Finite Markov Chains and Algorithmic Applications

Olle Häggström

Matematisk statistik, Chalmers tekniska högskola och Göteborgs universitet

CAMBRIDGE
UNIVERSITY PRESS

PUBLISHED BY THE PRESS SYNDICATE OF THE UNIVERSITY OF CAMBRIDGE
The Pitt Building, Trumpington Street, Cambridge, United Kingdom

CAMBRIDGE UNIVERSITY PRESS
The Edinburgh Building, Cambridge CB2 2RU, UK
40 West 20th Street, New York, NY 10011-4211, USA
477 Williamstown Road, Port Melbourne, VIC 3207, Australia
Ruiz de Alarcón 13, 28014, Madrid, Spain
Dock House, The Waterfront, Cape Town 8001, South Africa

http://www.cambridge.org

First published 2002

Printed in the United Kingdom at the University Press, Cambridge

Typeface Times 10/13pt. *System* LaTeX 2_ε [DBD]

A catalogue record of this book is available from the British Library

ISBN 0 521 81357 3 hardback
ISBN 0 521 89001 2 paperback

Contents

1

Basics of probability theory

The majority of readers will probably be best off by taking the following piece of advice:

Skip this chapter!

Those readers who have previously taken a basic course in probability or mathematical statistics will already know everything in this chapter, and should move right on to Chapter 2. On the other hand, those readers who lack such background will have little or no use for the telegraphic exposition given here, and should instead consult some introductory text on probability. Rather than being read, the present chapter is intended to be a collection of (mostly) definitions, that can be consulted if anything that looks unfamiliar happens to appear in the coming chapters.

□ □ □ □

Let Ω be any set, and let Σ be some appropriate class of subsets of Ω, satisfying certain assumptions that we do not go further into (closedness under certain basic set operations). Elements of Σ are called **events**. For $A \subseteq \Omega$, we write A^c for the **complement** of A in Ω, meaning that

$$A^c = \{s \in \Omega : s \not\in A\} .$$

A **probability measure** on Ω is a function $\mathbf{P} : \Sigma \to [0, 1]$, satisfying

(i) $\mathbf{P}(\emptyset) = 0$.

(ii) $\mathbf{P}(A^c) = 1 - \mathbf{P}(A)$ for every event A.

(iii) If A and B are disjoint events (meaning that $A \cap B = \emptyset$), then $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B)$. More generally, if $A_1, A_2, \ldots$ is a countable sequence of disjoint events ($A_i \cap A_j = \emptyset$ for all $i \neq j$), then $\mathbf{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbf{P}(A_i)$.

Note that (i) and (ii) together imply that $\mathbf{P}(\Omega) = 1$.

If A and B are events, and $\mathbf{P}(B) > 0$, then we define the **conditional probability of** A **given** B, denoted $\mathbf{P}(A \mid B)$, as

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

The intuitive interpretation of $\mathbf{P}(A \mid B)$ is as how likely we consider the event A to be, given that we know that the event B has happened.

Two events A and B are said to be **independent** if $\mathbf{P}(A \cap B) = \mathbf{P}(A)\mathbf{P}(B)$. More generally, the events $A_1, \ldots, A_k$ are said to be independent if for any $l \leq k$ and any $i_1, \ldots, i_l \in \{1, \ldots, k\}$ with $i_1 < i_2 < \cdots < i_l$ we have

$$\mathbf{P}\left(A_{i_1} \cap A_{i_2} \cap \cdots \cap A_{i_l}\right) = \prod_{n=1}^{l} \mathbf{P}(A_{i_n}) .$$

For an infinite sequence of events $(A_1, A_2, \ldots)$, we say that $A_1, A_2, \ldots$ are independent if $A_1, \ldots, A_k$ are independent for any k.

Note that if $\mathbf{P}(B) > 0$, then independence between A and B is equivalent to having $\mathbf{P}(A \mid B) = \mathbf{P}(A)$, meaning intuitively that the occurrence of B does not affect the likelihood of A.

A **random variable** should be thought of as some random quantity which depends on chance. Usually a random variable is real-valued, in which case it is a function $X : \Omega \to \mathbf{R}$. We will, however, also consider random variables in a more general sense, allowing them to be functions $X : \Omega \to S$, where S can be any set.

An event A is said to be **defined in terms of the random variable** X if we can read off whether or not A has happened from the value of X. Examples of events defined in terms of the random variable X are

$$A = \{X \leq 4.7\} = \{\omega \in \Omega : \ X(\omega) \leq 4.7\}$$

and

$$B = \{X \text{ is an even integer}\} .$$

Two random variables are said to be independent if it is the case that whenever the event A is defined in terms of X, and the event B is defined in terms of Y, then A and B are independent. If $X_1, \ldots, X_k$ are random variables, then they are said to be independent if $A_1, \ldots, A_k$ are independent whenever each A_i is defined in terms of X_i. The extension to infinite sequences is similar: The random variables $X_1, X_2, \ldots$ are said to be independent if for any sequence

$A_1, A_2, \ldots$ of events such that for each i, A_i is defined in terms of X_i, we have that $A_1, A_2, \ldots$ are independent.

A distribution is the same thing as a probability measure. If X is a real-valued random variable, then the **distribution** μ_X of X is the probability measure on $\mathbf{R}$ satisfying $\mu_X(A) = \mathbf{P}(X \in A)$ for all (appropriate) $A \subseteq \mathbf{R}$. The distribution of a real-valued random variable is characterized in terms of its **distribution function** $F_X : \mathbf{R} \to [0, 1]$ defined by $F_X(x) = \mathbf{P}(X \leq x)$ for all $x \in \mathbf{R}$.

A distribution μ on a finite set $S = \{s_1, \ldots, s_k\}$ is often represented as a vector $(\mu_1, \ldots, \mu_k)$, where $\mu_i = \mu(s_i)$. By the definition of a probability measure, we then have that $\mu_i \in [0, 1]$ for each i, and that $\sum_{i=1}^{k} \mu_i = 1$.

A sequence of random variables $X_1, X_2, \ldots$ is said to be **i.i.d.**, which is short for **independent and identically distributed**, if the random variables

(i) are independent, and

(ii) have the same distribution function, i.e., $\mathbf{P}(X_i \leq x) = \mathbf{P}(X_j \leq x)$ for all i, j and x.

Very often, a sequence $(X_1, X_2, \ldots)$ is interpreted as the evolution in time of some random quantity: X_n is the quantity at time n. Such a sequence is then called a **random process** (or, sometimes, **stochastic process**). Markov chains, to be introduced in the next chapter, are a special class of random processes.

We shall only be dealing with two kinds of real-valued random variables: **discrete** and **continuous** random variables. The discrete ones take their values in some finite or countable subset of $\mathbf{R}$; in all our applications this subset is (or is contained in) $\{0, 1, 2, \ldots\}$, in which case we say that they are **nonnegative integer-valued** discrete random variables.

A **continuous** random variable X is a random variable for which there exists a so-called **density function** $f_X : \mathbf{R} \to [0, \infty)$ such that

$$\int_{-\infty}^{x} f_X(x)dx = F_X(x) = \mathbf{P}(X \leq x)$$

for all $x \in \mathbf{R}$. A very well-known example of a continuous random variable X arises by letting X have the Gaussian density function $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-((x-\mu)^2)/2\sigma^2}$ with parameters μ and $\sigma > 0$. However, the only continuous random variables that will be considered in this text are the **uniform** [0, 1] ones, which have density function

$$f_X(x) = \begin{cases} 1 & \text{if } x \in [0, 1] \\ 0 & \text{otherwise} \end{cases}$$

and distribution function

$$F_X(x) = \int_{-\infty}^{x} f_X(x)dx = \begin{cases} 0 & \text{if } x \leq 0 \\ x & \text{if } x \in [0,1] \\ 1 & \text{if } x \geq 1\,. \end{cases}$$

Intuitively, if X is a uniform $[0,1]$ random variable, then X is equally likely to take its value anywhere in the unit interval $[0,1]$. More precisely, for every interval I of length a inside $[0,1]$, we have $\mathbf{P}(X \in I) = a$.

The **expectation** (or **expected value**, or **mean**) $\mathbf{E}[X]$ of a real-valued random variable X is, in some sense, the "average" value we expect from x. If X is a continuous random variable with density function $f_X(x)$, then its expectation is defined as

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

which in the case where X is uniform $[0,1]$ reduces to

$$\mathbf{E}[X] = \int_0^1 x\,dx = \frac{1}{2}\,.$$

For the case where X is a nonnegative integer-valued random variable, the expectation is defined as

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} k\mathbf{P}(X = k)\,.$$

This can be shown to be equivalent to the alternative formula

$$\mathbf{E}[X] = \sum_{k=1}^{\infty} \mathbf{P}(X \geq k)\,. \tag{1}$$

It is important to understand that the expectation $\mathbf{E}[X]$ of a random variable can be infinite, even if X itself only takes finite values. A famous example is the following.

Example 1.1: The St Petersburg paradox. Consider the following game. A fair coin is tossed repeatedly until the first time that it comes up tails. Let X be the (random) number of heads that come up before the first occurrence of tails. Suppose that the bank pays 2^X roubles depending on X. How much would you be willing to pay to enter this game?

According to the classical theory of hazard games, you should agree to pay up to $\mathbf{E}[Y]$, where $Y = 2^X$ is the amount that you receive from the bank at the end of the game. So let's calculate $\mathbf{E}[Y]$. We have

$$\mathbf{P}(X = n) = \mathbf{P}(n \text{ heads followed by 1 tail}) = \left(\frac{1}{2}\right)^{n+1}$$

for each n, so that

$$\begin{aligned}\mathbf{E}[Y] &= \sum_{k=1}^{\infty} k\mathbf{P}(Y=k) = \sum_{n=0}^{\infty} 2^n \mathbf{P}(Y=2^n) \\ &= \sum_{n=0}^{\infty} 2^n \mathbf{P}(X=n) = \sum_{n=0}^{\infty} 2^n \left(\frac{1}{2}\right)^{n+1} \\ &= \sum_{n=0}^{\infty} \frac{1}{2} = \infty\,.\end{aligned}$$

Hence, there is obviously something wrong with the classical theory of hazard games in this case.

Another important characteristic, besides $\mathbf{E}[X]$, of a random variable X, is the **variance Var**$[X]$, defined by

$$\mathbf{Var}[X] = \mathbf{E}[(X-\mu)^2] \quad \text{where } \mu = \mathbf{E}[X]\,. \tag{2}$$

The variance is, thus, the mean square deviation of X from its expectation. It can be computed either using the defining formula (2), or by the identity

$$\mathbf{Var}[X] = \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \tag{3}$$

known as **Steiner's formula**.

There are various linear-like rules for working with expectations and variances. For expectations, we have

$$\mathbf{E}[X_1 + \cdots + X_n] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n] \tag{4}$$

and, if c is a constant,

$$\mathbf{E}[cX] = c\mathbf{E}[X]\,. \tag{5}$$

For variances, we have

$$\mathbf{Var}[cX] = c^2\mathbf{Var}[X] \tag{6}$$

and, *when $X_1, \ldots, X_n$ are independent,*[1]

$$\mathbf{Var}[X_1 + \cdots + X_n] = \mathbf{Var}[X_1] + \cdots + \mathbf{Var}[X_n]\,. \tag{7}$$

Let us compute expectations and variances in some simple cases.

Example 1.2 Fix $p \in [0, 1]$, and let

$$X = \begin{cases} 1 & \text{with probability } p \\ 0 & \text{with probability } 1-p\,. \end{cases}$$

[1] Without this requirement, (7) *fails* in general.

Such an X is called a **Bernoulli (p) random variable**. The expectation of X becomes $\mathbf{E}[X] = 0 \cdot \mathbf{P}(X = 0) + 1 \cdot \mathbf{P}(X = 1) = p$. Furthermore, since X only takes the values 0 and 1, we have $X^2 = X$, so that $\mathbf{E}[X^2] = \mathbf{E}[X]$, and

$$\begin{aligned} \mathbf{Var}[X] &= \mathbf{E}[X^2] - (\mathbf{E}[X])^2 \\ &= p - p^2 = p(1-p) \end{aligned}$$

using Steiner's formula (3).

Example 1.3 Let Y be the sum of n independent Bernoulli (p) random variables $X_1, \ldots, X_n$. (For instance, Y may be the number of heads in n tosses of a coin with heads-probability p.) Such a Y is said to be a **binomial (n, p) random variable**. Then, using (4) and (7), we get

$$\mathbf{E}[Y] = \mathbf{E}[X_1] + \cdots + \mathbf{E}[X_n] = np$$

and

$$\mathbf{Var}[Y] = \mathbf{Var}[X_1] + \cdots + \mathbf{Var}[X_n] = np(1-p)\,.$$

Variances are useful, e.g., for bounding the probability that a random variable deviates by a large amount from its mean. We have, for instance, the following well-known result.

Theorem 1.1 (Chebyshev's inequality) *Let X be a random variable with mean μ and variance σ^2. For any $a > 0$, we have that the probability $\mathbf{P}[|X - \mu| \geq a]$ of a deviation from the mean of at least a, satisfies*

$$\mathbf{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}\,.$$

Proof Define another random variable Y by setting

$$Y = \begin{cases} a^2 & \text{if } |X - \mu| \geq a \\ 0 & \text{otherwise.} \end{cases}$$

Then we always have $Y \leq (X - \mu)^2$, so that $\mathbf{E}[Y] \leq \mathbf{E}[(X - \mu)^2]$. Furthermore, $\mathbf{E}[Y] = a^2 \mathbf{P}(|X - \mu| \geq a)$, so that

$$\begin{aligned} \mathbf{P}(|X - \mu| \geq a) &= \frac{\mathbf{E}[Y]}{a^2} \\ &\leq \frac{\mathbf{E}[(X - \mu)^2]}{a^2} \\ &= \frac{\mathbf{Var}[X]}{a^2} = \frac{\sigma^2}{a^2}\,. \end{aligned}$$

□

Chebyshev's inequality will be used to prove a key result in Chapter 9 (Lemma 9.3). A more famous application of Chebyshev's inequality is in the proof of the following very famous and important result.

Theorem 1.2 (The Law of Large Numbers) *Let $X_1, X_2, \ldots$ be i.i.d. random variables with finite mean μ and finite variance σ^2. Let M_n denote the average of the first n X_i's, i.e., $M_n = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, for any $\varepsilon > 0$, we have*

$$\lim_{n\to\infty} \mathbf{P}(|M_n - \mu| \geq \varepsilon) = 0\,.$$

Proof Using (4) and (5) we get

$$\mathbf{E}[M_n] = \frac{1}{n}(\mu + \cdots + \mu) = \mu\,.$$

Similarly, (6) and (7) apply to show that

$$\mathbf{Var}[M_n] = \frac{1}{n^2}(\sigma^2 + \cdots + \sigma^2) = \frac{\sigma^2}{n}\,.$$

Hence, Chebyshev's inequality gives

$$\mathbf{P}(|M_n - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}$$

which tends to 0 as $n \to \infty$. □