Math minimalia

Alessandro Verri

Abstract

Make sure you have pen and paper while you go through these notes and do try to solve each and every exercise on your own. If you only need to refresh your knowledge, what you find here is probably going to be sufficient. If you experience difficulties, instead, you may want to go back to some manual of basic math or attend one of the many courses you find online.

Contents

1 Linear spaces

Linear spaces are possibly the most important ingredient for both machine learning and digital signal processing. A theoretical grasp of the concepts in this section is mandatory.

Mathematical structure: definition and closure property of a linear space

Linear dependency and independency: linear span, subspaces, dimension, and bases

Euclidean structure: scalar product and its properties

$*\quad*\quad*\quad*\quad*$

Theoretical minimum

Let V be a set of elements, often called *vectors*, and $\mathbb{F}$ a field (like $\mathbb{R}$ or $\mathbb{C}$). We define two operations. The first, called *vector sum*, is a map from $V \times V \to V$ which sends a pair of vectors $\mathbf{u}$ and $\mathbf{v} \in V$ into another vector $\mathbf{w} \in V$, or

$$\mathbf{w} = \mathbf{u} + \mathbf{v}$$

The second, called *scalar multiplication*, is a map from $\mathbb{F} \times V \to V$ which sends a number $\alpha \in \mathbb{F}$ and a vector $\mathbf{v} \in V$ into another vector $\mathbf{w} \in V$, or

$$\mathbf{w} = \alpha\mathbf{v}$$

In what follows we assume that the field $\mathbb{F}$ is $\mathbb{R}$.
The set V is a *linear space* iff

1. For all $\mathbf{u}$ and $\mathbf{v} \in V$, $\mathbf{u} + \mathbf{v} = \mathbf{v} + \mathbf{u}$;
2. For all $\mathbf{u}, \mathbf{v}$ and $\mathbf{w} \in V$, $\mathbf{u} + (\mathbf{v} + \mathbf{w}) = (\mathbf{u} + \mathbf{v}) + \mathbf{u}$;
3. There exists a vector $\mathbf{0} \in V$, the *zero* vector, such that for all $\mathbf{v} \in V$; $\mathbf{v} + \mathbf{0} = \mathbf{v}$
4. For all $\mathbf{v} \in V$ there exists an *inverse* vector $-\mathbf{v}$ such that $\mathbf{v} + (-\mathbf{v}) = \mathbf{0}$;
5. For all α and $\beta \in \mathbb{R}$ and $\mathbf{v} \in V$, $\alpha(\beta\mathbf{v}) = (\alpha\beta)\mathbf{v}$;
6. For all $\mathbf{v} \in V$, $1\mathbf{v} = \mathbf{v}$;
7. For all $\alpha \in \mathbb{R}$ and $\mathbf{u}$ and $\mathbf{v} \in V$, $\alpha(\mathbf{u} + \mathbf{v}) = \alpha\mathbf{u} + \alpha\mathbf{v}$;
8. For all α and $\beta \in \mathbb{R}$ and $\mathbf{v} \in V$, $(\alpha + \beta)\mathbf{v} = \alpha\mathbf{v} + \beta\mathbf{v}$.

Observation 1.1. *One necessary element*
If $\mathbf{0} \notin V$, the set V cannot be a linear space!

Observation 1.2. *Vectors or not?*
Keep in mind that picturing vectors of a linear space as arrows can occasionally be misleading!

Definition 1.1. *Linear independence*
Let V be a linear space. The vectors $\mathbf{v}_1, \ldots \mathbf{v}_k \in V$ are *linearly independent* if

$$\alpha_1\mathbf{v_1} + \ldots \alpha_k\mathbf{v}_k = \mathbf{0} \quad \text{iff} \quad \alpha_1 = \cdots = \alpha_k = \mathbf{0},$$

and linearly dependent otherwise.

Definition 1.2. *Subspace*
A subset W of a linear space V over $\mathbb{R}$ is a subspace of V if W is a linear space over $\mathbb{R}$.

Observation 1.3. *Once again, the* $\mathbf{0}$ *is in*
Like for all good linear spaces, the $\mathbf{0}$ vector must belong to any subspace!

Definition 1.3. *Linear span*
The set W of all the linear combinations of $\mathbf{v}_1, \ldots, \mathbf{v}_k \in V$

$$W = \mathit{Span}\{\mathbf{v_1}, \ldots, \mathbf{v_k}\} = \{t_1\mathbf{v}_1 + \cdots + t_k\mathbf{v}_k : t_1, \ldots, t_k \in \mathbb{R}\}$$

is called *linear span* generated by $\mathbf{v}_1, \ldots, \mathbf{v}_k \in V$.

Fact 1.1. *A linear span is a subspace*
For all possible choices of $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$ the linear span W is a subspace of V.

Fact 1.2. *One too many*
If

$$W = Span\{\mathbf{v_1}, \dots, \mathbf{v_k}\} = \{t_1\mathbf{v}_1 + \cdots + t_k\mathbf{v}_k : t_1, \dots, t_k \in \mathbb{R}\}$$

then any set of $k+1$ vectors in W cannot be linearly independent.

Observation 1.4. *Not a word about linear independency*
We have not said anything about the linear independence of $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$. Intuitively the linear span is going to be as rich as the $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$ are diverse. If the $\mathbf{v}_1, \dots, \mathbf{v}_k$ are actually multiple one of the other, for example, the linear span W consists of only one vector, say $\mathbf{v}_1$. Consequently, if we take $k+1$ vectors in W, all of them will be multiple of $\mathbf{w}_1$.

Fact 1.3. *Dimension of a subspace and basis*
The linear span of k linearly independent vectors has a crucial minimal property. If the k vectors spanning W are linearly independent, a proper subset of the k vectors cannot span W. In this case we say that the *dimension* of W equals k (or $dim\ W = k$) and that the set $\mathbf{v}_1, \dots, \mathbf{v}_k \in V$ is a *basis* for W.

Fact 1.4. *Unique expansion for any fixed basis*
If $\mathbf{v}_1, \dots, \mathbf{v}_k$ is a basis for W, any $\mathbf{v}$ of W can always be written as

$$\mathbf{w} = t_1\mathbf{v}_1 + \cdots + t_k\mathbf{v}_k$$

for exactly one choice of the real coefficients $t_1, \dots, t_k \in \mathbb{R}$.

Definition 1.4. *Scalar product*
A *scalar product* is a map $\langle\cdot,\cdot\rangle$: $V \times V \to \mathbb{R}$ such that:

1. $\forall$ $\mathbf{u}$ and $\mathbf{v} \in V$ $\langle\mathbf{u},\mathbf{v}\rangle = \langle\mathbf{v},\mathbf{u}\rangle$;
2. $\forall$ $\mathbf{u} \in V$ $\langle\mathbf{u},\mathbf{u}\rangle \geq 0$ with $\langle\mathbf{u},\mathbf{u}\rangle = 0$ iff $\mathbf{u} = \mathbf{0}$;
3. $\forall$ $\mathbf{u}, \mathbf{v}$ and $\mathbf{w} \in V$ and $\forall\alpha$, $\beta \in \mathbb{R}$, $\langle\alpha\mathbf{u} + \beta\mathbf{v}, \mathbf{w}\rangle = \alpha\langle\mathbf{u},\mathbf{w}\rangle + \beta\langle\mathbf{v},\mathbf{w}\rangle$.

Observation 1.5. *Bilinearity*
Combining symmetry with linearity in the first argument one can easily see that the scalar product is also linear in the second argument.

Observation 1.6. *Replacing $\mathbb{R}$ with $\mathbb{C}$*
If $\mathbb{C}$ is used instead, the definition must be modified by introducing the complex conjugate as follows

1. $\forall$ $\mathbf{u}$ and $\mathbf{v} \in V$ $\langle\mathbf{u},\mathbf{v}\rangle = \langle\mathbf{v},\mathbf{u}\rangle^*$;
2. $\forall$ $\mathbf{u} \in V$ $\langle\mathbf{u},\mathbf{u}\rangle \geq 0$ with $\langle\mathbf{u},\mathbf{u}\rangle = 0$ iff $\mathbf{u} = \mathbf{0}$;
3. $\forall$ $\mathbf{u}, \mathbf{v}$ and $\mathbf{w} \in V$ and $\forall\alpha$, $\beta \in \mathbb{C}$, $\langle\alpha\mathbf{u} + \beta\mathbf{v}, \mathbf{w}\rangle = \alpha\langle\mathbf{u},\mathbf{w}\rangle + \beta\langle\mathbf{v},\mathbf{w}\rangle$.

Observation 1.7. *Norm and distance induced by the scalar product*

- $\|\mathbf{v}\| = \sqrt{\langle\mathbf{v},\mathbf{v}\rangle}$ is the *norm* of $\mathbf{v}$ (often called *length*)
- $\|\mathbf{u} - \mathbf{v}\|$ is the distance between $\mathbf{u}$ and $\mathbf{v}$, with $\|\mathbf{u} - \mathbf{v}\| = \mathbf{0}$ iff $\mathbf{u} = \mathbf{v}$.

Fact 1.5. *Three important properties*

Homogeneity $\forall\alpha \in \mathbb{R}$ and $\mathbf{v} \in V$ $\|\alpha\mathbf{v}\| = |\alpha|\|\mathbf{v}\|$.

Triangular inequality $\forall\mathbf{u}$ and $\mathbf{v} \in V$ $\|\mathbf{u} + \mathbf{v}\| \leq \|\mathbf{u}\| + \|\mathbf{v}\|$.

Cauchy-Schwartz inequality $\forall\mathbf{u}$ and $\mathbf{v} \in V$ $|\langle\mathbf{u},\mathbf{v}\rangle| \leq \|\mathbf{u}\|\|\mathbf{v}\|$. □

Through the scalar product we can measure the *angle* between two non zero vectors and obtain several useful properties.

Fact 1.6. *Angle and orthogonality*
Through the Cauchy-Schwartz inequality, the angle θ between two non zero vectors $\mathbf{u}$ and $\mathbf{v} \in V$ can be defined as

$$\theta = \arccos\left(\frac{\langle \mathbf{u}, \mathbf{v} \rangle}{\|\mathbf{u}\|\|\mathbf{v}\|}\right).$$

If $\langle \mathbf{u}, \mathbf{v} \rangle = 0$, $\theta = \pi/2$ then $\mathbf{u}$ and $\mathbf{v}$ are *orthogonal*. Interestingly, the angle can be defined independently of the dimension of V.

Theorem 1.1. *Pythagora*
If $\mathbf{u}$ and $\mathbf{v}$ are orthogonal, then

$$\|\mathbf{u} + \mathbf{v}\|^2 = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$$

Proof

$$\|\mathbf{u}+\mathbf{v}\|^2 = \langle \mathbf{u}+\mathbf{v}, \mathbf{u}+\mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + \langle \mathbf{u}, \mathbf{v} \rangle + \langle \mathbf{v}, \mathbf{u} \rangle + \langle \mathbf{v}, \mathbf{v} \rangle = \langle \mathbf{u}, \mathbf{u} \rangle + 0 + 0 + \langle \mathbf{v}, \mathbf{v} \rangle = \|\mathbf{u}\|^2 + \|\mathbf{v}\|^2$$

Notice that the same results hold true for an arbitrary number of mutually orthogonal vectors. We will see that Pythagora's theorem is true even for right triangles with infinite legs!
□

In a Euclidean linear space we can define *orthonormal* bases, or bases which consist of *mutually orthogonal vectors of unit length*. The coefficients describing $\mathbf{v}$ in an orthonormal basis can be simply computed by taking the scalar product between $\mathbf{v}$ and each basis vector, operation which corresponds to the orthogonal projection of $\mathbf{v}$ on each basis vector. Let us obtain this result explicitly.

Observation 1.8. *A remarkable property*
Assume we are given an orthonormal basis for a space V with $dim(V) = n$. We know we can write any $\mathbf{v} \in V$ as a linear combination of the n basis vectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$ each multiplied by an appropriate coefficient a_j with $j = 1, \ldots, n$ or

$$\mathbf{v} = a_1\mathbf{v}_1 + \cdots + a_n\mathbf{v}_n.$$

The value of each coefficient can actually be computed very easily.
Taking the scalar product on both sides with $\mathbf{v}_j$, through the orthonormality of the basis vectors we obtain for all j

$$\langle \mathbf{v}, \mathbf{v}_j \rangle = a_j.$$

Fact 1.7. *Scalar product in components*
If $\mathbf{u}$ and $\mathbf{v}$ are expressed through an orthonormal basis as $\mathbf{u} = (u_1 \ \ldots \ u_n)^\top$ and $\mathbf{v} = (v_1 \ \ldots \ v_n)^\top$, we have that

$$\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top\mathbf{v} = \mathbf{v}^\top\mathbf{u} = u_1v_1 + \ldots u_nv_n$$

The norm of $\mathbf{u}$ can be written

$$\|\mathbf{u}\|^2 = \mathbf{u}^\top\mathbf{u} = u_1^2 + \ldots u_n^2,$$

a further example of Pythagora's theorem for a right triangle with n legs!

Exercises

Exercise 1.1. $\mathbb{R}^3$
Verify that the set $\mathbb{R}^3$ which consists of all the ordered triplets of real numbers

$$\mathbf{v} = (v_1 \ \ v_2 \ \ v_3)^\top$$

is a linear space with

1. the *vector sum* defined as $\mathbf{u} + \mathbf{v} = (u_1 + v_1 \ \ u_2 + v_2 \ \ u_3 + v_3)^\top$
2. the *scalar multiplication* as $\alpha\mathbf{v} = (\alpha v_1 \ \ \alpha v_2 \ \ \alpha v_3)^\top$.

Exercise 1.2. *The set of polynomials of degree n with real coefficients*
Verify that $\mathbb{P}_n$, the set of polynomials of degree n with real coefficients, is a linear space if for every $\mathbf{u}$ and $\mathbf{v} \in \mathbb{P}_n$ and $\alpha \in \mathbb{R}$

1. the *vector sum* defined as $\mathbf{u}+\mathbf{v}=(u_0+v_0)+(u_1+v_1)x+\ldots(u_n+v_n)x^n$
2. the *scalar multiplication* as $\alpha\mathbf{v}=\alpha(v_0+v_1x+\ldots v_nx^n)$.

Exercise 1.3. *Linear dependency*
Show that $\mathbf{u}=(1\ \ 0\ \ 0)^\top$, $\mathbf{v}=(0\ \ 1\ \ 0)^\top$, $\mathbf{w}=(0\ \ 0\ \ 1)^\top$, and $\mathbf{z}=(1\ \ 1\ \ 1)^\top\in\mathbb{R}^3$ are pairwise linearly independent but linearly dependent as a set.

Exercise 1.4. *Linearly dependent and independent polynomials of degree 1*
Which pair of the polynomials $\mathbf{u}=1+x$, $\mathbf{v}=2+x$, and $\mathbf{w}=4+2x$, if any, is linearly independent?

Exercise 1.5. *A fake example*
Show that the set W of all the triplets of the form $(1,t,t)$ with $t\in\mathbb{R}$ is not a subspace of $\mathbb{R}^3$.

Exercise 1.6. *A subspace of* $\mathbb{R}^3$
Show that the set W of all the triplets of the form $(t\ \ -2t\ \ t)^\top$ with $t\in\mathbb{R}$ is a subspace of $\mathbb{R}^3$. Then show that W is a *proper* subspace of $\mathbb{R}^3$ by finding a triplet of $\mathbb{R}^3$ not in W.

Exercise 1.7. *Another subspace of* $\mathbb{R}^3$
Show that the set W of all the triplets of the form $(s,t,0)$ with $s,t\in\mathbb{R}$ is a subspace of $\mathbb{R}^3$. Then show that W is a proper subspace of $\mathbb{R}^3$ by finding a triplet of $\mathbb{R}^3$ not in W.

Exercise 1.8. *A subspace of the linear space of polynomials of degree 2*
Show that the set W of all the polynomials of degree 2 of the form $\mathbf{v}=v_0+v_2x^2$ is a subspace of the linear space of all polynomials of degree 2. Then show that W is a proper subspace of by finding a polynomial of degree 2 not in W.

Exercise 1.9. *Dimension of* $\mathbb{R}^3$ *and possible bases*
Clearly, the triplets $\mathbf{e}_1=(1\ \ 0\ \ 0)^\top$, $\mathbf{e}_2=(0\ \ 1\ \ 0)^\top$, and $\mathbf{e}_3=(0\ \ 0\ \ 1)^\top$ span $\mathbb{R}^3$. The dimension of $\mathbb{R}^3$, therefore, is 3 and $\mathbf{e}_1,\mathbf{e}_2$, and $\mathbf{e}_3$ form a basis for $\mathbb{R}^3$. Furthermore, for any triplet $\mathbf{w}=(a\ \ b\ \ c)^\top\in\mathbb{R}^3$ we have

$$\mathbf{w}=a\,\mathbf{e}_1+b\,\mathbf{e}_2+c\,\mathbf{e}_3.$$

Therefore, the unique choice of the coefficients are $t_1=a$, $t_2=b$, and $t_3=c$.
Show that also the triplets $\mathbf{f}_1=(1\ \ 1\ \ 0)^\top$, $\mathbf{f}_2=(1\ \ 0\ \ 1)^\top$, and $\mathbf{f}_3=(0\ \ 1\ \ 1)^\top$ span $\mathbb{R}^3$ and form a basis.

Exercise 1.10. *Always the same vector*
Verify that the triplet

$$(a\ \ b\ \ c)^\top$$

in the $\mathbf{f}_1,\mathbf{f}_2$, and $\mathbf{f}_3$ basis and the triplet

$$\left(\frac{a+b-c}{2}\ \ \frac{a+c-b}{2}\ \ \frac{b+c-a}{2}\right)^\top$$

in the standard basis correspond to the same vector $\mathbf{v}\in V$.

Exercise 1.11. *Dimension of the linear space of polynomials of degree 2 and possible bases*
Clearly, the polynomials $\mathbf{e}_1=1$, $\mathbf{e}_2=x$, and $\mathbf{e}_3=x^2$ span the set. Therefore, the dimension is 3 and $\mathbf{e}_1,\mathbf{e}_2$, and $\mathbf{e}_3$ form a basis. Show that also the polynomials $\mathbf{f}_1=1+x$, $\mathbf{f}_2=1+x^2$, and $\mathbf{f}_3=x+x^2$ span the set and form a basis.

2 Linear maps

Not even a superficial understanding of machine learning and signal processing is possible unless you are at ease with linear maps, or transformations, between linear spaces. Here is what you should know.

Mathematical structure: definition, range and kernel of a linear map

Main properties: composition and matrix representation of linear maps

$*\quad*\quad*\quad*\quad*$

Theoretical minimum

Everything follows from the linearity assumption.

Definition 2.1. *Linear map (or transformation)*
Let V and W be two linear spaces over the field $\mathbb{R}$. A map $\mathcal{L} : V \to W$ is *linear* iff

$$\forall \mathbf{u}, \mathbf{v} \in V \quad \text{and} \quad \forall c \in \mathbb{R} \quad \mathcal{L}(\mathbf{u}+\mathbf{v}) = \mathcal{L}(\mathbf{u}) + \mathcal{L}(\mathbf{v}) \quad \text{and} \quad \mathcal{L}(c\,\mathbf{u}) = c\,\mathcal{L}(\mathbf{u})$$

Definition 2.2. *Range and kernel*
The **range of the linear map** $\mathcal{L} : V \to W$, $\mathcal{L}(V)$, is the set of the images of all $v \in V$. Clearly $\mathcal{L}(V) \subseteq W$. The **kernel or null space of** $\mathcal{L}$, $N(\mathcal{L})$, is the set of all $v \in V$ mapped by $\mathcal{L}$ to $\mathbf{0} \in W$. Clearly, $N(\mathcal{L}) \subseteq V$.

Fact 2.1. *Once linear, always linear*
If V, W, and Z are linear spaces, and $\mathcal{L} : V \to W$ and $\mathcal{M} : W \to Z$ linear maps, the map

$$\mathcal{P} = \mathcal{M} \circ \mathcal{L} : V \to Z$$

is well-defined and linear. Indeed, forall $\mathbf{u}$ and $\mathbf{v} \in V$ and $c \in \mathbb{R}$ we have

$$\mathcal{P}(\mathbf{u}+\mathbf{v}) = \mathcal{M} \circ \mathcal{L}(\mathbf{u}+\mathbf{v}) = \mathcal{M}(\mathcal{L}(\mathbf{u}) + \mathcal{L}(\mathbf{v})) = \mathcal{M} \circ \mathcal{L}(\mathbf{u}) + \mathcal{M} \circ \mathcal{L}(\mathbf{v}) = \mathcal{P}\mathbf{u} + \mathcal{P}\mathbf{v}$$

$$\mathcal{P}(c\,\mathbf{u}) = \mathcal{M} \circ \mathcal{L}(c\,\mathbf{u}) = \mathcal{M}(c\,\mathcal{L}(\mathbf{u})) = c\,\mathcal{M} \circ \mathcal{L}(\mathbf{u}) = c\,\mathcal{P}(\mathbf{u})$$

Fact 2.2. *One-to-one linear maps*
A linear map $\mathcal{L} : V \to W$ for which $N(\mathcal{L}) = \mathbf{0}$ is injective. A linear map $\mathcal{L} : V \to W$ for which $\mathcal{L}(V) = W$ is surjective. If a linear map $\mathcal{L} : V \to W$ is both injective and surjective there exists a unique linear map $\mathcal{L}^{-1} : W \to V$ such that for all $v \in V$ $\mathcal{L}^{-1} \circ \mathcal{L}v = v$ and for all $w \in W$ $\mathcal{L} \circ \mathcal{L}^{-1}w = w$.

Exercises

Exercise 2.1. *A few examples*
Verify that the *identity map*, the *null map*, the *multiplication by a fixed constant*, a *set of m linear equations in n unknowns*, and the *scalar product with a fixed vector* (the last between Euclidean linear spaces) are all examples of linear maps.

Exercise 2.2. *Range and kernel*
Prove that $\mathcal{L}(V)$ is a subspace of W and that $N(\mathcal{L})$ is a subspace of V.

Exercise 2.3. *More on kernel*
Define input and output space and determine the kernel for the *identity map*, the *null map*, the *multiplication by a fixed constant*, a *set of linear equations*, and the *scalar product with a fixed vector*.

Exercise 2.4. *Rank-nullity theorem*
Verify that, in the examples of the previous exercise, if $\mathcal{L} : V \to W$ with V n-dimensional and W m-dimensional that

$$dim\ \mathcal{L}(V) + dim\ N(\mathcal{L}) = n$$

Exercise 2.5. *Matrix representation of a linear transformation*
Pick the standard basis for $\mathbb{R}^2$, $\mathbf{e}_1 = (1\ \ 0)^\top$ and $\mathbf{e}_2 = (0\ \ 1)^\top$ and the linear transformation $\mathcal{L} : \mathbb{R}^2 \to \mathbb{R}^2$ for which

$$\mathbf{u}_1 = \begin{pmatrix} 3 \\ 4 \end{pmatrix} \ \rightarrow\ \mathcal{L}(\mathbf{u}_1) = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \quad \text{and} \quad \mathbf{u}_2 = \begin{pmatrix} 1 \\ 1 \end{pmatrix} \ \rightarrow\ \mathcal{L}(\mathbf{u}_2) = \begin{pmatrix} -1 \\ 1 \end{pmatrix}$$

Show that the action of $\mathcal{L}$ on $\mathbf{v}$ can be represented by the *row-column* product between the matrix

$$L = \begin{pmatrix} -5 & 4 \\ 3 & -2 \end{pmatrix}$$

where the columns are the ordered images of the input basis vectors, and the components of $\mathbf{v}$ in the input basis.

Exercise 2.6. *Matrix representations are not unique!*
Show that **if the input is written in the $(\mathbf{u}_1, \mathbf{u}_2)$ basis and the output in the standard basis**, the action of the same linear transformation of the previous exercise can be expressed in terms of the matrix

$$L' = \begin{pmatrix} 1 & -1 \\ 1 & 1 \end{pmatrix}$$

the columns of which are the ordered images of the input basis vectors.

Exercise 2.7. *Don't be fooled by the identity matrix*
Show that if you use $\mathbf{u}_1$ and $\mathbf{u}_2$ as input basis and their images $\mathcal{L}(\mathbf{u}_1) = (1\ \ 1)^\top$ and $\mathcal{L}(\mathbf{u}_2) = (-1\ \ 1)^\top$ as output basis, the matrix representing the linear map $\mathcal{L}$ is the identity matrix I.

3 Linear algebra

Far from doing justice to linear algebra, we restrict immediately our attention to what is absolutely needed for machine learning and digital signal processing. Our emphasis is on computation. Here is the list of thinks you should know.

Matrix representation: row-column product, matrix rank and kernel

Orthogonal transformations: basic manipulations

Eigenvalues and eigenvectors: the case of symmetric matrices

Positive (semi)definite matrices: definition and properties

Singular Value Decomposition: effective rank and kernel of a matrix

$*\quad *\quad *\quad *\quad *$

Theoretical minimum

In what follows we tacitly assume that a linear map $\mathcal{L} : V \to W$ is represented by a matrix obtained by using standard bases in both V and W.

Fact 3.1. *Composing linear maps through matrix multiplication (and a simple sanity check)*
The linearity of the composition of two linear maps can be easily seen in terms of matrix representation. If the linear maps $\mathcal{L} : \mathbb{R}^n \to \mathbb{R}^m$ and $\mathcal{M} : \mathbb{R}^m \to \mathbb{R}^p$ are represented in the standard bases by the $m \times n$ matrix L and by the $p \times m$ matrix M respectively, the linear map $\mathcal{P} = \mathcal{L} \circ \mathcal{M} : \mathbb{R}^n \to \mathbb{R}^p$ is represented by the $p \times n$ matrix

$$P = ML = \begin{pmatrix} M_{11} & \dots & M_{1m} \\ \dots & \dots & \dots \\ M_{p1} & \dots & M_{pm} \end{pmatrix} \begin{pmatrix} L_{11} & \dots & L_{1n} \\ \dots & \dots & \dots \\ L_{m1} & \dots & L_{mn} \end{pmatrix} = \begin{pmatrix} P_{11} & \dots & P_{1n} \\ \dots & \dots & \dots \\ P_{p1} & \dots & P_{pn} \end{pmatrix}$$

obtained through the usual row-column product. Establishing whether, or not, the composition of two linear maps is well defined boils down to a simple check: since the dimensionality of the input of the external map must be equal to the dimensionality of the output of the internal map, **the number of columns of the external matrix must be equal to the number of rows of the internal matrix**. In addition, **the product matrix has the same number of rows of the external matrix and the same number of column of the internal matrix**.

Definition 3.1. *Linear transformation of a kind*
If V and W are Euclidean linear spaces, an *orthogonal map* $\mathcal{O}$ is a linear map which preserves the scalar product, that is,

$$\forall \mathbf{u}, \mathbf{v} \in V \quad \langle \mathbf{u}, \mathbf{v} \rangle = \langle \mathcal{O}(\mathbf{u}), \mathcal{O}(\mathbf{v}) \rangle$$

Thus, an orthogonal transformation preserves lengths and corresponds to a *rotation* (possibly coupled with a mirroring operation).

Fact 3.2. *A very useful decomposition*
A symmetric $n \times n$ matrix M can always be rewritten as

$$M = Q \Lambda Q^\top$$

where Q is an $n \times n$ orthogonal matrix and Λ an $n \times n$ diagonal matrix. The elements on the diagonal of Λ are the eigenvalues of M (each present as many times as its degeneracy) and Q the orthogonal transformation (rotation of the n axes) needed to render M.

Fact 3.3. *Characteristic equation*
The eigenvalues of an $n \times n$ matrix A can be found as the solutions to the equation of n-th degree for λ

$$Det(\lambda I - A) = 0$$

If A is symmetric, the n solutions - each counted as many times as its degeneracy - are real numbers.

Fact 3.4. *Eigenvalues and eigenvectors of a symmetric matrix*
For a symmetric $n \times n$ matrix M, λ is an eigenvalue and $\mathbf{x}^* \in \mathbb{R}^n$ a corresponding eigenvector if

$$M\mathbf{x}^* = \lambda\mathbf{x}^*$$

Since

$$\langle M\mathbf{x}^*, M\mathbf{x}^*\rangle = \lambda^2\langle \mathbf{x}^*, \mathbf{x}^*\rangle = \lambda^2\|\mathbf{x}^*\|^2 \geq 0$$

we have that **all the eigenvalues of a symmetric matrix are real**. Furthermore, we have that for any pair $\mathbf{x}^*$ and $\mathbf{y}^*$

$$\langle \mathbf{x}^*, M\mathbf{y}^*\rangle = \langle M\mathbf{x}^*, \mathbf{y}^*\rangle$$

Therefore, if $\mathbf{x}^*$ and $\mathbf{y}^*$ are eigenvectors associated to distinct eigenvalues λ and μ we have

$$\langle \mathbf{x}^*, M\mathbf{y}^*\rangle = \langle \mathbf{y}^*, M\mathbf{x}^*\rangle \quad \rightarrow \quad \langle \mathbf{x}^*, \mu\mathbf{y}^*\rangle = \langle \mathbf{y}^*, \lambda\mathbf{x}^*\rangle \quad \rightarrow \quad \mu\langle \mathbf{x}^*, \mathbf{y}^*\rangle = \lambda\langle \mathbf{x}^*, \mathbf{y}^*\rangle$$

Since $\lambda \neq \mu$ the last equality is true only if $\mathbf{x}^*$ and $\mathbf{y}^*$ are orthogonal and $\langle \mathbf{x}^*, \mathbf{y}^*\rangle = 0$. Therefore, **eigenvectors associated to distinct eigenvalues of a symmetric matrix are orthogonal!**

Fact 3.5. *Positive (semi)definite matrices*
A symmetric matrix M the eigenvalues of which are all non-negative is positive semidefinite. If the eigenvalues are all strictly positive is positive definite. If $M = A^\top A$ for some matrix A, M is positive semidefinite (the closest thing to a square number in the matrix world). If A is square and full rank, M is positive definite.

Observation 3.1. *A geometric interpretation*
For a (semi)positive definite $n \times n$ matrix $A^\top A$ in the decomposition

$$A^\top A = Q\Lambda Q^\top$$

the diagonal matrix Λ has as many strictly positive values as the strictly positive eigenvalues of $A^\top A$ (each always repeated as many times as its degeneracy). But there is more. The equation

$$\mathbf{x}^\top A^\top A\mathbf{x} = c^2 \quad \forall \mathbf{x} \in \mathbb{R}^n \text{ with } \|\mathbf{x}\| = 1 \quad \text{and} \quad c > 0$$

describes a hyper-ellipsoid in $\mathbb{R}^n$. If $A^\top A$ is positive definite, the mutually orthogonal directions of the principal axes correspond to the eigenvectors (and the width of each principal axis are proportional to the reciprocal of the square root of the corresponding eigenvalues). If some eigenvalues are equal to 0 the hyper-ellipsoid is degenerate. If we choose the unit eigenvectors as a basis (that is, we rotate the axes according to the orthogonal matrix Q) the matrix reduces to Λ, revealing the diagonal nature of the underlying linear transformation.

Fact 3.6. *The beauty of SVD*
The Singular Value Decomposition is a matrix decomposition which generalises the decomposition $A^\top A = Q\Lambda Q^\top$ to a generic $n \times m$ matrix A. Assume we can write

$$AV = U\Sigma$$

with V an $m \times m$ orthonormal matrix corresponding to a suitably ordered basis in the m-dimensional row space of A, U an $n \times n$ orthonormal matrix corresponding to a suitably ordered basis in the n-dimensional column space of A, and Σ an $n \times m$ matrix with the only non zero entries on the diagonal with $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ with r the rank of A.

We can then write

$$A = U\Sigma V^\top \quad \text{and} \quad A^\top = V\Sigma^\top U^\top$$

with

$$A^\top A = V\Sigma^\top U^\top U\Sigma V^\top = V(\Sigma^\top\Sigma)V^\top \quad \text{and} \quad AA^\top = U\Sigma V^\top V\Sigma^\top U^\top = U(\Sigma\Sigma^\top)U^\top$$

since $U^\top U = I$ and $V^\top V = I$ ($n \times n$ and $m \times m$ identity matrices respectively). So, starting from the decomposition $AV = U\Sigma$, we reached the conclusion that V and U are the orthogonal transformations needed to render $A^\top A$ and $AA^\top$ in diagonal form. In addition, the diagonal elements of Σ are the square root of the eigenvalues of $A^\top A$ and $AA^\top$.

Exercises

Exercise 3.1. *Not necessarily a square matrix*

$$L_1 = \begin{pmatrix} 3 & -2 \\ -5 & 4 \\ 1 & -1 \end{pmatrix}$$

and

$$L_2 = \begin{pmatrix} 3 & -2 & 1 \\ -5 & 4 & -1 \end{pmatrix}$$

are matrices representing linear maps, between $\mathbb{R}^2$ to $\mathbb{R}^3$ and $\mathbb{R}^3$ to $\mathbb{R}^2$ respectively in the standard bases. Verify the rank-nullity theorem for both L_1 and L_2.

Exercise 3.2. *Invertible maps*
From the rank-nullity theorem it follows that a linear map $\mathcal{L} : V \to V$ is one-to-one if the kernel is the null space. Which of the following matrices represents a one-to-one linear map?

$$M_1 = \begin{pmatrix} 1 & 1 & 0 \\ 1 & 1 & 1 \\ 1 & 0 & 1 \end{pmatrix} \quad M_2 = \begin{pmatrix} 1 & 2 \\ 2 & 2 \end{pmatrix} \quad M_3 = \begin{pmatrix} 0 & 0 \\ 1 & 2 \end{pmatrix}$$

Exercise 3.3. *Finding the inverse of an orthogonal matrix*
Verify that if O is an orthogonal matrix, then $O^{-1} = O^\top$.

Exercise 3.4. *Square of any matrix?*
For an $n \times m$ matrix A, form the product $A^\top A$ and $AA^\top$. Find the dimension of the symmetric matrices $A^\top A$ and $AA^\top$ and prove that for any $\mathbf{x}$ of appropriate dimensionality

$$\mathbf{x}^\top A^\top A\mathbf{x} \geq 0 \quad \text{and} \quad \mathbf{x}^\top AA^\top\mathbf{x} \geq 0$$

Exercise 3.5. *Singular value decomposition*
Find the singular value decomposition for the matrix

$$A = \begin{pmatrix} 4 & 4 \\ -3 & 3 \end{pmatrix}$$

Exercise 3.6. *Once more (for a rank deficient matrix)*
Find the singular value decomposition for the matrix

$$A = \begin{pmatrix} 1 & 2 \\ 3 & 6 \end{pmatrix}$$

4 Real-valued functions of a real variable

If you don't master the basic notions of calculus for functions of one variable you are bound to be lost. Here is the list of thinks you should know.

Functions: domain, range, main properties, graph skills

Derivative: notion and computational aspects

Optimisation: finding the minimum of a smooth function

Integration by parts: no Fourier without it!

* * * * *

Theoretical minimum

Definition 4.1. *Domain and range*
The *domain* of a real-valued function f over the reals is the largest set $D \subseteq \mathbb{R}$ over which f can be defined. The *range* of f is the set $R \subseteq \mathbb{R}$ of the values taken by f.

Observation 4.1. *Notation for a real-valued function f of a real variable x*
These are all equivalent: $f : D \to \mathbb{R}$, $y = f(x)$, $x \to f(x)$, $x \to y$.

Observation 4.2. *Injective and strictly monotone functions*
Restrict the domain of a function f so that it is injective and look at its graph. You should immediately see that the function f in the restricted domain is strictly monotone. Make sure you are at ease with the fact that these two properties are actually equivalent.

Observation 4.3. *Inverse function*
The inverse of an invertible function f can be obtained by mirror symmetry of f with respect to the bisectrix of the first and third quadrant. Verify that $(e^x, \ln x)$ as well as $(x^2, \sqrt{x})$ are inverse to one another.

Observation 4.4. *Are log functions all the same?*
If a and b are two bases, $\log_b x$ is $\log_a x$ times the constant $\log_b a$. Indeed, from $y = \log_a x$, since

$$x = a^y \quad \to \quad \log_b x = \log_b(a^y) \quad \to \quad \log_b x = y \log_b a$$

we obtain

$$\log_b x = \log_b a \log_a x$$

Observation 4.5. *Periodic functions*
A function f is periodic if, for some T,

$$f(x+T) = f(x) \ \ \forall x$$

Notice that if a function is periodic from some T, then the same function is periodic for values multiple of T. The **period** of a periodic function is the **smallest** T for which a function is periodic. For $\sin x$, for example, the period is 2π while for $\tan x$ is π.

Observation 4.6. *Two important limits*
From the McLaurin's series

$$\sin x = x - \frac{x^3}{3!} + \frac{x^5}{5!} - \cdots = \sum_{n=0}^{+\infty} (-1)^n \frac{x^{2n+1}}{(2n+1)!} \quad \text{and} \quad e^x = 1 + x + \frac{x^2}{2!} + \cdots = \sum_{n=0}^{+\infty} \frac{x^n}{n!}$$

we obtain

$$\begin{aligned} \lim_{x\to 0} \frac{\sin x}{x} &= \lim_{x\to 0} \frac{x + o(x^3)}{x} = 1 \\ \lim_{x\to 0} \frac{e^x - 1}{x} &= \lim_{x\to 0} \frac{1 + o(x) - 1}{x} = 1 \end{aligned}$$

Fact 4.1. *Derivative properties*
Let $f'(x)$ denote the derivative of f at x for all f which can be differentiated at x. By applying the definition of derivative we find

$$(f(x) + g(x))' = f'(x) + g'(x)$$

and

$$(f(x) \times g(x))' = f(x) \times g'(x) + f'(x) \times g(x)$$

For higher order derivative, it suffices to say that the second derivative is the first derivative of the first derivative!
The derivative of a composite function $h(x) = g(f(x))$ is computed from the outer function by applying the *chain rule*, or

$$\frac{\mathrm{d}h(x)}{\mathrm{d}x} = \frac{\mathrm{d}g(f(x))}{\mathrm{d}x} = \left.\frac{\mathrm{d}g(t)}{\mathrm{d}t}\right|_{t=f(x)} \times \frac{\mathrm{d}f(x)}{\mathrm{d}x}$$

For example if $h(x) = e^{\sin x}$ since $f(x) = \sin x$ and $g(t) = e^t$ we have

$$\frac{\mathrm{d}e^{\sin x}}{\mathrm{d}x} = \left.\frac{\mathrm{d}e^t}{\mathrm{d}t}\right|_{t=\sin x} \times \frac{\mathrm{d}\sin x}{\mathrm{d}x} = \left.e^t\right|_{t=\sin x} \times \cos x = e^{\sin x}\cos x$$

Fact 4.2. *Integration by parts*
If f and g are smooth functions we have

$$\int_a^b f(x)g'(x)\mathrm{d}x = \left.f(x)g(x)\right|_a^b - \int_a^b g(x)f'(x)\mathrm{d}x$$

Exercises

Exercise 4.1. *Not necessarily the entire real line*
Determine domain and range for the real functions $x^n, \sin x, \cos x, \tan x, e^x$ and $\ln x$. Plot each function accordingly.

Exercise 4.2. *Function composition*
Determine whether $g(f(x))$ for $f(x) = \sin x$ and $g(x) = \cos x$, and for $f(x) = \ln x$ and $g(x) = x^2$ are well defined functions.

Exercise 4.3. *Injective functions*
Determine the largest subset of the domain on which the functions $x^n, \sin x, \cos x, \tan x, e^x$ and $\ln x$ are injective.

Exercise 4.4. *From one to many*
Given the graph of an arbitrary function $f(x)$ defined over the entire real line

$f(-x)$: is obtained by mirroring the f graph w.r.t. the Y-axis

$-f(-x)$: is obtained by mirroring the f graph w.r.t. the Y-axis and the X-axis

$f(|x|)$: graph of f for $x > 0$ and its mirror image w.r.t. the Y-axis for $x < 0$

$|f(x)|$: graph of f for all x for which $f(x) > 0$, its mirror image w.r.t. the Y-axis for all x for which $f(x) < 0$

$f(a^2x)$: shrink if $a^2 > 1$ (or stretch if $a^2 < 1$) the f graph along the X-axis leaving 0 in place

$a^2f(x)$: shrink if $a^2 > 1$ (or stretch if $a^2 < 1$) the f graph along the Y-axis leaving 0 in place

$f(x + x_0)$: shift the f graph to the left by x_0 if $x_0 > 0$ (or to the right otherwise)

$f(x) + x_0$: shift the f graph upward by x_0 if $x_0 > 0$ (or downward otherwise)

Make sure you are able to draw each of these graphs by heart!

Exercise 4.5. *First derivative (or why it is called calculus)*
Compute the first derivative of the functions $x^n, \sin x, \cos x, e^x$ and $\ln x$.

Exercise 4.6. *Unique global minimum and maximum*
Determine maxima and minima for $f(x) = x(1-x)$ in the interval $[0,1]$ and for $g(x) = e^x - x$ in the interval $[-1,1]$. How can you be sure that the point in which the derivative equals zero is a maximum for f and a minimum for g? How about the minimum for f and the maximum for g? Check the correctness of the obtained results plotting $f(x)$ and $g(x)$ in the two cases.

Exercise 4.7. *Local and global*
Find the local and global minima and maxima of the function $f : \mathbb{R} \to \mathbb{R}$

$$f(x) = 6x^4 - 10x^3 - 6x^2 + 15x$$

Exercise 4.8. *Multiple global minima and maxima*
Determine the minima and the maxima of the functions $f(x) = \sin x$ over the real line.

Exercise 4.9. *Integration by parts*
Compute the integral

$$\int_1^e \ln x \, \mathrm{d}x.$$

Exercise 4.10. *Two simple integrals*
Verify that

$$\int_0^{2\pi} x \cos x \, \mathrm{d}x = 0 \quad \text{and} \quad \int_0^{2\pi} \sin x \cos x \, \mathrm{d}x = 0$$

5 Multivariate functions

Functions in machine learning may depend on hundreds of thousands variables. Signals are often multidimensional. At least a few notions of multivariate functions won't hurt. Here is the list of thinks you should know.

Gradient: partial derivatives and directional derivative

Constrained optimisation: finding the minimum of a smooth function with the Lagrange multiplier technique

$$* \quad * \quad * \quad * \quad *$$

Theoretical minimum

The *gradient* generalises the notion of derivative to the case of multivariate functions. For a smooth function $f : \mathbb{R}^d \to \mathbb{R}$ the gradient at the point $x = (x_1 \;\; x_2 \;\; \dots \;\; x_d)^\top \in \mathbb{R}^d$ is an d-dimensional vector defined as

$$\nabla f(x) = \left(\frac{\partial f}{\partial x_1} \; \frac{\partial f}{\partial x_2} \; \cdots \; \frac{\partial f}{\partial x_d} \right)^\top$$

The i-th component of the gradient, with $i = 1, \dots, d$, is the partial derivative of f along the direction i. In general the derivative of f in the direction of the unit vector $n = (n_1 \;\; n_2 \;\; \dots \;\; n_d)^\top$ is given by

$$\frac{\partial f}{\partial n} = \langle \nabla f(x), n \rangle = \nabla f(x)^\top n$$

Observation 5.1. *Always orthogonal*
Clearly, $\forall n$ of unit norm and $\forall x \in D \subseteq \mathbb{R}^d$

$$\left| \nabla f(x)^\top n \right| \leq \| \nabla f(x) \|$$

Consequently, the *maximal change* of a function f is in the *direction of the gradient*, while a function f *does not change* in the direction(s) *orthogonal to the gradient*.

Observation 5.2. *Sufficient conditions for extremal points*
In one dimension a point x_0 at which the $f'(x_0) = 0$ and $f''(x_0) > 0$ ($f''(x_0) < 0$) is a minimum (maximum). In higher dimension a point x_0 at which $\nabla f(x_0) = 0$ and all the eigenvalues of the Hessian matrix $H(x_0)$ are strictly positive (negative) is a minimum (maximum).

Exercises

Exercise 5.1. *Lagrange multiplier*
Find the minimum of the function $f : \mathbb{R}^2 \to \mathbb{R}$, $f(x, y) = x^2 - y$ in the region

$$g(x, y) = x^2 + y^2 \leq 25$$

6 Theory of probability

Addressing machine learning without some basic knowledge of probability leads to wrongful interpretation and poor understanding of the basics. Here is the list of thinks you should know.

Random variables: what they are

Probability: joint, marginal, and conditional probability

Expected values: expected value and variance

* * * * *

Theoretical minimum

Let S be the sample space. A *random variable* X is a *measurable* real-valued function defined on S

$$X : S \to \mathbb{R}$$

Fact 6.1. *Markov inequality*
Let X be a positive valued random variable distributed according to a density f. Then $\forall a > 0$

$$P\{X \geq a\} \leq \frac{\mathbb{E}[X]}{a}$$

Proof
Since $xf(x) \geq 0$ we have

$$\mathbb{E}[X] = \int_0^{+\infty} xf(x)\mathrm{d}x \geq \int_a^{+\infty} xf(x)\mathrm{d}x$$

Multiplying and dividing by a since $x/a > 1$ we obtain

$$\mathbb{E}[X] \geq a \int_a^{+\infty} \frac{x}{a} f(x)\mathrm{d}x \geq a \int_a^{+\infty} f(x)\mathrm{d}x$$

and thus

$$P\{X \geq a\} = \int_a^{+\infty} f(x)\mathrm{d}x \leq \frac{\mathbb{E}[X]}{a}$$

Fact 6.2. *Chebyshev inequality*
Let X be a random variable with finite expected value μ and finite variance σ^2. Then $\forall \epsilon > 0$

$$P\{|X - \mu| \geq \epsilon\} \leq \frac{\sigma^2}{\epsilon^2}$$

Proof
We apply Markov inequality to the non-negative random variable $(X - \mu)^2$ with $a = \epsilon^2$ and noticing that

$$\forall \epsilon > 0, P\{|X - \mu| \geq \epsilon\} = P\{|X - \mu|^2 \geq \epsilon^2\}$$

Fact 6.3. *Weak law of large numbers*
Let X_i with $i = 1, 2, \ldots, n$ be independent and identically distributed random variables with $\mathbb{E}[X_i] = \mu$ and finite variance σ^2. Then

$$\forall \epsilon > 0 \lim_{n \to \infty} P\left\{\left|\frac{1}{n}\sum_i X_i - \mu\right| \geq \epsilon\right\} = 0$$

Proof
Since

$$\mathbb{E}\left[\frac{1}{n}\sum_i X_i\right] = \mu \quad \text{and} \quad Var\left(\frac{1}{n}\sum_i X_i\right) = \frac{\sigma^2}{n}$$

from Chebyshev inequality for $k = \epsilon$ we obtain

$$P\left\{\left|\frac{1}{n}\sum_i X_i - \mu\right| \geq \epsilon\right\} \leq \frac{\sigma^2}{n\epsilon^2} \underset{n \to \infty}{\to} 0$$

Exercises

Exercise 6.1. *Number of heads in three coin tosses*
Toss a coin three times. The number X of heads is a random variable. In this case S is the set of the possible triplets

$$S = \{HHH, HHT, HTH, HTT, THH, THT, TTH, TTT\}$$

and thus $X : S \to \{0, 1, 2, 3\}$. Determine the probability of X being equal to 0, 1, 2, and 3 assuming that the coin is fair and, then, loaded with $p(H) = 1/3$.

Exercise 6.2. *Balls in an urn*
Draw 3 balls at random from an urn which contains 3 red, 4 white, and 5 blue balls. The random variable X counts the number of red balls which have been drawn, Y the number of white balls. The sample space S is the set of all possible triplets, or cases. We have that $X : S \to \{0, 1, 2, 3\}$ and $Y : S \to \{0, 1, 2, 3, 4\}$. The possible cases are $\binom{12}{3} = 220$. Denoting with (x, y) the event that $X = x$ and $Y = y$, for the favourable cases we have

$$\begin{array}{llll}
(0,0) \to \binom{5}{3} = 10 & (0,1) \to \binom{4}{1}\binom{5}{2} = 40 & (0,2) \to \binom{4}{2}\binom{5}{1} = 30 & (0,3) \to \binom{4}{3} = 4 \\
(1,0) \to \binom{3}{1}\binom{5}{2} = 30 & (1,1) \to \binom{3}{1}\binom{4}{1}\binom{5}{1} = 60 & (1,2) \to \binom{3}{1}\binom{4}{2} = 18 & \\
(2,0) \to \binom{3}{2}\binom{5}{1} = 15 & (2,1) \to \binom{3}{2}\binom{4}{1} = 12 & & \\
(3,0) \to \binom{3}{3} = 1 & & &
\end{array}$$

The joint and the marginals probabilities are given by

$$\begin{array}{lllll|l}
p(0,0) = 10/220 & p(0,1) = 40/220 & p(0,2) = 30/220 & p(0,3) = 4/220 & & p_X(0) = 84/220 \\
p(1,0) = 30/220 & p(1,1) = 60/220 & p(1,2) = 18/220 & & & p_X(1) = 108/220 \\
p(2,0) = 15/220 & p(2,1) = 12/220 & & & & p_X(2) = 27/220 \\
p(3,0) = 1/220 & & & & & p_X(3) = 1/220 \\
\hline
p_Y(0) = 56/220 & p_Y(1) = 112/220 & p_Y(2) = 48/220 & p_Y(3) = 4/220 & &
\end{array}$$

Compute the conditional probabilities $p(X|Y = 0)$ and $p(Y|X = 1)$ and explain why you are not surprised that $\forall i, j \; p(i, j) \neq p_X(i)p_Y(j)$.

Exercise 6.3. *Expected value of* heads *in three tosses of a fair coin*

Exercise 6.4. *Expected value of a function of a random variable*
Let $Y = X^2$. Compute $\mathbb{E}[X^2]$ if for the random variable X you have

$$P(X = -1) = 0.2, \;\; P(X = 0) = 0.5 \;\text{ e }\; P(X = 1) = 0.3$$

Exercise 6.5. *Variance of the expected value of* heads *in three tosses of a fair coin*

Exercise 6.6. *Sample average and variance*
If X_i for $i = 1, \ldots, n$ are random variables identically and independently distributed with expected value μ and variance σ^2 define the *sample average* μ_n and the *sample variance* σ_n^2 as

$$\mu_n = \frac{\sum_i X_i}{n} \quad \text{e} \quad \sigma_n^2 = \frac{\sum_i (X_i - \mu_n)^2}{n-1}$$

and compute $\mathbb{E}[\mu_n]$, $Var(\mu_n)$ and $\mathbb{E}[\sigma_n^2]$.

7 Complex numbers

Complex numbers are needed to understand the basics of how to analyse and process digital signals. Here is the list of things you should know.

Cartesian and polar representation: real and imaginary part, phase and module, complex conjugate of a complex number

Algebraic manipulation: sum, product and how to choose the appropriate representation of complex numbers

Roots of unity: no Fast Fourier Transform without it!

* * * * *

Theoretical minimum

If i is the *imaginary unit*, a number such that $i^2 = -1$, we write a complex number $z \in \mathbb{C}$ as

$$z = a + ib$$

with $a \in \mathbb{R}$ the *real part* and $b \in \mathbb{R}$ the *imaginary part.*

Observation 7.1. *Argand representation*
A complex number $z = a + ib$ can be represented as an ordered pair of real numbers (a, b) in the complex plane $\mathbb{C}$ (see figure 1). The horizontal axis is the (usual) real axis. If $0 \leq \theta < 2\pi$ is the angle formed by the complex number z with the real axis we have

$$a = |z| \cos\theta \text{ and } b = |z| \sin\theta \quad \text{with } \theta = \arctan\frac{b}{a} \text{ if } a \neq 0, \text{ or } \theta = \frac{\pi}{2} \text{ otherwise}$$

Observation 7.2. *Phase and module of a complex number*
The angle θ is called the *phase* of the complex number z. For all real positive numbers $\theta = 0$, while for real negative numbers $\theta = \pi$. For $z = 0$ the phase is not defined.
The number $z^* = a - ib$, obtained by changing the sign of the imaginary part of $z = a + ib$, is the *complex conjugate* of z. For all $z \in \mathbb{C}$,

$$zz^* = a^2 + b^2$$

is a real non-negative number. The quantity $|z| = \sqrt{a^2 + b^2}$ is the *module* of z.

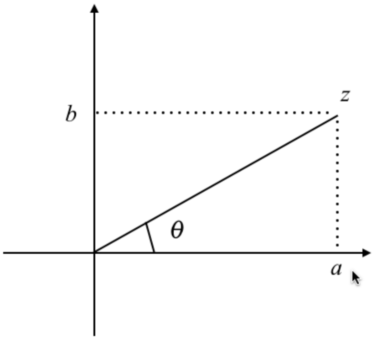

Figure 1: See text.

Observation 7.3. *Polar representation*
Since $z(\theta) = \cos\theta + i\sin\theta$ we have

$$\frac{\mathrm{d}z(\theta)}{\mathrm{d}\theta} = -\sin\theta + i\cos\theta = iz(\theta)$$

Integrating both sides of the equality

$$\frac{\mathrm{d}z(\theta)}{z(\theta)} = i\mathrm{d}\theta$$

we obtain

$$\ln z(\theta) = i\theta + C \tag{1}$$

for some constant C. Since $z(0) = 1$, we can conclude that $C = 0$. Applying the exponential to both sides of equation (1) we finally obtain

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

In general, for an arbitrary $z = a + ib \in \mathbb{C}$, we have $z = \rho\, e^{i\theta}$ with $\rho = |z|$.

Through the rules of usual algebra, if $z = a + ib$ and $w = c + id$, we find

$$\begin{aligned} z + w &= (a + ib) + (c + id) = (a + c) + i(b + d) \\ zw &= (a + ib)(c + id) = (ac - bd) + i(ad + bc) \end{aligned}$$

Observation 7.4. *Sum easier than product*
The real and imaginary part of the sum of two complex numbers is the sum of their real and imaginary parts respectively. If $z = a + ib$ and $w = c + id$ we find

$$z + w = (a + ib) + (c + id) = (a + c) + i(b + d)$$

In the polar representation, no simple formula can be found for the sum of two complex numbers.

Observation 7.5. *Product easier than sum*
The module of the product of two complex numbers $z_1 = \rho_1 e^{i\theta_1}$ and $z_2 = \rho_2 e^{i\theta_2}$ is the product of their modules, $\rho_1\rho_2$, while the phase is the sum of their phases, $\theta_1 + \theta_2$, since

$$zw = \rho_1 e^{i\theta_1} \rho_2\, e^{i\theta_2} = \rho_1\rho_2\, e^{i\theta_1} e^{i\theta_2} = \rho_1\rho_2\, e^{i(\theta_1+\theta_2)}$$

In the Argand representation no simple formula for the product can be found. If $z = a + ib$ and $w = c + id$ we find

$$zw = (a + ib)(c + id) = (ac - bd) + i(ad + bc)$$

Exercises

Exercise 7.1. *Some simple facts on complex numbers*
Prove that

1. $\forall z \quad |z| = |z^*|$
2. $\forall z \quad (z^*)^* = z$
3. $|z| = 0$ iff $z = 0$

Exercise 7.2. *A few minutes in the imaginary gym*
Let $z = a + ib$ and $w = c + id$. Compute the real and imaginary part of zw^* and z/w.

Exercise 7.3. *Some more time in the imaginary gym*
Find the polar representation of 2, -3, $4i$, $-5i$, $1 + i$, and i^i.

Exercise 7.4. *Roots of unity*
Compute the n roots of the equation

$$z^n = 1$$

and show that they define a regular polygon of n sides inscribed in the unit circle in the complex plane with one of the vertices lying in $(1, 0)$.