Chapter 1. Ideas and Definitions

This chapter introduces the basic definitions of Markov chains and related concepts.

Historical notes

The study of Markov chains originated from fundamental questions about probability and dependent events. While early probability theory focused on independent trials (as in Jacob Bernoulli's 1713 *Ars Conjectandi*), Andrey A. Markov (1856-1922) revolutionized the field by studying systems with memory of their current state.

Markov, a Russian mathematician from St. Petersburg, advocated for a rigorous and analytical approach to Probability Theory. His work emerged from a scientific dispute with Pavel A. Nekrasov, who believed independence was essential for probabilistic laws, a view stemming from his philosophical perspective. Markov extended Bernoulli's law of large numbers to dependent variables through his study of chain-linked probabilities. Markov chains not only resolved the debate with Nekrasov, but also found unexpected applications, such as in linguistic analysis. For example, he famously analyzed vowel and consonant transitions in Pushkin's *Eugene Onegin*, revealing dependencies in natural language.

An early example featuring interesting limit behaviors (similar to the law of large numbers) and embodying Markov's pioneering ideas was introduced by a very young George Pólya.

Example 0.1 (Pólya's urn)**.** An urn contains m_1 balls of color 1, m_2 balls of color 2, and so on up to m_N balls of color N. At each step, a ball is drawn (and removed from the urn) at random, and its color is observed. Afterwards, ℓ balls of the sampled color are added to the urn, together with n balls of each of the non-sampled colors.

The independent case corresponds to $\ell = 1$, $n = 0$ (or $\ell = n + 1$ more in general). But by elementary methods, one can compute a limit behavior (as the number t of iterations of the same procedure grows), which depends on the parameters $m_1, \dots, m_N, \ell, n$.

Since their inception, Markov chains have become a cornerstone of modern mathematics. Beyond statistics and probability, many (if not most) of state-of-the-art results in Dynamical Systems and Ergodic Theory have been established using Markov chains techniques. Direct applications of Markov chains are found in discrete groups, combinatorics, geometric flows theories, and in general they provide a substantial tool to interpret and predict behaviors of random and deterministic dynamics.

A partial list of subjects with very explicit applications of Markov chains includes: decision processes, Monte Carlo methods, statistical mechanics, quantum mechanics, computational physics, hidden Markov models, reinforcement learning, queueing theory, network traffic analysis, operations research, chemical kinetics, molecular dynamics, protein folding, financial modeling, stock market prediction, game theory, economic modeling, supply chain optimization, control systems, robotics, image processing, computer vision, artificial intelligence, autonomous systems, natural language processing, speech recognition, text generation, bioinformatics, genomics, disease modeling, climate and weather prediction, ecological modeling, population dynamics, evolutionary biology, phylogenetic analysis, medical imaging, brain network analysis, cognitive modeling, behavioral economics, social network analysis, search engines, recommendation systems, customer behavior analysis, blockchain, consensus algorithms,

distributed systems, fault tolerance, reliability engineering, signal processing, cryptography, energy systems, transportation systems, personalized learning systems, sports analytics and prediction, resource allocation, cybersecurity, anomaly detection, trajectory optimization.

Foundational Examples

Before giving precise definitions, let's dive into some examples that may suggest which mathematical objects are needed to properly define Markov chains, and which problems may be worth investigating.

Random Walks

Consider a person traveling between towns around the world. Each day, the person goes to the airport and buys a ticket to the destination with the best next-day weather forecast, among the towns that are reachable in less than four hours of travel. If we denote by X_t their position after t days of travel, and by $\mathbf{X} = (X_t)_{t\in\mathbb{N}}$ their entire path, we might be interested in the probabilistic distribution and properties of $\mathbf{X}$.

Let $p^t_{x,y}$ denote the probability that, on day t, y is the town with the best weather forecast among all those reachable from x in less than four hours (in particular, this value is 0 if x and y are more than four hours away). It's clear that this person's random trip exhibits the following key **Markov property**: No matter how town x was reached at time t, the probability of flying to y the next day is $p^t_{x,y}$. In probabilistic notation, we can express this property as follows: For any day t, any pair of towns x, y, and any sequence of towns $x_0, x_1, \dots, x_{t-1}$ visited before day t:

$$\mathbb{P}(X_{t+1} = y \mid X_0 = x_0, X_1 = x_1, \dots, X_t = x) = \mathbb{P}(X_{t+1} = y \mid X_t = x) =: p^t_{x,y} \tag{1}$$

It is also clear that if we know the initial distribution μ_x, corresponding to the probability that $X_0 = x$, we can compute the probability of each finite path:

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_t = x_t) = \mu_{x_0} p^0_{x_0,x_1} \cdots p^{t-1}_{x_{t-1},x_t} \tag{2}$$

From this example, we have all the ingredients needed to define Markov chains.

- The **state space** S: This is the set of towns, the space in which our traveler's position X_t takes its values.
- The **transition rules** $p^t_{x,y}$: The probabilities of moving from town $x \in S$ at time (day) t to town $y \in S$ at time $t+1$. For each t and x, $p^t_{x,\cdot} \geq 0$ represents a probability on y: $\sum_y p^t_{x,y} = 1$.
- The **initial distribution** μ: This is a probability on the state space S, representing the fact that our traveler may start at a random position (for instance, because they have been traveling randomly for some time before we start observing at time 0).
- As usual, we also need a **probability space** $(\Omega, \mathcal{F}, \mathbb{P})$ to realize our random travel, but this is *hidden* in the notation. Ω may include much more information than just the traveler's paths – for instance, the weather, random factors that influence the travel time between towns each day, and so on. In this respect, we can informally consider the σ-algebra $\mathcal{F}_t$, comprising all events observable up to day t (e.g. the weather).

We are thus left with a first problem, with an intuitive solution (but which may require some attention if S is not finite)

A. Give a mathematical definition of a Markov chain as an infinite random sequence $\mathbf{X} = (X_t)_{t\in\mathbb{N}}$, so that its distribution corresponds to the intuitive idea of randomly hopping on S with given transition probabilities.

Computer Algorithms

Modern computers executing *any* algorithm provide a profound realization of Markov chains. The processor has access to two resources:

- The memory (in a typical run on a laptop, this means the CPU registers, cache, and RAM; but it could represent different storage in other situations), which typically stores information in binary form.
- A source of randomness (e.g., a random or pseudorandom number generator), which provides independent stochastic inputs at each step.

The state space S represents the possible memory configurations. On a standard computer[1], we have $S = \{0, 1\}^N$. The configuration X_t of the memory at time t contains all the information we can access.

A single time step may correspond to a single processor clock cycle (on the order of 10^{-9} seconds) or, if we take a simplified view, the time it takes to perform a predefined block of operations in the algorithm (for instance a single computational step in a high-level programming language).

The noise is given by the intrinsic behavior of random number generator devices, as a sequence of independent random variables Z_t (in a suitable space of randomness, which on a computer is necessarily finite).

$$X_{t+1} = F(t, X_t, Z_t) \tag{3}$$

The Markov property is inherent in the fact that $F(\cdot)$ cannot depend on the past history of the memory, such as X_{t-1}, since that information is lost (or rather, what survives of it is encoded in X_t).

Example 0.2. As a concrete example, consider a randomized sorting algorithm that iterates the following steps:

1. If the list is already sorted, do nothing. Otherwise:
2. With probability p, operate a random transposition to the list.
3. With probability $1 - p$, make a random sort operation (a transposition that orders two elements).

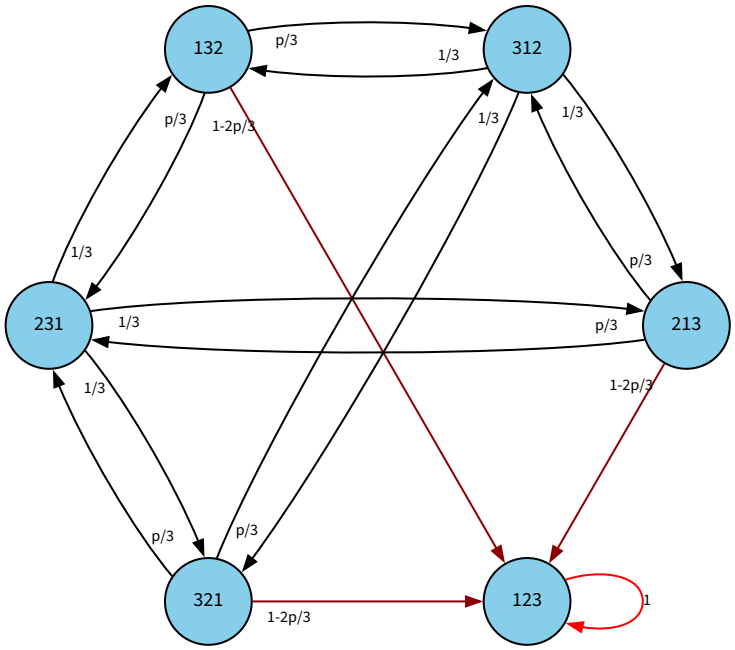

Figure 1

[1] In simplified models (for instance, if we do not want to consider rounding errors), memory can be considered a product of discrete and continuous models.

We can easily see what happens in practice if we sort just 3 elements with this algorithm. Permutations that are two transpositions away from the identity will jump with equal probability to one of their neighbors (regardless of whether step 2 or step 3 is applied; each neighbor has the same probability here). Permutations that are one transposition away from the identity will select one of their neighbors with probability $p/3$ (if step 2 is selected), but there is an additional probability $1-p$ to jump to the identity. On the identity, step 1 always applies.

The algorithm example suggests an interesting question. Let $A \subset S$ represent the set of memory states that define the termination condition of the algorithm. This is usually very explicit, e.g. A consists of states where a counter exceeds a certain value, or a `for` or `while` loop terminates. For instance, in example Example 0.2, the natural choice is $A = \{(1,2,3)\}$, the singleton of the sorted list.

If the *cost* (e.g. the actual physical time) of changing from the state $x \in S$ to $y \in S$ is $c(x,y)$, then we want to estimate $\sum_{t=0}^{\tau_A - 1} c(X_t, X_{t+1})$, where $\tau_A \in \mathbb{N}$ is the random time which equals the first time t such that X_t *touches* A (in Example 0.2, the first time the list is sorted). So some examples of interesting questions to investigate are:

B. Estimate the distribution of $\sum_{t=0}^{\tau_A - 1} c(X_t, X_{t+1})$. For instance, for $c \equiv 1$, this equals the number of steps needed to reach A. A classical problem (for a given target set A) is to minimize $\mathbb{E}[\tau_A]$ over all choices of transition probabilities (algorithms) with some given property. C. Estimate the probability of hitting a set A (for instance the solution to our problem, the sorted list), before hitting a set B (for instance, an overflow).

Exercise 0.1. Check that if $\mathbf{X} = (X_t)_{t\in\mathbb{N}}$ is obtained as Equation 3, for a given function F and a sequence $(Z_t)_{t\in\mathbb{N}}$ of independent random variables, then the property Equation 1 holds.

Financial Models

Markov chains are widely used in financial modeling to represent the stochastic evolution of asset prices, interest rates, and other economic variables. A simple, yet illustrative, example is a discrete-time model for a stock price.

Let Y_t represent the price of a stock at time t (where t might represent days, weeks, or some other discrete time interval). A standard approach is to assume a random dynamics for the *return* R_t of the stock. This means

$$Y_{t+1} = Y_t(1 + R_t)$$

where for the sake of simplicity one may assume that R_t can take on only a finite number of values, say $r_1, r_2, \dots, r_n$. For example, we may have $r_1 = -0.02$, $r_2 = 0.01$ and $r_3 = 0.05$ representing respectively, for each time period, a 2% loss, a 1% gain and a 5% gain. The transition probabilities of R_t would then describe the likelihood of each return given the current return, but it is reasonable to model them as depending on the stock price Y_t:

$$p_{i,j}(x) = \mathbb{P}(R_{t+1} = r_j \mid R_t = r_i, Y_t = x)$$

In this case, R_t is not a Markov chain, in the sense that Equation 1 does not hold (for R_t). However it is easy to check that the pair $X_t = (Y_t, R_t)$ does satisfy Equation 1, so X_t is a Markov chain.

This example shows that, from a modeling point of view, an interesting part is to actually *detect* which variables constitute a Markov chain in a given random dynamics. For the model described here, a well-known *fair pricing* problem is the following

D. Find a probability measure $\mathbb{Q}$ which is absolutely continuous w.r.t. the original probability $\mathbb{P}$ and such that Y_t is a martingale.

Language Models

A language model can be formally defined as a probability distribution over sequences of tokens (words, or characters, or a string of characters). Markov chain models provide a fundamental framework for these distributions through so-called n-gram approximations. Clearly it is too naive to assume that single tokens feature the Markov property. For instance, let's take single words as tokens. Let us consider the sentence "I want to learn stochastic processes. I will start reading about"; there is high probability that the next words will be "Markov chains". This is certainly not the case if we only have to infer how to continue the sentence from the word "about". In other words, conditioning on several tokens before the present one, modifies the probability distribution differently than conditioning on a single past token, a straightforward violation of Equation 1.

To approach this problem, we can make *n-gram approximations*. Let V be our vocabulary (set of possible tokens). An **n-th order Markov chain** model assumes the probability of each token to depend only on the previous n tokens: for $t \geq n$, $w, w_1, \dots w_t \in V$

$$\mathbb{P}(W_{t+1} = w \mid W_1 = w_1, \dots W_t = w_t) = \mathbb{P}(W_{t+1} = w \mid W_{t-n+1} = w_{t-n+1}, \dots, W_t = w_t) \tag{4}$$

In this case, to get a Markov chain, we can choose $S = V^n$, in other words the Markov property Equation 1 is recovered once we consider n-tuples of tokens as states. However in practical cases, for instance when we want to infer the next token from a previous string of tokens, this is not very effective unless we take n enormous. If we read a detective novel, and we read the sentence "The detective understood that the culprit is", the next word may very much depend on some detail written at the beginning of the book. So n may need to be as long as our text corpus. Mathematically, this forces us to enter the dangerous territory where $S = V^{\mathbb{N}}$, so we have an uncountable state space and an extra problem is added to our list:

E. Extend mathematical definitions of Markov chains when S is a generic measurable space.

Practically, on the other hand, we can still take n very large, and in this case we are left with the daunting task of estimating the transition probabilities for n-tuples of tokens from our corpus of text. Modern neural language models extend this concept through distributed representations that can capture longer-range dependencies.

https://youtu.be/KZeIEiBrT_w

Markov Chains: Definitions

Let us dig into the mathematical theory. Hereafter $(\Omega, \mathcal{F}, \mathbb{P})$ is a given probability space.

Definition 0.1 (Markov chain on a countable state space)**.** Let S be a non-empty finite or countable set. A **Markov chain $\mathbf{X} = (X_t)_{t\in\mathbb{N}}$** is a sequence of random variables $X_t \colon \Omega \to S$ such that, for $t \in \mathbb{N}$ and $x \in S$

$$\mathbb{P}(X_{t+1} = x \mid (X_s)_{s\leq t}) = \mathbb{P}(X_{t+1} = x \mid X_t) \tag{5}$$

S is called the **state space** of the Markov chain. The quantities $p^t_{x,y} := \mathbb{P}(X_{t+1} = y \mid X_t = x)$ are called the **transition probabilities** and Equation 5 is called the **Markov property**.

The Markov chain is called **homogeneous** (or time-homogeneous) if the transition probabilities $p^t_{x,y} \equiv p_{x,y}$ do not depend on the time parameter t.

Exercise 0.2. Check that transition probabilities are indeed probabilities in the following sense: for each $t \geq 0$, $x \in S$, $p^t_{x,\cdot}$ defines a probability on S:

- $p^t_{x,y} \geq 0$.
- $\sum_{y\in S} p^t_{x,y} = 1$.

Definition 0.2 (Skorohod space on a countable state space)**.** Let S be a finite or countable set and let $(p^t_{x,y})_{t\geq 0;x,y\in S}$ be transition probabilities. The **Skorohod space** associated to $(p^t_{x,y})$ is the path space

$$D(S) := \left\{\mathbf{x} \in S^{\mathbb{N}} \,:\, p^t_{x_t,x_{t+1}} > 0, \text{ for all } t \in \mathbb{N}\right\}$$

$D(S)$ is naturally equipped with the σ-algebra it inherits as a subspace of $S^{\mathbb{N}}$.

Definition 0.3. For S, $p^t_{x,y}$ as above, the **multi-step transition probabilities** are defined for $s \leq t$ as

$$p^{(s,t)}_{x,y} := \sum_{z_{s+1},\dots z_{t-1}\in S} p^s_{x,z_{s+1}} p^{s+1}_{z_{s+1},z_{s+2}} \cdots p^{t-1}_{z_{t-1},y} \tag{6}$$

in particular $p^{(t,t+1)}_{x,y} = p^t_{x,y}$ and $p^{(t,t)}_{x,y} = \delta_x(\{y\})$.

If $p^t_{x,y} \equiv p_{x,y}$ is time-homogeneous, then $p^{(s,t)}$ only depends on $t-s$, and thus we denote in this case $p^{(t)}_{x,y} \equiv p^{(s,s+t)}$ regardless of $s \geq 0$.

Exercise 0.3. Verify that the multi-step transition probabilities are indeed consistent with their name, namely if $\mathbf{X}$ is a Markov chain with transition probabilities $(p^t_{x,y})$ then

$$\mathbb{P}(X_t = y \mid X_s = x) = p^{(s,t)}_{x,y}$$

Exercise 0.4. Verify that the multi-step transition probabilities satisfy the Chapman-Kolmogorov equations (also known as the semigroup property): for $s \leq t \leq u$

$$p^{(s,u)}_{x,z} = \sum_{y\in S} p^{(s,t)}_{x,y} p^{(t,u)}_{y,z}$$

Let S be a finite or countable set[2], and $\mathbf{X}$ a Markov process with:

- **initial distribution** $\mu \in \mathcal{P}(S)$, a probability measure on S.
- **transition probabilities** $(p^t_{x,y})_{t\geq 0;x,y\in S}$.

Then the law of $\mathbf{X}$ on $D(S)$ is uniquely identified by μ and $(p^t_{x,y})$ since

$$\mathbb{P}(X_0 = x_0, X_1 = x_1, \dots, X_t = x_t) = \mu_{x_0} p^0_{x_0,x_1} \cdots p^{t-1}_{x_{t-1},x_t}$$

for all $t \geq 0$ and $x_0, \dots, x_t \in S$. By Kolmogorov's theorem, this readily identifies a unique probability measure on $D(S)$.

Remark. In general, we think that the transition probabilities are fixed once and for all, while the initial condition may change. It comes therefore handy the following notation:

- For $x \in S$, $\mathbb{P}_x$ denotes the probability $\mathbb{P}_x(\cdot) = \mathbb{P}(\cdot \mid X_0 = x)$ (where we silently assumed $\mathbb{P}(X_0 = x) > 0$).
- More in general, for μ a probability measure on S

$$\mathbb{P}_\mu := \sum_x \mu_x \mathbb{P}_x$$

[2]We do not need S to be finite or countable. But for technical reasons (applying Kolmogorov's theorem), one still needs S to be a reasonable measurable space.

This has to be interpreted in the linear sense, e.g.

$$\mathbb{E}_\mu[f(X_1, X_2)] = \sum_x \mu_x \mathbb{E}_x[f(X_1, X_2)]$$

Notice in particular that $\mathbb{P}_{\delta_x} \equiv \mathbb{P}_x$. The point of this notation is that we are usually just concerned with the *law of* $\mathbf{X}$, and **under** $\mathbb{P}_\mu$**,** $\mathbf{X}$ **is a Markov chain with the same transition probabilities (as under the original** $\mathbb{P}$**), and initial distribution** μ. To be precise, if we start with a Markov chain $\mathbf{X}\colon \Omega \to S^{\mathbb{N}}$ over the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and for the sake of simplicity we take $\mathbb{P}(X_0 = x) > 0$ for all $x \in S$, then for each probability $\mu \in \mathcal{P}(S)$ we can consider the space $(\Omega, \mathcal{F}, \mathbb{P}_\mu)$, where $\mathbf{X}$ features the same transition probabilities, but initial distribution μ.

Remark. If $\mathbf{Y}$ is not time-homogeneous, we can define the Markov chain $\mathbf{X}$ with state space $\mathbb{N} \times S$, by $X_t := (t, Y_t)$ and transition time-homogeneous probabilities $q_{(s,x),(t,y)} = \mathbf{1}_{t,s+1}\, p^s_{x,y}$, where $p^s_{x,y}$ are the transition probabilities of $\mathbf{Y}$.

In view of the last remark, we will almost exclusively consider homogeneous Markov chains hereafter. This is not completely general but simplifies the notation and allows stronger results.

Markov operators

Hereafter we fix the transition probabilities $(p_{x,y})$, and an associated Markov chain $\mathbf{X}$.

Let $\mathcal{B}(S)$ be the space of all bounded (measurable if S is uncountable) functions $f\colon S \to \mathbb{R}$. We define a linear operator

$$\begin{aligned} &P\colon \mathcal{B}(S) \to \mathcal{B}(S) \\ &(Pf)(x) := \sum_{y \in S} p_{x,y} f(y) = \mathbb{E}_x[f(X_1)] = \mathbb{E}[f(X_{t+1}) \mid X_t = x], \qquad t \geq 0, x \in S \end{aligned}$$

Similarly for $\mu \in \mathcal{P}(S)$ a probability measure on S, we define

$$(\mu P)_x := \sum_{y \in S} \mu_y p_{y,x} = \mathbb{P}_\mu(X_1 = x)$$

Exercise 0.5. Check that (μP) defines a probability measure on S.

Exercise 0.6. Check that P is a contraction $\sup_x (Pf)(x) \leq \sup_x f(x)$.

Remark. If S is finite, the operator P is nothing but the operator represented by the matrix $(p_{x,y})$. It operates on functions to the right, and on probability measures to the left. For S countable, this stays true, just sums become series, just we need to be sure that f or μ are bounded to define $Pf(x)$ and $(\mu P)_x$ in this case.

Then the Chapman-Kolmogorov equation becomes the simple fact that the multi-step transition probabilities $p^{(t)}_{x,y}$ represent the entries of the power P^t of P, as indeed $p^{(1)}_{x,y} = p_{x,y}$ and $p^{t+1}_{x,y} = \sum_z p^t_{x,z} p_{z,y}$.

Markov chains: Notation

Let us recap the notation we have seen so far, since there are several objects living in different spaces, and that may be confusing.

- $(\Omega, \mathcal{F}, \mathbb{P})$ is a given probability space. The exact nature of this space is irrelevant and non-canonical. We want to avoid even mentioning which space Ω we take exactly. This is the same attitude one may have in Geometry, when one wants to define a manifold as such, not necessarily specifying a parametrization or a specific atlas.
- The time is assumed discrete, we usually denote the time with the letters $s, t, u \in \mathbb{N}$.
- S is the state space, the space where the Markov chain takes value. Elements in S are denoted x, y, z etc.
- $\mathbf{X} \colon \Omega \to S^{\mathbb{N}}$ is the actual *random* chain. It is a sequence $(X_0, X_1, ...)$ of (measurable) functions $X_t \colon \Omega \to S$.
- For each t, X_t can be thought as a random element of S. It has therefore a distribution μ_t on S, defined by $\mu_{t,x} := \mathbb{P}(X_t = x)$.
- The transition probabilities are given by $p_{x,y} = \mathbb{P}_x(X_1 = y)$, which can be interpreted as the entries of the Markov operator P. We can start from a Markov chain $\mathbf{X}$ and define it initial distribution μ and transition probabilities. Or we can start from a $\mu \in \mathcal{P}(S)$ and the transition probabilities, and construct a Markov chain $\mathbf{X}$.
- We have at the same time a random trajectory on S, and a deterministic trajectory[3] on $\mathcal{P}(S)$
  - A random sequence of elements on S: $X_0, X_1, \dots, X_t, \dots$.
  - A deterministic sequence of elements of $\mathcal{P}(S)$: $\mu_0 = \mu, \mu_1 = \mu P, \dots, \mu_t = \mu P^t$.
- Ultimately we can represent the a Markov chain either with a (possibly infinite) matrix with entries $p_{x,y}$ (i.e. a matrix with non-negative entries and such that each row sums to 1) or with a weighted directed graph (where and edge (x, y) is present if $p_{x,y} > 0$, and the sum of the weights of the arrows outgoing from each vertex is 1).

Example 0.3. The operator P represented by the matrix

$$P = \begin{pmatrix} 0.0 & 0.4 & 0.0 & 0.1 & 0.5 \\ 0.0 & 0.2 & 0.3 & 0.5 & 0.0 \\ 0.0 & 0.0 & 0.6 & 0.0 & 0.4 \\ 0.7 & 0.0 & 0.0 & 0.1 & 0.2 \\ 0.0 & 0.8 & 0.0 & 0.2 & 0.0 \end{pmatrix}$$

corresponds to the directed graph

Additional exercises

Exercise 0.7. A football team plays in a championship consisting of eleven games. Their performance strongly depends on their morale:

- At each match. they always have probability $1/3$ to draw.
- If they won the last match, they will win again with probability $1/2$ (and lose with probability $1/6$ regardless of the previous matches).
- If they lost the last match, they will lose again with probability $1/2$ (and win with probability $1/6$ regardless of the previous matches).
- If the last match was a draw, they will lose or win with probability $1/3$ (regardless of the previous matches).

a. Compute the probability that they will win the third match if they won/lost/draw their first.
b. Compute the probability that they will win the last match if they won/lost/draw their first.

[3] While knowing $\mathbf{X}$ is exactly equivalent to knowing $X_0, X_1, \dots,$, knowing the law of $\mathbf{X}$ (a probability on $D(S)$), contains more information than the sequence $\mu_0, \mu_1, \dots$ of the laws of $X_0, X_1, \dots$.

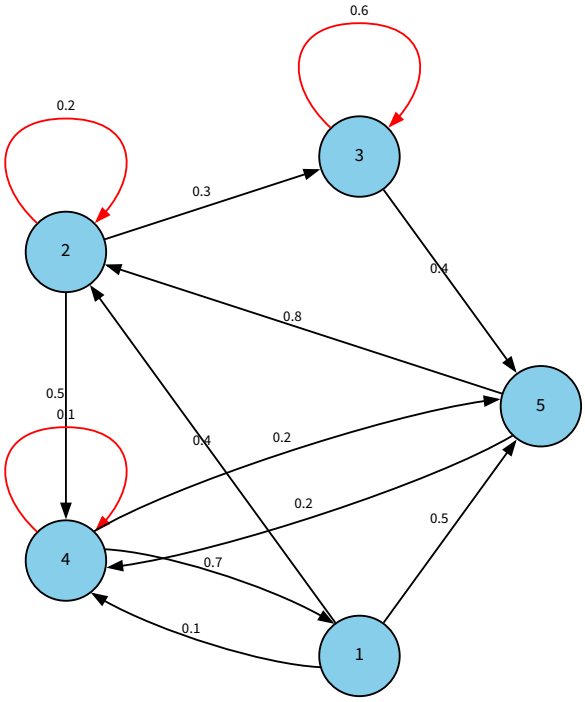

Figure 2

Solution

The matrix P with entries $p_{x,y}$ where $x, y \in S := \{W, D, L\}$ is given componentwise

$$P := \begin{pmatrix} 1/2 & 1/3 & 1/6 \\ 1/3 & 1/3 & 1/3 \\ 1/6 & 1/3 & 1/2 \end{pmatrix}$$

By symmetry, it is clear that $p^{(n)}_{x,D} = p^{(n)}_{D,x} = 1/3$, and $a_n := p^{(n)}_{W,W} = p^{(n)}_{L,L}$ and $b_n := p^{(n)}_{L,W} = p^{(n)}_{W,L}$ with $a_n + 1/3 + b_n = 1$. So if we write $a_n = 1/3 + c_n$ and $b_n = 1/3 - c_n$

$$P^n := \begin{pmatrix} 1/3 + c_n & 1/3 & 1/3 - c_n \\ 1/3 & 1/3 & 1/3 \\ 1/3 - c_n & 1/3 & 1/3 + c_n \end{pmatrix}$$

and by recurrence we get $c_n = 3^{-n}/2$.

a. From the previous computation we get that the required probabilities are $p^{(2)}_{W,W} = 7/18$, $p^{(2)}_{D,W} = 1/3$, $p^{(2)}_{L,W} = 5/18$.
b. The intuitive idea that all these probabilities will converge quickly to $1/3$ (as computed) is clear: $p^{(10)}_{W,W} \approx 0.333342$, $p^{(2)}_{L,W} \approx 0.333325$.

Exercise 0.8. My watch got crazy. Instead of rotating clockwise all the time, it has a random behavior. Each hour, it rotates clockwise (that is it adds one hour) with probability $1/2$, and counterclockwise with probability $1/2$. At midnight it marked the correct time $x = 0$. What is the probability that this night (i.e. after 24 hours) it will mark the correct time again?

Solution

P is a circulant matrix with $p_{x,x+1} = p_{x,x-1} = 1/2$ (where ± 1 is understood $\pmod 1 2$), and all the other entries are 0. If $\eta := e^{i2\pi/N}$ where $N = 12$, we have that P has eigenvalues $(\eta^k + \eta^{-k})/2 = \cos(2\pi k/N)$ for $k = 0, \dots, N-1$. We can write $P = VDV^{-1}$ with diagonal with the aforementioned eigenvalues and

$$V_{j,k} = e^{i2\pi jk/N} \quad V^{-1}_{j,k} = e^{-i2\pi jk/N}/N \tag{7}$$

Since $P^n = VDV^{-1}$, we have $p^{(n)}_{x,y} = \frac{1}{N}\sum_{k=0}^{N-1} \cos(2\pi k/N)^n e^{i2\pi k(x-y)/N}$. In particular

$$p^{(n)}_{0,0} = \sum_{k=0}^{N-1} \cos(2\pi k/N)^n/N$$

and $N = 12$ and $n = 24$ we get $p^{(n)}_{0,0} = 1486675/8388608 \approx 0.177$.
We can also say that the watch needs to take as many steps clockwise as counterclockwise, $\pmod N$. This gives

$$p^{(n)}_{0,0} = \sum_{k\in\mathbb{Z}} \binom{n}{(n+kN)/2} 2^{-n}$$

where the binomial coefficient is interpreted as 0 if $n + KN/2 \notin \{0, 1, \dots, n\}$. This equals of course the same value computed before.

Exercise 0.9. Let $P = (p_{x,y})_{x,y\in S}$ be a matrix, with S finite. Prove that the following are equivalent:

- P is a transition probability (also called stochastic) matrix, i.e. $p_{x,y} \geq 0$ for all $x, y \in S$ and $\sum_y p_{x,y} = 1$ for all $x \in S$.
- For each non-negative function $f\colon S \to \mathbb{R}$, Pf is non-negative. Moreover $P\mathbf{1} = \mathbf{1}$, where $\mathbf{1}$ is the function identically equal to 1.
- For each probability distribution μ on S, μP is also a probability distribution.

What changes if S is countable but not finite?

Exercise 0.10. Check that if the transition probabilities $p_{x,y} = m_y$ do not depend on x, then regardless of the initial distribution of X_0, $X_1, X_2 \dots$ is an i.i.d. sequence.

Find an example of (for a specific initial condition), where $(X_0, X_1, \dots)$ are i.i.d. despite $p_{x,y}$ being dependent on x.

Abstraction

The standard Markov chain construction

Let us try to find the basic ingredients for a general definition of Markov chains from a measure-theoretic point of view. The main point is that we can take the state space S a general measurable space, although we will take it a Polish space equipped with its Borel σ-algebra, to avoid any regularity issue. Then the transition probabilities will be measurable maps $S \ni x \mapsto p(x,\cdot) \in \mathcal{P}(S)$. In other word, when the Markov chain is at a point $x \in S$, it will jump to a set $A \subset S$ with probability $p(x,A)$. Here *measurable* means that the map $x \mapsto p(x,A)$ is measurable as a map from S to $\mathbb{R}$. All the constructions and results provided here generalize to this case.

A more involved abstraction layer

One can try to fit Markov chains in a more general framework, where X_t leaves in a different space for each t, and the set Θ of times is just a partially ordered set. There are not many relevant usages of Θ beyond subset of $\mathbb{Z}$ and $\mathbb{R}$, a notable exception being compact subset in $\mathbb{C}$ ordered by inclusion.

Definition 0.4 (Measurable bundle)**.** $(E,\Theta,\pi,(\mathcal{G}_t)_{t\in\Theta})$ is a **measurable bundle** if:

- E is a non-empty set, called the **total space**.
- Θ is a non-empty set, called the **base**.
- $\pi\colon E \to \Theta$ is a surjective map.
- $\mathcal{G}_t$ is a σ-algebra on $\pi^{-1}(\{t\})$, for $t \in \Theta$.

A **section** is a map $\mathbf{X}\colon \Theta \to E$, which is a right-inverse of π, namely $\pi(X_t) = t$ for $t \in \Theta$. A section can equivalently be regarded as a collection $\mathbf{X} = (X_t)_{t\in\Theta}$ with $X_t \in \pi^{-1}(\{t\})$. The space of sections is denoted $D(E)$.

Definition 0.5 (Stochastic section)**.** Let $(\Omega,\mathcal{F})$ be a measurable space and $(E,\Theta,\pi,(\mathcal{G}_t)_{t\in\Theta})$ a measurable bundle. A map $\mathbf{X}\colon \Omega \to D(E)$ is a **stochastic section** if the map $\Omega \ni \omega \mapsto X_t(\omega) \in \pi^{-1}(\{t\})$ is $\mathcal{F}/\mathcal{G}_t$-measurable.

Definition 0.6 (Markov process)**.** Let $(\Omega,\mathcal{F},(\mathcal{F}_t)_{t\in\Theta},\mathbb{P})$ be a filtered probability space, and $(E,\Theta,\pi,(\mathcal{G}_t)_{t\in\Theta})$ a measurable bundle. A stochastic section $\mathbf{X}$ is a **Markov process** if

$$\mathbb{P}(X_t \in A \mid \mathcal{F}_s) = \mathbb{P}(X_t \in A \mid X_s) \qquad \forall s \le t, A \in \mathcal{G}_t$$

If we define $S_t := \pi^{-1}(\{t\})$, then the transition probabilities are maps $p^{(s,t)}\colon S_s \to \mathcal{P}(S_t)$ given by[4]

$$p^{(s,t)}(x,A) := \mathbb{P}(X_t \in A \mid X_s = x)$$

[4] We are assuming a minimal regularity of the space S_t, in order to guarantee that $\mathbb{P}(X_t \in A \mid X_s)$ is indeed a measurable function of X_s, so that we can write $\mathbb{P}(X_t \in A \mid X_s = x)$ unambigously.