Mathematical Methods in Data Science

Paul Breiding and Samantha Fairchild

Author's addresses:
Paul Breiding, University of Osnabrück + MPI MiS Leipzig, pbreiding@uni-osnabrueck.de.
Samantha Fairchild, University of Osnabrück + MPI MiS Leipzig, samantha.fairchild@mis.mpg.de

These lecture notes were written in the Summer Semester 2022, when the first author gave the class "Mathematische Grundlagen der Datenanalyse" at the University of Osnabrück.

The authors have been funded by the Deutsche Forschungsgemeinschaft
(DFG, German Research Foundation) – Projektnummer 445466444.

Contents

1 The Basics

Many mathematical methods in data analysis rely on linear algebra and probability. In the first two lectures we will recall basic concepts from these fields.

1.1 Linear Algebra

This lecture is based on the article *The Fundamental Theorem of Linear Algebra* by Gilbert Strang [Str93]. We will use the following notation:

$$A = (a_{ij}) \in \mathbb{R}^{m\times n} \text{ (resp. } \mathbb{C}^{m\times n})$$

is an $m \times n$ **matrix** with real (resp. complex) entries a_{ij} for $1 \leq i \leq m$, $1 \leq j \leq n$. The column vectors are

$$a_j := (a_{ij})_{i=1}^m.$$

A matrix $A \in \mathbb{R}^{m\times n}$ can be viewed as a **list** of vectors in $\mathbb{R}^m$ which we denote by

$$A = [a_1, \ldots, a_n].$$

For $x = (x_1, \ldots, x_n)^T \in \mathbb{R}^n$,

$$Ax = x_1 a_1 + \cdots + x_n a_n$$

is a **linear combination** of the columns of A. Other interpretations of A are

1. a list of n vectors in $\mathbb{R}^m$
2. a list of m vectors in $\mathbb{R}^n$
3. a linear map $\mathbb{R}^n \to \mathbb{R}^m$ given by $x \mapsto Ax$
4. a linear map $\mathbb{R}^m \to \mathbb{R}^n$ given by $y \mapsto A^T y$
5. a bilinear map $\mathbb{R}^n \times \mathbb{R}^m \to \mathbb{R}$ given by $(x, y) \mapsto y^T Ax$.

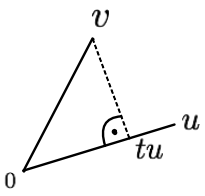

Figure 1.1: The meaning of the inner product between u and v is illustrated in this picture: let $t \in \mathbb{R}$ such that $v = tu + u'$, where u' is orthogonal to u. Then, $\langle u, v\rangle = u^T(tu + u') = tu^T u = t\langle u, u\rangle$. In particular, if $\langle u, u\rangle = \langle v, v\rangle = 1$, then $t = \langle u, v\rangle$ is the arccosine of the angle between u and v.

All of these viewpoints are best understood by considering four subspaces (two subspaces of $\mathbb{R}^n$ and two of $\mathbb{R}^m$).

Definition 1.1 (Four Subspaces)**.** Let $A \in \mathbb{R}^{m\times n}$. The **image** and **kernel** of A and A^T are

1. $\mathrm{Im}(A) := \{Ax \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$,
2. $\mathrm{Im}(A^T) := \{A^T y \mid y \in \mathbb{R}^m\} \subseteq \mathbb{R}^n$,
3. $\ker(A) := \{x \in \mathbb{R}^n \mid Ax = 0\} \subseteq \mathbb{R}^n$,
4. $\ker(A^T) := \{y \in \mathbb{R}^m \mid A^T y = 0\} \subseteq \mathbb{R}^m$

We give the $\mathbb{R}$-vector spaces $\mathbb{R}^n$ and $\mathbb{R}^m$ the structure of a Euclidean space by defining the positive definite form $\langle a, b\rangle := a^T b$. For $A = [a_1, \ldots, a_n]$, we have $A^T y = [\langle a_i, y\rangle]_{i=1}^n$.

Definition 1.2. Let $U, V \subseteq \mathbb{R}^n$ be subspaces. Then U **is perpendicular to** V (denoted $U \perp V$) when for all $u \in U$ and $v \in V$, $\langle u, v\rangle = 0$.

In the following we will denote $r = r(A)$ to be the **rank** of A.

Theorem 1.3. *Let $A \in \mathbb{R}^{m\times n}$. Then*

1. $\mathrm{Im}(A) \oplus \ker(A^T) = \mathbb{R}^m$,
2. $\mathrm{Im}(A) \perp \ker(A^T)$,
3. $\mathrm{Im}(A^T) \oplus \ker(A) = \mathbb{R}^n$,
4. $\mathrm{Im}(A^T) \perp \ker(A)$.

Proof of (1) and (2). From linear algebra, we know that

$$r(A) = \dim(\mathrm{Im}(A)) = \dim(\mathrm{Im}(A^T))$$

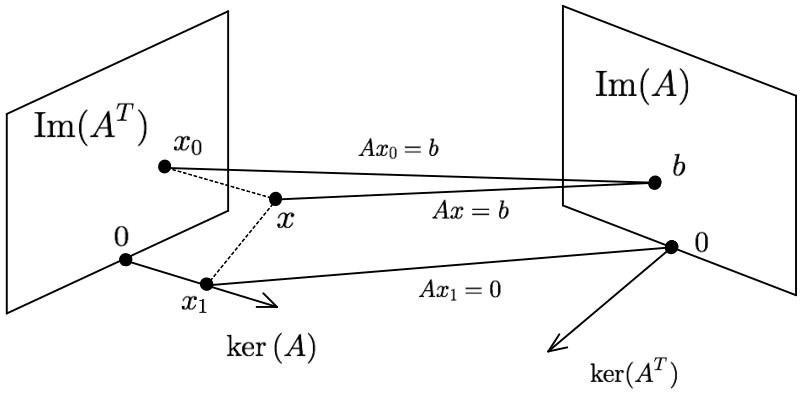

Figure 1.2: The situation when $b \in \mathrm{Im}(A)$: in this case, $Ax = b$ has a unique solution $x \in \mathrm{Im}(A^T)$ and the solution space for $Ax = b$ is $x + \ker(A)$.

and by the Rank-Nullity theorem

$$\dim(\mathrm{Im}(A^T)) + \dim(\ker(A^T)) = m.$$

Therefore

$$\dim(\mathrm{Im}(A)) + \dim(\ker(A^T)) = m.$$

Moreover for $y \in \ker(A^T)$ and $Ax \in \mathrm{Im}(A)$,

$$\langle y, Ax \rangle = y^T Ax = (A^T y)^T x = 0.$$

Thus $\mathrm{Im}(A) \perp \ker(A^T)$ and in particular $\mathrm{Im}(A) \cap \ker(A^T) = \{0\}$ and

$$\dim(\mathrm{Im}(A) + \ker(A^T)) = \dim(\mathrm{Im}(A)) + \dim(\ker(A^T)) = m.$$

Thus $\mathrm{Im}(A) \oplus \ker(A^T) = \mathbb{R}^m$. The proof of (3) and (4) follows similarly. □

We now want to understand the solution of the system of linear equations $Ax = b$ in the context of Theorem 1.3. Namely, let $b \in \mathrm{Im}(A)$ and let $r = \dim(\mathrm{Im}(A)) = \dim(\mathrm{Im}(A^T))$. First, we observe that $Ax = b$ has a solution $x \in \mathbb{R}^n$, if and only if $b \in \mathrm{Im}(A)$. Suppose that x is such a solution. This situation is depicted in Fig. 1.2. From Theorem 1.3 we know that $\mathrm{Im}(A^T) \oplus \ker(A) = \mathbb{R}^n$. So, there exist uniquely determined $x_0 \in \mathrm{Im}(A^T)$ and $x_1 \in \ker(A)$ with $x = x_0 + x_1$ and we have

$$b = Ax = A(x_0 + x_1) = Ax_0 + Ax_1 = Ax_0.$$

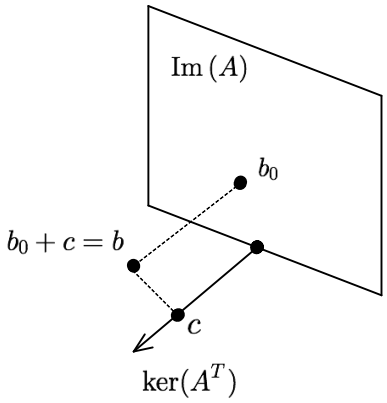

Figure 1.3: Visualization of the proof of Lemma 1.4: b_0 minimizes the distance from b to $\mathrm{Im}(A)$.

Therefore, $Ax = b$ has a **unique** solution in $\mathrm{Im}(A^T)$. Consequently, A restricted to $\mathrm{Im}(A^T)$ is a **linear isomorphism**.

When $b \notin \mathrm{Im}(A)$ there is no solution to $Ax = b$. We can however find the point $b_0 \in \mathrm{Im}(A)$ which minimizes the Euclidean distance $\|b - b_0\| = \sqrt{\langle b - b_0, b - b_0 \rangle}$. We use the notation

$$b_0 = \mathrm{argmin}_{y \in \mathrm{Im}(A)} \|b - y\|$$

to denote the argument (i.e. the value $y = b_0$) which minimizes the function $\|b - y\|$. The solution to this minimization problem and the fact that b_0 is uniquely determined is given by the next lemma.

Lemma 1.4. *Let $b \in \mathbb{R}^m$ and $A \in \mathbb{R}^{m \times n}$. The point $b_0 = \mathrm{argmin}_{y \in \mathrm{Im}(A)} \|b - y\|$ is determined by*

1. *the decomposition from Theorem 1.3, which gives $b = b_0 + c$ for $c \in \ker(A^T)$; or*
2. *$A^T b = A^T b_0$.*

Proof. (2.⇒1.) This direction follows because $e \in \ker(A^T)$. This also shows that b_0 is uniquely determined.

So now it suffices to prove (2.) Let $A = [a_1, \ldots, a_n]$. Since $b_0 \in \mathrm{Im}(A)$, set $b_0 = Ax_0$ for some x_0. Define the map $\phi(x) = Ax - b$. Suppose that we write the output vector in $\mathbb{R}^m$ by $\phi = [\phi_1, \ldots, \phi_m]^T$. Then, we minimize the scalar function $\|\phi(x)\|$ by taking the

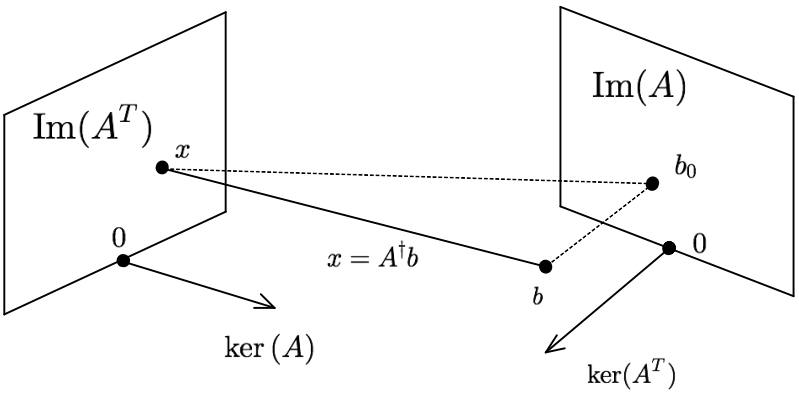

Figure 1.4: The pseudoinverse $A^\dagger \in \mathbb{R}^{m\times n}$ of $A \in \mathbb{R}^{m\times n}$ first orthogonally projects $b \in \mathbb{R}^n$ to $b_0 \in \mathrm{Im}(A)$ and then maps b_0 to the unique point $x \in \mathrm{Im}(A^T)$ with $Ax = b_0$.

derivative and setting it equal to zero. Namely we want to compute when the gradient $\frac{\mathrm{d}}{\mathrm{d}x}\|\phi(x)\| = \left[\frac{\partial}{\partial x_1}\|\phi(x)\|, \ldots, \frac{\partial}{\partial x_n}\|\phi(x)\|\right] \in \mathbb{R}^n$ is equal to zero. We compute

$$\frac{\partial}{\partial x_i}\|\phi(x)\| = \frac{1}{2\|\phi(x)\|}\sum_{j=1}^{m} 2\phi_j(x)\frac{\partial \phi_j}{\partial x_i}(x) = \frac{1}{\|\phi(x)\|}a_i^T(Ax-b).$$

If $\|\phi(x_0)\| = 0$, then $b_0 = b$ and we are done. Otherwise we must have that x_0 satisfies $A^T(Ax-b) = 0$. This implies $A^TAx_0 = A^Tb$, and so $A^Tb_0 = A^Tb$. □

Lemma 1.4 implies that the map which projects b to the point on $\mathrm{Im}(A)$ minimizing the distance to b in the Euclidean norm is linear: call it $\Pi_{\mathrm{Im}(A)}$. Furthermore, recall that A restricted to $\mathrm{Im}(A^T)$ is a linear isomorphism, hence invertible. Consequently, we have a well-defined linear map $(A|_{\mathrm{Im}(A^T)})^{-1} \circ \Pi_A : \mathbb{R}^m \to \mathbb{R}^n$, shown in Fig. 1.4. The matrix representation of this linear map is called the **pseudoinverse** of A.

Definition 1.5. Let $A \in \mathbb{R}^{m\times n}$. The **pseudoinverse** $A^\dagger \in \mathbb{R}^{n\times m}$ is the matrix such that

$$A^\dagger b = x$$

for $x \in \mathrm{Im}(A^T)$, $Ax = b_0$ and $b_0 = \mathrm{argmin}_{y\in\mathrm{Im}(A)}\|b-y\|$.

Note when consulting other texts sometimes A^+ is used instead $A^\dagger$.

Let us first notice two properties of the pseudoinverse, which follow immediately from the definition.

Corollary 1.6. *Let $A \in \mathbb{R}^{m\times n}$ and $A^\dagger \in \mathbb{R}^{n\times m}$ be its pseudoinverse.*

1. *If A is invertible, then $A^\dagger = A^{-1}$.*
2. *$AA^\dagger$ is the orthogonal projection onto $\mathrm{Im}(A)$; i.e., $AA^\dagger = \Pi_{\mathrm{Im}(A)}$.*

In the case when $A \in \mathbb{R}^{m\times n}$ has full rank, which means that $r(A) = \min\{m,n\}$, the pseudoinverse has the following properties.

Proposition 1.7. *Let $A \in \mathbb{R}^{m\times n}$ have full rank.*

1. *If $r(A) = n$,*
$$A^\dagger = (A^TA)^{-1}A^T$$
and $A^\dagger A = \mathbf{1}_n$. So A is left-invertible.
2. *If $r(A) = m$,*
$$A^\dagger = A^T(AA^T)^{-1}$$
and $AA^\dagger = \mathbf{1}_m$. So A is right-invertible.

Proof. Let $b \in \mathbb{R}^m$ and $A^\dagger b = x$. By Lemma 1.4 we have $A^TAx = A^Tb$, which implies

$$A^TAA^\dagger b = A^Tb.$$

Since $r(A) = n$, the matrix $A^TA \in \mathbb{R}^{n\times n}$ is invertible, so that

$$A^\dagger b = (A^TA)^{-1}A^Tb.$$

This shows $A^\dagger = (A^TA)^{-1}A^T$ and it also shows $A^\dagger A = (A^TA)^{-1}A^TA = \mathbf{1}_n$. For the second part, see Exercise 1.2. □

In closing of this lecture we want to discuss an important choice of bases for $\mathrm{Im}(A)$ and $\mathrm{Im}(A^T)$. Let $r := r(A)$. The matrix $A^TA \in \mathbb{R}^{n\times n}$ is symmetric. By the spectral theorem, there is a basis $\{v_1,\dots,v_n\}$ of eigenvectors of A^TA such that $\langle v_i, v_j\rangle = \delta_{i,j}$ (called an **orthonormal basis**). Let λ_i be the eigenvalue corresponding to v_i; i.e., $A^TAv_i = \lambda_i v_i$. Since A^TA is positive semidefinite, it has only real nonnegative eigenvalues. We can assume that $\lambda_1 \geq \cdots \geq \lambda_r > 0$ and $\lambda_{r+1} = \cdots = \lambda_n = 0$. We have

$$\mathrm{span}\{v_1,\dots,v_r\} = \mathrm{span}\{v_{r+1},\dots,v_n\}^\perp = \mathrm{Im}(A^T) = \ker(A)^\perp$$

(the last equality because of Theorem 1.3), so $\{v_1, \dots, v_r\}$ is an orthonormal basis for $\mathrm{Im}(A^T)$ and $\{v_{r+1}, \dots, v_n\}$ is an orthonormal basis for $\ker(A)$. Let

$$u_i := \frac{1}{\sqrt{\lambda_i}} A v_i, \quad i = 1, \dots, r.$$

Then, by construction we have

$$\langle u_i, u_j \rangle = \frac{1}{\sqrt{\lambda_i \lambda_j}} v_i^T A^T A v_j = \delta_{i,j},$$

which shows that $\{u_1, \dots, u_r\}$ is an orthonormal basis for $\mathrm{Im}(A)$. We have

$$A v_i = \sigma_i u_i, \quad \sigma_i = \sqrt{\lambda_i}.$$

For $U = [u_1, \dots, u_r] \in \mathbb{R}^{m \times r}$, $V = [v_1, \dots, v_r] \in \mathbb{R}^{n \times r}$ and $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$ we have

$$A = U \Sigma V^T.$$

This decomposition is called the **singular value decomposition** (SVD) of A and $\sigma_1, \dots, \sigma_r$ are called the **singular values** of A. The next theorem shows that the SVD is essentially unique.

Theorem 1.8. *Let $A \in \mathbb{R}^{m \times n}$ and $r = r(A)$. Then, there exist matrices $U \in \mathbb{R}^{m \times r}$ and $V \in \mathbb{R}^{n \times r}$ with $U^T U = V^T V = \mathbf{1}_r$ and uniquely determined numbers $\sigma_1, \dots, \sigma_r > 0$ such that*

$$A = U \Sigma V^T, \quad \Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r).$$

We have $\mathrm{Im}(A) = \mathrm{Im}(U)$ and $\mathrm{Im}(A^T) = \mathrm{Im}(V)$. If the σ_i are pairwise distinct and ordered $\sigma_1 > \ldots > \sigma_r > 0$, the matrices U and V are uniquely determined up to the signs of their columns.

Proof. Existence of the SVD and $\mathrm{Im}(A) = \mathrm{Im}(U)$ and $\mathrm{Im}(A^T) = \mathrm{Im}(V)$ follow from the discussion above. We have to show uniqueness of singular values, and in the case when the singular values are pairwise distinct uniqueness of U and V (up to sign). Suppose

$$A = U \Sigma V^T = \tilde{U} \tilde{\Sigma} \tilde{V}^T$$

are two SVDs of A with $\Sigma = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$ and $\tilde{\Sigma} = \mathrm{diag}(\tilde{\sigma}_1, \dots, \tilde{\sigma}_r)$. Then, we have

$$A A^T = U \Sigma V^T V \Sigma U^T = U \Sigma^2 U^T \qquad \text{and} \qquad A A^T = \tilde{U} \tilde{\Sigma} \tilde{V}^T \tilde{V} \tilde{\Sigma} \tilde{U}^T = \tilde{U} \tilde{\Sigma}^2 \tilde{U}^T,$$

because $V^TV = \tilde{V}^T\tilde{V} = \mathbf{1}_r$. Let us write $U = [u_1, \dots, u_r]$ and $\tilde{U} = [\tilde{u}_1, \dots, \tilde{u}_r]$. The above equations imply that $AA^T u_i = \sigma_i^2 u_i$ and $AA^T \tilde{u}_i = \tilde{\sigma}_i^2 \tilde{u}_i$ for $i = 1, \dots, r$. Using that $r = r(A) = r(AA^T)$ we conclude that both $\sigma_1, \dots, \sigma_r$ and $\tilde{\sigma}_1, \dots, \tilde{\sigma}_r$ are the nonzero eigenvalues of AA^T. Since eigenvalues are unique, this implies $\sigma_i = \tilde{\sigma}_i$, $i = 1, \dots, r$. Therefore, the singular values are uniquely determined.

Let us now assume that the σ_i are pairwise distinct. Then, since every σ_i is positive, also the σ_i^2 are pairwise distinct for $i = 1, \dots, r$. This means that the nonzero eigenvalues of AA^T are all simple, which implies that the eigenvector of σ_i is unique up to sign, hence $u_i = \pm\tilde{u}_i$. Repeating the same argument for A^TA shows that the columns of V and $\tilde{V}$ also coincide up to sign. □

An alternative definition of the SVD is $A = USV^T$ for $U \in \mathbb{R}^{m\times k}$ and $V \in \mathbb{R}^{n\times k}$ with $k = \min\{m,n\}$ and $U^TU = V^TV = \mathbf{1}_k$, and $S = \operatorname{diag}(\sigma_1, \dots, \sigma_r, 0, \dots, 0)$. This is sometimes called the **non-compact SVD**, while the decomposition in Theorem 1.8 is called **compact SVD**. The difference between the two is that the compact SVD involves orthonormal bases of $\operatorname{Im}(A)$ and $\operatorname{Im}(A^T)$, while the non-compact SVD appends orthonormal vectors from $\ker(A^T)$ and $\ker(A)$. The way one should think about the SVD (compact or non-compact) is that it provides particular orthonormal bases that reveal essential information about the matrix A.

The final result of this lecture is the connection between SVD and pseudoinverse.

Lemma 1.9. *Let $A \in \mathbb{R}^{m\times n}$ and $A = U\Sigma V^T$ be the SVD of A as in Theorem 1.8. Then,*

$$A^\dagger = V\Sigma^{-1}U^T.$$

Exercise 1.1. Show that $\sum_{j=1}^m \phi_j(x)\frac{\partial \phi_j}{\partial x_i}(x) = a_i^T(Ax - b)$ as in the proof of Lemma 1.4.

Exercise 1.2. Prove part 2 of Proposition 1.7.

Exercise 1.3. Prove Lemma 1.9.

Exercise 1.4. Consider $A = \begin{bmatrix} 1 & 0 & -2 \\ -\frac{1}{2} & 0 & 1 \end{bmatrix}$.

(a) Compute by hand a singular value decomposition $U\Sigma V^T$ and the pseudoinverse $A^\dagger$ of A.

(b) Now try to do the same using the `LinearAlgebra` library in `Julia` [BEKS17] (or any other numerical linear algebra implementation). Do you get what you expected? What happens if you compare the pseudoinverse obtained via the command `pinv` to the one obtained by taking $V\Sigma^{-1}U^T$?

1.2 Probability Theory

Using probability theory we can model uncertainty and randomness in data. The basic idea is to assign to an event A a probability $P(A) \in [0,1]$. It measures how likely it is that A happens.

There are two main interpretations of $P(A)$.

1. The first interpretation is that $P(A)$ should be approximately equal to the relative frequency of the event A happening in n experiments. That is, $P(A) \approx \frac{k}{n}$, where k is the number of times A happened in n experiments. Furthermore, as $n \to \infty$ the $\approx$ should become an equality. This point of view is called **frequentist probability**.
2. The second interpretation is that $P(A)$ is a value based on experience or knowledge inferred from data. In particular, this means that unlike in the frequentist's view $P(A)$ is not independent of the observed data and can be updated when new data is available. Furthermore, we can model incomplete information about determinstic processes. This point of view is called **Bayesian probability**.

For data analysis Bayesian probability is more relevant. However, both points are only interpretations of the abstract mathematical definitions in probability! We discuss the theory in this lecture. For more details see, for instance, the (freely available) textbook [Ash70].

Definition 1.10. Let Ω be a nonempty set and $\mathscr{A} \subset 2^\Omega$ be a subset of the power set of Ω. We call $\mathscr{A}$ a **σ-algebra**, if it satisfies the following properties

1. $\Omega \in \mathscr{A}$;
2. if $A \in \mathscr{A}$, then $\Omega \setminus A \in \mathscr{A}$;
3. if $A_n \in \Omega, n \in \mathbb{N}$, then $\bigcup_{n\in\mathbb{N}} A_n \in \Omega$.

Definition 1.11. A **probability space** is a triple $(\Omega, \mathscr{A}, P)$, where

1. Ω is a nonempty set,
2. $\mathscr{A} \subset 2^\Omega$ is a σ-algebra, and
3. $P : \mathscr{A} \to [0,1]$ is a probability measure. This means that

$$P(\Omega) = 1 \qquad \text{and} \qquad P\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) = \sum_{n\in\mathbb{N}} P(A_n), \text{ if } A_i \cap A_j = \emptyset \text{ for } i \neq j.$$

Every set $A \in \mathscr{A}$ is called an **event**, Ω is called the **space of events**, and $P(A)$ is the **probability** of A. The map P is called a **(probability) distribution**.

The restriction that $\mathscr{A}$ is a σ-algebra is crucial: without this assumption one can create paradoxa like the Banach-Tarski-Paradox. However, if Ω is discrete or even finite we can always take $\mathscr{A} = 2^\Omega$ as σ-algebra. In the case $\Omega = \mathbb{R}$ we have the **Borel σ-algebra**. This is the smallest σ-algebra (by inclusion) that contains every interval in $\mathbb{R}$.

Definition 1.12. Let $\mathscr{A}$ be the Borel σ-algebra in $\mathbb{R}$. We call a function $g : \mathbb{R} \to \mathbb{R}$ **measurable**, if for all $A \in \mathscr{A}$ we have $g^{-1}(A) \in \mathscr{A}$.

Example 1.13. Let $\Omega = \{0,1\}$ and $\mathscr{A} = \{\emptyset, \{0\}, \{1\}, \Omega\} = 2^\Omega$. Suppose $P(\{1\}) = p$. Then, we have

$$P(\{0\}) = P(\Omega) - P(\{1\}) = 1 - p.$$

This probability distribution is called **Bernoulli distribution** with parameter p. It models the probability of an experiment with two outcomes.

Often Ω is complicated, but at the same time we don't want to know every information about events in Ω, just some particular pieces of information. This motivates the definition of random variables.

Definition 1.14. A **random variable** X is a map $X : (\Omega', \mathscr{A}', P') \to (\Omega, \mathscr{A}, P)$ between probability spaces, such that for all events $A \in \mathscr{A}$ it holds that

$$X^{-1}(A) \in \mathscr{A} \qquad \text{and} \qquad P(A) = P'(X^{-1}(A)).$$

We also write $P(X \in A) := P'(X^{-1}(A))$ and call it the **probability distribution** of X.

If $\Omega = \mathbb{R}$ and $\mathscr{A}$ is the Borel σ-algebra, we call X a **continuous real random variable**. If $\Omega \subset \mathbb{R}$ is discrete and $\mathscr{A} = 2^\Omega$, we call X a **discrete real random variable**.

The definition of a random variable X is rather technical. What is it good for? The definition of a probability space in Definition 1.11 introduces the probability measure of *sets* in Ω. By contrast, one should think of a random variables as *random elements* in Ω. Often Ω is $\mathbb{R}$ or $\mathbb{R}^n$ so that a random variable $X \in R$ represents a random real number and $X \in R^n$ is a random real vector.

Example 1.15. Suppose that Ω is the set of all coin tosses. Let $X : \Omega \to \{0,1\}$ be a random variable with $P(X = 0) = P(X = 1) = \frac{1}{2}$. Then, $P(X = 0)$ can be interpreted as the probability that the coin lands on heads, and $P(X = 1)$ as the probability that the coin lands on tails.

Given a continous real random variable $X \in \mathbb{R}$ every measurable function $g : \mathbb{R} \to \mathbb{R}$ induces another random variable $Y := g(X)$ with $P(Y \in A) := P(X \in g^{-1}(A))$.

In the following, we fix a probability space $(\Omega, \mathscr{A}, P)$. Let $A, B \in \mathscr{A}$. We want to describe the probability of A in the situation when we already know that B has happened. This probability is denoted $P(A \mid B)$. It is reasonable to require $P(A \mid B)$ to be proportional to $P(A \cap B)$ and $P(B \mid B) = 1$. This motivates the following definition.

Definition 1.16. Assume $P(B) > 0$. The **conditional probability** of A given B is

$$P(A \mid B) := \frac{P(A \cap B)}{P(B)}.$$

Example 1.17. Let $\Omega = \{1, \ldots, 6\}$, $A = \{2\}$ and $B = \{2, 4, 6\}$. Suppose that $P(\{k\}) = \frac{1}{6}$ for $k = 1, \ldots, 6$. Then:

$$P(A) = \frac{1}{6} \qquad \text{and} \qquad P(A \mid B) = \frac{1}{3}.$$

In other words, when all the 6 numbers are equally likely, it is more likely to draw number 2, if we know that only even numbers will be drawn.

Theorem 1.18 (Bayes' theorem)**.** *Let $A, B \in \mathscr{A}$ with $P(A), P(B) > 0$. Then,*

$$P(A \mid B) = P(B \mid A) \cdot \frac{P(A)}{P(B)}.$$

Proof. By Definition 1.16 we have $P(A \mid B) = \frac{P(A \cap B)}{P(B)}$ and $P(B \mid A) = \frac{P(A \cap B)}{P(A)}$. This implies $P(A \mid B)P(B) = P(B \mid A)P(A)$, from which the statement follows. □

There is an interesting consequence of Bayes' theorem. Namely, $P(A \mid B) > P(A)$, if and only if $P(B \mid A) > P(B)$. In other words, B makes A more likely, if and only if A makes B more likely.

Definition 1.19. Let $A, B \in \mathscr{A}$ with $P(B) \neq 0$. We call A and B **independent**, if

$$P(A \mid B) = P(A).$$

We call two continuous (resp. discrete) real random variables X and Y **independent**, if

$$P(X \in A \text{ and } Y \in A) = P(X \in A)P(Y \in A)$$

for all events $A \in \mathscr{A}$. We say that a sequence of continuous (resp. discrete) real random variables $(X_n)_{n \in \mathbb{N}}$ are **independent and identically distributed** (abbreviated as **i.i.d.**), if the X_n are pairwise independent and all have the same probability distribution.

Let now $X \in \mathbb{R}$ be a real random variable. If X is discrete and $X(\Omega) = \{x_1, \ldots, x_n\}$ is the range of discrete values that X can admit, its probability distribution P is completely determined by the values $P(X = x_i)$ for $i = 1, \ldots, n$. If X is continuous, the probability distribution of X is not so easy to describe. In many cases, however, the probability distribution can be given by a so-called probability density.

Definition 1.20. Let $X \in \mathbb{R}$ be a continuous real random variable. An integrable function $f : \mathbb{R} \to \mathbb{R}_{\geq 0}$ is called a **probability density** of X, if for all events A we have

$$P(X \in A) = \int_A f(x)\,\mathrm{d}x.$$

In particular, $\int_{\mathbb{R}} f(x)\,\mathrm{d}x = 1$. If $X = (X_1, \ldots, X_n) \in \mathbb{R}^n$ is a continuous real random vector, we call a function $f : \mathbb{R}^n \to \mathbb{R}_{\geq 0}$ with

$$P(X \in A) = \int_A f(x)\,\mathrm{d}x$$

the **joint density** of the X_i, or simply the probability density of X.

Again recall all random variables have a probability distribution, but not all all random variables have densities.

The interpretation of a probability density is that $f(x)$ measures the "infinitesimal probability" of $x \in \mathbb{R}$. We will often denote the probability density by $P_X(x) := f(x)$ or $P(x) := f(x)$. The only time this becomes confusing is when we have the singleton $A = \{x_0\}$, in which case the probability of any single event occurring for a continuous random variable is always zero:

$$P(\{x_0\}) = \int_{\{x_0\}} P(x)\,dx = \int_{\{x_0\}} f(x)\,dx = 0,$$

while $P(x)$ does not need to be zero.

Suppose now that $X \in \mathbb{R}^n$ is a continuous random variable with a density. While the probability that $X = x$ for a single point $x_0 \in \mathbb{R}^n$ is zero, we can still express the conditional probability distribution given $X = x$.

Definition 1.21. Let $(X, Y) \in \mathbb{R}^n \times \mathbb{R}^m$ be a random variable with a probability density $P_{(X,Y)}$ and $x_0 \in \mathbb{R}^n$. The **conditional density** of Y given $X = x$ is

$$P_{Y|X=x}(y) = \frac{P_{(X,Y)}(x, y)}{P_X(x)}.$$

We write $Y \mid X = x$ for the random variable with this density.

To see that the right-hand side in Definition 1.21 is indeed a density observe that

$$P_X(x) = \int_{\mathbb{R}^m} P_{(X,Y)}(x,y)\,\mathrm{d}y.$$

We also have Bayes' theorem for conditional densities.

Theorem 1.22 (Bayes' theorem for densities)**.** *Let $(X,Y) \in \mathbb{R}^n \times \mathbb{R}^m$ be a random variable with a probability density $P_{(X,Y)}$ and $x \in \mathbb{R}^n$ and $y \in \mathbb{R}^m$. Then:*

$$P_{Y|X=x}(y) = P_{X|Y=y}(x) \cdot \frac{P_Y(y)}{P_X(x)}.$$

Proof. This follows immediately from Definition 1.21. □

Next, we introduce several important properties of real random variables.

Definition 1.23. Let $X \in \{x_1,\ldots,x_n\}$ be a discrete real random variable. The **expected value** of X is

$$\mathbb{E}X := \sum_{i=1}^{n} x_i \cdot P(X = x_i).$$

If $X \in \mathbb{R}$ is a continuous real random variable with a density P its **expected value** is

$$\mathbb{E}X := \int_{\mathbb{R}} x \cdot P(x)\,\mathrm{d}x.$$

In both cases, the **variance** is defined as

$$\mathrm{Var}(X) := \mathbb{E}(X - \mathbb{E}X)^2.$$

The **standard deviation** is $s(X) := \sqrt{\mathrm{Var}(X)}$. Let X and Y be two continuous (resp. discrete) real random variables. The **covariance** of X and Y is

$$\mathrm{Cov}(X,Y) := \mathbb{E}(X - \mathbb{E}X)(Y - \mathbb{E}Y).$$

Lemma 1.24. *Let X be a continuous (resp. discrete) real random variable.*

1. *Let $g : \mathbb{R} \to \mathbb{R}$. If $X \in \{x_1,\ldots,x_n\}$ is discrete, then*

$$\mathbb{E}g(X) = \sum_{i=1}^{n} g(x_i) \cdot P(X = x_i).$$

2. *If X is continuous with density P and g is measurable, then*

$$\mathbb{E}\,g(X) = \int_{\mathbb{R}} g(x)\cdot P(x)\,\mathrm{d}x,$$

provided $\int_{\mathbb{R}} |g(x)|\cdot P(x)\,\mathrm{d}x < \infty$.

Proof. We denote the random variable $Z := g(X)$. In the discrete case we set $z_i = g(x_i)$. Then, $P(Z = g(x_i)) = P(X = x_i)$ and therefore

$$\mathbb{E}Z = \sum_{i=1}^{n} z_i \cdot P(Z = z_i) = \sum_{i=1}^{n} g(x_i)\cdot P(Z = g(x_i)) = \sum_{i=1}^{n} g(x_i)\cdot P(X = x_i).$$

The continuous case requires some ideas from measure theory, which we skip here. We refer to [Ash70, Section 3, Theorem 2] for a proof. □

Lemma 1.25 (Linearity of the expected value)**.** *Let X and Y be two real random variables with finite expected expectation: $\mathbb{E}X, \mathbb{E}Y < \infty$. Then, for all $a, b \in \mathbb{R}$ we have*

$$\mathbb{E}(aX + bY) = a\,\mathbb{E}X + b\,\mathbb{E}Y.$$

Proof. See, e.g., [Ash70, Section 3.3]. See also Exercise 1.5. □

Example 1.26. The following list of random variables describes important distributions.

1. **Bernoulli distribution:** $X \in \{0,1\}$ and $P(X = 0) = p$.

   We write $X \sim \mathrm{Ber}(p)$.

2. **Binomial distribution:** $X \in \{1,\dots,n\}$ and $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

   We write $X \sim \mathrm{Bin}(n,p)$.

   Notice that $P(X = k) = P(\{\#i \mid Z_i = 0, 1 \le i \le n\})$ for $Z_1,\dots,Z_n \sim \mathrm{Ber}(p)$ i.i.d.

3. **Discrete uniform distribution:** $X \in \{a_1,\dots,a_n\}$ and $P(X = k) = \frac{k}{n}$.

   We write $X \sim \mathrm{Unif}(\{a_1,\dots,a_n\})$.

4. **Continuous uniform distribution:** $X \in [a,b]$ and $P(A) = \int_A \frac{1}{b-a}\,\mathrm{d}x$.

   We write $X \sim \mathrm{Unif}([a,b])$.

5. **Normal distribution:** $X \in \mathbb{R}$ and

$$P(A) = \frac{1}{\sqrt{2\pi\sigma^2}} \int_A \exp\left(-\tfrac{1}{2\sigma^2}(x-\mu)^2\right)\mathrm{d}x,$$

   where $\sigma^2 > 0$ and $\mu \in \mathbb{R}$.

   We write $X \sim N(\mu, \sigma^2)$.

6. **Multivariate normal distribution:** $X \in \mathbb{R}^n$ and

$$P(A) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \int_A \exp(-\tfrac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))\,\mathrm{d}x$$

for $\Sigma \in \mathbb{R}^{n\times n}$ positive semi-definite and $\mu \in \mathbb{R}^n$.

We write $X \sim N(\mu, \Sigma)$.

We further write $N(x \mid \mu, \Sigma) = \frac{1}{\sqrt{2\pi \det(\Sigma)}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu))$ for the density of X.

The next lemma explains why Σ is called **covariance matrix**.

Lemma 1.27. *Let $X \sim N(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then,*

$$\mathbb{E}X = \mu \qquad \textit{and} \qquad \mathrm{Var}(X) = \sigma^2.$$

Furthermore, if $Y \sim N(\nu, \Sigma)$ with $\nu \in \mathbb{R}^n$ and Σ positive semi-definite, we have

$$\mathrm{Cov}(Y_i, Y_j) = \Sigma_{i,j}$$

for all $1 \le i, j \le n$.

Exercise 1.5. Prove Lemma 1.25. **Hint:** Use Lemma 1.24 for $g(X,Y) = X+Y$, the fact that $P(X \in A) = P((X,Y) \in A \times \mathbb{R})$ and $P(Y \in B) = P((X,Y) \in \mathbb{R} \times B)$, and that for every random variable $\mathbb{E}|X| < \infty$ if and only if $\mathbb{E}X < \infty$ (see [Ash70, Eq. (3.1.7)]).

Exercise 1.6. Prove Lemma 1.27.

Exercise 1.7. Let $\Omega := \{x_1, \dots, x_n\}$ and $p_1, \dots, p_n \ge 0$ with $p_1 + \cdots + p_n = 1$. Prove that the following algorithm generates a random variable $X \in \Omega$ with $P(X = x_i) = p_i$:

1. define the numbers $w_k := \sum_{i=1}^k p_i$, $1 \le k \le n$, and $w_0 := 0$ and $w_{n+1} := 1$;
2. draw $Y \sim \mathrm{Unif}([0,1])$

   (for instance, in `Julia` one can draw Y using the command `rand()`);
3. let k such that $w_{k-1} \le Y < w_k$;
4. return x_k.

Bibliography

[Ash70] Robert B. Ash. *Basic probability theory*. John Wiley & Sons, Inc., New York-London-Sydney, 1970.

[BEKS17] Jeff Bezanson, Alan Edelman, Stefan Karpinski, and Viral Shah. Julia: A fresh approach to numerical computing. *SIAM Review*, 59(1):65–98, 2017.

[Chu97] Fan R. K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI, 1997.

[Chu10] Fan Chung. Graph theory in the information age. *Notices Amer. Math. Soc.*, 57(6):726–732, 2010.

[KB09] Tamara G. Kolda and Brett W. Bader. Tensor decompositions and applications. *SIAM Rev.*, 51(3):455–500, 2009.

[Str93] Gilbert Strang. The fundamental theorem of linear algebra. *Amer. Math. Monthly*, 100(9):848–855, 1993.

[UV20] André Uschmajew and Bart Vandereycken. Geometric methods on low-rank matrix and tensor manifolds. In *Handbook of variational methods for nonlinear geometric data*, pages 261–313. Springer, Cham, [2020] ©2020.