$

□

12.5 Linear prediction

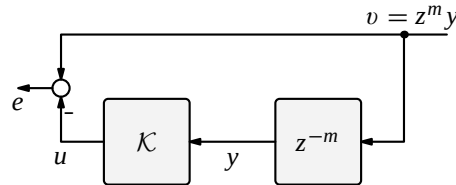


Figure 12.5: m -step ahead predictor for y

An important and interesting application is *prediction* or *forecasting*. The problem is to predict the future behavior of the phenomenon that is observed as accurately as possible given the observations up to the present time. See Fig. 12.5 (here we used z^{-m} to denote the m -step delay operator).

Application–17: Given zero mean WSS process y find a causal FIR m -step ahead predictor $u = \mathcal{K}(y)$ of given order p ,

$$u[n] = \underbrace{[k_0 \ \cdots \ k_{p-1}]}_{k'} \begin{bmatrix} y[n] \\ \vdots \\ y[n - p + 1] \end{bmatrix}, \quad n \in \mathbb{Z}$$

that minimizes the m -step ahead prediction error $E(y[n + m] - u[n])^2$.

This is Application 15 for $v[n] = y[n + m]$. So we can specialize Lemma 12.2.1:

Theorem 12.5.1 (Linear prediction). The $k_* \in \mathbb{R}^p$ solves Application 17 iff

$$\begin{bmatrix} r_y[0] & r_y[1] & \cdots & r_y[p-1] \\ r_y[1] & r_y[0] & \ddots & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ r_y[p-1] & r_y[p-2] & \cdots & r_y[0] \end{bmatrix} \begin{bmatrix} k_0 \\ k_1 \\ \vdots \\ k_{p-1} \end{bmatrix} = \begin{bmatrix} r_y[m] \\ r_y[m+1] \\ \vdots \\ r_y[p+m-1] \end{bmatrix}.$$

Example 12.5.2 (Second order FIR filter). Suppose

$$r_y[\tau] = a^{|\tau|}$$

where $|a| \leq 1$ (otherwise it is not a covariance function) and try a second order FIR filter. The Wiener-Hopf normal equations become

$$\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} \begin{bmatrix} k_0 \\ k_1 \end{bmatrix} = \begin{bmatrix} a^m \\ a^{m+1} \end{bmatrix}$$

so that

$$\begin{bmatrix} k_0 \\ k_1 \end{bmatrix} = \frac{1}{1-a^2} \begin{bmatrix} 1 & -a \\ -a & 1 \end{bmatrix} \begin{bmatrix} a^m \\ a^{m+1} \end{bmatrix} = \frac{1}{1-a^2} \begin{bmatrix} a^m - a^{m+2} \\ 0 \end{bmatrix} = \begin{bmatrix} a^m \\ 0 \end{bmatrix}.$$

Perhaps surprisingly the best linear prediction m steps ahead is simply the last available $y[t]$ scaled by a factor a^m . Note that for $a = 1$ the covariance matrix is singular but still $k_0 = a^m, k_1 = 0$ is a solution of the Wiener-Hopf equations (Lemma (12.2.1)).

The number a determines the temporal relatedness in the process. For $a = 0$ the covariance function is that of white noise — i.e. there is no temporal relatedness — and it is not surprising to find that the linear predictor then is zero: $\hat{y}[t+m] := u[t] = 0$. In contrast if $a = 0.99$ then the covariance function decreases very slowly so we expect that prediction is good, that is, that the prediction error is small

$$e[t] := y[t+m] - u[t] = y[t+m] - a^m y[t].$$

The power (variance) of the prediction error is

$$\begin{aligned} E(e^2[t]) &= E(y[t+m] - a^m y[t])^2 = E(y[t+m]^2 - 2a^m y[t+m]y[t] + a^{2m} y[t]^2) \\ &= r[0] - 2a^m r[m] + a^{2m} r[0] = 1 - 2a^{2m} + a^{2m} \\ &= 1 - a^{2m}. \end{aligned}$$

For $a = 1$ the prediction error is in fact zero. This is the somewhat special case that $r[\tau] = r[0]$ which according to Cauchy-Schwarz is the case iff $y[\tau] = y[0]$ “with probability 1” i.e. that all realizations are constant. $\square$

An alternative, more systems approach to linear prediction is done next.

Example 12.5.3. Consider again the m -step ahead prediction problem of y and that y is zero-mean WSS with covariance function $r_v[\tau] = r_y[\tau] = a^{-|\tau|}$. This implies that v can be seen as the output of a system

$$v = \frac{1-a^2}{1-az^{-1}}w$$

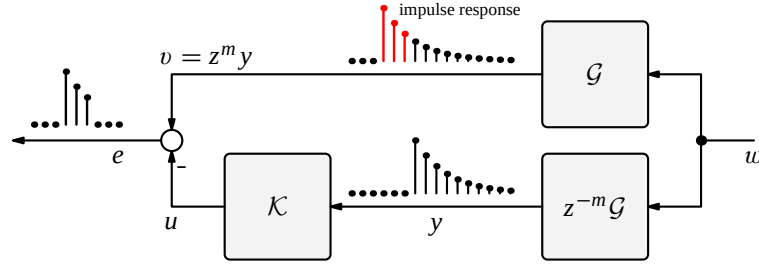


Figure 12.6: Equivalent filtering configuration

in which w is zero-mean unit variance white noise (see Example 7.13.2). According to Lemma 7.13.1 the power of e that we aim to minimize equals the ℓ^2 -norm squared of the mapping from w to e ,

$$\text{var}(e[n]) = \|(1 - \mathcal{K}z^{-m})\mathcal{G}\|_{\ell^2}^2.$$

If we are to minimize with respect to causal $\mathcal{K}$ then the first m entries of the “closed loop” impulse response can not be affected by $\mathcal{K}$. The best we can do is kill the tail. From Fig. 12.6 it will be clear that to kill the tail we need

$$K(z) = a^m.$$

That is, the optimal $\mathcal{K}_*$ is

$$u = \mathcal{K}(y) : \quad u[t] = a^m y[t].$$

This is the same filter as what we found in Example 12.5.2. The difference is that here we optimized over *any* causal $\mathcal{K}$ while in Example 12.5.2 we considered FIR systems. Owing to the specific structure of the covariance function r_v the two results however are the same. Usually the answers differ. $\square$

12.6 IIR Wiener filter — causal linear prediction

In Signal Processing FIR systems are often set apart from IIR (infinite impulse response) systems. In this section we cover Wiener filtering for causal IIR systems, in other words causal LTI systems. We only deal with prediction in this section and we assume to have a bicausal, bistable whitening filter for v , that is, that we can write v as

$$v = \mathcal{H}(w)$$

for some zero-mean unit variance white noise process and bistable, bicausal $\mathcal{H}$. According to Lemma 7.15.1 (and the comment immediately following that lemma) effectively every WSS v has such a representation. Given such a representation it is fairly straightforward to derive an explicit representation for the optimum m -step ahead predictor.

Application–18: Let $y = z^{-m}v = \mathcal{H}(w)$ with w a zero-mean unit-variance white noise process and $\mathcal{H}$ a bicausal, bistable LTI with impulse response h in $\ell^2(\mathbb{N})$. Determine the m -step ahead predictor $u = \mathcal{K}(y)$ of y by minimizing

$$\min_{u=\mathcal{K}(y), \text{causal}} \mathbb{E}(v[n] - u[n])^2.$$

Figure 12.6 illustrates that the impulse response of mapping from w to the prediction error $e = v - \mathcal{K}(y)$ can not be affected by filter $\mathcal{K}$ at $n = 0, 1, \dots, M - 1$. That is, whatever causal predictor (optimal or not) we take the mean square error (=variance=power) must equal or exceed

$$h[0]^2 + h^2[1] + \dots + h^2[M - 1].$$

Lemma 12.6.1. Let $H(z)$ be a bistable, biproper transfer function with impulse response $h \in \ell^2(\mathbb{N})$. Then filter $\mathcal{K}_*$ solves Application 18 iff

$$K_*(z) = \frac{h[M] + h[M + 1]z^{-1} + h[M + 2]z^{-2} + \dots}{H(z)}.$$

It is stable system and renders the MMSE equal to

$$\mathbb{E} e^2[n] = h[0]^2 + h^2[1] + \dots + h^2[M - 1].$$

If $H(z)$ is rational, $H(z) = B(z)/A(z)$ with $B(z), A(z)$ polynomials in z^{-1} , then the optimal filter $\mathcal{K}_*$ has rational transfer function

$$K_*(z) = \frac{z^M R_M(z)}{B(z)} \quad (12.3)$$

where $R_M(z)$ is the M th order remainder in the expansion

$$\frac{B(z)}{A(z)} = h[0] + h[1]z^{-1} + h[2]z^{-2} + \dots + h[M - 1]z^{1-M} + \frac{R_M(z)}{A(z)} \quad (12.4)$$

with remainder R_M of the form $R_M(z) = z^{-M}(r[0] + r[1]z^{-1} + \dots)$.

Proof. Do it yourself. ■

Lemma 12.6.2 (prediction error is white). For $M = 1$ the prediction error $e[t] := y[t + 1] - u[t]$ is a zero mean white noise process.

Proof. See Problem 12.6. ■

Expansion 12.4 is an example of “division with remainder” and is easily obtained using long division.

Example 12.6.3. By way of example consider

$$v = \frac{1}{1 - az^{-1}} w \quad (12.5)$$

We derive the M -step predictor formally. We have $A(z) = 1 - az^{-1}$ and $B(z) = 1$, then long division gives us that

$$\begin{aligned} \frac{B(z)}{A(z)} &= \frac{1}{1 - az^{-1}} \\ &= 1 + az^{-1} + a^2 z^{-2} + \dots + a^{M-1} z^{1-M} + \frac{z^{-M} a^M}{1 - az^{-1}}. \end{aligned} \quad (12.6)$$

For the sake of stability we need to assume that $|a| < 1$. Apparently we have $h[i] = a^i$, $i = 0, 1, \dots$ and $R_M(z) = z^{-M}a^M$. This yields as m -step predictor (12.3)

$$\hat{y}[t + M] := u[t] = a^M y[t]. \quad (12.7)$$

□

More general IIR wiener filtering can be solved using spectral factorization (see Section 11.5).

12.7 Standard biased estimate of covariances — least squares

This course is not about estimation of covariances but we have to mention at least *one* method. It is the standard biased estimate and it is the vantage point for many variations. Given a finite sequence of N samples (extended with zeros on both sides)

$$w = (\dots, 0, 0, 0, w[1], w[2], \dots, w[N], 0, 0, 0, \dots)$$

assumed to come from a zero-mean WSS process, the *standard biased estimate* of

$$r_w[\tau] = \text{cov}(w[n + \tau], w[n])$$

is

$$\hat{r}_w[\tau] = \frac{1}{N} \sum_{n \in \mathbb{Z}} w[n + \tau] w[n].$$

Of course the summation is a finite one because only a finite number of samples $w[n]$ are nonzero. The estimate of the $p \times p$ covariance *matrix*

$$R_w^{p \times p} := \begin{bmatrix} r_w[0] & r_w[1] & \cdots & r_w[p-1] \\ r_w[1] & r_w[0] & \ddots & r_w[p-2] \\ \vdots & \ddots & \ddots & \vdots \\ r_w[p-1] & r_w[p-2] & \cdots & r_w[0] \end{bmatrix}$$

can now elegantly be formulated as

$$\hat{R}_w^{p \times p} = \frac{1}{N} W' W, \quad W := \begin{bmatrix} w[1] & 0 & \cdots & \cdots & 0 \\ w[2] & w[1] & \ddots & \cdots & \vdots \\ \vdots & w[2] & \ddots & 0 & \vdots \\ \vdots & \vdots & \ddots & w[1] & 0 \\ \vdots & \vdots & \ddots & w[2] & w[1] \\ w[N] & \vdots & \ddots & \vdots & w[2] \\ 0 & w[N] & \ddots & \vdots & \vdots \\ \vdots & 0 & \ddots & \vdots & \vdots \\ \vdots & \vdots & \ddots & w[N] & \vdots \\ 0 & \cdots & \cdots & 0 & w[N] \end{bmatrix} \in \mathbb{R}^{(N+p-1) \times p}.$$

Here we recognize the least squares normal equations that we encountered in Chapter 9. An important reason to use the standard biased estimate is that estimated covariance matrices are by construction nonnegative definite (which actual covariance matrices are as well).

Give two two finite realizations,

$$\begin{aligned} w &= (\cdots, 0, 0, 0, w[1], w[2], \dots, w[N], 0, 0, 0, \dots), \\ x &= (\cdots, 0, 0, 0, x[1], x[2], \dots, x[N], 0, 0, 0, \dots) \end{aligned}$$

the *crosscovariance*

$$r_{wx}[\tau] = \text{cov}(w[n + \tau], x[n])$$

is similarly estimated as

$$\hat{r}_{wX}[\tau] = \frac{1}{N} \sum_{n \in \mathbb{Z}} w[n + \tau] x[n].$$

12.8 Matched filter and the signal-to-noise ratio

This section show that the inequality of Cauchy-Schwarz has interesting practical implications.

Definition 12.8.1 (signal-to-noise-ratio (SNR)). Let $y = v + \eta$ with v a desired signal and η noise. The *signal-to-noise ratio (SNR)* is informally defined as the ratio of the power of the desired signal and noise

$$\text{SNR} = \frac{\|v\|^2}{\|\eta\|^2}$$

with $\|x\|^2 := \mathbb{E} |x|^2$.

The SNR seems to come in different forms. Some use instead $\frac{\|v+\eta\|^2}{\|\eta\|^2}$ and (Mallat, 1998, p. 426) defines it as

$$\frac{\|v\|^2}{\|v - \hat{v}\|^2}$$

where $\hat{v}$ is any estimator of v thus making the SNR a property of an estimator (together with the process v). In Mallat's version of the SNR the wiener filtering problem can equivalently be seen as the problem of maximizing the SNR. Back to the SNR of Definition 12.8.1.

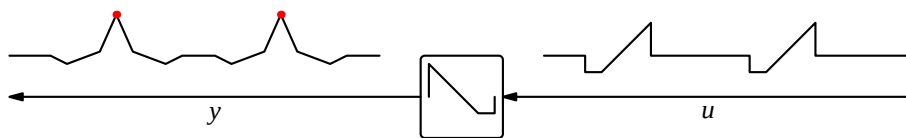


Figure 12.7: Matched filter for template $x = \nearrow$

Somewhat informally we define:

Application—19: Find a discrete filter $y = \mathcal{H}(u)$ that in real time can detect the presence of a known finite template $x : \{0, 1, \dots, M\} \rightarrow \mathbb{R}$ in a discrete signal $u : \mathbb{Z} \rightarrow \mathbb{R}$.

The *matched filter* tries to solve this problem as indicated in Fig. 12.7. The impulse response h of the system is taken to be the template x reversed, modulo scaling and translation,

$$h[t] = x[M - t]$$

with $M \in \mathbb{Z}$ such that h is causal. The result is that $y[M]$ is the dot product u and template x :

$$y[M] = (h * u)[M] = \sum_k h[M - k]u[k] = \sum_k x[k]u[k] = x \cdot u.$$

Then by time-invariance $y[n + M]$ is the dot product of $u[\cdot + n]$ and x . Cauchy-Schwarz tells us that the dot product (or any inner product for that matter) when normalized is at most 1 in absolute value. This would suggest a nonlinear version of the matched filter.

Lemma 12.8.2 (Normalized matched filter). Let $x : \{0, 1, M\} \rightarrow \mathbb{R}$ be a finite template and define

$$h[t] = \frac{x[M - t]}{|x|}$$

then the output y of the nonlinear finite dimensional system

$$y = \mathcal{H}(u) : \quad y[n] = \frac{h[M]u[n - M] + \cdots h[1]u[n - 1] + h[0]u[n]}{\sqrt{u^2[n - M] + \cdots + u^2[n - 1] + u^2[n]}}$$

at any $n \in \mathbb{Z}$ satisfies $-1 \leq y[n] \leq 1$ and equality is achieved iff

$$\begin{bmatrix} u[n - M] & \cdots & u[n - 1] & u[n] \end{bmatrix} = \lambda \begin{bmatrix} x[0] & \cdots & x[M - 1] & x[M] \end{bmatrix}$$

for some nonzero $\lambda \in \mathbb{R}$.

Proof. Do it yourself. ■

This assumes that the signal u is noise free and that the templates x in u do not overlap.

```
x=[-1 -1 -1 -1:6]; % whatever
h=fliplr(x);
u=[0 0 0 x 0 0 0]; % well...
y=filter(h,1,u); % should peak at size(h)+3
```

12.9 Matched filters for noisy signals

A beautiful and useful extension is if the desired signal (series of known templates) is additively corrupted by noise. Still we want to pinpoint where the templates occur. Again somewhat informally we define:

Application–20: Suppose $u = v + \eta$ with η zero mean white noise and v a signal that contains a series of finite templates x . Find an LTI system $y = \mathcal{H}(u)$ that maximizes the signal-to-noise ratio

$$\min_{\mathcal{H}} \frac{|\mathcal{H}(x)|^2}{\mathbb{E} |\mathcal{H}(\eta)|^2}.$$

The problem is stylized a bit to allow an elegant solution. Whatever the optimal system is, the answer is not unique for if $\mathcal{H}_*$ is a solution then so is any scalar multiple. The central result is this static version.

Lemma 12.9.1. Let $x : \{0, 1, \dots, M\} \rightarrow \mathbb{R}$ be given and suppose $\eta[0], \eta[1], \dots, \eta[M]$ is a zero-mean white noise (finite) process. Then

$$\frac{|d_0x[0] + d_1x[1] + \dots + d_Mx[M]|^2}{\mathbb{E}(d_0\eta[0] + d_1\eta[1] + \dots + d_M\eta[M])^2}$$

is minimal iff d is a nonzero multiple of x ,

$$(d_0, d_1, \dots, d_M) = \lambda(x[0], x[1], \dots, x[M]), \quad \lambda \neq 0.$$

Proof. Since η is white noise we have $\mathbb{E}(d_0\eta[0] + d_1\eta[1] + \dots + d_M\eta[M])^2 = \mathbb{E}(\eta^2)(d_0^2 + d_1^2 + \dots + d_M^2) = \mathbb{E}(\eta^2)|d|^2$. Now by Cauchy-Schwarz

$$\frac{|d_0x[0] + d_1x[1] + \dots + d_Mx[M]|^2}{\mathbb{E}(d_0\eta[0] + d_1\eta[1] + \dots + d_M\eta[M])^2} \leq \frac{|d|^2|x|^2}{\mathbb{E}(\eta^2)|d|^2} = \frac{|x|^2}{\mathbb{E}(\eta^2)}$$

with equality holding iff d and x are aligned (see Theorem 5.6.1). ■

In an LTI system $y = \mathcal{H}(x)$ the indices of the weights $h[i]$ for x are reversed compared to the weights d_i in the above lemma,

$$y[t] = h[0]x[t] + h[1]x[t-1] + \dots + h[M]x[t-M].$$

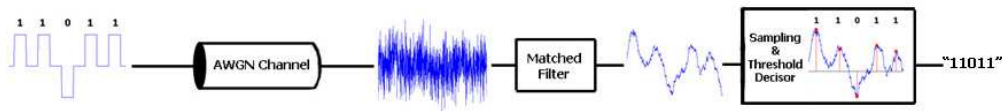
So optimal is to take as impulse response h the flipped d of Lemma 12.9.1 i.e. the flipped template x , modulo scaling and translation. Typically one takes

$$h[t] = x[M-t], \quad t = 0, 1, \dots, M$$

as this makes the LTI system causal (needed for real time implementations). The presence of a template x in a signal u then is captured only after the last sample of the template is available. That makes sense.

The following text and example is taken from Wikipedia. Matched filters are commonly used in radar, in which a signal is sent out, and we measure the reflected signals, looking for something similar to what was sent out. Pulse compression is an example of matched filtering. Two-dimensional matched filters are commonly used in image processing, e.g., to improve SNR for X-ray pictures.

Example 12.9.2 (digital communications). In the context of a communication system which sends binary messages from the transmitter to the receiver across a noisy channel, a matched filter can be used to detect the transmitted pulses within a received signal. Imagine we want to



send the sequence “11011000100” coded in polar Return-to-zero (RZ) through a certain channel. Mathematically, a sequence in RZ code can be described as a sequence of pulses, each pulse being weighted by +1 if the bit to send is “1” and by “-1” if the bit to send is “0”. When going

through the noisy channel additive white Gaussian noise (AWGN) is added to the signal. Before sampling, the received signal is filtered with a matched filter. In this case, the filter should be matched to a RZ (Return-to-zero) pulse (equivalent to a "1" coded in RZ code). Precisely, the impulse response of the ideal matched filter should be one pulse

$$h[t] = \text{RZ pulse}$$

□

In this example from WIKIPEDIA the noise is assumed Gaussian. Since we are looking for LTI filters the assumed Gaussianity is not needed. That the noise is white is sufficient. Generalizations to non-white noise exist as well (see Exercise 12.4).

12.10 Problems

12.1 Consider the system (also considered in Problem 7.1):

$$y[n] - 1.97y[n-1] + 0.98y[n-2] = w[n]$$

with w zero-mean unit-variance white noise.

- Compute the poles of the system.
- Determine the 1-step ahead linear predictor $u = \mathcal{K}(y)$ of y .
- Generate 1000 samples of y and plot y and $y[n+1] - u[n]$.
- Compute the MMSE (exact) $E(y[n+1] - u[n])^2$ and compare it with the above error signal.
- Use the 1000 samples to estimate the necessary (cross)covariances (Section 12.7) needed to compute the optimal FIR filter $u = \mathcal{K}_p(y)$ of order $p = 50$ and compare the so computed impulse response with the impulse response of $u = \mathcal{K}(y)$ obtained in part 12.1b of this problem.

12.2 Does anything change in the solution (Lemma 12.1.1) of Application 14 if the zero-mean assumption is removed.

12.3 Smoothing?

12.4 Nonwhite Matched filter..

12.5 Why is the covariance matrix in (12.2) equal to $R_b + R_\eta$?

12.6 Prove Lemma 12.6.2.

12.7 Prove Lemma 12.6.1.

13 JPEG and 2D-Fourier

In this chapter we will explore ways of representing images (finite rectangular 2D-signals) in different bases with the aim of lossy compression. We summarize the JPEG-1992 standard.

13.1 Images as matrices

We consider only monochrome images in shades of grey (often called black-and-white images) and the images are assumed to be rectangles. Such images can be identified with a matrix of color intensities. For example,

$$\begin{bmatrix} 0.0 & 0.1 & 0.2 & 0.3 \\ 0.4 & 0.5 & 0.6 & 0.7 \\ 0.8 & 0.9 & 1.0 & 0.97 \end{bmatrix} \longleftrightarrow \begin{array}{cccc} \blacksquare & \blacksquare & \blacksquare & \blacksquare \\ \blacksquare & \blacksquare & \blacksquare & \blacksquare \\ \blacksquare & \blacksquare & \blacksquare & \blacksquare \end{array}$$

Here 0 corresponds to a black pixel and 1 corresponds to a white pixel, and any value in between, $0 < f_{ij} < 1$, is some form of grey. Other conventions exist, for instance where the entries f_{ij} are integer numbers from $\{0, 1, \dots, 255\}$ with 0 being black and 255 white. For analysis it will be convenient to allow every real number, such as $f_{ij} = 20.34$ and $f_{ij} = -\pi$ and for the most part we need not worry about what number corresponds to what level of grey. (One could take any $f_{ij} \geq 1$ to represent white and any $f_{ij} \leq 0$ to represent black, but different choices of saturation can be dreamed up).

13.2 Introduction: bases for 2×2 images

What is a basis for $\mathbb{R}^{2 \times 2}$? The natural basis is

$$e_1 := \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}, \quad e_2 := \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix}, \quad e_3 := \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix}, \quad e_4 := \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}. \quad (13.1)$$

We can equip $\mathbb{R}^{2 \times 2}$ with inner product

$$\langle f, g \rangle = \sum_{i=1}^2 \sum_{j=1}^2 f_{ij} g_{ij} \quad (13.2)$$

which is just a recasting of the dot product on $\mathbb{R}^4$. In this inner product the natural basis (13.1) is an orthonormal basis. Other bases exist as well, of course. Here is one,

$$v_1 = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \quad v_2 = \begin{bmatrix} 1 & 0 \\ 1 & 0 \end{bmatrix}, \quad v_3 = \begin{bmatrix} 1 & 1 \\ 0 & 0 \end{bmatrix}, \quad v_4 = \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix}.$$

This basis is not orthonormal in the standard inner product 13.2. It can be turned into an orthonormal basis using the Gram-Schmidt procedure or QR -factorization, but for this simple case

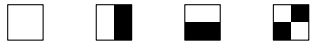
we can do it manually as well,

$$\hat{e}_1 = \begin{bmatrix} 0.5 & 0.5 \\ 0.5 & 0.5 \end{bmatrix}, \quad \hat{e}_2 = \begin{bmatrix} 0.5 & -0.5 \\ 0.5 & -0.5 \end{bmatrix}, \quad \hat{e}_3 = \begin{bmatrix} 0.5 & 0.5 \\ -0.5 & -0.5 \end{bmatrix}, \quad \hat{e}_4 = \begin{bmatrix} 0.5 & -0.5 \\ -0.5 & 0.5 \end{bmatrix}. \quad (13.3)$$

By definition of basis, no information is lost if we move from one basis to the other. However, if we want to compress or otherwise modify the image by removing basis elements that are “not important” then the choice of basis does matter. The natural basis (13.1) is probably not suitable as they are the same modulo rotation / reflection:



The orthonormal basis (13.3) holds more promise:



We have by orthonormality that any $f \in \mathbb{R}^{2 \times 2}$ has Fourier series

$$f = \langle f, \hat{e}_1 \rangle \hat{e}_1 + \langle f, \hat{e}_2 \rangle \hat{e}_2 + \langle f, \hat{e}_3 \rangle \hat{e}_3 + \langle f, \hat{e}_4 \rangle \hat{e}_4.$$

Since $\hat{e}_1$ is constant we might want to remove the corresponding entry $\langle f, \hat{e}_1 \rangle \hat{e}_1$ if it is edge detection we are after. To the human eye the fourth basis element $\hat{e}_4$ (when scaled down) may be hardly visible so for compression the last entry $\langle f, \hat{e}_4 \rangle \hat{e}_4$ might be a good candidate for removal.

13.3 Bases functions for images (2D-signals)

How do we generalize the idea of the previous section to large $n \times n$ images? The following little example gives a hint.

Example 13.3.1 (Construction of basis). Let $e_1 = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$ and $e_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$. This is the standard basis of $\mathbb{R}^2$. The standard basis 13.1 for $\mathbb{R}^{2 \times 2}$ can be written as

$$\begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix} = e_1 e'_1, \quad \begin{bmatrix} 0 & 1 \\ 0 & 0 \end{bmatrix} = e_1 e'_2, \quad \begin{bmatrix} 0 & 0 \\ 1 & 0 \end{bmatrix} = e_2 e'_1, \quad \begin{bmatrix} 0 & 0 \\ 0 & 1 \end{bmatrix} = e_2 e'_2.$$

□

The example suggests (correctly) that we can construct orthonormal bases for $\mathbb{R}^{n \times k}$ given *any* two bases for $\mathbb{R}^n$ and $\mathbb{R}^k$. We formulate and prove this result actually for the vector space of “continuous images” $\mathbb{L}^2([a, b] \times [a, b])$. The proof for our discrete images is essentially the same (easier). For simplicity we consider square images only.

Theorem 13.3.2. If $\{e_1, e_2, \dots\}$ is an orthonormal basis of $\mathbb{L}^2[a, b]$ then

$$\phi_{ij}(t, s) := e_i(t)e_j(s), \quad i \in \mathbb{N}, j \in \mathbb{N}$$

is an orthonormal basis of $\mathbb{L}^2([a, b] \times [a, b])$ with inner product

$$\langle x, y \rangle = \int_{t=a}^{t=b} \int_{s=a}^{s=b} x(t, s) \overline{y(t, s)} \, ds \, dt.$$

Proof. First the ϕ_{ij} is an orthonormal set:

$$\begin{aligned}\langle \phi_{ij}, \phi_{kn} \rangle &= \int_a^b \int_a^b \phi_{ij}(t, s) \overline{\phi_{kn}(t, s)} \, ds \, dt \\ &= \int_a^b \int_a^b e_i(t) e_j(s) \overline{e_k(t) e_n(s)} \, ds \, dt \\ &= \int_a^b e_i(t) \overline{e_k(t)} \, dt \int_a^b e_j(s) \overline{e_n(s)} \, ds\end{aligned}$$

which by orthogonality of the e_j is nonzero iff $i = k$ or $j = n$ in which case it is 1. This proves orthonormality of the ϕ_{ij} .

To show completeness we use Theorem 5.17.2 (and that our space is a Hilbert space): suppose $g \in \mathbb{L}^2([a, b] \times [a, b])$ is orthogonal to all ϕ_{ij} . Then

$$0 = \langle g, \phi_{ij} \rangle = \iint_{t,s} g(t, s) \overline{\phi_{ij}(t, s)} \, dt \, ds = \int_s \overline{e_j(s)} \underbrace{\int_t g(t, s) \overline{e_i(t)} \, dt}_{h_i(s)} \, ds \quad \forall i.$$

Thus $h_i \perp e_j$ for all j which by completeness of the basis implies $h_i(s) = 0$ (for almost all s). Since $0 = h_i(s) = \int_t g(t, s) \overline{e_i(t)} \, dt$ for all i this, in turn, again by completeness of the e_i implies that $g(t, s) = 0$ (for almost all s, t). ■

13.4 Two-dimensional Fourier series

An immediate application of Theorem 13.3.2 and Lemma 6.4.2 is that

$$\phi_{kn}(t, s) = e^{i2\pi(kt+ns)} \quad k, n \in \mathbb{Z}$$

is an orthonormal basis of $\mathbb{L}^2([0, 1] \times [0, 1])$. Any $f \in \mathbb{L}^2([0, 1] \times [0, 1])$ therefore has an expansion

$$f(t, s) = \sum_{k,n \in \mathbb{Z}} \hat{f}_{kn} e^{i2\pi(kt+ns)}$$

and by orthonormality the 2D Fourier coefficients readily follow as

$$\hat{f}_{k,n} = \langle f, \phi_{k,n} \rangle = \int_0^1 \int_0^1 f(t, s) e^{-i2\pi(kt+ns)} \, ds \, dt.$$

A way to visualize the basis elements $\phi_{kn}(t, s) = e^{i2\pi(kt+ns)}$ is to plot $\operatorname{Re} \phi_{kn}(t, s)$ and $\operatorname{Im} \phi_{kn}(t, s)$, where white means 1 and black means -1 , see Fig. 13.1. Clearly for $k = n = 0$ we have a constant level (a white square) and with increasing k and n high-frequency ripples show up.

13.5 Cosine transform

Ordinary Fourier series of one-dimensional $f \in \mathbb{L}^2[0, 1]$ has difficulty modeling discontinuities (Gibbs phenomenon). Discontinuities typically appear at the boundary of its domain

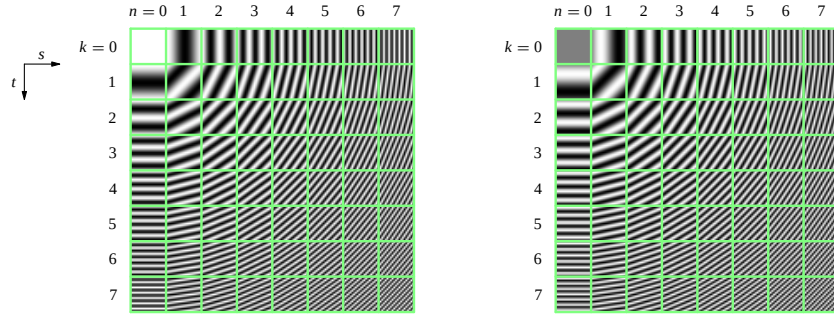
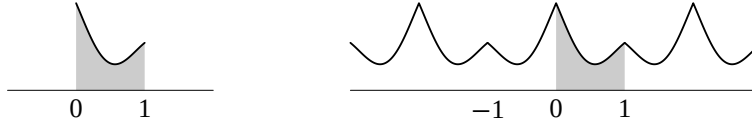


Figure 13.1: $\text{Re } \phi_{kn}(t, s)$ and $\text{Im } \phi_{kn}(t, s)$ for $k, n \in \{0, 1, 2, \dots, 7\}$ (black=-1; white=1)

($t = 0$, $t = 1$) because Fourier assumes that the function is periodically extended. An easy and effective way to remove this troublesome discontinuity at the boundary is to evenly extend the function $f : [0, 1] \rightarrow \mathbb{R}$ to the entire real axis, as explained by this figure:



The even extension has double the period and is continuous if f is, so the Gibbs phenomenon at the edges (here 0 and 1) disappears. Since the extended function by construction is even, the Fourier series becomes even and real,

$$f(t) = \sum_{k=0}^{\infty} \hat{f}_k \cos(\pi kt), \quad \hat{f}_0 = \int_0^1 f(t) dt, \quad \hat{f}_k = 2 \int_0^1 f(t) \cos(\pi kt) dt \quad (k \neq 0).$$

The same idea applies to images. An image can be evenly extended, see Fig. 13.2, rendering the Fourier series even and real (even in both directions),

$$f(t, s) = \sum_{k, n \in \mathbb{N}} \hat{f}_{kn} \cos(\pi kt) \cos(\pi ns).$$

The new basis functions $\cos(\pi kt) \cos(\pi ns)$ are shown in Fig. 13.3.

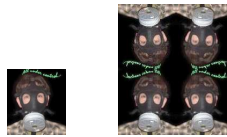


Figure 13.2: An even extension

13.6 Discrete two-dimensional cosine transform

We briefly want to mention that in applications (e.g. JPEG) where we work with discrete signals, often the DCT-II (discrete cosine transformation, type 2) is used. The one-dimensional version

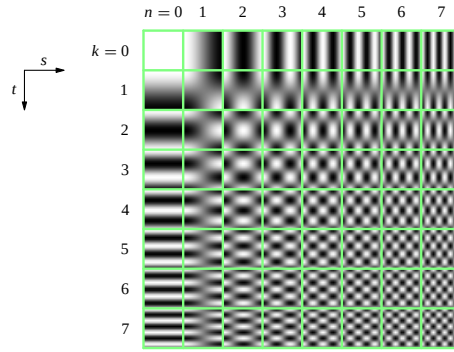
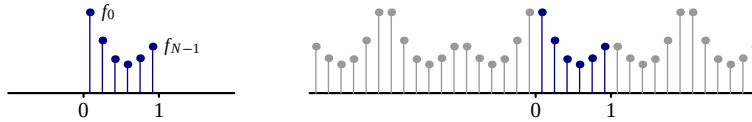


Figure 13.3: $\cos(\pi kt) \cos(\pi ns)$ $k, n \in \{0, 1, 2, \dots, 7\}$ (black=-1; white=1)

is explained by this figure:



Seen as samples of a continuous-time signal, the first sample f_0 is taken at (continuous) time $\frac{1}{2N}$ with N the number of samples. In two dimensions the DCT-II expansion becomes

$$f_{t,s} = \sum_{0 \leq k, n < N} \hat{f}_{kn} \cos(\pi k \frac{t+1/2}{N}) \cos(\pi n \frac{s+1/2}{N}), \quad t, s = \{0, 1, \dots, N-1\}. \quad (13.4)$$

Multi-dimensional versions of the FFT exist that can compute the Fourier coefficients $\hat{f}_{kn}$ efficiently. This way we have a finite (N^2 elements) basis $\cos(\pi k \frac{t+1/2}{N}) \cos(\pi n \frac{s+1/2}{N})$. They look very similar to that of Fig. 13.3.

13.7 JPEG

Application–21: Find a efficient representation (small file & easy to encode and decode) of images, possibly lossy (=at the cost of some loss of image quality).

The JPEG standard of 1992 achieves this by storing the 2D Fourier coefficients (cosine transform actually) of the image, whereby removing high frequency components and coefficients that are negligible small. The following is largely based on Wikipedia. A jpeg-1992 decoding algorithm consists of:

1. *Color space transformation and downsampling.* The three color channels (RGB=red, green, blue) are obviously correlated. Handling each channel independent of the others would therefore not be efficient. JPEG-1992 exploits the correlation by converting the RGB-space to another space (also three channels). We will not deal with these matters here. We only deal with monochrome images. After color space transformation we end up with three other channels (Y,Cb,Cr). The human eye is more sensitive to Y than to (Cb,Cr) so downsampling (Cb,Cr) is an option. Again we are not going into details.

2. Split each channel into a series of 8×8 blocks of 64 pixels. The remainder of the steps is to be performed on all 8×8 blocks. Suppose the 8×8 block is (in grey-scale with entries in $\{0, 1, \dots, 255\}$):

$$Y = \begin{bmatrix} 52 & 55 & 61 & 66 & 70 & 61 & 64 & 73 \\ 63 & 59 & 55 & 90 & 109 & 85 & 69 & 72 \\ 62 & 59 & 68 & 113 & 144 & 104 & 66 & 73 \\ 63 & 58 & 71 & 122 & 154 & 106 & 70 & 69 \\ 67 & 61 & 68 & 104 & 126 & 88 & 68 & 70 \\ 79 & 65 & 60 & 70 & 77 & 68 & 58 & 75 \\ 85 & 71 & 64 & 59 & 55 & 61 & 65 & 83 \\ 87 & 79 & 69 & 68 & 65 & 76 & 78 & 94 \end{bmatrix}$$


3. Subtract 128 (which roughly centers the image around zero)

$$F := Y - 128 = \begin{bmatrix} -76 & -73 & -67 & -62 & -58 & -67 & -64 & -55 \\ -65 & -69 & -73 & -38 & -19 & -43 & -59 & -56 \\ -66 & -69 & -60 & -15 & 16 & -24 & -62 & -55 \\ -65 & -70 & -57 & -6 & 26 & -22 & -58 & -59 \\ -61 & -67 & -60 & -24 & -2 & -40 & -60 & -58 \\ -49 & -63 & -68 & -58 & -51 & -60 & -70 & -53 \\ -43 & -57 & -64 & -69 & -73 & -67 & -63 & -45 \\ -41 & -49 & -59 & -60 & -63 & -52 & -50 & -34 \end{bmatrix}$$

4. Apply discrete cosine transform type II to F and round to nearest integer,

$$\hat{F} = \begin{bmatrix} -415 & -30 & -61 & 27 & 56 & -20 & -2 & 0 \\ 4 & -22 & -61 & 10 & 13 & -7 & -9 & 5 \\ -47 & 7 & 77 & -25 & -29 & 10 & 5 & -6 \\ -49 & 12 & 34 & -15 & -10 & 6 & 2 & 2 \\ 12 & -7 & -13 & -4 & -2 & 2 & -3 & 3 \\ -8 & 3 & 2 & -6 & -2 & 1 & 4 & 2 \\ -1 & 0 & 0 & -2 & -1 & -3 & 4 & -1 \\ 0 & 0 & -1 & -4 & -1 & 0 & 1 & 2 \end{bmatrix}$$

Here the matrix $\hat{F}$ is the matrix of Fourier coefficients $\hat{f}_{kn}$ (see Eqn. (13.4)) ordered corresponding to the matrix of basis functions (see Fig. 13.3).

5. The human eye is good at seeing small differences in brightness over a relatively large area, but not so good at distinguishing the exact strength of a high frequency brightness variation. This fact allows one to get away with greatly reducing the amount of information in the high frequency components. This is done by simply dividing each component in the frequency domain by a constant for that component, and then rounding to the nearest integer. This is the main lossy operation in the whole process. As a result of this, it is typically the case that many of the higher frequency components are rounded to zero, and many of the rest become small positive or negative numbers, which take many fewer bits to store.

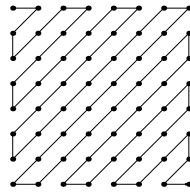
The *quantization matrix* to divide with is (for our example)

$$Q = \begin{bmatrix} 16 & 11 & 10 & 16 & 24 & 40 & 51 & 61 \\ 12 & 12 & 14 & 19 & 26 & 58 & 60 & 55 \\ 14 & 13 & 16 & 24 & 40 & 57 & 69 & 56 \\ 14 & 17 & 22 & 29 & 51 & 87 & 80 & 62 \\ 18 & 22 & 37 & 56 & 68 & 109 & 103 & 77 \\ 24 & 35 & 55 & 64 & 81 & 104 & 113 & 92 \\ 49 & 64 & 78 & 87 & 103 & 121 & 120 & 101 \\ 72 & 92 & 95 & 98 & 112 & 100 & 103 & 99 \end{bmatrix}$$

The numbers go up along the diagonal (except the last entry!), expressing that we are less sensitive to high frequencies. Determine the element-wise quotient and round the result to nearest integer:

$$\hat{G} := \hat{F} ./ Q = \begin{bmatrix} -26 & -3 & -6 & 2 & 2 & -1 & 0 & 0 \\ 0 & -2 & -4 & 1 & 1 & 0 & 0 & 0 \\ -3 & 1 & 5 & -1 & -1 & 0 & 0 & 0 \\ -4 & 1 & 2 & -1 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}$$

6. Reorder the matrix $\hat{G}$ into a row of numbers following this pattern:



For our examples this results in

26, 3, 0, 3, 2, 6, 2, 4, 1, 4, 1, 1, 5, 1, 2, 1, 1, 1, 2, 0, 0, 0, 0, 0, 1, 1, 0, 0, ...

with many zeros at the end.

7. Zip the numbers (typically Run-length encoding and Huffman)
8. Save the zipped number to a file, say, `bla.jpg`.

Whenever we download a jpeg-1992 image our computer reverses process (decoding) as follows:

1. Unzip the numbers, reorder them in a matrix $\hat{G}$.
2. Multiply $\hat{G}$ elementwise with the quantization matrix

$$\hat{F} := \hat{G} .* Q = \begin{bmatrix} -416 & -33 & -60 & 32 & 48 & -40 & 0 & 0 \\ 0 & -24 & -56 & 19 & 26 & 0 & 0 & 0 \\ -42 & 13 & 80 & -24 & -40 & 0 & 0 & 0 \\ -56 & 17 & 44 & -29 & 0 & 0 & 0 & 0 \\ 18 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

3. Determine inverse cosine transform F of $\hat{F}$, round the result and add 128. For our example this would result in

$$Y_{\text{decoded}} = \img alt="A small, dark, square image representing the decoded image Y_decoded. It shows a very faint, blurry version of the original image." data-bbox="340 174 374 204"/>$$

while the original is

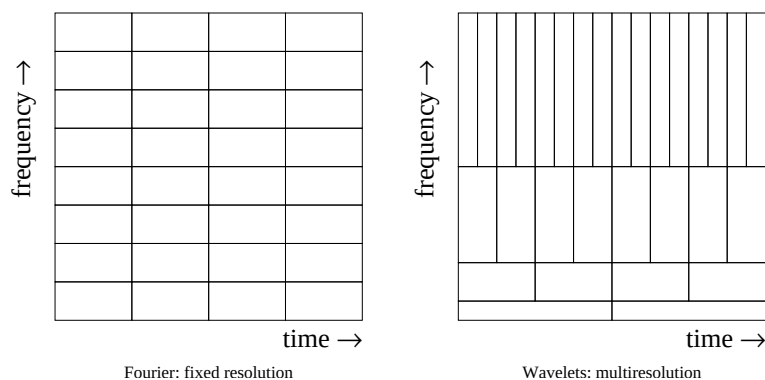
$$Y_{\text{original}} = \img alt="A small, dark, square image representing the original image Y_original. It shows a very faint, blurry version of the original image." data-bbox="336 236 374 267"/>$$

14 A very brief introduction to wavelets



This chapter deals with some minor snippets of wavelet theory (if these notes are ever going to be expanded—probably not—then this chapter should be expanded).

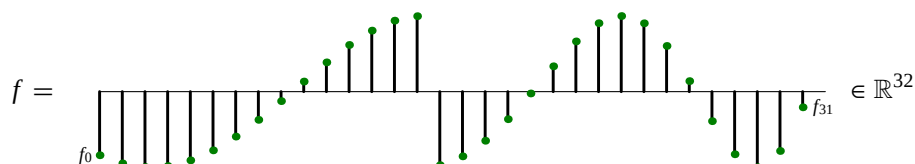
The example of jpeg clearly demonstrates that one should choose an “appropriate” basis for the problem at hand. Fourier (the basis for jpeg) successful as it may be has the serious drawback that its basis elements are not local in time (space). One would like to have a basis that is both local in time and local in frequency. Indeed this is what music scores do, isn’t it? Music scores specify *when* and how *long* a note should be played and at what *pitch* (frequency). Within the Fourier Kingdom one can enforce finite time by restricting the time/space interval over which the Fourier transform is taken (JPEG works per 8×8 block). An alternative is to use *wavelets* which allows for *multiresolution* analysis. Multiresolution is often depicted as follows



It expresses that we want good time resolution for high-frequency events, see Section 14.6 for a few more details.

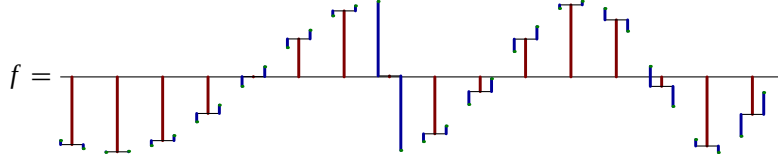
14.1 Multiresolution approximation and detail—Haar

We are going to consider finite discrete signals $f \in \mathbb{R}^N$ and we will represent them by a series of stems or lollipops such as



and the first entry is denoted f_0 (and the last f_{N-1}). The above signal is mostly smooth with a very sharp transition in the middle. (Chopped Fourier series f_* would not model f well because

of Gibbs.) We can separate local fluctuations and more long-term behavior by determining pair-wise average and deviation:



Here the red bars are the pair-wise averages

$$\tilde{f}^1 := \frac{1}{2} [f_0 + f_1 \quad f_2 + f_3 \quad \dots \quad f_{N-2} + f_{N-1}] \in \mathbb{R}^{N/2}. \quad (14.1)$$

and these form an *approximation* of f at double the scale and the blue line are defined as the deviation from the average

$$\tilde{d}^1 := \frac{1}{2} [f_0 - f_1 \quad f_2 - f_3 \quad \dots \quad f_{N-2} - f_{N-1}] \in \mathbb{R}^{N/2}. \quad (14.2)$$

They represent fluctuations or “detail” and indeed in our example the blue line are mostly small, except at the point where f fluctuates heavily.

Clearly the mapping from

$$f \in \mathbb{R}^N \rightarrow (\tilde{f}^1, \tilde{d}^1) \in (\mathbb{R}^{N/2}, \mathbb{R}^{N/2})$$

is invertible (we assume N even). We could continue with the approximation $\tilde{f}^1$ of f and decompose it in a similar manner

$$\tilde{f}^1 \rightarrow (\tilde{f}^2, \tilde{d}^2)$$

after which could decompose $\tilde{f}^2$,

$$\tilde{f}^2 \rightarrow (\tilde{f}^3, \tilde{d}^3)$$

etcetera. With each decomposition we get a coarser approximation (at double the scale = at half the number of samples) but since the mapping $\tilde{f}^k \rightarrow (\tilde{f}^{k+1}, \tilde{d}^{k+1})$ is invertible—assuming length of $\tilde{f}^k$ is even—we can for every k uncover f uniquely from

$$(\tilde{f}^k, \tilde{d}^k, \tilde{d}^{k-1}, \dots, \tilde{d}^1).$$

Example 14.1.1. A more dramatic example is shown in Fig. 14.1. The signal $f \in \mathbb{R}^{512}$ is partly smooth with a clear discontinuity somewhere in the middle. In the first half some noise appears and in the second half some variation in frequency (and noise) is visible. The noise in the first half only shows up in the finest layer of detail $\tilde{d}^1$. In $\tilde{d}^2$ it is already quite small and at coarser levels it is hardly visible. The discontinuity half way shows up in all levels of detail $\tilde{d}^k$. The fluctuations in frequency in the second half appears on all layers of detail. The figure in bottom half was reconstructed from $(\tilde{f}^4, \tilde{d}^3, \tilde{d}^2, \tilde{d}^1)$ after removal of 80% of the “smallest” detail coefficients. (In the following section we explain what “smallest” means.) The reconstructed $\hat{f}$ can handle discontinuities in f and noisy terms (Fourier cannot). A draw back is that it is piecewise constant (other wavelet transforms should do better in this respect.) $\square$

14.2 Orthogonal Haar basis

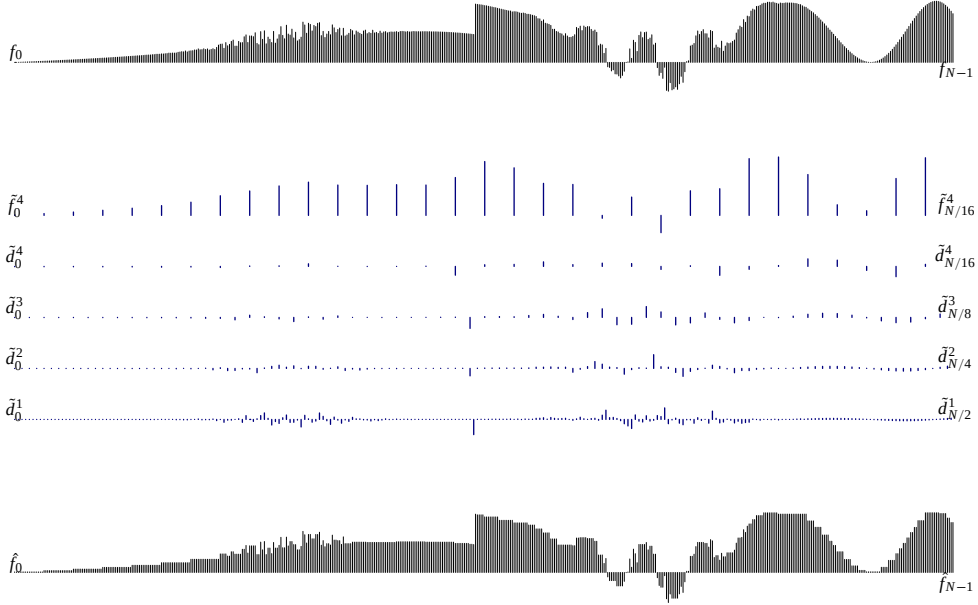


Figure 14.1: Top: Original $f \in \mathbb{R}^N$ with $N = 2^9 = 512$; Middle: coarse signal $\tilde{f}^4 \in \mathbb{R}^{N/16}$ and wavelet coefficients $\tilde{d}^4 \in \mathbb{R}^{N/16}$, $\tilde{d}^3 \in \mathbb{R}^{N/8}$, $\tilde{d}^2 \in \mathbb{R}^{N/4}$, $\tilde{d}^1 \in \mathbb{R}^{N/2}$. Bottom: reconstruction $\hat{f}$ of f obtained after setting 80% of the smallest wavelet coefficients $\tilde{d}_n^k$ to zero

The transformation from f to $\begin{bmatrix} \tilde{f}^1 & \tilde{d}^1 \end{bmatrix}$ described in the previous section is “almost” norm preserving:

$$\| \begin{bmatrix} f^1 & d^1 \end{bmatrix} \|^2 = \frac{1}{2} \|f\|^2.$$

This follows readily from (14.1) and (14.2). This is not ideal. If f would have been constant for instance then $d^1 = 0$ and so the approximation (which we feel is error free) has norm $\|f^1\| = \frac{1}{2}\sqrt{2}\|f\|$. For comparison of energy of the approximation f^1 and details d^1, d^2 it is beneficial to have a truly norm preserving transformation. This is why we redefine (and we loose the tildes now).

$$f^1 := \frac{1}{2}\sqrt{2} \begin{bmatrix} f_0 + f_1 & f_2 + f_3 & \cdots & f_{N-2} + f_{N-1} \end{bmatrix} \in \mathbb{R}^{N/2}. \quad (14.3)$$

and

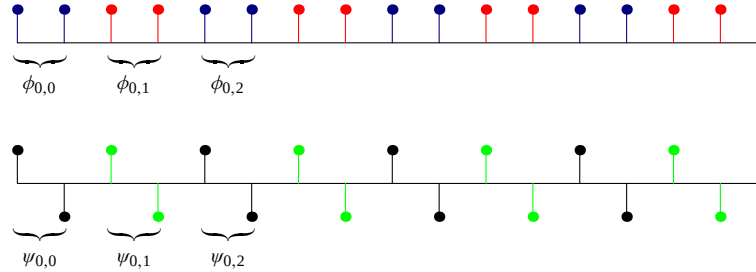
$$d^1 := \frac{1}{2}\sqrt{2} \begin{bmatrix} f_0 + d_1 & f_2 + f_3 & \cdots & f_{N-2} + f_{N-1} \end{bmatrix} \in \mathbb{R}^{N/2}. \quad (14.4)$$

Lemma 14.2.1 (Unitary transformation—Haar?). The transformation $\begin{bmatrix} \mathcal{H} \\ \mathcal{G} \end{bmatrix} : \mathbb{R}^{2M} \rightarrow \begin{bmatrix} \mathbb{R}^M \\ \mathbb{R}^M \end{bmatrix}$ that sends f to (f^1, d^1) is a *unitary* operator and we have by repeated application of this transformation that

$$\|f\| = \|(f^k, d^k, \dots, d^1)\|.$$

□

Viewing it from a different angle, the transformation from f to (f^1, d^1) can be seen as determining the expansion of f in the orthonormal basis (and now we hope you are not color blind) $\phi_{0,n}, \psi_{0,n}$:



Each basis element $\phi_{0,n}$ and $\psi_{0,n}$ has two neighboring nonzero entries, all its other entries being zero. Since the $\phi_{0,n}$ do no “overlap” and $\psi_{0,n}$ do not overlap it is clear that they are all mutually orthogonal. The function¹ $\phi_{0,n}$ is what one would call a *scaling function* and the $\psi_{0,n}$ are the *wavelets* and the numbers d_m^k the *wavelet coefficients*.

14.3 Down and up-sampling

The (f^1, d^1) can be easily computed using convolution with first order FIR filters with respective impulse responses $h = \{\frac{\sqrt{2}}{2}, \frac{\sqrt{2}}{2}\}$ and $g = \{\frac{\sqrt{2}}{2}, -\frac{\sqrt{2}}{2}\}$,

$$h * f, \quad g * f$$

followed by a factor 2 downsampling (subsampling?) which means removal of every second entry, see Fig. 14.2. The reconstruction of f from (f^1, d^1) can be computed by first upsampling the signals (f^1, d^1) by a factor 2,

$$f^1 \rightarrow [f_0^1 \ 0 \ f_1^1 \ 0 \ \dots], \quad d^1 \rightarrow [d_0^1 \ 0 \ d_1^1 \ 0 \ \dots],$$

and then passing them through filters with (in this case) the same FIR, and adding up the two outcomes, see Fig. 14.2.

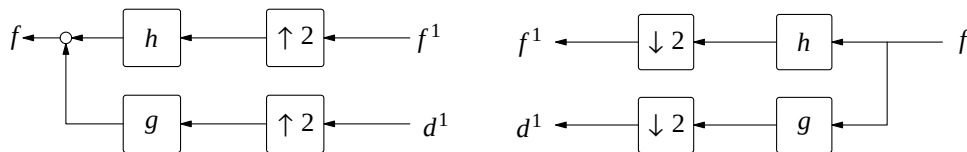


Figure 14.2: f^1 equals $h * f$ downsampled; d^1 equals $g * f$ downsampled

14.4 Haar basis (continuous time)

¹continuous time only??

A continuous-time version of the wavelets $\psi_{0,n}$ introduced at the end of Section 14.2 is

$$\psi(t) := \begin{cases} 1 & 0 \leq t < 1 \\ -1 & 1 \leq t < 2 \\ 0 & \text{otherwise} \end{cases} \quad (14.5)$$

By scaling and translating this wavelet

$$\psi_{j,n} := \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j}\right), \quad j, n \in \mathbb{Z}$$

we get a orthonormal sequence that in fact is a basis of $\mathbb{L}^2(\mathbb{R})$ [no proof], see Fig. 14.3.

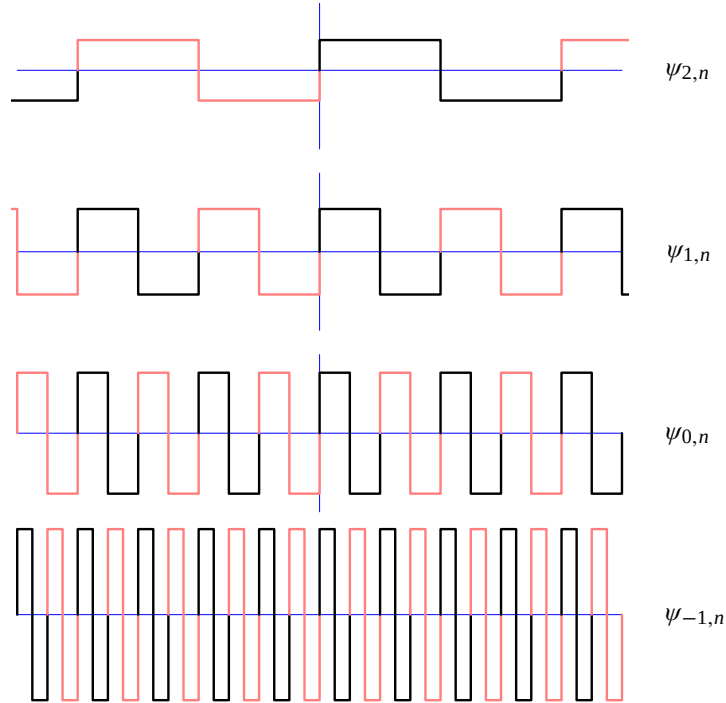


Figure 14.3: The Haar basis $\psi_{j,n}(t) := \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j}\right)$ for various $j, n \in \mathbb{Z}$

14.5 Generalizations

Are there *mother wavelets* ψ with the property that

- $\psi_{j,n} := \frac{1}{\sqrt{2^j}} \psi\left(\frac{t - 2^j n}{2^j}\right)$ with $j, n \in \mathbb{Z}$ spans the entire $\mathbb{L}^2(\mathbb{R})$
- $\psi_{j,n}$ are orthonormal
- $\psi_{j,n}$ have finite support

- have a user prescribed number M of vanishing moments, that is, $\langle t^m, \psi_{j,n} \rangle = 0$ for $m = 0, 1, \dots, M - 1$.

The answer to this is yes, and given the number of vanishing moments M , the one with smallest support is attributed to Daubechies..... For $M = 1$ the resulting mother wavelet is in fact (14.5) considered in the preceding section. Indeed $\langle t^0, \psi \rangle = 0$ for that ψ . Fig. 14.4 shows what happens if we redo Example 14.1.1 but now for Daubechies wavelets of order $M = 4$. The reconstructed $\hat{f}$ is much smoother (where needed).

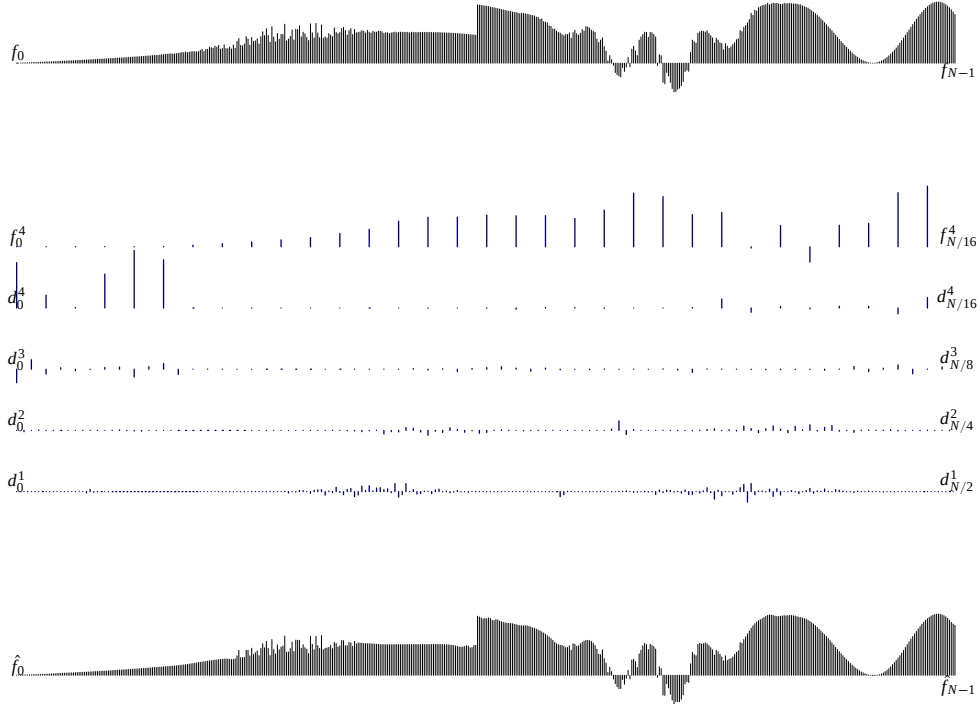


Figure 14.4: Top: Original $f \in \mathbb{R}^N$ with $N = 2^9 = 512$; Middle: coarse signal $f^4 \in \mathbb{R}^{N/16}$ and wavelet coefficients $d^4 \in \mathbb{R}^{N/16}$, $d^3 \in \mathbb{R}^{N/8}$, $d^2 \in \mathbb{R}^{N/4}$, $d^1 \in \mathbb{R}^{N/2}$. Bottom: reconstruction $\hat{f}$ of f obtained after setting 80% of the smallest wavelet coefficients d_n^k to zero

14.6 Some more snippets

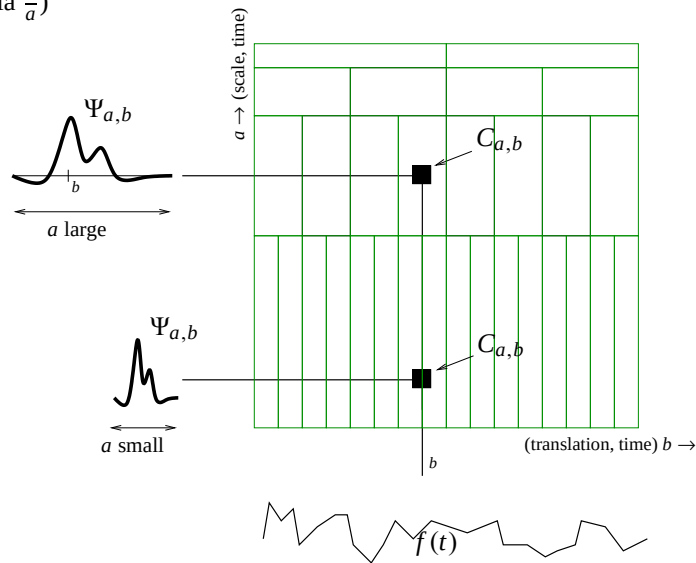
There is also a continuous time wavelet where one considers continuous scaling and translations ($a \in \mathbb{R}, b \in \mathbb{R}$)

$$\Psi_{a,b}(t) = \frac{1}{\sqrt{a}} \Psi\left(\frac{t-b}{a}\right)$$

of a wavelet Ψ . In principle then we compute at any $a, b \in \mathbb{R}$ the *continuous wavelet transform*

$$C_{a,b} := \langle f, \Psi_{a,b} \rangle := \int_{-\infty}^{\infty} f(t) \overline{\Psi_{a,b}(t)} dt$$

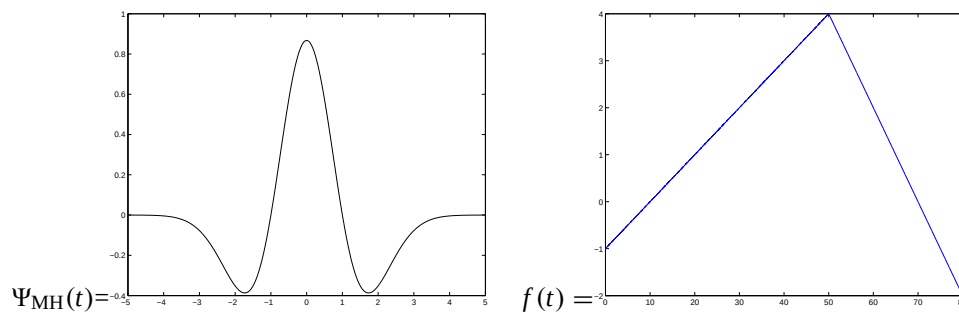
The time scale a takes over the role of frequency (frequencies are not explicitly accounted for, only implicitly via $\frac{1}{a}$)



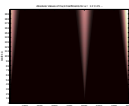
Example 14.6.1 (Mexican hat). Consider the “Mexican hat” wavelet

$$\Psi_{\text{MH}}(t) := \frac{2}{\sqrt{3}}\pi^{-\frac{1}{4}}(1-t^2)e^{-\frac{t^2}{2}}$$

and some piecewise linear $f(t)$:



The mexican hat has zero correlation with linear functions: $\langle t, \Psi_{\text{MH}} \rangle = 0$ so this wavelet should be good at pinpointing where the discontinuity in the derivative of f occurs. Here $C_{a,b}$:



□

Some comments:

- Choice of wavelet depends on what features to be extracted
 - Mexican-hat localized in time (and trend-insensitive)
 - Morlet is localized in frequency

- *Daubechies-N* have compact support; have zero correlation with polynomials of degree $N - 1$; are orthogonal.. ($N = 1, 2, 3, 4, \dots$)
- ...?
- Want to have short basis functions (for discontinuity detection). Want long ones (for detailed freq. analysis). This requires short high freq basis fncs and long low freq ones.
- Free Matlab toolbox: WVELAB. Commercial toolbox: WVELET
- Can deal with discontinuities and sharps spikes
- With wavelets we process the data at *different* time scales
 - Large windows (a large): coarse features
 - Small windows (a small): small, local features

A Matrices and Matlab

A.1 Matrices and vectors

In this appendix we'll review some standard notions of linear algebra.

A real *matrix* is an element of $\mathbb{R}^{n \times m}$ for certain $n, m \in \mathbb{N}$. Matrices are denoted by capital letters and their elements are by lower case letters with two subscript indices. The first index is the row index, the second the column index, for example

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdots & a_{1m} \\ a_{21} & a_{22} & a_{23} & \cdots & a_{2m} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & a_{n3} & \cdots & a_{nm} \end{bmatrix} \in \mathbb{C}^{n \times m}.$$

The (*complex conjugate*) *transpose* A' of a matrix A is defined as

$$A' = \begin{bmatrix} \bar{a}_{11} & \bar{a}_{21} & \cdots & \bar{a}_{m1} \\ \bar{a}_{12} & \bar{a}_{22} & \cdots & \bar{a}_{m2} \\ \bar{a}_{13} & \bar{a}_{23} & \cdots & \bar{a}_{m3} \\ \vdots & \vdots & \ddots & \vdots \\ \bar{a}_{1n} & \bar{a}_{2n} & \cdots & \bar{a}_{mn} \end{bmatrix} \in \mathbb{C}^{m \times n}.$$

It is convenient to think of the transpose A' as the result of reflecting A in its diagonal (and in case the entries are complex you should also take its complex conjugate). The k th column of a matrix A is denoted by A_{*k} and, similarly, A_{r*} means its r th row. A *column vector* is a matrix with only one column ($m = 1$) and a *row vector* is a matrix with only one row ($n = 1$). It is common to associate elements of $\mathbb{R}^n$ with *column* vectors. This is a choice not a necessity.

The *zero matrix* in whatever dimension $n \times m$ is usually denoted simply as 0 ; the square $n \times n$ *identity matrix* is denoted by I_n or simply by I and its k th column usually is written as e_k ,

$$0 = \begin{bmatrix} 0 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & 0 \end{bmatrix}, \quad I = \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 0 & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 \end{bmatrix}, \quad e_k = \begin{bmatrix} \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \end{bmatrix}.$$

We assume familiarity with the common matrix addition and matrix multiplication. One particular result that will return in various forms is this:

Lemma A.1.1. Let $A \in \mathbb{R}^{n \times m}$. The following statements are equivalent.

1. $Ax = 0$ has a unique solution ($x = 0$)
2. $A'A$ is a square nonsingular matrix

3. $|Ax| > 0$ for every $x \neq 0$
4. A has full column rank
5. The columns of A form a linear independent set of $\mathbb{R}^n$
6. $A\mathbb{R}^m$ has dimension m
7. All eigenvalues of $A'A$ are real and strictly positive
8. MATLAB: $[E, V] = \text{eig}(A'A); \text{all}(E > 0)$
9. MATLAB: $\det(A'A) > 0$
10. MATLAB: $[n, m] = \text{size}(A); m = \text{rank}(A)$
11. MATLAB: $[Q, R] = \text{qr}(A); (\text{prod}(\text{diag}(R)) \sim 0) \ \& \ (n \geq m)$

Proof. Consult whatever book on linear algebra. ■

A.2 Positive definite matrices

A real matrix P is *symmetric* if $P = P'$. A real or complex matrix P is *Hermitian* if $P = P'$. In particular symmetric and Hermitian matrices are square.

Definition A.2.1 (Positive definite and semi-positive definite). A Hermitian matrix $P = P' \in \mathbb{C}^{n \times n}$ is *positive definite*, notation $P > 0$, if

$$x'Px > 0 \quad \text{for all nonzero } x \in \mathbb{C}^n.$$

It is said to be *positive semi-definite* (also known as *nonnegative definite*), notation $P \geq 0$, if

$$x'Px \geq 0 \quad \text{for all } x \in \mathbb{C}^n.$$

□

It is an interesting fact that Hermitian matrices have real eigenvalues only. Indeed, if $Px = \lambda x$ with $x \neq 0$, then

$$\lambda x'x = x'(\lambda x) = x'(Px) = (Px)'x = (\lambda x)'x = \bar{\lambda} x'x$$

so that $\lambda = \bar{\lambda}$ i.e. that λ is real.

Lemma A.2.2 (Positive definite matrix). Let $P = P' \in \mathbb{C}^{n \times n}$. The following statements are equivalent.

1. $P > 0$.
2. $\det(P_{1:k, 1:k}) > 0$ for all $k \in \{1, 2, \dots, n\}$.
3. All eigenvalues of P are > 0 .
4. There is a nonsingular matrix T such that $P = T'T$.
5. P is nonsingular and $P = T'T$ for some (possibly nonsquare) T .

6. *Cholesky factorization*: there is a (unique) upper-triangular matrix T with diagonal entries > 0 such that $P = T'T$.
7. There is a $T = T' > 0$ such that $P = T^2$.
8. $V'PV > 0$ if V is square nonsingular.
9. There is a full column rank matrix T such that $P = T'T$
10. For any split of P ,

$$P = \begin{bmatrix} P_{11} & P_{12} \\ P'_{12} & P_{22} \end{bmatrix}$$

with P_{11} square (hence P_{22} square) we have that

$$P_{11} > 0 \quad \text{and} \quad P_{22} - P'_{12}P_{11}^{-1}P_{12} > 0.$$

(that is, both P_{11} and its *Schur complement* $P_{22} - P'_{12}P_{11}^{-1}P_{12}$ are positive definite)

11. $P_{22} > 0$ and $P_{11} - P_{12}P_{22}^{-1}P_{21} > 0$.

Proof. Consult any advanced book on linear algebra. ■

A.3 Trace

Before we can handle the MIMO case we need to document some easy facts about *trace*.

Definition A.3.1 (Trace). The trace $\text{tr}(C)$ of a square matrix $C \in \mathbb{C}^{n \times n}$ is defined as the sum of its diagonal entries, $\text{tr}(C) = \sum_{j=1}^n c_{jj}$. ■

Lemma A.3.2 (Trace). For any $A, B \in \mathbb{C}^{n \times m}$ the products AB' and $B'A$ are square and we have

$$\text{tr } AB' = \overline{\text{tr } BA'}.$$

Proof. $\text{tr } AB' = \sum_{j=1}^n (\sum_{k=1}^m a_{jk} \bar{b}_{kj}) = \sum_{k=1}^m (\sum_{j=1}^n a_{jk} \bar{b}_{kj}) = \overline{\text{tr } BA'}$. ■

In particular for real valued matrices A, B of compatible dimension we have $\text{tr } AB = \text{tr } BA$ which is interesting since the dimensions of the matrices AB and BA may differ.

A.4 Matlab

Basic numbers and matrices:

```
z=2+i;           % complex number
v=[3; 5];        % column vector with v(2)=5
w=[3 5];         % row vector with w(2)=5
A=[1 2; 4 5];    % 2*2 matrix with A(2,1)=4

x=v';           % (complex conjugate) transpose
B=A(:,1);       % first column of A
C=A(2,:);       % second row of A
```

```

A+[4 5; 8 9];      % matrix sum
A*x;               % matrix product
A^10;              % matrix power
[A A; w 3 3];      % a 3*4 matrix

I=eye(n);           % n*n identity matrix
A=zeros(3,5);       % 3*5 zero matrix
A=ones(3,5);        % 3*5 matrix with A(i,j)=1

[n,m]=size(A);      % number of rows and columns

```

Matrix equations and inverses:

```

b=[11; 10];        % column vector
x=A\b;              % A*x=b solved neatly
y=b'/A;             % y*A=b' solved neatly
inv(A);             % inverse matrix
A^(-1);             % inverse matrix
det(A);             % determinant
rank(A);            % rank (inherently difficult numerically)
V=null(A);          % then im(V)=ker(A)

```

Matrix norms (of which we only do the Euclidean norm, for matrices the Euclidean norm is known as the Frobenius norm):

```

norm(A,1);          % 1-norm
norm(A,"inf");       % max-norm
norm(A,"fro");       % Euclidean / Frobenius norm
sqrt(trace(A'A))     % Euclidean / Frobenius norm (same)
norm(A,2)           % largest singular value

```

Matrix decompositions:

```

[T,Lambda]=schur(A) % Schur decomposition, A=T*Lambda*inv(T)
[T,Lambda]=eig(A)   % Eigenvalue decomposition, A=T*Lambda*inv(T)
[Q,R]=qr(A)         % QR-factorization, A=QR

```

To check whether or not a Hermitian matrix is positive definite:

```

function y=ispos(A) % y=1 if A>0, else y=0
% assumes A=A'      % no error checking (terrible programming)
n=max(size(A));
y=A(1,1)>0;          % y=1 iff A(1,1)>0
if (n>1) & (y==1)
    C=A(2:n,2:n)-A(2:n,1)/A(1,1)*A(1,2:n);
    y=ispos(C);      % not very efficient programming
end

```

B Bibliography

Monson H. Hayes. *Statistical Digital Signal Processing and Modeling*. John Wiley & Sons, 1996.

David G. Luenberger. *Optimization by Vector Space Methods*. Wiley & Sons, 1969.

Stéphane Mallat. *A Wavelet Tour of Signal Processing*. Academic Press, 1998.

Karen Saxe. *Beginning Functional Analysis*. Undergraduate texts in mathematics. Springer-Verlag, New York, 2002.

Nicholas Young. *An Introduction to Hilbert Space*. Cambridge University Textbooks, 1988.

K. Zhou, J. C. Doyle, and K. Glover. *Robust and Optimal Control*. Prentice Hall, Englewood Cliffs, 1996. ISBN 0-13-456567-3.

B Bibliography

Index

- $C([a, b], \mathbb{R})$, 30
- QR-factorization, 11
- $\perp$
 - of sets, 6
- $\sharp$, 27
- σ , 73
- (co)inner-outer, 124
- addition
 - vector, 15
- additive inverse, 15
- adjoint, 10, 43
- affine variety, 7
- algebraic Riccati equation, 126
- aligned
 - vectors, 36
- antistable, 61
- ARE, 126
- associative
 - additive, 15
 - multiplication, 15
- autocorrelation, 73
- basis, 19
- Bessel's inequality, 47
- best approximation, 9
 - on inner product space, 39
- bilateral Laplace, 60
- bistable, 115
- Cauchy Schwarz, 36
- causal, 61
- Cholesky factorization, 161
- closed, 47
- closed under addition, 6
- closed under scaling, 7
- co-inner, 119
- codomain, 23
- column vector, 159
- commutative
 - additive, 15
- complete, 47
- complex vector space, 17
- conjugate, 63
- conjugate reflexive, 34
- continuous wavelet transform, 157
- continuous functions
 - symbol for, 30
- controllability Gramian, 102
- converge, 47
- convolution, 24
- correlation, 68
- correlation coefficient, 68
- covariance, 68
 - matrix, 78
- covariance function, 73
- covariance matrix, 69
- crosscovariance, 137
- DFT, 51
- dimension, 19
- distributive, 15
 - multiplication, 15
- dot product, 5
- DTFT, 58
- eigenvalue
 - of operator, 27
- Euclidean norm, 5
- expected value, 67
- factorization
 - QR, 12
- filter, 121
- finite dimensional, 19
- forecasting, 133
- Fourier coefficients, 56
- Fourier series, 24, 47
- Fourier transform, 24, 51
- Fredholm equation, 24
- Gram-matrix, 40
- Gram-Schmidt, 49
- Hamiltonian, 106
- Hardy space, 61

Index

- Hermitian, 160
- Hilbert matrix, 91
- Hilbert spaces, 47
- homogeneous, 34
- identity matrix, 159
- image, 23
 - matrix, 7
- independent
 - linear, 19
- infinite dimensional, 19
- inner, 111
- inner product, 33, 68
 - complex-, 34
- kernel
 - matrix, 7
- Laplace
 - bilateral, 60
- Legendre polynomials, 42
- Levinson-Durbin, 98
- LHP, 63
- linear, 23
- linear in first argument, 33
- linear predictive coding, 95
- linearly independent, 19
- linear combination, 19
- LPC, 95
- matched filter, 138
- matrix, 27, 159
 - as representation of mapping, 27
 - covariance-, 78
- mean, 67, 73
- mother wavelets, 155
- multiplication
 - scalar-vector, 15
- nonnegative definite, 160
- norm, 34
 - definition of, 48
 - Euclidean (on $\mathbb{R}$), 5
- normal equations, 10, 41, 44
 - stochastic, 71
- open, 47
- operator norm, 62
- origin, 7, 15
- orthogonal, 5, 69
 - complement, 6
 - matrix, 12
 - orthogonal complement, 6
 - orthogonal projection, 13
 - orthonormal sequence, 42
- Parseval, 42
- Parseval's identity, 47
- perpendicular, 5
 - sets, 6
- PFE, 123
- Plancherel's theorem, 59
- positive definite, 33, 34, 160
- positive semi-definite, 160
- power, 57, 77
- power spectral density, 76
- prediction, 133
- probability density function, 67
- probability distribution, 67
- projection, 13
- QR-factorization, 12
- range, 23
- real vector space, 15
- reflection coefficients, 98
- reflexive, 33
- region of convergence (ROC), 60
- return difference inequality, 114
- RHP, 63
- Riccati equation, 126
 - stabilizing solution, 126
- RMS, 89
- root mean square, 89, 90
- rotation, 26
- row vector, 159
- scalars, 17
- scaling function, 154
- Schur complement, 161
- second moment, 67
- span, 19
- spectral
 - density, 76
- spectral factor
 - rational, 115
- spectral factorization, 114
- stability
 - AR system, 96
- stabilizing, 117
 - solution of Riccati eqn., 126
- stable, 61, 62

- $\mathbb{L}^2$, 62
- anti-, 61
- standard biased estimate, 137
- standard deviation, 73
- standard $\mathbb{H}^2$ problem, 125
- state transition matrix, 101
- std, 73
- stochastic process, 72
- subspace, 6, 18
- symmetric, 160

- Toeplitz, 97
- trace, 161
- transformation
 - Fourier, 24
- transpose, 9
 - complex-conjugate, 159
- triangle inequality, 34

- uncorrelated, 69
- unitary, 58, 59, 153

- Vandermonde matrix, 89
- variance, 67
- vectors, 15
 - complex vector space, 17
- vector space
 - complex, 17
 - real, 15

- wavelets, 154
- wavelet coefficients, 154
- whitening filter, 81
- white noise, 75, 77
- wide sense stationary (WSS), 73
- Wiener-Hopf, 130
- Wiener filtering, 129

- zero vector, 15