Markov Chains

Adam Kelly (ak2316@cam.ac.uk)

October 14, 2021

A stochastic process is said to have the 'Markov property' if, conditional on its present value, the future is independent of the past.

This is a *restrictive* assumption, but we do end up with a useful model with a rich mathematical theory, which we shall study in this course.

This article constitutes my notes for the 'Markov Chains' course, held in Michaelmas 2021 at Cambridge. These notes are *not a transcription of the lectures*, and differ significantly in quite a few areas. Still, all lectured material should be covered.

Contents

§1 The Markov Property

§1.1 What is a Markov Chain?

Let S be a countable set (the set of possible 'states'), and let X_n be a sequence of random variables taking values in S.

Definition 1.1 (Markov Chain)
The sequence of random variables X_n is a **Markov chain** if it satisfies the **Markov property**

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

The Markov chain is said to be **homogeneous** if for all $i, j \in S$ the conditional probability $\mathbb{P}(X_{n+1} = j \mid X_n = i)$ is independent of n.

In this course we are only going to study homogeneous Markov chains.

§2 Introduction

For this whole course, I will be a finite or countable set. All of our random variables will also be defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$.

Definition 2.1 (Markov Chain)
A **stochastic process** $(X_n)_{n\geq 0}$ is called a **Markov chain** if for all $n \geq 0$ and all $x_0, \dots, x_{n+1} \in I$, we have[a]

$$\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, \dots, X_0 = x_0) = \mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n).$$

[a] We assume here that we are not conditioning on a zero probability event.

Remark. This definition gives a *discrete time* Markov chain. It is possible to define a continuous time Markov chain, but we won't worry about that for now.

If $\mathbb{P}(X_{n+1} = y \mid X_n = x)$ for all $x, y \in I$ is independent of n, then X is called **time-homogeneous**. Otherwise, it is **time-inhomogeneous**. In this course, we will only study time-homogeneous Markov chains.

We will write $P(x, y) = \mathbb{P}(X_1 = y \mid X_0 = x)$, where $x, y \in I$. We call P a **stochastic matrix**, because

$$\sum_{y \in I} P(x, y) = 1,$$

that is, the sum of each row is 1.

Remark. The index set does not have to be $\mathbb{N}$, it could be say $\{0, 1, \dots, N\}$ for $N \in \mathbb{N}$.

So to characterize a Markov chain, we need this matrix P, giving the probability of passing from a state x to a state y. We call this matrix the **transition matrix** of X.

Definition 2.2 (Markov)
We say that X is Markov(λ, P) if X_0 has distribution λ and P is the transition matrix. That is,

(i) $\mathbb{P}(X_0 = x_0) = \lambda_{x_0}$, $x_0 \in I$,

(ii) $\mathbb{P}(X_{n+1} = x_{n+1} \mid X_n = x_n, \dots, X_0 = x_0) = P(x_n, x_{n+1}) = P_{x_n x_{n+1}}$.

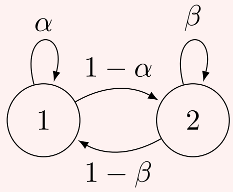
We usually represent a Markov chain by its diagram corresponding to the allowed transitions.

Example 2.3 (Diagram of a Markov Chain)
Let $\alpha, \beta \in (0, 1)$. We consider the matrix

$$P = \begin{pmatrix} \alpha & 1-\alpha \\ 1-\beta & \beta \end{pmatrix}.$$

This is a transition matrix on two states which we can call 1 and 2. Here α is the probability of staying at 1, and $1-\alpha$ is the probability of moving from state 2 when at state 1.

A diagram of this is given below. This is a directed graph with the relevant probabilities labelling each edge.

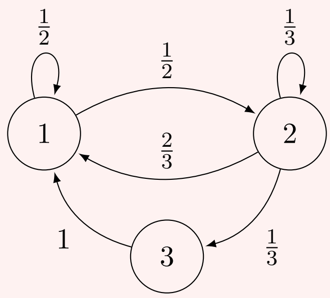


Example 2.4

Suppose that we have the transition matrix

$$P = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} & 0 \\ \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 \end{pmatrix}.$$

This is a transition matrix on three states and corresponds with the diagram below.



Theorem 2.5

The process X is Markov(λ, P) if and only if for all $n \geq 0$ and all $x_0, \dots, x_n \in I$ we have

$$\mathbb{P}(X_0 = x_0, \dots, X_n = x_n) = \lambda_{x_0} P(x_0, x_1) \cdots P(x_{n-1}, x_n).$$

Proof. First suppose that X is Markov(λ, P). Then

$$\begin{aligned} \mathbb{P}(X_n = x_n, \dots, X_0 = x_0) &= \mathbb{P}(X_n = x_n \mid X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &\quad \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &= P(x_{n-1}, x_n) \cdot \mathbb{P}(X_{n-1} = x_{n-1}, \dots, X_0 = x_0) \\ &= P(x_0, x_1) \dots P(x_{n-1}, x_n) \mathbb{P}(X_0 = x_0) \\ &= \lambda_{x_0} P(x_0, x_1) \cdots P(x_{n-1}, x_n), \end{aligned}$$

as required.

Now suppose that the property holds. Then $n = 0$ gives $\mathbb{P}(X_0 = x_0) = \lambda_{x_0}$, so our base case holds. Then

$$\begin{aligned}\mathbb{P}(X_n = x_)n \mid X_{n-1} = x_{n-1}, \dots, X_0 = x_0) &= \frac{\lambda_{x_0} P(x_0, x_1) \cdots P(x_{n-1}, x_n)}{\lambda_{x_0} P(x_0, x_1) \cdots P(x_{n-2}, x_{n-1})} \\ &= P(x_{n-1}, x_n)\end{aligned}$$

□

Now we are going to define some useful notation.

Definition 2.6 (δ_i-mass)

For $i \in I$, the **δ_i-mass** of i is defined as $\delta_{ij} = \mathbb{1}(i = j)$.

Recall the notion of independence for random variables. Let $X_1, \dots, X_n$ be discrete random variables. They are *independent* if for all $x_1, \dots, x_n \in I$, we have

$$\mathbb{P}(X_1 = x_1, \dots, X_n = x_n) = \prod_{i=1}^{n} \mathbb{P}(X_i = x_i).$$

We have a similar notion for sequences of random variables. We say a sequence $(X_n)_{n \geq 0}$ is *independent* if for all $i_1 < i_2 < \cdots < i_k$ and all $x_1, \dots, x_k$,

$$\mathbb{P}(X_{i_1} = x_1, \dots, X_{i_k} = x_k) = \prod_{j=1}^{k} \mathbb{P}(X_{i_j} = x_j).$$

If $X = (X_n)_{n \geq 0}$ and $Y = (Y_n)_{n \geq 0}$ are two sequences of random variables, they are independent if for all k, m and $i_1 < \cdots < i_k$, $j_1 < \cdots j_m$ we have

$$\begin{aligned}&\mathbb{P}(X_{i_1} = x_1, \dots, X_{i_k} = x_k, Y_{j_1} = y_1, \dots, Y_{j_m} = y_m) \\ &= \mathbb{P}(X_{i_1} = x_1, \dots, X_{i_k} = x_k) \cdot \mathbb{P}(Y_{j_1} = y_1, \dots, Y_{j_m} = y_m)\end{aligned}$$

Theorem 2.7 (Simple Markov Property)

Suppose that $X \sim \text{Markov}(\lambda, P)$. Fix $m \in \mathbb{N}$ and $i \in I$. Conditional on $X_m = i$, the process $(X_{m+n})_{n \geq 0}$ is $\text{Markov}(\delta_i, P)$ and it is independent of $X_0, \dots, X_m$.

Proof. We have

$$\mathbb{P}\left(X_{m+n} = x_{m+n}, \dots, X_m = x_m \mid X_m = i\right) = \frac{\mathbb{P}(X_{m+n} = x_{m+n}, \dots, X_m = x_m)\delta_{i x_m}}{\mathbb{P}(X_m = i)}.$$

We can rewrite the numerator as

$$\begin{aligned}&\mathbb{P}\left(X_{m+n} = x_{m+n}, \dots, X_m = x_m\right) \\ &= \sum_{x_0, \dots, x_{m-1} \in I} \mathbb{P}\left(X_{m+n} = x_{m+n}, \dots, X_m = x_m, X_{m-1} = x_{m-1}, \dots, X_0 = x_0\right) \\ &= \sum_{x_0, \cdots, x_{m-1}} \lambda_{x_0} P\left(x_0, x_1\right) \cdots P(x_{m-1}, x_m) P(x_m, x_{m+1}) \cdots P(x_{m+n-1}, x_{m+n}) \\ &= P\left(x_m, x_{m+1}\right) \cdots P\left(x_{m+n-1}, x_{m+n}\right) \mathbb{P}\left(X_m = x_m\right).\end{aligned}$$

Substituting this back into our original expression, we get

$$\mathbb{P}\left(X_{m+n}=x_{m+n},\ldots,X_m=x_m \mid X_m=i\right)=\delta_{ix_m}P(x_m,x_{m+1})\cdots P(x_{m+n-1},x_{m+n}),$$

showing that $(X_{m+n})_{n\geq 0}$ is Markov(δ_i,P) conditional on $X_m=i$. □

Remark. Informally, this theorem says 'past and future are independent given the present'.

§3 MISSING LECTURE 2

§4 Lecture 3

Example 4.1

Given the transition matrix

$$P=\begin{pmatrix}0 & 1 & 0\\ 0 & 1/2 & 1/2\\ 1/2 & 0 & 1/2\end{pmatrix},$$

we want to find $p_{11}(n)$. The eigenvalues of this matrix are $1, i/2$ and $-i/2$. We write $i/2=(\cos\pi/2+i\sin\pi/2)/2$, and then we can write the general form for $p_{11}(n)$ as

$$p_{11}(n)=\alpha+\beta\cdot\left(\frac{1}{2}\right)^n\cos\left(\frac{n\pi}{2}\right)+\gamma\cdot\left(\frac{1}{2}\right)^n\sin\left(\frac{n\pi}{2}\right).$$

We can then compute by hand $p_{11}(0)=1$, $p_{11}(1)=0$ and $p_{11}(2)=0$. So, if you solve this system for α,β,γ, you get

$$p_{11}(n)=\frac{1}{5}+\left(\frac{1}{2}\right)^n\left(\frac{4}{5}\cos\left(\frac{n\pi}{2}\right)-\frac{2}{5}\sin\left(\frac{n\pi}{2}\right)\right).$$

§4.1 Communicating Classes

Definition 4.2

X is a Markov chain with transition matrix P and values in I. For $x,y\in I$ we say that x **leads to** y and write it $x\to y$ if

$$\mathbb{P}(X_m=y \text{ for some } n\geq 0)>0.$$

We say that x **communicates** with y and write $x\longleftrightarrow y$ if both $x\to y$ and $y\to x$.

Theorem 4.3

The following are equivalent:

(i) $x\to y$;

(ii) There exists a sequence of states $x=x_0,x_1,\ldots,x_k=y$ such that

$$P(x_0,x_1)P(x_1,x_2)\cdots P(x_{k-1},x_k)>0;$$

(iii) There exists $n \geq 0$ such that $p_{xy}(n) > 0$.

Proof. Trivial. □

Corollary 4.4

$\longleftrightarrow$ is an equivalence relation on I.

Proof. Trivial. □

Definition 4.5 (Communicating Classes)

The equivalence classes induced by $\longleftrightarrow$ on I are called **communicating classes**.

A communicating class C is **closed** if whenever $x \in C$ and $x \to y$ then $y \in C$.

A matrix P is called **irreducible** if it has a single communicating class, that is, for all $x, y \in I$ we have $x \longleftrightarrow y$.

A state x is called **absorbing** if $\{x\}$ is a closed class.

§4.2 Hitting Times

Definition 4.6

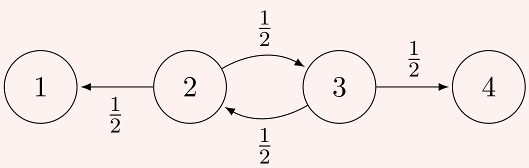
For $A \subseteq I$, we define T_A to be the **hitting time** of A, $T_A : \Omega \to \{0, 1, 2, \dots\} \cup \{\infty\}$, defined by $T_A(\omega) = \inf\{n \geq 0 : X_n(\omega) \in A\}$, where we take $\inf \emptyset = \infty$.

The **hitting probability** of A is $h^A : I \to [0, 1]$ such that $h_i^A = \mathbb{P}_i(T_A < \infty)$.

The **mean hitting time** of A is $k^A : I \to \mathbb{R}$ with $k_i^A = \mathbb{E}_i[T_A] = \sum_{n=0}^{\infty} n \cdot \mathbb{P}_i(T_a = n) + \infty \cdot \mathbb{P}_i(T_a = \infty)$.

Example 4.7

Consider the Markov chain in the diagram below.



We take $A = \{4\}$, and want to find $h_2^A = \mathbb{P}_2(T_A < \infty)$. We have

$$\begin{aligned} h_2^A &= \frac{1}{2} h_3^A \\ h_3^A &= \frac{1}{2} \cdot 1 + \frac{1}{2} h_2^A \\ \implies h_2^A &= \frac{1}{3}. \end{aligned}$$

If instead we took $B = \{1, 4\}$ and wanted to find k_2^B, we would get

$$\begin{aligned} k_2^B &= 1 + \frac{1}{2} k_3^B \\ k_3^B &= 1 + \frac{1}{2} k_2^B \\ \implies k_2^B &= 2. \end{aligned}$$

In the computations above, we really should check that this is a valid method (though it is quite intuitive).

Theorem 4.8

Let $A \subseteq I$. The vector $(h_i^A)_{i \in A}$ is the minimal non-negative solution to

$$h_i^A = \begin{cases} 1 & \text{if } i \in A, \\ \sum_j P(i,j) h_j^A & \text{if } i \notin A, \end{cases}$$

where minimality means that if $(x_i)_{i \in A}$ is another solution to the linear system, then $x_i \geq h_i^A$ for all i.

Proof. We first check that h_i does indeed solve this system. □