Université Paris Sciences et Lettres
&
Paris Dauphine

Linear Algebra and Markov Chains

Constantin Dalyac

Supervisor:
Cyril Labbé

June 28, 2018

Contents

Introduction

Markov chains were introduced by Andrei Markov in the early 20th century in an argument with Pavel Nekrasov, who claimed independence was necessary for the weak law of large numbers to hold. Markov managed to show in 1906 that under some conditions, the average of Markov chains will converge to a stationary distribution, thus proving a weak law of large numbers without the assumption of independence. Today, Markov chains are used in many domains, ranging from Biology and Physics to speech recognition. Google decided to model websites and links as Markov chains: using its mathematical properties was key in making it the most-used research engine in the world. We will see in the mathematical introduction that Markov chains can be described with matrices; a central aim of this paper is to use the tools of linear algebra in order to understand the different properties of Markov Chains, illustrating them with examples simulated with MATLAB. We will first explore the different characteristics of Markov chains and the way they evolve in time.

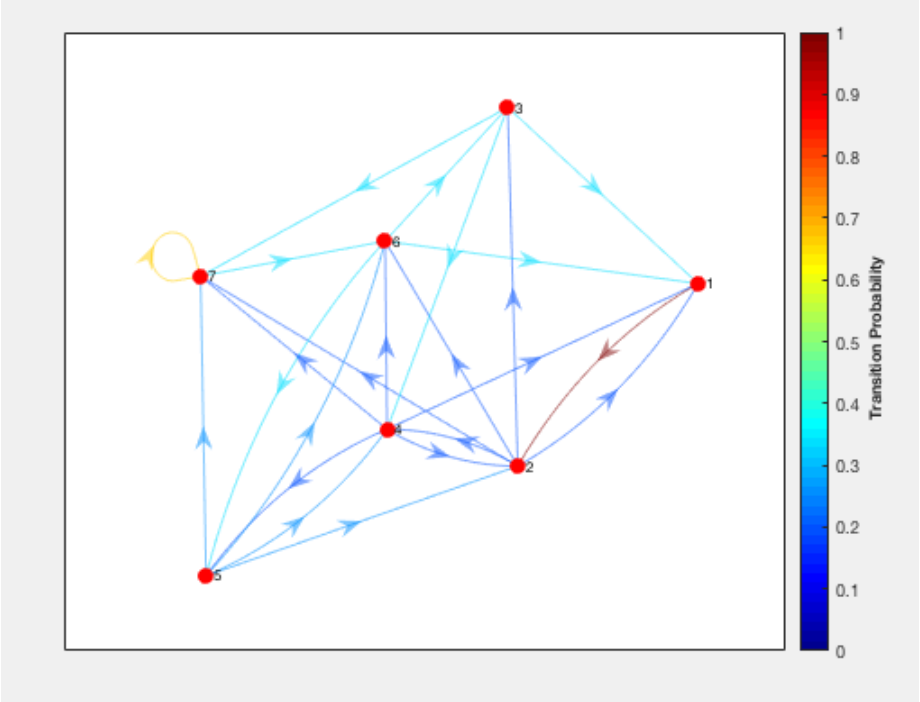


Figure 1: Example of a visualization of a Markov chain space. Vertices represents states that can be taken by a sequence of random variables. At each step, X_t moves to a different state. For example, if X_t is in state 7, X_{t+1} has 70% of chance of being state 7 and 30% of being state 6.

Part I
Matrix Representation of Markov Chains

Let Ω be a finite set of the form $\Omega = \{x_1, x_2, \ldots, x_N\}$. A finite Markov chain is a process which moves along the elements of Ω in the following manner: when at $x_i \in \Omega$, the next position is chosen according to a fixed probability distribution $P(x_i, \cdot)$. More precisely, a sequence of random variables $(X_0, X_1, \ldots)$ is a Markov chain with state space Ω and transition matrix P if for all $i, j \in [\![1; N]\!]$, all $t \geq 1$, and all events $H_{t-1} = \cap_{s=0}^{t-1}\{X_s = y_s\}$, $y_s \in \Omega$, satisfying $\mathbf{P}(H_{t-1} \cap \{X_t = x\}) > 0$, we have:

$$\mathbf{P}\{X_{t+1} = x_j \mid H_{t-1} \cap \{X_t = x_i\}\} = \mathbf{P}\{X_{t+1} = x_j \cap X_t = x_i\} = \mathbf{P}(x_i, x_j) \quad (1)$$

This equation is called the **Markov property**, meaning that the conditional probability of proceeding from state x_i to state x_j does not depend on the states preceding x_i. Hence the total information on the Markov chain is contained in a matrix $\mathrm{P} \in \mathscr{M}_N([0; 1])$. P is a stochastic matrix, i.e. its entries are all non-negative and

$$\sum_{y \in \Omega} P(x, y) = 1, \text{for all} \in \Omega.$$

Matrix representation enables us to use tools of linear algebra. Suppose we start at $t = 0$ in position x_2. We introduce a distribution vector of the form:

$$\mu_0 := (0, 1, 0, \ldots, 0),$$

where the j^{th} coordinate corresponds to the probability of presence at the state x_j. The following distribution μ_1 will then be given by the multiplication $\mu_1 = \mu_0 P$. By recurrence, we see that multiplying by P on the right updates the distribution by another step:

$$\mu_{t+1} = \mu_t P$$

and for any initial distribution μ_0,

$$\mu_t = \mu_0 P^t$$

If we multiply a column vector f by P on the left, thinking of f as a function on the state space Ω, then the x-th entry of the resulting vector is:

$$Pf(x_i) = \sum_{j=1}^{|\Omega|} P(x_i, x_j) f(x_j) = \sum_{j=1}^{|\Omega|} f(y) P_{x_i}\{X_1 = x_j\} = \mathbb{E}_{x_i}(f(X_1)$$

Hence multiplyng a column vector by P takes us from a function on the state space to the expected value of that function at the following time. A natural

question then rises: can we expect μ_t to converge to a certain distribution when t goes to infinity? And if it is the case, does the long-term distribution depend on the initial distribution μ_0 ?

1 Stationary Distributions: existence and uniqueness

We call a distribution π on Ω a ***stationary distribution*** if it satisfies the following equation:

$$\pi = \pi P.$$

It is stationary because updating the distribution by a step is done by multiplying π by P on the right, but $\pi P = \pi$ by definition, hence the distribution is unchanged. This does not imply that we "loose" the randomness in the process but it describes the fact that the probability of being in a certain state of Ω is fixed. Let us now show that under some assumptions, stationary distributions exist and are unique.

Definition 1.1. *A chain P is called **irreducible** if for any $i, j, \in [\![1; N]\!]$, there exists an integer t (possibly depending on i and j) such that $P^t(x_i, x_j) > 0$. This means it is possible to go from one state to any other using only transitions of positive probability.*

Definition 1.2. *Let $\mathscr{T}(x_i) := \{P^t(x_i, x_i) > 0\}$ be the set of times when it is possible for the chain to return to a starting position x_i. The period of the state x_i is defined to be the greatest common divisor of $\mathscr{T}(x_i)$.*

Lemma 1.1. *If P is irreducible, then gcd($\mathscr{T}(x_i)$)=gcd($\mathscr{T}(x_j)$, for all $i, j \in [\![1; N]\!]$.*

Proof. Let's choose x_i and x_j in Ω. P is irreducible therefore there exists r and l such that $P^r(x_i, x_j) > 0$ and $P^l(x_j, x_i) > 0$. Let $m := r + l$. Then $m \in \mathscr{T}(x_i) \cap \mathscr{T}(x_j)$ and $\mathscr{T}(x_i) \subset \mathscr{T}(x_j) - m$. Hence $\gcd\mathscr{T}(x_j)$ divides all elements of $\mathscr{T}(x_i)$, i.e. $\gcd\mathscr{T}(x_j) \leq \gcd\mathscr{T}(x_i)$. By a symmetric reasoning, we obtain $\gcd\mathscr{T}(x_i) \leq \gcd\mathscr{T}(x_j)$. □

For an irreducible chain, the period of the chain is defined to be the period which is common to all states. The chain will be called aperiodic if all states have period 1. If a chain is not aperiodic, we call it periodic.

Proposition 1.1. *Let P be the transition matrix of an irreducible Markov chain. There exists a unique probability distribution π satisfying $\pi = \pi P$.*

This proposition is proved by the Convergence Theorem, stated in the next part. The Convergence Theorem shows that if a Markov chain is irreducible and aperiodic, it converges in distribution to its unique stationary distribution. Moreover the theorem quantifies the speed of convergence to the stationary distribution.

Part II
Markov Chain Mixing

Since we are interested in quantifying the speed of convergence of Markov chains, we need to choose an appropriate metric for measuring the distance between distributions.

The ***total variation distance*** between two probability distributions μ and ν on Ω is defined by

$$\|\mu - \nu\|_{TV} = \max_{A \subset \Omega} |\mu(A) - \nu(A)| \tag{2}$$

2 The Convergence theorem

Theorem 1. *Suppose that P is irreducible and aperiodic, with stationary distribution π. Then there exists constants $\alpha \in$ (0,1) and $C >$ 0 such that*

$$\max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{TV} \leq C\alpha^t \tag{3}$$

Proof. See p.52 of [1]. □

In order to bound the maximal distance between $P^t(x_0, \cdot)$ and π, we define

$$d(t) := \max_{x \in \Omega} \|P^t(x, \cdot) - \pi\|_{TV}$$

. We also introduce a parameter which measures the time required by a Markov chain for the distance to stationarity to be small. The ***mixing time*** is defined by

$$t_{mix}(\epsilon) := min\{t : d(t) \leq \epsilon\} \text{ and } t_{mix} := t_{mix}(1/4).$$

3 Reversibility and Time Reversals

Tools of linear algebra can only be applied to Markov chains that are ***reversible***. We will therefore give the definition of reversibility then show how useful linear algebra can be in that case.

Suppose a probability π on Ω satisfies for all $i, j \in [\![1; N]\!]$

$$\pi(x_i)P(x_i, x_j) = \pi(x_j)P(x_j, x_i) \tag{4}$$

.

Theses equations are called the detailed balanced equations.

Proposition 3.1. *Let P be the transition matrix of a Markov chain with state space Ω. Any distribution π satisfying the detailed balanced equations is stationary for P.*

Proof. π is a stationary distribution i.i.f. $\pi = \pi P$. Let $\tilde{\pi} = \pi P$. Then for all $j \in [\![1; N]\!]$, $\tilde{\pi}_j = \sum_{i=1}^{N} P(x_i, x_j)\pi(x_i) = \sum_{i=1}^{N} P(x_j, x_i)\pi(x_j) = \pi(x_j)$ since P is stochastic. Hence $\tilde{\pi} = \pi$ □

Furthermore, when (2) holds,

$$\pi(x_0)P(x_0, x_1) \cdots P(x_{N-1}, x_N) = \pi(x_N)P(x_N, x_{N-1}) \cdots P(x_0, x_0)$$

which we can rewrite in the following suggestive form:

$$\mathbf{P}_\pi\{X_0 = x_0, \cdots, X_N = x_N\} = \mathbf{P}_\pi\{X_0 = x_N, \cdots, X_n = x_0\}$$

In other words, if a chain X_t satisfies (2) and has stationary initial distribution , then the distribution of $(X_0, X_1, \cdots, X_N)$ is the same as the distribution of $(X_N, X_{N-1}, \cdots, X_0)$. For this reason, a chain satisfying (2) is called ***reversible***.

4 Eigenvalues and relaxation time

We start by giving some facts about the eigenvalues of transition matrices:

Lemma 4.1. *Let P be the transition matrix of a finite Markov chain.*

1. *If λ is an eigenvalue of P, then $|\lambda| \leq 1$.*
2. *If P is irreducible, the vector space of eigenfunctions corresponding to the eigenvalue 1 is the one-dimensional space generated by the column vector $\mathbf{1} := (1, 1, \ldots, 1)^T$.*
3. *If P is irreducible and aperiodic, then -1 is not an eigenvalue of P.*

Proof. (A écrire) □

We denote by $\langle \cdot, \cdot \rangle$ the usual inner product on $\mathbb{R}^\Omega$, given by

$$\langle f, g \rangle = \sum_{x \in \Omega} f(x)g(x)$$

. We also define the inner product $\langle \cdot, \cdot \rangle_\pi$ as:

$$\langle f, g \rangle_\pi = \sum_{x \in \Omega} f(x)g(x)\pi(x) \tag{5}$$

Because we regard elements of $\mathbb{R}^\Omega$ as functions from Ω to $\mathbb{R}$, we will call eigenvectors of the matrix P eigenfunctions.

Lemma 4.2. *Let P be reversible with respect to π. The inner product space $(\mathbb{R}^\Omega, \langle \cdot, \cdot \rangle_\pi)$ has an orthonormal basis of real-valued eigenfunctions $\{f_j\}_{j=1}^{|\Omega|}$ corresponding to real eigenvalues $\{\lambda_j\}$.*

Proof. (A écrire) □

4.1 The relaxation time

For a reversible transition matrix P, we label the eigenvalues of P in decreasing order:

$$1 = \lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{|\Omega|} \geq -1.$$

We define $\lambda_\star := max\{|\lambda| : \lambda \text{ is an eigenvalue of P, } \lambda \neq 1\}$

The difference $\gamma_\star := 1-\lambda_\star$ is called ***the absolute spectral gap***.Lemma 4.1 implies that if P is periodic and irreducible, $\gamma_\star > 0$. The ***spectral gap*** of a reversible chain is defined by $\gamma := 1 - \lambda_2$.

The ***relaxation time*** t_{rel} of a reversible Markov chain with absolute spectral gap $\gamma_\star$ is defined to be

$$t_{rel} := \frac{1}{\gamma_\star}$$

Theorem 2. *Let P be the transition matrix of a reversible, irreducible Markov chain with state space Ω, and let $\pi_{min} := min_{x\in\Omega}\pi(x)$. Then*

$$(t_{rel} - 1)log(\frac{1}{2\epsilon}) \leq t_{mix}(\epsilon) \leq log(\frac{1}{\epsilon\pi_{min}})t_{rel}$$

We will now illustrate the previous definitions with two family of examples. The first one will be a Markov chain on a cyclic group, then we will see how it is linked to a random walk on a path.

Part III
Two examples of Markov chains

We decided to study the random walk on a cycle and on a segment.

5 Random walk on the n-cycle

Let $\omega = e^{2\pi i/n}$. The set $W_n := \{\omega, \omega^2, \cdots, \omega^{n-1}, 1\}$ represents the n roots of unity inscribed in the unit circle of the complex plane. We can therefore view simple random walk on the n-cycle as the random walk on the group $(W_n, \cdot)$ with increment distribution uniform on $\{\omega, \omega^{-1}\}$. This chain is clearly aperiodic and irreducible; according to the convergence theorem, there exists a unique stationary distribution, which is $\pi(x_i) = \frac{1}{n}$ as one would expect. Consider $P \in \mathscr{M}_N([0;1])$ the transition matrix of the random walk.

$$P = \begin{pmatrix} 0 & \frac{1}{2} & 0 & \cdots & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \frac{1}{2} & \ddots & & 0 \\ 0 & \ddots & \ddots & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & \ddots & 0 \\ 0 & \ddots & \ddots & \ddots & 0 & \frac{1}{2} \\ \frac{1}{2} & 0 & \cdots & 0 & \frac{1}{2} & 0 \end{pmatrix}$$

Let $f = \begin{pmatrix} f(\omega) \\ f(\omega^2) \\ \vdots \\ f(1) \end{pmatrix}$ be an eigenfunction of P with eigenvalue λ. It satisfies:

$$\forall k \in [\![0; n-1]\!], \qquad \lambda f(\omega^k) = Pf(\omega^k) = \frac{f(\omega^{k-1}) + f(\omega^{k+1})}{2}$$

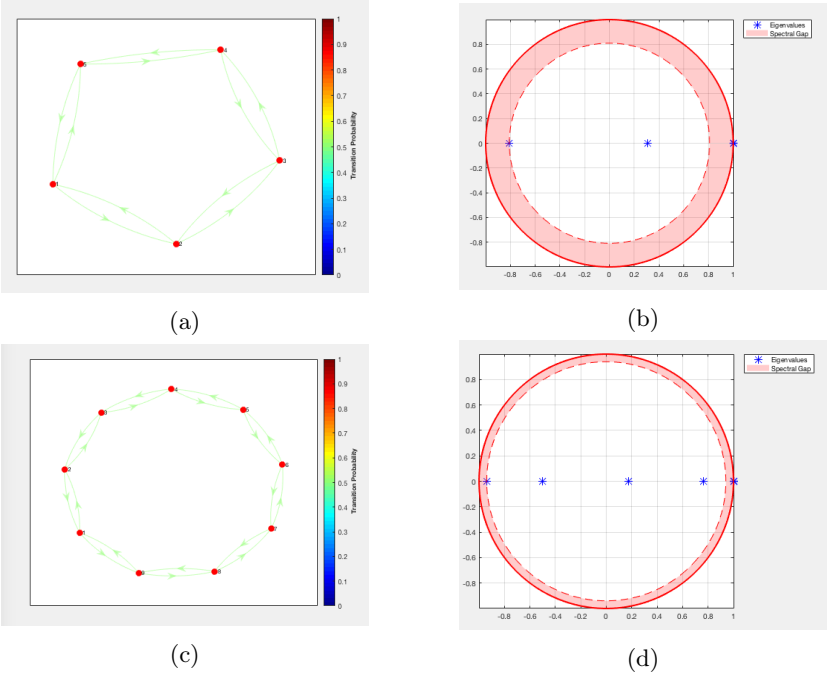
For $0 \leq j \leq n-1$, define $\phi_j(\omega^k) := \omega^{kj}$. Then

$$P\phi_j(\omega^k) = \frac{\phi_j(\omega^{k-1}) + \phi_j(\omega^{k+1})}{2} = \frac{\omega^{jk+j} + \omega^{jk-j}}{2} = \omega^{kj}(\frac{\omega^j + \omega^{-j}}{2}) = cos(\frac{2\pi j}{n})\phi_j(\omega^k),$$

hence ϕ_j is an eigenfunction of P associated to the eigenvalue $cos(\frac{2\pi j}{n})$.

We have $\lambda_2 = cos(2\pi/n) = 1 - \frac{4\pi^2}{2n^2} + O(n^{-4})$,, therefore the spectral gap is of order n^{-2} and the relaxation time is of order n^2.

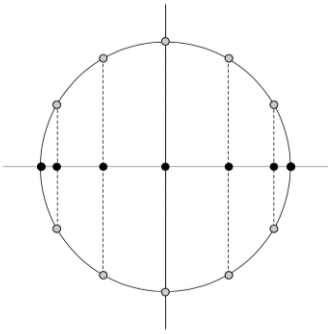
When $n = 2p$ is even, $cos(2\pi p/n) = -1$ and -1 is an eigenvalue so the absolute spectral gap is 0. That is because random walk on an even number of states is periodic; states can be classified as even or odd and each step always goes from one class to another.



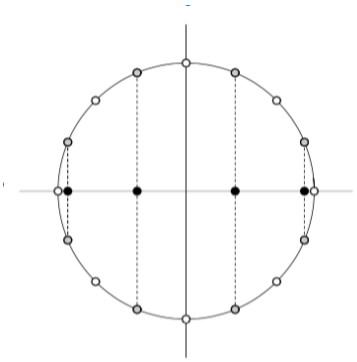
(a) (b)

(c) (d)

Figure 2: Matlab simulation of the random walk on the group $(W_5, \cdot)$ ((a) and (b)) and $(W_9, \cdot)$ ((c) and (d)). The eigenvalues are represented in blue in the complex plane, while the width of the red band represents the spectral gap. As $n \to \infty$, the spectral gap tends to 0 and relaxation time tends to ∞, as expected from our calculations.



(a) projection of the simple walk on the 12-cycle onto the real axis. We can see that for most of the projected points, except at the ends, there is the same probability of going to the left or to the right. Notice the reflecting boundary conditions.



(b) Projecting a random walk on the odd states of a 16-cycle gives a random walk on the 4-path, with holding probability of 1/2 at the endpoints.

6 Random walk on the path

The random walk on the path is closely linked to the random walk on the cycle. We first need to introduce the concept of projected chains.

Lemma 6.1. *Let Ω be the state space of a Markov chain (X_t) with transition matrix P. Let $\sim$ be an equivalence relation on Ω with equivalence classes $\Omega^\sharp = \{\bar{x} : x \in \Omega\}$, and assume that P satisfies*

$$P(x, \bar{y}) = P(x', \bar{y})$$

whenever $x \sim x'$. Then $\bar{X}_t$ is a Markov chain with state space $\Omega^\sharp$ and transition matrix $P^\sharp$ defined by $P^\sharp(\bar{x}, \bar{y}) := P(x, \bar{y})$.

We can therefore construct a new chain by taking equivalence classes for an equivalence relation compatible with the transition matrix in the sense of the Lemma. If we project the previous simple walk on the n-cycle onto the real axis, (see figure) we obtain a process which appears to be a random walk on the path.

The link between the eigenvectors and eigenvalues of the two chains is given by the following lemma:

Lemma 6.2. *With the same notations and conditions as in the previous lemma:*

Let $f : \Omega \to \mathbb{R}$ be an eigenfunction of P with eigenvalue λ which is constant on each equivalence class. Then the natural projection $f^\sharp : \Omega^\sharp \to \mathbb{R}$ of f, defined by $f^\sharp(\bar{x}) = f(x)$, is an eigenfunction of $P^\sharp$ with eigenvalue λ.

Proof. By computation: $(Pf^\sharp)(\bar{x}) = \sum_{\bar{y} \in \Omega^\sharp} P^\sharp(\bar{x}, \bar{y}) f^\sharp(\bar{y}) = \sum_{\bar{y} \in \Omega^\sharp} P(x, \bar{y}) f(y) = \sum_{\bar{y} \in \Omega^\sharp} \sum_{z \in \bar{y}} P(x, z) f(z) = \sum_{z \in \Omega} P(x, z) f(z) = (Pf)(x) = \lambda f(x) = \lambda f(\bar{x})$. □

Path with reflection at endpoints

Let P be the transition matrix of simple random walk on the 2(n-1)-cycle identified with random walk on the multiplicative group $W_{2(n-1)} = \{\omega, \omega^2, \cdots, \omega^{2n-1} = 1\}$, where $\omega = e^{\pi i/(n-1)}$. Now we choose the relation of equivalence as conjugation, i.e. $\omega^k \sim \omega^{-k}$. The equivalence respects the first lemma, and now if we identify each equivalence class with the projection of its elements on the real axis $v_k = cos(\pi k/(n-1))$, the projected chain is a simple random walk on the math with n vertices $W^\sharp = \{v_0, v_1, \cdots, v_{n-1}\}$. Note the reflecting boudary conditions; when at v_0, it moves with probability one to v_1.

According to the previous lemma and the calculation done in the previous part, the functions $f_j^\sharp : W^\sharp \to \mathbb{R}$ defined for all $j \in [\![0; n-1]\!]$ by

$$f_j^\sharp(v_k) = cos(\frac{\pi jk}{n-1})$$

are eigenfunctions of the projected walk, associated to the eigenvalue $\cos(\frac{\pi j}{(n-1)})$.

We have $\lambda_2 = cos(\pi/(n-1)) = 1 - \frac{\pi^2}{2(n-1)^2} + O(n^{-4})$, therefore the spectral gap is of order n^{-2} and the relaxation time is of order n^2, as in the simple random walk on the cycle.

Path with holding probability 1/2 at endpoints

How could change the initial chain in order to obtain, by projection, a random walk on the path such that on the endpoints there is a probability of 1/2 of staying on the same spot?

Consider $\omega = e^{\pi i/(2n)}$, and the simple random walk on the $(2n)$-element set identified with random walk on the multiplicative group $W_{odd} = \{\omega, \omega^3, \cdots, \omega^{2n-1}\}$, where at each step the current state is multiplied by an uniformly chosen element of $\{\omega^2, \omega^{-2}\}$.

Now the calculations we made for the n-cycle are the same here, and we find that the function $f_j : W_{odd} \to \mathbb{R}$ defined by

$$f_j(\omega^{2k+1}) = cos(\frac{\pi j(2k+1)}{2n})$$

is an eigenfunction with eigenvalue $cos(\frac{\pi j}{n})$.

Again, we identify the class of equivalence with the relation of conjugate, $\omega^{2k+1} \sim \omega^{-(2k+1)}$, and we identify each class of equivalence to the projections of its elements on the real axis. The projected chain is therefore a simple random walk on the path with n vertices $W^\sharp = \{u_0, u_1, \cdots, u_{n-1}\}$ and loops at the endpoints (see figure).

According to the previous lemma and the calculation done in the previous part, the functions $f_j^\sharp : W_{odd}^\sharp \to \mathbb{R}$ defined for all $j \in [\![0; n-1]\!]$ by

$$f_j^\sharp(u_k) = cos(\frac{\pi j(2k+1)}{2n})$$

are eigenfunctions of the projected walk, associated to the eigenvalue $\cos(\frac{\pi j}{n})$.

We have $\lambda_2 = cos(\pi/n) = 1 - \frac{\pi^2}{n^2} + O(n^{-4})$, therefore the spectral gap is of order n^{-2} and the relaxation time is of order n^2.

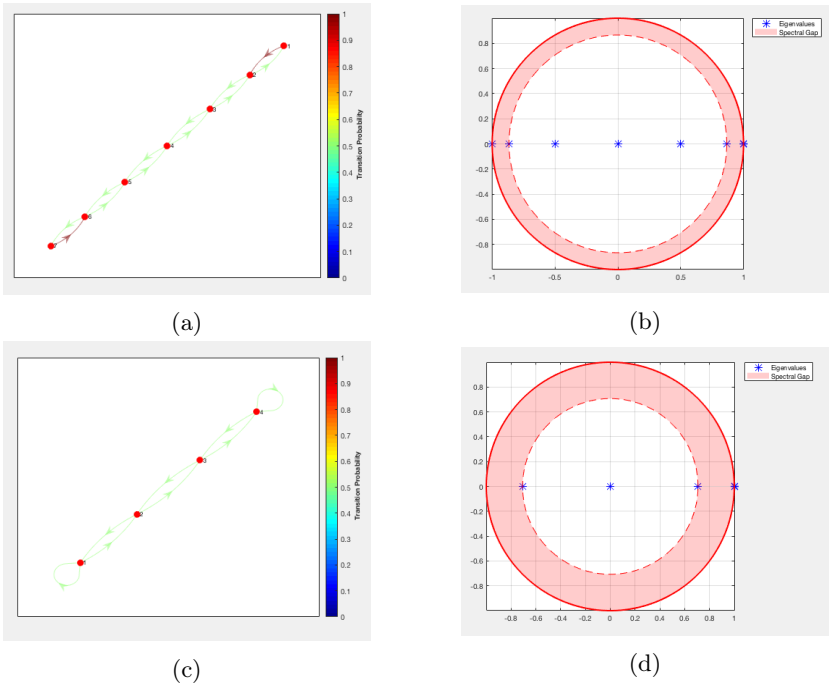


Figure 4: Matlab simulation of the random walk on the 7-path, i.e. the projected chain of the random walk on the 12-cycle ((a) and (b)) with reflection at the endpoints. ((c) and (d)) represent the calculations for the random walk on the 4-path as a projection of a random walk on the "odd" states of a 16-cycle. The eigenvalues are represented in blue in the complex plane, while the width of the red band represents the spectral gap; the simulations confirm our calculations.

Remerciement

Un grand merci à Cyril Labbé qui m'a suivi de près durant cette période, tout en me laissant le plaisir d'explorer les nombreuses facettes de cet objet mathématique si intéressant. J'ai le sentiment d'avoir joué les apprentis-chercheurs pendant ces quelques mois qui ont été baignés dans la bonne humeur; merci!

References

[1] David A. Levin, Yuval Peres, Elizabeth L. Wilmer. *Markov Chains and Mixing Times*. Vol 107,American Mathematical Soc., 2017.

[2] John Tsitsiklis. *Lecture 17: Markov Chains II*.
`https://www.youtube.com/watch?v=ZulMqrvP-Pkfeature=youtu.be`
MIT, 2011.