Mathematical Methods in Data Science

Paul Breiding and Samantha Fairchild

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

These lecture notes were written in the Summer Semester 2022, when the 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 that understanding the **geometry of data** is a key principle.

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

Contents

1 The Basics

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

1.1 Linear Algebra

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

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

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

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

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

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

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

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

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

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

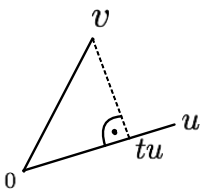


Figure 1.1: The meaning of the inner product between u and v is illustrated in this picture: let $t \in \mathbb{R}$ such that $v = tu + u'$, where u' is orthogonal to u. Then, $\langle u, v\rangle = u^T(tu + u') = tu^T u = t\langle u, u\rangle$. In particular, if $\langle u, u\rangle = \langle v, v\rangle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

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

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

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

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

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

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

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

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

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

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

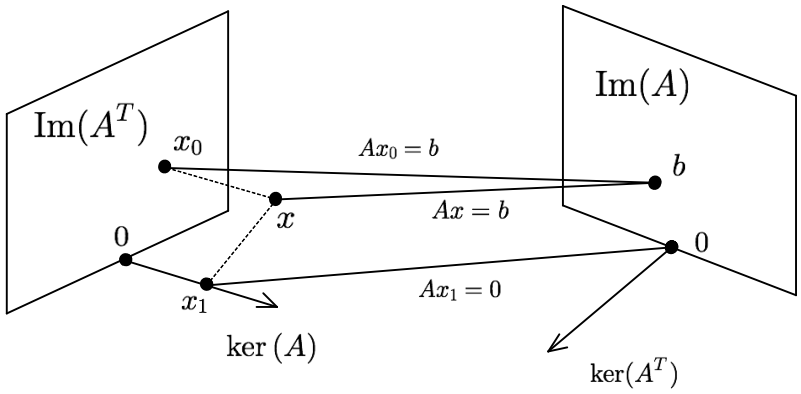


Figure 1.2: The situation when $b \in \mathrm{Im}(A)$: in this case, $Ax = b$ has a unique solution $x \in \mathrm{Im}(A^T)$ and the solution space for $Ax = b$ is $x + \ker(A)$.

and by the Rank-Nullity theorem

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

Therefore

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

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

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

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

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

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

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

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

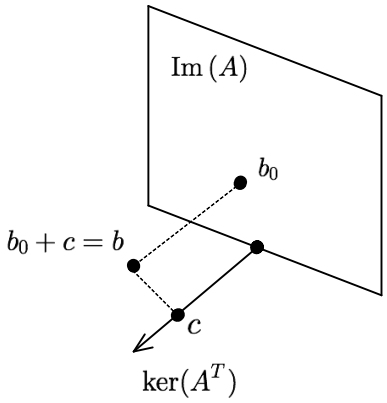


Figure 1.3: Visualization of the proof of Lemma 1.4: b_0 minimizes the distance from b to $\mathrm{Im}(A)$.

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

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

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

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

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

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

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

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

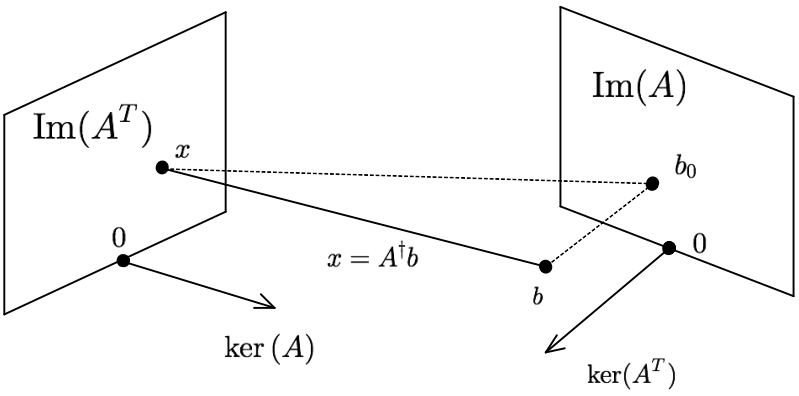


Figure 1.4: The pseudoinverse $A^\dagger \in \mathbb{R}^{m\times n}$ of $A \in \mathbb{R}^{m\times n}$ first orthogonally projects $b \in \mathbb{R}^n$ to $b_0 \in \mathrm{Im}(A)$ and then maps b_0 to the unique point $x \in \mathrm{Im}(A^T)$ with $Ax = b_0$.

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

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

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

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

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

$$A^\dagger b = x$$

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Then, by construction we have

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Exercise 1.2. Prove part 2 of Proposition 1.7.

Exercise 1.3. Prove Lemma 1.9.

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

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

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

1.2 Probability Theory

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

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

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

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

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

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

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

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

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

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

The restriction that $\mathscr{A}$ is a σ-algebra is crucial: without this assumption 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 $\Omega = \{1, \ldots, 6\}$, $A = \{2\}$ and $B = \{2, 4, 6\}$. Suppose that $P(\{k\}) = \frac{1}{6}$ for $k = 1, \ldots, 6$. Then:

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

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

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

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

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

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

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

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

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

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

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

Let now $X \in \mathbb{R}$ be a real random variable. If X is discrete and $X(\Omega) = \{x_1, x_2, \ldots\}$ 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 $X \in \{x_1, x_2, \ldots\}$ 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:

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

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

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

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

2. **Binomial distribution:** $X \in \{0,\ldots,n\}$ and $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$.

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

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

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

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

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

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

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

Exercise 1.5. Prove Lemma 1.24. **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. Prove Lemma 1.27.

Exercise 1.7. 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 1.8. 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 $Z \sim \mathrm{Pois}(\lambda)$, 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 recom mend [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 $\{u, v\}$ is **adjacent** to e. 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 edges adjacent to v denoted

$$\deg(v) := \{\{i, j\} \in E \mid i = v \text{ or } j = v\}.$$

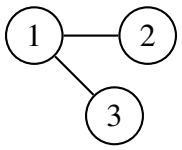
In the following, we will only consider graphs $G = (V,E)$ that have no isolated vertices; i.e., we assume

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

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 **undirected** graphs.

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



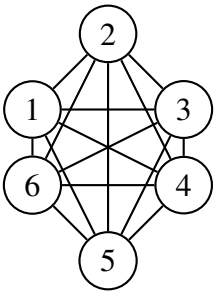
The adjacency matrix of this graph is

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

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. Let $v, w \in V$. A **path** from v to w is a collection of vertices

$$P = \{v_0, v_1, \ldots, v_D\},$$

such that $v_0 = v, v_D = w$, and $\{v_{i-1}, v_i\} \in E$ for all $1 \leq i \leq D$. The edges of P are

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

The **length** of P is $D = \#E(P)$. We say that G is **connected**, if for every $v, w \in V$ there is path from v to w in G.

Lemma 2.7. *Let A be the adjacency matrix of a graph $G = (V,E)$, and let $v, w \in V$. Then the number of paths 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.

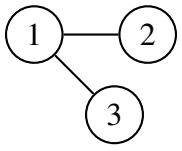
Definition 2.8. Let $G = (V,E)$ be a graph. The **Laplace Matrix** of G is

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

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 2.9. 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 2.10. 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 2.11. Another common definition of the Laplacian of a graph is $\mathscr{L} := T - A$, where T is as in Definition 2.10 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 2.12. *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. In the identification from Eq. (2.1.1) this means $e_u(u) = 1$ and $e_u(v) = 0$ for $u \neq v$. We compute for $u,v \in V$, and using the fact that T is symmetric so $T = T^T$

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

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

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

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

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

for $v \in V$.

Lemma 2.13. *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 2.12 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 2.14. The eigenvalues of L,

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

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

$$\lambda_G = \lambda_1.$$

Example 2.15. The Laplace matrix from Example 2.9 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 2.16. 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 2.17. *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 2.13, 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:

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

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 2.17 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 2.18. *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 2.17,

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

Furthermore, let us consider the vector with $f(u) = \sqrt{\deg(u)}$. Then, again by Theorem 2.17, 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

$$f = T^{1/2} e \in \ker L, \tag{2.1.4}$$

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

Next, we give the spectra of some example graphs.

Proposition 2.19. *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, but not a cycle, then $\lambda_k = 1 - \cos\frac{\pi k}{n-1}$.*
4. *If G is a cycle, then $\lambda_k = 1 - \cos\frac{2\pi k}{n-1}$.*

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 2.7.

Exercise 2.3. For the graph in Example 2.5, compute the number of paths of length 3 from vertex 1 to the vertex 2 .

Exercise 2.4. Prove Proposition 2.19.

2.2 The Spectrum of a Graph

In this lecture, 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 2.18 that these eigenvalues are nonnegative. The main goal of this lecture is to proof the following theorem.

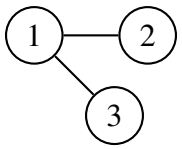
Theorem 2.20. *Let $G = (V,E)$ be a graph with $n = |V| \geq 2$, and let*

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

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



From Example 2.15 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 2.20 1.). We have one connected component corresponding to $\lambda_0 = 0 < \lambda_1$ (see Theorem 2.20 4.). Finally, $\lambda_2 = 2$ as G is bipartite (see Theorem 2.20 5.).

Let us now prove Theorem 2.20.

Proof of Theorem 2.20. 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 2.8), 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 2.19 that, if G is complete, $\lambda_G = \frac{n}{n-1}$. We show that otherwise $\lambda_G \leq 1$. If G is not complete, there exist $u, v \in V$ with $\{u, v\} \in E$. Let us define $f \in \mathscr{F}(V)$ with

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

Recall from Eq. (2.1.4) that $T^{1/2}e \in \ker L$, so that we can express λ_G using the Rayleigh quotient:

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

The function f satisfies

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

By Theorem 2.17, we have

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

This shows $\lambda_G \leq 1$.

For item 4. we first make the following observation: Let $f \in \ker L$. Then, by Theorem 2.17, we have

$$0 = \frac{\langle f, Lf\rangle}{\langle f, f\rangle} = \frac{1}{\sum_{u\in V} f(u)^2} \sum_{\{u,v\}\in E} \left(\frac{f(u)}{\sqrt{\deg(u)}} - \frac{f(v)}{\sqrt{\deg(v)}} \right).$$

Let us write $g := T^{-1/2} f$. Then, $g(u) = g(v)$ for all $\{u,v\} \in E$. 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 this path we have $g(u) = g(v)$, we also have $g(i) = g(j)$. Consequently, g is contant on this path.

Assume now that G is connected, then for every $i, j \in G$ we can find a path from i to j, so 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 $f \neq T^{1/2}e$ in the kernel of L. But then, there must exist vertices $i, j \in G$ such that G contains no path from i to j, hence G is not connected.

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

The proof of item 5. is based on the following observation that for all $g \in \mathscr{F}(V)$ and $u, v \in V$ we have $0 \leq (g(u) + g(v))^2 = g(u)^2 + 2g(u)g(v) + g(v)^2$, hence

$$(g(u) - g(v))^2 = g(u)^2 - 2g(u)g(v) + g(v)^2 \leq 2(g(x)^2 + g(y)^2).$$

Now, using again Theorem 2.17 and setting $g = T^{1/2} f$ we get, using the Rayleigh quotient

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

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

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

For the previous argument we have only used the inquality $0 \leq (g(u) + g(v))^2$. Thefore, $\lambda_{n-1} = 2$, 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 $H = (V', E')$, let V_1 and V_2 be the bipartite components of V'. We can 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 $\lambda_{n-1} = 2$. If on the other hand there exits a nonzero function $g \in \mathscr{F}(V)$ with $g(u) = -g(v)$ for all $\{u,v\} \in E$, then let $H = (V',E')$ be a connected component of G on which g does not vanish. We define the subsets $W_1 := \{w \in W \mid g(w) > 0\}$ and $W_2 := \{w \in W \mid g(w) < 0\}$. Then, $V' = W_1 \cup W_2$ and for all edges $\{u,v\} \in E'$ we must have $u \in W_1$ and $v \in W_2$. 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 such that $V_i = \{n_{i-1}+1, \ldots, n_i\}$ 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 union of the eigenvalues of the L_i. □

Theorem 2.20 shows how we can obtain information about the structure of a graph by computing its spectrum. Howecer, 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 2.22. Let $G = (V,E)$ be a graph. The **volume** of G is

$$\mathrm{vol}(G) := \sum_{v \in V} \deg(v).$$

Proposition 2.23. *Let $G = (V,E)$ be a graph, $n = |V|$, with spectrum λ_G, λ_{n-1} as in Theorem 2.20. Let $G_1 = (V_1,E_1)$ and $G_2 = (V_2,E_2)$ be two subgraphs with $V = V_1 \cup V_2$ and $V_1 \cap V_2 = \emptyset$. Denote*

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

Then,

$$\lambda_G \leq \varepsilon \frac{\mathrm{vol}(G)}{(\mathrm{vol}(G_1)+\varepsilon) \cdot (\mathrm{vol}(G_2)+\varepsilon)} \leq \lambda_{n-1}.$$

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.

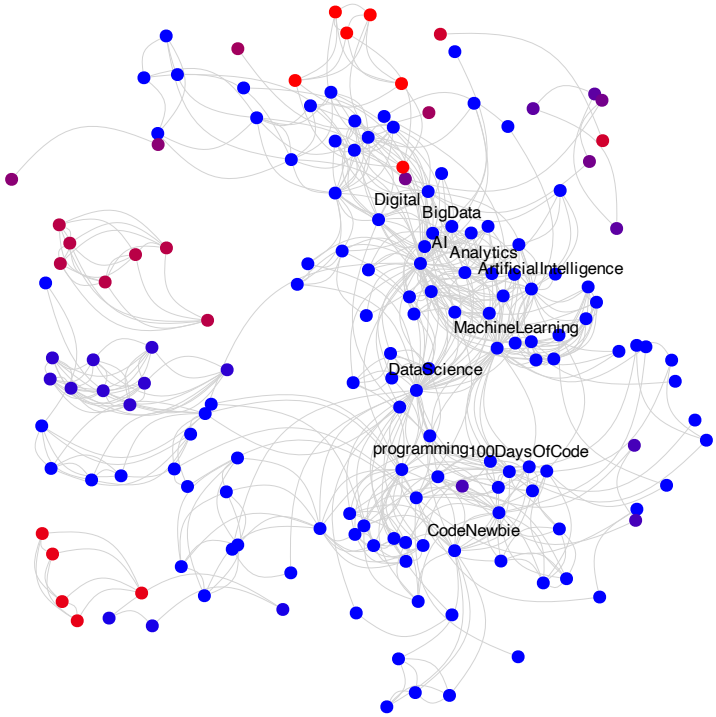


Figure 2.1: The network from Example 2.24. Different colors encode different connected components. The 10 vertices with highest degree are labelled. The labels are the hashtags the vertices correspond to.

Example 2.24. We illustrate Proposition 2.23 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 and only if the two corresponding hashtags appear together in at least one tweet. This gives the graph on $n = 166$ vertices that can be seen in Fig. 2.1. We have labelled the 10 vertices with highest degrees in this graph with their corresponding hashtag. There are multiple connected components, which is shown by the different colors of the vertices – one color per components.

We consider the blue component in Fig. 2.1 and call the underlying graph G. Fig. 2.2 shows the structure of G. The spectrum of G is shown in Fig. 2.3. We have $\lambda_G \approx 0.05$, so that we can't use Proposition 2.23 to conclude that G is well connected. The opposite

is true, as one can see in Fig. 2.2 that there are few edges between the red vertices and the blue vertices. In fact, the red vertices in Fig. 2.2 correspond to vertices, where the eigenvector of λ_G is positive, and orange vertices in Fig. 2.2 correspond to vertices, where the eigenvector of λ_G is negative. This partition of the vertices is similar as in the proof of Proposition 2.23.

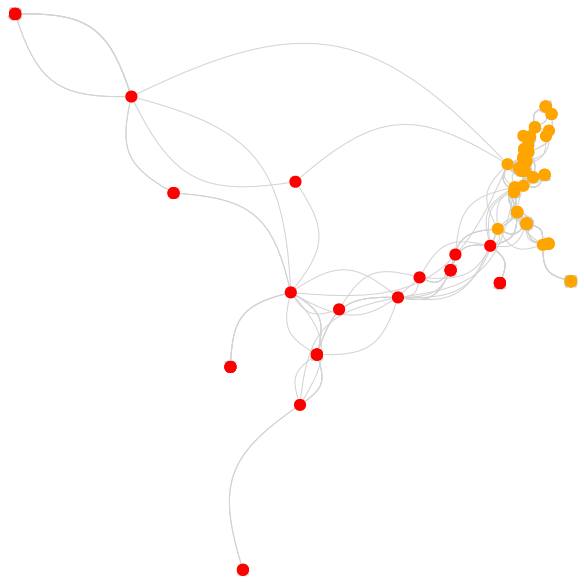


Figure 2.2: This graph shows the blue component in Fig. 2.1, called G. It is partitioned into two sets of vertices corresponding to whether the eigenfunction of λ_G is positive or negative. Compared to Fig. 2.1 the vertices are rearranged, so that one can see the two clusters better.

Let us also see how Proposition 2.23 is related to Theorem 2.17. The graph G is bipartite with components G_1 and G_2, if and only if $\mathrm{vol}(G_1) = \mathrm{vol}(G_2) = 0$. In this case, we get $2\varepsilon = \mathrm{vol}(G)$, so that the bound in Proposition 2.23 becomes $\lambda_G \leq 2 \leq \lambda_{n-1}$, similar to Theorem 2.20 5. Furthermore, the components G_1 and G_2 are disconnected, if and only if $\varepsilon = 0$, in which case we have $\lambda_G = 0$.

Proof of Proposition 2.23. Let us denote $m_i := \mathrm{vol}(G_i) + \varepsilon$. Observe that for $i \neq j$:

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

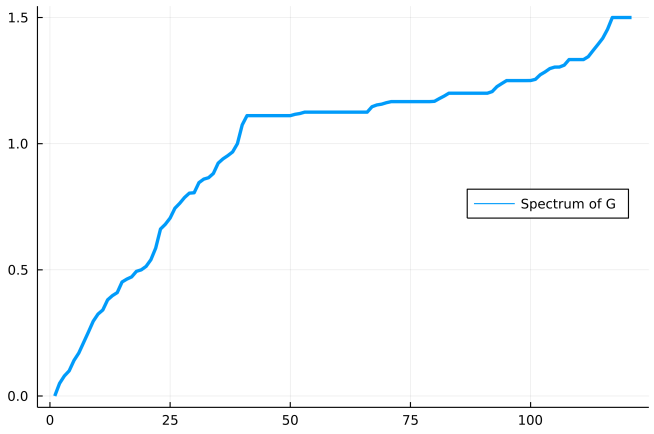


Figure 2.3: Spectrum of the graph from Fig. 2.2.

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

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,

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

and, by Theorem 2.17,

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

On the one hand, we have $\lambda_{n-1} = \max_{g\in\mathscr{F}(V)} \frac{\langle g, Lf\rangle}{\langle g, g\rangle}$, 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

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

□

The last result in this lecture relates λ_G to the lengths of paths between vertices.

Proposition 2.25. *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 path between vertices $u,v\in V$. Then,*

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

Proof. Let $f\in\mathscr{F}(V)$ be an eigenvector of λ_G with $\langle f, T^{1/2}e\rangle$. Such an eigenvector exists by Eq. (2.1.4) (and since $L(G)$ is symmetric). Write $g := T^{1/2}f$. Then,

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

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

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

We show that $\lambda_G \geq (D\cdot\mathrm{vol}(G))^{-1}$. Using Theorem 2.17 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

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

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

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

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 $g(u_0)g(v_0)<0$. □

Bibliography

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

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

[Chu97] Fan R. K. Chung. *Spectral graph theory*, volume 92 of *CBMS Regional Conference Series in Mathematics*. Published for the Conference Board of the Mathematical Sciences, Washington, DC; by the American Mathematical Society, Providence, RI. Available at `https://mathweb.ucsd.edu/ fan/research/revised.html`, 1997.

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

[GK12] Venkatesan Guruswami and Ravi Kannan. Computer science theory for the information age, 2012. Available at `https://www.cs.cmu.edu/ venkatg/teaching/CStheory-infoage/`.

[SS11] Thomas Sauerwald and He Sun. Spectral graph theory, 2011. Available at `https://resources.mpi-inf.mpg.de/departments/d1/teaching/ws11/SGT/`.

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