Mathematical Methods in Data Science

Paul Breiding, Alessio D'Alì and Samantha Fairchild

Author's addresses:
Paul Breiding, University of Osnabrück + MPI MiS Leipzig, pbreiding@uni-osnabrueck.de.
Alessio D'Alì, University of Osnabrück, alessio.dali@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 authors gave the class ”Mathematische Grundlagen der Datenanalyse” at the University of Osnabrück. The notes are intended for a course on an advanced Bachelor level. A main theme of these notes is studying the **geometry** of problems in data science.

The first and third 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:

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

is an $m \times n$ **matrix** with real (resp. complex) entries a_{ij} for $1 \le i \le m$, $1 \le j \le n$. 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 $(x, y) \mapsto y^T Ax$.

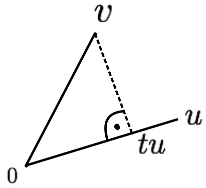


Figure 1.1: The meaning of the inner product between u and v is illustrated in this picture: let $t \in \mathbb{R}$ such that $v = tu + u'$, where u' is orthogonal to u. Then, $\langle u, v\rangle = u^T(tu+u') = tu^Tu = t\langle u, u\rangle$. In particular, if $\langle u, u\rangle = \langle v, v\rangle = 1$, 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. $\operatorname{Im}(A) := \{Ax \mid x \in \mathbb{R}^n\} \subseteq \mathbb{R}^m$,
2. $\operatorname{Im}(A^T) := \{A^Ty \mid y \in \mathbb{R}^m\} \subseteq \mathbb{R}^n$,
3. $\ker(A) := \{x \in \mathbb{R}^n \mid Ax = 0\} \subseteq \mathbb{R}^n$,
4. $\ker(A^T) := \{y \in \mathbb{R}^m \mid A^Ty = 0\} \subseteq \mathbb{R}^m$

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 $\langle a, b\rangle := a^Tb$. For $A = [a_1, \dots, a_n]$, we have $A^Ty = [\langle a_i, y\rangle]_{i=1}^n$.

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. $\operatorname{Im}(A) \oplus \ker(A^T) = \mathbb{R}^m$*,*

2. $\operatorname{Im}(A) \perp \ker(A^T)$*,*

3. $\operatorname{Im}(A^T) \oplus \ker(A) = \mathbb{R}^n$*,*

4. $\operatorname{Im}(A^T) \perp \ker(A)$*.*

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

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

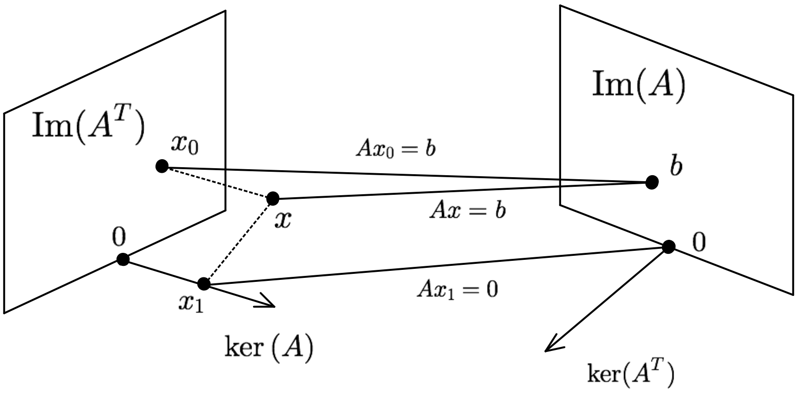


Figure 1.2: The situation when $b \in \mathrm{Im}(A)$: in this case, $Ax = b$ has a unique solution $x \in \mathrm{Im}(A^T)$ and the solution space for $Ax = b$ is $x + \ker(A)$.

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

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

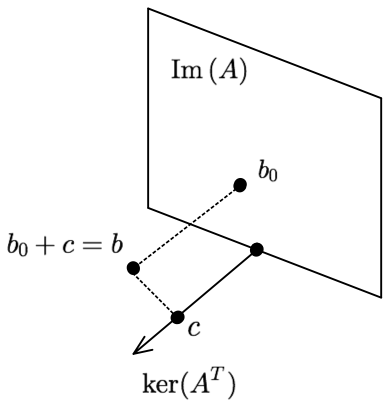


Figure 1.3: Visualization of the proof of Lemma 1.4: b_0 minimizes the distance from b to $\mathrm{Im}(A)$.

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

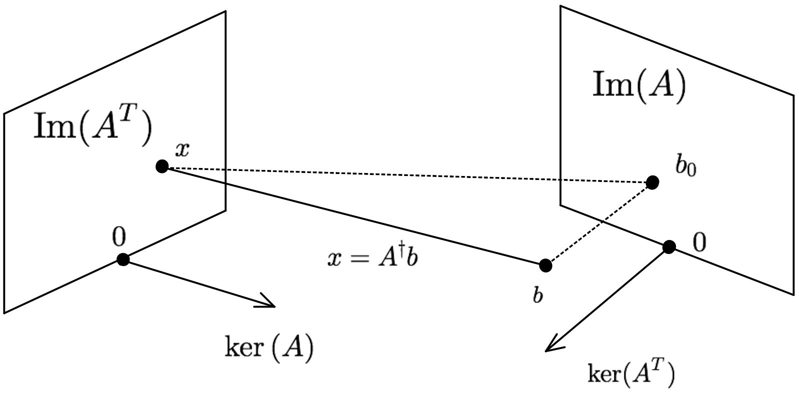


Figure 1.4: The pseudoinverse $A^\dagger \in \mathbb{R}^{m\times n}$ of $A \in \mathbb{R}^{m\times n}$ first orthogonally projects $b \in \mathbb{R}^n$ to $b_0 \in \mathrm{Im}(A)$ and then maps b_0 to the unique point $x \in \mathrm{Im}(A^T)$ with $Ax = b_0$.

derivative and setting it equal to zero. Namely we want to compute when the gradient $\frac{\mathrm{d}}{\mathrm{d}x}\|\phi(x)\| = \left[\frac{\partial}{\partial x_1}\|\phi(x)\|, \dots, \frac{\partial}{\partial x_n}\|\phi(x)\|\right] \in \mathbb{R}^n$ 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 $\sigma_1 > \dots > \sigma_r > 0$, 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

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

because $V^TV = \tilde{V}^T\tilde{V} = \mathbf{1}_r$. Let us write $U = [u_1,\ldots,u_r]$ and $\tilde{U} = [\tilde{u}_1,\ldots,\tilde{u}_r]$. The above equations imply that $AA^Tu_i = \sigma_i^2u_i$ and $AA^T\tilde{u}_i = \tilde{\sigma}_i^2\tilde{u}_i$ for $i = 1,\ldots,r$. Using that $r = r(A) = r(AA^T)$ we conclude that both $\sigma_1,\ldots,\sigma_r$ and $\tilde{\sigma}_1,\ldots,\tilde{\sigma}_r$ are the nonzero eigenvalues of AA^T. Since eigenvalues are unique, this implies $\sigma_i = \tilde{\sigma}_i$, $i = 1,\ldots,r$. 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 $i = 1,\ldots,r$. 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 $S = \operatorname{diag}(\sigma_1,\ldots,\sigma_r,0,\ldots,0)$. 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

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

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 a probability might not even exist. 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 $X \in \mathbb{R}$ represents a random real number and $X \in \mathbb{R}^n$ 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 $\Omega = \{1, \dots, 6\}, A = \{2\}$ and $B = \{2, 4, 6\}$. Suppose that $P(\{k\}) = \frac{1}{6}$ for $k = 1, \dots, 6$. 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 $X(\Omega) = \{x_1, x_2, \ldots\}$ 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 $x \in \mathbb{R}^n$. 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

$$P_X(x) = \int_{\mathbb{R}^m} P_{(X,Y)}(x,y)\,\mathrm{d}y, \tag{1.2.1}$$

since $P(X \in A) = P((X,Y) \in A \times \mathbb{R})$. Here, P_X is called the **marginal density**.

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 $X \in \{x_1, x_2, \dots\}$ be a discrete real random variable. The **expected value** of X is

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

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

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

In particular, $\mathrm{Var}(X) = \mathrm{Cov}(X,X)$.

Lemma 1.24 (Linearity of the expected value)**.** *Let X and Y be two real random variables with finite expected values: $\mathbb{E}X, \mathbb{E}Y < \infty$. 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. □

Linearity of the expected value implies

$$\operatorname{Var} X = \mathbb{E} X^2 - (\mathbb{E} X)^2 \qquad \text{and} \qquad \operatorname{Cov}(X,Y) = \mathbb{E} XY - \mathbb{E} X\, \mathbb{E} Y.$$

Lemma 1.25. *Let $X \in \mathbb{R}^n$ be a random variable and $g : \mathbb{R}^n \to \mathbb{R}$.*

1. *If $X \in \{x_1, x_2, \ldots\}$ is discrete, then*
$$\mathbb{E}\, g(X) = \sum_{i=1}^{\infty} g(x_i) \cdot P(X = x_i).$$
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 $\int_{\mathbb{R}^n} |g(x)| \cdot P(x)\, \mathrm{d}x < \infty$.

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

$$\mathbb{E} Z = \sum_{i=1}^{\infty} z_i \cdot P(Z = z_i) = \sum_{i=1}^{\infty} \sum_{k : g(x_k) = z_i} g(x_k) P(X = x_k) = \sum_{k=1}^{\infty} g(x_k) \cdot P(X = x_k).$$

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 implies the following expressions for covariance of random variables $(X,Y) \in \mathbb{R}^2$ with joint density P:

$$\operatorname{Cov}(X,Y) = \int_{\mathbb{R}} xy\, P(x,y)\, \mathrm{d}(x,y) - \mathbb{E} X\, \mathbb{E} Y.$$

Lemma 1.26. *Let $X \in \mathbb{R}^n$ and $Y \in \mathbb{R}^m$ be random variables, and suppose that Y has a density $P_Y(y)$ and that $(X \mid Y)$ has a density $P_{X|Y=y}(x)$. Then,*

$$\mathbb{E}_X X = \mathbb{E}_Y\, \mathbb{E}_{X|Y=y} X.$$

Proof. By Eq. (1.2.1), the density of X is given by $P_X(x) = \int_{\mathbb{R}^m} P_{X|Y=y}(x) P_Y(y)\, \mathrm{d}y$. This implies

$$\begin{aligned}
\mathbb{E}_X X &= \int_{\mathbb{R}^n} x \cdot P_X(x)\, \mathrm{d}x \\
&= \int_{\mathbb{R}^n} \int_{\mathbb{R}^m} x \cdot P_{X|Y=y}(x) \cdot P_Y(y)\, \mathrm{d}y\, \mathrm{d}x \\
&= \int_{\mathbb{R}^m} \left(\int_{\mathbb{R}^n} x \cdot P_{X|Y=y}(x)\, \mathrm{d}x \right) \cdot P_Y(y)\, \mathrm{d}y \\
&= \mathbb{E}_Y\, \mathbb{E}_{X|Y=y} X.
\end{aligned}$$

□

Example 1.27. 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:** $X \in \{0,\dots,n\}$ and $P(X=k)=\binom{n}{k}p^k(1-p)^{n-k}$.

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

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

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 $P(A)=\int_A \frac{1}{b-a}\,\mathrm{d}x$ for $A \subseteq [a,b]$.

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

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

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

   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

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

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

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

   We further write

   $$\Phi(x \mid \mu,\Sigma)=\tfrac{1}{\sqrt{(2\pi)^n\det(\Sigma)}}\exp(-\tfrac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)) \tag{1.2.2}$$

   for the density of X.

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

Lemma 1.28. *Let $X \sim N(\mu, \sigma^2)$ for $\mu \in \mathbb{R}$ and $\sigma^2 > 0$. Then,*

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

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

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

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

Lemma 1.29. *Let $X \sim N(\mu, \Sigma)$ and $Y \sim N(\nu, S)$ be independent Gaussian random variables in $\mathbb{R}^n$, and let $A, B \in \mathbb{R}^{m \times n}$ and $b \in \mathbb{R}^m$. Then,*

$$AX + BY + b \sim N(A\mu + B\nu + b, A\Sigma A^T + BSB^T).$$

Exercise 1.5. Prove Lemma 1.24 (for the continuous case you can assume that (X,Y) has a joint density). **Hint:** Use Lemma 1.25 for $g(X,Y) = X + Y$ 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. Let $\Omega := \{x_1, \ldots, x_n\}$ and $p_1, \ldots, p_n \geq 0$ 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 $w_k := \sum_{i=1}^k p_i$, $1 \leq k \leq n$, 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 $w_{k-1} \leq Y < w_k$;
4. return x_k.

Exercise 1.7. Prove Lemma 1.28.

Exercise 1.8. Prove Lemma 1.29. **Hint:** Prove first the case $B = 0$ by computing the density of $AX + b$. Then, use that

$$AX + BY = \begin{bmatrix} A & B \end{bmatrix} \begin{bmatrix} X \\ Y \end{bmatrix}.$$

Exercise 1.9. The element Caesium-137 has a half-life of about 30.17 years. In other words, a single atom of Caesium-137 has a 50 percent chance of surviving after 3.,17 years, a 25 percent chance of surviving after 60.34 years, and so on.

(a) Determine the probability that a single atom of Caesium-137 decays (i.e., does not survive) after a single day. How would you model the random variable X that takes the value 1 when the atom decays and 0 otherwise?

(b) Using `Julia`, simulate 1000 times the behaviour of a collection C of 10^6 Caesium-137 atoms in a single day. How would you model the following random variable?

$$Y = \# \text{ atoms in } C \text{ decaying after a single day}$$

(c) The *Poisson distribution* with parameter λ is a discrete probability distribution that is used to "model rare events". When $Z \sim \text{Pois}(\lambda)$, one has that

$$P\{Z = k\} = \frac{\lambda^k}{k!} e^{-\lambda}.$$

Plot the Poisson distribution with $\lambda = 10^6 \cdot p$, where p is the probability computed in part (a).

(d) Compare the empirical distribution in part (b) to the theoretical distribution in (c). Some `Julia` packages that might be useful: `Distributions`, `StatsPlots`.

2 Network Analysis

After the preliminaries we will now start the first chapter on mathematical methods in data science. Our first goal is to analyze structures of networks using spectral methods. We will mostly follow the book by Chung [Chu97], and the lecture notes by Guruswami and Kannan [GK12], and by Sauerwald and Sun [SS11]. For more context we also recommend [Chu10].

2.1 Graphs and the Laplace Matrix

In this section we follow the first chapter in [Chu97].

A network consist of a number of entities that are in relation to each other. Think of users in a social network that are connected, or airports for which there is a direct flight from one to another. The mathematical model for networks is a *graph*.

Definition 2.1. A graph $G = (V,E)$ is a pair consisting of a finite number of **vertices** given by

$$V = \{1,\dots,n\}$$

and a finite number of **edges** between pairs of vertices

$$E \subseteq \{\{i,j\} : i,j \in V, i \neq j\}.$$

When $v \in V$ and $e = \{u,v\} \in E$ for some $u \in V$ we say that u is **adjacent** to v. The **adjacency matrix** of G is

$$A(G) = (a_{ij}) \in \mathbb{R}^{n\times n}, \qquad \text{where } a_{ij} = \begin{cases} 1 & \{i,j\} \in E \\ 0 & \text{otherwise.} \end{cases}$$

The adjacency matrix $A(G)$ can be understood as a **data structure** for a graph.

Given a vertex $v \in V$, the **degree of** v is the number of vertices adjacent to v denoted

$$\deg(v) := \#\{u \in V \mid \{u,v\} \in E\}.$$

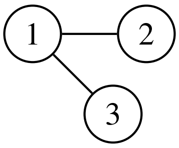
In the following, we will only consider graphs $G = (V,E)$ that have no isolated vertices; i.e., we assume

$$\deg(u) > 0, \qquad \text{for all } v \in V.$$

Isolated vertices do not contribute to the network structure we want to analyze, which is why we want to ignore them. Detecting isolated graphs from the adjacency matrix $A(G)$ is straightforward, so that we can remove columns and rows corresponding to isolated vertices from $A(G)$ immediately.

Remark 2.2. The notation $\{i, j\}$ is used to denote an unordered set, so in particular $\{i, j\} = \{j,i\}$ which means we will be working with **simple** and **undirected** graphs.

Example 2.3. Consider $G = (V,E)$ for $V = \{1,2,3\}$ and $E = \{\{1,2\},\{1,3\}\}$



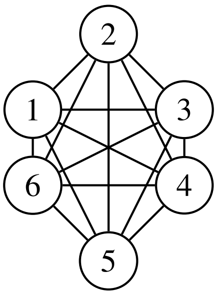
The adjacency matrix of this graph is

$$A(G) = \begin{pmatrix} 0 & 1 & 1 \\ 1 & 0 & 0 \\ 1 & 0 & 0 \end{pmatrix}.$$

The degrees are $\deg(1) = 2$ and $\deg(2) = \deg(3) = 1$.

Definition 2.4. We say $G = (V,E)$ is **complete** when $E = \{\{i, j\} : i, j \in V, i \neq j\}$.

Example 2.5. The following is a complete graph on 6 nodes.



Definition 2.6. A graph $G = (V,E)$ is said to be **bipartite** if the vertex set V can be subdivided into two disjoint subsets V_1 and V_2 so that every edge in E has an endpoint in V_1 and the other in V_2. If moreover every possible edge between V_1 and V_2 is present, G is a **complete bipartite** graph (note that there are several possible complete bipartite graphs on the same vertex set V).

Definition 2.7. A **walk** in G is a sequence of vertices

$$P = (v_0, v_1, \dots, v_D),$$

such that $\{v_{i-1}, v_i\} \in E$ for all $1 \leq i \leq D$. In this case, we say that P is a walk from v_0 to v_D. The edges of P are

$$E(P) = \{\{v_{i-1}, v_i\} \mid 1 \leq i \leq D\}.$$

The **length** of P is D. If $v_i \neq v_j$ for $i \neq j$ (no repeated vertices), we call P a **path**. We say that G is **connected**, if for every $v, w \in V$ there is a walk from v to w in G. A **connected component** of G is a maximal connected subgraph of G.

Lemma 2.8. *Let A be the adjacency matrix of a graph $G = (V,E)$, and let $v, w \in V$. Then the number of walks from v to w of length k is given by $(A^k)_{v,w}$.*

Proof. See Exercise 2.2. □

Next, we introduce the **Laplace matrix** or **Laplacian** of a graph G. We will see that its eigenvalues provide essential information about the network structure of G.

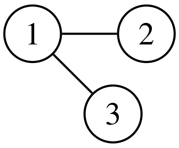
Definition 2.9. Let $G = (V,E)$ be a graph. The **Laplace Matrix** of G is

$$L(G) = (\ell_{ij}) \in \mathbb{R}^{|V| \times |V|},$$

where

$$\ell_{ij} = \begin{cases} 1 & i = j \\ \frac{-1}{\sqrt{\deg(i)\deg(j)}} & i \neq j \text{ and } \{i,j\} \in E \\ 0 & \text{otherwise} \end{cases}$$

Example 2.10. Consider the graph from Example 2.3 with $G = (V,E)$ and $V = \{1,2,3\}$ and $E = \{\{1,2\},\{1,3\}\}$:



Then

$$L(G) = \begin{pmatrix} 1 & -\frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} & 1 & 0 \\ -\frac{1}{\sqrt{2}} & 0 & 1 \end{pmatrix}$$

In the following, we fix a graph $G = (V,E)$ and denote $A := A(G)$ and $L := L(G)$.

Definition 2.11. We define the following diagonal matrix

$$T = (t_{uv}) \in \mathbb{R}^{|V|\times|V|}, \quad t_{uv} = \begin{cases} \deg(u) & u = v \\ 0 & \text{otherwise.} \end{cases}$$

Remark 2.12. Another common definition of the Laplacian of a graph is $\mathscr{L} := T - A$, where T is as in Definition 2.11 and A is the adjacency matrix of G. In fact, we have $L = T^{-1/2}\mathscr{L}T^{-1/2}$ (as shown in the next lemma). Compared to $\mathscr{L}$ our Laplacian is also called the **normalized Laplacian**. In our lecture we follow the definition in [Chu97] using L. In [Chu97, Section 1.2] Chung discusses that preferring L over $\mathscr{L}$ can be helpful in the context of stochastic processes - a topic that we will cover later in our lectures.

Lemma 2.13. *The following holds*

$$L = \mathbf{1}_{|V|} - T^{-1/2}AT^{-1/2}.$$

Proof. For $u \in V = \{1,\dots,n\}$ let $e_u = (0,\dots,0,1,0,\dots,0)^T$ the u-th standard basis vector. We compute for $u,v \in V$, and using the fact that T is symmetric so $T = T^T$

$$\begin{aligned}(T^{-1/2}AT^{-1/2})_{uv} &= e_u^T T^{-1/2}AT^{-1/2}e_v \\ &= (T^{-1/2}e_u)^T A(T^{-1/2}e_v) \\ &= \frac{1}{\sqrt{\deg(u)\deg(v)}}e_u^T Ae_v \\ &= \begin{cases} \frac{1}{\sqrt{\deg(u)\deg(v)}} & \{u,v\} \in E \\ 0 & \text{otherwise.} \end{cases}\end{aligned}$$

Since $\{u,u\} \notin E$, we conclude $L = \mathbf{1}_{|V|} - T^{-1/2}AT^{-1/2}$. □

The vector space $\mathbb{R}^{|V|}$ can be interpreted as the space of functions

$$\mathscr{F}(V) := \{f : V \to R\}$$

with the correspondence given by

$$x = (x_1, \ldots, x_{|V|}) \quad \leftrightarrow \quad f : V \to \mathbb{R},\, f(i) = x_i. \tag{2.1.1}$$

Then $L = L(G) \in \mathbb{R}^{|V|\times|V|}$ induces a linear mapping $\mathscr{F}(V) \to \mathscr{F}(V), f \mapsto Lf$. In this way, L is the linear map given by

$$Lf(v) = \sum_{j=1}^{n} \ell_{vj} f(j)$$

for $v \in V$.

Lemma 2.14. *The map L induced by the Laplacian of a graph $G = (V,E)$ is given by*

$$Lf(u) = \frac{1}{\sqrt{\deg(u)}} \sum_{v\in V:\{u,v\}\in E} \frac{f(u)}{\sqrt{\deg(u)}} - \frac{f(v)}{\sqrt{\deg(v)}}.$$

Proof. Let us write

$$g := T^{-1/2} f.$$

By Lemma 2.13 we have for $u \in V$:

$$Lf(u) = f(u) - (T^{-1/2}Ag)(u) = f(u) - \frac{1}{\sqrt{\deg(u)}} \sum_{v\in V} A_{uv} g(v).$$

Moreover,

$$\sum_{v\in V} A_{uv} g(v) = \sum_{v\in V:\{u,v\}\in E} g(v) = \frac{1}{\sqrt{\deg(v)}} \sum_{v\in V:\{u,v\}\in E} f(v).$$

This shows,

$$Lf(u) = f(u) - \sum_{v\in V:\{u,v\}\in E} \frac{f(v)}{\sqrt{\deg(u)\deg(v)}}. \tag{2.1.2}$$

We can write $\deg(v) = \sum_{u\in V:\{u,v\}\in E} 1$. Thus multiplying and dividing by $\deg(v)$ we can write

$$f(u) = \deg(u)\frac{f(u)}{\deg(u)} = \sum_{v\in V:\{u,v\}\in E} \frac{f(u)}{\deg(u)}. \tag{2.1.3}$$

Combining with Eq. (2.1.2) we then have

$$\begin{aligned}
Lf(u) &= f(u) - \sum_{v\in V:\{u,v\}\in E} \frac{f(v)}{\sqrt{\deg(u)\deg(v)}} && \text{(by Eq. (2.1.2))}\\
&= \sum_{v\in V:\{u,v\}\in E} \frac{f(u)}{\deg(u)} - \frac{f(v)}{\sqrt{\deg(u)\deg(v)}} && \text{(by Eq. (2.1.3))}\\
&= \frac{1}{\sqrt{\deg(u)}} \sum_{v\in V:\{u,v\}\in E} \frac{f(u)}{\sqrt{\deg(u)}} - \frac{f(v)}{\sqrt{\deg(v)}}
\end{aligned}$$

□

The Laplace Matrix is real and symmetric, $L = L^T$. Thus all eigenvalues of L are real.

Definition 2.15. The eigenvalues of L,

$$\lambda_0 \le \cdots \le \lambda_{|V|-1},$$

are called the **spectrum** of G. We define

$$\lambda_G = \lambda_1.$$

Example 2.16. The Laplace matrix from Example 2.10 has spectrum $0,1,2$.

The spectrum of a graph G encodes information about the structure of G as we will see in the following.

Definition 2.17. We define the following inner product on $\mathscr{F}(V)$:

$$\langle f,g\rangle := \sum_{u\in V} f(u)g(u).$$

We first investigate how L behaves relative to this inner product.

Theorem 2.18. *The **Rayleigh quotient** of L for $f \in \mathscr{F}(V)$ is*

$$\frac{\langle f,Lf\rangle}{\langle f,f\rangle} = \frac{1}{\sum_{u\in V} f(u)^2} \sum_{\{u,v\}\in E} \left(\frac{f(u)}{\sqrt{\deg(u)}} - \frac{f(v)}{\sqrt{\deg(v)}} \right)^2 .$$

Proof. By Lemma 2.14, we have

$$\langle f,Lf\rangle = \sum_{u\in V} f(u)\cdot Lf(u) = \sum_{u\in V} \frac{f(u)}{\sqrt{\deg(u)}} \sum_{v\in V:\{u,v\}\in E} \frac{f(u)}{\sqrt{\deg(u)}} - \frac{f(v)}{\sqrt{\deg(v)}}.$$

As above we set

$$g := T^{-1/2} f,$$

so that

$$\langle f, Lf \rangle = \sum_{u \in V} g(u) \sum_{v \in V : \{u,v\} \in E} g(u) - g(v).$$

We order the sum on the right as follows:

$$\begin{aligned} \langle f, Lf \rangle =& \frac{1}{2} \left(\sum_{u \in V} g(u) \sum_{v \in V : \{u,v\} \in E} g(u) - g(v) \right) \\ & - \frac{1}{2} \left(\sum_{v \in V} g(v) \sum_{u \in V : \{u,v\} \in E} g(u) - g(v) \right) \\ =& \sum_{\{u,v\} \in E} (g(u) - g(v))^2. \end{aligned} \tag{2.1.4}$$

Passing back to f coordinates, where we have $\sqrt{\deg(u)} g(u) = f(u)$, finally yields

$$\frac{\langle f, Lf \rangle}{\langle f, f \rangle} = \frac{1}{\sum_{u \in V} f(u)^2} \sum_{\{u,v\} \in E} \left(\frac{f(u)}{\sqrt{\deg(u)}} - \frac{f(v)}{\sqrt{\deg(v)}} \right)^2$$

as claimed. □

Theorem 2.18 shows that L defines a bilinear form $(f, g) \mapsto \langle f, Lg \rangle$ that is positive semi-definite. Consequently, the spectrum of G is always nonnegative. We give a formal proof for this observation.

Corollary 2.19. *We have $\lambda_i \geq 0$ for $i = 0, \ldots, |V| - 1$, and $\lambda_0 = 0$.*

Proof. Let λ be an eigenvalue of L with eigenvector $f \neq 0$. Then, by Theorem 2.18,

$$\lambda = \frac{\langle f, Lf \rangle}{\langle f, f \rangle} \geq 0.$$

Furthermore, let us consider the vector with $f(u) = \sqrt{\deg(u)}$. Then, again by Theorem 2.18, we have $\langle f, Lf \rangle = 0$, which shows that at least one eigenvalue is zero, so that $\lambda_0 = 0$. □

The proof shows that we always have

$$f = T^{1/2}e \in \ker L, \tag{2.1.5}$$

where $e \in \mathscr{F}(V)$ is the constant one function (in the identification from Eq. (2.1.1) this is $e = (1,\dots,1)$).

Next, we give the spectra of some example graphs.

Proposition 2.20. *Let G be a graph with $n = |V|$ vertices.*

1. *If G is the complete graph, then $\lambda_k = \frac{n}{n-1}$ for $k \geq 1$.*
2. *If G is a complete bipartite graph, then $\lambda_k = 1$ for $1 \leq k \leq |V|-2$ and $\lambda_{|V|-1} = 2$.*
3. *If G is a path, then $\lambda_k = 1 - \cos\frac{\pi k}{n-1}$.*
4. *If G is a cycle, then $\lambda_k = 1 - \cos\frac{2\pi k}{n}$.*

Exercise 2.1. Consider the complete graph on 6 vertices from Example 2.5. Construct the adjacency matrix and the Laplace matrix for this graph. What are the adjacency matrix and the Laplace matrix for a complete graph on n vertices?

Exercise 2.2. Prove Lemma 2.8.

Exercise 2.3. Let

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

Compute the $(1,1)$-entry of A^k for any $k \geq 1$ without computing the matrix power A^k explicitly.

Exercise 2.4. For the graph in Example 2.5, compute the number of paths of length 3 from vertex 1 to vertex 2.

Exercise 2.5. Prove Proposition 2.20.

2.2 The Spectrum of a Graph

In this lecture, G is a fixed graph with $n = |V|$ vertices and $L = L(G)$ is its Laplacian. Recall from the previous lecture that the spectrum of a graph $G = (V,E)$ is given by the eigenvalues of its Laplacian $L(G)$. We proved in Corollary 2.19 that these eigenvalues are nonnegative.

The main goal of this lecture is to prove the following theorem.

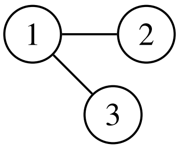
Theorem 2.21. *Let $G = (V,E)$ be a graph with $n = |V| \geq 2$, and let*

$$0 = \lambda_0 \leq \lambda_1 \leq \cdots \leq \lambda_{n-1}$$

be the spectrum of G. We also denote $\lambda_G := \lambda_1$. The following holds.

1. *$\lambda_0 + \lambda_1 + \cdots + \lambda_{n-1} = n$.*
2. *$\lambda_G \leq \frac{n}{n-1} \leq \lambda_{n-1}$.*
3. *If G is not complete, $\lambda_G \leq 1$. Otherwise, $\lambda_G = \frac{n}{n-1}$.*
4. *$\lambda_i = 0$ and $\lambda_{i+1} > 0$, if and only if G has exactly $i+1$ connected components.*
5. *We have $\lambda_{n-1} \leq 2$. Furthermore, $\lambda_{n-1} = 2$, if and only if there is a connected component in G that is bipartite.*
6. *The spectrum of G is the union of the spectra of its connected components.*

Example 2.22. Before we prove this theorem, let us recall the graph from Example 2.3:



From Example 2.16 we know that the spectrum of this graph is $\lambda_0 = 0, \lambda_1 = 1, \lambda_2 = 2$.

First, G is not complete, which can also be seen from $\lambda_1 = 1$ (see Theorem 2.21 1.). We have one connected component corresponding to $\lambda_0 = 0 < \lambda_1$ (see Theorem 2.21 4.). Finally, $\lambda_2 = 2$ as G is bipartite (see Theorem 2.21 5.).

Let us now prove Theorem 2.21.

Proof of Theorem 2.21. Let $L := L(G)$ be the Laplacian of G.

The first item follows because the diagonal entries of L are all equal to 1 (see Definition 2.9), so that $\lambda_0 + \lambda_1 + \cdots + \lambda_{n-1} = \mathrm{Trace}(L) = n$. Using that $\lambda_0 = 0$ this implies

$$n = \lambda_1 + \cdots + \lambda_{n-1} \geq (n-1)\lambda_G,$$

so that $\lambda_G \leq \frac{n}{n-1}$. In the same spirit,

$$n = \lambda_1 + \cdots + \lambda_{n-1} \leq (n-1)\lambda_{n-1},$$

so that $\lambda_{n-1} \geq \frac{n}{n-1}$. This proves the second item.

For the third item we recall from Proposition 2.20 that, if G is complete, $\lambda_G = \frac{n}{n-1}$. We show that otherwise $\lambda_G \leq 1$.

Recall from Eq. (2.1.5) that $T^{1/2}e \in \ker L$. As a consequence, using the Rayleigh quotient, λ_G can be written in the following way:

$$\lambda_G = \min_{g \in \mathscr{F}(V)\setminus\{0\}:\, \langle g, T^{1/2}e\rangle = 0} \frac{\langle g, Lg\rangle}{\langle g, g\rangle}. \tag{2.2.1}$$

If G is not complete, there exist $u, v \in V$ with $\{u,v\} \notin E$. Let us define $f \in \mathscr{F}(V)$ with

$$f(i) = \begin{cases} \sqrt{\deg(v)}, & \text{if } i = u \\ -\sqrt{\deg(u)}, & \text{if } i = v \\ 0, & \text{else.} \end{cases}$$

The function f satisfies

$$\langle f, T^{1/2}e\rangle = \sqrt{\deg(u)\deg(v)} - \sqrt{\deg(u)\deg(v)} = 0.$$

By Theorem 2.18, we have

$$\begin{aligned}\frac{\langle f, Lf\rangle}{\langle f, f\rangle} &= \frac{1}{\sum_{a\in V} f(a)^2} \sum_{\{a,b\}\in E} \left(\frac{f(a)}{\sqrt{\deg(a)}} - \frac{f(b)}{\sqrt{\deg(b)}} \right)^2 \\ &= \frac{1}{\deg(u) + \deg(v)} \left(\sum_{i\in V:\{u,i\}\in E} \frac{\deg(v)}{\deg(u)} + \sum_{i \in V:\{v,i\}\in E} \frac{\deg(u)}{\deg(v)} \right) \\ &= \frac{1}{\deg(u)+\deg(v)} \left(\deg(u)\frac{\deg(v)}{\deg(u)} + \deg(v)\frac{\deg(u)}{\deg(v)} \right) = 1.\end{aligned}$$

This shows $\lambda_G \leq 1$.

For item 4. we first make the following observation: let $f \in \ker L$. Then $\langle f, Lf\rangle = 0$. Writing $g := T^{-1/2}f$ and recalling Eq. (2.1.4), we infer that $g(u) = g(v)$ for all edges $\{u,v\} \in E$. Let now i, j be two vertices in G and P be a path in G from i to j. Since for all edges $\{u,v\}$ in G we have that $g(u) = g(v)$, it follows that $g(i) = g(j)$ and g is constant on the given path.

Assume now that G is connected: then, for every $i, j \in G$ we can find a path from i to j. It follows that g is a multiple of the constant one function e, and f is a multiple of $T^{1/2}e$. Consequently, 0 is a simple eigenvalue of L and $\lambda_1 > 0$. Conversely, if $\lambda_1 = 0$, then there exists a nonzero function f in the kernel of L which is not a multiple of $T^{1/2}e$. But then there must exist vertices $i, j \in G$ such that G contains no path from i to j, and hence G is not connected.

The statement for multiple connected components follows from this and item 6.

To prove item 5. note first that, for every $a, b \in \mathbb{R}$, one has that $(a-b)^2 \leq 2(a^2+b^2)$, and equality holds if and only if $b = -a$.

Now, setting $g = T^{-1/2}f$ and using again the expression of the Rayleigh quotient in Theorem 2.18, we get that

$$\lambda_{n-1} = \max_{f \in \mathscr{F}(V)\setminus\{0\}} \frac{\langle f, Lf \rangle}{\langle f, f \rangle} = \max_{f \in \mathscr{F}(V)\setminus\{0\}} \frac{1}{\sum_{u \in V} \deg(u) g(u)^2} \sum_{\{u,v\} \in E} (g(u) - g(v))^2.$$

Combining this with the above inequality and $\deg(u) = \sum_{v \in V : \{u,v\} \in E} 1$ yields

$$\lambda_{n-1} \leq \frac{2}{\sum_{u \in V} \deg(u) g(u)^2} \sum_{\{u,v\} \in E} (g(u)^2 + g(v)^2) = 2.$$

The only inequality used in the argument was $(g(u) - g(v))^2 \leq 2(g(u)^2 + g(v)^2)$. Therefore, $\lambda_{n-1} = 2$ if and only if there is a function $g \in \mathscr{F}(V) \setminus \{0\}$ with $g(u) = -g(v)$ for all $\{u, v\} \in E$. If G has a bipartite component $H = (V', E')$, we can write $V' = V_1' \sqcup V_2'$ and choose

$$g(u) = \begin{cases} 1, & \text{if } u \in V_1' \\ -1, & \text{if } u \in V_2' \\ 0, & \text{if } u \in V \setminus V' \end{cases}$$

to see that $\lambda_{n-1} = 2$. Conversely, if there exists a nonzero function $g \in \mathscr{F}(V)$ with $g(u) = -g(v)$ for all $\{u, v\} \in E$, let $H = (V', E')$ be a connected component of G on which g does not vanish. We define the subsets of vertices

$$W_1 := \{w \in W \mid g(w) > 0\} \quad \text{and} \quad W_2 := \{w \in W \mid g(w) < 0\}.$$

Then $V' = W_1 \cup W_2$ and, for every edge $\{u, v\} \in E'$, the endpoints u and v must lie in different W_i's. Therefore, H is bipartite.

Finally, for the last item we denote the connected components of G by $G_1, \ldots, G_k$. Let us write $G_i = (V_i, E_i)$, so that $V = \bigcup_{i=1}^k V_i$. We can reenumerate the vertices to have $V_i = \{n_{i-1}+1,\dots,n_i\}$ with $0 = n_0 < n_1 < \cdots < n_k = n$. Let also L_i be the Laplacian of G_i. Then, the Laplace matrix of G is a block diagonal matrix:

$$L(G) = \begin{bmatrix} L_1 & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & L_k \end{bmatrix}.$$

This shows that the eigenvalues of $L(G)$ are given by the eigenvalues of the L_i. □

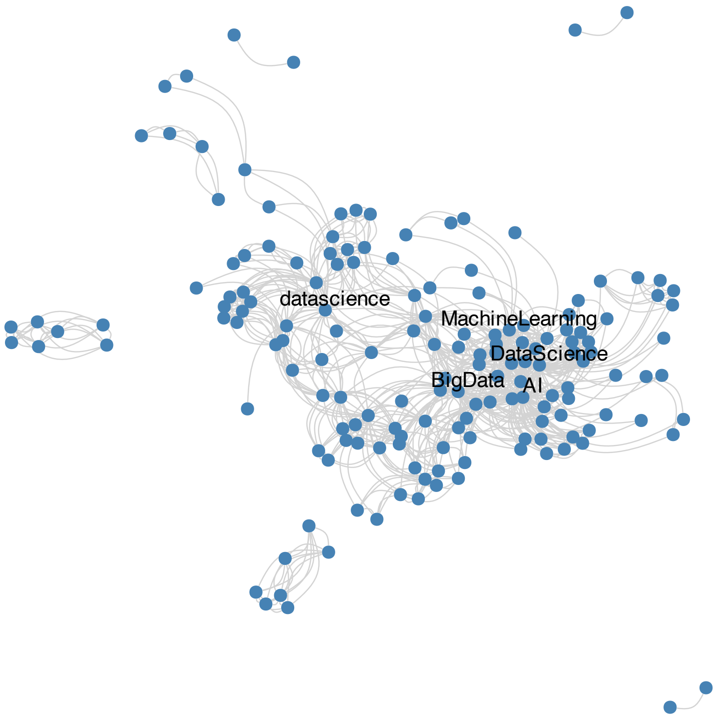


Figure 2.1: The network from Example 2.25. One can see 7 connected components. The 5 vertices with highest degree are labelled. The labels are the hashtags the vertices correspond to.

Theorem 2.21 shows how we can obtain information about the structure of a graph by computing its spectrum. However, often networks are almost disconnected or almost bipartite rather than having exactly this property. Such a scenario is also reflected in the spectrum. We need another definition for formulating results in this direction.

Definition 2.23. Let $G=(V,E)$ be a graph. The **volume** of G is

$$\operatorname{vol}(G) := \sum_{v\in V} \deg(v) = 2|E|.$$

Proposition 2.24. *Let $G=(V,E)$ be a graph, $n=|V|$, with λ_G, λ_{n-1} as in Theorem 2.21. Let $G_1=(V_1,E_1)$ and $G_2=(V_2,E_2)$ be two subgraphs with $V=V_1\cup V_2$, $V_1\cap V_2=\emptyset$ and $E_j=\{\{u,v\}\in E \mid u,v\in V_j\}, j=1,2$. Denote*

$$\varepsilon := \frac{|\{\{u,v\}\in E \mid u\in V_1, v\in V_2\}|}{|E|}.$$

the fraction of edges between V_1 and V_2. Then,

$$\lambda_G \le \varepsilon \frac{2}{(\operatorname{vol}(G_1)/|E|+\varepsilon)\cdot(\operatorname{vol}(G_2)/|E|+\varepsilon)} \le \lambda_{n-1}.$$

The meaning of the proposition is that, if λ_G is large, G can't be almost disconnected, and if λ_{n-1} is small, G can't be almost bipartite.

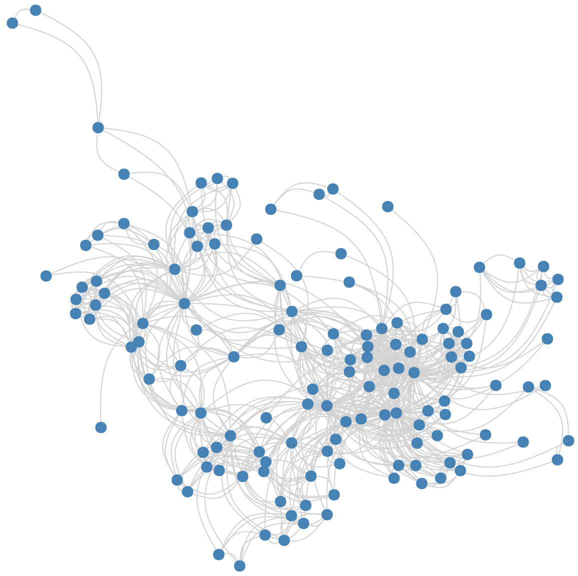


Figure 2.2: This graph shows the big component in Fig. 2.1, called G.

Let us also see how Proposition 2.24 is related to Theorem 2.18. The graph G is bipartite with components G_1 and G_2, if and only if $\operatorname{vol}(G_1)=\operatorname{vol}(G_2)=0$ and $\varepsilon=1$. In this case, the bound in Proposition 2.24 becomes $\lambda_G \leq 2 \leq \lambda_{n-1}$, similar to Theorem 2.21 5. Furthermore, the components G_1 and G_2 are disconnected, if and only if $\varepsilon = 0$, in which case we have $\lambda_G = 0$.

Example 2.25. We illustrate Proposition 2.24 in an example. We generate a graph with the following data. Using the `Julia` package `Twitter.jl` we download the 500 most recent tweets featuring the hashtag `#DataScience`. The vertices in this graph are all hashtags used in these tweets. We add an edge between two vertices if the two corresponding hashtags appear together in at least one tweet. This gives the graph on $n = 171$ vertices that can be seen in Fig. 2.1. We have labelled the 5 vertices with highest degrees in this graph with their corresponding hashtag.

The graph in Fig. 2.1 has 7 connected components. We consider the big component and call the underlying graph G. Fig. 2.2 shows G and Fig. 2.3 shows the spectrum of G.

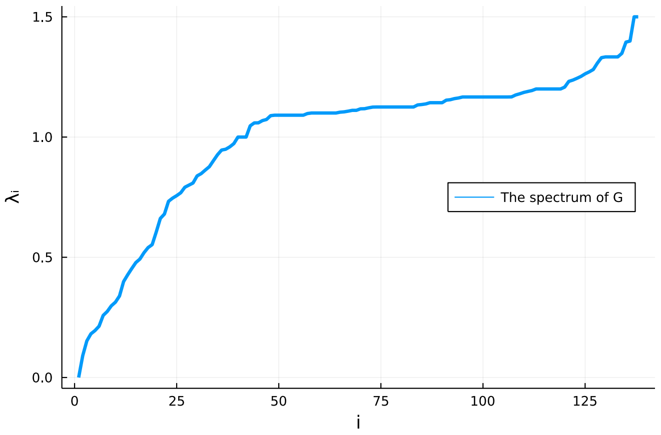


Figure 2.3: Spectrum of the graph from Fig. 2.2. The x-axis represents the index of the λ_i and the y-axis their numerical value. Instead of plotting discrete points we have plotted a piecewise linear curve connecting the discrete values $\lambda_0, \ldots, \lambda_{n-1}$.

Fig. 2.3 shows that λ_G is small. Following Proposition 2.24 we therefore anticipate[1] that $G = (V, E)$ has two components $G_1 = (V_1, E_1)$ and $G_2 = (V_1, E_1)$ such that there are few edges between them. The proof of Proposition 2.24 motivates the following clustering method: let f be the eigenfunction of λ_G. Then, we take

$$V_1 := \{v \in V \mid f(v) > 0\} \quad \text{and} \quad V_2 := \{v \in V \mid f(v) < 0\}.$$

[1] Strictly speaking, in Proposition 2.24 we only showed an upper bound for λ_G in terms of the number of connected edges ε, so that λ_G could be small while ε is big.

These two clusters are shown in Fig. 2.4. The vertices in V_1 are shown in red and the vertices in V_2 are blue. The two clusters indeed seem to give two relatively separated components of G.

We can use the same approach for analyzing the largest eigenvalue λ_{n-1}. Here, however, two cluters are not enough, because the eigenfunction g of λ_{n-1} defines *three* clusters: one cluster, where g is negative, one, where g is negative, and a third cluster, where g is zero. We show these three components in Fig. 2.5, where we have plotted the components corresponding to where g is zero in yellow. We see that the yellow vertices form the major part of G and that the blue and red vertices give a bipartite subgraph. The analysis of three components is the topic of Exercise 2.7.

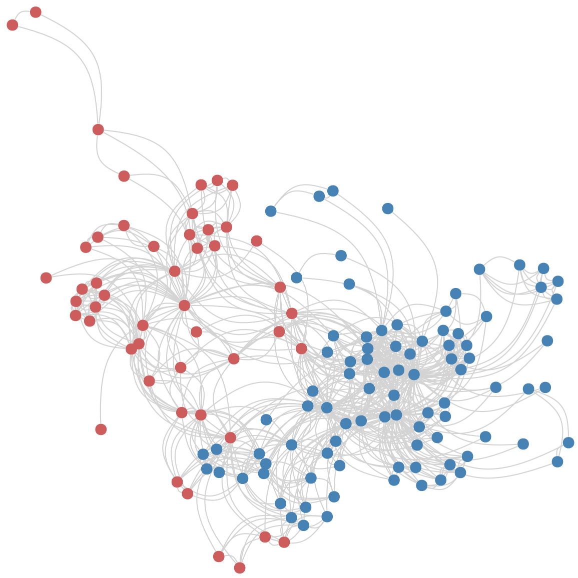


Figure 2.4: We have partitioned the vertices of the graph in Fig. 2.2 into two classes of vertices corresponding to whether the eigenfunction of λ_G is positive or negative.

Proof of Proposition 2.24. Let us denote $m_i := \mathrm{vol}(G_i) + \varepsilon$. Observe that for $i \neq j$:

$$\sum_{u\in V_i} \deg(u) = \sum_{u\in V_i} \left(\sum_{v\in V_i\colon \{u,v\}\in E} 1 + \sum_{v\in V_j\colon \{u,v\}\in E} 1 \right) = \mathrm{vol}(G_i) + \varepsilon|E| = m_i.$$

This also shows $m_1 + m_2 = \mathrm{vol}(G)$.

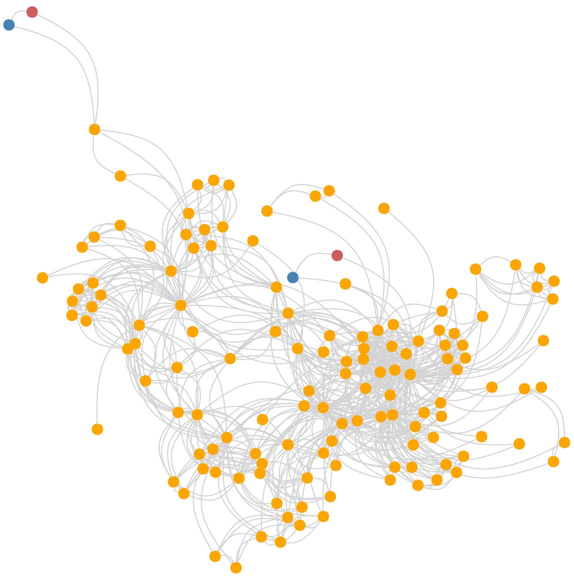


Figure 2.5: We have partitioned the vertices of the graph in Fig. 2.2 into three classes of vertices corresponding to whether the eigenfunction of λ_{n-1} is positive, negative or zero. The class corresponding to zero is yellow.

Let us define the function

$$f(u) = \begin{cases} m_2\sqrt{\deg(u)}, & \text{if } u \in V_1 \\ -m_1\sqrt{\deg(u)}, & \text{if } u \in V_2 \end{cases}.$$

Then,

$$\langle f, T^{1/2}e\rangle = m_2 \sum_{u\in V_1} \deg(u) - m_1 \sum_{u\in V_2} \deg(u) = m_2m_1 - m_1m_2 = 0,$$

and, by Theorem 2.18,

$$\begin{aligned} \frac{\langle f, Lf\rangle}{\langle f, f\rangle} &= \frac{1}{\sum_{u\in V} f(u)^2} \sum_{\{u,v\}\in E} \left(\frac{f(u)}{\sqrt{\deg(u)}} - \frac{f(v)}{\sqrt{\deg(v)}}\right)^2 \\ &= \varepsilon|E| \frac{(m_1+m_2)^2}{m_1m_2^2 + m_1^2m_2}. \end{aligned}$$

On the one hand, we have $\lambda_{n-1} = \max_{g \in \mathscr{F}(V)\setminus\{0\}} \frac{\langle g,Lg\rangle}{\langle g,g\rangle}$, and on the other hand, as in Eq. (2.2.1) we have that $\lambda_G = \min_{g \in \mathscr{F}(V)\setminus\{0\}:\, \langle g,T^{1/2}e\rangle = 0} \frac{\langle g,Lg\rangle}{\langle g,g\rangle}$. This shows that

$$\lambda_G \leq \frac{\langle f,Lf\rangle}{\langle f,f\rangle} = \varepsilon|E|\frac{\mathrm{vol}(G)}{m_1 m_2} \leq \lambda_{n-1}.$$

Finally, $|E|\frac{\mathrm{vol}(G)}{m_1 m_2} = 2(\mathrm{vol}(G_1)/|E|+\varepsilon)^{-1}(\mathrm{vol}(G_2)/|E|+\varepsilon)^{-1}$. □

The next two results underline that λ_G should be understood as a measure of connectivity for G.

Proposition 2.26. *Let $G = (V,E)$ be a connected graph and let* $\mathrm{diam}(G)$ *be the* ***diameter*** *of G; i.e.,* $\mathrm{diam}(G)$ *is the maximal length of all shortest paths between vertices $u, v \in V$. Then,*

$$\lambda_G \geq \frac{1}{\mathrm{diam}(G)\cdot \mathrm{vol}(G)}.$$

Proof. Let $f \in \mathscr{F}(V)$ be an eigenvector of λ_G with $\langle f,T^{1/2}e\rangle = 0$. Such an eigenvector exists by Eq. (2.1.5) (and since $L(G)$ is symmetric). Write $g := T^{-1/2}f$. Then,

$$0 = \langle f,T^{1/2}e\rangle = \langle T^{1/2}g,T^{1/2}e\rangle = \sum_{u\in V} \deg(u)g(u). \tag{2.2.2}$$

Let $v_0 \in V$ with $|g(v_0)| = \max_{v\in V}|g(v)|$. Eq. (2.2.2) implies that there exists $u_0 \in V$ with $g(u_0)g(v_0) < 0$ (i.e., they have opposite sign). If P is a shortest path from u_0 to v_0 of length $D > 0$, then

$$\frac{1}{D\cdot \mathrm{vol}(G)} \geq \frac{1}{\mathrm{diam}(G)\cdot \mathrm{vol}(G)}.$$

We show that $\lambda_G \geq (D\cdot \mathrm{vol}(G))^{-1}$. Using Theorem 2.18 we have

$$\begin{aligned}\lambda_G = \frac{\langle f,Lf\rangle}{\langle f,f\rangle} &= \frac{1}{\sum_{u\in V}\deg(u)g(u)^2}\sum_{\{u,v\}\in E}(g(u)-g(v))^2\\ &\geq \frac{1}{\mathrm{vol}(G)\,g(v_0)^2}\sum_{\{u,v\}\in E(P)}(g(u)-g(v))^2.\end{aligned}$$

Let us now denote the edges in P by $\{v_i, v_{i+1}\}$ for $i = 0,\ldots,D-1$, where $v_D = u_0$. We define the following vectors

$$\begin{aligned}&a = (1,\ldots,1)^T \in \mathbb{R}^D \quad \text{and}\\ &b = (g(v_1)-g(v_0), g(v_2)-g(v_1),\ldots,g(v_D)-g(v_{D-1}))^T \in \mathbb{R}^D.\end{aligned}$$

Then, we can use the Cauchy-Schwartz inequality to deduce that

$$D \cdot \sum_{\{u,v\} \in E(P)} (g(u) - g(v))^2 = \|a\|^2 \cdot \|b\|^2 \geq (a^T b)^2 = (g(v_0) - g(u_0))^2.$$

It follows that

$$\lambda_G \geq \frac{1}{D \cdot \mathrm{vol}(G)} \frac{(g(v_0) - g(u_0))^2}{g(v_0)^2},$$

and we have $(g(v_0) - g(u_0))^2 \geq g(v_0)^2$, because $g(u_0)g(v_0) < 0$. □

The final result of this lecture is a classical result in combinatorics known as Kirchhoff's theorem. From Theorem 2.21 we know that, if $G)$ is connected, $L(G)$ has rank $n-1$. Therefore, all $(n-1) \times (n-1)$ submatrices of $L(G)$ are invertible, and so have nonzero determinant. The next theorem shows that this determinant counts spanning trees in G. Recall that a tree is a graph that has no circles, and that a spanning tree τ is a tree-subgraph of $G = (V,E)$, such that the vertices of τ are V.

Theorem 2.27. *Let $G = (V,E)$ be connected, $L = L(G)$ be the Laplacian of G and $u \in V$. Let L_u be the submatrix of L that is obtained by removing the u-th row and u-th columns from L. Then,*

$$\det(L_u) = \frac{\#\, \textit{spanning trees in } G}{\prod_{v \in V : v \neq u} \deg(v)}.$$

Remark 2.28. Let $\mathscr{L} = T - A$ as in Remark 2.12. Theorem 2.27 implies that $\det(\mathscr{L}_u)$ is the number of spanning trees in G for every $u \in V$.

Proof. We follow the proof in [AZ18].

Let us define the matrix $S = (s_{eu}) \in \mathbb{R}^{|V| \times |E|}$ indexed by vertices times edges with

$$s_{u\{i,j\}} = \begin{cases} 0, & \text{if } u \neq i, u \neq j \\ \frac{1}{\sqrt{\deg u}}, & \text{if } u = i < j \\ \frac{-1}{\sqrt{\deg u}}, & \text{if } u = i > j. \end{cases}$$

It follows from Theorem 2.18 that $\langle S^T f, S^T f \rangle = \langle f, Lf \rangle$, which shows that $L = SS^T$. Therefore, if S_u denotes the matrix that is obtained from $S \in \mathbb{R}^{|E| \times (|V|-1)}$ by removing the u-th row, we have $L_u = S_u S_u^T$. As before, we let $n = |V|$. The Cauchy-Binet formula [HJ92, Sec. 0.8.7] implies

$$\det(L_u) = \sum_{\substack{B \in \mathbb{R}^{(n-1)\times(n-1)}: \\ B \text{ is a submatrix of } S_u}} \det(B)^2. \tag{2.2.3}$$

Let us consider a fixed $(n-1)\times(n-1)$ submatrix of S_u and call it B. The columns of B are labelled by $n-1$ edges in G. Let these edges be $E_1,\dots,E_{n-1}$, and let τ be the subgraph of G spanned by the E_i.

If τ is not a spanning tree, it must contain a circle; i.e., a walk $(v_0,v_1,\dots,)$ with $v_D = v_0$, $D \le n-1$. After relabeling we can assume that $E_i = \{v_{i-1},v_i\}$ for $1\le i\le D$. But then we can find the following linear combination of the first D columns of B:

$$\begin{bmatrix} \frac{1}{\sqrt{\deg}(v_0)} \\ \frac{-1}{\sqrt{\deg}(v_1)} \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix} + \begin{bmatrix} 0 \\ \frac{-1}{\sqrt{\deg}(v_1)} \\ \frac{1}{\sqrt{\deg}(v_2)} \\ \vdots \\ 0 \\ 0 \end{bmatrix} + \cdots + \begin{bmatrix} \frac{-1}{\sqrt{\deg}(v_0)} \\ 0 \\ 0 \\ \vdots \\ 0 \\ \frac{-1}{\sqrt{\deg}(v_D)} \end{bmatrix} = 0.$$

Therefore, the columns of B are linearly dependent and we have $\det(B) = 0$.

If, on the other hand, τ is a spanning tree, there must a exist a vertex $v_1 \in V\setminus\{u\}$ such that v_1 has degree 1 in τ. After relabeling we can assume that $v_1 \in E_1$ and $v_1 \notin E_2$. After deleting v_1 from τ, we get a tree with $n-2$ edges, for which we can repeat this argument. Therfore, after a permutation of the rows and columns, we can bring B into an upper triangular form with diagonal given by $(\deg(v)^{\frac{-1}{2}})_{v\in V\setminus\{u\}}$. This shows

$$\det(B)^2 = \frac{1}{\prod_{v\in V: v\neq u}\deg(v)}.$$

Since all spanning trees must be obtained as some choice of $n-1$ edges, this shows that the right hand side in Eq. (2.2.3) is equal to the number of spanning trees in G divided by $(\prod_{v\in V:v\neq u}\deg(v))^{-1}$. □

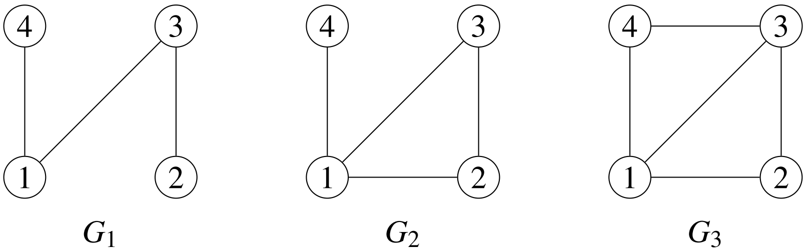
Exercise 2.6. It follows from Theorem 2.21 that G has exactly k connected components, if and only if $\dim\ker L(G) = k$. Can you determine the k components from $\ker L$?

Exercise 2.7. State and prove a version of Proposition 2.24 involving three pairwise distinct components of a graph. Prove bounds for λ_G and λ_{n-1} for a tripartite graph. **Hint:** Consider a function that is positive on the first components, negative on the second, and zero on the third component.

Exercise 2.8. Use `Twitter.jl` to generate a network, and analyze it using the spectral methods from this lecture.

Exercise 2.9. Compute the number of spanning trees in a complete graph.

Exercise 2.10. Let G_1, G_2, G_3 be the following graphs on vertex set $\{1,2,3,4\}$:



(a) List all spanning trees for G_1, G_2 and G_3. (You can either draw them or list their edges.) How many do you get?

(b) Compute the number of spanning trees for G_1, G_2 and G_3 using Theorem 2.27.

2.3 Markov Processes in Networks

In this section we partially follow the lecture notes by Guruswami and Kannan [GK12], and by Sauerwald and Sun [SS11].

In some situations the complexity of a network makes it computationally infeasible to treat all vertices and edges of the corresponding graph at once. As an alternative one can explore the network vertex by vertex using **random walks**. In this lecture we analyze networks using a special type of random walks, namely **Markov processes**. Again, eigenvalues will play a central role.

As before, $G = (V,E)$ denotes a graph with $V = \{1,\ldots,n\}$. We adopt the point of view from Eq. (2.1.1) that identifies vectors in $\mathbb{R}^n$ with functions $V \to \mathbb{R}$.

Definition 2.29. A **Markov process** X on G is a sequence of random variables

$$X_0, X_1, \ldots \in V,$$

called the **steps** of the walk, such that for all $i \geq 1$ we have:

1. $P(X_i = u \mid X_{i-1} = v, X_{i-2} = v_{i-2}, \ldots, X_0 = v_0) = P(X_i = u \mid X_{i-1} = v)$;
2. $P(X_i = u \mid X_{i-1} = v) > 0$ only if $\{u,v\} \in E$ or $u = v$ (i.e., remaining at the current vertex is allowed);
3. $P(X_i = u \mid X_{i-1} = v)$ does not depend on i.

The first item means that the probability law of the i-th step only depends on the position of the $(i-1)$-th step, but is independent of what happened before. The third item means that the probability law of a step does not depend on the number of steps that have passed. The second item means that we can only progress along edges in the graph or stand still. In the following, we will denote

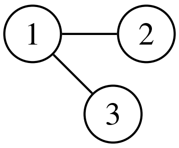
$$P(u \mid v) := P(X_i = u \mid X_{i-1} = v) \text{ for } i \geq 1.$$

Definition 2.30. Let X be a Markov process in G. The **transition matrix** of X is

$$P = (p_{uv}) \in \mathbb{R}^{n \times n}$$

with $p_{uv} = P(u \mid v)$. The p_{uv} are called **transition probabilities**.

Example 2.31. Consider $G = (V,E)$ for $V = \{1,2,3\}$ and $E = \{\{1,2\},\{1,3\}\}$:



Suppose the we start at the vertex 3. We have $P(2 \mid 3) = 0$, because there is no edge between 2 and 3. We can either move to 1 or stand still. This means that, if $p := P(1 \mid 3)$, then $P(3 \mid 3) = 1 - p$. Similarly, starting at 2 we can't move to 3. The transition matrix of any Markov process on G therefore has the form:

$$P = \begin{bmatrix} r & q & p \\ s & 1-q & 0 \\ 1-r-s & 0 & 1-p \end{bmatrix}$$

with $0 \leq p,q,r,s$ and $p,q,r+s \leq 1$.

Let us make a simple but important observation, implicitly used in Example 2.31.

Lemma 2.32. *Let $e = (1,\ldots,1) \in \mathbb{R}^n$ and let $P = (p_{uv}) \in \mathbb{R}^{n \times n}$ be the transition matrix of a Markov process on G. Then, e is an eigenvector of P^T with eigenvalue 1:*

$$P^T e = e.$$

Proof. We have $(P^T e)(v) = \sum_{u \in V} p_{uv} = \sum_{u \in V} P(u \mid v) = 1$. □

Definition 2.33. We call X a **uniform** Markov process, if its transition probabilities are

$$p_{uv} = P(u \mid v) = \begin{cases} \frac{1}{\deg(v)}, & \text{if } u \neq v \text{ and } \{u,v\} \in E \\ 0, & \text{else.} \end{cases}$$

That is, $P = AT^{-1}$, where A is the adjacency matrix of G and T is as in Definition 2.11.

Let us denote

$$\mathscr{F}_+(V) := \{f \in \mathscr{F}(V) \mid f(u) \geq 0 \text{ for all } u \in V\}. \tag{2.3.1}$$

A probability distribution f in V is given by a function $f \in \mathscr{F}_+(V)$ such that $\langle f, e\rangle = 1$, where as before e is the constant one function on V.

Lemma 2.34. *Let X be a Markov process on G with transition matrix P. Let also $i \geq 0$ and $f_i := (P(X_i = 1), \ldots, P(X_i = n))^T \in \mathscr{F}_+(V)$ be the probability distribution of the i-th step of X. Then,*

$$f_{i+k} = P^k f_i$$

for all $k \geq 0$.

Proof. Let $u \in V$. Then, we have

$$\begin{aligned}(Pf_i)(u) = \sum_{v \in V} p_{uv} \cdot f_i(v) &= \sum_{v \in V} P(X_{i+1} = u \mid X_i = v) \cdot P(X_i = v) \\ &= P(X_{i+1} = u) = f_{i+1}(u).\end{aligned}$$

This shows $Pf_i = f_{i+1}$. Consequently,

$$P^k f_i = P^{k-1}(Pf_i) = P^{k-1} f_{i+1} = \cdots = Pf_{i+k-1} = f_{i+k}.$$

□

Lemma 2.34 should be interpreted as follows: $(P^k)_{uv}$ is the probability that, if the Markov process starts at v, it reaches u after k steps.

Example 2.35. In Example 2.31 we take the transition matrix of the uniform process:

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

This means that, if we start at 3, the next vertex is almost surely 1. Then, from 1 with probability $\frac{1}{2}$ we either go back to 3 or go to 2. We also have

$$P^2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & \frac{1}{2} & \frac{1}{2} \\ 0 & \frac{1}{2} & \frac{1}{2} \end{bmatrix}.$$

The last column of P^2 shows that, starting from 3, after 2 steps we either are at 2 or 3, both with probability $\frac{1}{2}$.

Definition 2.36. Let X be a Markov process on G with transition matrix $P = (p_{uv})$.

1. We call a probability distribution $\pi : V \to \mathbb{R}$ **stationary** (with respect to X), if
$$P\pi = \pi.$$
2. A probability distribution $f : V \to \mathbb{R}$ is called **reversible** (with respect to X), if
$$f(v)p_{uv} = f(u)p_{vu}$$
for all $u, v \in V$.

We will spend a great part of the remainder of this lecture to give conditions under which stationary distributions exist and are unique. In the literature, a stochastic process that has a stationary distribution is also called an **ergodic** process. First, we give a sufficient condition for π being a stationary distribution of a Markov process X.

Proposition 2.37. *Let X be a Markov process on G and $\pi : V \to \mathbb{R}$ be a probability distribution reversible with respect to X. Then, π is a stationary distribution of X.*

Proof. Let P denote the transition matrix of X. For all $u \in V$ we have

$$(P\pi)(u) = \sum_{v \in V} p_{uv}\pi(v) = \sum_{v \in V} p_{vu}\pi(u) = \pi(u)\,(P^T e)(u) = \pi(u),$$

because $P^T e = e$ by Lemma 2.32. □

Example 2.38. The probability distribution with

$$\pi(u) = \frac{\deg(u)}{\mathrm{vol}(G)}$$

is reversible with respect to the transition matrix of the uniform process from Definition 2.33. For instance, if we consider the transition matrix P from Example 2.35, $P\pi = \pi$ for $\pi = (\frac{1}{2}, \frac{1}{4}, \frac{1}{4})^T$.

Definition 2.39. Let X be a Markov process on G with transition matrix P.

1. X is called **aperiodic**, if
$$\gcd(\{k \in \mathbb{N} \mid (P^k)_{uu} > 0\}) = 1$$
for all $u \in V$.

2. X is called **irreducible**, if for all $u, v \in V$ there exists $k \in \mathbb{N}$ with
$$(P^k)_{uv} > 0.$$

Recall from Lemma 2.34 that $(P^k)_{uv}$ is the probability of being at u after k steps given that we have started from v. Therefore, if on G there exists an irreducible Markov process, G must be connected. On the other hand, if G is connected, this does not imply that any Markov process on G is irreducible: on any graph we can always define the process which does not move with probability one.

We now come to the main theorem of this lecture.

Theorem 2.40. *Let X be an aperiodic and irreducible Markov process on G. Let P be the transition matrix of X. Then:*

1. *X has a unique stationary distribution π.*
2. *$\lim_{k\to\infty} P^k = \pi e^T$, where $e = (1, \dots, 1)^T \in \mathbb{R}^n$.*
3. *For all probability distributions $f : V \to \mathbb{R}$ we have $\lim_{k\to\infty} P^k f = \pi$.*

We need two auxiliary results for the proof of Theorem 2.40. We prove them first, and then we prove Theorem 2.40 towards the end of this section.

Proposition 2.41 (The Perron-Frobenius Theorem)**.** *Let $A = (a_{ij}) \in \mathbb{R}^{n\times n}$ be a matrix with $a_{ij} > 0$ for all i, j and $A^T e = e$. Then, 1 is a simple eigenvalue of A and all other eigenvalues λ satisfy $|\lambda| < 1$.*

Proof. Since $A^T e = e$, the transpose of A has eigenvalue 1, and so A has eigenvalue 1. Fix now $k \geq 1$ and let $M := A^k$. Let us write the entries of M as m_{ij}. Since $M^T e = e$ and since the m_{ij} are all positive, we must have
$$\max_{1\leq i,j\leq n} \|m_{ij}\| < 1. \tag{2.3.2}$$

If in the Jordan normal form of A we have a Jordan block of the form

$$B = \begin{bmatrix} \lambda & 1 & & \\ & \ddots & \ddots & \\ & & \ddots & 1 \\ & & & \lambda \end{bmatrix},$$

then B^k has $k\lambda^{k-1}$ on the off-diagonal. Since by Eq. (2.3.2) the entries of M must be bounded, there can be such a Jordan block only if $|\lambda| < 1$. This implies that 1 is a simple eigenvalue.

Let now $\lambda \in \mathbb{C}, \lambda \neq 1$, be another eigenvalue of A^T. Let $x = (x_1, \dots, x_n)^T \in \mathbb{C}^n$ be a corresponding eigenvector. Then, we have $\lambda x_i = a_{1i}x_1 + \cdots + a_{ni}x_n$ for all $1 \leq i \leq n$. We have $1 = a_{1i} + \cdots + a_{ni}$, because $A^T e = e$. Since the a_{ji} are positive, this implies that λx_i is a convex combination of the entries of x. Let us denote by $C \subset \mathbb{C}$ the convex hull of the x_i. Since x must be linearly independent of e, there must be indices i and j with $x_i \neq x_j$. Therefore, C is not a single point. Moreover, since the a_{1i} are all strictly positive, λx_i lies in the relative interior of C. From this we get

$$|\lambda x_i| < \max_{1 \leq j \leq n} |x_j|,$$

which implies $|\lambda| < 1$. □

The second result we need is a lemma from commutative algebra. For this, we recall that a **semigroup** in $\mathbb{N}$ is a subset $S \subset \mathbb{N}$, such that for all $s, r \in S$ also $r + s \in S$.

Lemma 2.42. *Let $S \subset \mathbb{N}$ be semigroup and suppose that* $\gcd(S) = 1$ *(i.e., the greatest common divisor of all elements in S is 1). Then, $\mathbb{N} \setminus S$ is finite.*

Proof. We refer to [RGS09]. □

We can now prove Theorem 2.40.

Proof of Theorem 2.40. The proof is based on the fact that under the hypothesis of the theorem there exists $M \in \mathbb{N}$ with $(P^m)_{uv} > 0$ for all $m \geq M$ and $u, v \in V$. We show this at the end of this proof. Let us first see how the statement of the theorem follows.

We know that $P^T e = e$ from Lemma 2.32, and so $(P^m)^T e = e$ for all $m \geq 1$. Therefore, if $(P^m)_{uv} > 0$ for all $u, v \in V$, Proposition 2.41 implies that P^m has 1 as simple eigenvalue and that all other eigenvalues λ of P^m satisfy $|\lambda| < 1$. Since μ is an eigenvalue of P, if and only if $\lambda = \mu^m$ is an eigenvalue of P^m, we see that P has 1 as simple eigenvalue and that all other eigenvalues of P have absolute value strictly less than 1. It follows that P has a unique right-eigenvector $\pi \in \mathscr{F}(V)$ with eigenvalue 1. This is not yet enough to prove that π is a stationary distribution. For this, we have to show $\pi \in \mathscr{F}_+(V)$ (i.e., π is nonnegative). We do this next.

Since $\lim_{k\to\infty} \mu^k = 0$ for all eigenvalues $\mu \neq 1$ of P, we see that P^k converges to a matrix P^* of rank one. Let us write $P^* = xy^T$. Since $P^k\pi = \pi$ and $(P^k)^T e = e$ for every k, we also have $P^*\pi = \pi$ and $(P^*)^T e = e$. This shows

$$\langle y, \pi\rangle x = \pi \quad \text{and} \quad \langle x, e\rangle y = e.$$

First, this shows that $\langle y, \pi\rangle \neq 0$ and $\langle x, e\rangle \neq 0$. After rescaling y, we can assume that $\langle x, e\rangle = 1$, so that $y = e$. Then, $\pi = \langle y, \pi\rangle x = \langle e, \pi\rangle x$. Furthermore, after rescaling π, we can assume that

$$\langle e, \pi\rangle = 1. \tag{2.3.3}$$

We get

$$\lim_{k\to\infty} P^k = \pi e^T. \tag{2.3.4}$$

On the one hand, this proves the second item of Theorem 2.40. On the other hand, since π is the limit of the columns of P^k, we also have $\pi \in \mathscr{F}_+(V)$. Together with Eq. (2.3.3) this shows that π indeed gives a stationary distribution. Moreover, Eq. (2.3.4) shows that for all probability distributions $f : V \to \mathbb{R}$ we have

$$\lim_{k\to\infty} P^k f = P^* f = \pi e^T f = \pi,$$

because $e^T f = \langle e, f\rangle = 1$.

It remains to show the claim above. For this we pick $u \in V$ and define

$$S_u := \{t \in \mathbb{N} \mid (P^t)_{uu} > 0\}.$$

Recall from Lemma 2.34 that $(P^t)_{uu}$ is the probability of having a walk with t steps that starts and ends at u. If $\alpha, \beta \in S_u$, then we must have $\alpha + \beta \in S_u$, since we can simply join two walks starting and ending at u. This shows that S_u is a semigroup. Since X is aperiodic, $\mathbb{N} \setminus S_u$ is finite by Lemma 2.42. So, there exists $M_u \in \mathbb{N}$ with $(P^m)_{uu} > 0$ for all $m \geq M_u$. We set

$$M' := \max_{u\in V} M_u.$$

Let now $u, v \in V$, $u \neq v$, and define for $t \geq 1$:

$$\ell_{uv}^t := P(X_t = u \text{ and } X_i \neq u \text{ for } i < t \mid X_0 = v). \tag{2.3.5}$$

This is the probability of arriving at u for the first time after t steps, when we have started at v. For every $k \geq 1$ we have

$$(P^k)_{uv} = \sum_{t=1}^{k} \ell_{uv}^t \cdot (P^{k-t})_{uu}.$$

Since X is irreducible, there exists $r \in \mathbb{N}$ such that $(P^r)_{uv} > 0$. Therefore, $\ell_{uv}^1, \ldots, \ell_{uv}^r$ can't all be equal to zero, so that $\sum_{t=1}^{r} \ell_{uv}^t > 0$. We set

$$M := M' + r$$

and get for $m \geq M$:

$$(P^m)_{uv} = \sum_{t=1}^{m} \ell_{uv}^t \cdot (P^{m-t})_{uu} \geq \sum_{t=1}^{r} \ell_{uv}^t \cdot (P^{m-t})_{uu} > \delta \sum_{t=1}^{r} \ell_{uv}^t > 0,$$

where $\delta := \min_{m-R \leq k < m} (P^k)_{uu} > 0$. □

We have shown in Theorem 2.40 that aperiodic and irreducible Markov chains on a graph have a unique stationary distribution. Let us now turn our point of view upside down and start with probability distribution $\pi : V \to \mathbb{R}$ on a connected graph $G = (V, E)$. Can we find a Markov process whose stationary distribution is π? The answer is yes! And we can use this process to sample from π by simply starting at a vertex $v \in V$ and following the random walk along the graph. This procedure is known as **Metropolis-Hastings algorithm** and it is based on the next theorem.

Theorem 2.43. *Let $G = (V, E)$ be a connected graph with $n = |V|$ and let $\pi : V \to \mathbb{R}$ be a probability distribution with $\pi(u) > 0$ for all $u \in V$. Let $d > \max_{v \in V} \deg(v)$. Define $P = (p_{uv}) \in \mathbb{R}^{n \times n}$ with*

$$p_{uv} = \begin{cases} \frac{1}{d} \min\{1, \frac{\pi(u)}{\pi(v)}\}, & \text{if } u \neq v \text{ and } \{u, v\} \in E \\ 0, & \text{if } u \neq v \text{ and } \{u, v\} \notin E \\ 1 - \sum_{i \neq v} p_{iv}, & \text{if } u = v. \end{cases}$$

Then, P is the transition matrix of an aperiodic and irreducible Markov process with unique stationary distribution π.

The proof of the theorem is Exercise 2.11.

One interesting property of the transition matrix in Theorem 2.43 is that it only involves ratios of the entries of π. This is useful, when we only know π up to scaling. We give an example of such a situation.

Example 2.44. Let $G=(V,E)$ be a graph. We call a function $F\colon V\to\{1,\ldots,k\}$ a k-**coloring** of G. Let us call a coloring **admissible**, if $F(u)\neq F(v)$ for all $\{u,v\}\in E$ (the minimal k, such that G has an admissible k-coloring is called the chromatic number of G). Counting the number of admissible k-colorings in G is a hard problem. Nevertheless, we can still sample uniformly a random k-coloring from the set of all admissible k-colorings in G using Theorem 2.43. For this, we assume $k\geq\max_{v\in V}\deg(v)+2$, and define the graph $\hat{G}:=(\hat{V},\hat{E})$ with

$$\begin{aligned}\hat{V}&:=\{F\colon V\to\{1,\ldots,k\}\mid F \text{ is admissible}\},\\ \hat{E}&:=\{\{F,G\}\subset\hat{V}\mid F \text{ and } G \text{ differ in exactly one vertex } u\}.\end{aligned}$$

Then, G is connected (see, e.g., [FV07]) and the number of all admissible k-colorings is $|\hat{V}|$. Using Theorem 2.43 we can sample the probability distribution $\pi: V\to R$ with

$$\pi(u)=\frac{1}{|\hat{V}|}\text{ for all } u\in\hat{V}$$

without knowing $|\hat{V}|$.

Lastly, one thing that we did not discuss in this lecture is the number of steps it takes such that an aperiodic and irreducible Markov process X is somehow close to its stationary distribution π. One way to measure convergence is using the **total variation distance** $d(t):=\max_{u,v\in V}|P(X_t=u\mid X_0=v)-\pi(u)|$. For a given $\delta>0$ the minimal t such that $d(t)<\delta$ is called **mixing time** of the process. We refer to [Chu97, Chapter 1] for more details. In practice, however, it often suffices to take a fixed number of steps.

Exercise 2.11. Prove Theorem 2.43. **Hint:** Use Proposition 2.37.

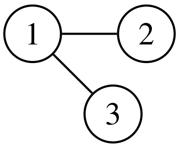
Exercise 2.12. Let G be a graph and X be the uniform Markov process on G. Show:

1. X is aperiodic, if and only if G is not bipartite.
2. X is irreducible, if and only if G is connected.

Exercise 2.13. Prove that the following algorithm implements the Markov process from Example 2.44. Suppose that at the i-th step we have the admissible coloring F_i. Then we do the following:

1. Choose $(u,c) \sim \mathrm{Unif}(V \times \{1,\dots,k\})$.
2. For all $v \in V \setminus \{u\}$ set $F_{i+1}(v) = F_i(v)$, and $F_{i+1}(u) = c$.
3. If F_{i+1} is not an admissible coloring, go back to 1. Otherwise, return F_{i+1}.

Exercise 2.14. Implement the algorithm from Exercise 2.13 for the graph



using 4 colors.

2.4 Centrality Measures

Let $G = (V,E)$ be a graph. In Section 2.2 we interpreted the values $f(u)$, $u \in V$, of a function $f \in \mathscr{F}(V)$ as an indicator of belonging to a class of vertices. For instance, we assigned to a vertex $u \in V$ one of two labels depending on whether $f(u) \geq 0$ or $f(u) < 0$. In Section 2.3 we considered $f \in \mathscr{F}_+(V)$ to be a probability distribution on the set of vertices (recall from Eq. (2.3.1) that we denote by $\mathscr{F}_+(V)$ the vector space of nonnegative functions $V \to \mathbb{R}$).

In this last section on networks we take a third perspective. We regard a function $f \in \mathscr{F}_+(V)$ as a measure of importance and call f a **centrality measure**. Centrality measures are used to assess the role of a vertex in a network. If $f(u) > f(v)$, we interpret this as u playing a more important role for the network structure than v.

We start with one of the most relevant centrality measures: **Page-Rank** [PBMW98].

Definition 2.45. Let $G = (V,E)$ be a graph. We call a function $c_R \in \mathscr{F}_+(V)$ satisfying

$$c_R(u) = \sum_{v \in V : \{u,v\} \in E} \frac{c_R(v)}{\deg(v)}$$

a Page-Rank of G.

If c_R is a Page-Rank of G, so is $\lambda \cdot c_R$ for every $\lambda \geq 0$. The next result gives conditions when c_R exists and is unique up to such a scaling.

Proposition 2.46. *Let $G = (V, E)$ be a graph. If G is connected and not bipartite, a Page-Rank of G exists and is unique up to scaling.*

Proof. We can write the defining equation in Definition 2.45 as

$$Pc_R = c_R,$$

where P is the transition matrix of the uniform Markov process X on G (see Definition 2.33). It follows from Exercise 2.11 that X is aperiodic and irreducible, and then Theorem 2.40 implies that X has a unique stationary process; i.e., a unique solution $c_R \in \mathscr{F}_+(V)$ with $Pc_R = c_R$ and $\langle e, c_R \rangle = 1$. □

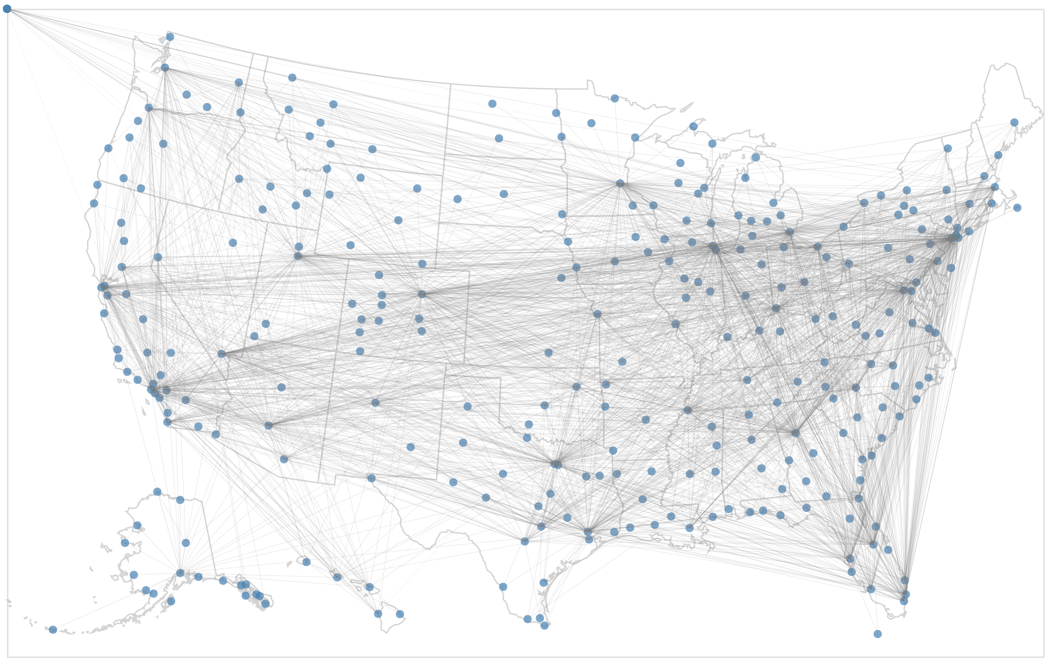


Figure 2.6: The network from Example 2.48. The blue vertices represent 305 airports in the US. There is an edge between two airports if there is a flight from one to the other. The vertex on the top left is Adak Airport.

Remark 2.47. We can use the transition matrix of any irreducible and aperiodic Markov process to define a corresponding Page-Rank.

Example 2.48. This example is based on the Graphs lecture in the Data Science course by Huda Nassar[2].

We compute Page-Rank for the airport dataset from the `VegaDatasets.jl`[3] package. From this dataset we compute a graph G with $n = 305$ vertices and 5668 edges. The vertices represent airports in the US, and there is an edge between two airports if there is a flight from one of the two airports to the other. The graph is shown in Fig. 2.6.

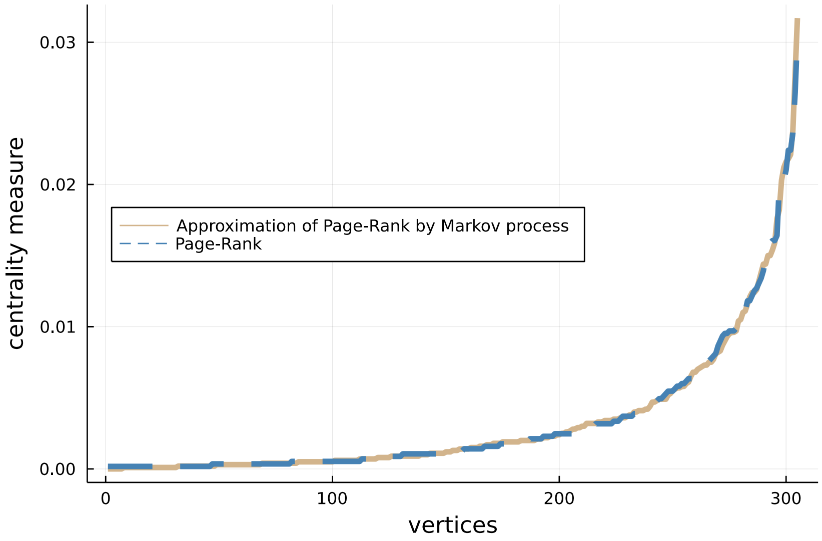


Figure 2.7: Page-Rank (blue) and approximated Page-Rank (brown) for the network from Fig. 2.6.

For computing Page-Rank we set up the transition matrix P of the uniform Markov process on G and then compute an eigendecomposition of P. We also approximate Page-Rank by using the third item in Theorem 2.40. For $1 \leq i \leq N$ we sample a random airport v_0 and then we start a random walk at v_0 using the transition probabilities in P. After m steps we record the locations v_i. We approximate Page-Rank by the empirical distribution function

$$f(u) = \frac{1}{N}|\{i \mid u = v_i\}|.$$

The result of an experiment with $N = 10^4$ and $m = 20$ is shown in Fig. 2.8 and Fig. 2.7.

Next, we introduce several other commonly used centrality measures.

[2] `https://github.com/JuliaAcademy/DataScience`

[3] `https://github.com/queryverse/VegaDatasets.jl`

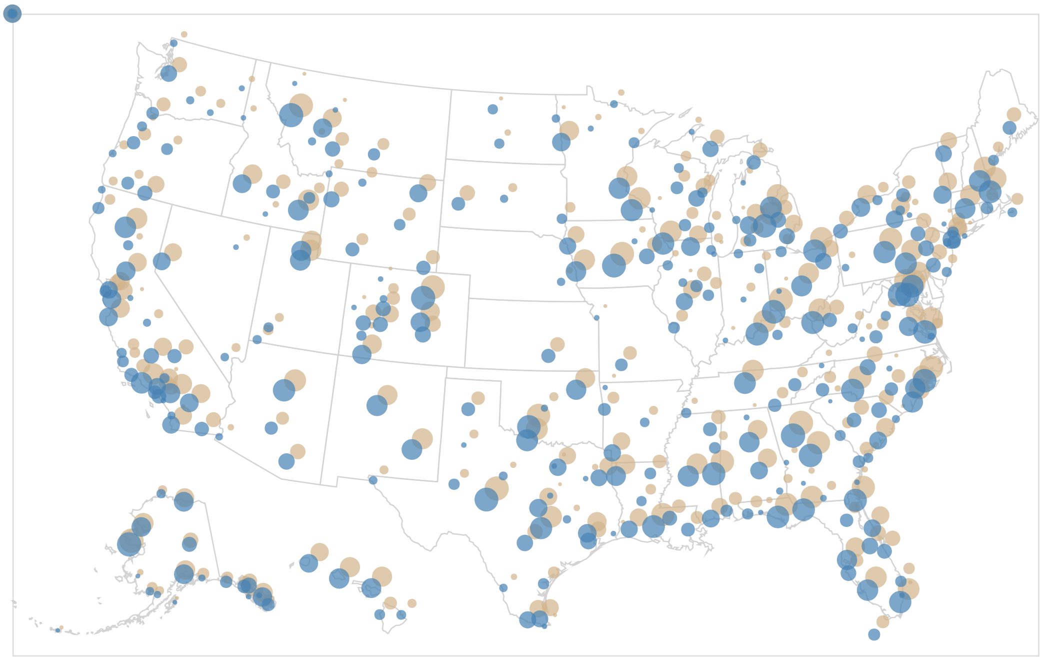


Figure 2.8: Page-Rank (blue) and approximated Page-Rank (brown) for the network from Fig. 2.6. The size of the circles corresponds to the respective centrality measures.

Definition 2.49. Let $G = (V, E)$ be a graph and $u \in V$ be a vertex.

1. The **degree centrality** of u is
$$c_D(u) := \deg(u).$$

2. The **closeness centrality** of u is
$$c_C(u) := \frac{1}{\sum_{v \in V} \operatorname{dist}(u, v)},$$
where $\operatorname{dist}(u, v)$ is the length of a shortest path from v to u.

3. The **harmonic centrality** of u is
$$c_H(u) := \sum_{v \in V} \frac{1}{\operatorname{dist}(u, v)}.$$

4. Let $\sigma_{x,y}$ be the number of shortest paths from x to y, and $\sigma_{x,y}(u)$ the number of shortest paths from x to y passing through u. The **betweenness centrality** of u is
$$c_B(u) := \sum_{x,y \in V \setminus \{u\}, x \neq y} \frac{\sigma_{x,y}(u)}{\sigma_{x,y}}.$$

5. The **Markov centrality** of u is

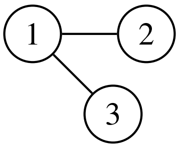
$$c_M(u) := \frac{1}{\sum_{v \in V \setminus \{u\}} \mathbb{E}\, \tau(u,v)},$$

where $\tau(u,v)$ is the minimal t such that a uniform Markov process X starting at v arrives at u for the first time after t steps:

$$\tau(u,v) = \min\{t \mid (X_t = u \mid X_0 = v)\};$$

i.e., $\mathbb{E}\, \tau(u,v) = \sum_{t=0}^{\infty} t \cdot \ell_{u,v}^t$ and $\ell_{u,v}^t = P(X_t = u \text{ and } X_i \neq u \text{ for } i < t \mid X_0 = v)$ is as in Eq. (2.3.5).

Example 2.50. Consider the graph from Example 2.3:



We compute the centrality measures from Definition 2.49 for the three vertices. In all cases 1 will have the largest measure. This can be interpreted as 1 taking the most important role in the network.

The degree centralities of the vertices are $c_D(2) = c_D(3) = 1$ and $c_D(1) = 2$. The closeness centralities are

$$C_C(1) = \frac{1}{\text{dist}(1,2) + \text{dist}(1,3)} = \frac{1}{1+1} = \frac{1}{2}$$

and

$$C_C(2) = \frac{1}{\text{dist}(1,2) + \text{dist}(2,3)} = \frac{1}{1+2} = \frac{1}{3}.$$

Similarly, $C_C(3) = \frac{1}{3}$. For harmonic centrality we have

$$C_H(1) = \frac{1}{\text{dist}(1,2)} + \frac{1}{\text{dist}(1,3)} = 1 + 1 = 2$$

and

$$C_H(2) = \frac{1}{\text{dist}(1,2)} + \frac{1}{\text{dist}(2,3)} = 1 + \frac{1}{2} = \frac{3}{2}$$

and $C_H(3) = \frac{3}{2}$ due to symmetry. The betweenness centralities are

$$C_B(1) = \frac{\sigma_{2,3}(1)}{\sigma_{2,3}} = \frac{1}{1} = 1$$

and $C_B(2) = C_B(3) = 0$ as there are no shortest paths from 1 to 3 passing through 2, and no shortest paths from 1 to 2 passing through 3. Finally, we compute the Markov centralities of the three vertices. First, we compute $c_M(1)$:

$$c_M(1) = \frac{1}{\mathbb{E}\,\tau(1,2) + \mathbb{E}\,\tau(1,3)}.$$

Starting from either 2 or 3 the next vertex must always be 1, which means that we have $\mathbb{E}\,\tau(1,2) = \mathbb{E}\,\tau(1,3) = 1$. Consequently, $c_M(1) = \frac{1}{2}$. Next, we compute $c_M(2)$. We have

$$c_M(2) = \frac{1}{\mathbb{E}\,\tau(2,1) + \mathbb{E}\,\tau(2,3)}.$$

We have $\mathbb{E}\,\tau(2,1) = \sum_{t=0}^{\infty} t \cdot \ell_{2,1}^t$. First, we observe that we can only pass from 1 to 2 in an odd number of steps. This implies $\ell_{2,1}^{2k} = 0$ for all k. If the number of steps is $2k+1$, then this means we have moved from 1 to 3 k times, before moving from 1 to 2. The probability of this event is $\ell_{2,1}^{2k+1} = \frac{1}{2^{k+1}}$. Therefore,

$$\mathbb{E}\,\tau(2,1) = \sum_{k=0}^{\infty} \frac{2k+1}{2^{k+1}} = 3.$$

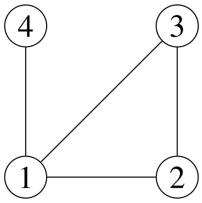
Furthermore, for moving from 3 to 2 we always need an even number of steps and we can argue as above to get

$$\mathbb{E}\,\tau(2,3) = \sum_{k=0}^{\infty} \frac{2k}{2^k} = 4.$$

This shows $c_M(2) = \frac{1}{3+4} = \frac{1}{7}$. Due to symmetry, also $c_M(3) = \frac{1}{7}$.

Exercise 2.15. How would you define Page-Rank for a directed graph? Analyze the data from Example 2.48 using your ideas.

Exercise 2.16. Compute the centrality measures from this section for the following graph:



3 Machine Learning

This chapter is based on [DFO20, Part II].

What is machine learning? Machine learning is a subarea of data analysis. The fundamental motivation is to develop algorithms that *automatically* extract information from datasets. Here we mean “automatic” in the sense that we want to find general methods which can be used in special situations so that we don’t need to create an algorithm for each individual scenario.

The automation of an algorithm happens through the analysis of *training data*. It is also said that one “learns” from the data. The data is understood as numeric vectors.

3.1 Data, Models, and Learning

The mathematical abstraction of a problem in machine learning is:

$$\begin{aligned}&\text{Given } (x_1,y_1),\dots,(x_n,y_n) \in \mathbb{R}^D \times \mathbb{R}^N,\\ &\text{find a function } f:\mathbb{R}^D \to \mathbb{R}^N \text{ so that}\\ &\quad 1.\ f(x_i) \approx y_i \text{ for } 1 \le i \le n\\ &\quad 2.\ \text{for each new data point } (x,y),\ f(x) \approx y.\end{aligned} \tag{3.1.1}$$

In real-world data we always have to assume the presence of noise, so that modeling $f(x) = y$ is not realistic. Using instead approximations allows for more flexibility. The exact meaning of $\approx$ depends on the problem, and is usually measured with a **loss function**. We will come back to this in Definition 3.5 below.

Definition 3.1. In Eq. (3.1.1):

1. We call the x_i **input data**, or **attributes**.
2. The y_i are called **labels**, **output-variables**, or **response-variables**.

Example 3.2. We are given the following data $(x,y) \in \mathbb{R}^3 \times \mathbb{R}$:

$x^{(1)}$ = degree	$x^{(2)}$ = current city	$x^{(3)}$ = age	y = yearly income
MSc	Osnabrück	36	60.145 €
PhD	Osnabrück	24	72.541 €
BSc	Hannover	31	58.901 €
MSc	Bremen	29	61.005 €

This dataset does not consist of numerical values. We can convert the data into numerical values as follows (though other ways are also possible):

$x^{(1)}$ = degree	$x^{(2)}$ = latitude	$x^{(3)}$ = longitude	$x^{(4)}$ = age	y = yearly income
2	52,28	8,05	36	60.145
3	52,28	8,05	24	72.541
1	52,38	9,73	31	58.901
2	53,1	8,8	29	61.005

Definition 3.3. Variables which have a continuous domain of values are called **continuous variables**. Variables with a discrete domain of values are called **discrete variables** or **categorical variables**.

In Eq. (3.1.1) it is usually not helpful to consider the class of all functions. The selected functions we use are called a **model**. The choice of model is dependent on context, and can be constructed either from an analyst or automatically.

Definition 3.4. Consider given data $(x_1,y_1),\ldots,(x_n,y_n) \in \mathbb{R}^D \times \mathbb{R}^N$.

1. A **deterministic model** is a function $f_\theta : \mathbb{R}^D \to \mathbb{R}^N$ depending on $\theta \in \mathbb{R}^P$.
2. A **statistical model** is a conditional probability distribution for $(y \mid x) \in \mathbb{R}^N$ that depends on a parameter $\theta \in \mathbb{R}^P$.

In the following, we will always consider statistical models with a density that we denote by $P_\theta(y \mid x)$. A key assumption for statistical models is that the n pairs in the training data are **chosen independently**.

The goal of machine learning is to determine a parameter θ that accurately describes the data in the model. We also say that we **learn** the parameter θ. Ideally, we have chosen a model and learned a parameter that predicts well on unseen data. For this reason, we split the data into **training data** and **test data**. The training data is used to learn the parameter. The role of the test data is to simulate unseen data. We assess the quality of prediction of our model by evaluating a **quality function** on the test data. This last step is called **validation**.

Thus, at the core approaching a machine learning problem consists of four steps.

Algorithm 3.1: Core steps of solving a machine learning problem.

```
Select a model;
Split the data into training and test data;
Learn parameters;
Validation.
```

Let us first consider the validation step. The common definition of a quality functions that can be used for validation is **empirical risk**.

Definition 3.5. Given data $(x_1,y_1),\dots,(x_n,y_n)$ and the model $f_\theta : \mathbb{R}^D \to \mathbb{R}^N$ (resp. P_θ), let $\ell : \mathbb{R}^N \times \mathbb{R}^N \to \mathbb{R}$ be a function called the **loss function**. Then the empirical risk depending on ℓ is

$$R(\theta) := \frac{1}{n}\sum_{i=1}^{n} \ell(y_i, f_\theta(x_i)) \quad \text{resp.} \quad \frac{1}{n}\sum_{i=1}^{n} \mathbb{E}_{\hat{y}_i \sim P_\theta(y|x_i)}\, \ell(y_i, \hat{y}_i).$$

When there is a discrepancy in the value of the quality function between training and test data, we speak of **overfitting**. Overfitting means that the model fits the training data well, but does not accurately predict the test data.

The simplest way for model selection is to split the data by randomly assigning the data points to either training or test data, learn a parameter θ, and validate θ on the test data using a quality function $Q(\theta)$. The model for which Q is minimized is chosen. A more sophisticated way is **cross-validation**. Here, we randomly split the data into k parts $D_1 \cup \dots \cup D_k$ with $D_i \cap D_j \neq \emptyset$ for $i \neq j$. Then, for every $1 \leq i \leq k$ we learn a parameter θ_i using $\bigcup_{j \neq i} D_j$ as training data and return $Q := \frac{1}{k}\sum_{i=1}^{k} R(\theta_i)$, where $R(\theta_i)$ is the risk (or any other quality function) for the test data D_i. As before, we choose the model for which Q is minimized. Therefore, cross-validation is a method for model selection, but not for parameter learning (although the process for cross-validation involves the computation of parameters).

For splitting the data into training and test data we can proceed as before and randomly assign training or test labels. Usually between 50%–80% of the data is used for training. The random choice in this step, however, should be independent of the random choices in the model selection step.

After a model is chosen and data is prepared, we learn parameters. In the deterministic model we utilize a technique called **Empirical Risk Minimization (ERM)**. This means computing a parameter θ^* which minimizes the empirical risk $R(\theta)$ on the training data, namely $\theta^* \in \operatorname{argmin}_\theta R(\theta)$. For the statistical model we can use **Maximum-Likelihood Estimation (MLE)** or **Maximum a-Posteriori Estimation (MAP)**. They correspond to maximizing the following functions.

Definition 3.6. Let $(x_1,y_1),\dots,(x_n,y_n)$ be data points and P_θ be a statistical model.

1. The **likelihood function** is the probability of observing the response variables given the input data: $L(\theta) := \prod_{i=1}^n P_\theta(y_i \mid x_i)$. The **log-likelihood function** is
$$l(\theta) := \sum_{i=1}^n \log P_\theta(y_i \mid x_i).$$
2. Let us denote $X := \begin{bmatrix} x_1 & \dots & x_n \end{bmatrix}^T \in \mathbb{R}^{n\times D}$ and $Y := \begin{bmatrix} y_1 & \dots & y_n \end{bmatrix}^T \in \mathbb{R}^{n\times N}$; i.e., X and Y have input data and response variables as their rows. Suppose that the parameter θ is a random variable, and that $(\theta \mid X,Y)$ has a probability density P. The **posteriori function** is
$$\alpha(\theta) = P(\theta \mid X,Y).$$

The motivation for the definition of l is that in order to maximize L, we can also maximize l. The latter is often simpler.

In particular, in maximum a-posteriori estimation we model the parameter θ as random. Taking the point of view of Bayesian probability allows us to use prior information for modelling the random variable θ. The choice of probability distribution for θ is therefore called the prior distribution. For instance, θ could be a normal distribution around a mean value that we have observed often. Going one step further, we can also keep θ random, so that our model for θ allows fluctuations. The response variable then has a conditional distribution $(y \mid x, \theta)$, and we can use Eq. (1.2.1) to get $P(y \mid x,X,Y) = \int_{\theta\in\mathbb{R}^P} P_\theta(y \mid x)\cdot P(\theta \mid X,Y)\,\mathrm{d}\theta$. These two approaches are summarized under the name **Bayesian machine learning**.

Finally, let us briefly come back to model selection. Suppose that we have to choose among models $M_1,\dots,M_r$. In Bayesian machine learning we can place a prior $P(M)$

on the choice of model. For instance, we could define $P(M_i) = \frac{1}{r}$ for $1 \leq i \leq r$. This would correspond to choosing a model uniform at random. Then, we have the posteriori function $P(M \mid X,Y)$, which we use for maximum a-posteriori estimation. It follows from Bayes theorem for densities (Theorem 1.22) that

$$\operatorname{argmax}_{1\leq i\leq r} P(M_i \mid X,Y) = \operatorname{argmax}_{1\leq i\leq r} P(M_i)\cdot P(X,Y \mid M_i)$$

and $P(X,Y \mid M_i)$ is the probability of having the data giving the model, also called **evidence** of the model. It is the marginal density, where the random variable θ has been integrated out: $P(X,Y \mid M_i) = \int_{\theta\in\mathbb{R}^P} P(X,Y \mid \theta)\cdot P(\theta \mid M_i)\,\mathrm{d}\theta$, by Eq. (1.2.1). Here, $P(\theta \mid M_i)$ is the prior distribution of θ for the model M_i and $P(X,Y \mid \theta)$ the joint density of (X,Y) given θ.

3.2 Nonlinear Regression and Neural Networks

In this lecture we consider a specific machine learning model for continuous input data and continuous response variables, namely linear regression.

Definition 3.7. The **linear model** $f_\theta : \mathbb{R}^D \to \mathbb{R}$ is given by the following function depending on the parameter $\theta = (\theta_0,\theta_1,\ldots,\theta_D)^T \in \mathbb{R}^{D+1}$:

$$f_\theta(x) = x^T\theta' + \theta_0,$$

for $\theta' = (\theta_1,\ldots,\theta_D)^T \in \mathbb{R}^D$. The **quadratic loss** is

$$\ell(y,\hat{y}) = (y-\hat{y})^2.$$

In this case the ERM is called the **least squares problem** or **linear regression**.

The empirical risk $R(\theta)$ (see Definition 3.5) for the quadratic loss is also called **mean squared error (MSE)**. Alternatively, one also considers the **root mean squared error (RMSE)** defined by $\mathrm{RMSE}(\theta) := \sqrt{R(\theta)}$.

We again consider training data $(x_1,y_1),\ldots,(x_n,y_n) \in \mathbb{R}^D \times \mathbb{R}$. Recall that we denote

$$X := \begin{bmatrix} x_1 & \ldots & x_n \end{bmatrix}^T \in \mathbb{R}^{n\times D} \quad\text{and}\quad Y := \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix}^T \in \mathbb{R}^n. \tag{3.2.1}$$

We also denote

$$\Omega = \begin{bmatrix} 1 & \cdots & 1 \\ x_1 & \cdots & x_n \end{bmatrix}^T \in \mathbb{R}^{n\times(D+1)}, \tag{3.2.2}$$

called **feature matrix**.

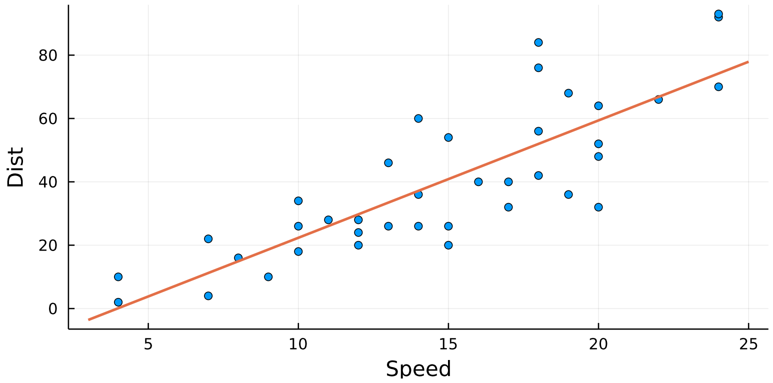


Figure 3.1: The picture shows 75% of the data in the dataset `cars` from the `RDatasets` package, and the result of a linear regression for this data.

Theorem 3.8. *Giving the setting from Definition 3.7, let $Y \in \mathbb{R}^n$ be as in Eq.* (3.2.1) *and Ω be the feature matrix. Recall that $r(\Omega)$ denotes the rank of the matrix Ω.*

1. *When $r(\Omega) < D+1$, then $\operatorname{argmin}_\theta R(\theta)$ has infinitely many solutions.*
2. *When $r(\Omega) = D+1$, then $\operatorname{argmin}_\theta R(\theta)$ has a unique solution*

$$\theta^* := \Omega^\dagger Y,$$

where $\Omega^\dagger = (\Omega^T\Omega)^{-1}\Omega^T$ is the pseudoinverse of Ω as in Definition 1.5.

Proof. The empirical risk is

$$R(\theta) = \frac{1}{n}\sum_{i=1}^n (f_\theta(x_i) - y_i)^2 = \frac{1}{n}\sum_{i=1}^n (x_i^T\theta' + \theta_0 - y_i)^2 = \frac{1}{n}(\Omega\theta - Y)^T(\Omega\theta - Y).$$

Let $z := \Omega\Omega^\dagger Y$. Recall from Corollary 1.6 that $z = \operatorname{argmin}_{w\in \operatorname{Im}(\Omega)}(w-Y)^T(w-Y)$. If $r(\Omega) < D+1$, then $\Omega\theta = z$ has infinitely many solutions. If $r(\Omega) = D+1$, by Proposition 1.7, $\Omega^\dagger\Omega = \mathbf{1}_{D+1}$, so that $\theta = \Omega^\dagger Y$ is uniquely determined. □

Example 3.9. The `RDatasets`[1] provides the data set `cars` that features the variables $x =$`speed` and $y =$`dist`. The dataset is shown in Fig. 3.1. We expect a linear relation of the form $y = ax + b$ to hold. We estimate a and b using Theorem 3.8 and a subset ot 75% of the data for training. The result is shown in Fig. 3.1.

[1] `https://github.com/JuliaStats/RDatasets.jl`

Unfortunately, the least-squares parameter often tends to overfitting; i.e., it works well on the training data but not on the test data. The reason is that the quadratic loss function allows too much flexibility for the parameter. To prevent overfitting we can incorporate a **regularizing term**.

Definition 3.10. Let $A \in \mathbb{R}^{(D+1)\times(D+1)}$ be an invertible matrix, and let $\lambda \in \mathbb{R}$. The **regularized loss** for the linear model is given by

$$\ell(\hat{y}, y) = (y - \hat{y})^2 + \lambda \left\|A\theta\right\|^2.$$

The special case of Definition 3.10 given by the case $A = \mathbf{1}_{D+1}$ is called **Tikhonov regularization**. The corresponding regression problem is called **ridge regression**. The parameter λ is here not considered a parameter of the model in the sense of step 3. of Algorithm 3.1, but a parameter which is chosen in advance or tuned later. Such parameters are called **hyperparameters**.

Theorem 3.11. *Let $\Omega \in \mathbb{R}^{n\times(D+1)}$ be the feature matrix from Eq.* (3.2.2) *and $Y \in \mathbb{R}^n$ be as in Eq.* (3.2.1)*. Giving the setting from Definition 3.10 and Eq.* (3.2.3)*, for almost every λ there is a unique solution $\theta^* = \operatorname{argmin}_\theta R(\theta)$ given by*

$$\theta^* = (\Omega^T\Omega + n\lambda A^T A)^{-1}\Omega^T Y.$$

Notice that Theorem 3.11 shows that the regularized linear models not only helps against overfitting, but also in the case that Ω does not have rank $D+1$. This is especially useful in the case when there is too little data with $n \leq D$ or when the data lives in a lower dimensional subspace.

Proof of Theorem 3.11. The risk of the reglarized loss is given by

$$R(\theta) = \frac{1}{n}\left\|\Omega\theta - Y\right\|^2 + \lambda \left\|A\theta\right\|^2 = \frac{1}{n}(\Omega\theta - Y)^T(\Omega\theta - Y) + \lambda\theta A^T A\theta. \qquad (3.2.3)$$

We compute the derivative of $R(\theta)$ using Exercise 1.1 and set it equal to zero:

$$\nabla R(\theta) = \frac{2}{n}\Omega^T(\Omega\theta - Y) + 2\lambda A^T A\theta = 0.$$

Which implies

$$(\Omega^T\Omega + n\lambda A^T A)\theta = \Omega^T Y.$$

Notice that $\det(\Omega^T\Omega + n\lambda A^T A)$ is a polynomial in λ of degree $D+1$ with at most $D+1$ real zeroes. So for almost every λ the matrix $\Omega^T\Omega + n\lambda A^T A$ is invertible and thus

$$\theta = (\Omega^T\Omega + n\lambda A^T A)^{-1}\Omega^T Y,$$

which also must be the minimum. □

Let us now consider regression problems using statistical models. Let $f_\theta : \mathbb{R}^D \to \mathbb{R}$ be a function that depends on a parameter θ. We handle the joint probability density

$$P_\theta(y \mid x) = \Phi(y \mid f_\theta(x), \sigma^2),$$

where Φ is the density of the normal distribution as in Eq. (1.2.2). This means that we have $y = f_\theta(x) + \varepsilon$ for $\varepsilon \sim N(0, \sigma^2)$. Note that through this statistical model we model (measurement) errors in the data. The variance σ^2 is not considered a parameter in this model, but it is a hyperparameter. We will consider a model that has σ^2 as a parameter in Proposition 3.14 below.

We now again consider linear regression for $\theta = (\theta_0, \theta_1, \ldots, \theta_D)^T \in \mathbb{R}^{D+1}$. That is $f_\theta(x) = x^T\theta' + \theta_0$, where $\theta' = (\theta_1, \ldots, \theta_D)$.

Theorem 3.12. *Let $\Omega \in \mathbb{R}^{n\times(D+1)}$ be the feature matrix as defined in Eq.* (3.2.2). *Let $Y \in \mathbb{R}^n$ be as in Eq.* (3.2.1). *The maximum likelihood estimation in the above model for linear regression*

1. *is not unique if $r(\Omega) < D+1$,*
2. *is uniquely determined by $\theta_{\mathrm{ML}} = \Omega^\dagger Y$ when $r(\Omega) = D+1$.*

Note that Theorem 3.12 shows that the risk minimization in the deterministic model and in the statistical model give the same answer.

Proof of Theorem 3.12. Following Definition 3.6 the log-likelihood-function is

$$\begin{aligned} l(\theta) &= \sum_{i=1}^{n} \log P_\theta(y_i|x_i) \\ &= -\frac{n}{2}\log(2\sigma^2) + \sum_{i=1}^{n} \frac{1}{2\sigma^2}(x_i^T\theta' - \theta_0 - y_i)^2 \\ &= -\frac{n}{2}\log(2\sigma^2) + \frac{1}{2\sigma^2}\left\|\Omega\theta - Y\right\|^2. \end{aligned}$$

The first term does not involve θ. As in the proof of Theorem 3.11, we proved that $\operatorname{argmin}_\theta \|\Omega\theta - Y\|^2$ has a unique solution exactly when $r(\Omega) = D+1$, and then the solution is $\theta_{\text{ML}} = \Omega^\dagger Y$. □

The generalization from linear regression is **nonlinear regression**, which is defined through the following statistical model for $\phi : \mathbb{R}^D \to \mathbb{R}^P$ and $\theta = (\theta_1, \dots, \theta_P) \in \mathbb{R}^P$:

$$P_\theta(y \mid x) = \Phi(y \mid \theta^T \phi(x), \sigma^2).$$

For every such model we have a corresponding **feature matrix**, which we denote by

$$\Omega = \begin{bmatrix} \phi(x_1) & \dots & \phi(x_n) \end{bmatrix}^T \in \mathbb{R}^{n \times P}. \tag{3.2.4}$$

This notation is not in conflict with Eq. (3.2.2): in Eq. (3.2.2) we see the special case of Eq. (3.2.4) for linear regression. We note the following.

- In linear regression we have $P = D+1$ and
$$\phi(x) = \begin{bmatrix} 1 \\ x \end{bmatrix}.$$
- For $D = 1$ and
$$\phi(x) = \begin{bmatrix} 1 \\ x \\ \vdots \\ x^{P-1} \end{bmatrix}$$
we have **polynomial regression**.

The maximum likehood estimate for nonlinear regression is given next.

Theorem 3.13. *Let Ω be the feature matrix in Eq.* (3.2.4) *and $Y \in \mathbb{R}^n$ be as in Eq.* (3.2.1). *The maximum likelihood estimator in nonlinear regression is*

1. *not unique when $r(\Omega) < P$.*
2. *uniquely determined by $\theta_{\text{ML}} = \Omega^\dagger Y$ when $r(\Omega) = P$.*

Proof. See Exercise 3.1 □

In Theorem 3.12 and Theorem 3.13 we have the variance σ^2 provided. Alternatively we can also model the parameter.

Proposition 3.14. *If we model the variance σ^2 in Theorem 3.13 instead as a parameter, then this gives the MLE $(\theta_{\mathrm{ML}}, \sigma^2_{\mathrm{ML}})$ with $\theta_{\mathrm{ML}} = \Omega^{\dagger} Y$, and $\sigma^2_{\mathrm{ML}} = \frac{1}{n} \|Y - \Omega\theta_{\mathrm{ML}}\|^2$.*

Proof. See Exercise 3.3 □

Similar to linear regression the MLE tends towards **overfitting**. Our solution in the deterministic model was to introduce a penalty term $\lambda \|A\theta\|^2$. In the Bayesian approach θ itself is random, and instead of a penalty term we prescribe a choice of θ called a **prior**. This then leads to a maximum a-posteriori estimator as in Definition 3.6.

The next theorem computes the MAP for the statistical model $(y \mid x) \sim N(\theta^T \phi(x), \sigma^2)$ using a Gaussian prior $N(\mu, \Sigma)$. The theorem shows that in this case MAP can be understood as the statistical analogue of regularization.

Theorem 3.15. *Let $\mu \in \mathbb{R}^P$ and $\Sigma \in \mathbb{R}^{P\times P}$ be positive Given the statistical model above, then for the prior $\theta \sim N(\mu, \Sigma)$ we have the MAP*

$$\theta_{\mathrm{MAP}} = (\Omega^T\Omega + \sigma^2\Sigma^{-1})^{-1}(\Omega^T Y + \sigma^2\Sigma^{-1}\mu),$$

where Ω is the feature matrix, $Y \in \mathbb{R}^n$ is as in Eq. (3.2.1)*, and under the assumption that $\Omega^T\Omega + \sigma^2\Sigma^{-1}$ is invertible.*

Proof. Let $\alpha(\theta) = P(\theta \mid X, Y)$ be defined as in Definition 3.6. By Bayes' Theorem for densities (Theorem 1.22),

$$P(\theta \mid X, Y) = P(\theta)\frac{P(Y \mid X, \theta)}{P(Y \mid X)}.$$

Therefore setting $c = -\log P(Y \mid X)$ which is constant in θ, we have

$$\log P(\theta \mid X, Y) = \log P(\theta) + \log P(Y \mid X, \theta) + c.$$

Recall that

$$P(\theta) = \Phi(\theta \mid \mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^n \det\Sigma}} \exp\left(-\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu)\right). \tag{3.2.5}$$

Furthermore, as the n pairs in the training data are assumed to be independent, we have

$$P(Y \mid X, \theta) = \prod_{i=1}^{n} \Phi(y_i \mid \theta^T\phi(x_i), \sigma^2) = \frac{1}{\sqrt{(2\sigma^2)^n}} \exp\left(-\frac{1}{2\sigma^2}\|Y - \Omega\theta\|^2\right). \tag{3.2.6}$$

Eq. (3.2.5) and Eq. (3.2.6) together yield

$$\log P(\theta \mid X,Y) = -\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu) - \frac{1}{2\sigma^2}\|Y-\Omega\theta\|^2 + c', \qquad (3.2.7)$$

where $c' = c + \log\frac{1}{\sqrt{(2\pi)^n \det\Sigma}} + \log\frac{1}{\sqrt{2\sigma^2}^n}$ is independent of θ.

The logarithm is strictly monotone, so we can instead of maximizing $\alpha(\theta)$ also maximize $\log\alpha(\theta) = \log P(\theta \mid X,Y)$. Taking the derivatives with respect to θ of Eq. (3.2.7):

$$\frac{\mathrm{d}}{\mathrm{d}\theta}\log\alpha(\theta) = -\Sigma^{-1}(\theta-\mu) + \frac{1}{\sigma^2}\Omega^T(Y-\Omega\theta).$$

Thus, if we set this derivative equal to zero

$$\Sigma^{-1}\theta - \Sigma^{-1}\mu = \frac{1}{\sigma^2}\Omega^T Y - \frac{1}{\sigma^2}\Omega^T\Omega\theta,$$

which implies

$$(\Omega^T\Omega + \sigma^2\Sigma^{-1})\theta = \frac{1}{\sigma^2}\Omega^T Y + \sigma^2\Sigma^{-1}\mu.$$

So we conclude

$$\theta = (\Omega^T\Omega + \sigma^2\Sigma^{-1})^{-1}(\Omega^T Y + \sigma^2\Sigma^{-1}\mu)$$

and this must also be the minimizer θ_{MAP}. □

Example 3.16. We compute the ML and MAP estimator using Theorem 3.12 and Theorem 3.15 for the data in Example 3.9. A sample from these statistical models is shown in Fig. 3.2. The sample points are connected by lines to plot a piecewise linear function.

In the discussion after Definition 3.6 we observed that instead of finding θ deterministically through MLE or MAP, we can also compute the distribution of θ given the training data. In the context of regression this approach is called **Bayesian regression**. We will assume as before that $\theta \sim N(\mu,\Sigma)$ and $(y \mid x,\theta) \sim N(\phi(x)^T\theta,\sigma^2)$. The goal of Bayesian regression is to sample from the posteriori distribution with density $\alpha(\theta) = P(\theta \mid X,Y)$. In this concrete case we can explicitly compute the posteriori distribution.

Theorem 3.17. *In the setting above we have*

$$(\theta \mid X,Y) \sim N(m,S),$$

where $S = (\sigma^{-2}\Omega^T\Omega + \Sigma^{-1})^{-1}$ *and* $m = S(\sigma^{-2}\Omega^T Y + \Sigma^{-1}\mu)$.

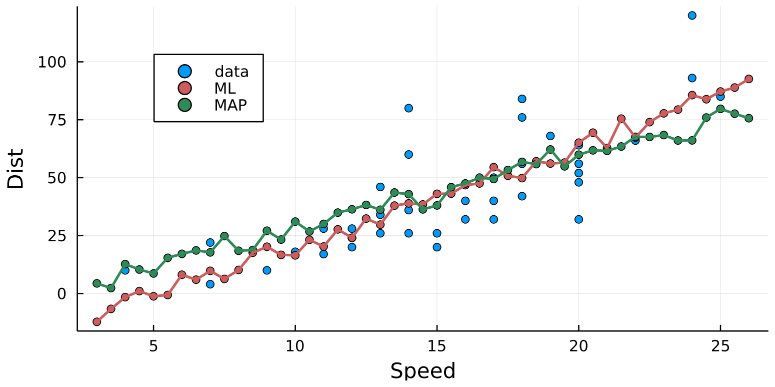


Figure 3.2: The picture shows ML and MAP for the data in Fig. 3.1. For the plot we have sampled points and connected them by lines to illustrate the statistical model. The red points show the ML estimator, the green points the MAP estimator.

In fact, Theorem 3.17 implies Theorem 3.15, because the density of a Gaussian is maximized at the expected value; i.e., $m = \theta_{\text{MAP}}$.

Proof of Theorem 3.17. Recall from Eq. (3.2.7) that

$$\log P(\theta \mid X, Y) = -\frac{1}{2}(\theta-\mu)^T\Sigma^{-1}(\theta-\mu) - \frac{1}{2\sigma^2}\,\|Y-\Omega\theta\|^2 + c,$$

where c is some constant, which is independent of θ. Let us write

$$Q := (\theta-\mu)^T\Sigma^{-1}(\theta-\mu) + \frac{1}{\sigma^2}(Y-\Omega\theta)^T(Y-\Omega\theta),$$

so that $P(\theta \mid X, Y) = \exp(-\frac{1}{2}Q + c')$. We expand to find

$$Q = \theta^T\Sigma^{-1}\theta - 2\mu^T\Sigma^{-1}\theta + \frac{1}{\sigma^2}(\theta^T\Omega^T\Omega\theta - 2Y^T\Omega\theta) + c',$$

where c' is independent of θ. Setting

$$A = \frac{1}{\sigma^2}\Omega^T\Omega + \Sigma^{-1} \quad \text{and} \quad a = \frac{1}{\sigma^2}\Omega^T Y + \Sigma^{-1}\mu$$

gives

$$Q = \theta^T A\theta - 2a^T\theta + c' = (\theta-b)^T A(\theta-b) + c'',$$

where $Ab = a$ and c'' is independent of θ. Notice

$$b = A^{-1}a = \Big(\frac{1}{\sigma^2}\Phi^T\Phi + \Sigma^{-1}\Big)^{-1}\Big(\frac{1}{\sigma^2}\Phi^T Y + \Sigma^{-1}\mu\Big).$$

This shows that the density function of θ given X,Y is

$$P(\theta \mid X,Y) = \exp\Big(-\frac{1}{2}Q + c'\Big) = C\exp\Big(-\frac{1}{2}(\theta - m)^T S^{-1}(\theta - m)\Big)$$

with C independent from $\theta \in \mathbb{R}^P$. Since $P(\theta \mid X,Y)$ is a density, it must integrate to 1 and we must therefore have $C = 1/\sqrt{(2\pi)^P \det(S)}$. □

For the validation step we compute the marginal distribution of the response variable $y \in \mathbb{R}$ given an input variable $x \in \mathbb{R}^D$ and the training data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}$.

Proposition 3.18. *Giving training data* $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}$ *and an additional data point* $(x,y) \in \mathbb{R}^D \times \mathbb{R}$*, the distribution of* $(y \mid x, X, Y)$ *is*

$$(y \mid x, X, Y) \sim N(\phi^T(x)m,\, \phi^T(x)S\phi(x) + \sigma^2)$$

where S and m are as in Theorem 3.17, and X,Y are as in Eq. (3.2.1).

Proof. See Exercise 3.4. □

The discussion in this section centered around models with (non-)linear functions $\mathbb{R}^D \to \mathbb{R}$. Given a nonlinear function $\phi : \mathbb{R}^D \to \mathbb{R}^P$, a straightforward generalization of this setting is to multivariate models

$$f_\theta : \mathbb{R}^D \to \mathbb{R}^N,\, x \mapsto \begin{bmatrix} \phi(x)^T \theta^{(1)} \\ \vdots \\ \phi(x)^T \theta^{(N)} \end{bmatrix}, \quad \theta = \begin{bmatrix} \theta^{(1)} & \cdots & \theta^{(N)} \end{bmatrix} \in \mathbb{R}^{P \times N}, \tag{3.2.8}$$

and using quadratic loss $\ell(y, \hat{y}) = ||y - \hat{y}||^2$ for the deterministic model or the multivariate Gaussian distribution $P_\theta(y \mid x) = \Phi(y \mid f_\theta(x), \sigma^2 \mathbf{1}_N)$ for the statistical model. Since the Euclidean norm satisfies $||y - \hat{y}||^2 = \sum_{i=1}^N (y_i - \hat{y}_i)^2$ for vectors $y = (y_1, \ldots, y_N)^T$ and $\hat{y} = (\hat{y}_1, \ldots, \hat{y}_N)^T$, maximization with respect to parameters can be done for each entry of $f_\theta : \mathbb{R}^D \to \mathbb{R}^N$ separately. Therefore, the estimators obtained in this section can be used for each entry of $f_\theta : \mathbb{R}^D \to \mathbb{R}^N$ separately.

Another generalization of linear regression models is given by neural networks. They are obtained by iterating nonlinear regression models.

Definition 3.19. Let $L > 0$ and $N_0, N_1, \ldots, N_L > 0$. Let $g_{\theta_i} : \mathbb{R}^{N_{i-1}} \to \mathbb{R}^{N_i}$ be nonlinear regression models, $1 \leq i \leq L$. Denote $D := N_0, N := N_L$. We call The model

$$f_\theta = (g_{\theta_L} \circ \cdots \circ g_{\theta_1}) : \mathbb{R}^D \to \mathbb{R}^N$$

a **neural network** of depth L. The function g_{θ_i} is called the i-th **layer** of the network.

In this definition, the linear and nonlinear regression models above are neural networks of depth 1. For depth larger than 1, however, we have no analytic expression for MLE or MAP. Instead, we have to use methods from optimization for computing them.

Neural networks often emcompass combinations of linear functions with so-called **activation functions**. This means that in Definition 3.19 the i-th layer is given by a function of the form $g_{\theta_i} = \sigma_i(f_{\theta_i}(x))$, where f_{θ_i} is a multivariate nonlinear regression model as in Eq. (3.2.8) and σ_i is a nonlinear activation function. For instance, activation functions for $z \in \mathbb{R}^k$ are the **ReLu function** $\sigma(z) = (\max(0, z_i))_{1\leq i\leq k}$ the **sigmoid function** $\sigma(z) = (\frac{1}{1+\exp(-z_i)})_{1\leq i\leq k}$, or the **softmax function**

$$\sigma(z) = \left(\frac{\exp(z_i)}{\sum_{j=1}^{k}\exp(z_j)}\right)_{1\leq i\leq k}. \tag{3.2.9}$$

I.e., the softmax function is a smooth version of the max function.

Example 3.20. We use a neural network of depth 2 for the data in Example 3.9. The inner functions should be $\sigma \circ f_1$ and $\sigma \circ f_2$, where σ is the ReLu activation function and $f_1 : \mathbb{R}^1 \to \mathbb{R}^2$ and $f_2 : \mathbb{R}^2 \to \mathbb{R}^1$ are linear. Unlike in the previous examples we don't have a closed form for the optimal parameters. Instead, we minimize the empirical risk using optimization methods. The result of the computation is shown in Fig. 3.3. We can see from the figure that the neural network computed an estimator that is piecewise linear. This indicates that there could be a hidden latent variable describing the data with two different linear models depending on whether speed is small or large.

In Example 3.20 we used a neural network to obtain a predictor for a continuous response variable. We can also use a neural network for statistical models of categorical variables. Suppose that we are given data $(x_1, y_1), \ldots, (x_n, y_n) \in \mathbb{R}^D \times \mathbb{R}^N$, where the response variables are categorical and take values in the finite set $\{c_1, \ldots, c_k\}$. Let $f_\theta : \mathbb{R}^D \to \mathbb{R}^k$ be a neural network such that $\sum_{j=1}(f_\theta(z))_j = 1$ and $(f_\theta(z))_i \geq 0$ for all $1 \leq i \leq k$ and $z \in \mathbb{R}^D$. We get a **statistical model for categorical variables** $P_\theta(y \mid x)$ by setting

$$P_\theta((y \mid x) = c_i) := (f_\theta(x))_i.$$

For instance, we can choose in the last layer the soft-max from Eq. (3.2.9) as activation function. One should interpret soft-max as smooth version of the max function (hence the name). The soft-max function is often used for **classification problems**. The idea is to choose the label, which has the highest probability predicted by the model.

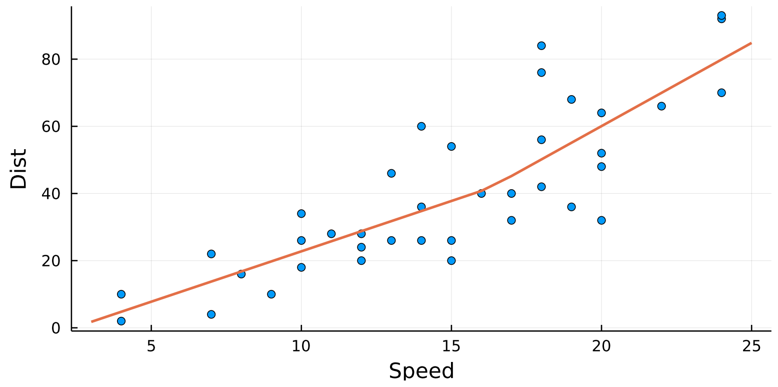


Figure 3.3: The picture shows a function describing the data in Fig. 3.1 that was computed using a neural network.

Exercise 3.1. Prove Theorem 3.13. **Hint:** Adapt the proof of Theorem 3.12.

Exercise 3.2. Reformulate and prove Theorem 3.8 for linear models $\mathbb{R}^D \to \mathbb{R}^N$ and the quadratic loss $\ell(y,\hat{y}) = ||y-\hat{y}||^2$.

Exercise 3.3. Prove Proposition 3.14.

Exercise 3.4. Prove Proposition 3.18. **Hint:** By Eq. (1.2.1) the marginal density is given by $P(y \mid x, X, Y) = \int_{\mathbb{R}^P} P(y \mid x, \theta) \cdot P(\theta \mid X, Y) \, \mathrm{d}\theta$.

Exercise 3.5. From the `RDatasets` package in `Julia` load the `pressure` data set. This data set contains the variables `temperature` and `pressure`, which give the values of pressure of mercury depending on temperature. The **Antoine Equation** is a simple model for this dependency:

$$\log(\texttt{pressure}) = a - \frac{b}{\texttt{temperature}}.$$

Set up and solve a regression problem to estimate a and b.

3.3 Support Vector Machines

At the end of the last section we discussed how to use neural networks for classification problems. Another machine learning method for classification are **support vector machines (SVM)**.

We discuss SVMs in the context of **binary classification**. In this setting the labels $y_1, \dots, y_n$ are elements in $\{1,-1\}$ and we want to find a model that fits the data $(x_1,y_1),\dots,(x_n,y_n) \in \mathbb{R}^D \times \{-1,1\}$. We will work here with finding a deterministic model as in Definition 3.4 given by

$$f_\theta : \mathbb{R}^D \to \{-1,1\}.$$

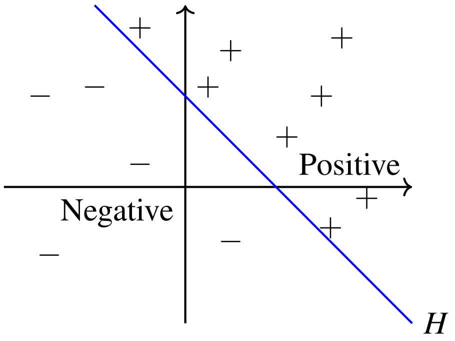
The generalization to $k \geq 2$ categories is then given by taking several of such functions, so that we map into $\{-1,1\}^\ell$ for $2^\ell \geq k$.

SVMs provide the following family of parametrized functions:

$$f_\theta(x) = \operatorname{sgn}(\langle a,x\rangle + b) \quad \text{for} \quad \theta = (a,b) \in \mathbb{R}^D \times \mathbb{R},\ a \neq 0. \tag{3.3.1}$$

Thus, the idea behind SVMs is to find a hyperplane $H = \{x \in \mathbb{R}^D \mid \langle a,x\rangle + b = 0\}$ and split the data into on each side of H. Notice that we can define the same plane H if we take $\langle a,a\rangle = 1$, since $(\lambda a, \lambda b)$ and (a,b) define the same hyperplane.

Example 3.21. Suppose that our input data is contained in the plane $\mathbb{R}^2$. Consider the line $H = \{x = (x_1,x_2) \in \mathbb{R}^2 \mid x_1 + x_2 - 2 = 0\}$ shown in blue below. Then H splits the plane in the following two regions.



We classify data in $\mathbb{R}^2$ by assigning them a plus or a minus sign. The data points on the upper right of the plane are labeled with a plus, because for them $x_1 + x_2 - 2 > 0$. The other data points are labeled with a minus sign, because here $x_1 + x_2 - 2 < 0$.

The plan for the rest of this lecture is now to find a suitable of way computing parameters for the model Eq. (3.3.1). To do so, we first make two small observations.

Lemma 3.22. *Let $x \in \mathbb{R}^D$ and $\theta = (a,b) \in \mathbb{R}^D \times \mathbb{R}$. For $y \in \{-1,1\}$, we have $y = f_\theta(x)$ if and only if $y(\langle a,x\rangle + b) > 0$.*

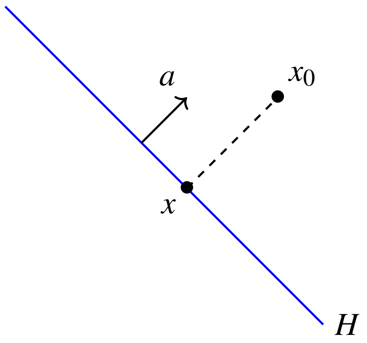


Figure 3.4: Geometrical setup in the proof for Lemma 3.23.

Proof. In the case that $f_\theta(x) = -1$, we have $\langle a,x\rangle + b < 0$ and $y = -1$, which implies $y(\langle a,x\rangle + b) > 0$. Similarly when $f_\theta(x) = 1$, then $\langle a,x\rangle + b > 0$ and $y = 1$ so that in this case we also have $y(\langle a,x\rangle + b) > 0$. □

We can see from Example 3.21 that a hyperplane H which separates the data doesn't need to be unique. SVMs select the Hyperplane which maximizes the distance to the data. The following lemma helps us to formulate the right optimization problem.

Lemma 3.23. *Let $x \in \mathbb{R}^D$ and $H = \{z \in \mathbb{R}^D \mid \langle a,z\rangle + b = 0\}$ for $\langle a,a\rangle = 1$. The Euclidean distance from x to H is $y(\langle a,x\rangle + b)$ where $y = f_\theta(x)$.*

Proof. The geometrical setup of this proof is depicted in Fig. 3.4. Let $z \in H$ be the point which minimizes the distances to x. So we can write $x = z + \varepsilon r a$ where $r = \|x - z\| \geq 0$ and $\varepsilon \in \{-1, 1\}$. Since $z \in H$ we compute

$$\langle a,x\rangle + b = \langle a, z + ra\rangle + b = \langle a,z\rangle + \varepsilon r\langle a,a\rangle + b = 0 + \varepsilon r.$$

Note if $\varepsilon = -1$, then $y = -1$ and therefore $y(\langle a,x\rangle + b) = y\varepsilon r = r$. Similarly if $\varepsilon = 1$ then $y = 1$ and therefore $y(\langle a,x\rangle + b) = y\varepsilon r = r$.

□

In order to compute the hyperplane that maximizes the distance to the data, we can now write it in terms of solving the following optimization problem:

$$\begin{aligned} &\max_{\theta=(a,b)} \; r \\ &\text{s.t. } y_k(\langle a,x_k\rangle + b) \geq r, \text{ for } k = 1,\ldots,n, \\ &\quad \text{and } \langle a,a\rangle = 1, r \geq 0. \end{aligned} \tag{3.3.2}$$

The optimization problem of Eq. (3.3.2) is often solved via an equivalent problem by normalizing the value of r. Namely if $a' = \frac{a}{r}$ and $b' = \frac{b}{r}$, then the constrains now ask for $y_k(\langle a', x_k\rangle + b') \geq 1$. Then in this case $\|a'\| = \frac{1}{r}$, so we can either maximize $\frac{1}{\|a'\|}$, or minimize $\|a'\|$. This leads to the following parameter finding problem.

Definition 3.24. The **Hard Margin SVM** is given by the optimization problem

$$\begin{aligned} &\min_{\theta=(a,b)} \|a\|^2 \\ &\text{s.t. } y_k(\langle a, x_k\rangle + b) \geq 1, \text{ for } k = 1, \ldots, n. \end{aligned}$$

An issue with Hard Margin SVM when working with noisy data is that it does not allow for outliers. To compensate for this, we will introduce a **slack variable** ξ_k.

Definition 3.25. The **Soft Margin SVM** is given by the optimization problem

$$\begin{aligned} &\min_{a,b,\xi} \|a\|^2 + C\sum_{k=1}^{n} \xi_k \\ &\text{s.t. } y_k(\langle a, x_k\rangle + b) \geq 1 - \xi_k, \text{ for } \xi_k \geq 0, k = 1, \ldots, n. \end{aligned}$$

The parameter C is called the **regularization parameter**.

Here, the regularization parameter is not taken as a parameter of the model but as a hyperparameter. Moreover, soft Margin SVM can understood as empirical risk minimization. This is proved in the next proposition.

Proposition 3.26. *Consider the* ***Hinge loss*** $\ell(y, \hat{y}) = \max\{0, 1 - y \cdot \hat{y}\}$. *The Soft Margin SVM can be derived from ERM with respect to a regularized Hinge loss.*

Proof. Exercise 3.6. □

The SVMs in Definition 3.24 and Definition 3.25 both fall in the category of **Primal SVMs**. Another formulation is the **Dual SVM** which we now work toward defining. We first define the Lagrange function for Soft Margin SVM:

$$\mathscr{L}(a, b, \xi, \alpha, \beta) = \|a\|^2 + C\sum_{k=1}^{n} \xi_k - \sum_{k=1}^{n} \alpha_k \left(y_k(\langle a, x_k\rangle + b) - (1 - \xi_k)\right) - \sum_{k=1}^{n} \beta_k \xi_k, \tag{3.3.3}$$

where $\alpha_k, \beta_k \geq 0$. The KKT conditions say that the optimum occurs when

$$\frac{\partial \mathscr{L}}{\partial a} = 0, \quad \frac{\partial \mathscr{L}}{\partial b} = 0, \quad \frac{\partial \mathscr{L}}{\partial \xi} = 0.$$

We will write these equations by rewriting $\mathscr{L}$ using

$$u := (\tfrac{1}{2}\alpha_k y_k)_{k=1}^n, \quad v := (\langle a, x_k\rangle + b)_{k=1}^n, \quad e = (1,\dots,1)^T \in \mathbb{R}^n,$$

which gives

$$\mathscr{L} = \langle a,a\rangle + C\langle \xi, e\rangle - 2\langle u,v\rangle - \langle \alpha+\beta, \xi\rangle + \langle \alpha, e\rangle.$$

Therefore

$$\begin{aligned}
0 &= \frac{\partial L}{\partial a} = 2a - 2\sum_{k=0}^n u_k x_k && \Rightarrow \quad a = \sum_{k=1}^n u_k x_k \\
0 &= \frac{\partial L}{\partial b} = -2\langle u, e\rangle && \Rightarrow \quad 0 = \sum_{k=1}^n u_k \\
0 &= \frac{\partial L}{\partial \xi} = Ce - (\alpha+\beta) && \Rightarrow \quad Ce = \alpha+\beta \quad \Rightarrow \quad \alpha_i \le C
\end{aligned} \tag{3.3.4}$$

Remark 3.27. The equation $\frac{\partial \mathscr{L}}{\partial a} = 0$ is the meaning behind the name Support Vector Machine. Namely, the x_i with $u_k \neq 0$ and equivalently $\alpha_k \neq 0$ are the **support** of the vector a.

First, observe that from Eq. (3.3.4) it follows that.

$$\langle u,v\rangle = \sum_{k=1}^n u_k\langle a,x_k\rangle + b\sum_{k=1}^n u_k = \sum_{k=1}^n u_k\langle a,x_k\rangle = \langle \sum_{k=1}^n u_k x_k, \sum_{k=1}^n u_k x_k\rangle.$$

Now, we put the equations in Eq. (3.3.4) into $\mathscr{L}$ and get

$$\begin{aligned}
\mathscr{L} &= \langle a,a\rangle + C\langle \xi, e\rangle - 2\langle u,v\rangle - \langle \alpha+\beta, \xi\rangle + \langle \alpha, e\rangle \\
&= \langle \sum_{k=1}^n u_k x_k, \sum_{k=1}^n u_k x_k\rangle - 2\langle u,v\rangle + \langle \alpha, e\rangle \\
&= -u^T G u + \langle \alpha, e\rangle \quad \text{where} \quad G = (\langle x_k, x_\ell\rangle)_{k,\ell=1}^n, \text{ and } 0 \le \alpha_i \le C
\end{aligned}$$

This all leads to the following definition of the Dual SVM.

Definition 3.28. The **Dual SVM** is given by the optimization problem

$$\begin{aligned}
&\max_\alpha -u^T G u + \sum_{k=1}^n \alpha_k \\
&\text{s.t.} \sum_{k=1}^n \alpha_k y_k = 0, \text{ and } 0 \le \alpha_i \le C,
\end{aligned}$$

where $u := (\frac{1}{2}\alpha_k y_k)_{k=1}^n$ and $G = (\langle x_k, x_\ell\rangle)_{k,\ell=1}^n$.

Remark 3.29. *Why a maximum?* The solution for the optimization problem for the Primal SVM is given by

$$\min_{a,b,\xi}\max_{\alpha,\beta}\mathscr{L}(a,b,\xi,\alpha,\beta). \tag{3.3.5}$$

But the function in Definition 3.28 is independent of a,b,ξ, and β, so that we can remove the minimization step.

When the Dual SVM problem has been solved, we can use the following result to compute direction of the hyperplane a and the offset b.

Proposition 3.30. *Let α be an optimal solution from the Dual SVM. Then, we have optimal values for the Soft Margin SVM from Definition 3.25 by setting*

1. *$a^* := \sum_{k=1}^n u_k x_k$, where $u_k = y_k\alpha_k$;*
2. *b^* is the median value of $y_k - \langle a^*, x_k\rangle$ for all k with $\alpha_k \neq 0$.*

Proof. The formula for a^* follows from Eq. (3.3.4) Using that we have $\alpha+\beta = C$ in the optimal value, the Lagrange function from Eq. (3.3.3) becomes

$$\mathscr{L} = \|a\|^2 - \sum_{k=1}^n \alpha_k\left(y_k(\langle a,x_k\rangle + b) - 1\right).$$

By Eq. (3.3.5) we are maximizing the Lagrangian over α. Therefore,

$$y_k(\langle a,x_k\rangle + b) - 1 \leq 0 \quad \Rightarrow \quad \alpha_k = 0. \tag{3.3.6}$$

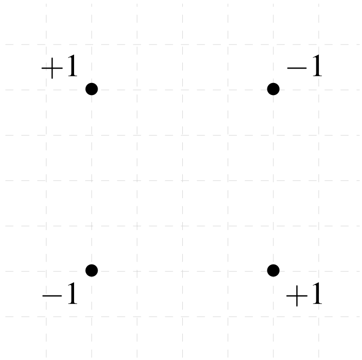
This implies

$$\mathscr{L} = \|a\|^2 - \sum_{k\colon \alpha_k>0} \alpha_k\left(y_k(\langle a,x_k\rangle + b) - 1\right) = \|a\|^2 - C\sum_{k\colon \alpha_k>0}^{n}\left(y_k(\langle a,x_k\rangle + b) - 1\right),$$

as the term in the middle is maximized for setting all $\alpha_k = C$. Now, by Eq. (3.3.6) the summands on the right are all nonnegative. Using that $|y_k| = 1$ for all k we get that b^* is a point on the real line that minimizes $\sum_{k\colon \alpha_k>0}|y_k - \langle a^*, x_k\rangle - b|$. The median of $y_k - \langle a^* x_k\rangle$ for $\alpha_k \neq 0$ is a minimizer (see Exercise 3.7). □

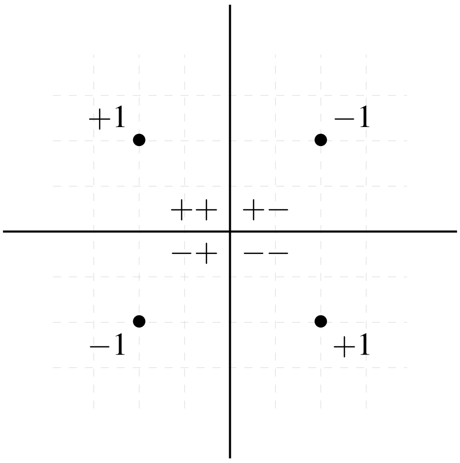
It is not always possible to find a suitable hyperplane using Soft-Margin SVM. We give an example of a situation where this occurs, and potential solutions.

Example 3.31. There is no hyperplane separating the following 4 points into 2 classes. Even using Soft-Margin SVM the 2 classes can't be well separated.

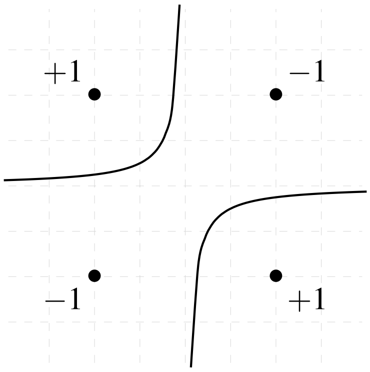


We can think of two solutions:

Solution 1: Combine multiple hyperplanes. The problem with this first solution is that we have multiple labels for each class.



Solution 2: Separate through curved hypersurfaces. In the below example we separate the data via a hyperbola given by a polynomial of degree 2.



For solution 2 above the idea is to combine an SVM with a **feature map**

$$\phi : \mathbb{R}^D \to \mathbb{R}^M$$

(as we did in nonlinear Regression), and compute an SVM for the modified data points $(\phi(x_1), y_1), \dots, (\phi(x_n), y_n)$. In this situation the Dual SVM is well suited. In particular, in the formulation of Definition 3.28 only the inner product between input variables occurs. So it is enough to know the value of the inner products $\langle \phi(x_k), \phi(x_\ell) \rangle$. These are parametrized through **positive semi-definite matrices**.

Lemma 3.32. *Let $G \in \mathbb{R}^{n\times n}$. Then G is positive semi-definite if and only if there exists M and vectors $z_1, \dots, z_n \in \mathbb{R}^M$ with $G = (\langle z_i, z_j \rangle)_{i,j=1}^n$.*

Proof. Let $Z \in \mathbb{R}^{M\times n}$ be the matrix whose columns are the z_i. If $G = Z^T Z$, for every $w \in \mathbb{R}^n \setminus \{0\}$ we then have $w^T G w = (Zw)^T Zw \geq 0$. On the other hand, if G is positive semi-definite, we can find a Cholesky-decomposition $G = Z^T Z$. □

In the context of SVMs, the positive definite matrices are also called **kernels**. The function

$$\kappa(x,y) := \langle \phi(x_k), \phi(x_\ell) \rangle$$

is called **kernel map**. By Proposition 3.30 we have the optimal value $a^* = \frac{1}{2}\sum_{i=1}^n u_i \phi(x_i)$. Let us define

$$\psi(x) := \frac{1}{2}\sum_{i=1}^n u_i\, \kappa(x_i, x).$$

Then, we have $\psi(x) = \langle a^*, \phi(x) \rangle$ by linearity. The optimal value for b is then the median of $y_k - \psi(x_k)$ for $\alpha_k \neq 0$. Moreover, we can evaluate the model from Eq. (3.3.1) as $f_\theta(x) = \operatorname{sgn}(\psi(x) + b)$. All this leads to the following algorithm.

Algorithm 3.2: Binary classification by Dual SVM.

```
1 Input: Training data $(x_1, y_1), \dots, (x_n, y_n) \in \mathbb{R}^D \times \{-1, 1\}$, a kernel map
      $\kappa(x,y)$, and a regularization parameter $C$.
2 Output: A function $f : \mathbb{R}^D \to \{-1, 1\}$ of the form $f(x) = \operatorname{sgn}(\langle a, \phi(x) \rangle + b)$.
3 Compute the kernel matrix $G = (\kappa(x_i, x_j))_{1\leq i,j\leq n}$;
4 Using $C$ and $G$ solve the Dual SVM problem from Definition 3.28 for $\alpha \in \mathbb{R}^n$;
5 Define the function $\psi(x) := \frac{1}{2}\sum_{i=1}^n y_i \alpha_i\, \kappa(x_i, x)$;
6 Take $b$ as the median of $\{y_k - \psi(x_k) \mid \alpha_k \neq 0.\}$;
7 Return $f(x) = \operatorname{sgn}(\psi(x) + b)$.
```

An important example of a kernel is the **polynomial kernel**.

Lemma 3.33. *Consider the feature map $\phi(x) = (x_1^{i_1} \cdots x_D^{i_D})_{(i_1,\dots,i_D)\in I_D}$, where the index set is $I_D := \{(i_1,\dots,i_D) \in \mathbb{N}^D \mid 0 \le i_1 + \cdots + i_D \le d\}$. I.e., $\phi(x)$ gives all monomials in x of degree at most d. Then for $u, v \in \mathbb{R}^D$*

$$\langle \phi(u), \phi(v) \rangle = (\langle u, v \rangle + 1)^d.$$

Proof. We write $u_0 := 1$ and $v_0 := 1$. Then

$$\begin{aligned}\langle \phi(u), \phi(v) \rangle = \phi(u)^T \phi(v) &= \sum_{0 \le i_0 + \cdots + i_D = d} u_0^{i_0} u_1^{i_1} \cdots u_D^{i_D} v_0^{i_0} v_1^{i_1} \cdots v_D^{i_D} \\ &= \sum_{0 \le i_0 + \cdots + i_D = d} (u_0 v_0)^{i_0} (u_1 v_1)^{i_1} \cdots (u_D v_D)^{i_D} v \\ &= (u_0 v_0 + u_1 v_1 + \cdots + u_D v_D)^d = (\langle u, v \rangle + 1)^d.\end{aligned}$$

□

Example 3.34. We load the `FashionMNIST`[2] dataset. It contains images of several fashion items. In this example we consider images of coats and sandals (8 of them are shown in Fig. 3.5). Each image is given by gray values for 28 × 28 pixels. Thus, we have data in $\mathbb{R}^D$ with $D = 28^2 = 784$.

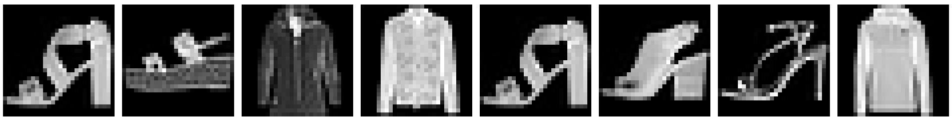


Figure 3.5: Images of coats and sandals from the `FashionMNIST` dataset.

The goal is to obtain a function that classifies the data. The `FashionMNIST` dataset contains 6000 images of coats and 6000 images sandals. We pick a subset of $n = 50$ images as training data and another subset of 50 images as test data.

First, we use an SVM with parameter $C = 5$ and kernel map $\kappa(x_1, x_2) = \langle x_1, x_2 \rangle$. Next, we use a neural network with three layers given by $\mathbb{R}^D \to \mathbb{R}^{20}$, $\mathbb{R}^{20} \to \mathbb{R}^{15}$ and $\mathbb{R}^{20} \to \mathbb{R}^2$. The inner layers feature as ReLu actication function. For the last layer we use the softmax activation function from Eq. (3.2.9). This way, the neural network

[2] `https://github.com/zalandoresearch/fashion-mnist`

produces a probability distribution for the two classes. We use the **crossentropy** loss function $\ell(p,q) = -p_1 \log(q_1) - p_2 \log(q_2)$ (see, e.g., [Mur13, Section 8.3.1]). The classifier chooses the class which has the largest probability.

The results of the two classifiers evaluated on the test data are shown in Fig. 3.6.

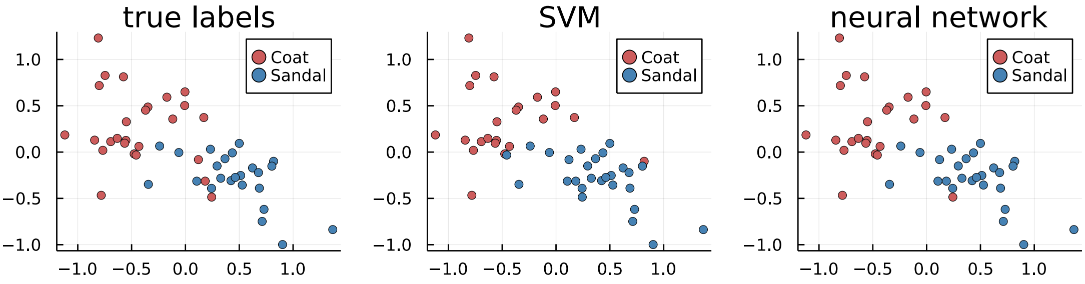


Figure 3.6: The results of the two classifiers from Example 3.34 evaluated on the test data. The SVM based classfier identfies images 90% correctly. The neural network based classfier identfies images 96% correctly.

Exercise 3.6. Prove Proposition 3.26.

Exercise 3.7. Let $w_1, \dots, w_n \in \mathbb{R}$. Prove that the median of the w_i minimizes the aggregated distances $d(v) = \sum_{i=1}^{n} |w_i - v|$.

Exercise 3.8. The `MLDatasets.jl`[3] package provides the `MNIST` dataset[4]. This dataset contains images of handwritten digits. Load the training data for images of zeros and ones and implement an algorithm that learns to separate these two classes. After the learning step let your algorithm predict the labels of test data points.

3.4 Principal Component Analysis

In this lecture we consider **Principal Component Analysis (PCA)**, which is one of the central methods for **dimensionality reduction**. This means the following:

Given data pairs $x_1, \dots, x_n \in \mathbb{R}^D$, with the help of PCA we model the data so that we reduce the number of parameters that the data describes. This can have several motivations. For instance, we could be interested in **data compression**, and reducing

[3] `https://github.com/JuliaML/MLDatasets.jl`

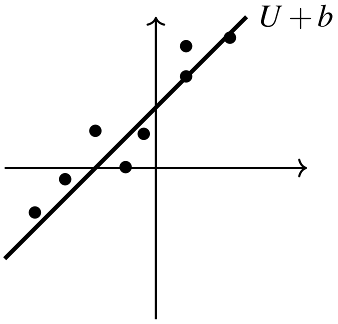
[4] `http://yann.lecun.com/exdb/mnist/`

parameters would reduce the memory for storing the data. Another motivation is to interpret the parameters as geometric information, so that the goal here would be to **learn the shape of the data**.

For PCA we will not be taking into account response variables $y_1,\dots,y_n \in \mathbb{R}^N$. Learning without using response variables is called **unsupervised learning**. By contrast, the settings from the previous sections are summarized as **supervised learning**.

The basic idea of PCA is to find a linear space $U \subseteq \mathbb{R}^D$ of dimension $d \ll D$ and a vector $b \in \mathbb{R}^D$ so that $x_1,\dots,x_n$ lay “nearby” $U+b$.

Example 3.35. Here is a small graphic when $D=2$ and $d=1$. The line is represented by $U+b$, and the points are the points $x_1,\dots,x_8$.



For the moment, we take d as a fixed input parameter. We will discuss later how this parameter can be chosen.

As depicted in Example 3.35, the assumption in PCA is that the data centers around a low-dimensional linear supspace, and the goal is to determine this subspace. This assumption, however, is not always fulfilled. Data points can also lie on a **nonlinear** subspace. For this reason, we again work with a nonlinear feature map $\phi : \mathbb{R}^D \to \mathbb{R}^M$, and we set

$$z_i := \phi(x_i), \quad 1 \le i \le n.$$

For instance, if we take the polynomial feature map from Lemma 3.33, then finding a linear subspace for the z_i means finding a low-dimensional **algebraic variety** describing the data; i.e., the vanishing set of a system of polynomial equations.

Example 3.36. We consider a sample of $n=100$ points from the unit circle in $\mathbb{R}^2$ plus Gaussian noise. The point sample is shown in Fig. 3.7. We apply PCA combined with the nonlinear feature map $\phi((x_1,x_2)) = (1,x_1,x_2,x_1x_2,x_1^2,x_2^2) \in \mathbb{R}^6$. By doing so, we find that the data lies in a hyperplane in $\mathbb{R}^6$. The equation of this hyperplane yields approximately the nonlinear equation $x_1^2 + x_2^2 - 1 = 0$.

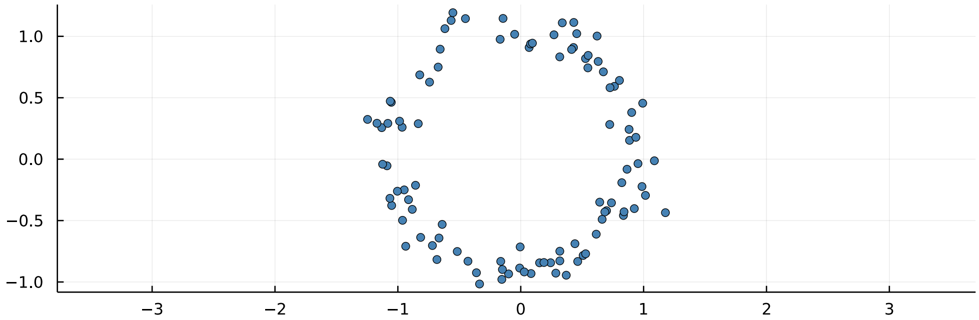


Figure 3.7: A noisy sample of $n = 100$ points from the unit circle.

In the following, we assume that the data $x_1, \dots, x_n$ are independent samples from an (unknown) random variable $x \in \mathbb{R}^D$, and we set $z := \phi(x)$, so that $z_1, \dots, z_n$ are independent samples from z.

We also denote the **expected value** $\mu := \mathbb{E}z \in \mathbb{R}^M$ and the **covariance matrix** of $z = (z^{(1)}, \dots, z^{(M)})^T$:

$$\Sigma := \begin{bmatrix} \operatorname{Cov}(z^{(1)}, z^{(1)}) & \cdots & \operatorname{Cov}(z^{(1)}, z^{(M)}) \\ & \ddots & \\ \operatorname{Cov}(z^{(M)}, z^{(1)}) & \cdots & \operatorname{Cov}(z^{(M)}, z^{(M)}) \end{bmatrix} \in \mathbb{R}^{M \times M}.$$

By Exercise 3.9, Σ is positive semi-definite, and thus diagonalizable with orthogonal matrices and has eigenvalues $\lambda_1 \geq \cdots \geq \lambda_M \geq 0$.

In practice we rarely know Σ exactly, because we only have a finite data sample available. Thus we need to approximate Σ through the **empirical covariance matrix**.

Definition 3.37. Let $z_1, \dots, z_n \in \mathbb{R}^M$.

1. The **empirical average** is given by the arithmetic mean $\bar{z} = \frac{1}{n}(z_1 + \cdots + z_n)$.
2. The **empirical covariance matrix** is $S = (s_{ij}) \in \mathbb{R}^{M \times M}$ with

$$s_{ij} = \frac{1}{n} \sum_{k=1}^{n} ((z_k)_i - \bar{z}_i)((z_k)_j - \bar{z}_j).$$

Remark 3.38. A common approach is to **standardize the data**. This means to replace for every k the i-th entry $(z_k)_i$ by $(z_k)_i/\sqrt{s_{ii}}$, and then using the modified data for PCA.

As before, let us denote the feature matrix (see Eq. (3.2.4))

$$\Omega = \begin{bmatrix}\phi(x_1) & \dots & \phi(x_n)\end{bmatrix}^T = \begin{bmatrix} z_1 & \dots & z_n\end{bmatrix}^T \in \mathbb{R}^{n\times P}.$$

Writing also $e = (1,\dots,1)^T \in \mathbb{R}^n$, then we can write

$$\bar{z} = \frac{1}{n}\Omega^T e, \quad \text{and} \quad S = \frac{1}{n}(\Omega^T - \bar{z}e^T)(\Omega^T - \bar{z}e^T)^T. \tag{3.4.1}$$

In particular, the second equation together with Lemma 3.32 shows that the empirical covariance S, like Σ, is also positive semi-definite.

We will consider two interpretations of “nearby”. The first is to select U as the space which maximizes the variance in the sense of the following definition.

Definition 3.39. Let $z \in \mathbb{R}^M$ be a random variable with expectation $\mu := \mathbb{E}z \in \mathbb{R}^M$. Let $1 \le d < M$. A **space of maximal variance** of dimension d is a linear subspace $U \subset \mathbb{R}^M$ of dimension $d = \dim(U)$ such that $U \in \operatorname{argmax}_{U:\dim(U)=d} \mathbb{E}\left\|P_U(z-\mu)\right\|^2$, where P_U is the orthogonal projection onto U.

We have the following theorem.

Theorem 3.40. *Let $\lambda_1 \ge \dots \ge \lambda_M \ge 0$ be the eigenvalues of the covariance matrix $\Sigma \in \mathbb{R}^{M\times M}$, and let $u_i \in \mathbb{R}^M$ be the eigenvectors associated to λ_i so that $\langle u_i, u_j\rangle = \delta_{ij}$. Then*

$$U = \operatorname{span}\{u_1,\dots,u_d\}$$

is a space of maximal variance of dimension d. Furthermore, U is uniquely determined in the case that $\lambda_d > \lambda_{d+1}$.

As mentioned above, in practice we do not have Σ available and instead use the eigenvectors of S for computing U.

Before we prove this theorem, let us discuss for a brief moment the choice of d. The statement of Theorem 3.40 implies that d should be chosen, such that the data spreads out in the direction of the eigenvectors for $\lambda_1,\dots,\lambda_d$, but not in the directions of the eigenvectors for $\lambda_{d+1},\dots,\lambda_M$. One choice for such a d is to take $\lambda_d > 0$ and $\lambda_{d+1} \approx 0$. Alternatively, we can choose d such that λ_d/λ_{d+1} or $\lambda_d - \lambda_{d+1}$ is maximized.

Proof of Theorem 3.40. Let U be a subspace of dimension d and $u_1, \dots, u_d$ an orthonormal basis of U (that is $\langle u_i, u_j \rangle = \delta_{ij}$). Also let $A = \begin{bmatrix} u_1 & \dots & u_d \end{bmatrix} \in \mathbb{R}^{M \times d}$. By Corollary 1.6, the orthogonal projection onto U is given by $P_U = AA^\dagger$ and, by Proposition 1.7, we have $A^\dagger = (A^TA)^{-1}A^T = A^T$. Therefore,

$$P_U = AA^T,$$

and so

$$\begin{aligned} \|P_U(z-\mu)\|^2 &= \left\|AA^T(z-\mu)\right\|^2 \\ &= (AA^T(z-\mu))^T(AA^T(z-\mu)) \\ &= (z-\mu)^TAA^TAA^T(z-\mu) \\ &= (z-\mu)^TAA^T(z-\mu); \end{aligned}$$

the last line because $A^TA = \mathbf{1}_d$. Now, using that $A^TA = \sum_{i=1}^d u_iu_i^T$ this gives

$$\|P_U(z-\mu)\|^2 = \sum_{i=1}^d (z-\mu)^T u_iu_i^T(z-\mu) = \sum_{i=1}^d u_i^T(z-\mu)(z-\mu)^Tu_i.$$

Using Exercise 3.9 and that the expected value is linear we get

$$\mathbb{E}\,\|P_U(z-\mu)\|^2 = \sum_{i=1}^d u_i^T\,\mathbb{E}[(z-\mu)(z-\mu)^T]u_i = \sum_{i=1}^d u_i^T\Sigma u_i.$$

Thus, a space of maximal variance $U = \operatorname{span}\{u_1, \dots, u_d\}$ maximizes the scalar function $\sum_{i=1}^d u_i^T\Sigma u_i$. This function is maximized by taking the eigenvectors of the largest eigenvalues of Σ. We prove this.

We maximize with Lagrange Multipliers. Let

$$\mathscr{L}(u_1, \dots, u_d, \ell_{ij}) = \sum_{i=1}^d u_i^T\Sigma u_i - \sum_{1 \le i \le j \le d} (u_i^Tu_j - \delta_{ij})\ell_{ij}.$$

We compute

$$\frac{\partial \mathscr{L}}{\partial u_j} = 2\Sigma u_j - 2u_j\ell_{jj} - \sum_{i<j} u_j\ell_{ij} = 0 \quad 1 \le j \le d. \tag{3.4.2}$$

$$\frac{\partial \mathscr{L}}{\partial \ell_{ij}} = u_i^Tu_j - \delta_{ij} = 0 \quad 1 \le i \le d. \tag{3.4.3}$$

From Eq. (3.4.2) for $j = 1$ it follows that $\Sigma u_1 = \ell_{11} u_1$ thus $\ell_{11} = \lambda_1$. Similarly,

$$2u_1^T \Sigma u_2 - \ell_{12} = 0 \Rightarrow \ell_{12} = 0 \Rightarrow \Sigma u_2 = \ell_{22} u_2 \Rightarrow \ell_{22} = \lambda_2.$$

By continuing this process, we see that $\sum_{i=1}^d u_i^T \Sigma u_i = \sum_{i=1}^d \lambda_i$. This expression is maximized by taking $\lambda_1 \geq \ldots \geq \lambda_d \geq \lambda_{d+1} \geq \ldots \geq \lambda_M$. Finally, if $\lambda_d > \lambda_{d+1}$, then U is uniquely determined as the sum of the eigenspaces for $\lambda_1, \ldots, \lambda_d$. □

Definition 3.41. The $u_1, \ldots, u_d$ in Theorem 3.40 are called **principal components** of Σ.

In the first approach we computed the principle components of S, thus obtaining $U \subset \mathbb{R}^D$. The data would then be compressed by applying the map

$$z_i \mapsto P_U(z_i - \bar{z}) + \bar{z}.$$

In the second approach, instead of maximizing the variance, we want to minimize the squared distance to the data points.

$$\sum_{i=1}^n \|(z_i - \bar{z}) - P_U(z_i - \bar{z})\|^2. \tag{3.4.4}$$

The next theorem shows that, although this approach is conceptually different from maximizing the variance, we get the same minimizer as in Theorem 3.40 (when replacing the covariance matrix Σ by the empirical covariance matrix S).

Theorem 3.42. *Let $\lambda_1, \ldots, \lambda_M \geq 0$ be the eigenvalues of the empirical covariance matrix S. Let u_i be eigenvectors of λ_i so that $\langle u_i, u_j \rangle = \delta_{ij}$. Then $U = \operatorname{span}\{u_1, \ldots, u_d\}$ minimizes the squared distance in Eq.* (3.4.4). *If $\lambda_d > \lambda_{d+1}$, the subspace U is uniquely determined.*

Proof. Let $u_1, \ldots, u_M$ be an orthonormal basis of $\mathbb{R}^M$; i.e., $\langle u_i, u_j \rangle = \delta_{ij}$. Let $w_i = z_i - \bar{z}$. As in the proof of Theorem 3.40 let $A = [u_1, \ldots, u_d] \in \mathbb{R}^{M \times d}$ so that $P_U = AA^T$. Then, w_i is the i-th column of $W := \Omega^T - \bar{z}e^T$. By Eq. (3.4.1), we have

$$WW^T = nS.$$

We note that $AA^T = \sum_{i=1}^d u_i u_i^T$ and $\mathbf{1}_M = \sum_{i=1}^M u_i u_i^T$. Therefore, we also have

$$W - AA^T W = \Big(\sum_{i=d+1}^M u_i u_i^T \Big) W.$$

Observe that

$$W - P_U W = W - AA^T W = \begin{bmatrix} w_1 - P_U(w_1) & \dots & w_n - P_U(w_n) \end{bmatrix}.$$

Therefore, Eq. (3.4.4) can be rewritten as $\mathrm{Trace}((W - P_U W)^T (W - P_U W))$. We get

$$\mathrm{Trace}((W - P_U W)^T (W - P_U W)) = \sum_{i=d+1}^{M} \mathrm{Trace}\left(u_i u_i^T W W^T\right) = n \sum_{i=d+1}^{M} u_i^T S u_i.$$

As in the proof from Theorem 3.40 we show that the u_i must be eigenvectors of S so that $U = \mathrm{span}\{u_1, \dots, u_d\}$ for $Su_i = \lambda_i u_i$ and $\lambda_1 \geq \dots \geq \lambda_M \geq 0$. The uniqueness statement follows as in the proof of Theorem 3.40. □

In both cases we need to compute eigenvectors of the covariance matrix S. In the case that $n \ll M$, so that the number of data points is significantly smaller than the dimension of the data, then we can proceed as follows. As in the proof of Theorem 3.42, let $W := \Omega^T - \bar{x}e^T$, so that the empirical covariance matrix is $S = \frac{1}{n} W W^T$ by Eq. (3.4.1). We assume noisy data, so that the rank of W can be assumed to n. Let also

$$W = UDV^T$$

be the **singular value decomposition** (see Theorem 1.8) of W, where

$$U \in \mathbb{R}^{M \times n}, \quad V \in \mathbb{R}^{n \times n}, \quad D = \mathrm{diag}(\sigma_1, \dots, \sigma_n)$$

with singular values $\sigma_1 \geq \dots \geq \sigma_n \geq 0$. Then, we have

$$S = \frac{1}{n} U D^2 U^T \in \mathbb{R}^{M \times M}.$$

This shows that, using PCA, we actually compute a low-dimensional representation of the data using SVD. We can exploit this: Consider the eigendecomposition

$$W^T W = V D^2 V^T \in \mathbb{R}^{n \times n}.$$

Since the i-th columns of U and V are related by $u_i = Wv_i / \|Wv_i\|$, this shows that it suffices to compute an eigendecomposition of an $n \times n$-matrix, rather than a decomposition of an $M \times M$-matrix. Moreover, $W^T W$ can be computed using only the kernel map.

Lemma 3.43. *Let $\kappa(x_1,x_2) = \langle\phi(x_1),\phi(x_2)\rangle$ be the kernel map for ϕ, and let the kernel matrix be $G = (\kappa(x_k,x_\ell))_{1\le k,\ell\le n}$. Then, for $W = \Omega^T - \bar{z}e^T$ we have*

$$W^TW = (\mathbf{1}_n - \tfrac{1}{n}ee^T)G(\mathbf{1}_n - \tfrac{1}{n}ee^T),$$

where, as before, $e = (1,\dots,1)^T \in \mathbb{R}^n$.

Proof. We get from Eq. (3.4.1) that $W = \Omega^T(\mathbf{1}_n - \frac{1}{n}ee^T)$. This implies,

$$W^TW = (\mathbf{1}_n - \tfrac{1}{n}ee^T)\Omega\Omega^T(\mathbf{1}_n - \tfrac{1}{n}ee^T),$$

and we have $\Omega\Omega^T = G$. □

In particular, this lemma shows that W^TW will always have rank at most $n-1$.

Example 3.44. We load the `CIFAR10`[5] dataset. It contains images of various items. We select $n = 50$ images of trucks for PCA (6 of them are shown in Fig. 3.8).

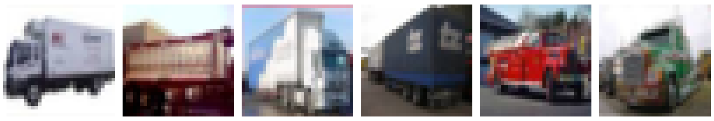


Figure 3.8: Images of trucks from the `CIFAR10` dataset.

Each image is given as RGB values for 32×32 pixels. We therefore have data in $\mathbb{R}^D$ for $D = 32^2\cdot 3 = 3072$. We want to apply PCA using $\kappa(x_1,x_2) = (\langle x_1,x_2\rangle + 1)^4$ as kernel map. By Lemma 3.33, this corresponds to the feature map $\phi(x)$ that evaluates all monomials of degree at most 4 at $x\in\mathbb{R}^D$. There are $\binom{D+4}{4} = 3,722,945,108,225 > 3\cdot 10^{12}$ such monomials. Evaluating ϕ directly is not an option. Instead, we use Lemma 3.43 to compute and eigendecomposition of W^TW. This gives eigenvectors $\lambda_1 \ge \dots \ge \lambda_n$. We compute the relative differences $d_i := (\lambda_i - \lambda_{i-1})/\lambda_i$ for $2\le i\le n$. Their values are shown in Fig. 3.9. We know that W^TW always has one zero eigenvalue. This is the point on the top left of Fig. 3.9. The second point on the top left indicates that there could be polynomial of degree 4 approximately vanishing on the data. For instance, the trucks are photographed from different angles, so that the equation could describe rotational symmetry (this would yield a polynomial equation of degree 2: $g(x) = 0$. If $g(x)$ is small compared to other eigenvalues, we expect that $g(x)^2$ is even smaller in comparison; i.e., taking degree 4 polynomials gives a clearer separation of eigenvalues).

[5] `http://www.cs.toronto.edu/ kriz/cifar.html`

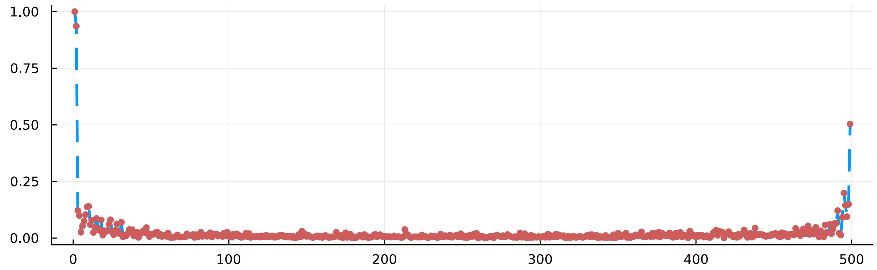


Figure 3.9: The distribution of the relative differences $d_: = (\lambda_i - \lambda_{i-1})/\lambda_i$ for the eigenvalues of W^TW in Example 3.44.

Let us now assume the following statistical setting for PCA. Recall that we have data points $x_1,\dots,x_n$ chosen independently from x, where $x \in \mathbb{R}^D$ is a random variable. Let $d \leq D$. We assume that there is a Gaussian **latent variable** $\zeta \sim N(\nu, B)$ with $\nu \in \mathbb{R}^D$ and $B \in \mathbb{R}^{D\times D}$ positive definite, such that

$$x = A\zeta + b + \varepsilon, \tag{3.4.5}$$

where $A \in \mathbb{R}^{D\times d}$ is a matrix of rank d, and we have a Gaussian noise $\varepsilon \sim N(0, \sigma^2 \mathbf{1}_D)$. That is,

$$(x \mid \zeta) \sim N(A\zeta + b, \sigma^2 \mathbf{1}_D).$$

In particular, we do not work with nonlinear feature maps here.

Proposition 3.45. *In the statistical setting above we have*

$$x \sim N(A\nu + b, ABA^T + \sigma^2 \mathbf{1}_D).$$

Proof. This follows from Lemma 1.29. □

We can use Proposition 3.45 for maximum likelihood, maximum a-posteriori estimation (see Definition 3.6), or for taking the Bayesian perspective where we assume priors for the parameters $(A,b) \in \mathbb{R}^{D\times d} \times \mathbb{R}^D$. Alternatively, we can also compute the distribution of ζ given a data point x. We can use this last approach to use data for updating the distribution of ζ, and then generating synthetic data points by sampling $(x \mid \zeta)$.

Theorem 3.46. *Suppose that we have a prior $\zeta \sim N(\nu, B)$. Then, the posterior distribution of ζ given x is*

$$(\zeta \mid x) \sim N(m, C)$$

with covariance matrix $C = (\sigma^{-2}A^TA + B^{-1})^{-1}$ and $m = C(\sigma^{-2}A^T(x-b) + B^{-1}\nu)$.

Proof. By Bayes' theorem for densities (Theorem 1.22) we have

$$\log P(\zeta \mid x) = \log P(x \mid \zeta) + \log P(\zeta) + c,$$

where c does not depend on ζ. By assumption, we have $(x \mid \zeta) \sim N(A\zeta + b, \sigma^2 \mathbf{1}_D)$ and $\zeta \sim N(v, B)$, so that

$$\log P(\zeta \mid x) = -\frac{1}{2\sigma^2}\|A\zeta - (x-b)\|^2 - \frac{1}{2}(\zeta - v)^T B^{-1}(\zeta - v) + c'$$

where c' does not depend on y This expression has the same form as in Eq. (3.2.7), and we have computed the conditional density in Eq. (3.2.7) in the proof of Theorem 3.17. Following the proof gives us $(\zeta \mid x) \sim N(m, C)$ with covariance matrix and expected value as stated. □

Example 3.47. We load $n = 500$ images of sneakers from the `FashionMNIST` dataset. A subset of these images is shown in Fig. 3.10.

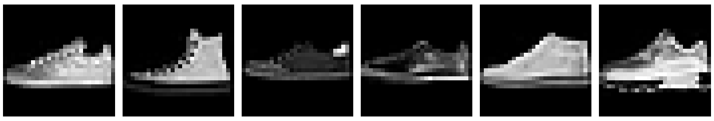


Figure 3.10: Images of sneakers from the `FashionMNIST` dataset.

As in Example 3.34 we have data points in $\mathbb{R}^{784}$ and we apply PCA to this data using a subspace of dimension $d = 31$. This results in a matrix $A \in \mathbb{R}^{784 \times 31}$ and a vector $b \in \mathbb{R}^{31}$. Then, we consider the statistical model from Eq. (3.4.5) with $\zeta \sim N(0, 2 \cdot \mathbf{1}_d)$ and $\varepsilon \sim N(0, 0.001 \cdot \mathbf{1}_d)$. We use this model to generate synthetic images of sneakers. A sample is shown in Fig. 3.11. In this setting, we call Eq. (3.4.5) a **generative model**.

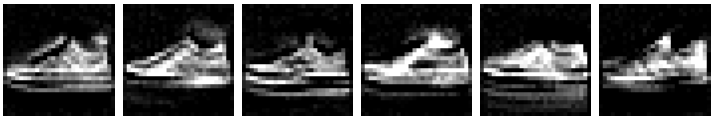


Figure 3.11: Synthetic images of sneakers sampled from the statistical model in Eq. (3.4.5).

Finally, we load one additional data point x and use Theorem 3.46 to compute the posterior distribution for ζ given x. Fig. 3.12 shows the image corresponding to x together with an image generated from this model.

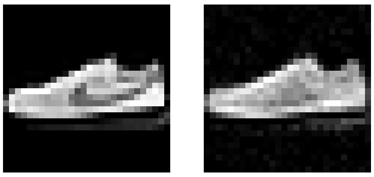


Figure 3.12: The left picture shows a datapoint from the `FashionMNIST` dataset. The picture on the right shows a synthetic image that was sampled using the posterior distribution in Theorem 3.46 given the image on the left.

Exercise 3.9. Let $z \in \mathbb{R}^M$ be a random variable with $\mu := \mathbb{E}z \in \mathbb{R}^M$. Show that the covariance matrix of z is given by $\Sigma = \mathbb{E}(z-\mu)(z-\mu)^T$. Use this to show that Σ is positive semi-definite.

Exercise 3.10. Take again the `MNIST` dataset from the `MLDatasets.jl` package and load the training data for pictures of ones and zeros. Use PCA to reduce the number of parameters representing these pictures. Then, load a point x from the test data set and compute the posterior distribution for $(\zeta \mid x)$ in Theorem 3.46 using x. Use the posterior distribution to generate synthetic data.

Exercise 3.11. Consider the function $f(\Sigma) = \Sigma^{-1}$, where $\Sigma \in \mathbb{R}^{n\times n}$ is invertible.

1. Prove that f is differentiable at Σ. **Hint:** Formulate f as a rational function in the entries of Σ.
2. Show that $\frac{\partial f}{\partial \Sigma_{ij}} = -\Sigma^{-1} e_i e_j^T \Sigma^{-1}$, where e_k is the k-th standard basis vector in $\mathbb{R}^n$. **Hint:** Differentiate both sides of $\Sigma\Sigma^{-1} = \mathbf{1}_n$.

Exercise 3.12. In the setting of Proposition 3.45 we have the likelihood equation

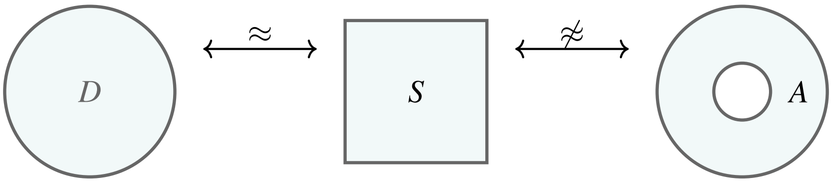
$$L(A,b) = \prod_{i=1}^{n} \Phi(x_i \mid b, AA^T + \sigma^2 \mathbf{1}_D).$$

Compute the MLE for (A,b) by maximizing L.

4 Topological Data Analysis

The goal of topological data analysis (TDA) is to learn the **topology** of data sets. For this one assumes that the data points are (possibly noisy) samples from an unknown geometric object in $\mathbb{R}^D$ called **model**.

The topology of the model gives an idea of the **shape of data** up to continuous deformations. A simple example is the unit disc $D = \{(x,y) : x^2 + y^2 \leq 1\} \subseteq \mathbb{R}^2$. The disc can be continuously deformed into the square $S = \{(x,y) : \max\{|x|,|y|\} \leq 1\}$ (denoted $D \approx S$). This cannot be done if we consider instead the unit circle with a hole cut out, for example the annulus $A = \{(x,y) : 1/2 \leq x^2 + y^2 \leq 1\}$:



The geometric reason why we cannot continuously deform A into S is that A has a **hole**, while S does not have a hole. Therefore, a central aspect in TDA is learning the number of holes of a model. We will give an explicit definition of what it means to have an ***n*-dimensional hole**, which we can do with the help of linear algebra, and **homology**.

Computing the number of n-dimensional holes can be used to classify and separate data. For example, see the *Blue Brain Project* [Hes20]. In other applications of TDA can be used to understand the relationship between data points. We will show that the number of 0-dimensional holes is equal to the number of connected components. For instance, in robotics the model could be the state space of a robot (the states into which the robot can move), and then the number of connected components of the is relevant to understand if the robot can change from any state to another state.

The fundamental idea in TDA is to assign a **simplicial complex** to the data, and then compute the topology through this simplicial complex.

4.1 Simplicial Complexes

In this section we will mostly follow the first chapter in [Bot20].

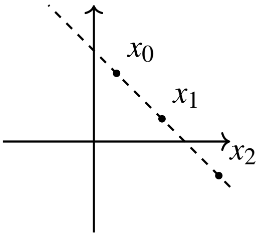
First we assume that our data points are given as points in $\mathbb{R}^D$. In this case, we can work with **geometric simplices.**

Definition 4.1. A collection of $n+1$ points $\{x_0,\dots,x_n\} \subset \mathbb{R}^D$ are called **affinely independent** if for all $t_0,t_1,\dots,t_n \in \mathbb{R}$ with $\sum_{i=0}^n t_i = 0$,

$$\sum_{i=0}^n t_i x_i = 0 \quad \implies \quad t_0 = \cdots = t_n = 0.$$

That is, $\{x_1 - x_0,\dots,x_n - x_0\}$ are linearly independent.

Example 4.2. In the graph below, $\{x_0,x_1\}$ are affinely independent because they define a unique line. The points $\{x_0,x_1x_2\}$ are collinear and not affinely independent.

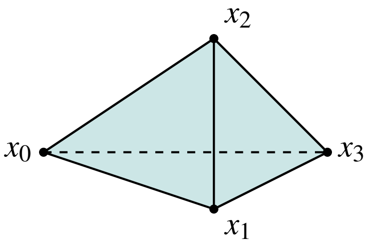


Definition 4.3. Let $\mathscr{P} := \{x_0,\dots,x_n\} \subseteq \mathbb{R}^D$ be affinely independent. The ***n*-simplex** that is spanned by the x_i is

$$\Delta = \Delta(\mathscr{P}) := \left\{ x \in \mathbb{R}^D \;\middle|\; x = \sum_{i=0}^n t_i x_i,\, \sum_{i=0}^n t_i = 1,\, t_i \geq 0 \right\}.$$

The **dimension** of the simplex is $\dim \Delta(\mathscr{P}) := n$.

Example 4.4. Below is a **tetrahedron** Δ, which is the 3-simplex through $\{x_0,x_1,x_2,x_3\}$. The sides of the tetrahedron are **triangles**, for example the right side is the 2-simplex defined by $\{x_1,x_2,x_3\}$. The sides of a simplex are themselves always simplices.



Definition 4.5. Let Δ be the n-simplex spanned by $n+1$ affinely independent points $\{x_0, x_1, \ldots x_n\} \subseteq \mathbb{R}^D$. Let $I \subsetneq \{1, \ldots, n\}$. If $|I| = p$, then the simplex σ_I spanned by the points $\{x_i \mid i \in I\}$ is called p-**face** of Δ. When $|I| = n-1$, we say σ_I is a **facet** of Δ. The **boundary** of Δ is the set of all its facets, denoted $\Gamma(\Delta)$.

Simplices are the building blocks of simplicial complexes. The next definition introduces simplicial complexes as unions of simplices with a particular structure.

Definition 4.6. A **simplicial complex** K is a finite collection of simplices with the property that

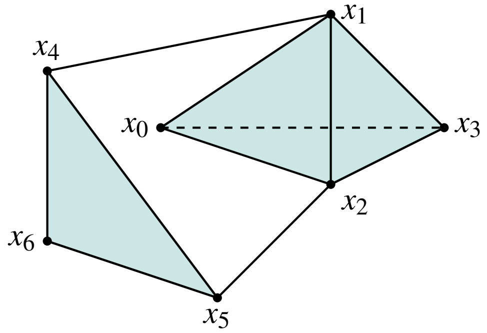
1. for all $\Delta \in K$ and all faces $\sigma \in \Delta$, we also have $\sigma \in K$;
2. for $\Delta, \sigma \in K$ with $\Delta \cap \sigma \neq \emptyset$, then $\Delta \cap \sigma$ is a face of both Δ and σ.

The **dimension** of K is $\dim K := \max_{\Delta \in K} \dim \Delta$. The p-**skeleton** of K is

$$K^{(p)} := \{\Delta \in K \mid \dim(\Delta) \leq p\}. \tag{4.1.1}$$

The 0-skeleton is also called the **vertices** of K

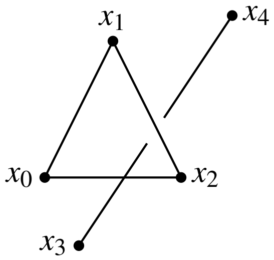
Example 4.7. We first give an example of a simplicial complex of dimension 3, given as the union of a tetrahedron, a triangle and two edges. The color indicates that the inside of a simplex also belongs to the simplicial complex.



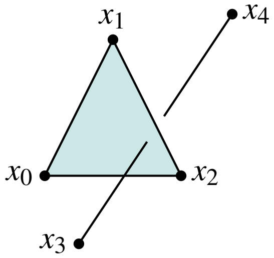
Next, we have a simplicial complex of dimension 1:

$$K = \{\{x_0\}, \{x_1\}, \{x_2\}, \{x_3\}, \{x_4\}, \{x_1, x_2\}, \{x_0, x_1\}, \{x_0, x_2\}, \{x_3, x_4\}\}.$$

It consists of 4 edges and 5 points. In particular K is not connected.



If we include also the convex hull of x_0,x_1,x_2, then we no longer have a simplicial complex, because the intersection of the line with the triangle is not a face.



We can uniquely identify every simplex Δ in a simplicial complex K through the vertices of Δ, namely $\alpha = \#\Delta^{(0)}$, where $\Delta^{(0)}$ is as in Eq. (4.1.1). The 0-skeleton α has the property that all subsets $\beta \subset \alpha$ are also identified as a simplex in K. This is the motivation of an abstract simplicial complex. Notice that the definition of an abstract simplicial complex generalizes the definition of a graph (Definition 2.1).

Definition 4.8. An abstract simplicial complex A on vertices $\{0,\dots,n\}$ is a collection of subsets $\alpha \subseteq \{0,\dots,n\}$ so that for all $\alpha \in A$ and for all $\beta \subseteq \alpha$, $\beta \in A$.

The dimension of $\alpha \in A$ is $\dim\alpha := |\alpha| - 1$, and $\dim A = \max_{\alpha\in A} \dim\alpha$.

Every geometric complex K defines an abstract complex A. We call K the **geometric realization** of A. We show next that the converse is also true.

Lemma 4.9. *Let A be an abstract simplicial complex of dimension* $\dim A = d$. *Then A has a geometric realization in* $\mathbb{R}^{2d+1}$.

Proof. Suppose that A is an abstract simplicial complex on $\{0,\dots,n\}$. By assumption, we have $|\alpha| \leq d+1$ for all $\alpha \in A$. Therefore, $|\alpha \cup \beta| \leq 2(d+1)$ for all $\alpha, \beta \in A$. We can thus find a function $h : \{0,\dots,n\} \to \mathbb{R}^{2d+1}$ so that $h(\alpha\cup\beta)$ is affinely independent for all $\alpha, \beta \in A$. For $\alpha \in A$ let Δ_α be the simplex spanned by $h(\alpha)$. We show that for all $\alpha, \beta \in A$ and $x \in \Delta_\alpha \cap \Delta_\beta$ we have $x \in \Delta_{\alpha\cap\beta}$.

Since $h(\alpha) = h(\alpha \cup \emptyset)$ is affinely independent, the 0-skeleton of Δ_α is exactly $h(\alpha)$. Similarly, the 0-skeleton of Δ_β is exactly $h(\beta)$. We can write

$$x = \sum_{i\in\alpha} t_i h(i) + \sum_{j\in\beta} s_j h(j)$$

with unique values $t_i \geq 0$ for $i \in \alpha$, $s_i \geq 0$ for $i \in \beta$, because $h(\alpha \cup \beta)$ is affinely independent. However $x \in \Delta_\alpha \cap \Delta_\beta$ must have $t_i = 0$ for $i \notin \beta$ and $s_j = 0$ for $j \notin \alpha$, so that $x = \sum_{i\in\alpha\cap\beta} t_i h(i) + \sum_{j\in\alpha\cap\beta} s_j h(j)$. It follows that $x \in \Delta_{\alpha\cap\beta}$. Therefore,

$$K = \{\Delta_\alpha \mid \alpha \in A\}$$

is a simplicial complex which gives a geometric realization of A in $\mathbb{R}^{2d+1}$. □

4.2 Assigning a Simplicial Complex to Data

Oftentimes Data has a natural interpretation as a simplicial complex. For example when considering a network of $n+1$ objects $\{0, \dots, n\}$. A simplicial complex A can be used to model the relationship between objects: $\alpha \in A$ is the case that all objects in α have a joint relationship.

In other situations, we must first associate to data points a simplicial complex. Let us assume again that our data is given as points in $\mathbb{R}^D$. Two main approaches for assigning a complex to our data is computing the **Čech-Complex** or the **Vietoris–Rips Complex**.

Definition 4.10. Let $\mathscr{P} := \{x_0, \dots, x_n\} \subset \mathbb{R}^D$ be a collection of data points. Let $r > 0$.

1. The **Čech Complex** of level r is

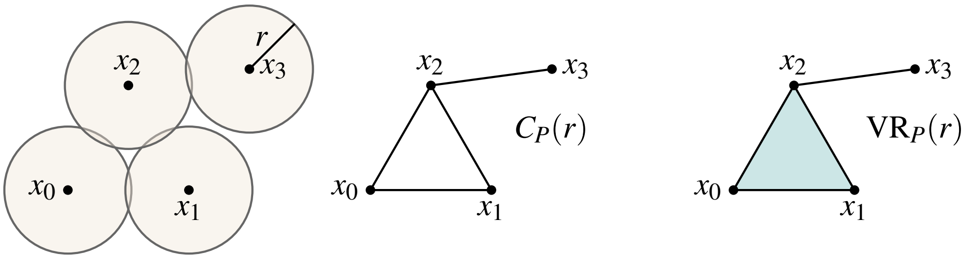
$$C_r(\mathscr{P}) := \Big\{ \alpha \subseteq \mathscr{P} \;\Big|\; \bigcap_{x\in\alpha} B_r(x) \neq \emptyset \Big\}$$

   where $B_r(x) = \Big\{ p \in \mathbb{R}^D \;\Big|\; \|x - p\| \leq r \Big\}$.

2. The **Vietoris–Rips Complex** of level r is

$$\mathrm{VR}_r(\mathscr{P}) := \Big\{ \alpha \subseteq \mathscr{P} \;\Big|\; \max_{x,x'\in\alpha} \|x - x'\| \leq 2r \Big\}.$$

Example 4.11. Consider $\mathscr{P} = \{x_0, x_1, x_2, x_3\} \subset \mathbb{R}^2$.



The **Čech Complex** and **Vietoris–Rips Complex** at levvel r can best be understood by drawing discs of radius r around each data point. For subset of data points $\{x_i \mid i \in I\}$, $I \subset \{0,1,2,3\}$, we add the corresponding simplex to $C_r(\mathscr{P})$ if the discs around $x_i, i \in I$, intersect in a common point. By constrast, we add the simplex to $\mathrm{VR}_r(\mathscr{P})$ if the discs intersect pairwise.

In the next section we will then develop strategies in the spirit of homology theory to define and count holes in simplicial complexes. This gives the following general algorithm for computing the topology of (discrete) datasets.

Algorithm 4.1: Computing the topology of discrete datasets in $\mathbb{R}^D$.

```
Input: Data $x_1, \dots, x_n \in \mathbb{R}^D$ and a positive real number $r$.
Compute the Čech Complex or Vietoris–Rips Complex at level $r$ for the data;
Compute the number of holes of the complex;
```

The approach in **persistent homology** is to let r range between a minimal and a maximal value instead of choosing a fixed number. Then, the number of holes that persist for a large portion of rs are considered to be signals coming from the data. The rest is considered noise. Here is a first simple version of the algorithm.

Algorithm 4.2: Persistent homology (simple version)

```
Input: Data $x_1, \dots, x_n \in \mathbb{R}^D$, two positive real numbers $r_{\min} < r_{\max}$ and a
        number $n \geq 1$.
For all $r = r_{\min} + \frac{k}{n}(r_{\min} - r_{\min})$, $k = 0, 1, \dots, n$, apply Algorithm 4.1 with input
    data $x_1, \dots, x_n$ and $r$;
```

The advantage of the Vietoris–Rips over the Čech complex is that it is easier to compute. For $\alpha \subseteq P$, then for Vietoris–Rips we must compute $\binom{|\alpha|}{2}$ distances, whereas the

Čech complex requires checking whether a system of polynomial inequalities has a solution. On the other hand, the Vietoris–Rips complex automatically fills in simplicies. For example $\{p,q\},\{p,u\},\{q,u\} \in \mathrm{VR}_P(r)$ implies $\{p,q,u\} \in \mathrm{VR}_P(r)$. Therefore the Vietoris–Rips complex cannot represent all simplicial complexes.

The trade-off between the Čech and the Vietoris–Rips complex is quantified in Proposition 4.12.

Proposition 4.12. *For $r > 0$, and $\mathscr{P} = \{x_0,\dots,x_n\} \subseteq \mathbb{R}^D$,*

$$C_P(r) \subseteq \mathrm{VR}_P(r) \subseteq C_P(2r).$$

Proof. Let $\alpha \subseteq \mathscr{P}$. If $\bigcap_{x\in\alpha} B_r(x) \neq \emptyset$, then for all $p,q \in \alpha$, $B_r(p) \cap B_r(q) \neq \emptyset$ and $\|p-q\| \leq 2r$. This shows the first inclusion.

For the second, we select $u,v \in \alpha$ and set $p := \frac{u+v}{2}$. Then for all $x \in \alpha$,

$$\|x-p\| \leq \|x-u\| + \|p-u\| = \|x-u\| + \frac{1}{2}\|u-v\| \leq 2r.$$

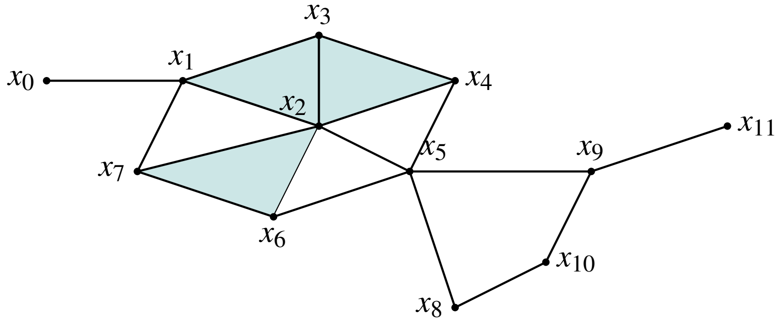
This implies $p \in \bigcap_{x\in\alpha} B_r(x)$. □

4.3 Holes of a Simplicial Complex in the Plane

We work towards defining holes in simplicial complexes. Let us first study the case of a geometric simplex $K \subset$ in the plane $\mathbb{R}^2$. Once we have understood this basic case, we will generalize the established concepts to simplices realized in higher dimension.

Let $\mathscr{P} = \{x_0,\dots,x_n\}$ be the vertices of K. In this case, a first meaningful definition of a hole is a triangle $\Delta = \{x_i,x_j,x_k\}$, such that the edges of Δ are present in K, but Δ itself is not. The next example, however, shows that this is not enough.

Example 4.13. In the following example there are 4 holes, but only three of them are triangles. The hole surrounded by x_5,x_8,x_9,x_{10} is a quadrilateral.



The example motivates the following definition for holes.

Definition 4.14. Let K be a simplicial complex on $\mathscr{P} = \{x_0, \dots, x_n\}$ in $\mathbb{R}^2$. Let σ be a subset of the vertices $\mathscr{P}$. We say σ is a **hole** of K if $\sigma \notin K$, but $K' = \{\alpha \in K : \alpha \subsetneq \sigma\}$ builds a circle.

With this definition the number of holes of $K \subset \mathbb{R}^2$ is invariant under continuous deformations of K. We have the following result for the number of holes:

Theorem 4.15. *Let K be a simplicial complex realized in $\mathbb{R}^2$ so that the 1-skeleton $K^{(1)}$ builds a connected graph. Then*

$$1 - \#\{\textit{holes of } K\} = \#\{\textit{vertices of } K\} - \#\{\textit{edges of } K\} + \#\{\textit{triangles of } K\}.$$

Proof. We proceed by induction on the number of edges. In the base case, there are 0 edges, so the complex consists only of a single vertex, and therefore has zero holes.

Now for the induction step, suppose K is a simplicial complex for which the assumption holds. We add an edge to K, which has two possibilities.

1. We add a vertex. Then in this case we do not create a circle,
   - $\#\{$holes of $K\}$ and $\#\{$triangles of $K\}$ stay the same;
   - $\#\{$vertices of $K\}$ and $\#\{$edges of $K\}$ each increase by 1.
2. We do not add any vertices. In this case, we create a circle.

   a) If in the case the circle has 3 edges, we add the triangle to get
      - $\#\{$holes of $K\}$ and $\#\{$vertices of $K\}$ stay the same;
      - $\#\{$edges of $K\}$ and $\#\{$triangles of $K\}$ each increase by 1.

   b) If we do not add a triangle, then
      - $\#\{$vertices of $K\}$ and $\#\{$triangles of $K\}$ stay the same;
      - $\#\{$holes of $K\}$ and $\#\{$edges of $K\}$ increase by 1.

In all cases, the equation above is preserved. □

Bibliography

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

[AZ18] Martin Aigner and Günter M. Ziegler. *Proofs from THE BOOK*. Springer Publishing Company, Incorporated, 6th edition, 2018.

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

[Bot20] Magnus Bakke Botnan. Topological data analysis, 2020. Lecture Notes.

[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. Available for download, 1997.

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

[DFO20] Marc Peter Deisenroth, A. Aldo Faisal, and Cheng Soon Ong. *Mathematics for Machine Learning*. Cambridge University Press, 2020.

[FV07] Alan Frieze and Eric Vigoda. A survey on the use of Markov chains to randomly sample colourings. *Combinatorica*, 01 2007.

[GK12] Venkatesan Guruswami and Ravi Kannan. Computer science theory for the information age, 2012. Lecture Notes.

[Hes20] Kathryn Hess. Topological adventures in neuroscience. In Gereon Quick Markus Szymik Marius Thaule Nils A. Baas, Gunnar E. Carlsson, editor, *Topological Data Analysis*. Springer International Publishing, 2020.

[HJ92] Richard A. Horn and Charles R. Johnson. *Matrix analysis*, volume 349. Cambridge University Press, Cambridge, 1992.

[Mur13] Kevin P. Murphy. *Machine learning : a probabilistic perspective*. MIT Press. Available for download., Cambridge, Mass. [u.a.], 2013.

[PBMW98] Larry Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the Web. In *Proceedings of the 7th International World Wide Web Conference*, pages 161–172, Brisbane, Australia, 1998.

[RGS09] José C. Rosales and Pedro A. García-Sánchez. *Numerical Semigroups*. Developments in Mathematics. Springer New York, 2009.

[SS11] Thomas Sauerwald and He Sun. Spectral graph theory, 2011. Lecture Notes.

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