Optimization and Computational Linear Algebra for Data Science
Lecture 6: Eigenvalues, eigenvectors and Markov chains

Léo Miolane · leo.miolane@gmail.com

June 19, 2019

Warning: *This material is not meant to be lecture notes. It only gathers the main concepts and results from the lecture, without any additional explanation, motivation, examples, figures...*

1 Eigenvalues and eigenvectors

Definition 1.1

Let $A \in \mathbb{R}^{n\times n}$. A **non-zero** *vector $v \in \mathbb{R}^n$ is said to be an* eigenvector *of A is there exists $\lambda \in \mathbb{R}$ such that*

$$Av = \lambda v.$$

The scalar λ is called the eigenvalue (of A) associated to v. The set

$$E_\lambda(A) = \{x \in \mathbb{R}^n \mid Ax = \lambda x\} = \mathrm{Ker}(A - \lambda \mathrm{Id})$$

is called the eigenspace of A associated to λ.

Remark 1.1. *Notice that $E_\lambda(A)$ is a subspace of $\mathbb{R}^n$: any (non-zero) linear combination of eigenvectors associated with the eigenvalue λ is also an eigenvector of A associated with λ.*

Proposition 1.1

Let $A \in \mathbb{R}^{n\times n}$. Suppose that A has an eigenvalue $\lambda \in \mathbb{R}$ and let $x \in \mathbb{R}^n$ be an eigenvector associated to λ. The following holds:

- *For all $\alpha \in \mathbb{R}$, $\alpha\lambda$ is an eigenvalue of the matrix αA and x is an associated eigenvector.*
- *For all $\alpha \in \mathbb{R}$, $\lambda + \alpha$ is an eigenvalue of the matrix $A + \alpha \mathrm{Id}$ and x is an associated eigenvector.*
- *For all $k \in \mathbb{N}$, λ^k is an eigenvalue of the matrix A^k and x is an associated eigenvector.*
- *If A is invertible then $1/\lambda$ is an eigenvalue of the matrix inverse A^{-1} and x is an associated eigenvector.*

Definition 1.2

The set of all eigenvalues of A is called the spectrum *of A and denoted by $\mathrm{Sp}(A)$.*

Proposition 1.2

A $n \times n$ matrix A admits at most n eigenvalues: $\#\mathrm{Sp}(A) \leq n$.

2 Diagonalizable matrices

Definition 2.1

A matrix $A \in \mathbb{R}^{n\times n}$ is said to be diagonalizable *if there exists a basis $(v_1, \dots, v_n)$ of $\mathbb{R}^n$ consisting of eigenvectors of A, i.e. such that there exists $\lambda_1, \dots, \lambda_n \in \mathbb{R}$, $Av_i = \lambda_i v_i$.*

Proposition 2.1

A matrix $A \in \mathbb{R}^{n\times n}$ is diagonalizable if and only if there exists an invertible $n \times n$ matrix P and a diagonal matrix $D = (\lambda_1, \dots, \lambda_n)$ such that

$$A = PDP^{-1}.$$

In this case, the i^{th} column of P is an eigenvector of A associated with the eigenvalue λ_i.

3 Application to Markov chains

3.1 First definitions and properties

A finite Markov chain is a process which moves among the elements of a finite set E in the following manner: when at $x \in E$, the next position is chosen according to a fixed probability distribution $P(x, \cdot)$. More formally:

Definition 3.1

A sequence of random variables $(X_0, X_1, \dots)$ is a Markov chain with state space E and transition matrix P if for all $t \geq 0$,

$$\mathbb{P}(X_{t+1} = y \,|\, X_0 = x_0, \dots, X_t = x_t) = P(x_t, y)$$

for all $x_0, \dots, x_t$ such that $\mathbb{P}(X_0 = x_0, \dots, X_t = x_t) > 0$.

The transition matrix P verifies therefore, for all $x \in E$,

$$\sum_{y\in E} P(x, y) = 1. \tag{1}$$

In order to simplify the notations, we will assume that $E = \{1, 2, \dots, n\}$ and write for all $i, j \in E$, $P_{i,j} = P(j, i)$. **Note that we switched here the order of i and j. This is not what is usually done in the literature, but this will allow us to be more coherent**. Such matrix is said to be stochastic:

Definition 3.2 (*Stochastic matrix*)

A matrix $P \in \mathbb{R}^{n\times n}$ is said to be stochastic *if:*

(i) $P_{i,j} \geq 0$ *for all* $1 \leq i, j \leq n$.

(ii) $\sum_{i=1}^{n} P_{i,j} = 1$, *for all* $1 \leq j \leq n$.

Let $(X_0, X_1, \dots)$ be a Markov chain on $\{1, \dots, n\}$ with transition matrix P. For $t \geq 0$ we will encode the distribution of X_t in the $1 \times n$ vector

$$x^{(t)} = (x_1^{(t)}, \dots x_n^{(t)}) = \big(\mathbb{P}(X_t = 1), \dots, \mathbb{P}(X_t = n)\big) \in \Delta_n$$

where Δ_n is the "n-simplex"

$$\Delta_n \stackrel{\text{def}}{=} \Big\{x \in \mathbb{R}^n \,\Big|\, \sum_{i=1}^{n} x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i\Big\}.$$

Proposition 3.1

For all $t \geq 0$

$$x^{(t+1)} = Px^{(t)} \quad \textit{and consequently,} \quad x^{(t)} = P^t x^{(0)}.$$

Proof. Let $i \in \{1, \dots, n\}$.

$$x_i^{(t+1)} = \mathbb{P}(X_{t+1} = i) = \sum_{j=1}^{n} \mathbb{P}(X_{t+1} = i | X_t = j)\mathbb{P}(X_t = j) = \sum_{i=1}^{n} P_{i,j} x_j^{(t)} = (x^{(t)} P)_i.$$

□

Corollary 3.1

Let P be a stochastic matrix. Then

- *For all $x \in \Delta_n$, $Px \in \Delta_n$.*
- *For all $t \geq 1$, P^t is stochastic.*

3.2 Invariant measures and the Perron-Frobenius Theorem

We will be interested in the distribution of X_t for t large, that is the limit of $x^{(t)} = x^{(0)} P^t$. As we will see, under suitable conditions on the matrix A, this

Definition 3.3

A vector $\mu \in \Delta_n$ is an invariant measure for the transition matrix P if $\mu = P\mu$, i.e.

$$\text{for all } j \in \{1, \dots, n\}, \; \mu_i = \sum_{j=1}^{n} P_{i,j} \mu_j.$$

Remark 3.1. *An invariant measure is an eigenvector of P with associated eigenvalue* 1.

Theorem 3.1 (*Perron-Frobenius, stochastic case*)

Let P be a stochastic matrix such that there exists $k \geq 1$ such that all the entries of P^k are strictly positive. Then the following holds:

(i) *1 is an eigenvalue of P and there exists an eigenvector $\mu \in \Delta_n$ associated to 1.*

(ii) *The eigenvectors associated to 1 are unique up to scalar multiple (i.e. $\mathrm{Ker}(P - \mathrm{Id}) = \mathrm{Span}(\mu)$).*

(iii) *For all $x \in \Delta_n$, $P^t x \xrightarrow[t \to \infty]{} \mu$.*

Theorem 3.1 is proved in the next section.

Corollary 3.2

Let P be a stochastic matrix such that there exists $k \geq 1$ such that all the entries of P^k are strictly positive. Then there exists a unique invariant measure μ and for all initial condition $x^{(0)} \in \Delta_n$,

$$x^{(t)} \xrightarrow[t \to \infty]{} \mu.$$

3.3 Proof of Theorem 3.1

We first prove the theorem in the case $k = 1$, when $P_{i,j} > 0$ for all i, j.

Lemma 3.1

The mapping

$$\begin{array}{rccc} \varphi : & \Delta_n & \to & \Delta_n \\ & x & \mapsto & Px \end{array}$$

is contracting for the ℓ_1-norm: there exists $c \in (0,1)$ such that for all $x, y \in \Delta_n$:

$$\|Px - Py\|_1 \leq c\|x - y\|_1.$$

Proof. First notice that φ is well-defined by Corollary 3.1. Let us write $\alpha \overset{\text{def}}{=} \min_{i,j} P_{i,j} \in (0,1)$. Let $x, y \in \Delta_n$. We will show that $\|Px - Py\|_1 \leq (1-\alpha)\|x - y\|_1$, i.e. $\|Pz\|_1 \leq \alpha\|z\|_1$ where $z = x - y$. Compute

$$\|Pz\|_1 = \sum_{i=1}^{n} |(Pz)_i| = \sum_{i=1}^{n} \Big| \sum_{j=1}^{n} P_{i,j} z_j \Big|.$$

Since $\sum_j z_j = 0$ we have $\sum_j (P_{i,j} - \alpha/n) z_j = \sum_j P_{i,j} z_j$. Hence

$$\|Pz\|_1 = \sum_{i=1}^{n} \Big| \sum_{j=1}^{n} (P_{i,j} - \alpha/n) z_j \Big| \leq \sum_{i=1}^{n} \sum_{j=1}^{n} (P_{i,j} - \alpha/n)|z_j| = \sum_{j=1}^{n} (1-\alpha)|z_j| = (1-\alpha)\|z\|_1.$$

$\square$

Using Lemma 3.1, Banach fixed point Theorem tells us that φ admits a unique fixed point μ on Δ_n (i.e. a unique $\mu \in \Delta_n$ such that $P\mu = \mu$) and that for all $x \in \Delta_n$, $P^t x \xrightarrow[t\to\infty]{} \mu$. This proves Theorem 3.1 in the case $k = 1$.

In the case $k > 1$ we simply apply the result for $k = 1$ to P^k.

This gives that there exists a unique $\mu \in \Delta_n$ such that $P^k \mu = \mu$. Multiplying by P on both sides leads to $P^k(P\mu) = P\mu$. Since $P\mu \in \Delta_n$ we obtain that $P\mu = \mu$ by uniqueness of μ. This proves (i). To prove (ii) we consider $x \in \mathbb{R}^n$ such that $Px = x$. By iteration we get $P^k x = x$ which implies (using the result on P^k) that $x \in (\mu)$. To prove (iii) we fix $\ell \in \{0, \dots, k-1\}$. Let $x \in \Delta_n$. By applying the point (iii) to P^k, we have

$$P^{kt} P^{\ell} x \xrightarrow[t\to\infty]{} \mu.$$

Since this holds for all $\ell \leq k - 1$ we obtain that $P^T x \xrightarrow[T\to\infty]{} \mu$ using the Euclidean division of T by k.

4 Example: Google's PageRank algorithm