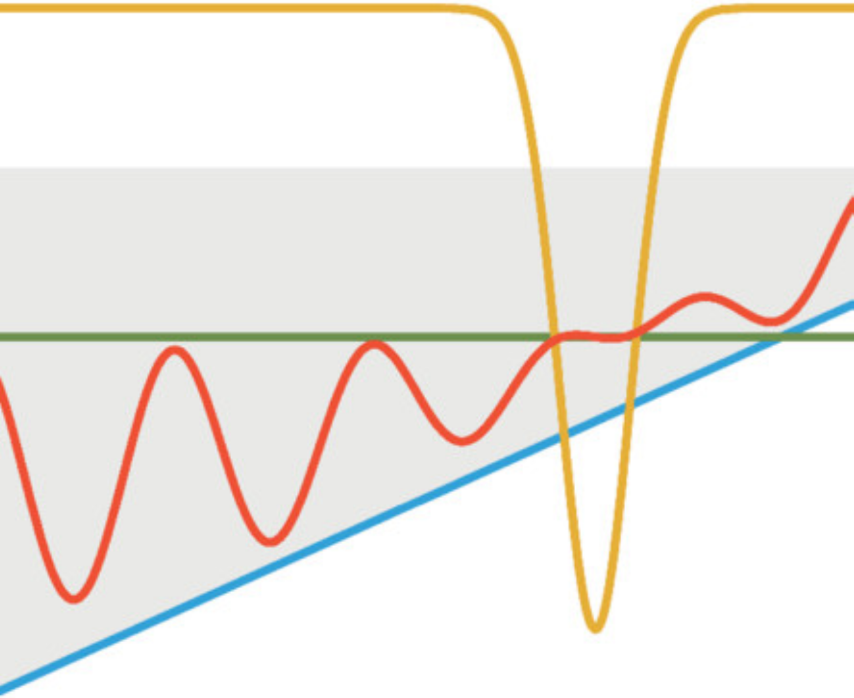
Bandit Algorithms

TOR LATTIMORE
CSABA SZEPESVÁRI



CAMBRIDGE
UNIVERSITY PRESS

University Printing House, Cambridge CB2 8BS, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

79 Anson Road, #06–04/06, Singapore 079906

Cambridge University Press is part of the University of Cambridge.

It furthers the University's mission by disseminating knowledge in the pursuit of education, learning, and research at the highest international levels of excellence.

www.cambridge.org

Information on this title: www.cambridge.org/9781108486828

DOI: 10.1017/9781108571401

First published 2020

Printed in the United Kingdom by TJ International, Padstow Cornwall

A catalogue record for this publication is available from the British Library.

Library of Congress Cataloging-in-Publication Data
Names: Lattimore, Tor, 1987- author. | Szepesvári, Csaba, author.
Title: Bandit algorithms / Tor Lattimore and Csaba Szepesvári.
Description: Cambridge ; New York, NY : Cambridge University Press, 2020. |
Includes bibliographical references and index.
Identifiers: LCCN 2019053276 (print) | LCCN 2019053277 (ebook) |
ISBN 9781108486828 (hardback) | ISBN 9781108571401 (epub)
Subjects: LCSH: Mathematical optimization. | Probabilities. |
Decision making–Mathematical models. | Resource allocation–Mathematical models. | Algorithms.
Classification: LCC QA402.5 .L367 2020 (print) | LCC QA402.5 (ebook) |
DDC 519.3–dc23
LC record available at https://lccn.loc.gov/2019053276
LC ebook record available at https://lccn.loc.gov/2019053277

ISBN 978-1-108-48682-8 Hardback

Contents

Preface

Multi-armed bandits have now been studied for nearly a century. While research in the beginning was quite meandering, there is now a large community publishing hundreds of articles every year. Bandit algorithms are also finding their way into practical applications in industry, especially in on-line platforms where data is readily available and automation is the only way to scale.

We had hoped to write a comprehensive book, but the literature is now so vast that many topics have been excluded. In the end we settled on the more modest goal of equipping our readers with enough expertise to explore the specialised literature by themselves, and to adapt existing algorithms to their applications. This latter point is important. Problems in theory are all alike; every application is different. A practitioner seeking to apply a bandit algorithm needs to understand which assumptions in the theory are important and how to modify the algorithm when the assumptions change. We hope this book can provide that understanding.

What is covered in the book is covered in some depth. The focus is on the mathematical analysis of algorithms for bandit problems, but this is not a traditional mathematics book, where lemmas are followed by proofs, theorems and more lemmas. We worked hard to include guiding principles for designing algorithms and intuition for their analysis. Many algorithms are accompanied by empirical demonstrations that further aid intuition.

We expect our readers to be familiar with basic analysis and calculus and some linear algebra. The book uses the notation of measure-theoretic probability theory, but does not rely on any deep results. A dedicated chapter is included to introduce the notation and provide intuitions for the basic results we need. This chapter is unusual for an introduction to measure theory in that it emphasises the reasons to use σ-algebras beyond the standard technical justifications. We hope this will convince the reader that measure theory is an important and intuitive tool. Some chapters use techniques from information theory and convex analysis, and we devote a short chapter to each.

Most chapters are short and should be readable in an afternoon or presented in a single lecture. Some components of the book contain content that is not really about bandits. These can be skipped by knowledgeable readers, or otherwise referred to when necessary. They are marked with a (🦘) because 'Skippy the Kangaroo' skips things.[1] The same mark is used for those parts that contain useful, but perhaps overly specific information for the first-time reader. Later parts will not build on these chapters in any substantial way. Most chapters end with a list of notes and exercises. These are intended to deepen intuition and highlight

[1] Taking inspiration from Tor's grandfather-in-law, John Dillon [Anderson et al., 1977].

the connections between various subsections and the literature. There is a table of notation at the end of this preface.

Thanks

We're indebted to our many collaborators and feel privileged that there are too many of you to name. The University of Alberta, Indiana University and DeepMind have all provided outstanding work environments and supported the completion of this book. The book has benefited enormously from the proofreading efforts of a large number of our friends and colleagues. We're sorry for all the mistakes introduced after your hard work. Alphabetically, they are: Aaditya Ramdas, Abbas Mehrabian, Aditya Gopalan, Ambuj Tewari, András György, Arnoud den Boer, Branislav Kveton, Brendan Patch, Chao Tao, Christoph Dann, Claire Vernade, Emilie Kaufmann, Eugene Ji, Gellért Weisz, Gergely Neu, Johannes Kirschner, Julian Zimmert, Kwang-Sung Jun, Lalit Jain, Laurent Orseau, Marcus Hutter, Michal Valko, Omar Rivasplata, Pierre Menard, Ramana Kumar, Roman Pogodin, Ronald Ortner, Ronan Fruit, Ruihao Zhu, Shuai Li, Toshiyuki Tanaka, Wei Chen, Yoan Russac, Yufei Yi and Zhu Xiaohu. We are especially grateful to Gábor Balázs and Wouter Koolen, who both read almost the entire book. Thanks to Lauren Cowles and Cambridge University Press for providing free books for our proofreaders, tolerating the delays and for supporting a freely available PDF version. Réka Szepesvári is responsible for converting some of our school figures to their current glory. Last of all, our families have endured endless weekends of editing and multiple false promises of 'done by Christmas'. Rosina and Beáta, it really is done now!

Notation

Some sections are marked with special symbols, which are listed and described below.

> 📝 This symbol is a note. Usually this is a remark that is slightly tangential to the topic at hand.

> ⚠ A warning to the reader.

> ☞ Something important.

> ⚗ An experiment.

Nomenclature and Conventions

A sequence $(a_n)_{n=1}^{\infty}$ is **increasing** if $a_{n+1} \geq a_n$ for all $n \geq 1$ and **decreasing** if $a_{n+1} \leq a_n$. When the inequalities are strict, we say **strictly increasing/decreasing**. The same terminology holds for functions. We will not be dogmatic about what is the range of $\operatorname{argmin}/\operatorname{argmax}$. Sometimes they return sets, sometimes arbitrary elements of those sets and, where stated, specific elements of those sets. We will be specific when it is non-obvious/matters. The infimum of the empty set is $\inf \emptyset = \infty$ and the supremum is $\sup \emptyset = -\infty$. The empty sum is $\sum_{i \in \emptyset} a_i = 0$ and the empty product is $\prod_{i \in \emptyset} a_i = 1$.

Landau Notation

We make frequent use of the Bachmann–Landau notation. Both were nineteenth century mathematicians who could have never expected their notation to be adopted so enthusiastically by computer scientists. Given functions $f, g : \mathbb{N} \to [0, \infty)$, define

$$f(n) = O(g(n)) \Leftrightarrow \limsup_{n \to \infty} \frac{f(n)}{g(n)} < \infty,$$

$$f(n) = o(g(n)) \Leftrightarrow \lim_{n \to \infty} \frac{f(n)}{g(n)} = 0,$$

$$
\begin{aligned}
f(n) = \Omega(g(n)) &\Leftrightarrow \liminf_{n\to\infty} \frac{f(n)}{g(n)} > 0,\\
f(n) = \omega(g(n)) &\Leftrightarrow \liminf_{n\to\infty} \frac{f(n)}{g(n)} = \infty,\\
f(n) = \Theta(g(n)) &\Leftrightarrow f(n) = O(g(n)) \text{ and } f(n) = \Omega(g(n)).
\end{aligned}
$$

We make use of the (Bachmann–)Landau notation in two contexts. First, in proofs where limiting arguments are made, we sometimes write lower-order terms using Landau notation. For example, we might write that $f(n) = \sqrt{n} + o(\sqrt{n})$, by which we mean that $\lim_{n\to\infty} f(n)/\sqrt{n} = 1$. In this case we use the mathematical definitions as envisaged by Bachmann and Landau. The second usage is to informally describe a result without the clutter of uninteresting constants. For better or worse, this usage is often a little imprecise. For example, we will often write expressions of the form: $R_n = O(m\sqrt{dn})$. Almost always what is meant by this is that there exists a **universal constant** $c > 0$ (a constant that does not depend on either of the quantities involved) such that $R_n \leq cm\sqrt{dn}$ for all (reasonable) choices of m, d and n. In this context we are careful *not* to use Landau notation to hide large lower-order terms. For example, if $f(x) = x^2 + 10^{100}x$, we will not write $f(x) = O(x^2)$, although this would be true.

Bandits

A_t	action in round t
k	number of arms/actions
n	time horizon
X_t	reward in round t
Y_t	loss in round t
π	a policy
ν	a bandit
μ_i	mean reward of arm i

Sets

$\emptyset$	empty set
$\mathbb{N}$, $\mathbb{N}^+$	natural numbers, $\mathbb{N} = \{0, 1, 2, \dots\}$ and $\mathbb{N}^+ = \mathbb{N} \setminus \{0\}$
$\mathbb{R}$	real numbers
$\bar{\mathbb{R}}$	$\mathbb{R} \cup \{-\infty, \infty\}$
$[n]$	$\{1, 2, 3, \dots, n-1, n\}$
2^A	the power set of set A (the set of all subsets of A)
A^*	set of finite sequences over A, $A^* = \bigcup_{i=0}^{\infty} A^i$
B_2^d	d-dimensional unit ball, $\{x \in \mathbb{R}^d : \\|x\\|_2 \leq 1\}$
$\mathcal{P}_d$	probability simplex, $\{x \in [0,1]^{d+1} : \\|x\\|_1 = 1\}$
$\mathcal{P}(A)$	set of distributions over set A
$\mathfrak{B}(A)$	Borel σ-algebra on A
$[x, y]$	convex hull of vectors or real values x and y

Functions, Operators and Operations

$\lvert A\rvert$	the cardinality (number of elements) of the finite set A
$(x)^+$	$\max(x, 0)$

$a \bmod b$	remainder when natural number a is divided by b
$\lfloor x \rfloor, \lceil x \rceil$	floor and ceiling functions of x
$\mathrm{dom}(f)$	domain of function f
$\mathbb{E}$	expectation
$\mathbb{V}$	variance
Supp	support of distribution or random variable
$\nabla f(x)$	gradient of f at x
$\nabla_v f(x)$	directional derivative of f at x in direction v
$\nabla^2 f(x)$	Hessian of f at x
$\vee, \wedge$	maximum and minimum, $a \vee b = \max(a, b)$ and $a \wedge b = \min(a, b)$
$\mathrm{erf}(x)$	$\frac{2}{\sqrt{\pi}} \int_0^x \exp(-y^2) dy$
$\mathrm{erfc}(x)$	$1 - \mathrm{erf}(x)$
$\Gamma(z)$	Gamma function, $\Gamma(z) = \int_0^\infty x^{z-1} \exp(-x) dx$
$\phi_A(x)$	support function $\phi_A(x) = \sup_{y \in A} \langle x, y \rangle$
$f^*(y)$	convex conjugate, $f^*(y) = \sup_{x \in A} \langle x, y \rangle - f(x)$
$\binom{n}{k}$	binomial coefficient
$\mathrm{argmax}_x\, f(x)$	maximiser or maximisers of f
$\mathrm{argmin}_x\, f(x)$	minimiser or minimisers of f
$\mathbb{I}\phi$	indicator function: converts Boolean ϕ into binary
$\mathbb{I}_B$	indicator of set B
$\mathrm{D}(P, Q)$	Relative entropy between probability distributions P and Q
$d(p, q)$	Relative entropy between $\mathcal{B}(p)$ and $\mathcal{B}(q)$

Linear Algebra

$e_1, \dots, e_d$	standard basis vectors of the d-dimensional Euclidean space
$\mathbf{0}, \mathbf{1}$	vectors whose elements are all zeros and all ones, respectively
$\det(A)$	determinant of matrix A
$\mathrm{trace}(A)$	trace of matrix A
$\mathrm{im}(A)$	image of matrix A
$\ker(A)$	kernel of matrix A
$\mathrm{span}(v_1, \dots, v_d)$	span of vectors $v_1, \dots, v_d$
$\lambda_{\min}(G)$	minimum eigenvalue of matrix G
$\langle x, y \rangle$	inner product, $\langle x, y \rangle = \sum_i x_i y_i$
$\lVert x \rVert_p$	p-norm of vector x
$\lVert x \rVert_G^2$	$x^\top G x$ for positive definite $G \in \mathbb{R}^{d \times d}$ and $x \in \mathbb{R}^d$
$\prec, \preceq$	Loewner partial order of positive semidefinite matrices: $A \preceq B$ ($A \prec B$) if $B - A$ is positive semidefinite (respectively, definite).

Distributions

$\mathcal{N}(\mu, \sigma^2)$	Normal distribution with mean μ and variance σ^2
$\mathcal{B}(p)$	Bernoulli distribution with mean p
$\mathcal{U}(a, b)$	uniform distribution supported on $[a, b]$
$\mathrm{Beta}(\alpha, \beta)$	Beta distribution with parameters $\alpha, \beta > 0$
δ_x	Dirac distribution with point mass at x

1 Introduction

Bandit problems were introduced by William R. Thompson in an article published in 1933 in *Biometrika*. Thompson was interested in medical trials and the cruelty of running a trial blindly, without adapting the treatment allocations on the fly as the drug appears more or less effective. The name comes from the 1950s, when Frederick Mosteller and Robert Bush decided to study animal learning and ran trials on mice and then on humans. The mice faced the dilemma of choosing to go left or right after starting in the bottom of a T-shaped maze, not knowing each time at which end they would find food. To study a similar learning setting in humans, a 'two-armed bandit' machine was commissioned where humans could choose to pull either the left or the right arm of the machine, each giving a random pay-off with the distribution of pay-offs for each arm unknown to the human player. The machine was called a 'two-armed bandit' in homage to the one-armed bandit, an old-fashioned name for a lever-operated slot machine ('bandit' because they steal your money).

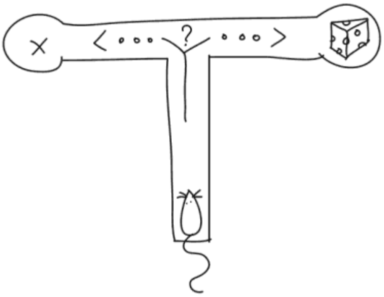


Figure 1.1 Mouse learning a T-maze.

There are many reasons to care about bandit problems. Decision-making with uncertainty is a challenge we all face, and bandits provide a simple model of this dilemma. Bandit problems also have practical applications. We already mentioned clinical trial design, which researchers have used to motivate their work for 80 years. We can't point to an example where bandits have actually been used in clinical trials, but adaptive experimental design is gaining popularity and is actively encouraged by the US Food and Drug Administration, with the justification that not doing so can lead to the withholding of effective drugs until long after a positive effect has been established.

While clinical trials are an important application for the future, there are applications where bandit algorithms are already in use. Major tech companies use bandit algorithms for configuring web interfaces, where applications include news recommendation, dynamic pricing and ad placement. A bandit algorithm plays a role in Monte Carlo Tree Search, an algorithm made famous by the recent success of AlphaGo.

Finally, the mathematical formulation of bandit problems leads to a rich structure with connections to other branches of mathematics. In writing this book (and previous papers), we have read books on convex analysis/optimisation, Brownian motion, probability theory, concentration analysis, statistics, differential geometry, information theory, Markov chains, computational complexity and more. What fun!

A combination of all these factors has led to an enormous growth in research over the last two decades. Google Scholar reports less than 1000, then 2700 and 7000 papers when searching for the phrase 'bandit algorithm' for the periods of 2001–5, 2006–10, and 2011–15, respectively, and the trend just seems to have strengthened since then, with 5600 papers coming up for the period of 2016 to the middle of 2018. Even if these numbers are somewhat overblown, they are indicative of a rapidly growing field. This could be a fashion, or maybe there is something interesting happening here. We think that the latter is true.

A Classical Dilemma

Imagine you are playing a two-armed bandit machine and you already pulled each lever five times, resulting in the following pay-offs (in dollars):

ROUND	1	2	3	4	5	6	7	8	9	10
LEFT	0		10	0		0				10
RIGHT		10			0		0	0	0	

Figure 1.2 Two-armed bandit

The left arm appears to be doing slightly better. The average pay-off for this arm is \$4, while the average for the right arm is only \$2. Let's say you have 10 more trials (pulls) altogether. What is your strategy? Will you keep pulling the left arm, ignoring the right? Or would you attribute the poor performance of the right arm to bad luck and try it a few more times? How many more times? This illustrates one of the main interests in bandit problems. They capture the fundamental dilemma a learner faces when choosing between uncertain options. Should one explore an option that looks inferior or exploit by going with the option that looks best currently? Finding the right balance between exploration and exploitation is at the heart of all bandit problems.

1.1 The Language of Bandits

A bandit problem is a sequential game between a **learner** and an **environment**. The game is played over n rounds, where n is a positive natural number called the **horizon**. In each round $t \in [n]$, the learner first chooses an action A_t from a given set $\mathcal{A}$, and the environment then reveals a reward $X_t \in \mathbb{R}$.

In the literature, actions are often also called 'arms'. We talk about **k-armed bandits** when the number of actions is k, and about **multi-armed bandits** when the number of arms is at least two and the actual number is immaterial to the discussion. If there are multi-armed bandits, there are also **one-armed bandits**, which are really two-armed bandits where the pay-off of one of the arms is a known fixed deterministic number.

Of course the learner cannot peek into the future when choosing their actions, which means that A_t should only depend on the **history** $H_{t-1} = (A_1, X_1, \ldots, A_{t-1}, X_{t-1})$. A **policy** is a mapping from histories to actions: A learner adopts a policy to interact with an environment. An environment is a mapping from history sequences ending in actions to rewards. Both the learner and the environment may randomise their decisions, but this detail is not so important for now. The most common objective of the learner is to choose actions that lead to the largest possible cumulative reward over all n rounds, which is $\sum_{t=1}^n X_t$.

The fundamental challenge in bandit problems is that the environment is unknown to the learner. All the learner knows is that the true environment lies in some set $\mathcal{E}$ called the **environment class**. Most of this book is about designing policies for different kinds of environment classes, though in some cases the framework is extended to include side observations as well as actions and rewards.

The next question is how to evaluate a learner. We discuss several performance measures throughout the book, but most of our efforts are devoted to understanding the **regret**. There are several ways to define this quantity. To avoid getting bogged down in details, we start with a somewhat informal definition.

Definition 1.1. The regret of the learner relative to a policy π (not necessarily that followed by the learner) is the difference between the total expected reward using policy π for n rounds and the total expected reward collected by the learner over n rounds. The regret relative to a set of policies Π is the maximum regret relative to any policy $\pi \in \Pi$ in the set.

The set Π is often called the **competitor class**. Another way of saying all this is that the regret measures the performance of the learner relative to the best policy in the competitor class. We usually measure the regret relative to a set of policies Π that is large enough to include the optimal policy for all environments in $\mathcal{E}$. In this case, the regret measures the loss suffered by the learner relative to the optimal policy.

Example 1.2. Suppose the action set is $\mathcal{A} = \{1, 2, \ldots, k\}$. An environment is called a **stochastic Bernoulli bandit** if the reward $X_t \in \{0, 1\}$ is binary valued and there exists a vector $\mu \in [0, 1]^k$ such that the probability that $X_t = 1$ given the learner chose action $A_t = a$ is μ_a. The class of stochastic Bernoulli bandits is the set of all such bandits, which are characterised by their mean vectors. If you knew the mean vector associated with the environment, then the optimal policy is to play the fixed action $a^* = \operatorname{argmax}_{a \in \mathcal{A}} \mu_a$. This means that for this problem the natural competitor class is the set of k constant polices $\Pi = \{\pi_1, \ldots, \pi_k\}$, where π_i chooses action i in every round. The regret over n rounds becomes

$$R_n = n \max_{a \in \mathcal{A}} \mu_a - \mathbb{E}\left[\sum_{t=1}^n X_t\right],$$

where the expectation is with respect to the randomness in the environment and policy. The first term in this expression is the maximum expected reward using any policy. The second term is the expected reward collected by the learner.

For a fixed policy and competitor class, the regret depends on the environment. The environments where the regret is large are those where the learner is behaving worse. Of course the ideal case is that the regret be small for all environments. The **worst-case regret** is the maximum regret over all possible environments.

One of the core questions in the study of bandits is to understand the growth rate of the regret as n grows. A good learner achieves sublinear regret. Letting R_n denote the regret over n rounds, this means that $R_n = o(n)$ or equivalently that $\lim_{n\to\infty} R_n/n = 0$. Of course one can ask for more. Under what circumstances is $R_n = O(\sqrt{n})$ or $R_n = O(\log(n))$? And what are the leading constants? How does the regret depend on the specific environment in which the learner finds itself? We will discover eventually that for the environment class in Example 1.2, the worst-case regret for any policy is at least $\Omega(\sqrt{n})$ and that there exist policies for which $R_n = O(\sqrt{n})$.

A large environment class corresponds to less knowledge by the learner. A large competitor class means the regret is a more demanding criteria. Some care is sometimes required to choose these sets appropriately so that *(a)* guarantees on the regret are meaningful and *(b)* there exist policies that make the regret small.

The framework is general enough to model almost anything by using a rich enough environment class. This cannot be bad, but with too much generality it becomes impossible to say much. For this reason, we usually restrict our attention to certain kinds of environment classes and competitor classes.

A simple problem setting is that of **stochastic stationary bandits**. In this case the environment is restricted to generate the reward in response to each action from a distribution that is specific to that action and independent of the previous action choices and rewards. The environment class in Example 1.2 satisfies these conditions, but there are many alternatives. For example, the rewards could follow a Gaussian distribution rather than Bernoulli. This relatively mild difference does not change the nature of the problem in a significant way. A more drastic change is to assume the action set $\mathcal{A}$ is a subset of $\mathbb{R}^d$ and that the mean reward for choosing some action $a \in \mathcal{A}$ follows a linear model, $X_t = \langle a, \theta \rangle + \eta_t$ for $\theta \in \mathbb{R}^d$ and η_t a standard Gaussian (zero mean, unit variance). The unknown quantity in this case is θ, and the environment class corresponds to its possible values ($\mathcal{E} = \mathbb{R}^d$).

For some applications, the assumption that the rewards are stochastic and stationary may be too restrictive. The world mostly appears deterministic, even if it is hard to predict and often chaotic looking. Of course, stochasticity has been enormously successful in explaining patterns in data, and this may be sufficient reason to keep it as the modelling assumption. But what if the stochastic assumptions fail to hold? What if they are violated for a single round? Or just for one action, at some rounds? Will our best algorithms suddenly perform poorly? Or will the algorithms developed be robust to smaller or larger deviations from the modelling assumptions?

An extreme idea is to drop all assumptions on how the rewards are generated, except that they are chosen without knowledge of the learner's actions and lie in a bounded set. If these are the only assumptions, we get what is called the setting of **adversarial bandits**. The trick to say something meaningful in this setting is to restrict the competitor class. The learner is not expected to find the best sequence of actions, which may be like finding a needle in a haystack. Instead, we usually choose Π to be the set of constant policies and demand it positively. There are many challenges. First of all, Netflix shows a long list of movies, so the set of possible actions is combinatorially large. Second, each user watches relatively few movies, and individual users are different. This suggests approaches such as low-rank matrix factorisation (a popular approach in 'collaborative filtering'). But notice this is not an offline problem. The learning algorithm gets to choose what users see and this affects the data. If the users are never recommended the AlphaGo movie, then few users will watch it, and the amount of data about this film will be scarce.

Network Routing

Another problem with an interesting structure is network routing, where the learner tries to direct internet traffic through the shortest path on a network. In each round the learner receives the start/end destinations for a packet of data. The set of actions is the set of all paths starting and ending at the appropriate points on some known graph. The feedback in this case is the time it takes for the packet to be received at its destination, and the reward is the negation of this value. Again the action set is combinatorially large. Even relatively small graphs have an enormous number of paths. The routing problem can obviously be applied to more physical networks such as transportation systems used in operations research.

Dynamic Pricing

In dynamic pricing, a company is trying to automatically optimise the price of some product. Users arrive sequentially, and the learner sets the price. The user will only purchase the product if the price is lower than their valuation. What makes this problem interesting is (a) the learner never actually observes the valuation of the product, only the binary signal that the price was too low/too high, and (b) there is a monotonicity structure in the pricing. If a user purchased an item priced at $10, then they would surely purchase it for $5, but whether or not it would sell when priced at $11 is uncertain. Also, the set of possible actions is close to continuous.

Waiting Problems

Every day you travel to work, either by bus or by walking. Once you get on the bus, the trip only takes 5 minutes, but the timetable is unreliable, and the bus arrival time is unknown and stochastic. Sometimes the bus doesn't come at all. Walking, on the other hand, takes 30 minutes along a beautiful river away from the road. The problem is to devise a policy for choosing how long to wait at the bus stop before giving up and walking to minimise the time to get to your workplace. Walk too soon, and you miss the bus and gain little information. But waiting too long also comes at a price.

While waiting for a bus is not a problem we all face, there are other applications of this setting. For example, deciding the amount of inactivity required before putting a hard drive into sleep mode or powering off a car engine at traffic lights. The statistical part of the waiting problem concerns estimating the cumulative distribution function of the bus arrival times from data. The twist is that the data is censored on the days you chose to walk before the bus arrived, which is a problem analysed in the subfield of statistics called survival analysis. The interplay between the statistical estimation problem and the challenge of balancing exploration and exploitation is what makes this and the other problems studied in this book interesting.

Resource Allocation

A large part of operations research is focussed on designing strategies for allocating scarce resources. When the dynamics of demand or supply are uncertain, the problem has elements reminiscent of a bandit problem. Allocating too few resources reveals only partial information about the true demand, but allocating too many resources is wasteful. Of course, resource allocation is broad, and many problems exhibit structure that is not typical of bandit problems, like the need for long-term planning.

Tree Search

The UCT algorithm is a tree search algorithm commonly used in perfect-information game-playing algorithms. The idea is to iteratively build a search tree where in each iteration the algorithm takes three steps: *(1)* chooses a path from the root to a leaf; *(2)* expands the leaf (if possible); *(3)* performs a Monte Carlo roll-out to the end of the game. The contribution of a bandit algorithm is in selecting the path from the root to the leaves. At each node in the tree, a bandit algorithm is used to select the child based on the series of rewards observed through that node so far. The resulting algorithm can be analysed theoretically, but more importantly has demonstrated outstanding empirical performance in game-playing problems.

1.3 Notes

1 The reader may find it odd that at one point we identified environments with maps from histories to rewards, while we used the language that a learner ‘adopts a policy’ (a map from histories to actions). The reason is part historical and part because policies and their design are at the center of the book, while the environment strategies will mostly be kept fixed (and relatively simple). On this note, strategy is also a word that sometimes used interchangeably with policy.

1.4 Bibliographic Remarks

As we mentioned in the very beginning, the first paper on bandits was by Thompson [1933]. The experimentation on mice and humans that led to the name comes from the paper by Bush and Mosteller [1953]. Much credit for the popularisation of the field must go to famous mathematician and statistician, Herbert Robbins, whose name appears on many of the works that we reference, with the earliest being: [Robbins, 1952]. Another early pioneer is Herman Chernoff, who wrote papers with titles like ‘Sequential Decisions in the Control of a Spaceship’ [Bather and Chernoff, 1967].

Besides these seminal papers, there are already a number of books on bandits that may serve as useful additional reading. The most recent (and also most related) is by Bubeck and Cesa-Bianchi [2012] and is freely available online. This is an excellent book and is warmly recommended. The main difference between their book and ours is that *(a)* we have the benefit of seven years of additional research in a fast-moving field and *(b)* our longer page limit permits more depth. Another relatively recent book is *Prediction, Learning and Games* by Cesa-Bianchi and Lugosi [2006]. This is a wonderful book, and quite comprehensive. But its scope is ‘all of’ online learning, which is so broad that bandits are not covered in great depth. We should mention there is also a recent book on bandits by Slivkins [2019]. Conveniently it covers some topics not covered in this book (notably Lipschitz bandits and bandits with knapsacks). The reverse is also true, which should not be surprising since our book is currently 400 pages longer. There are also four books on sequential design and multi-armed bandits in the Bayesian setting, which we will address only a little. These are based on relatively old material, but are still useful references for this line of work and are well worth reading [Chernoff, 1959, Berry and Fristedt, 1985, Presman and Sonin, 1990, Gittins et al., 2011].

Without trying to be exhaustive, here are a few articles applying bandit algorithms; a recent survey is by Bouneffouf and Rish [2019]. The papers themselves will contain more useful pointers to the vast literature. We mentioned AlphaGo already [Silver et al., 2016]. The tree search algorithm that drives its search uses a bandit algorithm at each node [Kocsis and Szepesvári, 2006]. Le et al. [2014] apply bandits to wireless monitoring, where the problem is challenging due to the large action space. Lei et al. [2017] design specialised contextual bandit algorithms for just-in-time adaptive interventions in mobile health: in the typical application the user is prompted with the intention of inducing a long-term beneficial behavioural change. See also the article by Greenewald et al. [2017]. Rafferty et al. [2018] apply Thompson sampling to educational software and note the trade-off between knowledge and reward. Sadly, by 2015, bandit algorithms still have not been used in clinical trials, as explicitly mentioned by Villar et al. [2015]. Microsoft offers a 'Decision Service' that uses bandit algorithms to automate decision-making [Agarwal et al., 2016].

2 Foundations of Probability (🦘)

This chapter covers the fundamental concepts of measure-theoretic probability, on which the remainder of this book relies. Readers familiar with this topic can safely skip the chapter, but perhaps a brief reading would yield some refreshing perspectives. Measure-theoretic probability is often viewed as a necessary evil, to be used when a demand for rigour combined with continuous spaces breaks the simple approach we know and love from high school. We claim that measure-theoretic probability offers more than annoying technical machinery. In this chapter we attempt to prove this by providing a non-standard introduction. Rather than a long list of definitions, we demonstrate the intuitive power of the notation and tools. For those readers with little prior experience in measure theory this chapter will no doubt be a challenging read. We think the investment is worth the effort, but a great deal of the book can be read without it, provided one is willing to take certain results on faith.

2.1 Probability Spaces and Random Elements

The thrill of gambling comes from the fact that the bet is placed on future outcomes that are uncertain at the time of the gamble. A central question in gambling is the fair value of a game. This can be difficult to answer for all but the simplest games. As an illustrative example, imagine the following moderately complex game: I throw a dice. If the result is four, I throw two more dice; otherwise I throw one dice only. Looking at each newly thrown dice (one or two), I repeat the same, for a total of three rounds. Afterwards, I pay you the sum of the values on the faces of the dice. How much are you willing to pay to play this game with me?

Many examples of practical interest exhibit a complex random interdependency between outcomes. The cornerstone of modern probability as proposed by Kolmogorov aims to remove this complexity by separating the randomness from the mechanism that produces the outcome.

Instead of rolling the dice one by one, imagine that sufficiently many dice were rolled before the game has even started. For our game we need to roll seven dice, because this is the maximum number that might be required (one in the first round, two in the second round and four in the third round. See Fig. 2.1). After all the dice are rolled, the game can be emulated by ordering the dice and revealing the outcomes sequentially. Then the value of the first dice in the chosen ordering is the outcome of the dice in the first round. If we see a four, we look at the next two dice in the ordering; otherwise we look at the single next dice.

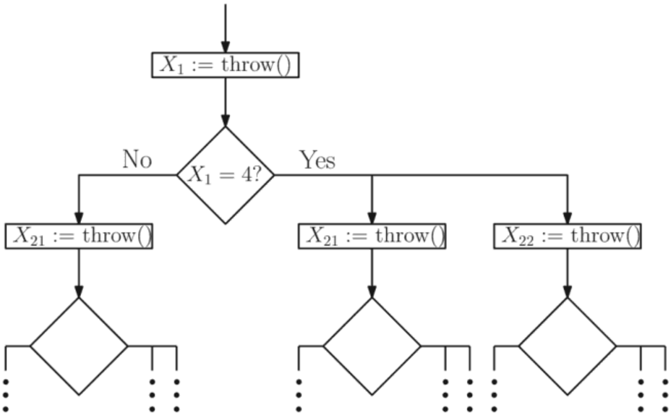

$X_1 := \text{throw}()$
No
$X_1 = 4?$
Yes
$X_{21} := \text{throw}()$
$X_{21} := \text{throw}()$
$X_{22} := \text{throw}()$

Figure 2.1 The initial phase of a gambling game with a random number of dice rolls. Depending on the outcome of a dice roll, one or two dice are rolled for a total of three rounds. The number of dice used will then be random in the range of three to seven.

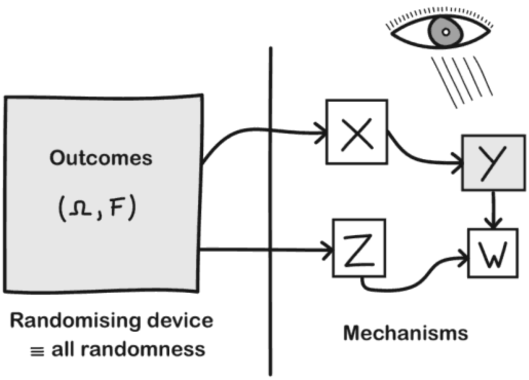


Figure 2.2 A key idea in probability theory is the separation of sources of randomness from game mechanisms. A mechanism creates values from the elementary random outcomes, some of which are visible for observers, while others may remain hidden.

By taking this approach, we get a simple calculus for the probabilities of all kinds of **events**. Rather than directly calculating the likelihood of each pay-off, we first consider the probability of any single outcome of the dice. Since there are seven dice, the set of all possible outcomes is $\Omega = \{1, \ldots, 6\}^7$. Because all outcomes are equally probable, the probability of any $\omega \in \Omega$ is $(1/6)^7$. The probability of the game pay-off taking value v can then be evaluated by calculating the total probability assigned to all those outcomes $\omega \in \Omega$ that would result in the value of v. In principle, this is trivial to do thanks to the separation of everything that is probabilistic from the rest. The set Ω is called the **outcome space**, and its elements are the **outcomes**. Fig. 2.2 illustrates this idea. Random outcomes are generated on the left, while on the right, various mechanisms are used to arrive at values; some of these values may be observed and some not.

There will be much benefit from being a little more formal about how we come up with the value of our artificial game. For this, note that the process by which the game gets its sequences. Let $(\Omega, \mathcal{F})$ be a measurable space, $\mathcal{X}$ be an arbitrary set and $\mathcal{G} \subseteq 2^{\mathcal{X}}$. A function $X : \Omega \to \mathcal{X}$ is called an **$\mathcal{F}/\mathcal{G}$-measurable map** if $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{G}$. Note that $\mathcal{G}$ need not be a σ-algebra. When $\mathcal{F}$ and $\mathcal{G}$ are obvious from the context, X is called a **measurable map**. What are the typical choices for $\mathcal{G}$? When X is real-valued, it is usual to let $\mathcal{G} = \{(a, b) : a < b$ with $a, b \in \mathbb{R}\}$ be the set of all open intervals. The reader can verify that if X is $\mathcal{F}/\mathcal{G}$-measurable, then it is also $\mathcal{F}/\sigma(\mathcal{G})$-measurable, where $\sigma(\mathcal{G})$ is the smallest σ-algebra that contains $\mathcal{G}$. This smallest σ-algebra can be shown to exist. Furthermore, it contains exactly those sets A that are in every σ-algebra that contains $\mathcal{G}$ (see Exercise 2.5). When $\mathcal{G}$ is the set of open intervals, $\sigma(\mathcal{G})$ is usually denoted by $\mathfrak{B}$ or $\mathfrak{B}(\mathbb{R})$ and is called the **Borel σ-algebra** of $\mathbb{R}$. This definition is extended to $\mathbb{R}^k$ by replacing open intervals with open rectangles of the form $\prod_{i=1}^k (a_i, b_i)$, where $a < b \in \mathbb{R}^k$. If $\mathcal{G}$ is the set of all such open rectangles, then $\sigma(\mathcal{G})$ is the Borel σ-algebra: $\mathfrak{B}(\mathbb{R}^k)$. More generally, the Borel σ-algebra of a topological space $\mathcal{X}$ is the σ-algebra generated by the open sets of $\mathcal{X}$.

Definition 2.2 (Random variables and elements). A **random variable** (**random vector**) on measurable space $(\Omega, \mathcal{F})$ is a $\mathcal{F}/\mathfrak{B}(\mathbb{R})$-measurable function $X : \Omega \to \mathbb{R}$ (respectively $\mathcal{F}/\mathfrak{B}(\mathbb{R}^k)$-measurable function $X : \Omega \to \mathbb{R}^k$). A **random element** between measurable spaces $(\Omega, \mathcal{F})$ and $(\mathcal{X}, \mathcal{G})$ is a $\mathcal{F}/\mathcal{G}$-measurable function $X : \Omega \to \mathcal{X}$.

Thus, random vectors are random elements where the range space is $(\mathbb{R}^k, \mathfrak{B}(\mathbb{R}^k))$, and random vectors are random variables when $k = 1$. Random elements generalise random variables and vectors to functions that do not take values in $\mathbb{R}^k$. The push-forward measure (or law) can be defined for any random element. Furthermore, random variables and vectors work nicely together. If $X_1, \dots, X_k$ are k random variables on the same domain $(\Omega, \mathcal{F})$, then $X(\omega) = (X_1(\omega), \dots, X_k(\omega))$ is an $\mathbb{R}^k$-valued random vector, and vice versa (Exercise 2.2). Multiple random variables $X_1, \dots, X_k$ from the same measurable space can thus be viewed as a random vector $X = (X_1, \dots, X_k)$.

Given a map $X : \Omega \to \mathcal{X}$ between measurable spaces $(\Omega, \mathcal{F})$ and $(\mathcal{X}, \mathcal{G})$, we let $\sigma(X) = \{X^{-1}(A) : A \in \mathcal{G}\}$ be the **σ-algebra generated by X**. The map X is $\mathcal{F}/\mathcal{G}$-measurable if and only if $\sigma(X) \subseteq \mathcal{F}$. By checking the definitions one can show that $\sigma(X)$ is a sub-σ-algebra of $\mathcal{F}$ and in fact is the smallest sub-σ-algebra for which X is measurable. If $\mathcal{G} = \sigma(\mathcal{A})$ itself is generated by a set system $\mathcal{A} \subset 2^{\mathcal{X}}$, then to check the $\mathcal{F}/\mathcal{G}$-measurability of X, it suffices to check whether $X^{-1}(\mathcal{A}) = \{X^{-1}(A) : A \in \mathcal{A}\}$ is a subset of $\mathcal{F}$. The reason this is sufficient is because $\sigma(X^{-1}(\mathcal{A})) = X^{-1}(\sigma(\mathcal{A}))$, and by definition the latter is $\sigma(X)$. In fact, to check whether a map is measurable, either one uses the composition rule or checks $X^{-1}(\mathcal{A}) \subset \mathcal{F}$ for a 'generator' $\mathcal{A}$ of $\mathcal{G}$.

Random elements can be combined to produce new random elements by composition. One can show that if f is $\mathcal{F}/\mathcal{G}$-measurable and g is $\mathcal{G}/\mathcal{H}$-measurable for σ-algebras $\mathcal{F}, \mathcal{G}$ and $\mathcal{H}$ over appropriate spaces, then their composition $g \circ f$ is $\mathcal{F}/\mathcal{H}$-measurable (Exercise 2.1). This is used most often for **Borel functions**, which is a special name for $\mathfrak{B}(\mathbb{R}^m)/\mathfrak{B}(\mathbb{R}^n)$-measurable functions from $\mathbb{R}^m$ to $\mathbb{R}^n$. These functions are also called **Borel measurable**. The reader will find it pleasing that all familiar functions are Borel. First and foremost, all continuous functions are Borel, which includes elementary operations such as addition and multiplication. Continuity is far from essential, however. In fact one is hard-pressed to construct a function that is not Borel. This means the usual operations are 'safe' when working with random variables.

Indicator Functions

Given an arbitrary set Ω and $A \subseteq \Omega$, the **indicator function** of A is $\mathbb{I}_A : \Omega \to \{0, 1\}$ given by

$$\mathbb{I}_A(\omega) = \begin{cases} 1, & \text{if } \omega \in A; \\ 0, & \text{otherwise.} \end{cases}$$

Sometimes A has a complicated description, and it becomes convenient to abuse notation by writing $\mathbb{I}\{\omega \in A\}$ instead of $\mathbb{I}_A(\omega)$. Similarly, we will often write $\mathbb{I}\{predicate(X, Y, \ldots)\}$ to mean the indicator function of the subset of Ω on which the predicate is true. It is easy to check that an indicator function $\mathbb{I}_A$ is a random variable on $(\Omega, \mathcal{F})$ if and only if A is measurable: $A \in \mathcal{F}$.

Why So Complicated?

You may be wondering why we did not define $\mathbb{P}$ on the power set of Ω, which is equivalent to declaring that all sets are measurable. In many cases this is a perfectly reasonable thing to do, including the example game where nothing prevents us from defining $\mathcal{F} = 2^\Omega$. However, beyond this example, there are two justifications not to have $\mathcal{F} = 2^\Omega$, the first technical and the second conceptual.

The technical reason is highlighted by the following surprising theorem according to which there does not exist a uniform probability distribution on $\Omega = [0, 1]$ if $\mathcal{F}$ is chosen to be the power set of Ω (a uniform probability distribution over $[0, 1]$, if existed, would have the property of assigning its length to every interval). In other words, if you want to be able to define the uniform measure, then $\mathcal{F}$ cannot be too large. By contrast, the uniform measure can be defined over the Borel σ-algebra, though proving this is not elementary.

THEOREM 2.3. *Let $\Omega = [0, 1]$, and $\mathcal{F}$ be the power set of Ω. Then there does not exist a measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that $\mathbb{P}([a, b]) = b - a$ for all $0 \leq a \leq b \leq 1$.*

The main conceptual reason of why not to have $\mathcal{F} = 2^\Omega$ is because then we can use σ-algebras represent information. This is especially useful in the study of bandits where the learner is interacting with an environment and is slowly gaining knowledge. One useful way to represent this is by using a sequence of nested σ-algebras, as we explain in the next section. One might also be worried that the Borel σ-algebra does not contain enough measurable sets. Rest assured that this is not a problem and you will not easily find a non-measurable set. For completeness, an example of a non-measurable set will still be given in the notes, along with a little more discussion on this topic.

A second technical reason to prefer the measure-theoretic approach to probabilities is that this approach allows for the unification of distributions on discrete spaces and densities on continuous ones (the uninitiated reader will find the definitions of these later). This unification can be necessary when dealing with random variables that combine elements of both, e.g. a random variable that is zero with probability $1/2$ and otherwise behaves like a standard Gaussian. Random variables like this give rise to so-called “mixed continuous and discrete distributions”, which seem to require special treatment in a naive approach to probabilities, yet dealing with random variables like these are nothing but ordinary under the measure-theoretic approach.

From Laws to Probability Spaces and Random Variables

A big ‘conspiracy’ in probability theory is that probability spaces are seldom mentioned in theorem statements, despite the fact that a measure cannot be defined without one. Statements are instead given in terms of random elements and constraints on their joint probabilities. For example, suppose that X and Y are random variables such that

$$\mathbb{P}(X \in A, Y \in B) = \frac{|A \cap [6]|}{6} \cdot \frac{|B \cap [2]|}{2} \qquad \text{for all } A, B \in \mathfrak{B}(\mathbb{R}), \tag{2.1}$$

which represents the joint distribution for the values of a dice ($X \in [6]$) and coin ($Y \in [2]$). The formula describes some constraints on the probabilistic interactions between the outputs of X and Y, but says nothing about their domain. In a way, the domain is an unimportant detail. Nevertheless, one *must* ask whether or not an appropriate domain exists at all. More generally, one may ask whether an appropriate probability space exists given some constraints on the joint law of a collection $X_1, \dots, X_k$ of random variables. For this to make sense, the constraints should not contradict each other, which means there is a probability measure μ on $\mathfrak{B}(\mathbb{R}^k)$ such that μ satisfies the postulated constraints. But then we can choose $\Omega = \mathbb{R}^k$, $\mathcal{F} = \mathfrak{B}(\mathbb{R}^k)$, $\mathbb{P} = \mu$ and $X_i : \Omega \to \mathbb{R}$ to be the ith coordinate map: $X_i(\omega) = \omega_i$. The push-forward of $\mathbb{P}$ under $X = (X_1, \dots, X_k)$ is μ, which by definition is compatible with the constraints.

A more specific question is whether for a particular set of constraints on the joint law there exists a measure μ compatible with the constraints. Very often the constraints are specified for elements of the cartesian product of finitely many σ-algebras, like in Eq. (2.1). If $(\Omega_1, \mathcal{F}_1), \dots, (\Omega_n, \mathcal{F}_n)$ are measurable spaces, then the cartesian product of $\mathcal{F}_1, \dots \mathcal{F}_n$ is

$$\mathcal{F}_1 \times \cdots \times \mathcal{F}_n = \{A_1 \times \cdots \times A_n : A_1 \in \mathcal{F}_1, \dots, A_n \in \mathcal{F}_n\} \subseteq 2^{\Omega_1 \times \cdots \times \Omega_n}.$$

Elements of this set are known as **measurable rectangles** in $\Omega_1 \times \cdots \times \Omega_n$.

Theorem 2.4 (Carathéodory’s extension theorem). *Let $(\Omega_1, \mathcal{F}_1), \dots, (\Omega_n, \mathcal{F}_n)$ be measurable spaces and $\bar{\mu} : \mathcal{F}_1 \times \cdots \times \mathcal{F}_n \to [0, 1]$ be a function such that*

(a) $\bar{\mu}(\Omega_1 \times \cdots \times \Omega_n) = 1$*; and*
(b) $\bar{\mu}(\cup_{k=1}^{\infty} A_k) = \sum_{k=1}^{\infty} \bar{\mu}(A_k)$ *for all sequences of disjoint sets with* $A_k \in \mathcal{F}_1 \times \cdots \times \mathcal{F}_n$.

Let $\Omega = \Omega_1 \times \cdots \times \Omega_n$ and $\mathcal{F} = \sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$. Then there exists a unique probability measure μ on $(\Omega, \mathcal{F})$ such that μ agrees with $\bar{\mu}$ on $\mathcal{F}_1 \times \cdots \times \mathcal{F}_n$.

The theorem is applied by letting $\Omega_k = \mathbb{R}$ and $\mathcal{F}_k = \mathfrak{B}(\mathbb{R})$. Then the values of a measure on all cartesian products uniquely determines its value everywhere.

> It is not true that $\mathcal{F}_1 \times \mathcal{F}_2 = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$. Take, for example, $\mathcal{F}_1 = \mathcal{F}_2 = 2^{\{1,2\}}$. Then, $|\mathcal{F}_1 \times \mathcal{F}_2| = 1 + 3 \times 3 = 10$ (because $\emptyset \times X = \emptyset$), while, since $\mathcal{F}_1 \times \mathcal{F}_2$ includes the singletons of $2^{\{1,2\} \times \{1,2\}}$, $\sigma(\mathcal{F}_1 \times \mathcal{F}_2) = 2^{\{1,2\} \times \{1,2\}}$. Hence, six sets are missing from $\mathcal{F}_1 \times \mathcal{F}_2$. For example, $\{(1,1),(2,2)\} \in \sigma(\mathcal{F}_1 \times \mathcal{F}_2) \setminus \mathcal{F}_1 \times \mathcal{F}_2$.

The σ-algebra $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$ is called the **product σ-algebra** of $(\mathcal{F}_k)_{k \in [n]}$ and is also denoted by $\mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n$. The product operation turns out to be associative: $(\mathcal{F}_1 \otimes \mathcal{F}_2) \otimes$

$\mathcal{F}_3 = \mathcal{F}_1 \otimes (\mathcal{F}_2 \otimes \mathcal{F}_3)$, which justifies writing $\mathcal{F}_1 \otimes \mathcal{F}_2 \otimes \mathcal{F}_3$. As it turns out, things work out well again with Borel σ-algebras: for $p, q \in \mathbb{N}^+$, $\mathfrak{B}(\mathbb{R}^{p+q}) = \mathfrak{B}(\mathbb{R}^p) \otimes \mathfrak{B}(\mathbb{R}^q)$. Needless to say, the same holds when there are more than two terms in the product. The n-fold product σ-algebra of $\mathcal{F}$ is denoted by $\mathcal{F}^{\otimes n}$.

2.2 σ-Algebras and Knowledge

One of the conceptual advantages of measure-theoretic probability is the relationship between σ-algebras and the intuitive idea of 'knowledge'. Although the relationship is useful and intuitive, it is regrettably not quite perfect. Let $(\Omega, \mathcal{F})$, $(\mathcal{X}, \mathcal{G})$ and $(\mathcal{Y}, \mathcal{H})$ be measurable spaces and $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ be random elements. Having observed the value of X ('knowing X'), one might wonder what this entails about the value of Y. Even more simplistically, under what circumstances can the value of Y be determined exactly having observed X? The situation is illustrated in Fig. 2.3. As it turns out, with some restrictions, the answer can be given in terms of the σ-algebras generated by X and Y. Except for a technical assumption on $(\mathcal{Y}, \mathcal{H})$, the following result shows that Y is a measurable function of X if and only if Y is $\sigma(X)/\mathcal{H}$-measurable. The technical assumption mentioned requires $(\mathcal{Y}, \mathcal{H})$ to be a Borel space, which is true of all probability spaces considered in this book, including $(\mathbb{R}^k, \mathfrak{B}(\mathbb{R}^k))$. We leave the exact definition of Borel spaces to the next chapter.

LEMMA 2.5 (Factorisation lemma). *Assume that $(\mathcal{Y}, \mathcal{H})$ is a Borel space. Then Y is $\sigma(X)$-measurable ($\sigma(Y) \subseteq \sigma(X)$) if and only if there exists a $\mathcal{G}/\mathcal{H}$-measurable map $f : \mathcal{X} \to \mathcal{Y}$ such that $Y = f \circ X$.*

In this sense $\sigma(X)$ contains all the information that can be extracted from X via measurable functions. This is not the same as saying that Y can be deduced from X if and only if Y is $\sigma(X)$-measurable because the set of $\mathcal{X} \to \mathcal{Y}$ maps can be much larger than the set of $\mathcal{G}/\mathcal{H}$-measurable functions. When $\mathcal{G}$ is coarse, there are not many $\mathcal{G}/\mathcal{H}$-measurable functions with the extreme case occurring when $\mathcal{G} = \{\mathcal{X}, \emptyset\}$. In cases like this, the intuition that $\sigma(X)$ captures all there is to know about X is not true anymore (Exercise 2.6). The issue is that $\sigma(X)$ does not only depend on X, but also on the σ-algebra of $(\mathcal{X}, \mathcal{G})$ and that if $\mathcal{G}$ is coarse-grained, then $\sigma(X)$ can also be coarse-grained and not many functions will be $\sigma(X)$-measurable. If X is a random variable, then by definition $\mathcal{X} = \mathbb{R}$ and $\mathcal{G} = \mathfrak{B}(\mathbb{R})$, which is relatively fine-grained, and the requirement that f be measurable is less restrictive. Nevertheless, even in the nicest setting where $\Omega = \mathcal{X} = \mathcal{Y} = \mathbb{R}$ and

$$\begin{tikzcd}
(\Omega, \mathcal{F}) \arrow[r, "X"] \arrow[dr, "Y"'] & (\mathcal{X}, \mathcal{G}) \arrow[d, "f"] \\
& (\mathcal{Y}, \mathcal{H})
\end{tikzcd}$$

Figure 2.3 The factorisation problem asks whether there exists a (measurable) function f that makes the diagram commute.

$\mathcal{F} = \mathcal{G} = \mathcal{H} = \mathfrak{B}(\mathbb{R})$, it can still occur that $Y = f \circ X$ for some non-measurable f. In other words, all the information about Y exists in X but cannot be extracted in a measurable way. These problems only occur when X maps measurable sets in Ω to non-measurable sets in $\mathcal{X}$. Fortunately, while such random variables exist, they are never encountered in applications, which provides the final justification for thinking of $\sigma(X)$ as containing all that there is to know about any random variable X that one may ever expect to encounter.

Filtrations

In the study of bandits and other online settings, information is revealed to the learner sequentially. Let $X_1, \dots, X_n$ be a collection of random variables on a common measurable space $(\Omega, \mathcal{F})$. We imagine a learner is sequentially observing the values of these random variables. First X_1, then X_2 and so on. The learner needs to make a prediction, or act, based on the available observations. Say, a prediction or an act must produce a real-valued response. Then, having observed $X_{1:t} \doteq (X_1, \dots, X_t)$, the set of maps $f \circ X_{1:t}$ where $f : \mathbb{R}^t \to \mathbb{R}$ is Borel, captures all the possible ways the learner can respond. By Lemma 2.5, this set contains exactly the $\sigma(X_{1:t})/\mathfrak{B}(\mathbb{R})$-measurable maps. Thus, if we need to reason about the set of $\Omega \to \mathbb{R}$ maps available after observing $X_{1:t}$, it suffices to concentrate on the σ-algebra $\mathcal{F}_t = \sigma(X_{1:t})$. Conveniently, $\mathcal{F}_t$ is independent of the space of possible responses, and being a subset of $\mathcal{F}$, it also hides details about the range space of $X_{1:t}$. It is easy to check that $\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_n \subseteq \mathcal{F}$, which means that more and more functions are becoming $\mathcal{F}_t$-measurable as t increases, which corresponds to increasing knowledge (note that $\mathcal{F}_0 = \{\emptyset, \Omega\}$, and the set of $\mathcal{F}_0$-measurable functions is the set of constant functions on Ω).

Bringing these a little further, we will often find it useful to talk about increasing sequences of σ-algebras without constructing them in terms of random variables as above. Given a measurable space $(\Omega, \mathcal{F})$, a **filtration** is a sequence $(\mathcal{F}_t)_{t=0}^n$ of sub-σ-algebras of $\mathcal{F}$ where $\mathcal{F}_t \subseteq \mathcal{F}_{t+1}$ for all $t < n$. We also allow $n = \infty$, and in this case we define

$$\mathcal{F}_\infty = \sigma\left(\bigcup_{t=0}^{\infty} \mathcal{F}_t\right)$$

to be the smallest σ-algebra containing the union of all $\mathcal{F}_t$. Filtrations can also be defined in continuous time, but we have no need for that here. A sequence of random variables $(X_t)_{t=1}^n$ is **adapted** to filtration $\mathbb{F} = (\mathcal{F}_t)_{t=0}^n$ if X_t is $\mathcal{F}_t$-measurable for each t. We also say in this case that $(X_t)_t$ is $\mathbb{F}$-adapted. The same nomenclature applies if n is infinite. Finally, $(X_t)_t$ is $\mathbb{F}$-**predictable** if X_t is $\mathcal{F}_{t-1}$-measurable for each $t \in [n]$. Intuitively we may think of an $\mathbb{F}$-predictable process $X = (X_t)_t$ as one that has the property that X_t can be known (or 'predicted') based on $\mathcal{F}_{t-1}$, while a $\mathbb{F}$-adapted process is one that has the property that X_t can be known based on $\mathcal{F}_t$ only. Since $\mathcal{F}_{t-1} \subseteq \mathcal{F}_t$, a predictable process is also adapted. A **filtered probability space** is the tuple $(\Omega, \mathcal{F}, \mathbb{F}, \mathbb{P})$, where $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and $\mathbb{F} = (\mathcal{F}_t)_t$ is filtration of $\mathcal{F}$.

2.3 Conditional Probabilities

Conditional probabilities are introduced so that we can talk about how probabilities should be updated when one gains some partial knowledge about a random outcome. Let $(\Omega, \mathcal{F}, \mathbb{P})$

When we say that $X_1, \dots, X_n$ are independent random variables, we mean that they are mutually independent. Independence is always relative to some probability measure, even when a probability measure is not explicitly mentioned. In such cases the identity of the probability measure should be clear from the context.

2.5 Integration and Expectation

A key quantity in probability theory is the **expectation** of a random variable. Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random variable $X : \Omega \to \mathbb{R}$. The expectation X is often denoted by $\mathbb{E}[X]$. This notation unfortunately obscures the dependence on the measure $\mathbb{P}$. When the underlying measure is not obvious from context, we write $\mathbb{E}_{\mathbb{P}}$ to indicate the expectation with respect to $\mathbb{P}$. Mathematically, we define the expected value of X as its Lebesgue integral with respect to $\mathbb{P}$:

$$\mathbb{E}[X] = \int_\Omega X(\omega)\, \mathrm{d}\mathbb{P}(\omega).$$

The right-hand side is also often abbreviated to $\int X\, \mathrm{d}\mathbb{P}$. The integral on the right-hand side is constructed to satisfy the following two key properties:

(a) The integral of indicators is the probability of the underlying event. If $X(\omega) = \mathbb{I}\{\omega \in A\}$ is an indicator function for some $A \in \mathcal{F}$, then $\int X \mathrm{d}\mathbb{P} = \mathbb{P}(A)$.

(b) Integrals are linear. For all random variables X_1, X_2 and reals α_1, α_2 such that $\int X_1 \mathrm{d}\mathbb{P}$ and $\int X_2 \mathrm{d}\mathbb{P}$ are defined, $\int (\alpha_1 X_1 + \alpha_2 X_2) \mathrm{d}\mathbb{P}$ is defined and satisfies

$$\int_\Omega (\alpha_1 X_1 + \alpha_2 X_2)\, \mathrm{d}\mathbb{P} = \alpha_1 \int_\Omega X_1\, \mathrm{d}\mathbb{P} + \alpha_2 \int_\Omega X_2\, \mathrm{d}\mathbb{P}. \tag{2.5}$$

These two properties together tell us that whenever $X(\omega) = \sum_{i=1}^n \alpha_i \mathbb{I}\{\omega \in A_i\}$ for some n, $\alpha_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$, $i = 1, \dots, n$, then

$$\int_\Omega X \mathrm{d}\mathbb{P} = \sum_i \alpha_i \mathbb{P}(A_i). \tag{2.6}$$

Functions of the form X are called **simple functions**.

In defining the Lebesgue integral of some random variable X, we use (2.6) as the definition of the integral when X is a simple function. The next step is to extend the definition to non-negative random variables. Let $X : \Omega \to [0, \infty)$ be measurable. The idea is to approximate X from below using simple functions and take the largest value that can be obtained this way:

$$\int_\Omega X \mathrm{d}\mathbb{P} = \sup \left\{ \int_\Omega h\, \mathrm{d}\mathbb{P} : h \text{ is simple and } 0 \le h \le X \right\}. \tag{2.7}$$

The meaning of $U \le V$ for random variables U, V is that $U(\omega) \le V(\omega)$ for all $\omega \in \Omega$. The supremum on the right-hand side could be infinite, in which case we say the integral of X is not defined. Whenever the integral of X is defined, we say that X is **integrable** or, if the identity of the measure $\mathbb{P}$ is unclear, that X is integrable with respect to $\mathbb{P}$.

Integrals for arbitrary random variables are defined by decomposing the random variable into positive and negative parts. Let $X : \Omega \to \mathbb{R}$ be any measurable function. Then define $X^+(\omega) = X(\omega)\mathbb{I}\{X(\omega) > 0\}$ and $X^-(\omega) = -X(\omega)\mathbb{I}\{X(\omega) < 0\}$ so that $X(\omega) = X^+(\omega) - X^-(\omega)$. Now X^+ and X^- are both non-negative random variables called the **positive** and **negative** parts of X. Provided that both X^+ and X^- are integrable, we define

$$\int_\Omega X \mathrm{d}\mathbb{P} = \int_\Omega X^+ \mathrm{d}\mathbb{P} - \int_\Omega X^- \mathrm{d}\mathbb{P}.$$

Note that X is integrable if and only if the non-negative-valued random variable $|X|$ is integrable (Exercise 2.12).

> None of what we have done depends on $\mathbb{P}$ being a probability measure. The definitions hold for any measure, though for signed measures it is necessary to split Ω into disjoint measurable sets on which the measure is positive/negative, an operation that is possible by the **Hahn decomposition theorem**. We will never need signed measures in this book, however.

A particularly interesting case is when $\Omega = \mathbb{R}$ is the real line, $\mathcal{F} = \mathfrak{B}(\mathbb{R})$ is the Borel σ-algebra and the measure is the **Lebesgue measure** λ, which is the unique measure on $\mathfrak{B}(\mathbb{R})$ such that $\lambda((a, b)) = b - a$ for any $a \leq b$. In this scenario, if $f : \mathbb{R} \to \mathbb{R}$ is a Borel-measurable function, then we can write the Lebesgue integral of f with respect to the Lebesgue measure as

$$\int_\mathbb{R} f \, d\lambda.$$

Perhaps unsurprisingly, this almost always coincides with the improper Riemann integral of f, which is normally written as $\int_{-\infty}^{\infty} f(x)dx$. Precisely, if $|f|$ is both Lebesgue integrable and Riemann integrable, then the integrals are equal.

> There exist functions that are Riemann integrable and not Lebesgue integrable, and also the other way around (although examples of the former are more exotic than the latter).

The Lebesgue measure and its relation to Riemann integration is mentioned because when it comes to actually calculating the value of an expectation or integral, this is often reduced to calculating integrals over the real line with respect to the Lebesgue measure. The calculation is then performed by evaluating the Riemann integral, thereby circumventing the need to rederive the integral of many elementary functions. Integrals (and thus expectations) have a number of important properties. By far the most important is their linearity, which was postulated above as the second property in (2.5). To practice using the notation with expectations, we restate the first half of this property. In fact, the statement is slightly more general than what we demanded for integrals above.

Proposition 2.6. *Let $(X_i)_i$ be a (possibly infinite) sequence of random variables on the same probability space and assume that $\mathbb{E}[X_i]$ exists for all i and furthermore that $X = \sum_i X_i$ and $\mathbb{E}[X]$ also exist. Then*

$$\mathbb{E}[X] = \sum_i \mathbb{E}[X_i] .$$

This exchange of expectations and summation is the source of much magic in probability theory because it holds even if X_i are not independent. This means that (unlike probabilities) we can very often decouple the expectations of dependent random variables, which often proves extremely useful (a collection of random variables is dependent if they are not independent). You will prove Proposition 2.6 in Exercise 2.14. The other requirement for linearity is that if $c \in \mathbb{R}$ is a constant, then $\mathbb{E}[cX] = c\,\mathbb{E}[X]$ (Exercise 2.15).

Another important statement is concerned with independent random variables.

Proposition 2.7. *If X and Y are independent, then $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$.*

In general $\mathbb{E}[XY] \neq \mathbb{E}[X]\,\mathbb{E}[Y]$ (Exercise 2.18). Finally, an important simple result connects expectations of non-negative random variables to their tail probabilities.

Proposition 2.8. *If $X \geq 0$ is a non-negative random variable, then*

$$\mathbb{E}[X] = \int_0^\infty \mathbb{P}(X > x)\, dx.$$

The integrand in Proposition 2.8 is called the **tail probability function** $x \mapsto \mathbb{P}(X > x)$ of X. This is also known as the complementary cumulative distribution function of X. The **cumulative distribution function** (CDF) of X is defined as $x \mapsto \mathbb{P}(X \leq x)$ and is usually denoted by F_X. These functions are defined for all random variables, not just non-negative ones. One can check that $F_X : \mathbb{R} \to [0, 1]$ is increasing, right continuous and $\lim_{x\to-\infty} F_X(x) = 0$ and $\lim_{x\to\infty} F_X(x) = 1$. The CDF of a random variable captures every aspect of the probability measure $\mathbb{P}_X$ induced by X, while still being just a function on the real line, a property that makes it a little more human friendly than $\mathbb{P}_X$. One can also generalise CDFs to random vectors: if X is an $\mathbb{R}^k$-valued random vector, then its CDF is defined as the $F_X : \mathbb{R}^k \to [0, 1]$ function that satisfies $F_X(x) = \mathbb{P}(X \leq x)$, where, in line with our conventions, $X \leq x$ means that all components of X are less than or equal to the respective component of x. The pushforward $\mathbb{P}_X$ of a random element is an alternative way to summarise the distribution of X. In particular, for any real-valued, $f : \mathcal{X} \to \mathbb{R}$ measurable function,

$$\mathbb{E}[f(X)] = \int_{\mathcal{X}} f(x) \mathrm{d}\mathbb{P}_X(x)$$

provided that either the right-hand side, or the left-hand side exist.

2.6 Conditional Expectation

Conditional expectation allows us to talk about the expectation of a random variable given the value of another random variable, or more generally, given some σ-algebra.

EXAMPLE 2.9. Let $(\Omega, \mathcal{F}, \mathbb{P})$ model the outcomes of an unloaded dice: $\Omega = [6]$, $\mathcal{F} = 2^\Omega$ and $\mathbb{P}(A) = |A|/6$. Define two random variables X and Y by $Y(\omega) = \mathbb{I}\{\omega > 3\}$ and $X(\omega) = \omega$. Suppose we are interested in the expectation of X given a specific value of Y. Arguing intuitively, we might notice that $Y = 1$ means that the unobserved X must be either 4, 5 or 6, and that each of these outcomes is equally likely, and so the expectation of X given $Y = 1$ should be $(4+5+6)/3 = 5$. Similarly, the expectation of X given $Y = 0$ should be $(1+2+3)/3 = 2$. If we want a concise summary, we can just write that 'the expectation of X given Y' is $5Y + 2(1-Y)$. Notice how this is a random variable itself.

The notation for this conditional expectation is $\mathbb{E}\left[X \mid Y\right]$. Using this notation, in Example 2.9 we can concisely write $\mathbb{E}\left[X \mid Y\right] = 5Y + 2(1-Y)$. A little more generally, if $X : \Omega \to \mathcal{X}$ and $Y : \Omega \to \mathcal{Y}$ with $\mathcal{X}, \mathcal{Y} \subset \mathbb{R}$ and $|\mathcal{X}|, |\mathcal{Y}| < \infty$, then $\mathbb{E}[X \mid Y] : \Omega \to \mathbb{R}$ is the random variable given by $\mathbb{E}[X \mid Y](\omega) = \mathbb{E}[X \mid Y = Y(\omega)]$, where

$$\mathbb{E}[X \mid Y = y] = \sum_{x \in \mathcal{X}} x\,\mathbb{P}\left(X = x \mid Y = y\right) = \sum_{x \in \mathcal{X}} \frac{x\,\mathbb{P}\left(X = x, Y = y\right)}{\mathbb{P}\left(Y = y\right)}. \tag{2.8}$$

This is undefined when $\mathbb{P}(Y = y) = 0$ so that $\mathbb{E}[X \mid Y](\omega)$ is undefined on the measure zero set $\{\omega : \mathbb{P}(Y = Y(\omega)) = 0\}$.

Eq. (2.8) does not generalise to continuous random variables because $\mathbb{P}\left(Y = y\right)$ in the denominator might be zero for all y. For example, let Y be a random variable taking values on $[0, 1]$ according to a uniform distribution and $X \in \{0, 1\}$ be Bernoulli with bias Y. This means that the joint measure on X and Y is $\mathbb{P}\left(X = 1, Y \in [p, q]\right) = \int_p^q x dx$ for $0 \leq p < q \leq 1$. Intuitively it seems like $\mathbb{E}[X \mid Y]$ should be equal to Y, but how to define it? The mean of a Bernoulli random variable is equal to its bias so the definition of conditional probability shows that for $0 \leq p < q \leq 1$,

$$\begin{aligned}
\mathbb{E}[X = 1 \mid Y \in [p, q]] &= \mathbb{P}\left(X = 1 \mid Y \in [p, q]\right) \\
&= \frac{\mathbb{P}\left(X = 1, Y \in [p, q]\right)}{\mathbb{P}\left(Y \in [p, q]\right)} \\
&= \frac{q^2 - p^2}{2(q - p)} \\
&= \frac{p + q}{2}.
\end{aligned}$$

This calculation is not well defined when $p = q$ because $\mathbb{P}\left(Y \in [p, p]\right) = 0$. Nevertheless, letting $q = p + \varepsilon$ for $\varepsilon > 0$ and taking the limit as ε tends to zero seems like a reasonable way to argue that $\mathbb{P}\left(X = 1 \mid Y = p\right) = p$. Unfortunately this approach does not generalise to abstract spaces because there is no canonical way of taking limits towards a set of measure zero, and different choices lead to different answers.

Instead we use Eq. (2.8) as the starting point for an abstract definition of conditional expectation as a random variable satisfying two requirements. First, from Eq. (2.8) we see that $\mathbb{E}[X \mid Y](\omega)$ should only depend on $Y(\omega)$ and so should be measurable with respect to $\sigma(Y)$. The second requirement is called the 'averaging property'. For measurable $A \subseteq \mathcal{Y}$, Eq. (2.8) shows that

$$\begin{aligned}\mathbb{E}[\mathbb{I}_{Y^{-1}(A)}\mathbb{E}[X \mid Y]] &= \sum_{y\in A} \mathbb{P}\left(Y=y\right)\mathbb{E}[X \mid Y=y] \\ &= \sum_{y\in A}\sum_{x\in\mathcal{X}} x\,\mathbb{P}\left(X=x, Y=y\right) \\ &= \mathbb{E}[\mathbb{I}_{Y^{-1}(A)}X].\end{aligned}$$

This can be viewed as putting a set of linear constraints on $\mathbb{E}[X \mid Y]$ with one constraint for each measurable $A \subseteq \mathcal{Y}$. By treating $\mathbb{E}[X \mid Y]$ as an unknown $\sigma(Y)$-measurable random variable, we can attempt to solve this linear system. As it turns out, this can always be done: the linear constraints and the measurability restriction on $\mathbb{E}\left[X \mid Y\right]$ completely determine $\mathbb{E}[X \mid Y]$ except for a set of measure zero. Notice that both conditions only depend on $\sigma(Y) \subseteq \mathcal{F}$. The abstract definition of conditional expectation takes these properties as the definition and replaces the role of Y with a sub-σ-algebra.

Definition 2.10 (Conditional expectation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : \Omega \to \mathbb{R}$ be random variable and $\mathcal{H}$ be a sub-σ-algebra of $\mathcal{F}$. The conditional expectation of X given $\mathcal{H}$ is denoted by $\mathbb{E}[X \mid \mathcal{H}]$ and defined to be any $\mathcal{H}$-measurable random variable on Ω such that for all $H \in \mathcal{H}$,

$$\int_H \mathbb{E}[X \mid \mathcal{H}]\mathrm{d}\mathbb{P} = \int_H X\mathrm{d}\mathbb{P}. \tag{2.9}$$

Given a random variable Y, the conditional expectation of X given Y is $\mathbb{E}\left[X \mid Y\right] = \mathbb{E}\left[X \mid \sigma(Y)\right]$.

Theorem 2.11. *Given any probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a sub-$\sigma$-algebra $\mathcal{H}$ of $\mathcal{F}$ and a $\mathbb{P}$-integrable random variable $X : \Omega \to \mathbb{R}$, there exists an $\mathcal{H}$-measurable function $f : \Omega \to \mathbb{R}$ that satisfies* (2.9). *Further, any two $\mathcal{H}$-measurable functions $f_1, f_2 : \Omega \to \mathbb{R}$ that satisfy* (2.9) *are equal with probability one: $\mathbb{P}(f_1 = f_2) = 1$.*

When random variables X and Y agree with $\mathbb{P}$-probability one, we say they are **$\mathbb{P}$-almost surely** equal, which is often abbreviated to '$X = Y$ $\mathbb{P}$-a.s.', or '$X = Y$ a.s.' when the measure is clear from context. A related useful notion is the concept of **null sets**: $U \in \mathcal{F}$ is a null set of $\mathbb{P}$, or a $\mathbb{P}$-null set if $\mathbb{P}(U) = 0$. Thus, $X = Y$ $\mathbb{P}$-a.s. if and only if $X = Y$ agree except on a $\mathbb{P}$-null set.

> The reader may find it odd that $\mathbb{E}[X \mid Y]$ is a random variable on Ω rather than the range of Y. Lemma 2.5 and the fact that $\mathbb{E}[X \mid \sigma(Y)]$ is $\sigma(Y)$-measurable shows there exists a measurable function $f : (\mathbb{R}, \mathfrak{B}(\mathbb{R})) \to (\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ such that $\mathbb{E}[X \mid \sigma(Y)](\omega) = (f \circ Y)(\omega)$ (see Fig. 2.4). In this sense $\mathbb{E}[X \mid Y](\omega)$ only depends on $Y(\omega)$, and occasionally we write $\mathbb{E}[X \mid Y](y)$.

Returning to Example 2.9, we see that $\mathbb{E}\left[X \mid Y\right] = \mathbb{E}\left[X \mid \sigma(Y)\right]$ and $\sigma(Y) = \{\{1, 2, 3\}, \{4, 5, 6\}, \emptyset, \Omega\}$. Denote this set-system by $\mathcal{H}$ for brevity. The condition that

5 Can you think of a set that is not Borel measurable? Such sets exist, but do not arise naturally in applications. The classic example is the **Vitali set**, which is formed by taking the quotient group $G = \mathbb{R}/\mathbb{Q}$ and then applying the axiom of choice to choose a representative in $[0, 1]$ from each equivalence class in G. Non-measurable functions are so unusual that you do not have to worry much about whether or not functions $X : \mathbb{R} \to \mathbb{R}$ are measurable. With only a few exceptions, questions of measurability arising in this book are not related to the fine details of the Borel σ-algebra. Much more frequently they are related to filtrations and the notion of knowledge available having observed certain random elements.

6 There is a lot to say about why the sum, or the product of random variables are also random variables. Or why $\inf_n X_n$, $\sup_n X_n$, $\liminf_n X_n$, $\limsup_n X_n$ are measurable when X_n are. The key point is to show that the composition of measurable maps is a measurable map and that continuous maps are measurable and then apply these results (Exercise 2.1). For $\limsup_n X_n$, just rewrite it as $\lim_{m\to\infty} \sup_{n\geq m} X_n$; note that $\sup_{n\geq m} X_n$ is decreasing (we take suprema of smaller sets as m increases), hence $\limsup_n X_n = \inf_m \sup_{n\geq m} X_n$, reducing the question to studying $\inf_n X_n$ and $\sup_n X_n$. Finally, for $\inf_n X_n$ note that it suffices if $\{\omega : \inf_n X_n \geq t\}$ is measurable for any t real. Now, $\inf_n X_n \geq t$ if and only if $X_n \geq t$ for all n. Hence, $\{\omega : \inf_n X_n \geq t\} = \cap_n \{\omega : X_n \geq t\}$, which is a countable intersection of measurable sets, hence measurable (this latter follows by the elementary identity $(\cap_i A_i)^c = \cup_i A_i^c$).

7 The factorisation lemma, Lemma 2.5, is attributed to Joseph Doob and Eugene Dynkin. The lemma sneakily uses the properties of real numbers (think about why), which is another reason why what we said about σ-algebras containing all information is not entirely true. The lemma has extensions to more general random elements [Taraldsen, 2018, for example]. The key requirement in a way is that the σ-algebra associated with the range space of Y should be rich enough.

8 We did not talk about basic results like Lebesgue's dominated/monotone convergence theorems, Fatou's lemma or Jensen's inequality. We will definitely use the last of these, which is explained in a dedicated chapter on convexity (Chapter 26). The other results can be found in the texts we cite. They are concerned with infinite sequences of random variables and conditions under which their limits can be interchanged with Lebesgue integrals. In this book we rarely encounter problems related to such sequences and hope you forgive us on the few occasions they are necessary (the reason is simply because we mostly focus on finite time results or take expectations before taking limits when dealing with asymptotics).

9 You might be surprised that we have not mentioned **densities**. For most of us, our first exposure to probability on continuous spaces was by studying the normal distribution and its density

$$p(x) = \frac{1}{\sqrt{2\pi}} \exp(-x^2/2), \tag{2.10}$$

which can be integrated over intervals to obtain the probability that a Gaussian random variable will take a value in that interval. The reader should notice that $p : \mathbb{R} \to \mathbb{R}$ is Borel measurable and that the Gaussian measure associated with this density is $\mathbb{P}$ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ defined by

$$\mathbb{P}(A) = \int_A p \, d\lambda.$$

Here the integral is with respect to the Lebesgue measure λ on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. The notion of a density can be generalised beyond this simple setup. Let P and Q be measures (not necessarily probability measures) on arbitrary measurable space $(\Omega, \mathcal{F})$. The **Radon–Nikodym derivative** of P with respect to Q is an $\mathcal{F}$-measurable random variable $\frac{dP}{dQ} : \Omega \to [0, \infty)$ such that

$$P(A) = \int_A \frac{dP}{dQ}\, dQ \qquad \text{for all } A \in \mathcal{F}. \tag{2.11}$$

We can also write this in the form $\int \mathbb{I}_A dP = \int \mathbb{I}_A \frac{dP}{dQ} dQ$, $A \in \mathcal{F}$, from which we may realise that for any X P-integrable random variable, $\int X dP = \int X \frac{dP}{dQ} dQ$ must also hold. This is often called the **change-of-measure formula**. Another word for the Radon–Nikodym derivative $\frac{dP}{dQ}$ is the **density** of P with respect to Q. It is not hard to find examples where the density does not exist. We say that P is **absolutely continuous** with respect to Q if $Q(A) = 0 \implies P(A) = 0$ for all $A \in \mathcal{F}$. When $\frac{dP}{dQ}$ exists, it follows immediately that P is absolutely continuous with respect to Q by Eq. (2.11). Except for some pathological cases, it turns out that this is both necessary and sufficient for the existence of dP/dQ. The measure Q is σ-finite if there exists a countable covering $\{A_i\}$ of Ω with $\mathcal{F}$-measurable sets such that $Q(A_i) < \infty$ for each i.

THEOREM 2.13. *Let P, Q be measures on a common measurable space $(\Omega, \mathcal{F})$ and assume that Q is σ-finite. Then the density of P with respect to Q, $\frac{dP}{dQ}$, exists if and only if P is absolutely continuous with respect to Q. Furthermore, $\frac{dP}{dQ}$ is uniquely defined up to a Q-null set so that for any f_1, f_2 satisfying* (2.11)*, $f_1 = f_2$ holds Q-almost surely.*

Densities work as expected. Suppose that Z is a standard Gaussian random variable. We usually write its density as in Eq. (2.10), which we now know is the Radon–Nikodym derivative of the Gaussian measure with respect to the Lebesgue measure. The densities of 'classical' continuous distributions are almost always defined with respect to the Lebesgue measure.

10 In line with the literature, we will use $P \ll Q$ to denote that P is absolutely continuous with respect to Q. When P is absolutely continuous with respect to Q, we also say that Q **dominates** P.

11 A useful result for Radon–Nikodym derivatives is the **chain rule**, which states that if $P \ll Q \ll S$, then $\frac{dP}{dQ}\frac{dQ}{dS} = \frac{dP}{dS}$. The proof of this result follows from our earlier observation that $\int f dQ = \int f \frac{dQ}{dS} dS$ for any Q-integrable f. Indeed, the chain rule is obtained from this by taking $f = \mathbb{I}_A \frac{dP}{dQ}$ with $A \in \mathcal{F}$ and noting that this is indeed Q-integrable and $\int \mathbb{I}_A \frac{dP}{dQ} dQ = \int \mathbb{I}_A dQ$. The chain rule is often used to reduce the calculation of densities to calculation with known densities.

12 The Radon–Nikodym derivative unifies the notions of distribution (for discrete spaces) and density (for continuous spaces). Let Ω be discrete (finite or countable) and let ρ be the **counting measure** on $(\Omega, 2^\Omega)$, which is defined by $\rho(A) = |A|$. For any P on $(\Omega, \mathcal{F})$, it is easy to see that $P \ll \rho$ and $\frac{dP}{d\rho}(i) = P(\{i\})$, which is sometimes called the distribution function of P.

13 The Radon–Nikodym derivative provides another way to define the conditional expectation. Let X be an integrable random variable on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{H} \subset \mathcal{F}$ be a sub-σ-algebra and $\mathbb{P}|_{\mathcal{H}}$ be the restriction of $\mathbb{P}$ to $(\Omega, \mathcal{H})$. Define measure μ on $(\Omega, \mathcal{H})$ by $\mu(A) = \int_A X d\mathbb{P}|_{\mathcal{H}}$. It is easy to check that $\mu \ll \mathbb{P}|_{\mathcal{H}}$ and that $\mathbb{E}[X \mid \mathcal{H}] = \frac{d\mu}{d\mathbb{P}|_{\mathcal{H}}}$ satisfies Eq. (2.9). We note that the proof of the Radon–Nikodym theorem is nontrivial and that the existence of conditional expectations are more easily guaranteed via an 'elementary' but abstract argument using functional analysis.

14 The **Fubini–Tonelli theorem** is a powerful result that allows one to exchange the order of integrations. This result is needed for example for proving Proposition 2.8 (Exercise 2.19). To state it, we need to introduce **product measures**. These work as expected: given two probability spaces, $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$, the product measure $\mathbb{P}$ of $\mathbb{P}_1$ and $\mathbb{P}_2$ is defined as any measure on $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2)$ that satisfies $\mathbb{P}(A_1, A_2) = \mathbb{P}_1(A_1)\mathbb{P}_2(A_2)$ for all $(A_1, A_2) \in \mathcal{F}_1 \times \mathcal{F}_2$ (recall that $\mathcal{F}_1 \otimes \mathcal{F}_2 = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ is the product σ-algebra of $\mathcal{F}_1$ and $\mathcal{F}_2$). Theorem 2.4 implies that this product measure, which is often denoted by $\mathbb{P}_1 \times \mathbb{P}_2$ (or $\mathbb{P}_1 \otimes \mathbb{P}_2$) is uniquely defined. (Think about what this product measure has to do with independence.) The Fubini–Tonelli theorem (often just 'Fubini') states the following: let $(\Omega_1, \mathcal{F}_1, \mathbb{P}_1)$ and $(\Omega_2, \mathcal{F}_2, \mathbb{P}_2)$

be two probability spaces and consider a random variable X on the product probability space $(\Omega, \mathcal{F}, \mathbb{P}) = (\Omega_1 \times \Omega_2, \mathcal{F}_1 \otimes \mathcal{F}_2, \mathbb{P}_1 \times \mathbb{P}_2)$. If any of the three integrals $\int |X(\omega)| \, \mathrm{d}\mathbb{P}(\omega)$, $\int (\int |X(\omega_1, \omega_2)| \, \mathrm{d}\mathbb{P}_1(\omega_1)) \, \mathrm{d}\mathbb{P}_2(\omega_2)$, $\int (\int |X(\omega_1, \omega_2)| \, \mathrm{d}\mathbb{P}_2(\omega_2)) \, \mathrm{d}\mathbb{P}_1(\omega_1)$ is finite, then

$$\begin{aligned}\int X(\omega) \, \mathrm{d}\mathbb{P}(\omega) &= \int \Big(\int X(\omega_1, \omega_2) \, \mathrm{d}\mathbb{P}_1(\omega_1) \Big) \, \mathrm{d}\mathbb{P}_2(\omega_2) \\ &= \int \Big(\int X(\omega_1, \omega_2) \, \mathrm{d}\mathbb{P}_2(\omega_2) \Big) \, \mathrm{d}\mathbb{P}_1(\omega_1).\end{aligned}$$

15 For topological space X, the **support** of a measure μ on $(X, \mathfrak{B}(X))$ is

$$\mathrm{Supp}(\mu) = \{x \in X : \mu(U) > 0 \text{ for all neighborhoods } U \text{ of } x\}.$$

When X is discrete, this reduces to $\mathrm{Supp}(\mu) = \{x : \mu(\{x\}) > 0\}$.

16 Let X be a topological space. The weak* topology on the space of probability measures $\mathcal{P}(X)$ on $(X, \mathfrak{B}(X))$ is the coarsest topology such that $\mu \mapsto \int f d\mu$ is continuous for all bounded continuous functions $f : X \to \mathbb{R}$. In particular, a sequence of probability measures $(\mu_n)_{n=1}^{\infty}$ converges to μ in this topology if and only if $\lim_{n\to\infty} \int f d\mu_n = \int f d\mu$ for all bounded continuous functions $f : X \to \mathbb{R}$.

THEOREM 2.14. *When X is compact and Hausdorff and $\mathcal{P}(X)$ is the space of regular probability measures on $(X, \mathfrak{B}(X))$ with the weak* topology, then $\mathcal{P}(X)$ is compact.*

17 Mathematical terminology can be a bit confusing sometimes. Since $\mathbb{E}$ maps (certain) functions to real values, it is also called the **expectation operator**. ‘Operator’ is just a fancy name for a function. In **operator theory**, the study of operators, the focus is on operators whose domain is infinite dimensional, hence the distinct name. However, most results of operator theory do not hinge upon this property. If the image space is the set of reals, we talk about **functionals**. The properties of functionals are studied in yet another subfield of mathematics, **functional analysis**. The expectation operator is a functional that maps the set of $\mathbb{P}$-integrable functions (often denoted by $L^1(\Omega, \mathbb{P})$ or $L^1(\mathbb{P})$) to reals. Its most important property is linearity, which was stated as a requirement for integrals that define the expectation operator (Eq. (2.5)). In line with the previous comment, when we use $\mathbb{E}$, more often than not, the probability space remains hidden. As such, the symbol $\mathbb{E}$ is further abused.

2.8 Bibliographic Remarks

Much of this chapter draws inspiration from David Pollard’s *A user’s guide to measure theoretic probability* [Pollard, 2002]. We like this book because the author takes a rigourous approach, but still explains the ‘why’ and ‘how’ with great care. The book gets quite advanced quite fast, concentrating on the big picture rather than getting lost in the details. Other useful references include the book by Billingsley [2008], which has many good exercises and is quite comprehensive in terms of its coverage of the ‘basics’. These books are both quite detailed. For an outstanding shorter introduction to measure-theoretic probability, see the book by Williams [1991], which has an enthusiastic style and a pleasant bias towards martingales. We also like the book by Kallenberg [2002], which is recommended for the mathematically inclined readers who already have a good understanding of the basics. The author has put a major effort into organising the material so that redundancy is minimised and generality is maximised. This reorganisation resulted in quite a few original proofs, and the book is comprehensive. The factorisation lemma (Lemma 2.5) is stated in the book by Kallenberg [2002]

(Lemma 1.13 there). Kallenberg calls this lemma the 'functional representation' lemma and attributes it to Joseph Doob. Theorem 2.4 is a corollary of Carathéodory's extension theorem, which says that probability measures defined on semi-rings of sets have a unique extension to the generated σ-algebra. The remaining results can be found in either of the three books mentioned above. Theorem 2.14 appears as theorem 8.9.3 in the two-volume book by Bogachev [2007]. Finally, for something older and less technical, we recommend the philosophical essays on probability by Pierre Laplace, which was recently reprinted [Laplace, 2012].

2.9 Exercises

2.1 (Composing random elements) Show that if f is $\mathcal{F}/\mathcal{G}$-measurable and g is $\mathcal{G}/\mathcal{H}$-measurable for sigma algebras $\mathcal{F}, \mathcal{G}$ and $\mathcal{H}$ over appropriate spaces, then their composition, $g \circ f$ (defined the usual way: $(g \circ f)(\omega) = g(f(\omega))$, $\omega \in \Omega$), is $\mathcal{F}/\mathcal{H}$-measurable.

2.2 Let $X_1, \dots, X_n$ be random variables on $(\Omega, \mathcal{F})$. Prove that $(X_1, \dots, X_n)$ is a random vector.

2.3 (Random variable induced σ-algebra) Let $\mathcal{U}$ be an arbitrary set and $(\mathcal{V}, \Sigma)$ a measurable space and $X : \mathcal{U} \to \mathcal{V}$ an arbitrary function. Show that $\Sigma_X = \{X^{-1}(A) : A \in \Sigma\}$ is a σ-algebra over $\mathcal{U}$.

2.4 Let $(\Omega, \mathcal{F})$ be a measurable space and $A \subseteq \Omega$ and $\mathcal{F}_{|A} = \{A \cap B : B \in \mathcal{F}\}$.

(a) Show that $(A, \mathcal{F}_{|A})$ is a measurable space.
(b) Show that if $A \in \mathcal{F}$, then $\mathcal{F}_{|A} = \{B : B \in \mathcal{F}, B \subseteq A\}$.

2.5 Let $\mathcal{G} \subseteq 2^{\Omega}$ be a non-empty collection of sets and define $\sigma(\mathcal{G})$ as the smallest σ-algebra that contains $\mathcal{G}$. By 'smallest' we mean that $\mathcal{F} \in 2^{\Omega}$ is smaller than $\mathcal{F}' \in 2^{\Omega}$ if $\mathcal{F} \subset \mathcal{F}'$.

(a) Show that $\sigma(\mathcal{G})$ exists and contains exactly those sets A that are in every σ-algebra that contains $\mathcal{G}$.
(b) Suppose $(\Omega', \mathcal{F})$ is a measurable space and $X : \Omega' \to \Omega$ be $\mathcal{F}/\mathcal{G}$-measurable. Show that X is also $\mathcal{F}/\sigma(\mathcal{G})$-measurable. (We often use this result to simplify the job of checking whether a random variable satisfies some measurability property).
(c) Prove that if $A \in \mathcal{F}$ where $\mathcal{F}$ is a σ-algebra, then $\mathbb{I}\{A\}$ is $\mathcal{F}$-measurable.

2.6 (Knowledge and σ-algebras: a pathological example) In the context of Lemma 2.5, show an example where $Y = X$ and yet Y is not $\sigma(X)$ measurable.

Hint As suggested after the lemma, this can be arranged by choosing $\Omega = \mathcal{Y} = \mathcal{X} = \mathbb{R}$, $X(\omega) = Y(\omega) = \omega$, $\mathcal{F} = \mathcal{H} = \mathfrak{B}(\mathbb{R})$ and $\mathcal{G} = \{\emptyset, \mathbb{R}\}$ to be the trivial σ-algebra.

2.7 Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, $B \in \mathcal{F}$ be such that $\mathbb{P}(B) > 0$. Prove that $A \mapsto \mathbb{P}(A \mid B)$ is a probability measure over $(\Omega, \mathcal{F})$.

2.8 (Bayes law) Verify (2.2).

2.9 Consider the standard probability space $(\Omega, \mathcal{F}, \mathbb{P})$ generated by two standard, unbiased, six-sided dice that are thrown independently of each other. Thus, $\Omega = \{1, \dots, 6\}^2$, $\mathcal{F} = 2^{\Omega}$ and $\mathbb{P}(A) = |A|/6^2$ for any $A \in \mathcal{F}$ so that $X_i(\omega) = \omega_i$ represents the outcome of throwing dice $i \in \{1, 2\}$.

(a) Show that the events '$X_1 < 2$' and 'X_2 is even' are independent of each other.

(b) More generally, show that for any two events, $A \in \sigma(X_1)$ and $B \in \sigma(X_2)$, are independent of each other.

2.10 (Serendipitous independence) The point of this exercise is to understand independence more deeply. Solve the following problems:

(a) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Show that $\emptyset$ and Ω (which are events) are independent of any other event. What is the intuitive meaning of this?

(b) Continuing the previous part, show that any event $A \in \mathcal{F}$ with $\mathbb{P}(A) \in \{0, 1\}$ is independent of any other event.

(c) What can we conclude about an event $A \in \mathcal{F}$ that is independent of its complement, $A^c = \Omega \setminus A$? Does your conclusion make intuitive sense?

(d) What can we conclude about an event $A \in \mathcal{F}$ that is independent of itself? Does your conclusion make intuitive sense?

(e) Consider the probability space generated by two independent flips of unbiased coins with the smallest possible σ-algebra. Enumerate all pairs of events A, B such that A and B are independent of each other.

(f) Consider the probability space generated by the independent rolls of two unbiased three-sided dice. Call the possible outcomes of the individual dice rolls 1, 2 and 3. Let X_i be the random variable that corresponds to the outcome of the ith dice roll ($i \in \{1, 2\}$). Show that the events $\{X_1 \le 2\}$ and $\{X_1 = X_2\}$ are independent of each other.

(g) The probability space of the previous example is an example when the probability measure is uniform on a finite outcome space (which happens to have a product structure). Now consider any n-element, finite outcome space with the uniform measure. Show that A and B are independent of each other if and only if the cardinalities $|A|, |B|, |A \cap B|$ satisfy $n|A \cap B| = |A| \cdot |B|$.

(h) Continuing with the previous problem, show that if n is prime, then no non-trivial events are independent (an event A is **trivial** if $\mathbb{P}(A) \in \{0, 1\}$).

(i) Construct an example showing that pairwise independence does not imply mutual independence.

(j) Is it true or not that A, B, C are mutually independent if and only if $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(A)\,\mathbb{P}(B)\,\mathbb{P}(C)$? Prove your claim.

2.11 (Independence and random elements) Solve the following problems:

(a) Let X be a constant random element (that is, $X(\omega) = x$ for any $\omega \in \Omega$ over the outcome space over which X is defined). Show that X is independent of any other random variable.

(b) Show that the above continues to hold if X is almost surely constant (that is, $\mathbb{P}(X = x) = 1$ for an appropriate value x).

(c) Show that two events are independent if and only if their indicator random variables are independent (that is, A, B are independent if and only if $X(\omega) = \mathbb{I}\{\omega \in A\}$ and $Y(\omega) = \mathbb{I}\{\omega \in B\}$ are independent of each other).

(d) Generalise the result of the previous item to pairwise and mutual independence for collections of events and their indicator random variables.

2.12 Our goal in this exercise is to show that X is integrable if and only if $|X|$ is integrable. This is broken down into multiple steps. The first issue is to deal with the measurability of $|X|$. While a direct calculation can also show this, it may be worthwhile to follow a more general path:

(a) Any $f : \mathbb{R} \to \mathbb{R}$ continuous function is Borel measurable.

by reversing the process. To do this, we rearrange the $(F_t)_{t=1}^\infty$ sequence into a grid. For example:

$$\begin{array}{l} F_1, F_2, F_4, F_7, \cdots \\ F_3, F_5, F_8, \cdots \\ F_6, F_9, \cdots \\ F_{10}, \cdots \\ \vdots \end{array}$$

Letting $X_{m,t}$ be the tth entry in the mth row of this grid, we define $X_m = \sum_{t=1}^\infty 2^{-t} X_{m,t}$, and again one can easily check that with this choice the sequence $X_1, X_2, \ldots$ is independent and $\lambda_{X_t} = \mu$ is uniform for each t. □

3.1 Stochastic Processes

Let $\mathcal{T}$ be an arbitrary set. A **stochastic process** on probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a collection of random variables $\{X_t : t \in \mathcal{T}\}$. In this book $\mathcal{T}$ will always be countable, and so in the following we restrict ourselves to $\mathcal{T} = \mathbb{N}$. The first theorem is not the most general, but suffices for our purposes and is more easily stated than more generic alternatives.

THEOREM 3.2. *For each $n \in \mathbb{N}^+$, let $(\Omega_n, \mathcal{F}_n)$ be a Borel space and μ_n be a measure on $(\Omega_1 \times \cdots \times \Omega_n, \mathcal{F}_1 \otimes \cdots \otimes \mathcal{F}_n)$ and assume that μ_n and μ_{n+1} are related through*

$$\mu_{n+1}(A \times \Omega_{n+1}) = \mu_n(A) \qquad \textit{for all } A \in \Omega_1 \otimes \cdots \otimes \Omega_n. \tag{3.1}$$

Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random elements $X_1, X_2, \ldots$ with $X_t : \Omega \to \Omega_t$ such that $\mathbb{P}_{X_1,\ldots,X_n} = \mu_n$ for all n.

Sequences of measures $(\mu_n)_n$ satisfying Eq. (3.1) are called **projective**.

Theorem 3.1 follows immediately from Theorem 3.2. By assumption a random variable takes values in $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$, which is Borel. Then let $\mu_n = \otimes_{t=1}^n \mu$ be the n-fold product measure of μ with itself. That this sequence of measures is projective is clear, and the theorem does the rest.

3.2 Markov Chains

A Markov chain is an infinite sequence of random elements $(X_t)_{t=1}^\infty$ where the conditional distribution of X_{t+1} given $X_1, \ldots, X_t$ is the same as the conditional distribution of X_{t+1} given X_t. The sequence has the property that given the last element, the history is irrelevant to 'predict' the future. Such random sequences appear throughout probability theory and have many applications besides. The theory is too rich to explain in detail, so we give the basics and point towards the literature for more details at the end. The focus here is mostly on the definition and existence of Markov chains.

Let $(\mathcal{X}, \mathcal{F})$ and $(\mathcal{Y}, \mathcal{G})$ be measurable spaces. A **probability kernel** or **Markov kernel** from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{Y}, \mathcal{G})$ is a function $K : \mathcal{X} \times \mathcal{G} \to [0, 1]$ such that

(a) $K(x, \cdot)$ is a measure for all $x \in \mathcal{X}$; and
(b) $K(\cdot, A)$ is $\mathcal{F}$-measurable for all $A \in \mathcal{G}$.

The idea here is that K describes a stochastic transition. Having arrived at x, a process's next state is sampled $Y \sim K(x, \cdot)$. Occasionally, we will use the notation $K_x(A)$ or $K(A \,|\, x)$ rather than $K(x, A)$.

If K_1 is a $(\mathcal{X}, \mathcal{F}) \to (\mathcal{Y}, \mathcal{G})$ probability kernel and K_2 is a $(\mathcal{Y}, \mathcal{G}) \to (\mathcal{Z}, \mathcal{H})$ probability kernel, then the **product kernel** $K_1 \otimes K_2$ is the probability kernel from $(\mathcal{X}, \mathcal{F}) \to (\mathcal{Y} \times \mathcal{Z}, \mathcal{G} \otimes \mathcal{H})$ defined by

$$(K_1 \otimes K_2)(x, A) = \int_{\mathcal{Y}} \int_{\mathcal{Z}} \mathbb{I}_A((y, z)) K_2(y, dz) K_1(x, dy).$$

When P is a measure on $(\mathcal{X}, \mathcal{F})$ and K is a kernel from $\mathcal{X}$ to $\mathcal{Y}$, then $P \otimes K$ is a measure on $(\mathcal{X} \times \mathcal{Y}, \mathcal{F} \otimes \mathcal{G})$ defined by

$$(P \otimes K)(A) = \int_{\mathcal{X}} \int_{\mathcal{Y}} \mathbb{I}_A((x, y)) K(x, dy) dP(x).$$

There operations can be composed. When P is a probability measure on $\mathcal{X}$ and K_1 a kernel from $\mathcal{X}$ to $\mathcal{Y}$ and K_2 a kernel from $\mathcal{X} \times \mathcal{Y}$ to $\mathcal{Z}$, then $P \otimes K_1 \otimes K_2$ is a probability measure on $\mathcal{X} \times \mathcal{Y} \times \mathcal{Z}$. The following provides a counterpart of Theorem 3.2.

Theorem 3.3 (Ionescu–Tulcea). *Let $(\Omega_n, \mathcal{F}_n)_{n=1}^{\infty}$ be a sequence of measurable spaces and K_1 be a probability measure on $(\Omega_1, \mathcal{F}_1)$. For $n \geq 2$, let K_n be a probability kernel from $\prod_{t=1}^{n-1} \Omega_t$ to Ω_n. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and random elements $(X_t)_{t=1}^{\infty}$ with $X_t : \Omega \to \Omega_t$ such that $\mathbb{P}_{X_1, \ldots, X_n} = \bigotimes_{t=1}^{n} K_t$ for all $n \in \mathbb{N}^+$.*

A **homogeneous Markov chain** is a sequence of random elements $(X_t)_{t=1}^{\infty}$ taking values in **state space** $\mathcal{S} = (\mathcal{X}, \mathcal{F})$ and with

$$\mathbb{P}\left(X_{t+1} \in \cdot \,|\, X_1, \ldots, X_t\right) = \mathbb{P}\left(X_{t+1} \in \cdot \,|\, X_t\right) = \mu(X_t, \cdot) \qquad \text{almost surely,}$$

where μ is a probability kernel from $(\mathcal{X}, \mathcal{F})$ to $(\mathcal{X}, \mathcal{F})$ and we assume that $\mathbb{P}\left(X_1 \in \cdot\right) = \mu_0(\cdot)$ for some measure μ_0 on $(\mathcal{X}, \mathcal{F})$.

> The word 'homogeneous' refers to the fact that the probability kernel does not change with time. Accordingly, sometimes one writes 'time homogeneous' instead of homogeneous. The reader can no doubt see how to define a Markov chain where μ depends on t, though doing so is purely cosmetic since the state space can always be augmented to include a time component.

Note that if $\mu(x \,|\, \cdot) = \mu_0(\cdot)$ for all $x \in \mathcal{X}$, then Theorem 3.3 is yet another way to prove the existence of an infinite sequence of independent and identically distributed random variables. The basic questions in Markov chains resolve around understanding the evolution of X_t in terms of the probability kernel. For example, assuming that $\Omega_t = \Omega_1$ for all $t \in \mathbb{N}^+$, does the law of X_t converge to some fixed distribution as $t \to \infty$, and if so, how fast is this convergence? For now we make do with the definitions, but in the special case that $\mathcal{X}$ is finite, we will discuss some of these topics much later in Chapters 37 and 38.

3.3 Martingales and Stopping Times

Let $X_1, X_2, \dots$ be a sequence of random variables on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathbb{F} = (\mathcal{F}_t)_{t=1}^n$ a filtration of $\mathcal{F}$ and where we allow $n = \infty$. Recall that the sequence $(X_t)_{t=1}^n$ is $\mathbb{F}$-adapted if X_t is $\mathcal{F}_t$-measurable for all $1 \le t \le n$.

Definition 3.4. A $\mathbb{F}$-adapted sequence of random variables $(X_t)_{t \in \mathbb{N}_+}$ is a $\mathbb{F}$-adapted **martingale** if

(a) $\mathbb{E}[X_t \,|\, \mathcal{F}_{t-1}] = X_{t-1}$ almost surely for all $t \in \{2, 3, \dots\}$; and
(b) X_t is integrable.

If the equality is replaced with a less-than (greater-than), then we call $(X_t)_t$ a **supermartingale** (respectively, a **submartingale**).

The time index t need not run over $\mathbb{N}^+$. Very often t starts at zero instead.

Example 3.5. A gambler repeatedly throws a coin, winning a dollar for each heads and losing a dollar for each tails. Their total winnings over time is a martingale. To model this situation, let $Y_1, Y_2, \dots$ be a sequence of independent Rademacher distributions, which means that $\mathbb{P}(Y_t = 1) = \mathbb{P}(Y_t = -1) = 1/2$. The winnings after t rounds is $S_t = \sum_{s=1}^t Y_s$, which is a martingale adapted to the filtration $(\mathcal{F}_t)_{t=1}^\infty$ given by $\mathcal{F}_t = \sigma(Y_1, \dots, Y_t)$. The definition of super/sub-martingales (the direction of inequality) can be remembered by remembering that the definition favors the casino, not the gambler.

Can a gambler increase its expected winning by stopping cleverly? Precisely, the gambler at the end of round t can decide to stop ($\delta_t = 1$) or continue ($\delta_t = 0$) based on the information available to them. Denoting by $\tau = \min\{t : \delta_t = 1\}$ the time when the gambler stops, the question is whether by a clever choice of $(\delta_t)_{t \in \mathbb{N}}$, $\mathbb{E}[S_\tau]$ can be made positive. Here, $(\delta_t)_{t \in \mathbb{N}}$, a sequence of binary, $\mathbb{F}$-adapted random variables, is called a **stopping rule**, while τ is a stopping time with respect $\mathbb{F}$.

Note that the stopping rule is not allowed to inject additional randomness beyond what is already there in $\mathbb{F}$.

Definition 3.6. Let $\mathbb{F} = (\mathcal{F}_t)_{t \in \mathbb{N}}$ be a filtration. A random variable τ with values in $\mathbb{N} \cup \{\infty\}$ is a **stopping time** with respect to $\mathbb{F}$ if $\mathbb{I}\{\tau \le t\}$ is $\mathcal{F}_t$-measurable for all $t \in \mathbb{N}$. The σ-algebra at stopping time τ is

$$\mathcal{F}_\tau = \{A \in \mathcal{F}_\infty : A \cap \{\tau \le t\} \in \mathcal{F}_t \text{ for all } t\} .$$

> The filtration is usually indicated by writing 'τ is a $\mathbb{F}$-stopping time'. When the underlying filtration is obvious from context, it may be omitted. This is also true for martingales.

Using the interpretation of σ-algebras encoding information, if $(\mathcal{F}_t)_t$ is thought of as the knowledge available at time t, $\mathcal{F}_\tau$ is the information available at the random time τ. Exercise 3.7 asks you to explore properties of stopped σ-algebras; amongst other things, it asks you to show that $\mathcal{F}_\tau$ is in fact a σ-algebra.

Example 3.7. In the gambler example, the first time when the gambler's winnings hits 100 is a stopping time: $\tau = \min\{t : S_t = 100\}$. On the other hand, $\tau = \min\{t : S_{t+1} = -1\}$ is not a stopping time because $\mathbb{I}\{\tau = t\}$ is not $\mathcal{F}_t$-measurable.

Whether or not $\mathbb{E}[S_\tau]$ can be made positive by a clever choice of a stopping time τ is answered in the negative by a fundamental theorem of Doob:

Theorem 3.8 (Doob's optional stopping). *Let $\mathbb{F} = (\mathcal{F}_t)_{t\in\mathbb{N}}$ be a filtration and $(X_t)_{t\in\mathbb{N}}$ be an $\mathbb{F}$-adapted martingale and τ an $\mathbb{F}$-stopping time such that at least one of the following holds:*

(a) *There exists an $n \in \mathbb{N}$ such that $\mathbb{P}(\tau > n) = 0$.*
(b) *$\mathbb{E}[\tau] < \infty$, and there exists a constant $c \in \mathbb{R}$ such that for all $t \in \mathbb{N}$, $\mathbb{E}[|X_{t+1} - X_t| \,|\, \mathcal{F}_t] \le c$ almost surely on the event that $\tau > t$.*
(c) *There exists a constant c such that $|X_{t\wedge\tau}| \le c$ almost surely for all $t \in \mathbb{N}$.*

Then X_τ is almost surely well defined, and $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$. Furthermore, when (X_t) is a super/sub-martingale rather than a martingale, then equality is replaced with less/greater-than, respectively.

The theorem implies that if S_τ is almost-surely well defined then either $\mathbb{E}[\tau] = \infty$ or $\mathbb{E}[S_\tau] = 0$. Gamblers trying to outsmart the casino would need to live a very long life! One application of Doob's optional stopping theorem is a useful and a priori surprising generalisation of Markov's inequality to non-negative supermartingales.

Theorem 3.9 (Maximal inequality). *Let $(X_t)_{t=0}^{\infty}$ be a supermartingale with $X_t \ge 0$ almost surely for all t. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\sup_{t\in\mathbb{N}} X_t \ge \varepsilon\right) \le \frac{\mathbb{E}[X_0]}{\varepsilon}.$$

Proof Let A_n be the event that $\sup_{t\le n} X_t \ge \varepsilon$ and $\tau = (n+1) \wedge \min\{t \le n : X_t \ge \varepsilon\}$, where the minimum of an empty set is assumed to be infinite so that $\tau = n+1$ if $X_t < \varepsilon$ for all $0 \le t \le n$. Clearly τ is a stopping time and $\mathbb{P}(\tau \le n+1) = 1$. Then by Theorem 3.8 and elementary calculation,

$$\mathbb{E}[X_0] \ge \mathbb{E}[X_\tau] \ge \mathbb{E}[X_\tau \mathbb{I}\{\tau \le n\}] \ge \mathbb{E}[\varepsilon \mathbb{I}\{\tau \le n\}] = \varepsilon \mathbb{P}(\tau \le n) = \varepsilon \mathbb{P}(A_n),$$

where the second inequality uses the definition of the stopping time and the non-negativity of the supermartingale. Rearranging shows that $\mathbb{P}(A_n) \le \mathbb{E}[X_0]/\varepsilon$ for all $n \in \mathbb{N}$. Since $A_1 \subseteq A_2 \subseteq \ldots$, it follows that $\mathbb{P}(\sup_{t\in\mathbb{N}} X_t \ge \varepsilon) = \mathbb{P}(\cup_{n\in\mathbb{N}} A_n) \le \mathbb{E}[X_0]/\varepsilon$. □

Markov's inequality (which we will cover in the next chapter) combined with the definition of a supermartingale shows that

$$\mathbb{P}\left(X_n \geq \varepsilon\right) \leq \frac{\mathbb{E}[X_0]}{\varepsilon}. \tag{3.2}$$

In fact, in the above we have effectively applied Markov's inequality to the random variable X_τ (the need for the proof arises when the conditions of Doob's optional sampling theorem are *not* met). The maximal inequality is a strict improvement over Eq. (3.2) by replacing X_n with $\sup_{t\in\mathbb{N}} X_t$ at no cost whatsoever.

A similar theorem holds for submartingales. You will provide a proof in Exercise 3.8.

THEOREM 3.10. *Let $(X_t)_{t=0}^n$ be a submartingale with $X_t \geq 0$ almost surely for all t. Then for any $\varepsilon > 0$,*

$$\mathbb{P}\left(\max_{t\in\{0,1,\dots,n\}} X_t \geq \varepsilon\right) \leq \frac{\mathbb{E}[X_n]}{\varepsilon}.$$

3.4 Notes

1 Some authors include in the definition of a stopping time τ that $\mathbb{P}\left(\tau < \infty\right) = 1$ and call random times without this property **Markov times**. We do *not* adopt this convention and allow stopping times to be infinite with non-zero probability. Stopping times are also called **optional times**.

2 There are several notations for probability kernels depending on the application. The following are commonly seen and equivalent: $K(x, A) = K(A \mid x) = K_x(A)$. For example, in statistics a parametric family is often given by $\{\mathbb{P}_\theta : \theta \in \Theta\}$, where Θ is the parameter space and $\mathbb{P}_\theta$ is a measure on some measurable space $(\Omega, \mathcal{F})$. This notation is often more convenient than writing $\mathbb{P}(\theta, \cdot)$. In Bayesian statistics the posterior is a probability kernel from the observation space to the parameter space, and this is often written as $\mathbb{P}(\cdot \mid x)$.

3 There is some disagreement about whether or not a Markov chain on an uncountable state space should instead be called a **Markov process**. In this book we use Markov chain for arbitrary state spaces and discrete time. When time is continuous (which it never is in this book), there is general agreement that 'process' is more appropriate. For more history on this debate, see [Meyn and Tweedie, 2012, preface].

4 A topological space $\mathcal{X}$ is **Polish** if it is separable and there exists a metric d that is compatible with the topology that makes $(\mathcal{X}, d)$ a complete metric space. All Polish spaces are Borel spaces. We follow Kallenberg [2002], but many authors use **standard Borel space** rather than Borel space, and define it as the σ-algebra generated by the open sets of a Polish space.

5 In Theorem 3.2 it was assumed that each μ_n was defined on a Borel space. No such assumption was required for Theorem 3.3, however. One can derive Theorem 3.2 from Theorem 3.3 by using the existence of regular conditional probability measures when conditioning on random elements taking values in a Borel space (see the next note). Topological assumptions often creep into foundational questions relating to the existence of probability measures satisfying certain conditions, and pathological examples show these assumptions cannot be removed completely. Luckily, in this book we have no reason to consider random elements that do not take values in a Borel space.

(b) If $\tau = k$ for some $k \geq 1$, then $\mathcal{F}_\tau = \mathcal{F}_k$.

(c) If $\tau_1 \leq \tau_2$, then $\mathcal{F}_{\tau_1} \subset \mathcal{F}_{\tau_2}$.

(d) τ is $\mathcal{F}_\tau$-measurable.

(e) If (X_t) is $\mathbb{F}$-adapted, then X_τ is $\mathcal{F}_\tau$-measurable.

(f) $\mathcal{F}_\tau$ is the smallest σ-algebra such that all $\mathbb{F}$-adapted sequences (X_t) satisfy X_τ is $\mathcal{F}_\tau$-measurable.

3.8 Prove Theorem 3.10.

3.9 (Decomposing joint distributions) Let X and Y be random elements on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ taking values in measurable spaces $\mathcal{X}$ and $\mathcal{Y}$ respectively and assume that $\mathcal{X}$ is Borel. Show that $\mathbb{P}_{(X,Y)} = \mathbb{P}_Y \otimes \mathbb{P}_{X|Y}$ where $\mathbb{P}_{X|Y}$ denotes a regular conditional distribution of X and Y (the existence of which is guaranteed by Theorem 3.11).

4 Stochastic Bandits

The goal of this chapter is to formally introduce stochastic bandits. The model introduced here provides the foundation for the remaining chapters that treat stochastic bandits. While the topic seems a bit mundane, it is important to be clear about the assumptions and definitions. The chapter also introduces and motivates the learning objectives, and especially the regret. Besides the definitions, the main result in this chapter is the regret decomposition, which is presented in Section 4.5.

4.1 Core Assumptions

A **stochastic bandit** is a collection of distributions $\nu = (P_a : a \in \mathcal{A})$, where $\mathcal{A}$ is the set of available actions. The learner and the environment interact sequentially over n rounds. In each round $t \in \{1, \dots, n\}$, the learner chooses an action $A_t \in \mathcal{A}$, which is fed to the environment. The environment then samples a reward $X_t \in \mathbb{R}$ from distribution P_{A_t} and reveals X_t to the learner. The interaction between the learner (or policy) and environment induces a probability measure on the sequence of outcomes $A_1, X_1, A_2, X_2, \dots, A_n, X_n$. Usually the horizon n is finite, but sometimes we allow the interaction to continue indefinitely ($n = \infty$). The sequence of outcomes should satisfy the following assumptions:

(a) The conditional distribution of reward X_t given $A_1, X_1, \dots, A_{t-1}, X_{t-1}, A_t$ is P_{A_t}, which captures the intuition that the environment samples X_t from P_{A_t} in round t.

(b) The conditional law of action A_t given $A_1, X_1, \dots, A_{t-1}, X_{t-1}$ is $\pi_t(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$, where $\pi_1, \pi_2, \dots$ is a sequence of probability kernels that characterise the learner. The most important element of this assumption is the intuitive fact that the learner cannot use the future observations in current decisions.

A mathematician might ask whether there even exists a probability space carrying these random elements such that (a) and (b) hold. Specific constructions showing this in the affirmative are given in Section 4.6. These constructions are also valuable because they teach us important lessons about equivalent models. For now, however, we move on.

4.2 The Learning Objective

The learner's goal is to maximise the total reward $S_n = \sum_{t=1}^{n} X_t$, which is a random quantity that depends on the actions of the learner and the rewards sampled by the environment. This is not an optimisation problem for three reasons:

1 What is the value of n for which we are maximising? Occasionally prior knowledge of the horizon is reasonable, but very often the learner does not know ahead of time how many rounds are to be played.
2 The cumulative reward is a random quantity. Even if the reward distributions were known, then we require a measure of utility on distributions of S_n.
3 The learner does not know the distributions that govern the rewards for each arm.

Of these points, the last is fundamental to the bandit problem and is discussed in the next section. The lack of knowledge of the horizon is usually not a serious issue. Generally speaking it is possible to first design a policy assuming the horizon is known and then adapt it to account for the unknown horizon while proving that the loss in performance is minimal. This is almost always quite easy, and there exist generic approaches for making the conversion.

Assigning a utility to distributions of S_n is more challenging. Suppose that S_n is the revenue of your company. Fig. 4.1 shows the distribution of S_n for two different learners; call them A and B. Suppose you can choose between learners A and B. Which one would you choose? One choice is to go with the learner whose reward distribution has the larger expected value. This will be our default choice for stochastic bandits, but it bears remembering that there are other considerations, including the variance or tail behaviour of the cumulative reward, which we will discuss occasionally. In particular, in the situation shown on in Fig. 4.1, learner B achieves a higher expected reward than A. However B has a reasonable probability of earning less than the least amount that A can earn, so a risk-sensitive user may prefer learner A.

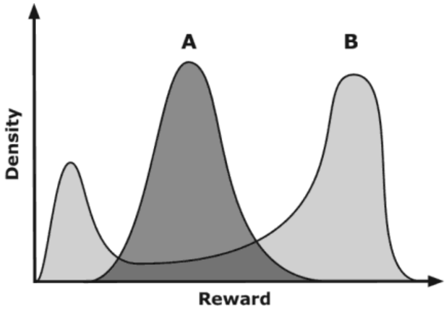


Figure 4.1 Alternative revenue distributions

4.3 Knowledge and Environment Classes

Even if the horizon is known in advance and we commit to maximising the expected value of S_n, there is still the problem that the bandit instance $\nu = (P_a : a \in \mathcal{A})$ is unknown. A policy that maximises the expectation of S_n for one bandit instance may behave quite badly on another. The learner usually has partial information about ν, which we represent by defining a set of bandits $\mathcal{E}$ for which $\nu \in \mathcal{E}$ is guaranteed. The set $\mathcal{E}$ is called the **environment class**. We distinguish between **structured** and **unstructured** bandits.

Unstructured Bandits

An environment class $\mathcal{E}$ is unstructured if $\mathcal{A}$ is finite and there exist sets of distributions $\mathcal{M}_a$ for each $a \in \mathcal{A}$ such that

$$\mathcal{E} = \{\nu = (P_a : a \in \mathcal{A}) : P_a \in \mathcal{M}_a \text{ for all } a \in \mathcal{A}\},$$

Table 4.1 Typical environment classes for stochastic bandits. Supp(P) is the (topological) support of distribution P. The kurtosis of a random variable X is a measure of its tail behaviour and is defined by $\mathbb{E}[(X-\mathbb{E}[X])^4]/\mathbb{V}[X]^2$. Subgaussian distributions have similar properties to the Gaussian and will be defined in Chapter 5.

Name	**Symbol**	**Definition**
Bernoulli	$\mathcal{E}^k_{\mathcal{B}}$	$\{(\mathcal{B}(\mu_i))_i : \mu \in [0,1]^k\}$
Uniform	$\mathcal{E}^k_{\mathcal{U}}$	$\{(\mathcal{U}(a_i,b_i))_i : a, b \in \mathbb{R}^k$ with $a_i \leq b_i$ for all $i\}$
Gaussian (known var.)	$\mathcal{E}^k_{\mathcal{N}}(\sigma^2)$	$\{(\mathcal{N}(\mu_i,\sigma^2))_i : \mu \in \mathbb{R}^k\}$
Gaussian (unknown var.)	$\mathcal{E}^k_{\mathcal{N}}$	$\{(\mathcal{N}(\mu_i,\sigma_i^2))_i : \mu \in \mathbb{R}^k$ and $\sigma^2 \in [0,\infty)^k\}$
Finite variance	$\mathcal{E}^k_{\mathbb{V}}(\sigma^2)$	$\{(P_i)_i : \mathbb{V}_{X\sim P_i}[X] \leq \sigma^2$ for all $i\}$
Finite kurtosis	$\mathcal{E}^k_{\mathrm{Kurt}}(\kappa)$	$\{(P_i)_i : \mathrm{Kurt}_{X\sim P_i}[X] \leq \kappa$ for all $i\}$
Bounded support	$\mathcal{E}^k_{[a,b]}$	$\{(P_i)_i : \mathrm{Supp}(P_i) \subseteq [a,b]\}$
Subgaussian	$\mathcal{E}^k_{\mathrm{SG}}(\sigma^2)$	$\{(P_i)_i : P_i$ is σ-subgaussian for all $i\}$

or, in short, $\mathcal{E} = \times_{a\in\mathcal{A}}\mathcal{M}_a$. The product structure means that by playing action a the learner cannot deduce anything about the distributions of actions $b \neq a$.

Some typical choices of unstructured bandits are listed in Table 4.1. Of course, these are not the only choices, and the reader can no doubt find ways to construct more, e.g. by allowing some arms to be Bernoulli and some Gaussian, or have rewards being exponentially distributed, or Gumbel distributed, or belonging to your favourite (non-)parametric family.

The Bernoulli, Gaussian and uniform distributions are often used as examples for illustrating some specific property of learning in stochastic bandit problems. The Bernoulli distribution is actually a natural choice. Think of applications like maximising click-through rates in a web-based environment. A bandit problem is often called a 'distribution bandit', where 'distribution' is replaced by the underlying distribution from which the pay-offs are sampled. Some examples are: Gaussian bandit, Bernoulli bandit or subgaussian bandit. Similarly we say 'bandits with X', where 'X' is a property of the underlying distribution from which the pay-offs are sampled. For example, we can talk about bandits with finite variance, meaning the bandit environment where the a priori knowledge of the learner is that all pay-off distributions are such that their underlying variance is finite.

Some environment classes, like Bernoulli bandits, are **parametric**, while others, like subgaussian bandits, are **non-parametric**. The distinction is the number of degrees of freedom needed to describe an element of the environment class. When the number of degrees of freedom is finite, it is parametric, and otherwise it is non-parametric. Of course, if a learner is designed for a specific environment class $\mathcal{E}$, then we might expect that it has good performance on all bandits $\nu \in \mathcal{E}$. Some environment classes are subsets of other classes. For example, Bernoulli bandits are a special case of bandits with a finite variance, or bandits with bounded support. Something to keep in mind is that we expect that it will be harder to achieve a good performance in a larger class. In a way, the theory of finite-armed stochastic bandits tries to quantify this expectation in a rigourous fashion.

Structured Bandits

Environment classes that are not unstructured are called structured. Relaxing the requirement that the environment class is a product set makes structured bandit problems much richer than the unstructured set-up. The following examples illustrate the flexibility.

EXAMPLE 4.1. Let $\mathcal{A} = \{1, 2\}$ and $\mathcal{E} = \{(\mathcal{B}(\theta), \mathcal{B}(1 - \theta)) : \theta \in [0, 1]\}$. In this environment class, the learner does not know the mean of either arm, but can learn the mean of both arms by playing just one. The knowledge of this structure dramatically changes the difficulty of learning in this problem.

EXAMPLE 4.2 (Stochastic linear bandit). Let $\mathcal{A} \subset \mathbb{R}^d$ and $\theta \in \mathbb{R}^d$ and

$$\nu_\theta = (\mathcal{N}(\langle a, \theta \rangle, 1) : a \in \mathcal{A}) \qquad \text{and } \mathcal{E} = \{\nu_\theta : \theta \in \mathbb{R}^d\}.$$

In this environment class, the reward of an action is Gaussian, and its mean is given by the inner product between the action and some unknown parameter. Notice that even if $\mathcal{A}$ is extremely large, the learner can deduce the true environment by playing just d actions that span $\mathbb{R}^d$.

EXAMPLE 4.3. Consider an undirected graph G with vertices $V = \{1, \ldots, |V|\}$ and edges $E = \{1, \ldots, |E|\}$. In each round the learner chooses a path from vertex 1 to vertex $|V|$. Then each edge $e \in [E]$ is removed from the graph with probability $1 - \theta_e$ for unknown $\theta \in [0, 1]^{|E|}$. The learner succeeds in reaching their destination if all the edges in their chosen path are present. This problem can be formalised by letting $\mathcal{A}$ be the set of paths and

$$\nu_\theta = \left(\mathcal{B}\left(\prod_{e \in a} \theta_e\right) : a \in \mathcal{A}\right) \qquad \text{and} \qquad \mathcal{E} = \{\nu_\theta : \theta \in [0, 1]^{|E|}\}.$$

☞ An important feature of structured bandits is that the learner can often obtain information about some actions while never playing them.

4.4 The Regret

In Chapter 1 we informally defined the regret as being the deficit suffered by the learner relative to the optimal policy. Let $\nu = (P_a : a \in \mathcal{A})$ be a stochastic bandit and define

$$\mu_a(\nu) = \int_{-\infty}^{\infty} x \, dP_a(x).$$

Then let $\mu^*(\nu) = \max_{a \in \mathcal{A}} \mu_a(\nu)$ be the largest mean of all the arms.

☞ We assume throughout that $\mu_a(\nu)$ exists and is finite for all actions and that $\operatorname{argmax}_{a \in \mathcal{A}} \mu_a(\nu)$ is non-empty. The latter assumption could be relaxed by carefully adapting all arguments using nearly optimal actions, but in practice this is never required.

The regret of policy π on bandit instance ν is

$$R_n(\pi, \nu) = n\mu^*(\nu) - \mathbb{E}\left[\sum_{t=1}^{n} X_t\right], \tag{4.1}$$

> Lemma 4.5 tells us that a learner should aim to use an arm with a larger suboptimality gap proportionally fewer times.

Note that the suboptimality gap for optimal arm(s) is zero.

Proof of Lemma 4.5 Since R_n is based on summing over rounds, and the right-hand side of the lemma statement is based on summing over actions, to convert one sum into the other one, we introduce indicators. In particular, note that for any fixed t we have $\sum_{a \in \mathcal{A}} \mathbb{I}\{A_t = a\} = 1$. Hence $S_n = \sum_t X_t = \sum_t \sum_a X_t \mathbb{I}\{A_t = a\}$, and thus

$$R_n = n\mu^* - \mathbb{E}[S_n] = \sum_{a \in \mathcal{A}} \sum_{t=1}^{n} \mathbb{E}[(\mu^* - X_t)\mathbb{I}\{A_t = a\}]. \tag{4.6}$$

The expected reward in round t conditioned on A_t is μ_{A_t}, which means that

$$\begin{aligned}
\mathbb{E}[(\mu^* - X_t)\mathbb{I}\{A_t = a\} \mid A_t] &= \mathbb{I}\{A_t = a\}\,\mathbb{E}[\mu^* - X_t \mid A_t] \\
&= \mathbb{I}\{A_t = a\}\,(\mu^* - \mu_{A_t}) \\
&= \mathbb{I}\{A_t = a\}\,(\mu^* - \mu_a) \\
&= \mathbb{I}\{A_t = a\}\,\Delta_a.
\end{aligned}$$

The result is completed by plugging this into Eq. (4.6) and using the definition of $T_a(n)$. □

The argument fails when $\mathcal{A}$ is uncountable because you cannot introduce the sum over actions. Of course the solution is to use an integral, but for this we need to assume $(\mathcal{A}, \mathcal{G})$ is a measurable space. Given a bandit ν and policy π define measure G on $(\mathcal{A}, \mathcal{G})$ by

$$G(U) = \mathbb{E}\left[\sum_{t=1}^{n} \mathbb{I}\{A_t \in U\}\right],$$

where the expectation is taken with respect to the measure on outcomes induced by the interaction of π and ν.

Lemma 4.6. *Provided that everything is well defined and appropriately measurable,*

$$R_n = \mathbb{E}\left[\sum_{t=1}^{n} \Delta_{A_t}\right] = \int_{\mathcal{A}} \Delta_a \, dG(a).$$

For those worried about how to ensure everything is well defined, see Section 4.7.

4.6 The Canonical Bandit Model (🦘)

In most cases the underlying probability space that supports the random rewards and actions is never mentioned. Occasionally, however, it becomes convenient to choose a specific probability space, which we call the **canonical bandit model**.

Finite Horizon

Let $n \in \mathbb{N}$ be the horizon. A policy and bandit interact to produce the outcome, which is the tuple of random variables $H_n = (A_1, X_1, \dots, A_n, X_n)$. The first step towards constructing a probability space that carries these random variables is to choose the measurable space. For each $t \in [n]$, let $\Omega_t = ([k] \times \mathbb{R})^t \subset \mathbb{R}^{2t}$ and $\mathcal{F}_t = \mathfrak{B}(\Omega_t)$. The random variables $A_1, X_1, \dots, A_n, X_n$ that make up the outcome are defined by their coordinate projections:

$$A_t(a_1, x_1, \dots, a_n, x_n) = a_t \qquad \text{and} \qquad X_t(a_1, x_1, \dots, a_n, x_n) = x_t.$$

The probability measure on $(\Omega_n, \mathcal{F}_n)$ depends on both the environment and the policy. Our informal definition of a policy is not quite sufficient now.

Definition 4.7. A **policy** π is a sequence $(\pi_t)_{t=1}^n$, where π_t is a probability kernel from $(\Omega_{t-1}, \mathcal{F}_{t-1})$ to $([k], 2^{[k]})$. Since $[k]$ is discrete, we adopt the notational convention that for $i \in [k]$,

$$\pi_t(i \mid a_1, x_1, \dots, a_{t-1}, x_{t-1}) = \pi_t(\{i\} \mid a_1, x_1, \dots, a_{t-1}, x_{t-1}).$$

Let $\nu = (P_i)_{i=1}^k$ be a stochastic bandit where each P_i is a probability measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. We want to define a probability measure on $(\Omega_n, \mathcal{F}_n)$ that respects our understanding of the sequential nature of the interaction between the learner and a stationary stochastic bandit. Since we only care about the law of the random variables (X_t) and (A_t), the easiest way to enforce this is to directly list our expectations, which are

(a) the conditional distribution of action A_t given $A_1, X_1, \dots, A_{t-1}, X_{t-1}$ is $\pi_t(\cdot \mid A_1, X_1, \dots, A_{t-1}, X_{t-1})$ almost surely.
(b) the conditional distribution of reward X_t given $A_1, X_1, \dots, A_t$ is P_{A_t} almost surely.

The sufficiency of these assumptions is asserted by the following proposition, which we ask you to prove in Exercise 4.2.

Proposition 4.8. *Suppose that $\mathbb{P}$ and $\mathbb{Q}$ are probability measures on an arbitrary measurable space $(\Omega, \mathcal{F})$ and $A_1, X_1, \dots, A_n, X_n$ are random variables on Ω, where $A_t \in [k]$ and $X_t \in \mathbb{R}$. If both $\mathbb{P}$ and $\mathbb{Q}$ satisfy* (a) *and* (b)*, then the law of the outcome under $\mathbb{P}$ is the same as under $\mathbb{Q}$:*

$$\mathbb{P}_{A_1, X_1, \dots, A_n, X_n} = \mathbb{Q}_{A_1, X_1, \dots, A_n, X_n}.$$

Next we construct a probability measure on $(\Omega_n, \mathcal{F}_n)$ that satisfies (a) and (b). To emphasise that what follows is intuitively not complicated, imagine that $X_t \in \{0, 1\}$ is Bernoulli, which means the set of possible outcomes is finite and we can define the measure in terms of a distribution. Let $p_i(0) = P_i(\{0\})$ and $p_i(1) = 1 - p_i(0)$ and define

$$p_{\nu\pi}(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^n \pi(a_t \mid a_1, x_1, \dots, a_{t-1}, x_{t-1}) p_{a_t}(x_t).$$

The reader can check that $p_{\nu\pi}$ is a distribution on $([k] \times \{0, 1\})^n$ and that the associated measure satisfies (a) and (b) above. Making this argument rigourous when (P_i) are not discrete requires the use of Radon–Nikodym derivatives. Let λ be a σ-finite measure on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ for which P_i is absolutely continuous with respect to λ for all i. Next, let $p_i = dP_i/d\lambda$ be the Radon–Nikodym derivative of P_i with respect to λ, which is a function

$p_i : \mathbb{R} \to \mathbb{R}$ such that $\int_B p_i \, d\lambda = P_i(B)$ for all $B \in \mathfrak{B}(\mathbb{R})$. Letting ρ be the counting measure with $\rho(B) = |B|$, the density $p_{\nu\pi} : \Omega \to \mathbb{R}$ can now be defined with respect to the product measure $(\rho \times \lambda)^n$ by

$$p_{\nu\pi}(a_1, x_1, \dots, a_n, x_n) = \prod_{t=1}^{n} \pi(a_t \,|\, a_1, x_1, \dots, a_{t-1}, x_{t-1}) p_{a_t}(x_t). \tag{4.7}$$

The reader can again check (more abstractly) that (a) and (b) are satisfied by the probability measure $\mathbb{P}_{\nu\pi}$ defined by

$$\mathbb{P}_{\nu\pi}(B) = \int_B p_{\nu\pi}(\omega)(\rho \times \lambda)^n(d\omega) \qquad \text{for all } B \in \mathcal{F}_n.$$

It is important to emphasise that this choice of $(\Omega_n, \mathcal{F}_n, \mathbb{P}_{\nu\pi})$ is not unique. Instead, all that this shows is that a suitable probability space does exist. Furthermore, if some quantity of interest depends on the law of H_n, by Proposition 4.8, there is no loss in generality in choosing $(\Omega_n, \mathcal{F}_n, \mathbb{P}_{\nu\pi})$ as the probability space.

A choice of λ such that $P_i \ll \lambda$ for all i always exists since $\lambda = \sum_{i=1}^{k} P_i$ satisfies this condition. For direct calculations, another choice is usually more convenient, e.g. the counting measure when (P_i) are discrete and the Lebesgue measure for continuous (P_i).

There is another way to define the probability space, which can be useful. Define a collection of independent random variables $(X_{si})_{s \in [n], i \in [k]}$ such that the law of X_{ti} is P_i. By Theorem 2.4 these random variables may be defined on $(\Omega, \mathcal{F})$, where $\Omega = \mathbb{R}^{nk}$ and $\mathcal{F} = \mathfrak{B}(\mathbb{R}^{nk})$. Then let $X_t = X_{tA_t}$, where the actions A_t are $\mathcal{F}_{t-1}$-measurable with $\mathcal{F}_{t-1} = \sigma(A_1, X_1, \dots, A_{t-1}, X_{t-1})$. We call this the **random table model**. Yet another way is to define $(X_{si})_{s,i}$ as above but let $X_t = X_{T_{A_t}(t), A_t}$. This corresponds to sampling a **stack of rewards** for each arm at the beginning of the game, giving rise to the reward-stack model. Each time the learner chooses an action, they receive the reward on top of the stack. All of these models are convenient from time to time. The important thing is that it does not matter which model we choose because the quantity of ultimate interest (usually the regret) only depends on the law of $A_1, X_1, \dots, A_n, X_n$, and this is the same for all choices.

Infinite Horizon

We never need the canonical bandit model for the case that $n = \infty$. It is comforting to know, however, that there does exist a probability space $(\Omega, \mathcal{F}, \mathbb{P}_{\nu\pi})$ and infinite sequences of random variables $X_1, X_2, \dots$ and $A_1, A_2, \dots$ satisfying (a) and (b). The result follows directly from the theorem of Ionescu–Tulcea (Theorem 3.3).

4.7 The Canonical Bandit Model for Uncountable Action Sets (🦘)

For uncountable action sets, a little more machinery is necessary to make things rigourous. The first requirement is that the action set must be a measurable space $(\mathcal{A}, \mathcal{G})$ and the collection of distribution $\nu = (P_a : a \in \mathcal{A})$ that defines a bandit environment must be a probability kernel from $(\mathcal{A}, \mathcal{G})$ to $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$. A policy is a sequence $(\pi_t)_{t=1}^n$, where π_t is a probability kernel from $(\Omega_{t-1}, \mathcal{F}_{t-1})$ to $(\mathcal{A}, \mathcal{G})$ with

$$\Omega_t = \prod_{s=1}^t (\mathcal{A} \times \mathbb{R}) \qquad \text{and} \qquad \mathcal{F}_t = \bigotimes_{s=1}^t (\mathcal{G} \otimes \mathfrak{B}(\mathbb{R}))\,.$$

The canonical bandit model is the probability measure $\mathbb{P}_{\nu\pi}$ on $(\Omega_n, \mathcal{F}_n)$ obtained by taking the product of the probability kernels $\pi_1, P_1, \ldots \pi_n, P_n$ and using Ionescu–Tulcea (Theorem 3.3), where P_t is the probability kernel from $(\Omega_{t-1} \times \mathcal{A}, \mathcal{F}_t \otimes \mathcal{G})$ to $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ given by $P_t(\,\cdot \mid a_1, x_1, \ldots, a_{t-1}, x_{t-1}, a_t) = P_{a_t}(\cdot)$.

We did not define $\mathbb{P}_{\nu\pi}$ in terms of a density because there may not exist a common dominating measure for either $(P_a : a \in \mathcal{A})$ or the policy. When such measures exist, as they usually do, then $\mathbb{P}_{\nu\pi}$ may be defined in terms of a density in the same manner as the previous section.

You will check in Exercise 4.5 that the assumptions on ν and π in this section are sufficient to ensure the quantities in Lemma 4.6 are well defined and that Proposition 4.8 continues to hold in this setting without modification. Finally, in none of the definitions above do we require that n be finite.

4.8 Notes

1 It is not obvious why the expected value is a good summary of the reward distribution. Decision makers who base their decisions on expected values are called risk-neutral. In the example shown on the figure above, a risk-averse decision maker may actually prefer the distribution labelled as A because occasionally distribution B may incur a very small (even negative) reward. Risk-seeking decision makers, if they exist at all, would prefer distributions with occasional large rewards to distributions that give mediocre rewards only. There is a formal theory of what makes a decision maker rational (a decision maker in a nutshell is rational if they do not contradict themself). Rational decision makers compare stochastic alternatives based on the alternatives' expected utilities, according to the von-Neumann–Morgenstern utility theorem. Humans are known not to do this. We are irrational. No surprise here.

2 The study of utility and risk has a long history, going right back to (at least) the beginning of probability [Bernoulli, 1954, translated from the original Latin, 1738]. The research can broadly be categorised into two branches. The first deals with describing how people actually make choices (**descriptive theories**), while the second is devoted to characterising how a rational decision maker should make decisions (**prescriptive theories**). A notable example of the former type is 'prospect theory' [Kahneman and Tversky, 1979], which models how people handle probabilities (especially small ones) and earned Daniel Khaneman a Nobel Prize (after the death of his long-time collaborator, Amos Tversky). Further descriptive theories concerned with alternative aspects of human decision-making include bounded rationality, choice strategies, recognition-primed decision-making and image theory [Adelman, 2013].

3 The most famous example of a prescriptive theory is the von Neumann–Morgenstern expected utility theorem, which states that under (reasonable) axioms of rational behaviour under uncertainty, a rational decision maker must choose amongst alternatives by computing the expected utility of the outcomes [Neumann and Morgenstern, 1944]. Thus, rational decision makers, under the chosen axioms, differ only in terms of how they assign utility to outcomes (i.e. rewards). Finance is another field where attitudes towards uncertainty and risk are important. Markowitz [1952] argues against expected return as a reasonable metric that investors would use. His argument is based on the (simple) observation that portfolios maximising expected returns will tend to have a single stock only (unless there are multiple stocks with equal expected returns, a rather unlikely outcome). He argues that such a complete lack of diversification is unreasonable. He then proposes that investors should minimise the variance of the portfolio's return subject to a constraint on the portfolio's expected return, leading to the so-called **mean-variance optimal portfolio choice theory**. Under this criteria, portfolios will indeed tend to be diversified (and in a meaningful way: correlations between returns are taken into account). This theory eventually won him a Nobel Prize in economics (shared with two others). Closely related to the mean-variance criterion are the 'value-at-risk' (VaR) and the 'conditional value-at-risk', the latter of which has been introduced and promoted by Rockafellar and Uryasev [2000] due to its superior optimisation properties. The distinction between the prescriptive and descriptive theories is important: human decision makers are in many ways violating rules of rationality in their attitudes towards risk.

4 We defined the regret as an expectation, which makes it unusable in conjunction with measures of risk because the randomness has been eliminated by the expectation. When using a risk measure in a bandit setting, we can either base this on the **random regret** or **pseudo-regret** defined by

$$\hat{R}_n = n\mu^* - \sum_{t=1}^{n} X_t. \qquad \text{(random regret)}$$

$$\bar{R}_n = n\mu^* - \sum_{t=1}^{n} \mu_{A_t}. \qquad \text{(pseudo-regret)}$$

While $\hat{R}_n$ is influenced by the noise $X_t - \mu_{A_t}$ in the rewards, the pseudo-regret filters this out, which arguably makes it a better basis for measuring the 'skill' of a bandit policy. As these random regret measures tend to be highly skewed, using variance to assess risk suffers not only from the problem of penalising upside risk, but also from failing to capture the skew of the distribution.

5 What happens if the distributions of the arms are changing with time? Such bandits are unimaginatively called **non-stationary** bandits. With no assumptions, there is not much to be done. Because of this, it is usual to assume the distributions change infrequently or drift slowly. We'll eventually see that techniques for stationary bandits can be adapted to this set-up (see Chapter 31).

6 The rigourous models introduced in Sections 4.6 and 4.7 are easily extended to more sophisticated settings. For example, the environment sometimes produces side information as well as rewards or the set of available actions may change with time. You are asked to formalise an example in Exercise 4.6.

4.9 Bibliographical Remarks

There is now a huge literature on stochastic bandits, much of which we will discuss in detail in the chapters that follow. The earliest reference that we know of is by Thompson [1933], who proposed an algorithm that forms the basis of many of the currently practical approaches in use today. Thompson

Let $T^*(n) = \sum_{t=1}^n \mathbb{I}\{\mu_{A_t} = \mu^*\}$ be the number of times an optimal arm is chosen. Prove or disprove each of the following statements:

(a) $\lim_{n\to\infty} \mathbb{E}[T^*(n)]/n = 1$.
(b) $\lim_{n\to\infty} \mathbb{P}(\Delta_{A_t} > 0) = 0$.

4.10 (ONE-ARMED BANDITS) Let $\mathcal{M}_1$ be a set of distributions on $(\mathbb{R}, \mathfrak{B}(\mathbb{R}))$ with finite means and $\mathcal{M}_2 = \{\delta_{\mu_2}\}$ be the singleton set with a Dirac at $\mu_2 \in \mathbb{R}$. The set of bandits $\mathcal{E} = \mathcal{M}_1 \times \mathcal{M}_2$ is called a **one-armed bandit** because, although there are two arms, the second arm always yields a known reward of μ_2. A policy $\pi = (\pi_t)_t$ is called a **retirement policy** if once action 2 has been played once, it is played until the end of the game. Precisely, if $a_t = 2$, then

$$\pi_{t+1}(2 \mid a_1, x_1, \dots, a_t, x_t) = 1 \text{ for all } (a_s)_{s=1}^{t-1} \text{ and } (x_s)_{s=1}^t.$$

(a) Let n be fixed and $\pi = (\pi_t)_{t=1}^n$ be any policy. Prove there exists a retirement policy $\pi' = (\pi'_t)_{t=1}^n$ such that for all $\nu \in \mathcal{E}$.

$$R_n(\pi', \nu) \le R_n(\pi, \nu).$$

(b) Let $\mathcal{M}_1 = \{\mathcal{B}(\mu_1) : \mu_1 \in [0,1]\}$ and suppose that $\pi = (\pi_t)_{t=1}^\infty$ is a retirement policy. Prove there exists a bandit $\nu \in \mathcal{E}$ such that

$$\limsup_{n\to\infty} \frac{R_n(\pi, \nu)}{n} > 0.$$

4.11 (FAILURE OF FOLLOW-THE-LEADER (I)) Consider a Bernoulli bandit with two arms and means $\mu_1 = 0.5$ and $\mu_2 = 0.6$.

(a) Using a horizon of $n = 100$, run 1000 simulations of your implementation of follow-the-leader on the Bernoulli bandit above and record the (random) pseudo regret, $n\mu^* - \sum_{t=1}^n \mu_{A_t}$, in each simulation.
(b) Plot the results using a histogram. Your figure should resemble Fig. 4.2.

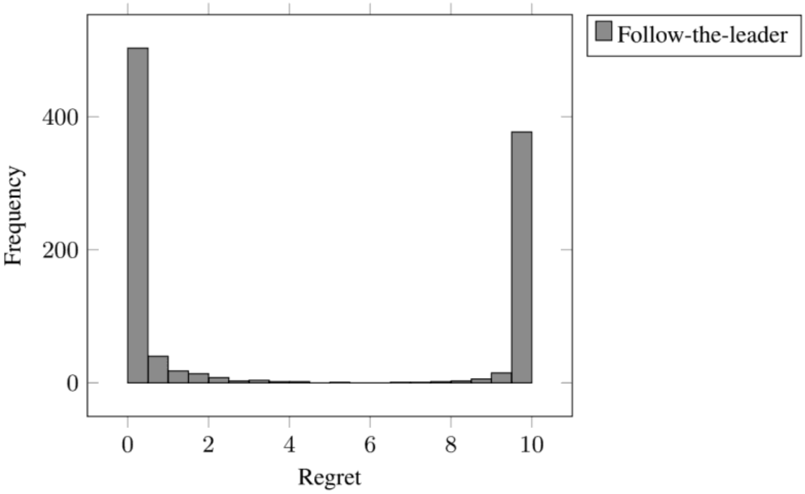


Figure 4.2 Histogram of regret for follow-the-leader over 1000 trials on a Bernoulli bandit with means $\mu_1 = 0.5$, $\mu_2 = 0.6$

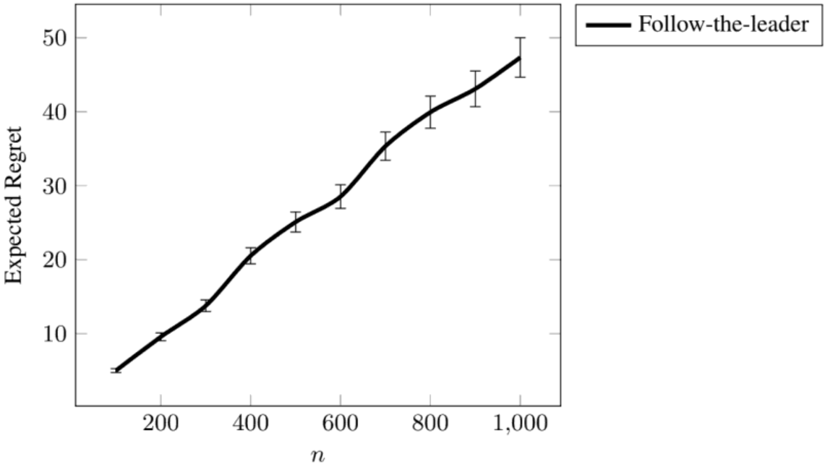


Figure 4.3 The regret for Follow-the-leader over 1000 trials on Bernoulli bandit with means $\mu_1 = 0.5$, $\mu_2 = 0.6$ and horizons ranging from $n = 100$ to $n = 1000$.

(c) Explain the results in the figure.

4.12 (Failure of follow-the-leader (ii)) Consider the same Bernoulli bandit as used in the previous question.

(a) Run 1000 simulations of your implementation of follow-the-leader for each horizon $n \in \{100, 200, 300, \ldots, 1000\}$.
(b) Plot the average regret obtained as a function of n (see Fig. 4.3). Because the average regret is an estimator of the expected regret, you should generally include error bars to indicate the uncertainty in the estimation.
(c) Explain the plot. Do you think follow-the-leader is a good algorithm? Why/why not?

5 Concentration of Measure

Before we can start designing and analysing algorithms, we need one more tool from probability theory, called **concentration of measure**. Recall that the optimal action is the one with the largest mean. Since the mean pay-offs are initially unknown, they must be learned from data. How long does it take to learn about the mean reward of an action? In this section, after introducing the notion of tail probabilities, we look at ways of obtaining upper bounds on them. The main point is to introduce subgaussian random variables and the Cramér–Chernoff exponential tail inequalities, which will play a central role in the design and analysis of the various bandit algorithms.

5.1 Tail Probabilities

Suppose that $X, X_1, X_2, \dots, X_n$ is a sequence of independent and identically distributed random variables, and assume that the mean $\mu = \mathbb{E}[X]$ and variance $\sigma^2 = \mathbb{V}[X]$ exist. Having observed $X_1, X_2, \dots, X_n$, we would like to estimate the common mean μ. The most natural estimator is

$$\hat{\mu} = \frac{1}{n} \sum_{i=1}^{n} X_i,$$

which is called the **sample mean** or **empirical mean**. Linearity of expectation (Proposition 2.6) shows that $\mathbb{E}[\hat{\mu}] = \mu$, which means that $\hat{\mu}$ is an **unbiased** estimator of μ. How far from μ do we expect $\hat{\mu}$ to be? A simple measure of the spread of the distribution of a random variable Z is its variance, $\mathbb{V}[Z] = \mathbb{E}\left[(Z - \mathbb{E}[Z])^2\right]$. A quick calculation using independence shows that

$$\mathbb{V}[\hat{\mu}] = \mathbb{E}\left[(\hat{\mu} - \mu)^2\right] = \frac{\sigma^2}{n}, \tag{5.1}$$

which means that we expect the squared distance between μ and $\hat{\mu}$ to shrink as n grows large at a rate of $1/n$ and scale linearly with the variance of X. While the expected squared error is important, it does not tell us very much about the distribution of the error. To do this we usually analyse the probability that $\hat{\mu}$ overestimates or underestimates μ by more than some value $\varepsilon > 0$. Precisely, how do the following quantities depend on ε?

$$\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) \quad \text{and} \quad \mathbb{P}(\hat{\mu} \leq \mu - \varepsilon).$$

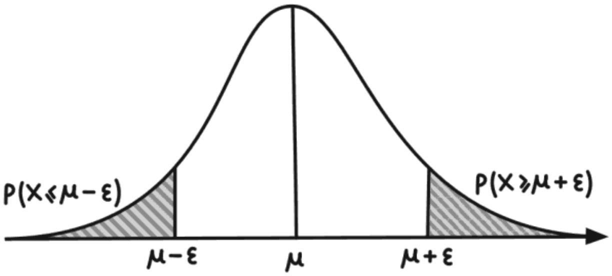


Figure 5.1 The figure shows a probability density, with the tails shaded indicating the regions where X is at least ε away from the mean μ.

The expressions above (as a function of ε) are called the **tail probabilities** of $\hat{\mu} - \mu$ (Fig. 5.1). Specifically, the first is called the upper tail probability and the second the lower tail probability. Analogously, $\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon)$ is called a two-sided tail probability.

5.2 The Inequalities of Markov and Chebyshev

The most straightforward way to bound the tails is by using **Chebyshev's inequality**, which is itself a corollary of **Markov's inequality**. The latter is one of the golden hammers of probability theory, and so we include it for the sake of completeness.

LEMMA 5.1. *For any random variable X and $\varepsilon > 0$, the following holds:*

(a) *(Markov):* $\mathbb{P}(|X| \geq \varepsilon) \leq \dfrac{\mathbb{E}[|X|]}{\varepsilon}$.

(b) *(Chebyshev):* $\mathbb{P}(|X - \mathbb{E}[X]| \geq \varepsilon) \leq \dfrac{\mathbb{V}[X]}{\varepsilon^2}$.

We leave the proof of Lemma 5.1 as an exercise for the reader. By combining (5.1) with Chebyshev's inequality, we can bound the two-sided tail directly in terms of the variance by

$$\mathbb{P}(|\hat{\mu} - \mu| \geq \varepsilon) \leq \frac{\sigma^2}{n\varepsilon^2}. \tag{5.2}$$

This result is nice because it was so easily bought and relied on no assumptions other than the existence of the mean and variance. The downside is that when X is well behaved, the inequality is rather loose. By assuming that higher moments of X exist, Chebyshev's inequality can be improved by applying Markov's inequality to $|\hat{\mu} - \mu|^k$, with the positive integer k to be chosen so that the resulting bound is optimised. This is a bit cumbersome, and thus instead we present the continuous analog of this, known as the Cramér-Chernoff method.

To calibrate our expectations on what improvement to expect relative to Chebyshev's inequality, let us start by recalling the **central limit theorem** (CLT). Let $S_n = \sum_{t=1}^{n}(X_t - \mu)$. The CLT says that under no additional assumptions than the existence of the variance, the limiting distribution of $S_n/\sqrt{n\sigma^2}$ as $n \to \infty$ is a Gaussian with mean zero and unit variance. If $Z \sim \mathcal{N}(0,1)$, then

$$\mathbb{P}(Z \geq u) = \int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx.$$

The integral has no closed-form solution, but is easy to bound:

$$\begin{aligned}\int_u^\infty \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{x^2}{2}\right) dx &\leq \frac{1}{u\sqrt{2\pi}} \int_u^\infty x \exp\left(-\frac{x^2}{2}\right) dx \\ &= \sqrt{\frac{1}{2\pi u^2}} \exp\left(-\frac{u^2}{2}\right), \end{aligned} \tag{5.3}$$

which gives

$$\begin{aligned}\mathbb{P}(\hat{\mu} \geq \mu + \varepsilon) &= \mathbb{P}\left(S_n/\sqrt{\sigma^2 n} \geq \varepsilon\sqrt{n/\sigma^2}\right) \approx \mathbb{P}\left(Z \geq \varepsilon\sqrt{n/\sigma^2}\right) \\ &\leq \sqrt{\frac{\sigma^2}{2\pi n\varepsilon^2}} \exp\left(-\frac{n\varepsilon^2}{2\sigma^2}\right). \end{aligned} \tag{5.4}$$

This is usually much smaller than what we obtained with Chebyshev's inequality (Exercise 5.3). In particular, the bound on the right-hand side of (5.4) decays slightly faster than the negative exponential of $n\varepsilon^2/\sigma^2$, which means that $\hat{\mu}$ rapidly concentrates around its mean.

⚠ An oft-taught rule of thumb is that the CLT provides a reasonable approximation for $n \geq 30$. We advise caution. Suppose that $X_1, \ldots, X_n$ are independent Bernoulli with bias $p = 1/n$. As n tends to infinity the distribution of $\sum_{t=1}^n X_t$ converges to a Poisson distribution with parameter 1, which does not look Gaussian at all.

The asymptotic nature of the CLT makes it unsuitable for designing bandit algorithms. In the next section, we derive finite-time analogs, which are only possible by making additional assumptions.

5.3 The Cramér-Chernoff Method and Subgaussian Random Variables

For the sake of moving rapidly towards bandits, we start with a straightforward and relatively fundamental assumption on the distribution of X, known as the **subgaussian** assumption.

DEFINITION 5.2 (Subgaussianity). A random variable X is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that $\mathbb{E}[\exp(\lambda X)] \leq \exp\left(\lambda^2\sigma^2/2\right)$.

An alternative way to express the subgaussianity condition uses the **moment-generating function** of X, which is a function $M_X : \mathbb{R} \to \mathbb{R}$ defined by $M_X(\lambda) = \mathbb{E}[\exp(\lambda X)]$. The condition in the definition can be written as

$$\psi_X(\lambda) = \log M_X(\lambda) \leq \frac{1}{2}\lambda^2\sigma^2 \qquad \text{for all } \lambda \in \mathbb{R}.$$

The function ψ_X is called the **cumulant-generating function**. It is not hard to see that M_X (or ψ_X) need not exist for all random variables over the whole range of real numbers. For example, if X is exponentially distributed and $\lambda \geq 1$, then